EVIDENTIAL STATISTICS, MODEL IDENTIFICATION, AND SCIENCE

EDITED BY: Mark Louis Taper, Jose Miguel Ponciano, Yukihiko Toquenaga
and Hidetoshi Shimodaira

PUBLISHED IN: Frontiers in Ecology and Evolution and Frontiers in Physiology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-406-0
DOI 10.3389/978-2-88974-406-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

EVIDENTIAL STATISTICS, MODEL IDENTIFICATION, AND SCIENCE

Topic Editors:
Mark Louis Taper, Montana State University System, United States
Jose Miguel Ponciano, University of Florida, United States
Yukihiko Toquenaga, University of Tsukuba, Japan
Hidetoshi Shimodaira, Kyoto University, Japan

Citation: Taper, M. L., Ponciano, J. M., Toquenaga, Y., Shimodaira, H., eds. (2022). Evidential Statistics, Model Identification, and Science. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-406-0

Table of Contents

frontiers | Frontiers in Ecology and Evolution

EDITORIAL
published: 25 April 2022
doi: 10.3389/fevo.2022.883456

Editorial: Evidential Statistics, Model Identification, and Science

Mark L. Taper[1,2]*, José M. Ponciano[2] and Yukihiko Toquenaga[3]

[1] Department of Ecology, Montana State University, Bozeman, MT, United States, [2] Department of Biology, University of Florida, Gainesville, FL, United States, [3] Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

Keywords: replication crisis, scientific inference, model misspecification, inference bias reduction, statistical foundations

Editorial on the Research Topic

Evidential Statistics, Model Identification, and Science

OPEN ACCESS

Edited and reviewed by:
Juergen Pilz,
University of Klagenfurt, Austria

****Correspondence:***
Mark L. Taper
markltaper@gmail.com

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal Frontiers in Ecology and Evolution

Received: *25 February 2022*
Accepted: *11 March 2022*
Published: *25 April 2022*

Citation:
Taper ML, Ponciano JM and Toquenaga Y (2022) Editorial: Evidential Statistics, Model Identification, and Science. Front. Ecol. Evol. 10:883456. doi: 10.3389/fevo.2022.883456

WHY THIS RESEARCH TOPIC

We have undertaken this Research Topic for several reasons: First to promote and disseminate the ideas and techniques of evidential statistics to ecologists and evolutionary biologists so that their research might benefit from the increased clarity that evidential thinking engenders. And, second to encourage statisticians to think how their own work relates to this emerging approach to the fundamental problems of statistics.

HOW TO READ THIS VOLUME

Selecting an optimal order to read the papers of this Research Topic requires decisions on the part of the reader. The papers are not ordered in any developmental fashion, but simply by the order that they were first published. Another difficulty is that there are two target audiences for this Research Topic: First, quantitative scientists, primarily ecologists, and evolutionary biologists, who might wish to apply evidential thinking to their own research; and second, statisticians who might be interested in furthering the technical development of evidential statistics.

Table 1 lays out the primary themes considered in each paper and identifies authorship abbreviations. Those readers who would like to begin with statistical principles, then move to applications, and conclude with more philosophical considerations might read the topic in the order of Dennis et al., Ponciano and Taper, Lele b, Taper et al., Shimodaira and Terada, Markatou and Sofikitou, Ferguson et al., Claeskens et al., Toquenaga and Gagné, Stewart and Blume, Jerde et al., Lele a, Brittan and Bandyopadhyay, Scheiner and Holt. For readers who might prefer to begin with philosophy, then move to application, and finish with technical details, a reasonable order might be: (Brittan and Bandyopadhyay, Scheiner and Holt, Jerde et al., Toquenaga and Gagné, Lele a, Stewart and Blume, Ferguson et al., Claeskens et al., Dennis et al., Ponciano and Taper, Lele b, Taper et al., Markatou and Sofikitou, Shimodaira and Terada).

WHAT IS EVIDENTIAL STATISTICS

Statistics is arguably the most powerful of all scientific instruments. For the last century, statistics has been dominated by two alternative approaches: Error statistics[1] and Bayesian statistics.

[1] By error statistics we mean that subcategory of frequentist statistics that uses error probabilities as the primary inferential quantity including Fisherian significance, null hypothesis significance testing, Neyman-Pearson hypothesis testing, and severe testing. The term classical statistics is sometimes applied to this grouping, but this can be considered a misnomer as Bayesian statistics predates these methods considerably.

Unfortunately, both approaches suffer from technical and philosophical problems (see Taper and Ponciano, 2016 for discussion). These problems make the instrument of statistics like the Hubble telescope before its optics were corrected in 1993: A fantastic tool not living up to its full potential.

We believe that the evidential approach can provide a similar technical correction to statistics. Evidential statistics is a cluster of statistical methods and approaches being developed to meet a set of desiderata or meta-criteria that were selected so as to impose desirable inferential properties on those methods (see Jerde et al., for a list of desiderata).

The central question for evidence is simple: Which of two models of reality is better supported by the data? More technically, evidence is a data-based estimate of the difference of the divergences of each of the distributions implicit in two models to the data distribution resulting from an unknown true generating process (see Lele, 2004; Taper et al.). Several salient features of the evidentialist perspective are immediately obvious: First, evidence is comparative, second, neither model is given a favored status, and third, that a "true" model is not assumed to be in the model set.

These guiding principles allows evidential statistics to draw on and refine elements from error statistics, likelihoodism, Bayesian statistics, information criteria, and robust methods to create an approach that smoothly incorporates model identification, model uncertainty, model comparison, parameter estimation, parameter uncertainty, pre-data control of error, post-data assessment of uncertainty, and post-data strength of evidence into a single coherent framework.

SOME IMPLICATIONS OF EVIDENTIAL STATISTICS FOR SCIENCE

The implications of evidential statistics for science are manifold. For brevity, we focus here on the impact an evidential approach could have on the replication crisis (Pashler and Wagenmakers, 2012). The replication crisis presents a profound challenge to both statistics and science. As more replication of scientific studies is attempted, it is being found that studies tend not to replicate at their nominal rates. This is undermining both trust in statistics by scientists and trust in science by the general population.

Virtually all models are to some degree misspecified (see Taper et al., for a technical definition of "misspecified"). Misspecification in itself is not a bad thing. A true model would be enormously complex and would be neither comprehensible nor estimable. What is dangerous is inference that doesn't acknowledge misspecification. With Neyman-Person Hypothesis testing (NPHT), error rates become distorted when both models are misspecified. Error rates can be less than, equal to, or greater than their nominal rates (Dennis et al.) making nominal rate replication extremely unlikely. Furthermore, under some reasonable model space geometries, a NPHT will select the wrong model with probabilities that go to 1 as sample size increases (Dennis et al.). In contrast, evidential model selection reliability seems in simulation to be estimated unbiasedly (Taper et al.,

TABLE 1 | Articles are listed left to right in publication order.

Thematic concern	Shimodaira and Terada	Scheiner and Holt	Jerde et al.	Dennis et al.	Brittan and Bandyopadhyay	Ponciano and Taper	Ferguson et al.	Claeskens et al.	Markatou and Sofikitou	Stuart and Blume	Lele a	Lele b	Toquenaga and Gagné	Taper et al.
Building evidence functions							•	•	•	•			•	•
Quantifying the uncertainty of evidence	•											•		•
Logic of statistical scientific inference		•	•	•	•	•								•
Application	•		•		•	•		•					•	•
Model space geometry	•			•		•							•	•
Comparative statistical inference				•	•	•					•	•		
Multiple comparisons and combining data	•		•			•								•
Model set misspecification	•			•		•							•	•

Inclusion in a paper of discussion of a topic is indicated by a bullet, •.

2019) and all evidential error rates go to 0 as sample size increases (Dennis et al.).

None of Fisherian significance (FS), null hypothesis significance tests (NHST), or NPHT can produce evidence for the null model (Dennis et al.). This is problematic because often it is the null which of scientific interest. Statisticians teach that "absence of evidence is not evidence of absence," but the need of scientists to say something about the null model forces this warning to be often ignored. In evidential statistics reference and alternative models are always correctly treated symmetrically (Dennis et al., Taper et al., Jerde et al.) for inference, although this does not imply that decision thresholds need to be symmetric.

When scientists, reviewers, and journals recognize that FS, NHST, and NPHT do not produce evidence for the null, a common response is publication bias, the tendency not to publish studies with attained $P < 0.05$ (Franco et al., 2014). This "file drawer problem" creates several biases in the literature. First, of course, is the lack of studies showing evidence for the null. More insidiously, because all tests are stochastic, a number of studies are published falsely showing significant evidence for the alternative (Type I errors). These are not balanced in the literature by the many studies in the file drawer.

The immense pressure on scientists to publish leads many, intentionally or unintentionally, into questionable research practices to avoid the file drawer problem. One of these is "cherry picking," the retroactive selection of data and/or statistics so as to achieve significance (Ioannidis, 2019). Another is HARKing, Hypothesizing After Results are Known (Kerr, 1998). Both have drastic effects on the replication crisis.

Evidential analysis gives scientists statistically correct language (Taper et al.) to speak about strong evidence for the null vs. the alternative, strong evidence for the alternative vs. the null, and evidence that doesn't clearly distinguish between the two models. All of which are of scientific interest. Even results that can't distinguish between models tell us where more data is needed. The results of any well-designed scientific study now have meaning and could potentially be publishable—regardless of significance.

Undertaken in an evidential statistics context, HARKing is a legitimate and even beneficial practice (Taper and Gogan, 2002). The evidence in HARKing has always been clear, although estimation of the uncertainty remained a problem (Taper and Lele, 2004). Bootstrapping of evidential comparisons now improves the understanding of the uncertainty of even HARKed results (Taper and Lele, 2011; Taper et al., 2019, Taper et al.).

COMMENTS ON THE ARTICLES

Shimodaira and Terada

At the heart of ecology is a search to better understand and characterize the relationship between species as well as that of a group of species and their environmental variables. On the other hand, a central topic in evolutionary studies is inferring the ancestral relationships of a set of extant species. In both cases, graph theory has become the theoretical foundation upon which the biological edifices in these two fields are constructed. In ecology, species are thought as nodes in a diagram and the relationships between species are represented as edges uniting any two nodes. In evolution, a phylogenetic binary tree is a diagram representing the evolutionary relationships among a set of extant species, which are shown as the tips (leaves) of the tree. Each interior node in the tree connects with three other nodes: two descendants and one ancestor.

The binary phylogenetic trees are called bifurcating trees because there are two branches leading out from each interior node. Proceeding from the present-day species of interest backwards in time under this binary framework eventually leads to a common ancestor, the root of the tree. In that context, one particular "tree topology" is one specific construction of the possible set of relationships among the species of interest and represents a single hypothesis about the ancestral relationships between these species, all the way back to their most recent common ancestor. How many such hypotheses can one posit with n species? With two species the answer is one, with three species the answer is three, with four it's fifteen, with five it's one hundred and five and in general, with n species it's $(2n-3)!/\left(2^{n-2}(n-2)!\right)$. For example, for six species, the number considered by Shimodaira and Terada one could posit 945 such trees.

In such setting, it quickly becomes obvious that good treatments of the statistical problems of multi-model selection and multiple hypotheses testing are key to making any progress in this area. Previously, the leading approach to deal with the problem of selecting among these models (hypotheses) the best representation of reality used NHST. This body of work was started by Kishino and Hasegawa (1989), and continued by Shimodaira (1998, 2002) and Shimodaira and Hasegawa (1999). Shimodaira and Terada now goes one step further and provides a novel methodology of shifting the phylogenetics question away from: "is a newly estimated tree topology significantly similar to the unknown, true species topology?" and instead ask: "from this set of models, which tree topology and group of models are significantly closer, in a KL distance spatial configuration sense, to the unknown, true topology?" To do so, Shimodaira and Terada estimate a spatial configuration of models in a three-dimensional model space, a geometrical construction very much like that of Ponciano and Taper. However, these two approaches differ in that while Shimodaira and Terada rely on a shifting combination of NHSTs and NPHTs for inference, Ponciano and Taper use a non-parametric self-entropy estimation to construct a model projection in a model space that can be used as the point to do a science-based examination of critical model attributes that allow a model to get closer to the generating process. The methodology of Ponciano and Taper is geared toward being coupled with uncertainty estimation and examining the strength of the evidence for a given model using the approach suggested by Taper et al. One should note that although (Shimodaira and Terada) are testing alternative hypotheses ($\mathrm{H}_0{:}\mu \in \mathrm{R}$ versus $\mathrm{H}_1{:}\mu \in R^c$), the tests are not standard NP tests. Truth does not lie in either hypothesis, but instead is being projected onto the manifold $R \cup R^c$. Further, the pseudo data being used to generate the distribution of the test statistic does not come from H_0, but is generated by a non-parametric bootstrap. Thus, the difference between the inference in Shimodaira and Terada and Taper et al. may be little more than the statistics they choose to present.

Scheiner and Holt

This paper takes the readers out of the weeds and forces them to look simultaneously at the trees and the forest. Deeply informed by both the history and the philosophy of science, the manuscript points out that evidential statistics formally only deals with the relationships among models and data; Scheiner and Holt then ask how evidential statistics can inform either the generation or the support for general and constitutive theories. Clearly it can because Peirce's abduction (Peirce, 1974) can be thought of as a conceptual adequacy measure for models, hypotheses, or theories, while modern abduction, i.e., inference to the best explanation (Haig, 2009) can be thought of as conceptual evidence for the same.

In an analogy to biological evolutionary theory, Scheiner and Holt discuss how model selection, an evidential process, can act as a selective force to winnow the models included in constitutive theories. Scheiner and Holt further suggest that pattern matching as well as Whewell's consilience and coherence (Forster and Wolfe, 1999) might possibly be utilized in formal procedures for quantifying the evidence supporting one theory over another.

Despite the excellence of this article, Scheiner and Holt do sin against science in suggesting that sometimes statistics is not necessary[2]. They claim for instance that if something never occurs then no statistics is necessary. To which a statistician would query, "never occurs in how many trials?" The evidential impact of something never occurring is very different in experiments of 1 trial, 4 trials, or 8 trials (see Jerde et al.). Because they are writing as theoreticians, Scheiner and Holt's sin is only venal. For theoreticians, statistics and even data, are always optional. The job of theoretical science is to construct alternative internally consistent possible worlds. The job of empirical science is to determine which of those possible worlds best describes the real world—and for that, statistics is always needed.

Jerde, Kraskura, Eliason, Csik, Stier, and Taper

Jerde et al. describe the motivation for, and the logic of, scientific inference using evidential statistics and demonstrate the utility of the evidential approach by tackling a long-standing controversial question in ecological physiology: How does standard metabolic rate (SMR) scale (intra-specifically) with individual body mass, and is this scaling similar among species? For fish, theoretical scaling rates of 0.67, 0.75, and 1.00 have been proposed. Empirical estimates of scaling coefficients vary tremendously among studies and generally all have large uncertainties leaving the theoretical question unprobed. Jerde et al. curate a large data set composed of a total of 1,456 observations in 55 separate trials on 12 species, all using current state of the art techniques for measuring SMR. The use of linear mixed effect models allowed (Jerde et al.) to combine all of these trials for inference.

Four suites of four models using random and fixed effects carefully explore the impacts of species, trial (within species), and temperature on the scaling of SMR with body mass. Model families were evaluated using the Schwarz information criterion (SIC, also known as the BIC). The SIC is a consistent criterion and the comparison of SIC values is an evidential procedure. Within and between model suites, evidence for specific values of the scaling coefficient were compared using profile ΔSIC curves. A ΔSIC value comparing two models >7 indicates strong evidence for the model with lower SIC.

Two model suites with a free parameter estimate of the metabolic scaling, separated themselves only by a ΔSIC of 1.5, were strongly differentiated from all others. Both had fixed effects for temperature and random effects (intercepts) for species. The best model had the log(weight) slope vary randomly across species (with modest variation), while the second-best model had a common slope over all species. In the best model the ML estimate for the mean scaling coefficient is 0.89 with a strong evidence profile ΔSIC interval spanning 0.82–0.99.

The evidence strongly indicates that none of the *a priori* theoretical scaling coefficients describe the scaling behavior in real fish.

Dennis, Ponciano, Taper, and Lele

Mathematics, and in particular probability, have long been intertwined with biology. The theoretician J. E. Cohen adroitly summarized the transcendence of the synergy between these fields with his essay "Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better" (Cohen, 2004). Key to the success of this interaction between these fields is the recognition that fundamental hypotheses in biology can be translated using the languages of mathematics, probability, and statistics into propositions than can be clearly probed. The increase in possibilities with such synergism is so dramatic that in some cases, it's as if a new portal to a field of scientific inquiry becomes available. Yet, becoming enamored with model construction and the phrasing of novel explanations of biological phenomena can sometimes obscure the analyst's vision and the realization that by its very human nature, mathematical models are limited constructs of biological processes. Mathematical models are indeed misspecifications of natural processes. Understanding the effects of model misspecification in our scientific inquiry should be paramount. This is the focus of Dennis et al. These authors assess analytically and numerically the performance of Neyman-Person Hypothesis testing (NPHT), Fisher significance testing (NHST), information criteria, and evidential statistics under model misspecification.

As mentioned above, evidential statistics seeks to quantify the strength of the evidence in the data for a reference model relative to another model. This goal is achieved through an evidence function, which is simply a statistic for comparing two models. Dennis et al.'s evidence function of choice was Schwarz Information Criterion, or SIC (Schwarz, 1978). The salient property of this and all evidence functions is that their associated probabilities of making a wrong model choice approach 0 as sample size increases. These probabilities, analogous to Type I and II errors in the Neyman-Pearson Hypothesis Testing

[2] In prepublication conversations on this point, we told the authors that they could say whatever they wanted in their paper, but that the final word would belong to the editors.

(NPHT) framework are in fact pre-data error rates. Royall (2000) showed that these probabilities measure the chances of obtaining weak misleading-evidence as well as strong misleading-evidence. Dennis et al. shows that in a context where both models are in fact mathematical misspecifications of reality, making the wrong model choice refers to deeming as best a model that is not the closest to the true generating process model. By the same token, misleading-evidence simply corresponds to obtaining observations that either weakly or strongly support a model other than the one that is the closest to the data-generating process.

Unlike the classic NPHT and Bayesian approaches, the Evidential Statistics paradigm provides sound guidelines to evaluate inferential errors when none of the proposed statistical models are a perfect representation of the natural, data-generating process. The NPHT framework depends critically on either the Null or the Alternative hypotheses being a perfect representation of the data generating mechanism and then fixes the Type I error probability irrespectively of sample size and thus problematically assesses the evidence *against* the null hypothesis and remains silent with respect to the evidence *for* the null hypothesis. The asymmetry of the NPHT error structure leads to difficulties in interpretation of hypotheses tests. The decision to pick an alternative model over a null hypothesis in and of itself is not controversial as it has some intuitively desirable statistical properties: for example, the probability to reject the null hypothesis given that the alternative is true converges to 0 as sample size increases. However, the probability of erroneously choosing the alternative when the null is true remains stuck at the chosen level alpha regardless of how large a sample size is collected. Matters get more complicated when it is considered that the original Neyman-Pearson theorem assumes that the data was generated under one of the two models but provides no guidance whatsoever in the event of model misspecification, a scenario commonly encountered in science. The fact that in scientific practice model comparison rarely stops at two models further muddying the interpretation of experimental results using the NPHT. To be fair, overconfidence in model selection procedures also results when the model misspecification is ignored in Bayesian Statistics (Yang and Zhu, 2018).

The evidential approach proposes fixing cutoff values for the evidence statistic, not the error probabilities. Under this concept of evidence, the value of a statistic like the likelihood ratio is evidence, not an error rate that is pre-set. Then, the evidential error probabilities both converge to 0 as sample size grows large. Finally, under this evidential statistics approach, the conclusion structure of say, a comparison between two models H_1 and H_2 has a trichotomy of outcomes: (*i*) strong evidence for H_1, (*ii*) weak or inconclusive evidence, and (*iii*) strong evidence for H_2.

Some, not all, information criteria commonly used for model selection are evidence functions. While the AIC only penalizes the likelihood function using the number of parameters, the SIC is also scaled by the sample size. As a result, as sample size increases, the error in deeming a model as "best" using the SIC statistics becomes vanishingly small. Dennis et al. show that this desirable property, called "Information consistency" is lacking in the AIC. Inconsistent criteria, such as the AIC, tend to overfit at all sample sizes. Hence, the AIC is not an evidence function because it is not information consistent.

Although all paradigms of statistical science (NPHT, Bayesian statistics, Evidential Statistics) have flaws (reviewed in Lele a, b), the Evidential Statistics paradigm possesses more desirable characteristics for the quantification of uncertainty and ultimately, for the design of inferential statements about the models' proximity to the true, generating process.

Brittan and Bandyopadhyay

Written by a pair of philosophers of science, Brittan and Bandyopadhyay provides a good entry into the Research Topic. Despite maintaining a high level of intellectual rigor, Brittan and Bandyopadhyay avoids getting bogged down in technical statistical detail. The authors review the logical structures for scientific evidence: Hypothetico-deductive testing, Popperian falsification and corroboration, Fisherian significance, Neyman-Pearson hypothesis testing, the severe testing of Mayo, Bayesian confirmation, and statistical evidence.

The authors are equal opportunity balloon poppers pointing out the limitation of all methodological approaches. Brittan and Bandyopadhyay focus on the strengths, weaknesses, and complementarity of statistical evidence and Bayesian confirmation. Contra the prevailing scientific mythos, Brittan and Bandyopadhyay demonstrate that Bayesian inference is "irreducibly personal." Bayesian methods do a good job of quantifying personal beliefs, and thus of informing personal decisions. Echoing Lele a; Brittan and Bandyopadhyay contend that non-informative priors are not objective and suffer from a variety of other problems. In contrast, statistical evidence does objectively quantify the relative support in data for specified pairs of models even though the models put forth for comparison may be generated subjectively.

Science is plagued by a suite of cognitive biases. Being aware of them can mitigate their impact. The authors note that each methodology works best to answer fairly narrow but different questions. Greater methodological self-consciousness on the part of scientists to match their choice of statistical approaches to match their scientific questions would promote scientific progress.

Brittan and Bandyopadhyay close on the same hopeful note and metaphor as do Scheiner and Holt. Despite the undeniable subjectivity of individual scientists, Science itself may achieve a "Darwinian Objectivity" when the mutational force of subjective scientific creativity is filtered by objective evidential model selection.

Ponciano and Taper

Information criteria have had a profound impact on modern science because they allow researchers to overcome the inadequacies of NPHT and tackle the multi-model selection process. Although model selection via information criteria gives the analyst an estimate of which probabilistic approximating models are closest to the generating process, information criterion comparison does not solve the problem of knowing how good the best model is. Indeed, the absolute distance to the generating process is not estimated through this process.

This caveat is all the more important when it is considered that in science, models are commonly misspecified. In this work, the authors resolve this shortcoming by designing a methodology to estimate a geometric representation of all the models under consideration along with the generating process. Such representation is a projection of all the models at hand into a two or three-dimensional space. As well, the location of the generating process in this representation is fully estimated. To estimate this model projection, the authors examined five key insights from Hirotsugu Akaike's original work. These insights reveal the deep, yet easy to grasp, geometrical nature of Akaike's formulation of the AIC. Ponciano and Taper extend Akaike's geometrical interpretation and propose visualizing all models at hand into a reduced space. This reduced space representation applies ordination techniques to the models themselves so that the analyst may see and estimate the divergence between each model and every other model including the generating process itself.

Ponciano and Taper's solution starts from the observation that while standard information criterion analysis considers only the divergences of each model from the generating process, the divergences amongst all approximating models, typically ignored, are indeed estimable. As a test bed for their ideas, the authors consider two ecological scenarios, one of them involving an individual-based model simulation framework that generates data to which different abundance models can be fitted and the second one involving structural equation models.

The authors also compare their approach to model averaging and show that model projection is not as sensitive as model averaging to the composition of the set of candidate models being investigated. Model averaging artificially favors redundance of model specification because the more models are developed in any given region of model space, the more heavily this particular region gets weighted. Furthermore, examining the resulting model space configuration can lead to an in-depth analysis of what are the model attributes that change from one model to the next that make it so that a model will get closer and closer to the generating process. This examination is the first step to explore models outside the bounds of the available model set, whereas by using model averaging, by definition, the analyst cannot do so.

Uncertainties around the estimation of model space estimation are yet not fully worked, but Taper et al. offers a first, non-parametric bootstrap approach to begin examining such question. Model projection methodology should be the starting point to do a science-based examination of critical model attributes that allow a model to get closer to the generating process (see also Toquenaga and Gagné). Finally, although Ponciano and Taper use the Kullback-Leibler, KL, divergence as the fundamental distance measure, the model projections methodology could be extended or adapted to any other metric.

Ferguson, Taper, Zenil-Ferguson, Jasieniuk, and Maxwell

There are a vast number of information criteria. Academic arguments about which is best are intense and often vitriolic. Ferguson et al. indicates that these arguments may be a tempest in teapot.

Seeking to improve model identification techniques for complex models with inter-dependent parameters, the authors modify Bozdogan's Information Complexity Criteria, ICC, to make them consistent and invariant to more kinds of transformations. To validate their suggested new criteria, Ferguson et al. perform a vast array of performance comparisons. Twenty-five information criteria are investigated: Two classical efficient criteria (AIC and AICc), two classical consistent criteria (BIC and BIC*), three forms of Bozdogan's ICC, and 18 new modifications of the ICC. All of these criteria were compared for their ability in attaining three different model selection goals: Selecting models with minimum prediction error, identifying the form of the generating model, and estimating the KL divergence to the generating process. All of this is done under 3 different classes of generating and approximating models, 3 different sample sizes, 3 different levels of process error, and 3 different levels of collinearity.

Ferguson et al. recommend one of their combined forms [BIC+2CvE(Ψ)] as achieving all measures of quality well under a broad range of modeling frameworks and having the theoretical advantage of being both scale invariant and consistent. However, it is important to note that No IC was best for *any* goal over all conditions and that All IC performed generally well for all goals.

Two important lessons should be taken from Ferguson et al.: First, much more attention needs to be paid to the uncertainty of model identification. And second, for these goals to be achieved sample sizes need to be larger in all model classes than is generally the case in ecology.

Claeskens, Cunen, and Hjort

Perhaps the most used statistical tools by ecologists are abundance count models. Simply counting the number of individuals of every species observed in a particular community is the point of entry to deeper studies aiming at understanding the generation and maintenance of organisms' diversity. Profound questions examining the processes driving ecological stability, resilience, resistance, invasion, and persistence all begin with being able to accurately ascertain organisms' abundances. In our joint decades of teaching and mentoring, time and again count models keep coming back as some of the main instruments of statistical inference sustaining masters' theses and PhD dissertations in biology, wildlife ecology and conservation. Ecologists are typically not only interested in estimating one or the other model parameters leading to particular predictions, but often see parameter estimation as the by-product of what they are typically after, which is understanding which hypothesized model components better represent the underlying natural processes generating the count data at hand.

Claeskens et al. propose and further elaborate on a methodology that may revolutionize the reaches of an ecology-driven statistical analyses and in particular, multi-model selection for models of count data. The main idea of the Focused Information Criterion (FIC) approach is to provide a model selection framework where the comparison and the ranking is formally defined according to the scientific quest at hand. Recognizing that different scientific teams might ask different focused questions of the same data and list of candidate models, Claeskens et al. design a methodology to focus the model selection process using different functions of the parameters of interest. When mainstream model selection tools are used in ecology and in a given scenario a model is chosen as the best model, practitioners are often left wondering why, in a specific scientific sense, such model is indeed the best model. FIC offers a theoretically sound methodology to obtain better, more precise estimates of a quantity of interest. For count models, such quantity is often the probability of a rare event occurring. As arbitrary or stale as it may sound at first, understanding and estimating accurately rare events in ecology has always been at the center of key explanations of diversity. Rarity, or "rare counts," have been for a long time (e.g., Patil and Taillie, 1982) hypothesized to be a critical component of explanations of how hyper-diverse communities can be maintained. Such was also the conclusion of one of the most recent and cited explanations of the maintenance of diversity in tropical forests published by Levi et al. (2019). As it turns out, the Focused Information Criterion of Claeskens et al., which seeks to minimize the bias and the variance of a quantity of interest, works particularly well for estimating the probability of rare events. In line with the rarity comments above, Claeskens et al. show as examples a situation where the focus of the inference is estimation of the probability of observing counts of a species above an arbitrary number. Importantly, the authors show how other information criteria like the BIC, although they may address the problem of determining which model is the closest to the true data generating mechanism, may not point toward the models that do the best job at estimating for instance, the tail of a distribution of counts. By allowing for a flexible specification of different foci of interest, Claeskens et al. provide a welcome addition to the toolbox of the evidentialist. This tool is not only conceptual but is crystallized in a practical, easy to use library for R users, the "fic" library.

Markatou and Sofiktou

Most of the papers summarized so far share a key point: a reliance on the Kullback-Leibler divergence as the main instrument to develop and exemplify the theory and practice of Evidential Statistics. A natural reaction of any statistician to such heavy reliance on a single metric should be to ponder what would happen if different metrics or distances are used. Can the desiderata of evidential statistics be kept under different measures of divergence between the generating process and any approximating model, or amongst models themselves? Would the theoretical and asymptotic warrants of evidential statistics hold under different distance measures? How can statisticians visualize the strength of evidence under different measures? How does a measure of strong evidence using the KL divergence translates to other scales of divergence? These and other questions are approached using philosophical and rigorous statistical techniques in the contribution by Markatou and Sofikitou. Importantly, Markatou and Sofikitou's contribution builds upon the pioneering concepts of model adequacy by Lindsay (2004) and evidence functions by Lele (2004). Notably, the authors propose an explanatory analysis tool called a standardized distance ratio plot that can be used to visualize the strength of evidence provided for or against hypotheses of interest using different divergence measures. Hence, this paper represents itself growth in the field and marks a clear path for future research. Indeed, of all the contributions in this special issue, this one is perhaps the one topic that is most ripe for further research and study. An open direction that seems promising is shining light on the behavior of different statistical divergence measures under model misspecification. Whenever we give seminars in statistics departments about evidential statistics, the question of usage of other divergence measures invariably comes up. We therefore encourage both, a close reading of this paper and thinking about building extensions to these results using Markatou and Sofikitou's work as the foundation.

Stuart and Blume

New statistical approaches often face resistance from empirical scientists. It can help acceptance if a new technique seems familiar. Stuart and Blume cleverly disguise an evidential procedure with the face of a *p*-value, something that virtually every working scientist is familiar with. It does look like a *p*-value in that the statistic can take on values of 0, 1, and everything in between. Stuart and Blume even strengthen the familiarity by calling it a SGPV or second-generation *p*-value.

Of course, a SGPV is not a *p*-value, it is not even a probability. The SGPV is better than a *p*-value. The question of interest is whether an unknown, but estimated, parameter is in an interval null or is outside of the interval null. A *p*-value or a null hypothesis significance test (NHST) can indicate that the parameter is likely outside the null, but neither can give you support that it is inside the null. Conversely, an equivalence test can give you support for the parameter being inside the interval but not for being outside the interval.

Evidence like, the procedure divides the range of possible value for the SGPV into 3 regions: The point SGPV = 0, which indicates strong evidence the parameter is in the interval null. The point SGPV = 1, which indicates strong evidence the parameter is not in the null. And, the region of all values in between, which indicate that the data are consistent with both hypotheses and which way the evidence is tipping.

Stuart and Blume also demonstrate another important evidential property. The SGPV is consistent; the probability of misleading evidence goes to 0 as sample size increases.

The SGPV is very flexible and can be applied retroactively to any scientific literature in which a statistical interval is published. Stuart and Blume claim that SGPV is applicable to any type of interval confidence, support, or credible. The authors spend the bulk of the paper demonstrating good statistical properties for the SGPV under a wide range of circumstances.

Lele a

It is undeniably true that State-Space Models (SMMs) or more generally, hierarchical statistical models, nowadays occupy a central role in ecology and evolution. SMMs are used to study the population dynamics of animals with complex life histories, to estimate abundances under detection limitations and heterogeneity (among individuals, across space, and in time). Entire statistical ecology books for graduate students and researchers alike with titles around "hierarchical models in ecology" now fill the electronic and physical bookshelves of modern ecologists and academicians. As well, social media with short instructionals, blogposts and even tweets by the authors of these books are consumed voraciously by graduate students needing to solve complex problems in the face of non-standard datasets. Software authors in turn, face the challenge of putting out for consumption accessible programs that can weather usage by anybody interested in applying a given hierarchical model. Over recent years, this high demand for accessible solutions to complex problems has facilitated the establishment of uncritical use of modern statistical machinery.

Lele a approaches the consequences of such uncritical use head-on by clearly illustrating with real-life examples the predicaments brought about by using non-informative Bayesian analysis. Indeed, non-informative Bayesian analysis tends to be nowadays the default setting under which complex statistical models in ecology are fitted. In the name of pragmatism, it is often argued that in modern, extensive big data sets the sample size is so large that the likelihood information "swamps" any prior effect and that effectively, the data will "speak for itself."

Lele a carefully delineates the flaws in such reasoning and vividly details how and why wildlife management decisions can vastly suffer from such uncritical use of Bayesian techniques. In particular, he shows that because of the lack of parameterization invariance of non-informative Bayesian Analysis, all subjective Bayesian inferences can be disguised as "objective," non-informative Bayesian inferences. Furthermore, cryptic biases can be introduced in the resulting analyses because the induced priors on functions of parameters are not non-informative.

Three other serious flaws are then discussed besides these two. However, even if the author had presented only these two problems, practitioners, ecologists and wildlife managers should take note, because if the results of an uncritical non-informative Bayesian analysis is subject to unstated and unqualified biases, it may be easily challenged in the legislature and in the court of law. For completeness, professor Lele emphasizes that hierarchical models can be and are analyzed using the likelihood and frequentist methods. That is, ***any*** Bayesian analysis can be transformed to a likelihood analysis by data cloning.

Lele b

Uncertainty is a fundamental part of any inference, but the depth of its complexity is often not adequately appreciated. This paper, Lele b, gives a surprisingly readable review of many of the issues involved with statistical uncertainty. Lele b begins with a short list, culled from the literature, of desirable features for uncertainty quantification procedures: (1) transformation invariance, (2) uncertainty measure reflect data informativeness, (3) ascertainability, and (4) diagnostic potential.

The first, transformation invariance, implies that the probability of an event occurring or not occurring is a reasonable measure of uncertainty. This of course requires understanding what probability is and the paper next discusses the two major definitions of probability used by statisticians and scientists alike: aleatory or frequency-based probability and epistemic or belief-based probability.

For adherents of frequentist statistics, data (i.e., data sets) are random realizations from a stochastic generating process. Consequently, estimates of parameters inherit stochasticity from the generating process through the stochasticity of data sets. The distribution of parameter estimates over an infinite number of random data sets is called the true sampling distribution of the parameter. One can estimate a parameters sampling distribution by bootstrap or analytic approximation. The estimated sampling distribution contains a great deal of information about the uncertainty of the procedure. Much of this uncertainty is captured by confidence intervals. While arguing for the utility of confidence intervals, Lele b points out they are often misinterpreted.

Lele b points out that the target of a confidence interval is to cover the true parameter, not to cover the parameter estimated in another experiment. Another common way that confidence intervals are misinterpreted is by failing to distinguish between unconditional/pre-data and conditional/post-data intervals. Both kinds of intervals are commonly used in the scientific literature. In separate sections Lele b returns to the questions of interval construction and interpretation from Bayesian and evidentialist perspectives.

As pointed out by Brittan and Bandyopadhyay "any adequate ('reliable') hypothesis must be both explanatory and predictive." It is only through the verification of predictions that the ascertainment of models or hypotheses is possible. Lele b takes this very seriously reviewing the representation of prediction uncertainty in all three inferential paradigms. Further, a new flexible approach to the calculation of an evidential predictive density is suggested and its advantages, both demonstrated and potential, are discussed.

The paper concludes by rehearsing the key features, strengths, and weaknesses of the characterization of uncertainty in the three paradigms in the light of the four desiderata. None is perfect, but overall, the evidentialist most closely conforms. All three paradigms require scientists to specify their models and whether inference should conditional or unconditional. Bayesian inference further requires the specification of priors, while evidence requires the specification of an evidence function. The last thing any reader wants to hear is that the quality of their scientific inference depends critically on the active choices they make—regardless of their statistical paradigm. Nevertheless, this is precisely the last thing that Lele b says.

Toquenaga and Gagné

Genetic sequencing is becoming an increasingly important tool in ecological and evolutionary studies. This trend has been accelerated by the new techniques of "next-generation sequencing," NGS. These sequencing procedures work by digesting a genetic sequence into many small fragments (called reads), sequencing the fragments, and then inferring the original sequence computationally. This is like the spy novel trope of pasting a shredded letter back together.

With the scientific opportunities, come many statistical challenges. There are many programs that make these calculations. Unfortunately, they don't agree—with each other and because many of the programs involve stochastic searches, even between multiple runs of the same program. Toquenaga and Gagné, use evidential principles to develop methods to choose among the many putative sequences offered by an array of sequencing software, to assess how good the proposed sequences are, and even to improve them.

The thinking in Toquenaga and Gagné is as follows: If multiple algorithms produce multiple sequences, each must be a model of the true sequence. If an appropriate function for measuring the divergence between these sequence models can be found, then the model projections in model space methods of Ponciano and Taper can be used to understand the relationships among the proposed models and even to a true sequence. The Levenshtein edit distance (Levenshtein, 1966), as a measure of the minimum number of changes needed to equate two sequences from finite alphabets, offers itself as an appropriate divergence.

Toquenaga and Gagné test this proposition by taking a known genetic sequence and randomly breaking it into a number of fragments (with potential overlap). The number and distribution of fragment sizes are set to mimic typical digestion results. In their test case, Toquenaga and Gagné are able to construct, using non-metric dimensional scaling, a two-dimensional map of the sequence estimates produced by the various sequencing programs compared by the authors. Their map correctly identifies the best-proposed sequence.

In this test case, one of the programs is able to correctly reconstruct the true sequence. However, such a felicitous occurrence may not be general. Usefully, Toquenaga and Gagné propose an approach that can suggest sequences likely to improve on the set of mistaken sequences. They do this by proposing new sequence models which are consensus sequences of existing models and seeing where they fit into the map.

Toquenaga and Gagné confirm their method with a parametric bootstrap based on a specified true sequence. Implicit in this is the potential to use similar bootstrapping to assess the uncertainty in sequence construction.

Taper, Lele, Ponciano, Dennis, and Jerde

Taper et al. develops themes from two other papers in this Research Topic. Dennis et al. show that in the presence of model misspecification Royall's universal bound on the strength of misleading evidence does not hold. Lele b reminds us that statical uncertainty comes in two forms: global/unconditional and local/conditional.

To Royall's regions of weak and strong evidence (Royall, 2000) the authors intersperse a third category, that of prognostic evidence. This is evidence not so weak as to be dismissed nor so strong as to be considered overwhelming. Thus, while evidence is itself continuous, useful descriptive categories for considering evidence are constructed.

Taper et al. show that even in the presence of model misspecification the uncertainty in model identification can be quantified in the form of non-parametric bootstrap confidence intervals on evidence. This decouples evidence and its uncertainty and allows scientists to consider both. The authors consider evidence (either prognostic or strong) for one model over another to be "secure" if the lower 5% confidence limit on the evidence is above the preset prognostic boundary, k_p.

To demonstrate the utility of this approach, Taper et al. make a detailed reanalysis of model selection in Grace and Keeley's (2006) classic structural equation modeling of post-fire diversity recovery in California shrublands. The use of evidence confidence intervals develops a much more nuanced understanding of which model components are likely to be robust and which are equivocal.

Technically, Taper et al. use an improved version of the EIC (see Kitagawa and Konishi, 2010). The improvements include: (1) bootstrapping of the ΔSIC rather than individual likelihoods to incorporate the effects of misspecification geometry. And (2) identification of components of EIC that correspond to global and local inference.

The paper finishes with an extended discussion of the interpretation of global and local inference in science.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and writing of this editorial. All authors contributed to the article and approved the submitted version.

FUNDING

JP was partially supported by two grants from the US National Institute of Health grant number to University of Florida NIH R01 GM103604 to JP and NIH 1R01GM117617 to Profs. Jason K. Blackburn and JP. JP was also partially supported by NSF Award Number 2052372 to University of Florida (PI: JP, Project title: Collaborative Research: Scaling Properties of Ecological Variation in Complex Dynamical Systems). YT was partly supported by JSPS KAKENHI Grant Numbers 17H04612 and 18K06410.

ACKNOWLEDGMENTS

We thank Brian Dennis and Subhash R. Lele for critically reading a draft of this editorial. We are grateful to all of the authors who contributed to this Research Topic and to the editors and reviewers who worked hard to clarify the presentations. Our understanding of evidence has been sharpened by years of discussions with Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Brian Dennis, Jake M. Ferguson, Christopher L. Jerde, Subhash R. Lele, and Kenichiro Shimatani.

REFERENCES

Cohen, J. E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol.* 2, 2017–2023. doi: 10.1371/journal.pbio.0020439

Forster, M. R., and Wolfe, A. (1999). *Conceptual Innovation and the Relational Nature of Evidence: The Whewell-Mill Debate*. Электронный ресур1. Режим доступа. Available online at: https://www.academia.edu/39209519/Conceptual_Innovation_and_the_Relational_Nature_of_Evidence_The_Whewell_Mill_Debate

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science* 345, 1502–1505. doi: 10.1126/science.1255484

Grace, J. B., and Keeley, J. E. (2006). A structural equation model analysis of postfire plant diversity in California shrublands. *Ecol. Appl.* 16, 503–514. doi: 10.1890/1051-0761(2006)016[0503:ASEMAO]2.0.CO;2

Haig, B. D. (2009). Inference to the best explanation: a neglected approach to theory appraisal in psychology. *Am. J. Psychol.* 122, 219–234.

Ioannidis, J. P. A. (2019). What have we (Not) learnt from millions of scientific papers with p values? *Am. Stat.* 73, 20–25. doi: 10.1080/00031305.2018.1447512

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2, 196–217. doi: 10.1207/s15327957pspr0203_4

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from dna-sequence data, and the branching order in Hominoidea. *J. Mole. Evolut*. 29, 170–179. doi: 10.1007/BF02100115

Kitagawa, G., and Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Ann. Inst. Stat. Math.* 62, 209–234. doi: 10.1007/s10463-009-0237-1

Lele, S. R. (2004). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper, and S. R. Lele (Chicago, IL: The University of Chicago Press), 191–216.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Doklady*. 10, 707–710.

Levi, T., Barfield, M., Barrantes, S., Sullivan, C., Holt, R. D., and Terborgh, J. (2019). Tropical forests can maintain hyperdiversity because of enemies. *Proc. Natl. Acad. Sci. U.S.A.* 116, 581–586. doi: 10.1073/pnas.1813211116

Lindsay, B. G. (2004). "Statistical distances as loss functions in assessing model adequacy," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations* eds M. L. Taper, and S. R. Lele (Chicago, IL: The University of Chicago Press), 439–488. doi: 10.7208/chicago/9780226789583.003.0014

Pashler, H., and Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253

Patil, G. P., and Taillie, C. (1982). Diversity as a concept and its measurement. *J. Am. Stat. Assoc.* 77, 548–561. doi: 10.1080/01621459.1982.10477845

Peirce, C. S. (1974). "Harvard lectures on pragmatism 1903," in *The Collected Papers of Charles Sanders Peirce*, Vol. 5, eds C. Hartshorne, and P. Weiss (Cambridge: Harvard University Press), 188–189.

Royall, R. M. (2000). On the probability of observing misleading statistical evidence. *J. Am. Stat. Assoc.* 95, 760–780. doi: 10.1080/01621459.2000.10474264

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Inst. Statist. Math.* 50, 1–13. Available online at: https://www.ism.ac.jp/editsec/aism/pdf/050_1_0001.pdf

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508. doi: 10.1080/10635150290069913

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201

Taper, M. L., and Gogan, P. J. P. (2002). The Northern Yellowstone elk: density dependence and climatic conditions. *J. Wildlife Manage.* 66, 106–122. doi: 10.2307/3802877

Taper, M. L., Lele, S. R., Ponciano, J.-M., and Dennis, B. (2019). Assessing the uncertainty in statistical evidence with the possibility of model misspecification using a non-parametric bootstrap. Available online at: https://arxiv.org/abs/1911.06421v1

Taper, M. L., and Lele. S. R. (2004). "The nature of scientific evidence: A forward-looking synthesis," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 527–551.

Taper, M. L., and Lele. S. R. (2011). "Evidence, evidence functions, and error probabilities," in Philosophy of Statistics, eds P. S. Bandyopadhyay and M. Forster (Elsevier), 513–532.

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

Yang, Z. H., and Zhu, T. Q. (2018). Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1854–1859. doi: 10.1073/pnas.1712673115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Ecology and Evolution

METHODS
published: 24 May 2019
doi: 10.3389/fevo.2019.00174

Selective Inference for Testing Trees and Edges in Phylogenetics

Hidetoshi Shimodaira[1,2]* *and Yoshikazu Terada*[2,3]

[1] *Graduate School of Informatics, Kyoto University, Kyoto, Japan,* [2] *Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan,* [3] *Graduate School of Engineering Science, Osaka University, Osaka, Japan*

Selective inference is considered for testing trees and edges in phylogenetic tree selection from molecular sequences. This improves the previously proposed approximately unbiased test by adjusting the selection bias when testing many trees and edges at the same time. The newly proposed selective inference p-value is useful for testing selected edges to claim that they are significantly supported if $p > 1-\alpha$, whereas the non-selective p-value is still useful for testing candidate trees to claim that they are rejected if $p < \alpha$. The selective p-value controls the type-I error conditioned on the selection event, whereas the non-selective p-value controls it unconditionally. The selective and non-selective approximately unbiased p-values are computed from two geometric quantities called signed distance and mean curvature of the region representing tree or edge of interest in the space of probability distributions. These two geometric quantities are estimated by fitting a model of scaling-law to the non-parametric multiscale bootstrap probabilities. Our general method is applicable to a wider class of problems; phylogenetic tree selection is an example of model selection, and it is interpreted as the variable selection of multiple regression, where each edge corresponds to each predictor. Our method is illustrated in a previously controversial phylogenetic analysis of human, rabbit and mouse.

Keywords: statistical hypothesis testing, multiple testing, selection bias, model selection, Akaike information criterion, bootstrap resampling, hierarchical clustering, variable selection

Edited by:
Rodney L. Honeycutt,
Pepperdine University, United States

Reviewed by:
Carlos Garcia-Verdugo,
Jardín Botánico Canario "Viera y Clavijo", Spain
Michael S. Brewer,
University of California, Berkeley, United States

****Correspondence:***
Hidetoshi Shimodaira
shimo@i.kyoto-u.ac.jp

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *31 January 2019*
Accepted: *30 April 2019*
Published: *24 May 2019*

Citation:
Shimodaira H and Terada Y (2019) Selective Inference for Testing Trees and Edges in Phylogenetics.
Front. Ecol. Evol. 7:174.
doi: 10.3389/fevo.2019.00174

1. INTRODUCTION

A phylogenetic tree is a diagram showing evolutionary relationships among species, and a tree topology is a graph obtained from the phylogentic tree by ignoring the branch lengths. The primary objective of any phylogenetic analysis is to approximate a topology that reflects the evolution history of the group of organisms under study. Branches of the tree are also referred to as edges in the tree topology. Given a rooted tree topology, or a unrooted tree topology with an outgroup, each edge splits the tree so that it defines the clade consisting of all the descendant species. Therefore, edges in a tree topology represent clades of species. Because the phylogenetic tree is commonly inferred from molecular sequences, it is crucial to assess the statistical confidence of the inference. In phylogenetics, it is a common practice to compute confidence levels for tree topologies and edges. For example, the bootstrap probability (Felsenstein, 1985) is the most commonly used confidence measure, and other methods such as the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999) and the multiscale bootstrap method (Shimodaira, 2002) are also often used. However, these conventional methods are limited in how well they address the issue of multiplicity when there are many alternative topologies and edges. Herein, we discuss a new approach, selective inference (SI), that is designed to address the issue of multiplicity.

For illustrating the idea of selective inference, we first look at a simple example of 1-dimensional normal random variable Z with unknown mean $\theta \in \mathbb{R}$ and variance 1:

$$Z \sim N(\theta, 1). \tag{1}$$

Observing $Z = z$, we would like to test the null hypothesis $H_0 : \theta \leq 0$ against the alternative hypothesis $H_1 : \theta > 0$. We denote the cumulative distribution function of $N(0,1)$ as $\Phi(x)$ and define the upper tail probability as $\bar{\Phi}(x) = 1 - \Phi(x) = \Phi(-x)$. Then, the ordinary (i.e., non-selective) inference leads to the p-value of the one-tailed z-test as

$$p(z) := P(Z > z \mid \theta = 0) = \bar{\Phi}(z). \tag{2}$$

What happens when we test many hypotheses at the same time? Consider random variables $Z_i \sim N(\theta_i, 1)$, $i = 1, \ldots, K_{\text{all}}$, not necessarily independent, with null hypotheses $\theta_i \leq 0$, where K_{true} hypotheses are actually true. To control the number of falsely rejecting the K_{true} hypotheses, there are several multiplicity adjusted approaches such as the family-wise error rate (FWER) and the false discovery rate (FDR). Instead of testing all the K_{all} hypotheses, selective inference (SI) allows for K_{select} hypotheses with $z_i > c_i$ for constants c_i specified in advance. This kind of selection is very common in practice (e.g., publication bias), and it is called as the *file drawer problem* by Rosenthal (1979). Instead of controlling the multiplicity of testing, SI alleviates it by reducing the number of tests. The mathematical formulation of SI is easier than FWER and FDR in the sense that hypotheses can be considered separately instead of simultaneously. Therefore, we simply write $z > c$ by dropping the index i for one of the hypotheses. In selective inference, the selection bias is adjusted by considering the conditional probability given the selection event, which leads to the following p-value (Fithian et al., 2014; Tian and Taylor, 2018)

$$p(z, c) := P(Z > z \mid Z > c, \theta = 0) = \bar{\Phi}(z)/\bar{\Phi}(c), \tag{3}$$

where $p(z)$ of Equation (2) is divided by the selection probability $P(Z > c \mid \theta = 0) = \bar{\Phi}(c)$. In the case of $c = 0$, this corresponds to the two-tailed z-test, because the selection probability is $\bar{\Phi}(0) = 0.5$ and $p(z, c) = 2p(z)$. For significance level α (we use $\alpha = 0.05$ unless otherwise stated), it properly controls the type-I error conditioned on the selection event as $P(p(Z, c) < \alpha \mid Z > c, \theta = 0) = \alpha$, while the non-selective p-value violates the type-I error as $P(p(Z) < \alpha \mid Z > c, \theta = 0) = \alpha/\bar{\Phi}(c) > \alpha$. The selection bias can be very large when $\bar{\Phi}(c) \ll 1$ (i.e., $c \gg 0$), or $K_{\text{select}} \ll K_{\text{all}}$.

Selective inference has been mostly developed for inferences after model selection (Taylor and Tibshirani, 2015; Tibshirani et al., 2016), particularly variable selection in regression settings such as lasso (Tibshirani, 1996). Recently, Terada and Shimodaira (2017) developed a general method for selective inference by adjusting the selection bias in the approximately unbiased (AU) p-value computed by the multiscale bootstrap method (Shimodaira, 2002, 2004, 2008). This new method can be used to compute, for example, confidence intervals of regression coefficients in lasso (**Figure 1**). In this paper, we apply this method to phylogenetic inference for computing proper confidence levels of tree topologies (dendrograms) and edges (clades or clusters) of species. As far as we know, this is the first attempt to consider selective inference in phylogenetics. Our selective inference method is implemented in software *scaleboot* (Shimodaira, 2019) working jointly with *CONSEL* (Shimodaira and Hasegawa, 2001) for phylogenetics, and it is also implemented in a new version of *pvclust* (Suzuki and Shimodaira, 2006) for hierarchical clustering, where only edges appeared in the observed tree are "selected" for computing p-values. Although our argument is based on the rigorous theory of mathematical statistics in Terada and Shimodaira (2017), a self-contained illustration is presented in this paper for the theory as well as the algorithm of selective inference.

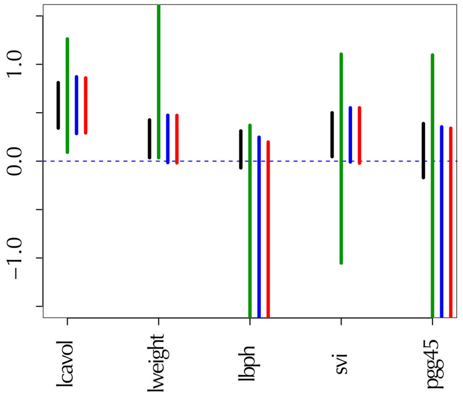


FIGURE 1 | Confidence intervals of regression coefficients for selected variables by lasso; see section 6.8 for details. All intervals are computed for confidence level $1 - \alpha$ at $\alpha = 0.01$. (Black) the ordinary confidence interval $[L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]$. (Green) the selective confidence interval $[L_j^{\text{model}}, U_j^{\text{model}}]$ under the selected model. (Blue) the selective confidence interval $[L_j^{\text{variable}}, U_j^{\text{variable}}]$ under the selection event that variable j is selected. (Red) the multiscale bootstrap version of selective confidence interval $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ under the selection event that variable j is selected.

Phylogenetic tree selection is an example of model selection. Since each tree can be specified as a combination of edges, tree selection can be interpreted as the variable selection of multiple regression, where edges correspond to the predictors of regression (Shimodaira, 2001; Shimodaira and Hasegawa, 2005). Because all candidate trees have the same number of model parameters, the maximum likelihood (ML) tree is obtained by comparing log-likelihood values of trees (Felsenstein, 1981). In order to adjust the model complexity by the number of parameters in general model selection, we compare Akaike Information Criterion (AIC) values of candidate models (Akaike, 1974). AIC is used in phylogenetics for selecting the substitution model (Posada and Buckley, 2004). There are several modifications of AIC that allow for model selection. These include the precise estimation of the complexity term known as Takeuchi Information Criterion (Burnham and Anderson, 2002; Konishi and Kitagawa, 2008), and adaptations for incomplete data (Shimodaira and Maeda, 2018) and covariate-shift data (Shimodaira, 2000). AIC and all these modifications are derived

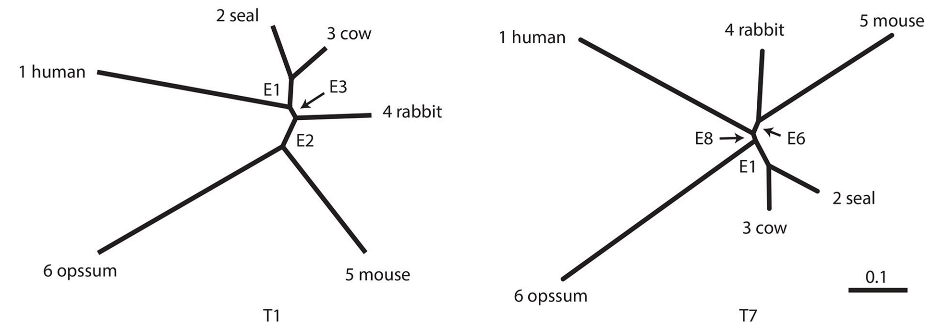


FIGURE 2 | Examples of two unrooted trees T1 and T7. Branch lengths represent ML estimates of parameters (expected number of substitutions per site). T1 includes edges E1, E2, and E3 and T7 includes E1, E6, and E8.

for estimating the expected Kullback-Leibler divergence between the unknown true distribution and the estimated probability distribution on the premise that the model is misspecified. When using regression model for prediction purpose, it may be sufficient to find only the best model which minimizes the AIC value. Considering random variations of dataset, however, it is obvious in phylogenetics that the ML tree does not necessarily represent the true history of evolution. Therefore, Kishino and Hasegawa (1989) proposed a statistical test whether two log-likelihood values differ significantly (also known as *Kishino-Hasegawa* test). The log-likelihood difference is often not significant, because its variance can be very large for non-nested models when the divergence between two probability distributions is large; see Equation (26) in section 6.1. The same idea of model selection test whether two AIC values differ significantly has been proposed independently in statistics (Linhart, 1988) and econometrics (Vuong, 1989). Another method of model selection test (Efron, 1984) allows for the comparison of two regression models with an adjusted bootstrap confidence interval corresponding to the AU *p*-value. For testing which model is better than the other, the null hypothesis in the model selection test is that the two models are equally good in terms of the expected value of AIC on the premise that both models are misspecified. Note that the null hypothesis is whether the model is correctly specified or not in the traditional hypothesis testing methods including the likelihood ratio test for nested models and the modified likelihood ratio test for non-nested models (Cox, 1962). The model selection test is very different from these traditional settings. For comparing AIC values of more than two models, a multiple comparisons method is introduced to the model selection test (Shimodaira, 1998; Shimodaira and Hasegawa, 1999), which computes the confidence set of models. But the multiple comparisons method is conservative by nature, leading to more false negatives than expected, because it considers the worst scenario, called the least favorable configuration. On the other hand, the model selection test (designed for two models) and bootstrap probability (Felsenstein, 1985) lead to more false positives than expected when comparing more than two models (Shimodaira and Hasegawa, 1999; Shimodaira, 2002). The AU *p*-value mentioned earlier has been developed for solving this problem, and we are going to upgrade it for selective inference.

2. PHYLOGENETIC INFERENCE

For illustrating phylogenetic inference methods, we analyze a dataset consisting of mitochondrial protein sequences of six mammalian species with $n = 3,414$ amino acids (n is treated as sample size). The taxa are labeled as 1=*Homo sapiens* (human), 2=*Phoca vitulina* (seal), 3=*Bos taurus* (cow), 4=*Oryctolagus cuniculus* (rabbit), 5=*Mus musculus* (mouse), and 6=*Didelphis virginiana* (opossum). The dataset will be denoted as $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The software package PAML (Yang, 1997) was used to calculate the site-wise log-likelihoods for trees. The mtREV model (Adachi and Hasegawa, 1996) was used for amino acid substitutions, and the site-heterogeneity was modeled by the discrete-gamma distribution (Yang, 1996). The dataset and evolutionary model are similar to previous publications (Shimodaira and Hasegawa, 1999; Shimodaira, 2001, 2002), thus allowing our proposed method to be easily compared with conventional methods.

The number of unrooted trees for six taxa is 105. These trees are reordered by their likelihood values and labeled as T1, T2, ..., T105. T1 is the ML tree as shown in **Figure 2** and its tree topology is represented as (((1(23))4)56). There are three internal branches (we call them as edges) in T1, which are labeled as E1, E2, and E3. For example, E1 splits the six taxa as {23|1456} and the partition of six taxa is represented as −++−−−, where +/− indicates taxa $1, \ldots, 6$ from left to right and ++ indicates the clade {23} (we set − for taxon 6, since it is treated as the outgroup). There are 25 edges in total, and each tree is specified by selecting three edges from them, although not all the combinations of three edges are allowed.

The result of phylogenetic analysis is summarized in **Table 1** for trees and **Table 2** for edges. Three types of *p*-values are computed for each tree as well as for each edge. BP is the bootstrap probability (Felsenstein, 1985) and AU is the approximately unbiased *p*-value (Shimodaira, 2002). Bootstrap probabilities are computed by the non-parametric bootstrap

TABLE 1 | Three types of p-values (BP, AU, SI) and geometric quantities (β_0, β_1) for the best 20 trees.

Tree	BP	AU	SI	β_0	β_1	Topology	Edges
T1†	0.559 (0.001)	0.752 (0.001)	0.372 (0.001)	-0.41 (0.00)	0.27 (0.00)	(((1(23))4)56)	E1, E2, E3
T2	0.304 (0.000)	0.467 (0.001)	0.798 (0.001)	0.30 (0.00)	0.22 (0.00)	((1((23)4))56)	E1 ,E2, E4
T3	**0.038** (0.000)	0.126 (0.002)	0.202 (0.003)	1.46 (0.01)	0.32 (0.00)	(((14)(23))56)	E1, E2, E5
T4	**0.014** (0.000)	0.081 (0.002)	0.124 (0.003)	1.79 (0.01)	0.40 (0.01)	((1(23))(45)6)	E1, E3, E6
T5	**0.032** (0.000)	0.127 (0.002)	0.199 (0.003)	1.50 (0.01)	0.36 (0.00)	(1((23)(45))6)	E1, E6, E7
T6	**0.005** (0.000)	**0.032** (0.002)	0.050 (0.002)	2.21 (0.02)	0.35 (0.01)	(1(((23)4)5)6)	E1, E4, E7
T7‡	**0.015** (0.000)	0.100 (0.003)	0.150 (0.003)	1.72 (0.01)	0.44 (0.01)	((1(45))(23)6)	E1, E6, E8
T8	**0.001** (0.000)	**0.011** (0.001)	**0.016** (0.002)	2.74 (0.03)	0.43 (0.02)	((15)((23)4)6)	E1, E4, E9
T9	**0.000** (0.000)	**0.001** (0.000)	**0.001** (0.000)	3.67 (0.09)	0.46 (0.04)	(((1(23))5)46)	E1, E3, E10
T10	**0.002** (0.000)	**0.022** (0.002)	**0.033** (0.002)	2.43 (0.02)	0.42 (0.01)	(((15)4)(23)6)	E1, E8, E9
T11	**0.000** (0.000)	**0.004** (0.001)	**0.006** (0.002)	3.14 (0.07)	0.51 (0.03)	(((14)5)(23)6)	E1, E5, E8
T12	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.78 (0.09)	0.41 (0.04)	(((15)(23))46)	E1, E9, E10
T13	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.001)	3.96 (0.19)	0.54 (0.09)	(1(((23)5)4)6)	E1, E7, E11
T14	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.66 (0.31)	0.65 (0.12)	((14)((23)5)6)	E1, E5, E11
T15	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.28 (0.34)	0.43 (0.11)	((1((23)5))46)	E1, E10, E11
T16	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.63 (0.04)	0.23 (0.01)	((((13)2)4)56)	E2, E3, E12
T17	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.81 (0.04)	0.22 (0.01)	((((12)3)4)56)	E2, E3, E13
T18	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.33 (0.10)	0.34 (0.03)	(((13)2)(45)6)	E3, E6, E12
T19	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.36 (0.11)	0.32 (0.04)	(((12)3)(45)6)	E3, E6, E13
T20	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.90 (0.12)	0.44 (0.05)	(((1(45))2)36)	E6, E8, E14

Standard errors are shown in parentheses. Boldface indicates significance (p < 0.05) for the null hypothesis that the tree is true (outside mode). For the rest of trees (T21, ..., T105), p-values are very small (p < 0.001). † T1 is the ML tree, i.e., the tree selected by the ML method based on the dataset of Shimodaira and Hasegawa (1999). ‡ T7 is presumably the true tree as suggested by later researches; see section 4.3.

resampling (Efron, 1979) described in section 6.1. The theory and the algorithm of BP and AU will be reviewed in section 3. Since we are testing many trees and edges at the same time, there is potentially a danger of selection bias. The issue of selection bias has been discussed in Shimodaira and Hasegawa (1999) for introducing the method of multiple comparisons of log-likelihoods (also known as *Shimodaira-Hasegawa test*) and in Shimodaira (2002) for introducing AU test. However, these conventional methods are only taking care of the multiplicity of comparing many log-likelihood values for computing just one p-value instead of many p-values at the same time. Therefore, we intend to further adjust the AU p-value by introducing the selective inference p-value, denoted as SI. The theory and the algorithm of SI will be explained in section 4 based on the geometric theory given in section 3. After presenting the methods, we will revisit the phyloegnetic inference in section 4.3.

For developing the geometric theory in sections 3 and 4, we formulate tree selection as a mathematical formulation known as *the problem of regions* (Efron et al., 1996; Efron and Tibshirani, 1998). For better understanding the geometric nature of the theory, the problem of regions is explained below for phylogenetic inference, although the algorithm is simple enough to be implemented without understanding the theory. Considering the space of probability distributions (Amari and Nagaoka, 2007), the parametric models for trees are represented as manifolds in the space. The dataset (or the empirical distribution) can also be represented as a "data point" X in the space, and the ML estimates for trees are represented as projections to the manifolds. This is illustrated in the visualization of probability distributions of **Figure 3A** using log-likelihood vectors of models (Shimodaira, 2001), where models are simply indicated as red lines from the origin; see section 6.2 for details. This visualization may be called as *model map*. The point X is actually reconstructed as the minimum full model containing all the trees as submodels, and the Kullback-Leibler divergence between probability distributions is represented as the squared distance between points; see Equation (27). Computation of X is analogous to the Bayesian model averaging, but based on the ML method. For each tree, we can think of a region in the space so that this tree becomes the ML tree when X is included in the region. The regions for T1, T2, and T3 are illustrated in **Figure 3B**, and the region for E2 is the union of these three regions.

In **Figure 3A**, X is very far from any of the tree models, suggesting that all the models are wrong; the likelihood ratio statistic for testing T1 against the full model is 113.4, which is highly significant as χ^2_8 (Shimodaira, 2001, section 5). Instead of testing whether tree models are correct or not, we test whether models are significantly better than the others. As seen in **Figure 3B**, X is in the region for T1, meaning that the model for T1 is better than those for the other trees. For convenience, observing X in the region for T1, we state that T1 is *supported* by the data. Similarly, X is in the region for E2 that consists of the three regions for T1, T2, T3, thus indicating that E2 is *supported* by the data. Although T1 and E2 are supported by the data, there is still uncertainty as to whether the true evolutionary history of lineages is depicted because the location of X fluctuates randomly. Therefore, statistical

TABLE 2 | Three types of p-values (BP, AU, SI) and geometric quantities (β_0, β_1) for all the 25 edges of six taxa.

Edge	BP	AU	SI	β_0	β_1	Clade
E1[†‡]	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	-3.87 (0.03)	0.16 (0.01)	−++−−−
E2[†]	0.930 (0.000)	**0.956** (0.001)	0.903 (0.001)	-1.59 (0.00)	0.12 (0.00)	++++−−
E3[†]	0.580 (0.001)	0.719 (0.001)	0.338 (0.001)	-0.39 (0.00)	0.19 (0.00)	+++−−−
E4	0.318 (0.000)	0.435 (0.001)	0.775 (0.001)	0.32 (0.00)	0.16 (0.00)	−+++−−
E5	**0.037** (0.000)	0.124 (0.002)	0.198 (0.002)	1.47 (0.01)	0.32 (0.00)	+−−+−−
E6[‡]	0.060 (0.000)	0.074 (0.001)	0.141 (0.002)	1.50 (0.00)	0.05 (0.00)	−−−++−
E7	**0.038** (0.000)	0.091 (0.002)	0.154 (0.002)	1.56 (0.01)	0.22 (0.00)	−++++−
E8[‡]	**0.018** (0.000)	0.068 (0.002)	0.110 (0.003)	1.80 (0.01)	0.31 (0.01)	+−−++−
E9	**0.003** (0.000)	**0.014** (0.001)	**0.023** (0.002)	2.48 (0.02)	0.27 (0.02)	+−−−+−
E10	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.72 (0.07)	0.29 (0.03)	+++−+−
E11	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.31 (0.10)	0.35 (0.03)	−++−+−
E12	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.68 (0.05)	0.17 (0.02)	+−+−−−
E13	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.90 (0.04)	0.15 (0.02)	++−−−−
E14	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.03 (0.09)	0.30 (0.04)	++−++−
E15	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.03 (0.13)	0.38 (0.06)	+−+++−
E16	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.44 (0.05)	0.12 (0.01)	−+−+−−
E17	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.70 (0.07)	0.19 (0.02)	++−+−−
E18	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.94 (0.09)	0.26 (0.04)	−+−++−
E19	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.23 (0.43)	0.57 (0.13)	−−++−−
E20	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.66 (0.29)	0.28 (0.09)	+−++−−
E21	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	6.38 (0.33)	0.24 (0.08)	−−+++−
E22	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.62 (0.21)	0.17 (0.07)	−−+−+−
E23	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.86 (0.43)	0.70 (0.13)	−+−−+−
E24	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.61 (0.17)	0.23 (0.04)	+−+−+−
E25	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	6.32 (0.71)	0.52 (0.20)	++−−+−

Standard errors are shown in parentheses. Boldface without underline indicates significance (p < 0.05) for the null hypothesis that the edge is true (outside mode). Boldface with underline indicates significance (p > 0.95) for the null hypothesis that the edge is not *true (inside mode).* [†] *Edges included in T1.* [‡] *Edges included in T7.*

confidence of the outcome needs to be assessed. A mathematical procedure for statistically evaluating the outcome is provided in the following sections.

3. NON-SELECTIVE INFERENCE FOR THE PROBLEM OF REGIONS

3.1. The Problem of Regions

For developing the theory, we consider $(m+1)$-dimensional multivariate normal random vector $\boldsymbol{Y}$, $m \geq 0$, with unknown mean vector $\boldsymbol{\mu} \in \mathbb{R}^{m+1}$ and the identity variance matrix $\boldsymbol{I}_{m+1}$:

$$\boldsymbol{Y} \sim N_{m+1}(\boldsymbol{\mu}, \boldsymbol{I}_{m+1}). \tag{4}$$

A region of interest such as tree and edge is denoted as $\mathcal{R} \subset \mathbb{R}^{m+1}$, and its complement set is denoted as $\mathcal{R}^C = \mathbb{R}^{m+1} \setminus \mathcal{R}$. There are K_{all} regions $\mathcal{R}_i$, $i = 1, \ldots, K_{\text{all}}$, and we simply write $\mathcal{R}$ for one of them by dropping the index i. Observing $\boldsymbol{Y} = \boldsymbol{y}$, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. This setting is called *problem of regions*, and the geometric theory for non-selective inference for slightly generalized settings (e.g., exponential family of distributions) has been discussed in Efron and Tibshirani (1998) and Shimodaira (2004). This theory allows arbitrary shape of $\mathcal{R}$ without assuming a particular shape such as half-space or sphere, and only requires the expression (29) of section 6.3.

The problem of regions is well described by geometric quantities (**Figure 4**). Let $\hat{\boldsymbol{\mu}}$ be the projection of $\boldsymbol{y}$ to the boundary surface $\partial\mathcal{R}$ defined as

$$\hat{\boldsymbol{\mu}} = \mathop{\arg\min}_{\boldsymbol{\mu} \in \partial\mathcal{R}} \|\boldsymbol{y} - \boldsymbol{\mu}\|,$$

and β_0 be the *signed distance* defined as $\beta_0 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\| > 0$ for $\boldsymbol{y} \in \mathcal{R}^C$ and $\beta_0 = -\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\| \leq 0$ for $\boldsymbol{y} \in \mathcal{R}$; see **Figures 4A,B**, respectively. A large β_0 indicates the evidence for rejecting $H_0 : \boldsymbol{\mu} \in \mathcal{R}$, but computation of p-value will also depend on the shape of $\mathcal{R}$. There should be many parameters for defining the shape, but we only need the *mean curvature* of $\partial\mathcal{R}$ at $\hat{\boldsymbol{\mu}}$, which represents the amount of surface bending. It is denoted as $\beta_1 \in \mathbb{R}$, and defined in (30).

Geometric quantities β_0 and β_1 of regions for trees (T1, ..., T105) and edges (E1, ..., E25) are plotted in **Figure 5**, and these values are also found in **Tables 1**, **2**. Although the phylogenetic model of evolution for the molecular dataset $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is different from the multivariate normal model (4) for $\boldsymbol{y}$, the multiscale bootstrap method of section 3.4 estimates β_0 and β_1 using the non-parametric bootstrap probabilities (section 6.1) with bootstrap replicates $\mathcal{X}^*_{n'}$ for several values of sample size n'.

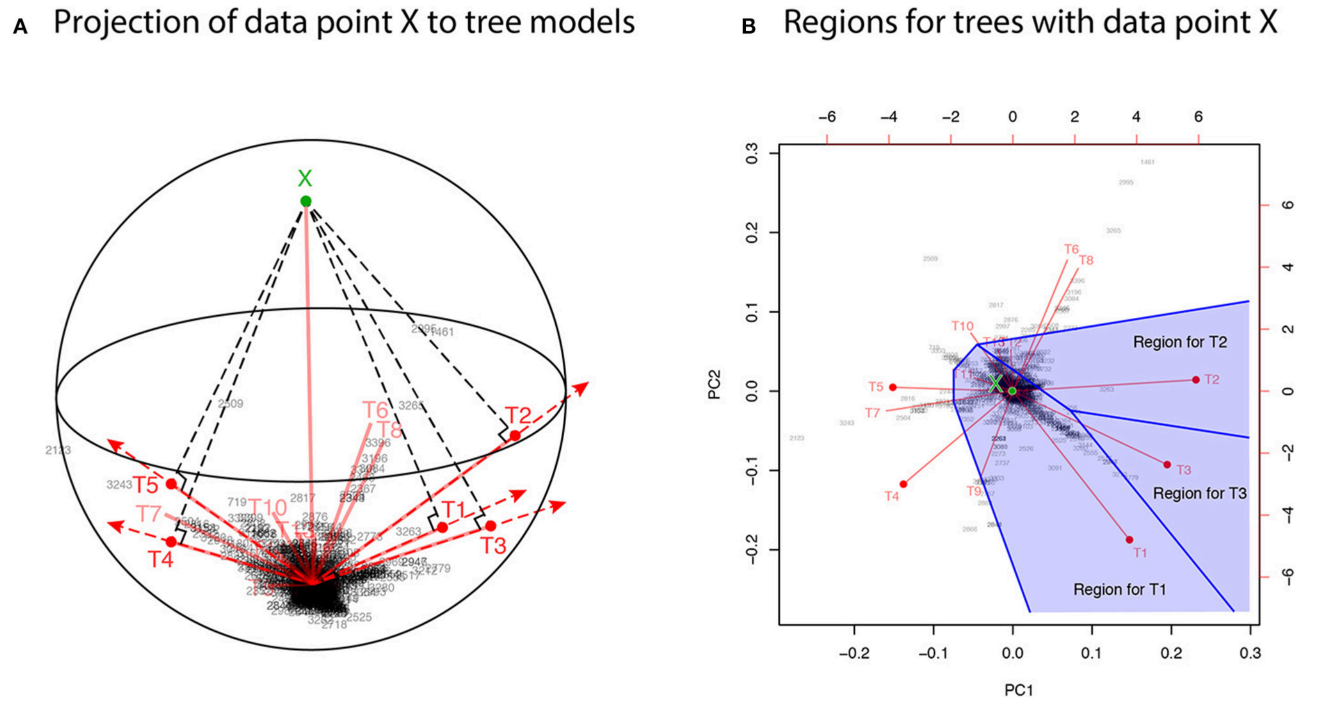


FIGURE 3 | Model map: Visualization of ML estimates of probability distributions for the best 15 trees. The origin represents the star-shaped tree topology (obtained by reducing the internal branches to zero length). Sites of amino acid sequences $t = 1, \ldots, n$ (black numbers) and probability distributions for trees T1, . . . ,T15 (red segments) are drawn by biplot of PCA. Auxiliary lines are drawn by hand. **(A)** 3-dimensional visualization using PC1, PC2, and PC3. The reconstructed data point X is also shown (green point). The ML estimates are represented as the end points of the red segments (shown by red points only for the best five trees), and they are placed on the sphere with the origin and X being placed at the poles. **(B)** The top-view of model map. Regions for the best three trees Ti, $i = 1, 2, 3$ (blue shaded regions) are illustrated; Ti will be the ML tree if X is included in the region for Ti.

3.2. Bootstrap Probability

For simulating (4) from $\boldsymbol{y}$, we may generate replicates $\boldsymbol{Y}^*$ from the bootstrap distribution (**Figure 4C**)

$$\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1}), \tag{5}$$

and define bootstrap probability (BP) of $\mathcal{R}$ as the probability of $\boldsymbol{Y}^*$ being included in the region $\mathcal{R}$:

$$\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) := P(\boldsymbol{Y}^* \in \mathcal{R}|\boldsymbol{y}). \tag{6}$$

$\mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ can be interpreted as the Bayesian posterior probability $P(\boldsymbol{\mu} \in \mathcal{R}|\boldsymbol{y})$, because, by assuming the flat prior distribution $\pi(\boldsymbol{\mu}) =$ constant, the posterior distribution $\boldsymbol{\mu}|\boldsymbol{y} \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1})$ is identical to the distribution of $\boldsymbol{Y}^*$ in (5). An interesting consequence of the geometric theory of Efron and Tibshirani (1998) is that BP can be expressed as

$$\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\beta_0 + \beta_1), \tag{7}$$

where $\simeq$ indicates the *second order asymptotic accuracy*, meaning that the equality is correct up to $O_p(n^{-1/2})$ with error of order $O_p(n^{-1})$; see section 6.3.

For understanding the formula (7), assume that $\mathcal{R}$ is a half space so that $\partial\mathcal{R}$ is flat and $\beta_1 = 0$. Since we only have to look at the axis orthogonal to $\partial\mathcal{R}$, the distribution of signed distance is identified as (1) with $\beta_0 = z$. The bootstrap distribution for (1) is $Z^* \sim N(z, 1)$, and bootstrap probability is expressed as $P(Z^* \leq 0|z) = \bar{\Phi}(z)$. Therefore, we have $\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) = \bar{\Phi}(\beta_0)$. For general $\mathcal{R}$ with curved $\partial\mathcal{R}$, the formula (7) adjusts the bias caused by β_1. As seen in **Figure 4C**, $\mathcal{R}$ becomes smaller for $\beta_1 > 0$ than $\beta_1 = 0$, and BP becomes smaller.

BP of $\mathcal{R}^C$ is closely related to BP of $\mathcal{R}$. From the definition,

$$\mathrm{BP}(\mathcal{R}^C|\boldsymbol{y}) = 1-\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) \simeq 1-\bar{\Phi}(\beta_0+\beta_1) = \bar{\Phi}(-\beta_0-\beta_1). \tag{8}$$

The last expression also implies that the signed distance and the mean curvature of $\mathcal{R}^C$ is $-\beta_0$ and $-\beta_1$, respectively; this relation is also obtained by reversing the sign of v in (29).

3.3. Approximately Unbiased Test

Although $\mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ may work as a Bayesian confidence measure, we would like to have a frequentist confidence measure for testing $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ against $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. The signed distance of $\boldsymbol{Y}$ is denoted as $\beta_0(\boldsymbol{Y})$, and consider the region $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}$ in which the signed distance is larger than the observed value $\beta_0 = \beta_0(\boldsymbol{y})$. Similar to (2), we then define an approximately unbiased (AU) p-value as

$$\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) := P(\beta_0(\boldsymbol{Y}) > \beta_0 \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}}) = \mathrm{BP}(\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}|\hat{\boldsymbol{\mu}}), \tag{9}$$

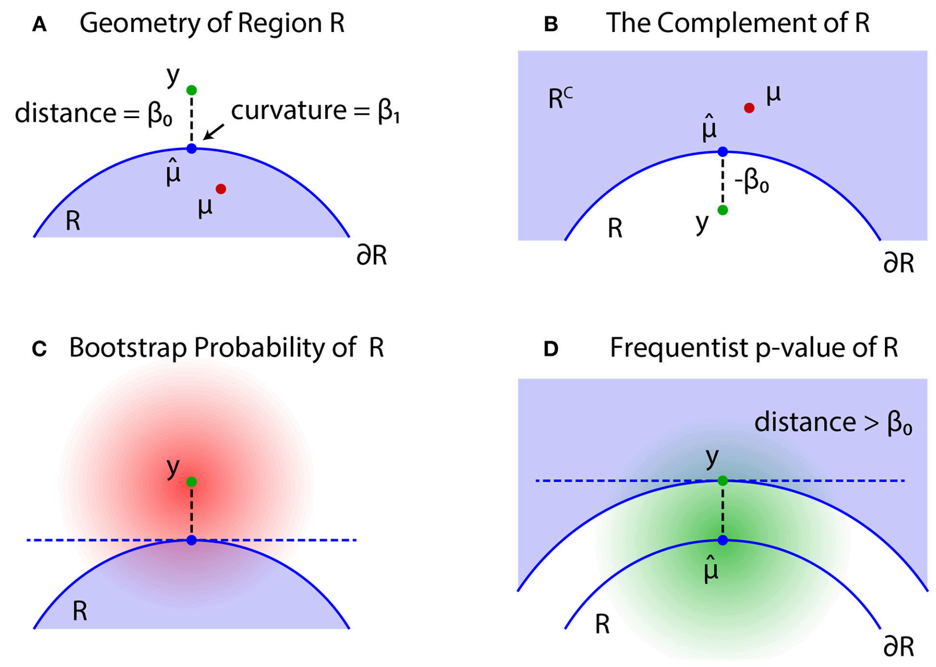


FIGURE 4 | Problem of regions. **(A)** $\beta_0 > 0$ when $\boldsymbol{y} \in \mathcal{R}^C$, then select the null hypothesis $\boldsymbol{\mu} \in \mathcal{R}$. **(B)** $\beta_0 \leq 0$ when $\boldsymbol{y} \in \mathcal{R}$, then select the null hypothesis $\boldsymbol{\mu} \in \mathcal{R}^C$. **(C)** The bootstrap distribution of $\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1})$ (red shaded distribution). **(D)** The null distribution of $\boldsymbol{Y} \sim N_{m+1}(\hat{\boldsymbol{\mu}}, \boldsymbol{I}_{m+1})$ (green shaded distribution).

where the probability is calculated for $\boldsymbol{Y} \sim N_{m+1}(\hat{\boldsymbol{\mu}}, \boldsymbol{I}_{m+1})$ as illustrated in **Figure 4D**. The shape of the region $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}$ is very similar to the shape of $\mathcal{R}^C$; the difference is in fact only $O_p(n^{-1})$. Let us think of a point $\boldsymbol{y}'$ with signed distance $-\beta_0$ (shown as $\boldsymbol{y}$ in **Figure 4B**). Then we have

$$\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) \simeq \mathrm{BP}(\mathcal{R}^C|\boldsymbol{y}') \simeq \bar{\Phi}(\beta_0 - \beta_1), \tag{10}$$

where the last expression is obtained by substituting $(-\beta_0, \beta_1)$ for (β_0, β_1) in (8). This formula computes AU from (β_0, β_1). An intuitive interpretation of (10) is explained in section 6.4.

In non-selective inference, p-values are computed using formula (10). If $\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) < \alpha$, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is rejected and the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$ is accepted. This test procedure is approximately unbiased, because it controls the non-selective type-I error as

$$P\big(\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{\mu} \in \partial\mathcal{R}\big) \simeq \alpha, \tag{11}$$

and the rejection probability increases as $\boldsymbol{\mu}$ moves away from $\mathcal{R}$, while it decreases as $\boldsymbol{\mu}$ moves into $\mathcal{R}$.

Exchanging the roles of $\mathcal{R}$ and $\mathcal{R}^C$ also allows for another hypothesis testing. AU of $\mathcal{R}^C$ is obtained from (9) by reversing the inequality as $\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) = \mathrm{BP}(\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) < \beta_0\}|\hat{\boldsymbol{\mu}}) = 1 - \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$. This is also confirmed by substituting $(-\beta_0, -\beta_1)$, i.e., the geometric quantities of $\mathcal{R}^C$, for (β_0, β_1) in (10) as

$$\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) \simeq \bar{\Phi}(-\beta_0 + \beta_1) \simeq 1 - \mathrm{AU}(\mathcal{R}|\boldsymbol{y}). \tag{12}$$

If $\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$ or equivalently $\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) > 1 - \alpha$, then we reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}$.

3.4. Multiscale Bootstrap

In order to estimate β_0 and β_1 from bootstrap probabilities, we consider a generalization of (5) as

$$\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \sigma^2 \boldsymbol{I}_{m+1}), \tag{13}$$

for a variance $\sigma^2 > 0$, and define multiscale bootstrap probability of $\mathcal{R}$ as

$$\mathrm{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) := P_{\sigma^2}(\boldsymbol{Y}^* \in \mathcal{R}|\boldsymbol{y}), \tag{14}$$

where P_{σ^2} indicates the probability with respect to (13).

Although our theory is based on the multivariate normal model, the actual implementation of the algorithm uses the non-parametric bootstrap probabilities in section 6.1. To fill the gap between the two models, we consider a non-linear transformation $\boldsymbol{f}_n$ so that the multivariate normal model holds at least approximately for $\boldsymbol{y} = \boldsymbol{f}_n(\mathcal{X}_n)$ and $\boldsymbol{Y}^* = \boldsymbol{f}_n(\mathcal{X}^*_{n'})$. An example of $\boldsymbol{f}_n$ is given in (25) for phylogenetic inference. Surprisingly, a specification of $\boldsymbol{f}_n$ is *not required* for computing p-values, but we simply assume the existence of such a transformation; this property may be called as "bootstrap trick." For phylogenetic inference, we compute the non-parametric bootstrap probabilities by (24) and substitute these values for (14) with $\sigma^2 = n/n'$.

For estimating β_0 and β_1, we need to have a scaling law which explains how BP_{σ^2} depends on the scale σ. We rescale (13) by multiplying σ^{-1} so that $\sigma^{-1}\boldsymbol{Y}^* \sim N_{m+1}(\sigma^{-1}\boldsymbol{y}, \boldsymbol{I}_{m+1})$ has the variance $\sigma^2 = 1$. $\boldsymbol{y}$ and $\mathcal{R}$ are now resaled by the factor σ^{-1}, which amounts to signed distance $\beta_0\sigma^{-1}$ and mean curvature

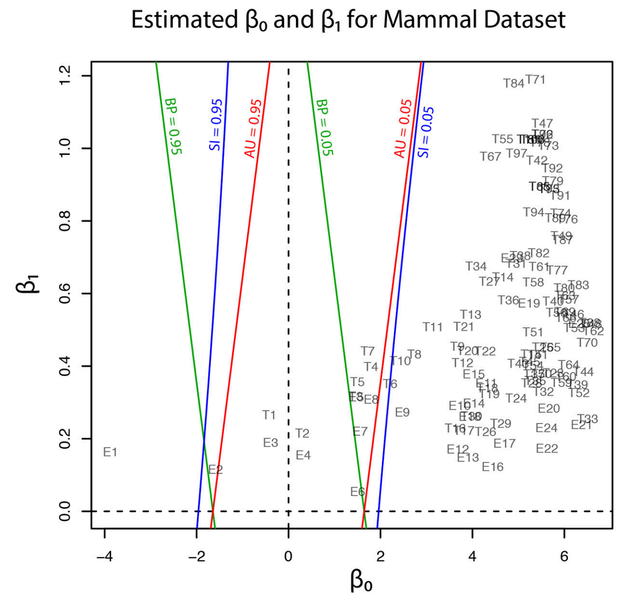


FIGURE 5 | Geometric quantities of regions (β_0 and β_1) for trees and edges are estimated by the multiscale bootstrap method (section 3.4). The three types of *p*-value (BP, AU, SI) are computed from β_0 and β_1, and their contour lines are drawn at $p = 0.05$ and 0.95.

$\beta_1\sigma$ (Shimodaira, 2004). Therefore, by substituting $(\beta_0\sigma^{-1}, \beta_1\sigma)$ for (β_0, β_1) in (7), we obtain

$$\text{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\beta_0\sigma^{-1} + \beta_1\sigma). \tag{15}$$

For better illustrating how BP_{σ^2} depends on σ^2, we define

$$\psi_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) := \sigma\bar{\Phi}^{-1}(\text{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y})) \simeq \beta_0 + \beta_1\sigma^2. \tag{16}$$

We can estimate β_0 and β_1 as regression coefficients by fitting the linear model (16) in terms of σ^2 to the observed values of non-parametric bootstrap probabilities (**Figure 6**). Interestingly, (10) is rewritten as $\text{AU}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\psi_{-1}(\mathcal{R}|\boldsymbol{y}))$ by formally letting $\sigma^2 = -1$ in the last expression of (16), meaning that AU corresponds to $n' = -n$. Although σ^2 should be positive in (15), we can think of negative σ^2 in $\beta_0 + \beta_1\sigma^2$. See section 6.5 for details of model fitting and extrapolation to negative σ^2.

4. SELECTIVE INFERENCE FOR THE PROBLEM OF REGIONS

4.1. Approximately Unbiased Test for Selective Inference

In order to argue selective inference for the problem of regions, we have to specify the selection event. Let us consider a selective region $\mathcal{S} \subset \mathcal{R}^{m+1}$ so that we perform the hypothesis testing only when $\boldsymbol{y} \in \mathcal{S}$. Terada and Shimodaira (2017) considered a general shape of $\mathcal{S}$, but here we treat only two special cases of $\mathcal{S} = \mathcal{R}^C$ and $\mathcal{S} = \mathcal{R}$; see section 6.6. Our problem is formulated as follows. Observing $\boldsymbol{Y} = \boldsymbol{y}$ from the multivariate normal model (4), we first check whether $\boldsymbol{y} \in \mathcal{R}^C$ or $\boldsymbol{y} \in \mathcal{R}$. If $\boldsymbol{y} \in \mathcal{R}^C$ and we are interested in the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$, then we may test it against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. If $\boldsymbol{y} \in \mathcal{R}$ and we are interested in the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$, then we may test it against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}$. In this paper, the former case ($\boldsymbol{y} \in \mathcal{R}^C$, and so $\beta_0 > 0$) is called as *outside mode*, and the latter case ($\boldsymbol{y} \in \mathcal{R}$, and so $\beta_0 \leq 0$) is called as *inside mode*. We do not know which of the two modes of testing is performed until we observe $\boldsymbol{y}$.

Let us consider the outside mode by assuming that $\boldsymbol{y} \in \mathcal{R}^C$, where $\beta_0 > 0$. Recalling that $p(z, c) = p(z)/\bar{\Phi}(c)$ in section 1, we divide $\text{AU}(\mathcal{R}|\boldsymbol{y})$ by the selection probability to define a selective inference *p*-value as

$$\text{SI}(\mathcal{R}|\boldsymbol{y}) := \frac{P(\beta_0(\boldsymbol{Y}) > \beta_0 \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}})}{P(\boldsymbol{Y} \in \mathcal{R}^C \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}})} = \frac{\text{AU}(\mathcal{R}|\boldsymbol{y})}{\text{BP}(\mathcal{R}^C|\hat{\boldsymbol{\mu}})}. \tag{17}$$

From the definition, $\text{SI}(\mathcal{R}|\boldsymbol{y}) \in (0, 1)$, because $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\} \subset \mathcal{R}^C$ for $\beta_0 > 0$. This *p*-value is computed from (β_0, β_1) by

$$\text{SI}(\mathcal{R}|\boldsymbol{y}) \simeq \frac{\bar{\Phi}(\beta_0 - \beta_1)}{\bar{\Phi}(-\beta_1)}, \tag{18}$$

where $\text{BP}(\mathcal{R}^C|\hat{\boldsymbol{\mu}}) = \bar{\Phi}(-\beta_1)$ is obtained by substituting $(0, \beta_1)$ for (β_0, β_1) in (8). An intuitive justification of (18) is explained in section 6.4.

For the outside mode of selective inference, *p*-values are computed using formula (18). If $\text{SI}(\mathcal{R}|\boldsymbol{y}) < \alpha$, then reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. This test procedure is approximately unbiased, because it controls the selective type-I error as

$$P\big(\text{SI}(\mathcal{R}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{Y} \in \mathcal{R}^C, \boldsymbol{\mu} \in \partial\mathcal{R}\big) \simeq \alpha, \tag{19}$$

and the rejection probability increases as $\boldsymbol{\mu}$ moves away from $\mathcal{R}$, while it decreases as $\boldsymbol{\mu}$ moves into $\mathcal{R}$.

Now we consider the inside mode by assuming that $\boldsymbol{y} \in \mathcal{R}$, where $\beta_0 \leq 0$. SI of $\mathcal{R}^C$ is obtained from (17) by exchanging the roles of $\mathcal{R}$ and $\mathcal{R}^C$.

$$\text{SI}(\mathcal{R}^C|\boldsymbol{y}) = \frac{\text{AU}(\mathcal{R}^C|\boldsymbol{y})}{\text{BP}(\mathcal{R}|\hat{\boldsymbol{\mu}})} \simeq \frac{\bar{\Phi}(-\beta_0 + \beta_1)}{\bar{\Phi}(\beta_1)}. \tag{20}$$

For the inside mode of selective inference, *p*-values are computed using formula (20). If $\text{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$, then reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}$. Unlike the non-selective *p*-value $\text{AU}(\mathcal{R}^C|\boldsymbol{y})$, $\text{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$ is *not* equivalent to $\text{SI}(\mathcal{R}|\boldsymbol{y}) > 1 - \alpha$, because $\text{SI}(\mathcal{R}|\boldsymbol{y}) + \text{SI}(\mathcal{R}^C|\boldsymbol{y}) \neq 1$. For convenience, we define

$$\text{SI}'(\mathcal{R}|\boldsymbol{y}) := \begin{cases} \text{SI}(\mathcal{R}|\boldsymbol{y}) & \boldsymbol{y} \in \mathcal{R}^C \\ 1 - \text{SI}(\mathcal{R}^C|\boldsymbol{y}) & \boldsymbol{y} \in \mathcal{R} \end{cases} \tag{21}$$

so that $\text{SI}' > 1 - \alpha$ implies $\text{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$. In our numerical examples of **Figure 5**, **Tables 1**, **2**, SI′ is simply denoted as SI. We do not need to consider (21) for BP and AU, because $\text{BP}'(\mathcal{R}|\boldsymbol{y}) = \text{BP}(\mathcal{R}|\boldsymbol{y})$ and $\text{AU}'(\mathcal{R}|\boldsymbol{y}) = \text{AU}(\mathcal{R}|\boldsymbol{y})$ from (8) and (12).

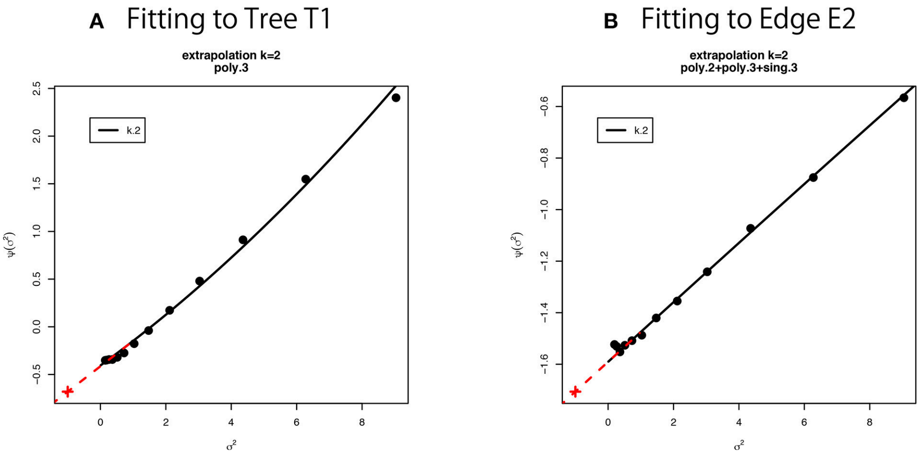


FIGURE 6 | Multiscale bootstrap for **(A)** tree T1 and **(B)** edge E2. $\psi_{\sigma^2}(\mathcal{R}|y)$ is computed by the non-parametric bootstrap probabilities for several $\sigma^2 = n/n'$ values, then β_0 and β_1 are estimated as the intercept and the slope, respectively. See section 6.5 for details.

4.2. Shortcut Computation of SI

We can compute SI from BP and AU. This will be useful for reanalyzing the results of previously published researches. Let us write BP $=$ BP$(\mathcal{R}|\boldsymbol{y})$ and AU $=$ AU$(\mathcal{R}|\boldsymbol{y})$. From (7) and (10), we have

$$\beta_0 = \tfrac{1}{2}\left(\bar{\Phi}^{-1}(\mathrm{BP}) + \bar{\Phi}^{-1}(\mathrm{AU})\right)$$
$$\beta_1 = \tfrac{1}{2}\left(\bar{\Phi}^{-1}(\mathrm{BP}) - \bar{\Phi}^{-1}(\mathrm{AU})\right).$$

We can compute SI from β_0 and β_1 by (18) or (20). More directly, we may compute

$$\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) = \frac{\mathrm{AU}}{\bar{\Phi}\left\{\frac{1}{2}\left(\bar{\Phi}^{-1}(\mathrm{AU}) - \bar{\Phi}^{-1}(\mathrm{BP})\right)\right\}}$$
$$\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) = \frac{1-\mathrm{AU}}{\bar{\Phi}\left\{\frac{1}{2}\left(\bar{\Phi}^{-1}(\mathrm{BP}) - \bar{\Phi}^{-1}(\mathrm{AU})\right)\right\}}.$$

4.3. Revisiting the Phylogenetic Inference

In this section, the analytical procedure outlined in section 2 is used to determine relationships among human, mouse, and rabbit. The question is: Which of mouse or human is closer to rabbit? The traditional view (Novacek, 1992) is actually supporting E6, the clade of rabbit and mouse, which is consistent with T4, T5, and T7. Based on molecular analysis, Graur et al. (1996) strongly suggested that rabbit is closer to human than mouse, thus supporting E2, which is consistent with T1, T2, and T3. However, Halanych (1998) criticized it by pointing out that E2 is an artifact caused by the *long branch attraction* (LBA) between mouse and opossum. In addition, Shimodaira and Hasegawa (1999) and Shimodaira (2002) suggested that T7 is not rejected by multiplicity adjusted tests. Shimodaira and Hasegawa (2005) showed that T7 becomes the ML tree by resolving the LBA using a larger dataset with more taxa. Although T1 is the ML tree based on the dataset with fewer taxa, T7 is presumably the true tree as indicated by later researches. With these observations in mind, we retrospectively interpret p-values in **Tables 1**, **2**.

The results are shown below for the two test modes (inside and outside) as defined in section 4.1. The extent of multiplicity and selection bias depends on the number of regions under consideration, thus these numbers are considered for interpreting the results. The numbers of regions related to trees and edges are summarized in **Table 3**; see section 6.7 for details.

In inside mode, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}_i^C$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}_i$ for $\boldsymbol{y} \in \mathcal{R}_i$ (i.e., $\beta_0 \leq 0$). This applies to the regions for T1, E1, E2, and E3, and they are *supported* by the data in the sense mentioned in the last paragraph of section 2. When H_0 is rejected by a test procedure, it is claimed that $\mathcal{R}_i$ is *significantly supported* by the data, indicating H_1 holds true. For convenience, the null hypothesis H_0 is said like E1 is not true, and the alternative hypothesis H_1 is said like E1 is true; then rejection of H_0 implies that E1 is true. This procedure looks unusual, but makes sense when both $\mathcal{R}_i$ and $\mathcal{R}_i^C$ are regions with nonzero volume. Note that selection bias can be very large in the sense that $K_{\mathrm{select}}/K_{\mathrm{all}} \approx 0$ for many taxa, and non-selective tests may lead to many false positives because $K_{\mathrm{true}}/K_{\mathrm{all}} \approx 1$. Therefore selective inference should be used in inside mode.

In outside mode, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}_i$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}_i^C$ for $\boldsymbol{y} \in \mathcal{R}_i^C$ (i.e., $\beta_0 > 0$). This applies to the regions for T2, ..., T105, and E4, ..., E25, and they are *not supported* by the data. When H_0 is rejected by a test procedure, it is claimed that $\mathcal{R}_i$ is rejected.

TABLE 3 | The number of regions for trees and edges. The number of taxa is $N = 6$.

	Inside mode		Outside mode	
	Tree	**Edge**	**Tree**	**Edge**
K_{select}	1	3	104	22
K_{true}	104	22	1	3
K_{all}	105	25	105	25

For convenience, the null hypothesis is said like T9 is true, and the alternative hypothesis is said like T9 is not true; rejection of H_0 implies that T9 is not true. This is more or less a typical test procedure. Note that selection bias is minor in the sense that $K_{\text{select}}/K_{\text{all}} \approx 1$ for many taxa, and non-selective tests may result in few false positives because $K_{\text{true}}/K_{\text{all}} \approx 0$. Therefore selective inference is not much beneficial in outside mode.

In addition to p-values for some trees and edges, estimated geometric quantities are also shown in the tables. We confirm that the sign of β_0 is estimated correctly for all the trees and edges. The estimated β_1 values are all positive, indicating the regions are convex. This is not surprising, because the regions are expressed as intersections of half spaces at least locally (**Figure 3B**).

Now p-values are examined in inside mode. (T1, E3) BP, AU, SI are all $p \leq 0.95$. This indicates that T1 and E3 are *not* significantly supported. There are nothing claimed to be definite. (E1) BP, AU, SI are all $p > 0.95$, indicating E1 is significantly supported. Since E1 is associated with the best 15 trees T1, ..., T15, some of them are significantly better than the rest of trees T16, ..., T105. Significance for edges is common in phylogenetics as well as in hierarchical clustering (Suzuki and Shimodaira, 2006). (E2) The results split for this presumably wrong edge. AU > 0.95 suggests E2 is significantly supported, whereas $\text{BP}, \text{SI} \leq 0.95$ are not significant. AU tends to violate the selective type-I error, leading to false positives or overconfidence in wrong trees/edges, whereas SI is approximately unbiased for the selected hypothesis. This overconfidence is explained by the inequality AU $>$ SI (meant SI′ here) for $\boldsymbol{y} \in \mathcal{R}$, which is obtained by comparing (12) and (20). Therefore SI is preferable to AU in inside mode. BP is safer than AU in the sense that BP $<$ AU for $\beta_1 > 0$, but BP is not guaranteed for controlling type-I error in a frequentist sense. The two inequalities ($\text{SI}, \text{BP} < \text{AU}$) are verified as relative positions of the contour lines at $p = 0.95$ in **Figure 5**. The three p-values can be very different from each other for large β_1.

Next p-values are examined in outside mode. (T2, E4, E6) BP, AU, SI are all $p \geq 0.05$. They are *not* rejected, and there are nothing claimed to be definite. (T8, T9, ..., T105, E9,..., E25) BP, AU, SI are all $p < 0.05$. These trees and edges are rejected. (T7, E8) The results split for these presumably true tree and edge. BP < 0.05 suggests T7 and E8 are rejected, whereas $\text{AU}, \text{SI} \geq 0.05$ are not significant. AU is approximately unbiased for controlling the type-I error when H_0 is specified in advance (Shimodaira, 2002). Since BP $<$ AU for $\beta_1 > 0$, BP violates the type-I error, which results in overconfidence in non-rejected wrong trees. Therefore BP should be avoided in outside mode. Inequality AU $<$ SI can be shown for $\boldsymbol{y} \in \mathcal{R}^C$ by comparing (10) and (18). Since the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is chosen after looking at $\boldsymbol{y} \in \mathcal{R}^C$, AU is not approximately unbiased for controlling the selective type-I error, whereas SI adjusts this selection bias. The two inequalities (BP $<$ AU $<$ SI) are verified as relative positions of the contour lines at $p = 0.05$ in **Figure 5**. AU and SI behave similarly (Note: $K_{\text{select}}/K_{\text{all}} \approx 1$), while BP is very different from AU and SI for large β_1. It is arguable which of AU and SI is appropriate: AU is preferable to SI in tree selection ($K_{\text{true}} = 1$), because the multiplicity of testing is controlled as $\text{FWER} = P(\text{reject any true null}) = P(\text{AU}(\mathcal{R}_{\text{true tree}}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{\mu} \in \mathcal{R}_{\text{true tree}}) \leq \alpha$. The FWER is multiplied by $K_{\text{true}} \geq 1$ for edge selection, and SI does not fix it either. For testing edges in outside mode, AU may be used for screening purpose with a small α value such as α/K_{true}.

5. CONCLUSION

We have developed a new method for computing selective inference p-values from multiscale bootstrap probabilities, and applied this new method to phylogenetics. It is demonstrated through theory and a real-data analysis that selective inference p-values are in particular useful for testing selected edges (i.e., clades or clusters of species) to claim that they are supported significantly if $p > 1 - \alpha$. On the other hand, the previously proposed non-selective version of approximately unbiased p-values are still useful for testing candidate trees to claim that they are rejected if $p < \alpha$. Although we focused on phylogenetics, our general theory of selective inference may be applied to other model selection problems, or more general selection problems.

6. REMARKS

6.1. Bootstrap Resampling of Log-Likelihoods

Non-parametric bootstrap is often time consuming for recomputing the maximum likelihood (ML) estimates for bootstrap replicates. Kishino et al. (1990) considered the resampling of estimated log-likelihoods (RELL) method for reducing the computation. Let $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the dataset of sample size n, where $\boldsymbol{x}_t$ is the site-pattern of amino acids at site t for $t = 1, \ldots, n$. By resampling $\boldsymbol{x}_t$ from $\mathcal{X}_n$ with replacement, we obtain a bootstrap replicate $\mathcal{X}^*_{n'} = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_{n'})$ of sample size n'. Although $n' = n$ for the ordinary bootstrap, we will use several $n' > 0$ values for the multiscale bootstrap. The parametric model of probability distribution for tree Ti is $p_i(\boldsymbol{x}; \boldsymbol{\theta}_i)$ for $i = 1, \ldots, 105$, and the log-likelihood function is $\ell_i(\boldsymbol{\theta}_i; \mathcal{X}_n) = \sum_{t=1}^n \log p_i(\boldsymbol{x}_t; \boldsymbol{\theta}_i)$. Computation of the ML estimate $\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}_i} \ell_i(\boldsymbol{\theta}_i; \mathcal{X}_n)$ is time consuming, so we do not recalculate $\hat{\boldsymbol{\theta}}^*_i = \arg\max_{\boldsymbol{\theta}_i} \ell_i(\boldsymbol{\theta}_i; \mathcal{X}^*_{n'})$ for bootstrap replicates. Define the site-wise log-likelihood at site t for tree Ti as

$$\xi_{ti} = \log p_i(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}_i), \quad t = 1, \ldots, n, \; i = 1, \ldots, 105, \tag{22}$$

so that the log-likelihood value for tree Ti is written as $\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) = \sum_{t=1}^n \xi_{ti}$. The bootstrap replicate of the log-likelihood value is approximated as

$$\ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'}) \approx \ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}^*_{n'}) = \sum_{t=1}^n w^*_t \xi_{ti}, \tag{23}$$

where w^*_t is the number of times $\boldsymbol{x}_t$ appears in $\mathcal{X}^*_{n'}$. The accuracy of this approximation as well as the higher-order term is given in Equations (4) and (5) of Shimodaira (2001). Once $\ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$, $i = 1, \ldots, 105$, are computed by (23), its ML tree is T$\hat{i}^*$ with $\hat{i}^* = \arg\max_{i=1,\ldots,105} \ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$.

The non-parametric bootstrap probability of tree Ti is obtained as follows. We generate B bootstrap replicates $\boldsymbol{X}^{*b}_{n'}$, $b = 1,\dots,B$. In this paper, we used $B = 10^5$. For each $\boldsymbol{X}^{*b}_{n'}$, the ML tree T$\hat{i}^{*b}$ is computed by the method described above. Then we count the frequency that Ti becomes the ML tree in the B replicates. The non-parametric bootstrap probability of tree Ti is computed by

$$\mathrm{BP}(\mathrm{T}i, n') = \#\{\hat{i}^{*b} = i,\ b = 1,\dots,B\}/B. \tag{24}$$

The non-parametric bootstrap probability of a edge is computed by summing BP(Ti, n') over the associated trees.

An example of the transformation $\boldsymbol{Y}^* = \boldsymbol{f}_n(\mathcal{X}^*_{n'})$ mentioned in section 3.4 is

$$\boldsymbol{Y}^* = \boldsymbol{V}_n^{-1/2}\boldsymbol{L}^*_{n'}, \tag{25}$$

where $\boldsymbol{L}^*_{n'} = (1/n')(\ell^*_1,\dots,\ell^*_{105})^T$ with $\ell^*_i = \ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$ and $\boldsymbol{V}_n$ is the variance matrix of $\boldsymbol{L}^*_n$. According to the approximation (23) and the central limit theorem, (13) holds well for sufficiently large n and n' with $m = 104$ and $\sigma^2 = n/n'$. It also follows from the above argument that $\mathrm{var}(\ell^*_i - \ell^*_j) \approx (n'/n)\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$, and thus the variance of log-likelihood difference is

$$\mathrm{var}\Big(\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) - \ell_j(\hat{\boldsymbol{\theta}}_j; \mathcal{X}_n)\Big) \approx \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2, \tag{26}$$

which gives another insight into the visualization of section 6.2, where the variance can be interpreted as the divergence between the two models; see Equation (27). This approximation holds well when the two predictive distributions $p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)$, $p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)$ are not very close to each other. When they are close to each other, however, the higher-order term ignored in (26) becomes dominant, and there is a difficulty for deriving the limiting distribution of the log-likelihood difference in the model selection test (Shimodaira, 1997; Schennach and Wilhelm, 2017).

6.2. Visualization of Probability Models

For representing the probability distribution of tree Ti, we define $\boldsymbol{\xi}_i := (\xi_{1i},\dots,\xi_{ni})^T \in \mathbb{R}^n$ from (22) for $i = 1,\dots,15$. The idea behind the visualization of **Figure 3** is that locations of $\boldsymbol{\xi}_i$ in $\mathbb{R}^n$ will represent locations of $p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)$ in the space of probability distributions. Let $D_{\mathrm{KL}}(p_i\|p_j)$ be the Kullback-Leibler divergence between the two distributions. For sufficiently small $(1/n)\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$, the squared distance in $\mathbb{R}^n$ approximates n times Jeffreys divergence

$$\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2 \approx n \times \Big(D_{\mathrm{KL}}(p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)\|p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)) + D_{\mathrm{KL}}(p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)\|p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)\Big) \tag{27}$$

for non-nested models (Shimodaira, 2001, section 6). When a model p_0 is nested in p_i, it becomes $\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_0\|^2 \approx 2n \times D_{\mathrm{KL}}(p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)\|p_0(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_0)) \approx 2 \times (\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) - \ell_0(\hat{\boldsymbol{\theta}}_0; \mathcal{X}_n))$. We explain three different visualizations of **Figure 7**. There are only minor differences between the plots, and the visualization is not sensitive to the details.

For dimensionality reduction, we have to specify the origin $\boldsymbol{c} \in \mathbb{R}^n$ and consider vectors $\boldsymbol{a}_i := \boldsymbol{\xi}_i - \boldsymbol{c}$. A naive choice would be the average $\boldsymbol{c} = \sum_{i=1}^{15} \boldsymbol{\xi}_i/15$. By applying PCA without centering and scaling (e.g., `prcomp` with option `center=FALSE, scale=FALSE` in R) to the matrix $(\boldsymbol{a}_1,\dots,\boldsymbol{a}_{15})$, we obtain the visualization of $\boldsymbol{\xi}_i$ as the axes (red arrows) of biplot in **Figure 7A**.

For computing the "data point" X in **Figure 3**, we need more models. Let tree T106 be the star topology with no internal branch (completely unresolved tree), and T107,..., T131 be partially resolved tree topologies with only one internal branch corresponding to E1,..., E25, whereas T1,..., T105 are fully resolved trees (bifurcating trees). Then define $\boldsymbol{\eta}_i := \boldsymbol{\xi}_{106+i}$, $i = 0,\dots,25$. Now we take $\boldsymbol{c} = \boldsymbol{\eta}_0$ for computing $\boldsymbol{a}_i = \boldsymbol{\xi}_i - \boldsymbol{\eta}_0$ and $\boldsymbol{b}_i = \boldsymbol{\eta}_i - \boldsymbol{\eta}_0$. There is hierarchy of models: $\boldsymbol{\eta}_0$ is the submodel nested in all the other models, and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3$, for example, are submodels of $\boldsymbol{\xi}_1$ (T1 includes E1, E2, E3). By combining these non-nested models, we can reconstruct a comprehensive model in which all the other models are nested as submodels (Shimodaira, 2001, Equation 10 in section 5). The idea is analogous to reconstructing the full model $y = \beta_1 x_1 + \cdots + \beta_{25} x_{25} + \epsilon$ of multiple regression from submodels $y = \beta_1 x_1 + \epsilon, \dots, y = \beta_{25} x_{25} + \epsilon$. Thus we call it as "full model" in this paper, and the ML estimate of the full model is indicated as the data point X; it is also said "super model" in Shimodaira and Hasegawa (2005). Let $\boldsymbol{B} = (\boldsymbol{b}_1,\dots,\boldsymbol{b}_{25}) \in \mathbb{R}^{n\times 25}$ and $\boldsymbol{d} = (\|\boldsymbol{b}_1\|^2,\dots,\|\boldsymbol{b}_{25}\|^2)^T \in \mathbb{R}^{25}$, then the vector for the full model is computed approximately by

$$\boldsymbol{a}_X = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{d}. \tag{28}$$

For the visualization of the best 15 trees, we may use only $\boldsymbol{b}_1,\dots,\boldsymbol{b}_{11}$, because they include E1 and two more edges from E2,...,E11. In **Figures 3**, **7B**, we actually modified the above computation slightly so that the star topology T106 is replaced by T107, the partially resolved tree corresponding to E1 (T107 is also said star topology by treating clade (23) as a leaf of the tree), and the 10 partially resolved trees for E2,..., E11 are replaced by those for (E1,E2),..., (E1,E11), respectively; the origin becomes the maximal model nested in all the 15 trees, and X becomes the minimal full model containing all the 15 trees. Just before applying PCA in **Figure 7B**, $\boldsymbol{a}_1,\dots,\boldsymbol{a}_{15}$ are projected to the space orthogonal to $\boldsymbol{a}_X$, so that the plot becomes the "top-view" of **Figure 3A** with $\boldsymbol{a}_X$ being at the origin.

In **Figure 7C**, we attempted a even simpler computation without using ML estimates for partially resolved trees. We used $\boldsymbol{B} = (\boldsymbol{a}_1,\dots,\boldsymbol{a}_{15})$ and $\boldsymbol{d} = (\|\boldsymbol{a}_1\|^2,\dots,\|\boldsymbol{a}_{15}\|^2)^T$, and taking the largest 10 singular values for computing the inverse in (28). The orthogonal projection to $\boldsymbol{a}_X$ is applied before PCA.

6.3. Asymptotic Theory of Smooth Surfaces

For expressing the shape of the region $\mathcal{R} \subset \mathbb{R}^{m+1}$, we use a local coordinate system $(\boldsymbol{u}, v) \in \mathbb{R}^{m+1}$ with $\boldsymbol{u} \in \mathbb{R}^m, v \in \mathbb{R}$. In a neighborhood of $\boldsymbol{y}$, the region is expressed as

$$\mathcal{R} = \{(\boldsymbol{u}, v) \mid v \le -h(\boldsymbol{u}),\ \boldsymbol{u} \in \mathbb{R}^m\}, \tag{29}$$

where h is a smooth function; see Shimodaira (2008) for the theory of non-smooth surfaces. The boundary surface $\partial\mathcal{R}$ is expressed as $v = -h(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^m$. We can choose the coordinates

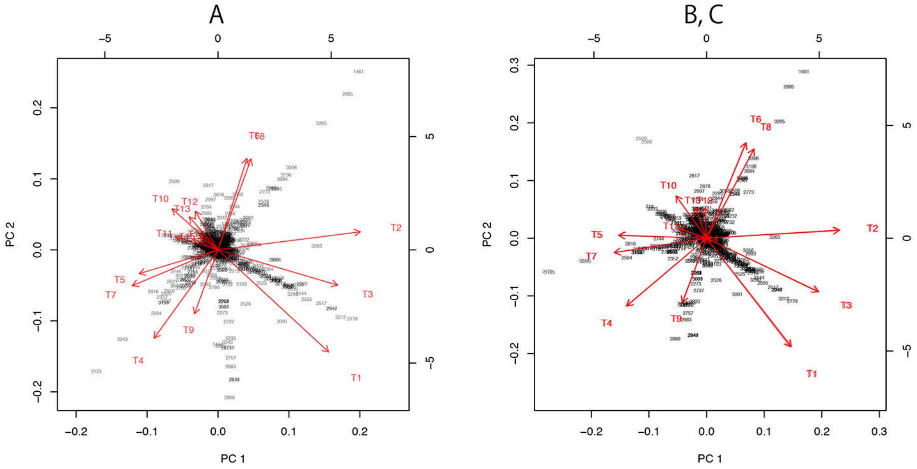


FIGURE 7 | Three versions the visualization of probability distributions for the best 15 trees drawn using different sets of models. **(A)** Only the 15 bifurcating trees. **(B)** 15 bifurcating trees + 10 partially resolved trees + 1 star topology. This is the same plot as **Figure 3B**. **(C)** 15 bifurcating trees + 1 star topology. Note that **(B,C)** are superimposed, since their plots are almost indistinguishable.

so that $\boldsymbol{y} = (\mathbf{0}, \beta_0)$ (i.e., $\boldsymbol{u} = (0, \dots, 0)$ and $v = \beta_0$), and $h(\mathbf{0}) = 0$, $\partial h/\partial u_i|_{\mathbf{0}} = 0$, $i = 1, \dots, m$. The projection now becomes the origin $\hat{\boldsymbol{\mu}} = (\mathbf{0}, 0)$, and the signed distance is β_0. The mean curvature of surface $\partial\mathcal{R}$ at $\hat{\boldsymbol{\mu}}$ is now defined as

$$\beta_1 = \frac{1}{2} \sum_{i=1}^{m} \left. \frac{\partial^2 h(\boldsymbol{u})}{\partial u_i \partial u_i} \right|_{\mathbf{0}}, \tag{30}$$

which is interpreted as the trace of the hessian matrix of h. When $\mathcal{R}$ is convex at least locally in the neighborhood, all the eigenvalues of the hessian are non-negative, leading to $\beta_1 \geq 0$, whereas concave $\mathcal{R}$ leads to $\beta_1 \leq 0$. In particular, $\beta_1 = 0$ when $\partial\mathcal{R}$ is flat (i.e., $h(\boldsymbol{u}) \equiv 0$).

Since the transformation $\boldsymbol{y} = f_n(\mathcal{X}_n)$ depends on n, the shape of the region $\mathcal{R}$ actually depends on n, although the dependency is implicit in the notation. As n goes larger, the standard deviation of estimates, in general, reduces at the rate $n^{-1/2}$. For keeping the variance constant in (4), we actually magnifying the space by the factor $n^{1/2}$, meaning that the boundary surface $\partial\mathcal{R}$ approaches flat as $n \to \infty$. More specifically, the magnitude of mean curvature is of order $\beta_1 = O_p(n^{-1/2})$. The magnitude of $\partial^3 h/\partial u_i \partial u_j \partial u_k$ and higher order derivatives is $O_p(n^{-1})$, and we ignore these terms in our asymptotic theory. For keeping $\boldsymbol{\mu} = O(1)$ in (4), we also consider the setting of "local alternatives," meaning that the parameter values approach a origin on the boundary at the rate $n^{-1/2}$.

6.4. Bridging the Problem of Regions to the *Z*-Test

Here we explain the problem of regions in terms of the z-test by bridging the multivariate problem of section 3 to the 1-dimensional case of section 1.

Ideal p-values are uniformly distributed over $p \in (0, 1)$ when the null hypothesis holds. In fact, $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) \sim U(0, 1)$ for $\boldsymbol{\mu} \in \partial\mathcal{R}$ as indicated in (11). The statistic $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y})$ may be called *pivotal* in the sense that the distribution does not change when $\boldsymbol{\mu} \in \partial\mathcal{R}$ moves on the surface. Here we ignore the error of $O_p(n^{-1})$, and consider only the second order asymptotic accuracy. From (10), we can write $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) \simeq \bar{\Phi}(\beta_0(\boldsymbol{Y}) - \beta_1(\boldsymbol{Y}))$, where the notation such as $\beta_0(\boldsymbol{Y})$ and $\beta_1(\boldsymbol{Y})$ indicates the dependency on $\boldsymbol{Y}$. Since $\beta_1(\boldsymbol{Y}) \simeq \beta_1(\boldsymbol{y}) = \beta_1$, we treat $\beta_1(\boldsymbol{Y})$ as a constant. Now we get the normal pivotal quantity (Efron, 1985) as $\bar{\Phi}^{-1}(\mathrm{AU}(\mathcal{R}|\boldsymbol{Y})) = \beta_0(\boldsymbol{Y}) - \beta_1 \sim N(0, 1)$ for $\boldsymbol{\mu} \in \partial\mathcal{R}$. More generally, it becomes

$$\beta_0(\boldsymbol{Y}) - \beta_1 \sim N(\beta_0(\boldsymbol{\mu}), 1), \quad \boldsymbol{\mu} \in \mathbb{R}^{m+1}. \tag{31}$$

Let us look at the z-test in section 1, and consider substitutions:

$$Z = \beta_0(\boldsymbol{Y}) - \beta_1, \quad \theta = \beta_0(\boldsymbol{\mu}), \quad c = -\beta_1. \tag{32}$$

The 1-dimensional model (1) is now equivalent to (31). The null hypothesis is also equivalent: $\theta \leq 0 \Leftrightarrow \beta_0(\boldsymbol{\mu}) \leq 0 \Leftrightarrow \boldsymbol{\mu} \in \mathcal{R}$. We can easily verify that AU corresponds to $p(z)$, because $p(z) = \bar{\Phi}(z) = \bar{\Phi}(\beta_0(\boldsymbol{y}) - \beta_1) \simeq \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$, which is expected from the way we obtained (31) above. Furthermore, we can derive SI from $p(z, c)$. First verify that the selection event is equivalent: $Z > c \Leftrightarrow \beta_0(\boldsymbol{Y}) - \beta_1 > -\beta_1 \Leftrightarrow \beta_0(\boldsymbol{Y}) > 0 \Leftrightarrow \boldsymbol{Y} \in \mathcal{R}^C$. Finally, we obtain SI as $p(z, c) = p(z)/\bar{\Phi}(c) \simeq \bar{\Phi}(\beta_0(\boldsymbol{y}) - \beta_1)/\bar{\Phi}(-\beta_1) \simeq \mathrm{SI}(\mathcal{R}|\boldsymbol{y})$.

6.5. Model Fitting in Multiscale Bootstrap

We have used thirteen σ^2 values from 1/9 to 9 (equally spaced in log-scale). This range is relatively large, and we observe a slight deviation from the linear model $\beta_0 + \beta_1\sigma^2$ in **Figure 6**. Therefore we fit other models to the observed values of ψ_{σ^2} as implemented in *scaleboot* package (Shimodaira, 2008). For example, poly.k model is $\sum_{i=0}^{k-1} \beta_i \sigma^{2i}$, and sing.3 model is $\beta_0 + \beta_1\sigma^2(1 + \beta_2(\sigma - 1))^{-1}$. In **Figure 6A**, poly.3 is the best model according to AIC

(Akaike, 1974). In **Figure 6B**, poly.2, poly.3, and sing.3 are combined by model averaging with Akaike weights. Then β_0 and β_1 are estimated from the tangent line to the fitted curve of ψ_{σ^2} at $\sigma^2 = 1$. In **Figure 6**, the tangent line is drawn as red line for extrapolating ψ_{σ^2} to $\sigma^2 = -1$. Shimodaira (2008) and Terada and Shimodaira (2017) considered the Taylor expansion of ψ_{σ^2} at $\sigma^2 = 1$ as a generalization of the tangent line for improving the accuracy of AU and SI.

In the implementation of *CONSEL* (Shimodaira and Hasegawa, 2001) and *pvclust* (Suzuki and Shimodaira, 2006), we use a narrower range of σ^2 values (ten σ^{-2} values: 0.5, 0.6, ..., 1.4). Only the linear model $\beta_0 + \beta_1\sigma^2$ is fitted there. The estimated β_0 and β_1 should be very close to those estimated from the tangent line described above. An advantage of using wider range of σ^2 in *scaleboot* is that the standard error of β_0 and β_1 will become smaller.

6.6. General Formula of Selective Inference

Let $\mathcal{H}, \mathcal{S} \subset \mathbb{R}^{m+1}$ be regions for the null hypothesis and the selection event, respectively. We would like to test the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{H}$ against the alternative $H_1 : \boldsymbol{\mu} \in \mathcal{H}^C$ conditioned on the selection event $\boldsymbol{y} \in \mathcal{S}$. We have considered the outside mode $\mathcal{H} = \mathcal{R}, \mathcal{S} = \mathcal{R}^C$ in (18) and the inside mode $\mathcal{H} = \mathcal{R}^C, \mathcal{S} = \mathcal{R}$ in (20). For a general case of $\mathcal{H}, \mathcal{S}$, Terada and Shimodaira (2017) gave a formula of approximately unbiased p-value of selective inference as

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) = \frac{\bar{\Phi}(\beta_0^{\mathcal{H}} - \beta_1^{\mathcal{H}})}{\bar{\Phi}(\beta_0^{\mathcal{S}} + \beta_0^{\mathcal{H}} - \beta_1^{\mathcal{H}})}, \tag{33}$$

where geometric quantities β_0, β_1 are defined for the regions $\mathcal{H}, \mathcal{S}$. We assumed that $\mathcal{H}$ and $\mathcal{S}^C$ are expressed as (29), and two surfaces $\partial\mathcal{H}, \partial\mathcal{S}$ are nearly parallel to each other with tangent planes differing only $O_p(n^{-1/2})$. The last assumption always holds for (18), because $\partial\mathcal{H} = \partial\mathcal{R}$ and $\partial\mathcal{S} = \partial\mathcal{R}^C$ are identical and of course parallel to each other.

Here we explain why we have considered the special case of $\mathcal{S} = \mathcal{H}^C$ for phylogenetic inference. First, we suppose that the selection event satisfies $\mathcal{S} \subset \mathcal{H}^C$, because a reasonable test would not reject H_0 unless $\boldsymbol{y} \in \mathcal{H}^C$. Note that $\boldsymbol{y} \in \mathcal{S} \subset \mathcal{H}^C$ implies $0 \le -\beta_0^{\mathcal{S}} \le \beta_0^{\mathcal{H}}$. Therefore, $\beta_0^{\mathcal{H}} + \beta_0^{\mathcal{S}} \ge 0$ leads to

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) \ge \mathrm{SI}(\mathcal{H}|\boldsymbol{y}), \tag{34}$$

where $\mathrm{SI}(\mathcal{H}|\boldsymbol{y}) := \mathrm{SI}(\mathcal{H}|\mathcal{H}^C, \boldsymbol{y})$ is obtained from (33) by letting $\beta_0^{\mathcal{H}} + \beta_0^{\mathcal{S}} = 0$ for $\mathcal{S} = \mathcal{H}^C$. The p-value $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$ becomes smaller as $\mathcal{S}$ grows, and $\mathcal{S} = \mathcal{H}^C$ gives the smallest p-value, leading to the most powerful selective test. Therefore the choice $\mathcal{S} = \mathcal{H}^C$ is preferable to any other choice of selection event satisfying $\mathcal{S} \subset \mathcal{H}^C$. This kind of property is mentioned in Fithian et al. (2014) as the monotonicity of selective error in the context of "data curving."

Let us see how these two p-values differ for the case of E2 by specifying $\mathcal{H} = \mathcal{R}_{\mathrm{E2}}^C$ and $\mathcal{S} = \mathcal{R}_{\mathrm{T1}}$. In this case, the two surfaces $\partial\mathcal{H}, \partial\mathcal{S}$ may not be very parallel to each other, thus violating the assumption of $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$, so we only intend to show the potential difference between the two p-values. The geometric quantities are $\beta_0^{\mathcal{H}} = -\beta_0^{\mathrm{E2}} = 1.59$, $\beta_1^{\mathcal{H}} = -\beta_1^{\mathrm{E2}} = -0.12$, $\beta_0^{\mathcal{S}} = \beta_0^{\mathrm{T1}} = -0.41$; the p-values are calculated using more decimal places than shown. SI of E2 conditioned on selecting T1 is

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) = \frac{\bar{\Phi}(1.59 + 0.12)}{\bar{\Phi}(-0.41 + 1.59 + 0.21)} = 0.448,$$

and it is very different from SI of E2 conditioned on selecting E2

$$\mathrm{SI}(\mathcal{H}|\boldsymbol{y}) = \frac{\bar{\Phi}(1.59 + 0.12)}{\bar{\Phi}(0.12)} = 0.097,$$

where $\mathrm{SI}'(\mathcal{R}_{\mathrm{E2}}^C|\boldsymbol{y}) = 1 - \mathrm{SI}(\mathcal{R}_{\mathrm{E2}}^C|\boldsymbol{y}) = 0.903$ is shown in **Table 2**. As you see, $\mathrm{SI}(\mathcal{H}|\boldsymbol{y})$ is easier to reject H_0 than $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$.

6.7. Number of Regions for Phylogenetic Inference

The regions $\mathcal{R}_i$, $i = 1, \dots, K_{\mathrm{all}}$ correspond to trees or edges. In inside and outside modes, the number of total regions is $K_{\mathrm{all}} = 105$ for trees and $K_{\mathrm{all}} = 25$ for edges when the number of taxa is $N = 6$. For general $N \ge 3$, they grow rapidly as $K_{\mathrm{all}} = (2N-5)!/(2^{N-3}(N-3)!)$ for trees and $K_{\mathrm{all}} = 2^{N-1} - (N+1)$ for edges. Next consider the number of selected regions K_{select}. In inside mode, regions with $\boldsymbol{y} \in \mathcal{R}_i$ are selected, and the number is counted as $K_{\mathrm{select}} = 1$ for trees and $K_{\mathrm{select}} = N - 3 = 3$ for edges. In outside mode, regions with $\boldsymbol{y} \notin \mathcal{R}_i$ are selected, and thus the number is K_{all} minus that for inside mode; $K_{\mathrm{select}} = K_{\mathrm{all}} - 1 = 104$ for trees and $K_{\mathrm{select}} = K_{\mathrm{all}} - (N-3) = 22$ for edges. Finally, consider the number of true null hypotheses, denoted as K_{true}. The null hypothesis holds true when $\boldsymbol{\mu} \notin \mathcal{R}_i$ in inside mode and $\boldsymbol{\mu} \in \mathcal{R}_i$ in outside mode, and thus K_{true} is the same as the number of regions with $\boldsymbol{y} \notin \mathcal{R}_i$ in inside mode and $\boldsymbol{y} \in \mathcal{R}_i$ in outside mode (These numbers do not depend on the value of $\boldsymbol{y}$ by ignoring the case of $\boldsymbol{y} \in \partial\mathcal{R}_i$). Therefore, $K_{\mathrm{true}} = K_{\mathrm{all}} - K_{\mathrm{select}}$ for both cases.

6.8. Selective Inference of Lasso Regression

Selective inference is considered for the variable selection of regression analysis. Here, we deal with prostate cancer data (Stamey et al., 1989) in which we predict the level of prostate-specific antigen (PSA) from clinical measures. The dataset is available in the R package *ElemStatLearn* (Halvorsen, 2015). We consider a linear model to the log of PSA (`lpsa`), with 8 predictors such as the log prostate weight (`lweight`), `age`, and so on. All the variables are standardized to have zero mean and unit variance.

The goal is to provide the valid selective inference for the partial regression coefficients of the selected variables by lasso (Tibshirani, 1996). Let n and p be the number of observations and the number of predictors. $\hat{\boldsymbol{M}}$ is the set of selected variables, and $\hat{\boldsymbol{s}}$ represents the signs of the selected regression coefficients. We suppose that regression responses are distributed as $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \tau^2 \boldsymbol{I}_n)$ where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\tau > 0$. Let e_i be the ith residual. Resampling the scaled residuals σe_i $(i = 1, \dots, n)$ with several values of scale σ^2, we can apply the multiscale bootstrap method described in section 4 for the selective inference in the regression problem. Here, we note that the target of the inference is the true partial regression coefficients:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\mu},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix. We compute four types of intervals with confidence level $1-\alpha = 0.95$ for selected variable j. $[L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]$ is the non-selective confidence interval obtained via t-distribution. $[L_j^{\text{model}}, U_j^{\text{model}}]$ is the selective confidence interval under the selected model proposed by Lee et al. (2016) and Tibshirani et al. (2016), which is computed by `fixedLassoInf` with `type="full"` in R package *selectiveInference* (Tibshirani et al., 2017). By extending the method of $[L_j^{\text{model}}, U_j^{\text{model}}]$, we also computed $[L_j^{\text{variable}}, U_j^{\text{variable}}]$, which is the selective confidence interval under the selection event that variable j is selected. These three confidence intervals are exact, in the sense that

$$P\left(\beta_j \in [L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]\right) = 1-\alpha,$$
$$P\left(\beta_j \in [L_j^{\text{model}}, U_j^{\text{model}}] \mid \hat{\boldsymbol{M}}, \hat{\boldsymbol{s}}\right) = 1-\alpha,$$
$$P\left(\beta_j \in [L_j^{\text{variable}}, U_j^{\text{variable}}] \mid j \in \hat{\boldsymbol{M}}, \hat{s}_j\right) = 1-\alpha.$$

Note that the selection event of variable j, i.e., $\{j \in \hat{\boldsymbol{M}}, \hat{s}_j\}$ can be represented as a union of polyhedra on $\mathbb{R}^n$, and thus, according to the polyhedral lemma (Lee et al., 2016; Tibshirani et al., 2016), we can compute a valid confidence interval $[L_j^{\text{variable}}, U_j^{\text{variable}}]$. However, this computation is prohibitive for $p > 10$, because all the possible combinations of models with variable j are considered. Therefore, we compute its approximation $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ by the multiscale bootstrap method of section 4 with much faster computation even for larger p.

We set $\lambda = 10$ as the penalty parameter of lasso, and the following model and signs were selected:

$$\hat{\boldsymbol{M}} = \{\texttt{lcavol}, \texttt{lweight}, \texttt{lbph}, \texttt{svi}, \texttt{pgg45}\},$$
$$\hat{\boldsymbol{s}} = (+,+,+,+,+).$$

The confidence intervals are shown in **Figure 1**. For adjusting the selection bias, the three confidence intervals of selective inference are longer than the ordinary confidence interval. Comparing $[L_j^{\text{model}}, U_j^{\text{model}}]$ and $[L_j^{\text{variable}}, U_j^{\text{variable}}]$, the latter is shorter, and would be preferable. This is because the selection event of the latter is less restrictive as $\{\hat{\boldsymbol{M}}, \hat{\boldsymbol{s}}\} \subseteq \{j \in \hat{\boldsymbol{M}}, \hat{s}_j\}$; see section 6.6 for the reason why larger selection event is better. Finally, we verify that $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ approximates $[L_j^{\text{variable}}, U_j^{\text{variable}}]$ very well.

AUTHOR CONTRIBUTIONS

HS and YT developed the theory of selective inference. HS programmed the multiscale bootstrap software and conducted the phylogenetic analysis. YT conducted the lasso analysis. HS wrote the manuscript. All authors have approved the final version of the manuscript.

FUNDING

This research was supported in part by JSPS KAKENHI Grant (16H02789 to HS, 16K16024 to YT).

ACKNOWLEDGMENTS

The authors appreciate the feedback from the audience of seminar talk of HS at Department of Statistics, Stanford University. The authors are grateful to Masami Hasegawa for his insightful comments on phylogenetic analysis of mammal species.

REFERENCES

Adachi, J., and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42, 459–468. doi: 10.1007/BF02498640

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Cont.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Amari, S.-I., and Nagaoka, H. (2007). *Methods of Information Geometry, Translations of Mathematical Monographs*, Vol. 191. Providence, RI: American Mathematical Society.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edn*. New York, NY: Springer-Verlag.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Stat. Soc. Ser. B (Methodol.).* 24, 406–424. doi: 10.1111/j.2517-6161.1962.tb00468.x

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552

Efron, B. (1984). Comparing non-nested linear models. *J. Am. Stat. Assoc.* 79, 791–803. doi: 10.1080/01621459.1984.10477096

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* 72, 45–58. doi: 10.1093/biomet/72.1.45

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13429–13434. doi: 10.1073/pnas.93.23.13429

Efron, B., and Tibshirani, R. (1998). The problem of regions. *Ann. Sta.* 26, 1687–1718. doi: 10.1214/aos/1024691353

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi: 10.1007/BF01734359

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv:1410.2597.*

Graur, D., Duret, L., and Gouy, M. (1996). Phylogenetic position of the order lagomorpha (rabbits, hares and allies). *Nature* 379:333. doi: 10.1038/379333a0

Halanych, K. M. (1998). Lagomorphs misplaced by more characters and fewer taxa. *Syst. Biol.* 47, 138–146. doi: 10.1080/106351598261085

Halvorsen, K. (2015). *ElemStatLearn: data sets, functions and examples from the book: "the elements of statistical learning, data mining, inference, and prediction" by trevor hastie, robert tibshirani and jerome friedman*. R package. Available online at: https://CRAN.R-project.org/package=ElemStatLearn

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29, 170–179. doi: 10.1007/BF02100115

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 30, 151–160. doi: 10.1007/BF02109483

Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York, NY: Springer Science & Business Media. doi: 10.1007/978-0-387-71887-3

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Stat.* 44, 907–927. doi: 10.1214/15-AOS1371

Linhart, H. (1988). A test whether two AIC's differ significantly. *South Afr. Stat. J.* 22, 153–161.

Novacek, M. J. (1992). Mammalian phytogeny: shaking the tree. *Nature* 356, 121–125. doi: 10.1038/356121a0

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808. doi: 10.1080/10635150490522304

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Schennach, S. M., and Wilhelm, D. (2017). A simple parametric model selection test. *J. Am. Stat. Assoc.* 112, 1663–1674. doi: 10.1080/01621459.2016.1224716

Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Ann. Inst. Stat. Math.* 49, 395–410. doi: 10.1023/A:1003140609666

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Inst. Stat. Math.* 50, 1–13. doi: 10.1023/A:1003483128844

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* 90, 227–244. doi: 10.1016/S0378-3758(00)00115-4

Shimodaira, H. (2001). Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Commun. Stat. Theory Methods* 30, 1751–1772. doi: 10.1081/STA-100105696

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508. doi: 10.1080/10635150290069913

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.* 32, 2616–2641. doi: 10.1214/009053604000000823

Shimodaira, H. (2008). Testing regions with nonsmooth boundaries via multiscale bootstrap. *J. Stat. Plan. Inference* 138, 1227–1241. doi: 10.1016/j.jspi.2007.04.001

Shimodaira, H. (2019). *Scaleboot: Approximately Unbiased p-Values via Multiscale Bootstrap*. R package version 1.0-0. Available online at: https://CRAN.R-project.org/package=scaleboot

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201

Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247. doi: 10.1093/bioinformatics/17.12.1246

Shimodaira, H., and Hasegawa, M. (2005). "Assessing the uncertainty in phylogenetic inference," in *Statistical Methods in Molecular Evolution*, Statistics for Biology and Health, ed R. Nielsen (New York, NY: Springer), 463–493. doi: 10.1007/0-387-27733-1_17

Shimodaira, H., and Maeda, H. (2018). An information criterion for model selection with missing data via complete-data divergence. *Ann. Inst. Stat. Math.* 70, 421–438. doi: 10.1007/s10463-016-0592-7

Stamey, T., Kabalin, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. Radical prostatectomy treted patients. *J. Urol.* 16, 1076–1083. doi: 10.1016/S0022-5347(17)41175-X

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542. doi: 10.1093/bioinformatics/btl117

Taylor, J., and Tibshirani, R. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7629–7634. doi: 10.1073/pnas.1507583112

Terada, Y., and Shimodaira, H. (2017). Selective inference for the problem of regions via multiscale bootstrap. *arXiv:1711.00949*.

Tian, X., and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Stat.* 46, 679–710. doi: 10.1214/17-AOS1564

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.).* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Stat. Assoc.* 111, 600–620. doi: 10.1080/01621459.2015.1108848

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. (2017). *SelectiveInference: Tools for Post-Selection Inference*. R package version 1.2.4. Available online at: https://CRAN.R-project.org/package=selectiveInference

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi: 10.2307/1912557

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372. doi: 10.1016/0169-5347(96)10041-0

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

PERSPECTIVE
published: 09 August 2019
doi: 10.3389/fevo.2019.00306

Evidential Statistics in Model and Theory Development

Samuel M. Scheiner[1]* *and Robert D. Holt*[2]

[1] *Division of Environmental Biology, National Science Foundation, Arlington, VA, United States,* [2] *Department of Biology, University of Florida, Gainesville, FL, United States*

Evidential statistics is an important advance in model and theory testing, and scientific reasoning in general, combining and extending key insights from other philosophies of statistics. A key desiderata in evidential statistics is the rigorous and objective comparison of alternative models against data. Scientific theories help to define the range of models which are brought to bear in any such assessment, including both tried and trusted models and risky novel models; such theories emerge from a kind of evolutionary process of repeated model assessment, where model selection is akin to natural selection acting both on the standing crop of genetic variation, and on novel mutations. The careful use of evidential statistics could play an important and as yet to be fulfilled role in the future development of scientific theories. We illustrate these ideas using examples from ecology and evolutionary biology.

Keywords: abduction, deduction, evidential statistics, induction, model, theory

Edited by:
Mark Louis Taper,
Montana State University System,
United States

Reviewed by:
Brian Dennis,
University of Idaho, United States
Jeff E. Houlahan,
University of New Brunswick Fredericton, Canada

****Correspondence:***
Samuel M. Scheiner
sscheine@nsf.gov

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *18 March 2019*
Accepted: *30 July 2019*
Published: *09 August 2019*

Citation:
Scheiner SM and Holt RD (2019) Evidential Statistics in Model and Theory Development. Front. Ecol. Evol. 7:306. doi: 10.3389/fevo.2019.00306

INTRODUCTION

Statistical inference aims at relating models to data and the empirical world, whether that model deals with an issue as simple as estimating the mean of a population or as complex as predicting millennial-scale changes in the global climate. There have been decades-long debates about the best way to make inferences (e.g., Neyman-Pearson error statistics vs. Bayesian approaches). This special feature highlights the approach called "evidential statistics," (Taper and Ponciano, 2016) which synthesizes prior approaches—error statistics, Bayesian statistics, information-based model selection, and likelihood approaches—to squarely focus on the comparative ability of alternative models or hypotheses for explaining an observed dataset. This approach to inference was sparked by Royall (1997) and Lele (2004), and the articles in this Special Issue highlight the rapid emergence and maturation of evidential statistics. We heartily concur with the value of such a synthesis of prior approaches, and the explicit emphasis on comparisons among alternative hypotheses or models as an essential component of scientific progress. Neither of us are card-carrying statisticians or philosophers of science; instead we are scientists interested in the conceptual basis of our discipline. Here we reflect on the need for intellectual flexibility by considering the role of statistical inference as a formal, mathematical procedure for refereeing the relationship between data, models, and theories, and place that in the context of the wider set of processes that scientists might use for theory development.

Scientists quest to obtain knowledge about the empirical world so as to understand its causal structure, and to use that causal structure for prediction as well as control and management. The inferential procedures employed to gain such knowledge should be "truth-tropic" (Lipton, 2004, p. 7). There are philosophers (e.g., Laudan, 1981) who reject the notion that science involves a kind of convergence toward an understanding of how nature works (conceived broadly), but we feel that most working scientists assume (or at least hope) that they are engaged in a "truth-tracking" enterprise (Roush, 2007). While models are the direct connection between data and specific conclusions drawn from those data, those models are embedded within larger conceptual frameworks, typically called theories. One role of theory is to help guide the creative formulation of novel models for comparison against any set of data. For example, we might construct a family of ecological niche models (ENMs, Holt, 2009; Peterson et al., 2011) to explain why saguaro cacti (*Carnegiea gigantea*) are common in parts of the Sonoran Desert, yet absent elsewhere with seemingly comparable climates. Those models would be embedded within, and get their warrant from, broader theories of ecology and evolution (Scheiner and Willig, 2011b, Scheiner and Mindell, 2019). The models might draw upon diverse data and models such as the physiology of plants with Crassulacean Acid Metabolism as their mode of photosynthesis, the geographic history of North America, and the phylogeny of the Cactaceae. A criterion for selecting among alternative ENMs might be the minimization of errors in predicting known occurrences from available distributional and environmental data.

Statistics is essential for testing models in the broad sense, examining their relationship with the empirical world, efforts that in turn contribute to the goal of crafting and testing more general theories. Building and testing theories relies on a variety of approaches, only some of which make explicit use of statistical inference. Evidential statistics aims at providing a systematic approach for assessing the relative informativeness of models, which depends upon available data and protocols—distinct from the personal beliefs embedded within Bayesian statistics—via objective metrics of evidence that ideally lead toward closer approximations of the "truth" as models continue to be refined and compared (Dennis et al., 2019). Theories are distillations of conclusions (Tukey, 1960) achieved collectively by scientists, carrying out such protocols repeatedly and objectively. Kuhn (1977, pp. 321–322) notes that the development of scientific theories must juggle qualities which at times may be contradictory, such as accuracy, consistency, simplicity, fruitfulness, and scope, to which Houlahan et al. (2017) add as an essential desideratum the successful prediction of novel states of the world.

Like any evolutionary process, theory development depends upon the availability of an array of alternative models for comparison, using both a standing crop of existing models that have proven useful in other contexts, and novel conceptual mutations. Evidential procedures are akin to natural selection culling genetic variants, favoring the fittest in the population at hand in a given environment. For example, in our saguaro cactus model, general climatic variables such as average rainfall or seasonal patterns in precipitation are doubtless important and would discriminate among many models, but a key idiosyncratic factor operating at the northern range limits appears to be the number of consecutive hours below freezing (MacArthur, 1972, p. 127), which can be strongly influenced by local topography. The fittest of the competing models would surely need to include this key observation.

This evolutionary perspective on theory development stems back to Popper (1972, p. 261) who states, "[T]he growth of our knowledge is the result of a process closely resembling what Darwin called 'natural selection,' that is, *the natural selection of hypotheses.*" In a sense, likelihood and related quantitative approaches provide fitness metrics for selecting some hypotheses over others based on evidence. Just as natural selection does not comprise all of evolution, knowledge development leading up to a general theory is more than just the accumulation of episodes of such evidence-based selection. Other processes, such as intellectual coherence, the generation of novel ideas, and the infusion of ideas across disciplinary boundaries, play roles comparable to mutation, gene flow, and recombination. A particular challenge is to articulate how the scientific community builds larger arenas of knowledge—theories—from more specific models grounded in evidence. Popper (1972, p. 262-3) suggests a kind of inverse evolutionary tree of knowledge emerging over time: "[T]he tree of knowledge [springs up] from countless roots which grow up into the air rather than down, and which ultimately, high up, tend to unite into one common stem." We now turn our attention to the relationship between models and theories, broadly conceived.

FROM MODELS TO THEORIES AND BACK AGAIN

Our approach to models and theories can be considered part of the Pragmatic View of the structure of scientific theory (Winther, 2012, 2015). The Pragmatic View combines formal components of mathematic axioms and associated models with less formal, non-mathematical components including concepts, metaphor, narrative, and analogy. The result is a pluralistic and pragmatic structure for scientific theory in which theory content is organized according to the research questions being asked (Love, 2010). Vandermeer (2018), in an encomium to Richard Levins, cogently remarks on why in biology, theory is not just a compilation of models: "Populations of organisms only approximately follow precise equations and theories about them thus cannot rely exclusively on models… [and] [m]athematical forms of models are tools, as Levins repeatedly expressed, 'to educate the intuition.'"

Scheiner and Willig (2008) proposed a hierarchical framework for organizing theories consisting of general theories, more narrow constitutive theories, and even more specific models. The three types of theories have different functions. General theories provide the conceptual framework within which theories and models are built and tested. They consist of a set of general principles—confirmed generalizations—that provide background assumptions. These principles may appear trivial, but that is only because they have been so thoroughly tested that they have become embedded in our background knowledge. Yet, they are often ignored when building models. For example, one of the general principles of the theory of ecology is that "Variation in the characteristics of organisms results in heterogeneity of ecological patterns and processes" (Scheiner and Willig, 2011a). It is a reminder that even though very many ecological models assume that all individuals within a species are identical, we know that this is an approximation. While violations of this assumption may not substantially change model predictions in some situations, in other cases relaxing this assumption even by a small amount can lead to marked changes (e.g., Kendall and Fox, 2003). The constitutive theories and models are not derived formally from general theories. Rather, general theories provide the background knowledge and general conceptual framework within which more specific theories and models are built. For more on this conceptualization of a theory hierarchy (the inverse knowledge tree of Popper, 1972), see Scheiner (2010) and Mindell and Scheiner (2019).

Constitutive theories are the workhorses in this framework and what most individuals would think of when asked to name or describe a theory. Their role is to organize models into larger entities. They consist of a set of propositions, which might arise inductively from a set of models (e.g., a constitutive theory of diversity gradients, Scheiner and Willig, 2005). Alternatively, the propositions might be conceived first and then used to guide model development (e.g., the theory of natural selection, Frank and Fox, 2019). For example, enemy-victim theory (Holt, 2011) includes, among others, three propositions: (1) The increased consumption generated by increased victim abundance in turn fuels an increase in the per capita growth rate (fitness) of the natural enemy population. (2) An increase in the victim population increases the rate of consumption by each individual natural enemy. (3) Consumption by the natural enemy implies mortality in the victim. Making simplifying assumptions about the functional forms for each of these (which in turn reflect models and theories about the component processes), along with ancillary assumptions (e.g., no direct density dependence), these propositions can be formalized as the classical Lotka-Volterra predator-prey model:

$$\frac{dP}{dt} = P[baN - m]$$
$$\frac{dN}{dt} = N[r - aP]$$

where P and N are the densities of predators and prey, respectively, the predator birth rate is given by baN, where a is the attack rate and b is the rate that prey biomass is converted into offspring, m is the predator death rate, and r is the prey birth rate. This model is just one particular instantiation of those propositions; many other versions are possible. These models then serve to link theories to data, which is where evidential statistics comes into play.

The framework is multilayered, and both general and constitutive theories can be nested and overlapping. For example, a model of the evolution of plasticity of *Drosophila melanogaster* body size in response to temperature is embedded within a constitutive theory of the evolution of phenotypic plasticity that draws upon the constitutive theory of evolution by natural selection, both in turn embedded within the theory of evolution (Scheiner, 2019), while also drawing upon constitutive theories within the theory of organisms (Zamer and Scheiner, 2014). Some of these constitutive theories include formalized mathematical models, but others do not.

Models can be both qualitative and quantitative in describing or predicting nature. In ecology and evolution we tend to think of dynamical mathematical models, systems of equations or computer rules linked by logical operators corresponding to assumptions about mechanisms at and across different levels of biological organization. A computer simulation, such as an individual-based model of population dynamics, might be an example. Models can also be qualitative; Charles Darwin's theory of evolution was almost entirely verbal and qualitative. There is a single, iconic tree-like figure in *On the Origin of Species* which displays the grand, overarching vision of a shared origin for all life in an instantly transparent manner—an elegant example of a graphical, non-mathematical model.

From models we deductively derive hypotheses that in turn make predictions. These predictions are often derived from a mathematical model, which are based on some expected distribution of parameter values (see other articles in this special feature). Those distributions are then compared to data (broadly defined). Whereas the model is general in the sense that it applies across a domain of interest, a hypothesis becomes a prediction when applied to a specific, empirical instance. That application, the collision of models and data, is where evidential statistics steps in.

THE RELATIONSHIP OF EVIDENTIAL STATISTICS TO MODELS AND CONSTITUTIVE THEORIES

Statistical methods shed light on the possible relative verisimilitude or falsity of a hypothesis, compared to coherently-specified alternative hypotheses. That hypothesis might be that a model parameter has a very specific value (e.g., in plant populations the relationship between the average mass per individual and the density of survivors should have a exponent of −3/2, Yoda et al., 1963), or it could be more general (e.g., the relationship between productivity and diversity is hump-shaped, VanderMeulen et al., 2001), or it could be qualitative (e.g., the mating system in this particular plant population will be gynodioecy). By inference, if the hypothesis is false then the model is inadequate in the sense that compared to some alternative model, the model in question does not correspond to the empirical world. The history of science is littered with failed models and hypotheses (e.g., phlogiston, the ether, epicycles, barnacles as larval stages of barnacle geese), and many scientific advances prove to be way stations toward a deeper understanding of the world (e.g., Newton's gravitational theory). But statistics does not have the same role (at least not so obviously) when it comes to constitutive or general theories.

Those theories are systems that organize models, data, concepts, and so forth [Box 3.2 in Pickett et al. (2007) describes the components of theories]. Considered as an organizational system, constitutive and general theories are never true or false. Rather, they are useful, not useful, or poorly structured, that is, conceptually fruitful or not. That is not to say that general theories (e.g., the theory of evolution) are not true; rather that the strength of the theory lies in the overall validity of its components, rather than a single assessment of the entire theory.

Within constitutive theories are families of models, and decisions need to be made as to which models to include or exclude. Sometimes that decision-making process is how well one model mirrors the empirical world relative to another model. Evidence based on the relationship of a hypothesis with data and the empirical world leads to inferences about the relative truth or falsity of the hypotheses generated by each model, a decision-making process mediated by statistics. But these decisions are only part of what goes into conclusions about the utility of a constitutive or general theory. A principle in a general theory (e.g., "The ecological properties of species are the result of evolution" from the theory of ecology, Scheiner and Willig, 2011a) comes from the accumulation of a multitude of individual observations and models. An individual model can be discarded without negating the more general theory. We might decide that a natural selection model of the frequency of third position codons in DNA is inapplicable, because third position codons evolve by drift (Kimura, 1968). That conclusion would not affect the status of the theory of evolution by natural selection.

Evaluating models, such as the predator-prey model given above, involves more than just comparing predictions with data. That model famously predicts predator-prey cycles, looking in some respects like real-world cycles (such as the lynx-snowshoe cycle of Canada). May (1973) pointed out, however, that these models are neutrally stable, and so are highly unlikely to describe real cycles that are persistent. Indeed, the model is structurally unstable, in that small deviations in model assumptions lead either to oscillations that blow up, or to a stable equilibrium. Structural stability should be a desideratum in all our model and theory construction. Yet, real organisms and communities are unlikely to exactly match any set of equations we are likely to concoct.

Models may be false, while still playing a vital role in the conceptual framework of ecological theory. We contrast structural stability (the robustness of model conclusions to small deviations in model assumptions) with the stability of model structure. For the predator-prey model, the essential structure of the model itself (a +/– interaction between two antagonists, a natural enemy and its victim) is applicable across many empirical systems (e.g., predator-prey, host-pathogen, and plant-herbivore). The Lotka-Volterra predator-prey model demonstrates that there is a tendency to oscillate inherent in such antagonistic interactions. This qualitative conclusion is robust across many variants of this basic model, although the details may differ (e.g., the oscillations may manifest as transients following a perturbation, rather than as permanent cycles). Because the Lotka-Volterra model makes such robust, qualitative predictions, it continues to play an important role in the conceptual framework of theoretical ecology, even though it is known to be literally false for all empirical predator-prey systems. The same can be said of the model of exponential growth, $dN/dt = rN$, where N is population size and r is the intrinsic rate of increase. It has been argued that the principle of exponential growth is one of the conceptual foundations of ecology (Pásztor et al., 2016), and Ginzburg and Colyvan (2004) state that "the whole body of the spectacularly successful evolutionary theory has Malthusian growth in its foundation." Yet essentially no populations, when examined closely, match this model—there are always age and stage structure effects, demographic and environmental stochasticity, genetic variation, spatial dynamics, and density-dependent feedbacks, at play. This sweeping generalization, however, does not vitiate the conceptual role of exponential growth as foundational in our discipline. In like manner Queller (2017), commenting on Ronald Fisher's fundamental theorem of evolution, notes that it leaves out many important drivers of evolutionary change, but nonetheless "demonstrate[s] the general value of simplifying and sacrificing a bit of accuracy in order to capture and highlight fundamental issues in a simple and elegant way." This highlighting is an essential role of theory—enhancing understanding.

If a theory is relatively narrow, encompassing just one or a few specific models, and all of those models fail, we would then discard the theory as not useful. For example, Arditi and Ginzburg (2012) argued that we should discard any theory of predation in which the rate that predators attack prey depends only on prey density but not predator density, as in the above Lotka-Volterra model, which illustrates what is called "prey-dependence." They compiled case studies that included formal estimates of a key parameter (m) measuring the strength of predator interference on foraging rates. While they did not do a formal meta-analysis of those estimates, if they had, statistical inference would likely have supported the conclusion that this effect of predator density needs to be incorporated into any predatory-prey model (but see Abrams, 2015). There is a large body of food web and network theory that simply assumes prey-dependence in trophic linkages (i.e., ignores predator density). It is not yet known if altering this assumption would merely tweak the rich body of conclusions drawn from this theory, or instead if the change would have revolutionary effects on ecological understanding.

The use of statistics to assess hypotheses and models involves both deductive and inductive reasoning. We deduce hypotheses/predictions from a model. If a prediction proves false, one or more aspects of the model may be concluded to be false, which is the basis of Popper's (1959) falsifiability criterion for scientific theories. We also use statistics as a form of inductive reasoning. With induction, we infer a general conclusion from particular instances. When we estimate a population parameter from an observed set of data (e.g., the mean weight of a population of *Drosophila melanogaster*), we are performing induction. A constitutive or general theory includes a set of confirmed generalizations—condensations and abstractions, ultimately, from a body of facts—that may include parameter estimates (e.g., the base-pair mutation rate), used in particular model comparisons. Evidential statistics (Taper and Ponciano, 2016) is based upon rigorous comparisons of the likelihood (broadly conceived) of two or more alternative models. But it does not specify where the set of alternative models come from in the first place. This is where constitutive and general theories come into play—representing a kind of closet collective Bayesianism, where the cumulative wisdom of scientists over time help define the range of models that are likely to be assessed against any given dataset (Longino, 2002), as well as providing a structure for the creation of novel models.

A third, less familiar, type of reasoning is abduction. The term was coined by Charles Peirce (Douven, 2017), who used it initially to encompass hypothesis generation, but later in a manner related to the idea of "inference to the best explanation." The basic notion is that one compares alternative models and accepts the one that best explains the evidence. What counts as "best" could be its likelihood (in the sense used in evidential statistics as articulated by the other papers in this special feature), but also can involve desiderata such as simplicity, unification across studies, structural stability, and so forth (Lipton, 2004). Many of these ideas about how one can build up from models to more general theories can be traced to Whewell's (1858) three criteria for theory confirmation: prediction, consilience (explaining phenomena of a different kind than those used to formulate the theory), and coherence (the simplification or unification of different phenomena without the need for *ad hoc* modification of the theory) (Forster and Wolfe, 1999; Snyder, 2019). Norton (in prep, https://www.pitt.edu/~jdnorton/papers/material_theory/9.%20Best%20Explanation%20Examples.pdf) argues that Darwin's entire theory (as expressed in *On the Origin of Species*) involves an extended inference to the best explanation, all without explicit statistical inference. To our knowledge, no philosopher of science has yet brought together the notion of inference to the best explanation, and the complementary but distinct concepts of confirmation and evidence articulated by Bandyopadhya et al. (2016). Mark Taper (pers. comm.) notes that one virtue of evidential statistics is that one keeps track not just of the "best" model, but other models that might prove useful in future investigations. Evidential statistics provides a clear path for comparing models against particular datasets; what is now needed is an articulation of higher-order protocols for assessing constitutive and general theories. Such protocols are presumably at play when a community of scientists converge on particular ways of understanding the world. The bridge from models to more general theories may be more loosely constructed in biology than in, say, quantum physics. As Vandermeer (2018, p. 4) cogently notes, "[In population biology] any model is only approximate with respect to the theory it intends to represent, and any theory is bolstered by its conformation, even if approximate, to multiple models."

The development of constitutive and general theories cannot be entirely shoe-horned into formal statistical inference, including evidential statistics, vital though that is for sifting hypotheses and models. Statistical inference alone is insufficient when dealing with the sculpting over time of scientific understanding, involving the concerted efforts of many scientific minds who collectively craft complex models or theories (Longino, 2002). The total weight of the evidence that bears on theory development includes not just the quantification of specific estimated parameters, or alternative functional forms of models, but also reflects our confidence in the logical structure and explanatory scope of the models that are derived from a constitutive theory, and whether the domain of that theory encompasses the specific instances under consideration. In some sense, constitutive and general theories rely upon a higher order of evidential support and logical considerations that may lie outside the specific scope of any given dataset. For example, when examining a particular trait, such as emergence of blindness in a cave fish in Kentucky, should our models invoke only natural selection, or also the accumulation of deleterious mutations and genetic drift? The answer to this question would likely depend on what has been learned about other cave fish worldwide. Taper and Ponciano (2016) use Gause's (1934) famed protozoan experiments to compare the relative evidentiary power of a suite of population dynamic models, such as the Ricker, Beverton-Holt, and Gompertz equations. Choosing this suite of models for comparison, and excluding others, implicitly involves a priori beliefs about the relevant drivers of population dynamics, presumably drawing on correspondences between this concrete empirical system and a wide array of somehow comparable systems, as well as more specific assumptions, such as: there is no spontaneous generation, the populations are closed to immigration and emigration so that local births and deaths entirely drive dynamics (this is ensured by the experimental setup), there are no time-lags in density dependence (which might occur with the buildup of toxins or waste products, or subtle stage-structure effects), and there are no hidden players such as viruses. These background assumptions help define the range of models to be compared explicitly, using the metrics of evidentiary statistics.

What is the role of evidential statistics in determining the relationship between models and theories where the latter are qualitative, rather than quantitative? For example, our explanation about the range of saguaro cacti includes information about the geographic history of the North and South American continents. We have models of the movements of the continents over geological history, but those models are not mathematical equations. Rather, we have inferred that history from a range of observations, only some of which include quantitative models. In modern systematics, a phylogeny is a quantitative model of a set of relationships among species (or higher taxa) in a clade. When multiple phylogenies are overlain on a map, the subsequent qualitative biogeographic patterns can be used to make inferences about the geological history of that region. It is possible to devise a formal inference process for making decisions about that history, but a formal process is not always necessary. Wegener's (1966) theory of continental drift was based, in part, on observing close phylogenetic relationships between South American and African species, as well as the fit of the shapes of the continents themselves. This process of bringing together models from multiple domains that all point to the same explanation is an illustration of the concept of "consilience" first championed by Whewell (1840). Ferguson et al. (2012) provide an example of how to devise statistical inference procedures when both predictions and data are qualitative. It strikes us that this may be one arena ripe for further analysis and formalization.

WHERE WILL EVIDENTIAL STATISTICS GO, AND HOW BEST CAN IT BE USED TO INFORM AND REFINE CONSTITUTIVE AND GENERAL THEORIES?

Evidential statistics is still a relatively new approach to linking data, models, and constitutive theories, but it promises to provide a clearer and more coherent way to assess the relative match of models to data, compared to competitors such as Neyman-Pearson testing or Bayesian analysis. Does the use of evidential statistics change if the purpose of a model is for understanding (e.g., why saguaro are confined to the Sonoran Desert) vs. prediction (e.g., what is the most likely global mean temperature in the year 2100)? Does this use change if the model is mechanistic vs. phenomenological? Are different evidence functions better suited for prediction vs. explanation? If one carries out multiple studies, each of which uses evidence functions, how can these best be brought together to examine broad-scale patterns across many systems? Maybe there is a straightforward, evidentiary-statistics version of meta-analysis (for a start, see Goodman, 1989). We use statistical inference to find the model that best fits the data. But the better fitting model may be "less true," in the sense of providing less understanding. A more "accurate" model can be the result of overfitting, especially if the model is phenomenological. Some types of statistical inference (e.g., the use of information criteria like AIC and BIC) try to correct for the inclusion of unnecessary parameters, but we also rely on logical reasoning and prior information to decide which parameters and functional forms are even appropriate to include, a process that is outside of statistical inference itself. For instance, a mathematical model must have units on each side of the equal sign that match; if not the model is, at best, nonsense. A number of evidence functions have been proposed in the literature, and presumably the class of such functions will grow with time. Are the criteria used to assess those functions part of evidentiary statistics, or in some sense outside of it?

If the goal is understanding, a very simple model may be appropriate. For example, we might ask whether saguaro abundance within its occupied range is controlled by intraspecific competition only, or also by interspecific competition with ferrocactus. We could build a very simple model of logistic growth without and with competition and use inferential statistics to ask which model is more consistent with observed densities across space and/or time. The model is not likely to be useful for making an accurate prediction of densities, but may nonetheless help uncover the presence of a particular ecological mechanism (e.g., competition). Simple models can illuminate essential elements of a system, even if statistical inference indicates that the model is very far from an accurate depiction of the empirical system. Depending on our goal, the most useful model could either be very simple (to highlight a single, essential feature) or very complex (to be as accurate as possible). In this case, our goal is not theory testing. Rather, the goal is to use an established theory to build a model for a specific instance so as to enhance understanding.

Prediction is important and indeed vital in the progress of science (Houlahan et al., 2017), but it does not outweigh other considerations in theory evaluation. After all, geocentric Ptolemaic astronomy did a fine job of predicting the movement of the planets for over 1,500 years, at the expense of more and more model complexity. Its supplanting by a gravity-driven, heliocentric theory, was driven, in part, by the latter model being both mechanistic and much simpler. The excellence of Ptolemaic astronomy as a predictive tool is not a very strong argument for hanging on to it as science moves forward. Newton's remarkable accomplishment in his *Principia Mathematica* was to explain an array of already known facts—Kepler's laws, tidal rhythms, the precession of the equinoxes—using just his three laws of motion plus the inverse-square law of gravitation. Novel predictions eventually emerged (e.g., the existence of Neptune), but such predictions were not required for the scientific community by-and-large to become enthusiastic champions of Newtonian mechanics. The super-computers of the future are likely to use vast neural networks, evolving arrays of code-based algorithms, and constant training with the flood of informatics they are constantly fed from arrays of sensors and surveys, and the like, to provide wonderful predictions of climate change and the weather, but this will not substitute for causal, theoretical understanding, often relying at its core on models that are not literally true.

WHEN STATISTICS ARE NOT NECESSARY

Sometimes statistical inference is not necessary for testing a theory, for example when a model is being used to explore if something is possible or not. The data are simply that some object or phenomenon exists or does not exist. The model either matches the data or it does not; no statistical inference is needed. For example, contra the "central dogma" we might have a theory that acquired characteristics can be inherited. For over a century, all of the data said that this theory was false. Then retroviruses were discovered showing that information can flow from RNA acquired from the environment back to DNA. For at least this narrow domain, the theory of the inheritance of acquired characteristics has been shown to be true. One might be able to shoehorn such examples into evidential statistics, but it is not clear that is necessary to understand the logic of scientific discovery in cases of this sort.

Even with a question that is less clear cut than simply "Does it exist?" statistical inference may be unnecessary. Statistical inference is about finding the informative signal within noisy data. For highly controlled experiments, the noise might be so small that the signal is immediately obvious. We know physiologists who say that if you need to use statistics, you really should refine your experimental methodology. Statisticians sometimes refer to this as the interocular trauma test, as in "it hits you between the eyes." Mark Taper (pers. comm.) ripostes "[Y]ou are still comparing the fit of data to models – it is just that the integration can be done by eye." Our evolutionary history has presumably fit us to be pretty good seat-of-the-pants statisticians, in that our past inferences have helped our ancestors survive and reproduce. But this decision process is not the same as the formal mathematics of statistical inference represented by evidential statistics.

CONCLUSION

Evidential statistics is an important advance in model and theory testing, and scientific reasoning in general, combining and extending key insights from other philosophies of statistics. We applaud the editors and authors of this special issue for crystallizing many important exciting themes swirling around the topic of evidential statistics. A scientist should use whichever tool is apt for the particular question at hand. Statistical inference itself is just one class of tools used in scientific inquiry that depends on quantitative data and mathematical reasoning. Other types of data and reasoning are sometimes more appropriate for a given question, such as qualitative data, and narrative or logical reasoning. We urge scientists to use as wide a range of tools as possible in the service of our quest to understand, predict, and manage our ever-fascinating, complex world.

AUTHOR CONTRIBUTIONS

The authors equally conceived of the content and wrote the paper.

FUNDING

RH was supported by the University of Florida Foundation. This manuscript is based on work done by SS while serving at the U.S. National Science Foundation. The views expressed in this paper do not necessarily reflect those of the National Science Foundation or the United States Government.

ACKNOWLEDGMENTS

We thank Mark Taper and Jeff Houlahan for their extensive and perceptive comments on earlier drafts that greatly helped to strengthen our presentation.

REFERENCES

Abrams, P. A. (2015). Why ratio dependence is (still) a bad model of predation. *Biol. Rev.* 90, 794–814. doi: 10.1111/brv.12134

Arditi, R., and Ginzburg, L. R. (2012). *How Species Interact*. New York, NY: Oxford University Press.

Bandyopadhya, P. S., Brittan G. Jr., and Taper, M. L. (2016). *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference*. Zurich: Springer International Publishing AG.

Dennis, B., Ponciano, J. M., Taper, M. L., and, Lele, S. R. (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Environ.* 7.

Douven, I. (2017). "Abduction," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Metaphysics Research Lab, Stanford University). Available online at: https://plato.stanford.edu/archives/sum2017/entries/abduction/

Ferguson, J. M., Taper, M. L., Guy, C. S., and Syslo, J. M. (2012). Mechanisms of coexistence between native bull trout (*Salvelinus confluentus*) and non-native lake trout (*Salvelinus namaycush*): inferences from pattern-oriented modeling. *Can. J. Fish. Aquatic Sci.* 69, 755–769. doi: 10.1139/f2011-177

Forster, M. R., and Wolfe, A. B. (1999). *Conceptual Innovation and the Relational Nature of Evidence: The Whewell-Mill Debate*. Available online at: http://philosophy.wisc.edu/forster/papers/Whewell1.pdf

Frank, S. A., and Fox, G. A. (2019). "The inductive theory of natural selection," in *The Theory of Evolution*, eds S. M. Scheiner and D. P. Mindell (Chicago: University of Chicago Press), 171–193.

Gause, G. F. (1934). *The Struggle for Existence*. Baltimore, MD: The Williams & Wilkins Co.

Ginzburg, L., and Colyvan, M. (2004). *Ecological Orbits: How Planets Move and Populations Grow*. Oxford: Oxford University Press.

Goodman, S. N. (1989). Meta-analysis and evidence. *Control. Clin. Trials* 10, 188–204. doi: 10.1016/0197-2456(89)90030-5

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19659–19665. doi: 10.1073/pnas.0905137106

Holt, R. D. (2011). "Natural enemy-victim interactions: do we have a unified theory, yet?" in *The Theory of Ecology*, eds S. M. Scheiner and M. R. Willig (Chicago: University of Chicago Press), 125–161.

Houlahan, J. E., McKinney, S. T., Anderson, T. M., and McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos* 126, 1–7. doi: 10.1111/oik.03726

Kendall, B. E., and Fox, G. A. (2003). Unstructured individual variation and demographic stochasticity. *Conserv. Biol.* 17, 1170–1172. doi: 10.1046/j.1523-1739.2003.02411.x

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0

Kuhn, T. S. (1977). *The Essential Tension*. Chicago: University of Chicago Press.

Laudan, L. (1981). A confutation of convergent realism. *Philos. Sci.* 48, 19–49. doi: 10.1086/288975

Lele, S. R. (2004). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago: University of Chicago Press), 191–216.

Lipton, P. (2004). *Inference to the Best Explanation, 2nd Edn*. London: Routledge.

Longino, H. E. (2002). *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Love, A. C. (2010). "Rethinking the structure of evolutionary theory for an extended synthesis," in *Evolution - The Extended Synthesis,* eds M. Pigliucci and G. B. Müller (Cambridge, MA: MIT Press), 403–441.

MacArthur, R. H. (1972). *Geographical Ecology*. Princeton, NJ: Princeton University Press.

May, R. M. (1973). *Stability and Complexity in Model Ecosystems*. Princeton, NJ: Princeton University Press.

Mindell, D. P., and Scheiner, S. M. (2019). "The theory of evolution," in *The Theory of Evolution,* eds S. M. Scheiner and D. P. Mindell (Chicago: University of Chicago Press), 1–22.

Pásztor, L., Bottak-Dukát, Z., Magyar, G., Czárán, T., and Meszena, G. (2016). *Theory-Based Ecology: A Darwinian Approach*. Oxford: Oxford University Press.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martinez-Meyer, E., Nakamura, M., et al. (2011). *Ecological Niches and Geographical Distributions*. Princeton, NJ: Princeton University Press.

Pickett, S. T. A., Kolasa, J., and Jones, C. G. (2007). *Ecological Understanding: The Nature of Theory and the Theory of Nature, 2nd Edn*. New York, NY: Elsevier.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Popper, K. R. (1972). "Evolution and the tree of knowledge," in *Objective Knowledge*, ed K. R. Popper (Oxford: The Clarendon Press), 133–143.

Queller, D. C. (2017). Fundamental theorems of evolution. *Am. Natural.* 189, 345–353. doi: 10.1086/690937

Roush, S. (2007). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. New York, NY: International Thompson Publishing.

Scheiner, S. M. (2010). Toward a conceptual framework for biology. *Q. Rev. Biol.* 85, 293–318. doi: 10.1086/655117

Scheiner, S. M. (2019). "The theory of the evolution of plasticity," in *The Theory of Evolution,* eds S. M. Scheiner and D. P. Mindell (Chicago: University of Chicago Press), 254–272.

Scheiner, S. M., and Mindell, D. P. (2019). *The Theory of Evolution*. Chicago: University of Chicago Press.

Scheiner, S. M., and Willig, M. R. (2005). Developing unified theories in ecology as exemplified with diversity gradients. *Am. Natural.* 166, 458–469. doi: 10.1086/444402

Scheiner, S. M., and Willig, M. R. (2008). A general theory of ecology. *Theor. Ecol.* 1, 21–28. doi: 10.1007/s12080-007-0002-0

Scheiner, S. M., and Willig, M. R. (2011a). "A general theory of ecology," in *The Theory of Ecology,* eds S. M. Scheiner and M. R. Willig (Chicago: University of Chicago Press), 3–19.

Scheiner, S. M., and Willig, M. R. (2011b). *The Theory of Ecology*. Chicago: University of Chicago Press.

Snyder, L. J. (2019). "William Whewell," in *Stanford Encyclopedia of Philosophy*, ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/spr2019/entries/whewell/

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

Tukey, J. W. (1960). Conclusions vs. decisions. *Technometrics* 2, 423–433. doi: 10.1080/00401706.1960.10489909

Vandermeer, J. (2018). "Objects of intellectual interest have real life impacts: the ecology (and more) of Richard Levins," in *The Truth is the Whole: Essays in Honor of Richard Levins,* eds T. Awebuch, M. S. Clark, and P. J. Taylor (Arlington, MA: The Pumping Station), 1–9.

VanderMeulen, M. A., Hudson, A. J., and Scheiner, S. M. (2001). Three evolutionary hypotheses for the hump-shaped productivity-diversity curve. *Evol. Ecol. Res.* 3, 379–392.

Wegener, A. (1966). *The Origin of Continents and Oceans, 4th Edn*. New York, NY: Dover Publications.

Whewell, W. (1840). *The Philosophy of the Inductive Sciences, Founded Upon Their Histories*. London: John W. Parker.

Whewell, W. (1858). *Novum Organon Renovatum*. London: J. W. Parker and Son.

Winther, R. (2012). Mathematical modeling in biology: philosophy and pragmatics. *Front. Plant Sci.* 3:102. doi: 10.3389/fpls.2012.00102

Winther, R. G. (2015). "The structure of scientific theories," in *Encyclopedia of Philosophy,* ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/spr2015/entries/structure-scientific-theories/

Yoda, K., Kira, T., Ogawa, H., and Hozumi, K. (1963). Self thinning in overcrowded pure stands under cultivated and natural conditions. *J. Biol. Osaka City Univ.* 14, 107–129.

Zamer, W. E., and Scheiner, S. M. (2014). A conceptual framework for organismal biology: linking theories, models, and data. *Integr. Comp. Biol.* 54, 736–756. doi: 10.1093/icb/icu075

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 20 September 2019
doi: 10.3389/fphys.2019.01166

Strong Evidence for an Intraspecific Metabolic Scaling Coefficient Near 0.89 in Fish

Christopher L. Jerde[1]*, Krista Kraskura[2], Erika J. Eliason[1,2], Samantha R. Csik[2], Adrian C. Stier[2] and Mark L. Taper[3,4]

[1] Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA, United States, [2] Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, Santa Barbara, CA, United States, [3] Department of Ecology, Montana State University, Bozeman, MT, United States, [4] Department of Biology, University of Florida, Gainesville, FL, United States

As an example of applying the evidential approach to statistical inference, we address one of the longest standing controversies in ecology, the evidence for, or against, a universal metabolic scaling relationship between metabolic rate and body mass. Using fish as our study taxa, we curated 25 studies with measurements of standard metabolic rate, temperature, and mass, with 55 independent trials and across 16 fish species and confronted this data with flexible random effects models. To quantify the body mass – metabolic rate relationship, we perform model selection using the Schwarz Information Criteria (ΔSIC), an established evidence function. Further, we formulate and justify the use of ΔSIC intervals to delineate the values of the metabolic scaling relationship that should be retained for further consideration. We found strong evidence for a metabolic scaling coefficient of 0.89 with a ΔSIC interval spanning 0.82 to 0.99, implying that mechanistically derived coefficients of 0.67, 0.75, and 1, are not supported by the data. Model selection supports the use of a random intercepts and random slopes by species, consistent with the idea that other factors, such as taxonomy or ecological or lifestyle characteristics, may be critical for discerning the underlying process giving rise to the data. The evidentialist framework applied here, allows for further refinement given additional data and more complex models.

Keywords: likelihood, evidence functions, SIC, standard metabolic rate, mixed effects models, metabolic scaling, evidentialist statistics

OPEN ACCESS

Edited by:
Maximino Aldana,
National Autonomous University
of Mexico, Mexico

Reviewed by:
Bartolo Luque,
Polytechnic University of Madrid,
Spain
Sigurd Einum,
Norwegian University of Science
and Technology, Norway

****Correspondence:***
Christopher L. Jerde
cjerde@ucsb.edu

Specialty section:
This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Physiology

Received: *25 April 2019*
Accepted: *28 August 2019*
Published: *20 September 2019*

Citation:
Jerde CL, Kraskura K, Eliason EJ,
Csik SR, Stier AC and Taper ML
(2019) Strong Evidence for an
Intraspecific Metabolic Scaling
Coefficient Near 0.89 in Fish.
Front. Physiol. 10:1166.
doi: 10.3389/fphys.2019.01166

INTRODUCTION

One of most contentious controversies in ecology is the scaling relationship between an organism's body mass and metabolic rate (Agutter and Wheatley, 2004; Isaac and Carbone, 2010; Glazier, 2018). Kleiber (1932) popularized the idea that contrary to a century of theory, a mammal's metabolic rate (MR) scales with body mass (BM) not as a power law with an exponent of $\beta = 0.67$, but as a power law with an exponent of $\beta = 0.75$. This relationship takes the form

$$\ln(MR) = \beta \times \ln(BM) + c \tag{1}$$

where β is the scaling relationship and c is an intercept from a liner regression. As a cornerstone of the metabolic theory of ecology (Brown et al., 2004), this 0.75 scaling relationship is used to link individual physiology to the observed patterns of communities and energy flows across landscapes. The 0.75 value has been mechanistically justified through hypotheses that maximize energy delivery to tissue in animals (West et al., 1997) and from xylem and phloem networks that transport water and nutrients in plants (Enquist and Niklas, 2001). However, the universality of the 0.75 value is eagerly disputed, with alternative hypotheses and empirical studies putting the scaling relationship commonly between 0.5 and 1 (Bokma, 2004; Glazier, 2018).

Intraspecific (within species) scaling has been proposed to differ from interspecific (between species) scaling and also different mechanisms may be responsible for different scaling relationships. Metabolic rates vary 2–3 fold across individuals of the same population and this variation is repeatable (Burton et al., 2011; Norin and Malte, 2011; Boldsen et al., 2013). Intraspecific scaling has received less attention than interspecific scaling, while even fewer studies have investigated scaling relationships within each tested individual as it grows (but see Norin and Gamperl, 2018). Both intraspecific and interspecific scaling are critical for linking species physiology to projections of population abundance (Kooijman, 1993) and predicting the impacts of climate change on species distributions (Sunday et al., 2010; Lindmark et al., 2018).

While the implications of deviations from the 0.75 scaling exponent are large, there is limited data available to accurately estimate the exponent. This is because measuring the metabolic rate of an individual is not a trivial experiment, let alone across a 10-fold range of body sizes from a population, at different temperatures, and/or across species (Lighton, 2018). To date, most studies have relied on either a limited study design (one species, many individuals, with fixed treatments of temperature; **Table 1**) or meta-analysis of mean metabolic rate data across studies using variable methods of measurement (Glazier, 2005). While the former can suffer from insufficient sample sizes, measurement error, and unaccounted for factors influencing the general relationship, the latter treats all studies equally and both approaches have ultimately been inconclusive as to the evidence supporting or refuting competing hypotheses (Glazier, 2018) with some concluding there is not a universal scaling constant (Bokma, 2004).

In this *Frontiers* Research Topic devoted to evidential statistics, model identification, and science, multiple contributions (Dennis et al., 2019) show how standard statistical approaches (such as Fisherian significant tests, Neyman-Pearson hypothesis testing, Akaike Information Criterion for multi-model inference) are misleading when models used for inference are misspecified. Model misspecification is arguably the case for most analyses, including ours, that seek to evaluate the evidence of a universal scaling relationship across a broad range of fish species, at different temperatures, and using studies, that have reliable data, but that were not necessarily designed to have a large range of body masses across which to regress metabolic rate. Here we demonstrate how an evidentialist approach can be applied to gain novel insight to the question, "What is the evidence for an intraspecific universal scaling relationship between fish body mass and metabolic rate?"

Scaling Relationships as Hypotheses for Fish

Multiple mechanisms have been put forth to justify $\beta = 0.67$, 0.75, and 1 scaling relationships. If the primary limitation for resources or waste removal is transport of chemicals across surfaces, then metabolic rate is predicted to scale with surface area with a relationship of 0.67. For example, Killen et al. (2010) found that highly active, pelagic fishes had a scaling relationship of 0.7 (SE 0.04), close to 0.67, which they attributed to a constraint in oxygen or fuel acquisition or waste removal across surface areas in these metabolically active fishes. However, the 0.67 scaling exponent is more commonly found in endotherms, mammals and birds, but rarely in ectotherms (White and Seymour, 2003; White et al., 2005).

If metabolic rate is primarily limited by the fractal nature of distribution networks (e.g., the internal transport networks of resources and wastes), then a scaling relationship of 0.75 is predicted (West et al., 1997). Previous synthesis of teleost fish found a scaling relationship of 0.79 (SE 0.11) (Clarke and Johnston, 1999), and with sufficient variability as to not exclude the 0.75 value used by Metabolic Theory of Ecology to explain broad ecological patterns (Brown et al., 2004). Similarly, Moses et al. (2008) showed metabolic scaling during ontogeny for seven fish species was 0.78 (SE 0.02), with some variability in slope estimates between species.

Metabolic rate is predicted to be directly proportional to body size (i.e., $\beta = 1$) when maintenance and routine activity costs are low and these demands can easily be met by both surface area and internal transport mechanisms. In the case of less active fish or those occupying deeper waters, individual metabolism has been demonstrated to scale nearly proportionally to body mass [i.e., scaling exponents approach 1 (Killen et al., 2010)].

Two more recent hypotheses work with the common observation that scaling exponents vary (e.g., Glazier, 2018). The metabolic-level boundaries (MLB) hypothesis of scaling (Glazier, 2008) states that any observed scaling exponent varies within the limits of 0.67 and 1, representing whether the mechanisms or processes that underlay the scaling relationship are predominantly limited by surface area constraints on fluxes of resources, waste and heat (0.67; e.g., gill surface area, internal transport limitation) or by volume (mass) constraints on energy demand or production of tissue (1; assuming energy demand is proportional to tissue size). Therefore, MLB also provides an explanation to variable scaling exponents of animals at different physiological states, or routine requirements. Alternatively, Dynamic Energy Budget (DEB) theory (Kooijman, 1993) provides a more recent approach predicting metabolic scaling relationships in all species irrespective to taxonomical classification; this approach is based solely on physical principles, and uses storage of nutrients (reserves increase with increasing structure) as a central mechanism explaining both intra- and inter species-specific scaling relationships (Maino et al., 2014). While both MLB and DEB would seemingly make the case that a universal scaling exponent does not exist and should consequently not be expected, they do not preclude a mean universal scaling exponent.

TABLE 1 | Overview of metabolic studies.

Citation	Species	*n*	Temp. (°C)	(Min, Max) weight (g)	Regression coefficient $\hat{\beta}$, (SE)	Trial
(1) Norin and Gamperl, 2018	Cunner (*Tautogoabrus adsperus*)	68	15	0.45, 4.61	0.92 (0.035)	1
Norin and Gamperl, 2018	Cunner (*Tautogoabrus adsperus*)	68	15	0.97, 7.94	0.98 (0.028)	2
Norin and Gamperl, 2018	Cunner (*Tautogoabrus adsperus*)	68	15	1.24, 13.2	0.89 (0.024)	3
Norin and Gamperl, 2018	Cunner (*Tautogoabrus adsperus*)	68	15	1.56, 15.56	0.83 (0.024)	4
Norin and Gamperl, 2018	Cunner (*Tautogoabrus adsperus*)	68	15	1.71, 19.46	0.79 (0.026)	5
(2) Auer et al., 2015	Brown Trout (*Salmo trutta*)	120	11.5	5.48, 16.12	0.61 (0.068)	6
(3) Behrens et al., 2017	Round Goby (*Neogobius melanostomus*)	8	15–17	43, 73	1.031 (0.24)	7
Behrens et al., 2017	Round Goby (*Neogobius melanostomus*)	8	15–17	35, 78	1.38 (0.16)	8
Behrens et al., 2017	Round Goby (*Neogobius melanostomus*)	8	15–17	36, 72	0.9 (0.17)	9
(4) Killen, 2014	Common Minnow (*Phoxinus phoxinus*)	13	10	0.72, 2.03	0.78 (0.27)	10
(5) Norin and Clark, 2017	Barramundi (*Lates calcarifer*)	24	29	23.1, 37.6	0.91 (0.17)	11
(6) McLean et al., 2018	Common Minnow (*Phoxinus phoxinus*)	123	13	0.68, 7.44	0.72 (0.07)	12
(7) Boldsen et al., 2013	European Eel (*Anguilla anguilla*)	24	20	184, 507	1.44 (0.25)	13
Boldsen et al., 2013	European Eel (*Anguilla anguilla*)	24	20	171, 504	1.05 (0.21)	14
(8) Kunz et al., 2016	Polar Cod (*Boreogadus saida*)	5	0	18.5, 27.4	0.81 (0.35)	15
Kunz et al., 2016	Polar Cod (*Boreogadus saida*)	5	3	16.1, 48.6	0.96 (0.1)	16
Kunz et al., 2016	Polar Cod (*Boreogadus saida*)	5	6	22.7, 32.8	1.06 (0.41)	17
Kunz et al., 2016	Polar Cod (*Boreogadus saida*)	6	8	11.4, 29.1	1.03 (0.3)	18
Kunz et al., 2016	Atlantic cod (*Gadus morhua*)	12	3	21.2, 105	0.97 (0.15)	19
Kunz et al., 2016	Atlantic cod (*Gadus morhua*)	10	8	45.7, 173.6	0.9 (0.15)	20
Kunz et al., 2016	Atlantic cod (*Gadus morhua*)	7	12	54.5, 149.1	1.1 (0.13)	21
Kunz et al., 2016	Atlantic cod (*Gadus morhua*)	5	16	83.2, 156.2	1.05 (0.18)	22
(9) Norin et al., 2016	Barramundi (*Lates calcarifer*)	60	29	23.08, 48.96	1.03 (0.13)	23
(10) Collins et al., 2016	Barramundi (*Lates calcarifer*)	20	30	153.9, 453.7	1.07 (0.14)	24
Collins et al., 2016	Barramundi (*Lates calcarifer*)	20	30	196.3, 390	1.19 (0.28)	25
(11) Khan et al., 2014	Hapuku Wreckfish (*Polyprion oxygeneios*)	8	12	88.2, 131.2	0.93 (0.45)	26
Khan et al., 2014	Hapuku Wreckfish (*Polyprion oxygeneios*)	8	15	105.3, 164.5	0.64 (0.44)	27
Khan et al., 2014	Hapuku Wreckfish (*Polyprion oxygeneios*)	8	18	146.1, 203.2	−0.21 (0.48)	28
Khan et al., 2014	Hapuku Wreckfish (*Polyprion oxygeneios*)	8	21	130.3, 188.6	0.61 (0.26)	29
Khan et al., 2014	Hapuku Wreckfish (*Polyprion oxygeneios*)	8	24	97.7, 131.6	1.2 (0.36)	30
(12) Khan et al., 2018a	Rainbow Trout (*Oncorhynchus mykiss*)	16	16	69.9, 120.2	0.87 (0.32)	31
(13) Khan et al., 2018b	Atlantic Salmon (*Salmo salar*)	25	14	39.1, 70.7	0.57 (0.22)	32
(14) Khan et al., 2015	Hapuku Wreckfish (*Polyprion oxygeneios*)	12	15	196.1, 324	0.84 (0.15)	33
(15) Khan et al., 2015	Hapuku Wreckfish (*Polyprion oxygeneios*)	12	21	114.5, 191	0.6 (0.2)	34
(16) Cooper et al., 2018	Three Spine Stickleback (*Gasterosteus aculeatus*)	31	12	0.46, 1.19	1.43 (0.39)	35
(17) McArley et al., 2017	Common Triplefin (*Forsterygion lapillum*)	20	15	1.59, 3.38	0.67 (0.19)	36
McArley et al., 2017	Common Triplefin (*Forsterygion lapillum*)	20	18	1.52, 3.81	0.82 (0.19)	37
McArley et al., 2017	Common Triplefin (*Forsterygion lapillum*)	23	21	1.54, 3.42	0.78 (0.15)	38
(18) McArley et al., 2018	Twister (*Bellapiscis medius*)	10	21	1.53, 3.98	0.94 (0.1)	39
McArley et al., 2018	Common Triplefin (*Forsterygion lapillum*)	10	21	1.27, 2.97	0.45 (0.16)	40
(19) Eliason et al., 2007	Rainbow Trout (*Oncorhynchus mykiss*)	24	8–14	381, 652.7	0.64 (0.74)	41
Eliason et al., 2007	Rainbow Trout (*Oncorhynchus mykiss*)	5	11–16	564.8, 3233.6	1.33 (0.3)	42
(20) Norin and Malte, 2011	Brown Trout (*Salmo trutta*)	33	15	20.7, 45.7	1.5 (0.18)	43
Norin and Malte, 2011	Brown Trout (*Salmo trutta*)	33	15	27.4, 55.1	1.19 (0.14)	44
Norin and Malte, 2011	Brown Trout (*Salmo trutta*)	33	15	37.7, 64.9	0.98 (0.18)	45
Norin and Malte, 2011	Brown Trout (*Salmo trutta*)	33	15	38.4, 68.2	1.11 (0.17)	46
(21) Norin and Malte, 2012	Brown Trout (*Salmo trutta*)	66	15	20.5, 57.7	1.09 (0.094)	47
(22) Nadler et al., 2016	Blue Green Puller (*Chromis viridis*)	16	29	1.3, 2.1	0.63 (0.3)	48
(23) Collins et al., 2013	Barramundi (*Lates calcarifer*)	9	26	172, 205	0.18 (1.03)	49
Collins et al., 2013	Barramundi (*Lates calcarifer*)	10	26	186, 221	2.06 (1.28)	50
Collins et al., 2013	Barramundi (*Lates calcarifer*)	10	26	169, 215	1.49 (0.78)	51
Collins et al., 2013	Barramundi (*Lates calcarifer*)	11	26	139, 244	0.65 (0.43)	52
Collins et al., 2013	Barramundi (*Lates calcarifer*)	9	26	184, 233	0.71 (0.54)	53
(24) Zhang et al., 2017	European Sea Bass (*Dicentrarchus labrax*)	11	16.5	48.1, 100.7	1.01 (0.18)	54
(25) Zhang et al., 2016	Atlantic Salmon (*Salmo salar*)	87	12	23.4, 57	1.15 (0.11)	55

Temperature and Other Factors

Temperature plays a critical role regulating individual metabolic rate in ectotherms such as fishes (Brett and Glass, 1973; Johnston and Dunn, 1987). The effects of temperature on the metabolic scaling relationship has been studied mechanistically (Gillooly et al., 2001) with syntheses showing low temperature sensitivity from resting measures of metabolism and a consistent metabolic scaling relationship (Clarke and Johnston, 1999, but see Lindmark et al., 2018).

Numerous ecological, physiological and lifestyle characteristics can influence metabolic rate and potentially affect scaling relationships. Metabolic rate in ectotherms is strongly dependent on physical and chemical characteristics of the water they live in, and consequently shows context-dependent variation (Killen et al., 2016). Therefore, habitat (abiotic factors), predation risk, activity level, food availability, and social status and behavioral traits, all can affect metabolic rates (for a review on variation of fish standard metabolic rate (SMR), see Metcalfe et al., 2016), thus also likely scaling parameters, especially intercept. For example, food availability affects growth rates and is linked to SMR variation in fish (Killen, 2014; Auer et al., 2015). Auer et al. (2018) demonstrated a strong dependence of SMR on individual ecology underlined by predation level, reproductive age and investment, longevity, and maximum body size (life-history traits). Many of these factors vary in unique combinations across populations of the same species (Eliason et al., 2011; Auer et al., 2018), therefore even within species we may expect variation in metabolic rate and its dependence on size.

Sources of Uncertainty and Measurement Error

Misspecification is a model that does not account for variables (i.e., temperature) or structural forms (i.e., random effects) that can lead to biased coefficients, misleading error terms, and unlimitedly wrong inferences about the generating process giving rise to the data (White, 1982). While temperature has been identified as a critical covariate for fish (Brett and Glass, 1973), other necessary covariates are less clear, but one should assume there is likely something missing. Additionally, as any model expands its inferential breadth beyond a single species, the model will become more complex either by adding fixed effects to measure species-level coefficients or by treating species as a random effect of the model from which to make inference across all fish. The advantage of using random effects to make broader inferences has been well recognized across ecology (Bolker et al., 2009). Such is the case when making population level inferences in resource selection functions from location data from multiple individuals (Gillies et al., 2006). However, more information on the species level traits may lead to better models and improved inferences.

The quality of the data will also impact inferences. One known source of uncertainty is measurement error – that is the errant measurement of observations, such as body mass. Farrell-Gray and Gotelli (2005) clearly showed that errant measurement of the predictor variable of mass biased the estimated slope parameter of the metabolic relationship and speculated that allometric exponents lower than 0.75 may be due simply to measurement error. The magnitude of the effect of measurement error in a predictor variable on the estimated slope of a linear regression is well known: $E(\hat{\beta}) = \lambda\beta$, where λ, the reliability coefficient, is the proportion of variation in the predictor variable not due to measurement error (Taper and Marquet, 1996; Cheng and Van Ness, 1999). The lower reliability the more biased the estimate. In **Box 1**, we evaluated the influence of measurement error for California spiny lobster (*Panulirus interruptus*), albeit not a fish, but find very little evidence for any bias due to measurement error from retained residual water. We assume going forward, that for fish, measurement error is not biasing our parameter estimates.

Measurement error in the response variable, metabolic rate in our study, leads to greater residual variability but no bias in the slope parameter. However, the added variability in the residual error can inflate our uncertainty surrounding the slope parameter leaving us unable to distinguish between potential hypotheses (competing models). Metabolic rate (MR) represents a sum of all chemical reactions that take place in an organism, and this may change drastically upon any intrinsic and extrinsic change, e.g., spontaneous activity, physiological disturbance, feeding, and even just circadian rhythms. To refine how MR varies as function of mass, it is a necessity that the data originate from animals at the same physiological states. Standard metabolic rate, SMR is defined as the subsistence metabolism to support body maintenance in a post-absorptive, resting state under thermally acclimated conditions (Chabot et al., 2016). True SMR is often impractical and challenging to measure in fishes, and so data often reflects routine metabolic rates, which alternatively may be perceived as a measurement error (in the response, Y axis) around individual SMR, which increases variability but does not bias the slope parameter. With a goal to minimize such variation, we developed specific experimental criteria for data to be included (see section "Data"). For a good overview of methods and approaches to metabolic scaling in animals see White and Kearney (2014).

MATERIALS AND METHODS

The general approach we implemented for this study is to: (1) include reliably collected SMR data based on recently published studies (200-present), (2) apply flexible, mixed effect linear models, and (3) employ an evidence function, the Schwarz Information Criterion (SIC), to evaluate the evidence for specified mechanistic hypotheses of the scaling relationship of β = 0.67, 0.75, 1, and β as an estimated, free parameter ($\hat{\beta}$).

Data

The approaches and technology used to measure fish metabolism have become more accurate, precise, and robust within the last 20 years (Nelson, 2016). We curated published data sets of individual fish metabolism comprised of fish that were: 1) post larval life stages, 2) in a post-absorptive state, meaning they were unfed for a minimum of 20 h prior to taking metabolic rate measurements, 3) with overnight metabolic rates (>12 h of automatic measurement), 4) with an acclimated water temperature for at least 7 days prior to the experiment, and 5) were at calm resting states.

BOX 1 | Measurement error in body mass of lobsters.

California spiny lobster (*Panulirus interruptus*) is commercially highly valued, and is ecologically important having a large effect on trophic dynamics and ecosystem resilience in kelp forests and rocky reef beds (Dunn et al., 2017; Caselle et al., 2018). Metabolic rate in ectotherms directly depends on animal's body size and temperature and represents the pace of nearly all biological processes. Meanwhile, MR varies within and among individuals (Glazier, 2005; White and Kearney, 2013; Norin and Gamperl, 2018). Lobsters are cumbersome to weigh, thus making them a good candidate to explore how measurement error in body mass may affect metabolic scaling.

Lobsters were collected by divers via SCUBA (CDFW Scientific Collection Permit #13746) and maintained in 110-gallon flow-through seawater tanks divided in half with perforated PVC. One individual was held in each half tank (24"L × 30"W × 18"H), and provided with 10" PVC cut in half to create structure and habitat. Lobsters were fed mussels (*Mytilus* spp.) *ad libitum* when not being used in respirometry trials. Animals were held at ambient temperatures and exposed to natural light.

To estimate measurement error, 45 lobsters were weighed three consecutive times. Before weighing, individual's dorsal side and tail were dried with a microfiber towel. The mass was measured to the nearest gram. Lobsters were fully submerged between repeat trials.

From the log transformed mass measurements (n = 45), the pooled error variance is 1.2×10^{-5} (SD 0.0035). We regressed the within individual standard deviation against the mean log(weight), but the slope was not different from zero and Levene's test did not indicate there is any heterogeneity. From inspection of the pooled error variance, there is very little variability in the individual measurements of body mass. Furthermore, regression revealed no trend in error variance as function of mean body mass.

Photo: Co-authors Krista Kraskura (left) and Samantha Csik (right) collect measurement error data on California spiny lobster.

For six lobsters, ranging in body mass from 175 to 2426 g, we conducted a more thorough drying by carefully removing water from the leg joints, carapace, and underside of the lobster abdomen, spending approximately double the time drying than the standard protocol called for. We regressed the mean log(weight) against the thoroughly dried log (weight) for the six lobsters. Expectedly, the intercept (0.05, SE 0.008) and slope (0.994, SE 0.0001) were statistically significant (p < 0.001), but the residual standard was very small (0.0033), indicating that *measurement error in mass is negligible.* Thus, for all regressions with log(weight) as a predictor variable, the reliability ratio will be effectively 1 and there will be no bias in estimated slopes due to measurement error.

Studies where species were manipulated, such as treatments to measure the effects of starvation on SMR, or where the study's authors noted substantial spontaneous activity were not included. Further, we ensured robust data analysis methods were used to calculate SMR following Chabot et al. (2016) and where SMR was measured at ecologically relevant temperature ranges for each species. Studies were not considered if they included surgical manipulations with the exception of non-invasive tagging (e.g., using passive integrated transponder (PIT) and visible implant elastomer tags). Data were not included if the study's methods lacked sufficient detail in any of the above criteria, the **Supplementary Data** online were not clear, or appeared to contain errors. All fish included were lab residents for at least 2 weeks before the SMR measurement took place.

Our database includes 25 studies, with 55 independent trials, across 16 fishes (**Figure 1**). **Table 1** details the sources of the data, species, trials identification, temperature under which the SMR measurements were collected, and sample sizes per trial. A total of n = 1456 observations are used in the study. Some studies where not designed or conducted to estimate the scaling relationship between individual fish SMR and body mass – a notable point we will return to in later sections.

Models

Linear Regression

Each trial (**Table 1**; n = 55) is an experiment of the metabolic scaling relationship of SMR to body mass. We applied linear regression to the log transformed SMR and body mass data for each trial. Because some of these studies were not designed to test this relationship, we expect the regression slope estimates to be variable and have large standard errors for those data sets with low sample size. Additionally, it is recommended to have a 4 to 10-fold range of fish body mass, but many trials and studies do not meet this recommendation. However, the data in totality has a range from 0.45 to 3233.6 g. We expect the distribution of slopes from trials to largely mirror the results found by Clarke and Johnston (1999).

Linear Mixed Effects Models

Using the lme4 package in the R statistical programing language (Bates et al., 2015), we tested four unique suites of model forms with combinations of fixed and random effects. For all models we included temperature (but see **Box 5**) and body mass as a fixed effect, and we treated trials within species as a nested effect. The first model suite allows intercepts to randomly vary among species. The second model suite, has fixed intercepts for each species with common slope, but does not assume a normal distribution of species' intercepts. With 16 unique species, this second approach adds significantly more parameters to estimate, but allows for inferential insights into the differences between species. The third model suite uses a random slope and random intercept by species. The correlation between the slope and intercept is estimated and not assumed to be independent. The fourth model suite uses a random slope with estimated intercepts for each species. The random slopes are interpreted as by-species deviations from the fixed effect slope.

For each of the four approaches, we evaluate the fixed effect slope of body mass as a free parameter and then constrained the slope to equal each of our underlying mechanistic hypotheses of 0.67, 0.75, and 1.

Analysis

All models were fit using Maximum Likelihood Estimation (MLE) and all analyses were conducted in the R statistical programing language (R Core Team, 2015).

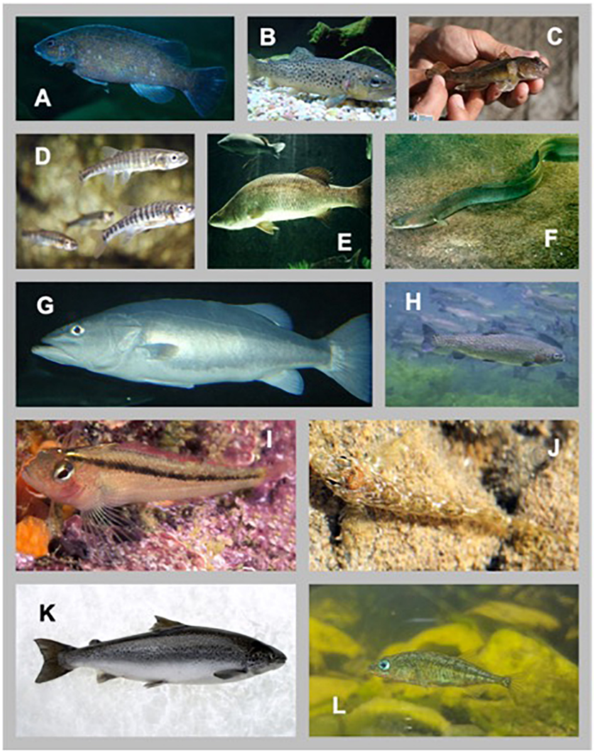


FIGURE 1 | Diversity of species used in this study. **(A)** Cunner (https://commons.wikimedia.org/wiki/File:Cunner.jpg; to Flickr, by Vhorvat), **(B)** Brown Trout (https://commons.wikimedia.org/wiki/File:Brown_trout.JPG; Zouavman Le Zouave), **(C)** Round Goby (https://www.michigan.gov/invasives/0,5664,7-324-68002_73845-368437--,00.html; David Copplestone), **(D)** Common Minnow (Subaqueous Vltava, Prague 2011, Czechia; Provided by Karelj), **(E)** Barramundi (https://commons.wikimedia.org/wiki/File:Barramundi.jpg provided by Nick Thorne), **(F)** European Eel (https://commons.wikimedia.org/wiki/File:Anguilla_anguilla.jpg; GerardM), **(G)** Hapuku Wreckfish (https://commons.wikimedia.org/wiki/File:Hapuka.jpg; Nholtzha), **(H)** Rainbow Trout (https://digitalmedia.fws.gov/digital/collection/natdiglib/id/2151 Eric Engbretson), **(I)** Common Triplefin (https://commons.wikimedia.org/wiki/File:Forsterygion_lapillum_(Common_triplefin).jpg; Ian Skipworth), **(J)** Twister (https://commons.wikimedia.org/wiki/File:Bellapiscis_medius_2.jpg; A.C. Tatarinov), **(K)** Atlantic Salmon (https://commons.wikimedia.org/wiki/File:CSIRO_ScienceImage_8062_Atlantic_salmon.jpg; Peter Whyte, CSIRO), **(L)** Three-spined Stickleback (https://commons.wikimedia.org/wiki/File:Three-spined_Stickleback_(Gasterosteus_aculeatus)_at_the_Palo_Alto_Junior_Museum_and_Zoo.jpg; Evan Baldonado/AquariumKids.com).

Strategy of Scientific Inference and Statistical Tactics

Classical hypothesis testing has been the backbone of scientific inference for nearly a century. Both the Fisherian and the Neyman-Pearson variants of hypothesis testing turn on the axle of a counterfactual argument. The argument stripped of probabilistic uncertainty runs like this: If we assume a particular model (generally called the null) is true then we can predict that a specific pattern should occur in our data. If the predicted pattern does not occur, then the null hypothesis cannot be true and something else must be.

This argument has worked well for science in tightly controlled situations where the predicted patterns are clear and the nature of the "something else" is unequivocal. But in more open situations, with more experiments, more models, more questions and variable amounts of data, the chain of hypotheses (multiple models) becomes harder to follow and the statistical adjustments required to maintain even the illusion of control of error rates become more cumbersome. Paradoxically, considering more models and asking more questions makes it harder to find support for any model or to answer any question.

One common approach to multimodal inference is the application of information criterion (Burnham and Anderson, 2004). Akaike's Information Criterion (AIC) is one such inductive inferential approach that is both widely recognized and applied (Akaike, 1981). The appeal of such an approach is to simultaneously assess competing hypotheses based on how well the models perform relative to each other through the likelihood function, but then discount the potential overfitting of models that have a large number of parameters.

User-defined thresholds demark ΔAIC values that constitute weak, strong, or very strong evidence for one model over the other. If parameters are estimated, the likelihood becomes a biased estimate of how close a model is to the generating process. The more parameters estimated, the greater this over optimism. Akaike (1973) initiated the use and study of information criteria, which correct for this bias. Information criteria have been enormously useful in analyzing biological data (see Burnham and Anderson, 2002). Many information criteria (the consistent criteria) fully meet all the criteria listed in **Box 2** and are evidence functions.

Evidence for one model over another is a function of the estimated relative discrepancy of any two models from the generating process and is measured by evidence functions. Evidence functions (**Box 2**) can take many forms (see Lele (2004), and Taper and Lele (2011) for technical and philosophical discussions, and Taper and Ponciano (2016) for a more general discussion). The Schwarz Information Criterion (SIC) often referred to as the Bayesian Information Criterion (BIC), when used to compare differences between competing models (ΔSIC) is an evidence function (Dennis et al., 2019). Similar to AIC, the SIC (Eq. 2) uses the maximum likelihood function (L) to evaluate the fit of the model to the data and uses a function of the amount of data (n) and the number of parameters (k) to penalize for overfitting (Burnham and Anderson, 2004).

$$\text{SIC} = \underbrace{\ln(n) * k}_{\substack{\text{Penalizes for over} \\ \text{fitting when there} \\ \text{is very few data} \\ \text{(n) or very many} \\ \text{parameters (k)}}} - \underbrace{2 * \ln(L)}_{\substack{\text{Measures how good} \\ \text{the data fit the} \\ \text{model via the} \\ \text{Likelihood function} \\ \text{(L)}}} \tag{2}$$

The SIC penalizes for model complexity more heavily than AIC and the error properties are aligned with the concept of evidence functions, whereas the AIC error properties are not (Dennis et al. this research topic). SIC is also commonly available in R packages (named the BIC). The criterion (Eq. 2) can be derived either in a Bayesian context (Schwarz, 1978) or in a

BOX 2 | What is an evidence function?
Evidence functions are based on nine desiderata (i.e., something that is desired or wanted) for statistical and philosophical properties with desirable and meaningful characteristics for scientific applications (Lele, 2004; Taper and Lele, 2011; Taper and Ponciano, 2016). Here, we attempt to translate those desired properties (D0 to D8) for scientists with emphasis on implications to applications.

D0: Evidence is measurable, does not require information about beliefs, and is made from confronting at least two models that represent scientific hypotheses with the data.

D1: Evidence functions measure how possible data under each model (at least two) match or are comparable to the observed data. Neither model may completely describe the process that generated the observed data, but the function can discriminate if one of the models is more likely to have generated the observed data.

D2: Evidence is continuous from virtually none to very strong, and measuring evidence should likewise be a continuous and not have a threshold like using α levels for hypothesis testing.

D3: Evidence must be arrived at in a reproducible way. If I do not describe processes by which I arrive at a conclusion, then it becomes difficult for someone else to follow the logic to get to that conclusion or challenge the underlying approach.

D4: Personal opinions, beliefs, or intentions cannot influence the evidence function in a hidden way and the process should be accessible to everybody. If a broader scientific audience does not understand what constitutes evidence, then the function cannot be used as evidence.

D5: Evidence functions do not change person to person (in contrast to Bayesian approaches with different personal priors).

D6: Evidence does not need to come from a single critical test (experiment). Evidence functions should have an explicit way of combining data sets to confront hypotheses and the process should be inherently dynamic with reevaluation as more data or better data are collected.

D7: The evidence should not change depending on the scale the data was collected and analyzed. Nor should evidence be sensitive on transformation of parameters. To give an example related to the metabolic scaling relationship research, if we allowed the appearance of plots to be evidence for the slope, then we could change our evidence by making one plot with one *x*-axis scale and another plot with different scale. One of the interests of this paper is how much difference there is among species in β. It should not make a difference to the evidence if this dispersion is parameterized as a variance or as a standard deviation.

D8: More data results in better inferences, but will only be as good as the completeness of the models/hypotheses tested. The model selected in any given analysis will, with more and more data collected, be the model closest to describing the process from which the data are observed. You can do no better in understanding the underlying process than the models contained within your suite of models evaluated.

frequentist context (Nishii, 1988) We adopt the SIC terminology throughout for model selection and evaluation of parameter uncertainty using ΔSIC intervals to avoid confusion of the evidentialist approach with Bayesian analysis and inference. The model with the lowest value of SIC is considered the best model and the evidence function, $\Delta\text{SIC}_{\text{ij}}$, is the pairwise difference formed by subtracting the SIC of a reference model i from the SIC of a competing model j. As an evidence function, $\Delta\text{SIC}_{\text{ij}}$ is continuous from negative infinity to infinity with the strength of evidence for the reference model over the competing model growing larger as the ΔSIC becomes positive and large. Commonly, when information criteria are used for model selection, the model in the model set with the lowest IC value is used as the reference model, and all ΔIC are therefore positive.

Given the hierarchical nature of mixed models several alternative effective sample sizes can be calculated (Jones, 2011); these methods adjust the sample size (*n*), used in the SIC calculation (Eq. 2) to the effective samples size to account for assumptions of non-independence in data. Which is most appropriate depends on the level in the hierarchy of inferential interest. Because the parameter of primary interest in this study is the fixed effect of body mass, the total sample size is the correct effective sample size to use (Lorah and Womack, 2019).

Instead of attempting to reject false models, the evidential approach seeks to assess which models are closer to the unknown natural generating process than other competing models. The support for one model does not in itself diminish support for other models. However, scientists may find themselves in the situation where several distinct models appear nearly as good. Given the data in hand, the scientist cannot strongly differentiate between the models in this set. In this case, all of these models should be retained in the scientist's thinking.

ΔSIC Intervals

SIC values can also be used to define uncertainty surrounding a parameter estimate – thus linking model selection to measures of uncertainty directly through the use of ΔSIC. Discussion of evidential intervals based on the likelihood ratio can be found in Royall (1997), while Bandyopadhyay et al. (2016) discuss ΔSIC evidential intervals. As with ΔAIC, there are some guidelines (suggestions) on what constitutes weak evidence or strong evidence for one model over another based on the value of ΔSIC. Raftery (1995) suggested that a ΔSIC (i.e., ΔBIC) values less than 2, 2 to 6, 6 to 10, and greater than 10 constitute weak, positive, strong, and very strong evidence, respectively. Such verbal partitioning of any information criterion is often desirable for interpretation, but rarely justified.

Box 3 provides a more intuitive probabilistic approach to selecting a value. From our more detailed example in **Box 3** using binomial probability model, it can be shown that at five consecutive heads, the probability of this occurring by chance is ∼0.03 with a ΔIC∼7. Building an uncertainty bound around a parameter value requires choosing a ΔSIC value, we use seven as our threshold for intervals, ΔSIC(7).

A ΔSIC interval for the metabolic scaling relationship (slope parameter) can be built for each trial or for the best selected model by calculating ΔSIC across the parameter space of the slope parameter. The ΔSIC is the difference of the SIC of the best model and the SIC of the same model with a fixed value of the slope parameter. The upper and lower bound of the ΔSIC interval occurs when ΔSIC = 7. **Figure 2** visually captures the process, where the parameter space of the slope parameter is on the x-axis and the ΔSIC is a function of this slope parameter. Expectedly, ΔSIC values greater than 7 would result in broader intervals. If we consider ΔSIC(7) as strong evidence, then the bound can be interpreted as *there is strong evidence that values of the scaling relationship outside of this range do not give rise to*

BOX 3 | Intuitions about evidence.

Fisherian significance tests (think *p*-values) and Neyman-Pearson hypothesis test (think α levels) rely on critical values. The confusion and convolution of these two statistical approaches have led applied scientists to misinterpretations of the strength of evidence against the null hypothesis. As Hubbard and Bayarri (2003) so state it, "This mass confusion, in turn, has rendered applications of classical statistical testing all but meaningless among applied researchers."

Multi-model inference using Information Criteria (IC) (e.g., AIC, SIC) have a continuous measure of evidence found in the difference (i.e., ΔAIC, ΔSIC) in values between the best model (hypothesis) and the reference model. However, communicating this strength of evidence has resulted in vagueness emerging from linguistic uncertainty (Elith et al., 2002). This is to say, applied scientists have created guidelines to discuss the strength of evidence. Maybe the most popular recommendation was provided by Burnham and Anderson (2002) for ΔAIC ($AIC_i - AIC_j$), where 0 > ΔAIC > 2, 4 > ΔAIC > 7, ΔAIC > 10, represent "substantial," "considerably less," and "essentially none" levels of evidence to support for retaining model *i* in the model set along with the best model *j*. Never minding the absence of what a value of 3 might indicate, some scientists have suggested different discretization of intervals (i.e., Burnham et al., 2011) adding to the apparent vagueness of what constitutes evidence on a continuous scale rather than a discrete critical test provided by *p*-values (Murtaugh, 2014).

To a certain extent that different scientists recognize different ΔIC levels as strong evidence represents differences in attitude about science as a whole and their specific research problem. This variation is no different from one scientist choosing a critical value of 0.05 for a hypothesis test and another scientist choosing 0.01. The clearest exposition for developing an intuition for evidence on a continuous scale (**Box 2**, D2) for an evidence function is in Royall (1997), which we recast here in terms of coin tosses.

Imagine that you are gambling with someone on their flipping of a coin and wonder if you are being cheated with a double-headed coin, or if the coin is fair. After the first coin toss results in a head you are not worried, yes there is a small amount of evidence for a double-headed coin, but it is just a single coin toss. Two heads in a row still happens frequently. With three heads in a row your suspicions are peaked. By four heads in a row you are having serious doubts. Five heads in a row pretty well convinces you that you are being cheated. And, after seeing eight heads in a row you are reaching for the derringer in your boot.

We can augment this example with calculations of the *p*-value of so many heads under the null model of a fair coin. Fisherian significance testing is generally the first inferential tool that we are taught so many of us will have developed intuitions on *p*-values. In the calculation of the *p*-values, the null model is the fair coin model. Evidence is often measured as a likelihood ratio. The table shows the ratio of the likelihood of the double headed coin model given the data to the likelihood of the fair coin model given the same data. We can scaffold these intuitions into greater understanding of the evidence contained in differences in information criteria, ΔIC = (2*Log(Likelihood ratio)). Selecting a specific IC, such as AIC or SIC, would introduce a penalty term for the number of parameters and amount of data (Eq. 2).

Consecutive heads	*p*-value	Likelihood ratio	ΔIC	Evidence intuition
1	0.5	2	1.39	Very weak
2	0.25	4	2.77	Weak
3	0.125	8	4.16	Marginal
4	0.063	16	5.55	Moderate
5	0.031	32	6.93	Strong
6	0.016	64	8.32	Very strong
7	0.008	128	9.70	Extremely strong
8	0.004	256	11.09	Overwhelming

Expectedly, there is a common trend between the *p*-value and ΔIC. As the evidence grows for a two-headed coin, the *p*-value gets smaller, while the ΔIC value increases. In Fisherian *p*-value testing, we would have selected a threshold for the observed data (say 0.05) that beyond which we would reject the null model (hypothesis) in favor of the alternative. Interpretation of *p*-values is generally not condoned as a strength of evidence. With the ΔIC, we have a gradient from which to draw our inferences.

We see at a *p*-value of 0.031, the ΔIC is 6.93. For our study, we selected ΔSIC(7) for our intervals – meaning models and values of the slope parameter within this bound should be retained for further consideration with more data. Models and values of the slope parameters outside this bound have strong evidence against those models giving rise to the observed data (relative to the best model) and can therefore be subsequently dismissed.

the observed data. For purpose of our study, we provide ΔSIC(7) intervals for each trial and for the best model. In practice, models with parameter values falling within the ΔSIC interval are cases where, given the data in hand, the scientist cannot strongly differentiate between the models within the bound, and all of these models should be retained and further scrutinized with additional data (**Box 2**, D6).

RESULTS

Using the slopes estimated for each trial (**Table 1**), the distribution of values with fitted normal curve is shown in **Figure 3**. The mean slope parameter value is 0.94 (SE 0.04), which is unexpectedly different than the 0.79 slope estimated from the synthesis provided by Clarke and Johnston (1999). One explanation for this difference is because many of the studies used in our analysis were initially conducted to test the SMR of similar body sized fish at different temperatures. As indicated by trial 28 (**Table 1**), small sample size ($n = 8$) can result in biologically unrealistic estimates ($\hat{\beta} = -0.21$).

The best model selected using ΔSIC came from model suite 3 with a random intercept and random slope, but with a common slope parameter of $\hat{\beta} = 0.89$ (SE 0.021). However, a common slope and random intercept model had a ΔSIC = 1.5, and is thus not strongly distinguishable from the best model. The correlation of random slope with random intercept was −0.86, indicating that as the intercept increases in value, the slope decreases in value. This correlation is likely due to noise.

The value of universal slope is consistent (0.87–0.89) across all model suites and there is strong evidence (ΔSIC > 7) against fixed mechanistic based values of the metabolic scaling rate of 0.67, 0.75, and 1 across all modeling suites. **Figure 2**, along with

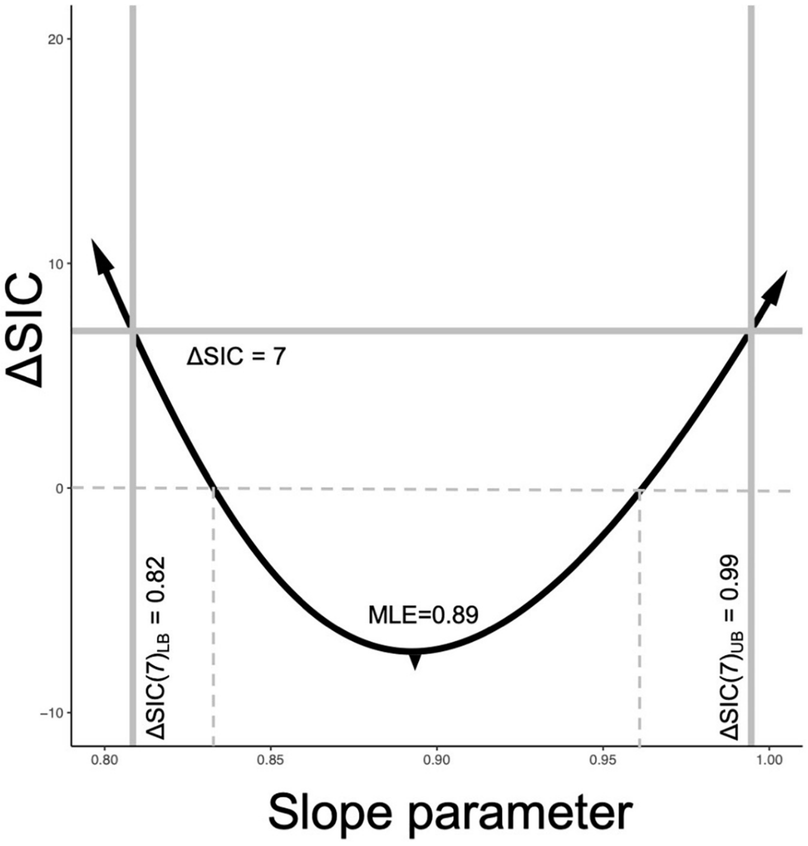


FIGURE 2 | SIC interval formulation. The black line is the ΔSIC as a function of the slope parameter space. The reference model is always the model with the estimated slope parameter. When ΔSIC = 7 (solid gray horizontal line intersects the ΔSIC), this defines the lower $\Delta SIC(7)_{LB}$ and upper $\Delta SIC(7)_{UB}$ of the information criterion interval. Values of the ΔSIC near the MLE can be negative values due to the penalization term (Eq. 2). This example is drawn from the best fit model of our study with an MLE for the slope parameter of $\hat{\beta} = 0.89$ with ΔSIC(7) = (0.82, 0.99). When the ΔSIC is negative, that is below the dashed line, the fixed slope models are favored, but weakly. When the ΔSIC is positive but less than 7, fitted slope model is favored, but weakly.

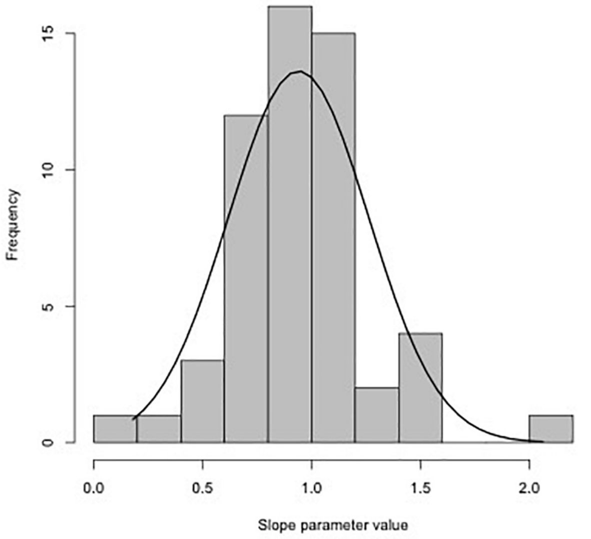


FIGURE 3 | Distribution of slopes estimated in **Table 1** for all 55 trials. Mean of the distribution is 0.94 (SE 0.04).

being a conceptualization of an ΔSIC(7) interval, is generated under the best model and the interval spans 0.81 to 0.99.

Figure 4 shows the ΔSIC(7) interval for each trial ordered by n*VAR(ln(weight)), from smallest values at the bottom to larger values at the top. This ordering is a regression experimental design component where few data points and/or small ranges in body mass result in small values indicating the lower precision of the slope parameter estimate. With exception of Cunner (Trial 3) where the ΔSIC(7) interval spans 0.81 to 0. 98, all other trials span at least one of the mechanistic hypotheses of 0.67, 0.75, or 1.

As outlined in the data section, all observations included in this study were collected under conditions to ensure data quality. However, not all studies were designed to estimate metabolic scaling relationship (a slope parameter) and some had few data points and/or did not cover a large breadth of fish body masses. The trials of Cunner, however, were designed for testing the metabolic scaling relationship and could potentially drive the overall value observed by the best model. As such, we conducted an additional analysis after removing the Cunner data and found the same estimate of the metabolic scaling relationship. See **Box 4** for more details. The metabolic scaling relationship of $\hat{\beta} = 0.87 - 0.89$ for fish has very little uncertainty, is robust

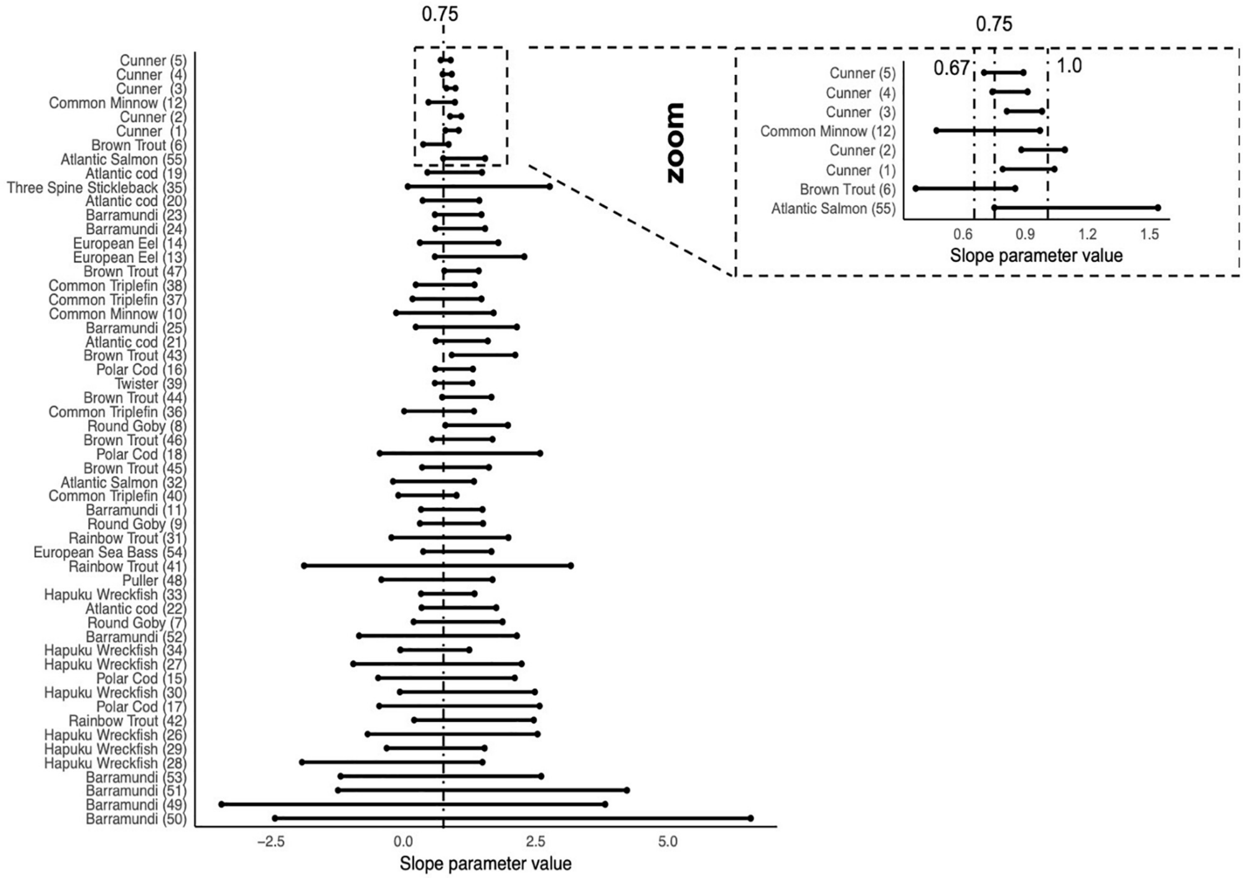


FIGURE 4 | ΔSIC(7) intervals for all trials ordered by n*VAR (Log(weight)). Trials with small n*VAR(Log(weight)) are expected to have wide intervals because the lack coverage of fish mass or have small samples sizes. As studies have larger n*VAR(Log(weight)), the ΔSIC(7) intervals become smaller and have the ability to exclude hypotheses of the slope, β = 0.67, 0.75, and 1. With the exception of the Cunner(3) trial, all other trials capture at least one of the hypotheses, the most common being β = 0.75, the dashed line in the figure. The zoom inset shows trials with relatively narrow ΔSIC(7) and dashed lines at β = 0.67, 0.75, and 1.0.

across models, and emerges when any trial or species is dropped from the analysis.

DISCUSSION

The evidence function (ΔSIC) approach we implemented here has led to selecting a best model; a mixed effect model with random slope and random intercept by species and an estimated correlation between random effects (**Table 2**, Model 9). However, we cannot dismiss the possibility that the model structure may only have a random species intercepts and common slope as witnessed by this alternative model having a ΔSIC = 1.5 (**Table 2**, Model 1). Models across all suites that represent mechanistic hypotheses of a scaling relationship of 0.67, 0.75, and 1 are dismissed with *very strong* evidence, ΔSIC > 8.4 (**Table 2**, **Box 2**). As such, our inference is that surface area limitations (β = 0.67), distribution network limitations (β = 0.75), and low cost demands on maintenance and routine activity (β = 1) are not exclusively driving the metabolic scaling relationship in fish.

However, the evidence for a $\hat{\beta} = 0.87$ to 0.89 universal scaling relationship is strong and presumably robust as indicated by similarity of the MLE for this parameter across all modeling suites and narrow bound of the ΔSIC(7) interval (**Figure 2**). Both fixed values are more than five standard deviations from the estimated common slope, and thus the chances are less than 1 in 1,000,000 that the common slope would have a β as small as 0.75 or as great as 1. If the data do come from the random slopes model, then it would be an extraordinary event for any species to have a β as low as 0.75, but perhaps as much as 6% of species might have a β as great 1. Accordingly, both DEB and MLB hypotheses warrant further consideration to determine the mechanism of metabolic scaling in fishes.

In many ways, the evidentialist approach is not that different from what is being applied in the multi-model literature, albeit with the meaningful caveat that an evidence function (**Box 2**) is being applied. The SIC is well studied, familiar to many, and also extractable from all the analyses we conducted in the R programing language. As such, the ΔSIC is readily accessible to scientists wishing to implement an

BOX 4 | Is it just cunner?

The Cunner study (Norin and Gamperl, 2018; *n* = 66 per trial for five trials) and the Common Minnow (McLean et al., 2018; *n* = 122 for one trial) both have large sample sizes compared to the other studies and were intentionally designed to estimate the metabolic scaling. Consequently, when we look at the span of ΔSIC(7) interval estimated for each trial as a function of the regression experimental design measure *n* * the variance of Log(weight) (**Figure 4**), we see the Cunner and the Common minnow studies have distinctly smaller ΔSIC(7) intervals. This raises the question, would our conclusion about the value of intraspecific scaling coefficient if the cunner study or the Common Minnow study were not included in our analysis.

We estimated the slope parameter under the best fit model and then calculated the resulting ΔSIC(7) interval by systematically withholding data by trial and then by species. For trials (**Figure Box 4.1**), they are ordered by value of *n* * the variance of Log(weight) from largest to smallest. For species (**Figure Box 4.2**), the ordering is alphabetical.

As expected, Cunner trials and the Common Minnow trial indeed do influence the MLE and the ΔSIC(7) intervals (**Figure Box 4.1**), but not so much as to capture the mechanistic hypotheses of 0.75 and 0.67 (dashed lines). However, the full model inference that the mechanistic hypothesis of metabolic scaling = 1 can be excluded from further consideration is sensitive to inclusion of some trials and species (**Figure Box 4.1** and **Figure Box 4.2**). In all trials, the value of $\hat{\beta} = 0.89$ is captured. Other trials with smaller values of *n* * the variance of Log(weight) have virtually no influence on the either the point estimate or the uncertainty measure.

The story is similar if we aggregate trials by species (**Figure Box 4.2**) and then systematically withhold all data from a species. Notably, withholding species data generally broadens the ΔSIC(7) interval with slight variation in the MLE that ranges from 0.89 to 0.9. Yet withholding a species from the analysis does not change the conclusion of the statistical inference that the slope of the metabolic scaling relationship is not 0.75 or 0.67. However, absence of Barramundi, Common Triplefin, Cunner, Hapuku Wreckfish, or Rainbow Trout results in a wider ΔSIC(7) interval that just captures the metabolic scaling of 1, and would, in the absence of any of these species, motivate further consideration of this mechanistic hypothesis.

While some of the trials were designed to test the metabolic scaling relationship, they do not unduly drive the conclusion. But maybe more importantly, the effect of many studies that are less suited to individually test the relationship (**Table 1**), together can provide meaningful insights into the metabolic scaling relationship.

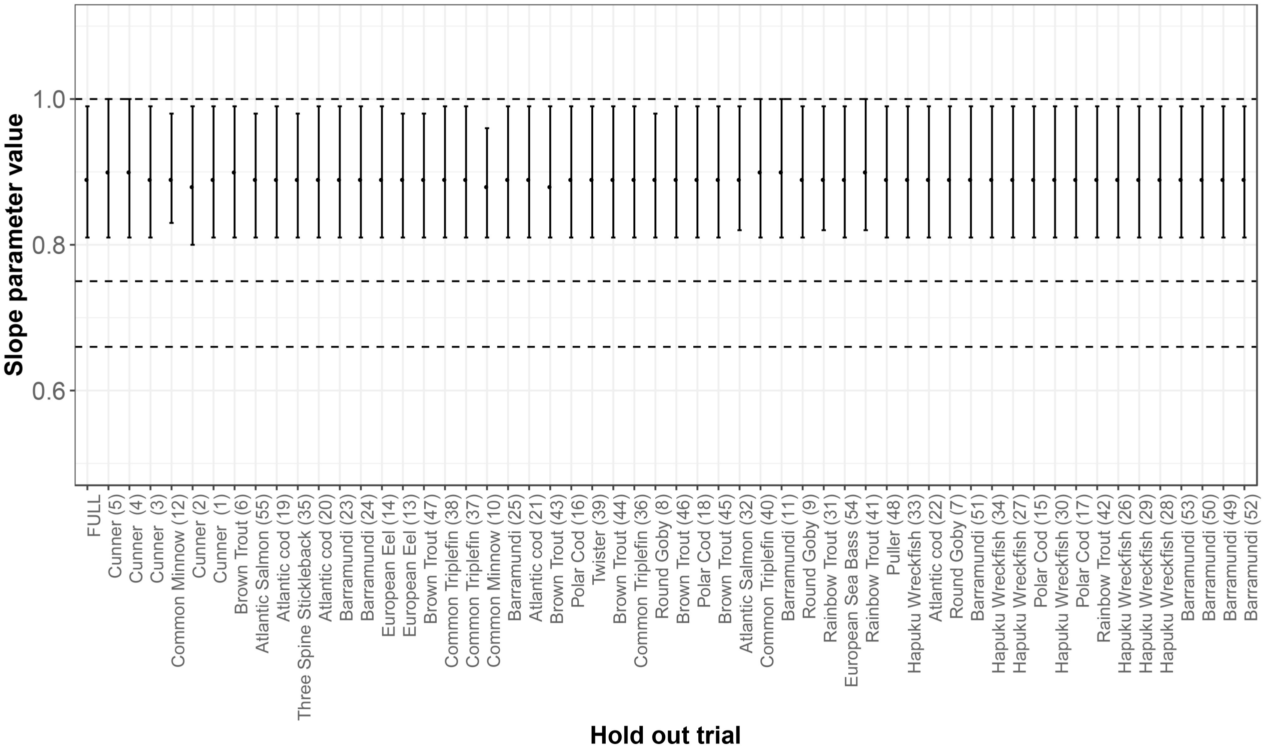


FIGURE BOX 4.1 | MLE of the slope parameter and ΔSIC(7) interval estimated by systematically withholding each trial. FULL is the MLE and interval with all data considered. Absence of any one data set does not drive our conclusion. However, absence of trial 4, 5, 11, 40, or 41 would suggest keeping the mechanistic hypothesis of metabolic scaling at 1 in the suite of models to be considered further.

evidentialist approach. While additional coding is required to produce ΔSIC intervals, this effort takes only elementary coding to automate. It must be noted, that the SIC for large sample sizes makes it difficult for new parameters to enter the model. In this analysis, our primary conclusion is that a model with β estimated as an extra free parameter is better than any of the models with β specified at any of the values of 0.67, 0.75, or 1.0. Thus the use of the SIC as a criterion as opposed to the AIC makes our conclusions conservative.

The other major contribution of the evidentialist approach underscored in this is the imperative to combine data sets such that evidence does not come from a single critical test, but rather from the accumulation of trials and critical

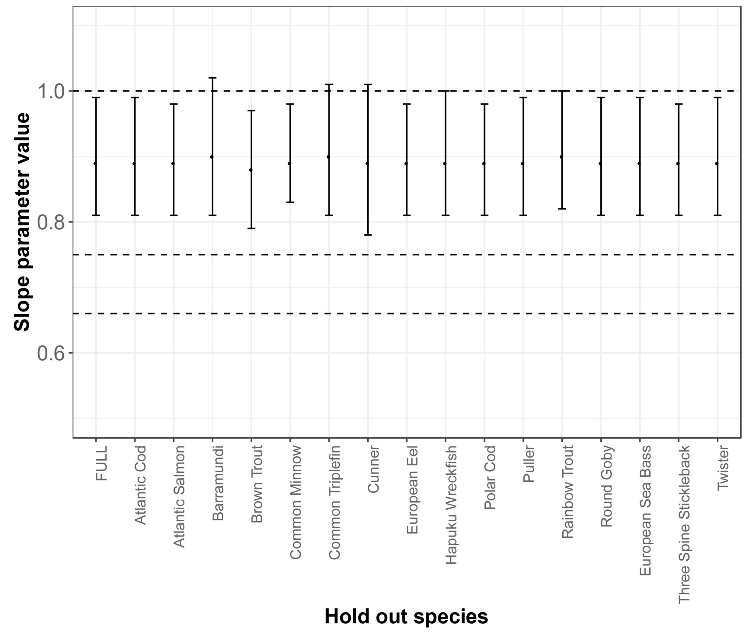


FIGURE BOX 4.2 | MLE of the slope parameter and ΔSIC(7) interval estimated by systematically withholding each species. FULL is the MLE and interval with all data considered. Absence of any one data set does not drive our conclusion. However, absence of Barramundi, Common Triplefin, Cunner, Hapuku Wreckfish, or Rainbow Trout would suggest keeping the mechanistic hypothesis of metabolic scaling at 1 in the suite of models to be considered further.

tests (See D6 of **Box 2**). Here we combined 55 trials across 16 species comprising 1456 observations. While this would normally form the basis of meta-analysis, this breadth of diverse data is desirable by allowing for a random effect of species to make our inferences across the population of fish species. If we look at each trial individually, we see that all but one trial (Cunner 4), captures one of the mechanistic hypotheses of 0.67, 0.75, or 1. In contrast when we look at the aggregate, none of these hypotheses are supported (**Box 4**).

Both the quantity and quality of metabolic rate data included in the metadata are important and can shape the conclusions of the study. Several extensive metadata analyses include mean metabolic rate values from close to a 100 or more species (e.g., Clarke and Johnston, 1999; White and Seymour, 2003; Glazier, 2005; Killen et al., 2010); however, the methods and quality of the data is not always rigorously considered. Metabolic rate is one of the most commonly investigated whole animal physiological performance metrics (Nelson, 2016), but different methods are more or less time and resource-intensive and can over-estimate SMR (Chabot et al., 2016). Furthermore, it is logistically challenging to obtain robust SMR measurements on many fish species, for example, large-bodied open ocean pelagic species or deep-sea fishes. Our study is unique because we only included standard metabolic rate data following specific and stringent criteria with each data point representing *individual* standard metabolic rate instead of reported species mean values. Future work could address how our (and others) conclusions change if the quality control criteria are relaxed.

There are many covariates that may be important predictors for species-specific scaling slopes and intercepts. While we tried to capture fishes across a broad latitudinal range with varying life histories, we did not examine life history factors such as species ecological activity (athletic vs. sedentary; Killen et al., 2010), growth rate, reproductive investment or strategy (e.g., fecundity), maximum body size, maximum age, or even environmental factors such as habitat (e.g., benthic vs. pelagic; freshwater vs. marine; Killen et al., 2010), or latitude (e.g., tropic vs. temperate vs. polar). Furthermore, temperature governs metabolism in ectotherms such as fish. Given this, all our models included temperature as an independent significant predictor of metabolic rates in fish (ΔSIC = 8.1 for best model compared to best model without temperature; **Box 5**). Recently, Lindmark et al. (2018) presented temperature-dependent intraspecific metabolic allometry, where MR increased with temperature to a lesser extent in larger fish. Furthermore, these effects scale to higher levels of organization, including from populations (population response-models), to ecosystems (MTE; Brown et al., 2004). We evaluated temperature effects and an interaction with log(weight) (See **Box 5**) with a ΔSIC = 7.2 compared to the best model. We can dismiss further consideration of an interaction of temperature with weight under the model suites evaluated. However, these temperature-size dependent effects on MR are mixed across and within species, and require more

TABLE 2 | Application of evidence functions using the Schwarz Information Criterion (SIC).

Model	Description	k	Log(L)	SIC	Δ SIC
Model Suite 1: Random Intercept models.					
Fixed effect: Weight, Temp; Random effect: Species; Nested effect: Trial					
1	β = Free*	6	73.9	−104.1	1.5
2	β = 0.67	5	−15.5	67.4	173
3	β = 0.75	5	42.1	−47.7	57.9
4	β = 1	5	35.7	−35	70.6
Model Suite 2: Estimated species intercept models					
Fixed effect: Weight, Temp and Species, Nested effect: Trial					
5	β = Free*	20	−69.8	−69.8	35.8
6	β = 0.67	19	96.8	96.8	202.4
7	β = 0.75	19	−16.3	−16.3	89.3
8	β = 1	19	−0.4	−0.4	105.2
Model Suite 3: Random intercepts with random slopes					
Fixed effect: Weight and Temp; Random effect: Species and Slope; Nested effect: Trial					
9	β = Random, Free*	8	81.9	−105.6	0
10	β = Random, 0.67	7	50.1	−49.2	56.4
11	β = Random, 0.75	7	64.6	−78.1	27.5
12	β = Random, 1	7	74.1	−97.2	8.4
Model Suite 4: Estimated species intercept models with random slopes					
Fixed effect: Weight, Temp and Species; Random effect: Slope; Nested effect: Trial					
13	β = Random, Free*	21	107.8	−62.6	43
14	β = Random, 0.67	20	45.2	55.3	160.9
15	β = Random, 0.75	20	78.6	−11.5	94.1
16	β = Random, 1	20	85.7	−25.6	80

R output with parameter values found in the* ***Supplementary Material**.*

research and metabolic scaling data from species in polar and tropical environments.

Norin and Gamperl (2018) provided a compelling study to measure allometric scaling for Cunner. It adhered to all the characteristics of a robust and well-designed study (White and Kearney, 2014) to estimate the scaling relationship, with ample breadth of fish mass, 68 observations per trial, and five trials (**Table 1**). What makes this study notable is their conclusion that no universal scaling relationship exists. We offer a few explanations for this apparent contradiction. Our inference is broadly applicable to fish, while theirs is limited to Cunner. Put simply, we are measuring evidence at a different inferential level for a universal scaling constant. If we look at the values of the SIC(7) intervals for all Cunner trials (**Figure 4**) they appear to be very similar. The intervals are {0.79, 1.04}, {0.88, 1.09}, {0.81, 0.98}, {0.74, 0.91}, and {0.7, 0.89}, and all SIC(7) intervals capture the values 0.88 and 0.89. Clearly our estimate of $\hat{\beta}$ = 0.89 from the best model with a random slope should be considered as a possible universal scaling for Cunner as well as other fish. As such, our results are consistent with Norin and Gamperl (2018), and their insightful suggestions about the need to consider species-specific scaling relationships when building fish population dynamic models that apply metabolic scaling exponents, should be heeded.

Scaling relationships are at times considered key tools for predicting the effects of global change on fisheries (e.g., Cheung et al., 2008), or as tools to estimate how abundant fish might be in the absence of fishing (e.g., Jennings and Blanchard, 2004). Therefore, variation in the scaling relationship between body size and metabolism have clear implications for how we predict fish populations will respond to changes in the environment or changes in body size distributions. As we move forward and seek to predict the consequences of changes in fish populations, the assumption of a universal scaling exponent, while attractive and generalizable may either under or overestimate a species sensitivity to changes in the environment. Given the evidence for species-specific variation in scaling relationships provided in our study, stock assessments seeking to integrate scaling relationships into forecasts may therefore benefit from species-specific values. While theoretical underpinnings have motivated application of a scaling relationship of β = 0.75, our data show that fisheries models that blindly adopt this parameter may be ultimately misleading.

We had some concern that the species distribution would be non-normally distributed, but there was no evidence from our analysis of this concern. However, those models may be useful for assessing the importance of species phylogenetics to metabolic scaling. The variance for the random species intercept model was 0.19 with a residual of 0.047. Similarly, from the random slopes model, the variance for the random intercept was 0.24, the random slope was 0.005, and the residual variance was 0.044 (see **Supplementary Material** for model outputs). Both measurement error in SMR and real inter-species variability contribute to the variability in $\hat{\beta}$. Variance components are notoriously difficult to tease apart, that is they are only weakly estimable (Ponciano et al., 2012). An estimate of the magnitude of measurement error in SMR would contribute greatly to the ability of further studies to accurately estimate the inter-specific variability in $\hat{\beta}$.

This study does not address the question of inter-specific metabolic scaling. This would entail a study of scaling of intra-specific intercepts with mean species body size. As we do not have accurate estimates of mean body size for these species, we cannot yet address this issue. Future work could use the random affects models or the estimated species intercepts models (model suites 2 and 4, **Table 2**) to evaluate if species relatedness and/or taxonomy are significant factors explaining species random effects variability.

Many of the studies used in this analysis were not designed to test the metabolic relationship, which is evident from the standard errors of the regression coefficients for individual trials (**Table 1**). However, under our data criteria, these studies had precise measurement of SMR, body mass, and temperature. The inclusion of these trials added unique

BOX 5 | Changes in SMR due to temperature and body mass.

Temperature has been thought to play a critical role regulating individual metabolic rate in fishes (Fry, 1947), where metabolic rates typically increase as temperature increases. As a consequence, all of the models we have considered so far have included a temperature effect. We can evaluate the effect of temperature more fully by considering six modification of models in suites 1 and 3 (**Table 2**). The first model (M17) is a random intercept model without inclusion of the temperature variable. The second model (M1) includes temperature (**Table 2**, Suite 1, Model 1), and the third model (M18) adds an interaction term of temperature with log(weight). These are all fixed slope models.

Including a log(weight) by temperature interaction is equivalent to saying that scaling of log(SMR) with log(weight) is itself a linear function of temperature. This is how we express it in the table below. The derivation of the standard error is discussed in the **Supplementary Material**.

The second group of models are built upon the random slopes model (**Table 2**, Suite 3, Model 9). The first model (M19) is absent temperature, the second model (M9) is the same as **Table 2**, Model 9 with an intercept defined by the temperature, and the third model (M20) has an interaction of temperature with log(weight). Using maximum likelihood fitting and extracting the SIC values, we can apply the same evidence function approach to evaluate the influence of temperature on intraspecific metabolic scaling. Model output is provided in the **Supplementary Material**.

Consistent with our previous model selection effort, M9 (**Table 2**), which includes temperature with a metabolic scaling coefficient (0.89), has the lowest SIC score. Models M17 and M19 without temperature include have ΔSIC > 7, which indicates temperature is a significant factor as the literature suggests. As observed previously, there is moderate evidence for M9 over M1, but not so much as to discourage future studies from considering a constant slopes model. Both M18 and M20 with interactions between temperature and log(weight) have ΔSIC > 7. Under the best model (M9), the expected metabolic scalings at 0°C, 15°C, and 30°C are 0.89, 0.9, and 0.91, respectively.

The conclusion from our focused study of temperature is that temperature is a critical factor to consider in modeling fish metabolic rate as there is *strong* evidence (**Box 3**) for including temperature in the intercept of the scaling relationship. Future work on evaluating the effect of temperature should expand the coverage of the temperature range with more polar and tropical fish species. Additional data at the endpoints of the temperature range will improve inferences about the scaling relationship and the evidence for, or against, a log(weight) by temperature interaction.

TABLE BOX 5.1 | Model selection using ΔSIC along with parameter estimates of for the metabolic scaling relationship. For models M18 and M20, the parameter estimate and standard error are a function of temperature.

Model	SIC	ΔSIC	$\hat{\beta}$	SE($\hat{\beta}$)
M17	−80.6	25	0.87	0.015
M1	−104.1	1.5	0.87	0.015
M18	−97.5	7.6	0.83 + 0.00257 (temp)	$\sqrt{0.0023 + (8.59 \times 10^{-6}) \times \text{temp}^2 + 2 \times -0.00013 \times \text{temp}}$
M19	−86.1	8.1	0.91	0.025
M9	−105.6	0	0.89	0.021
M20	−98.4	7.2	0.87 + 0.00106 (temp)	$\sqrt{0.0033 + (1.07 \times 10^{-5}) \times \text{temp}^2 + 2 \times -0.00017 \times \text{temp}}$

species to support the evaluation of a species random effect, which ultimately allows us to make inferences from this model across fish species. Given that some of these trials are ill-suited in themselves to critically test the metabolic relationship, due to low sample size or narrow range of body masses, this may be contributing to selection of the random slope model. Future studies that implement an evidentialist approach with additional data sets collected using appropriate experimental designs to uncover the allometric scaling relationship will likely reconcile if species requires a random slope.

Simulations to understand data requirements for robust analysis of interspecific metabolic scaling relationships suggest that the data should include 100–150 species spanning 3–4 orders of magnitude range in body size (White and Kearney, 2014). One approach to finding or estimating a universal intraspecific scaling constant is to take the average from the distribution of estimated slopes from each trial (e.g., **Figure 3** in the current study, 0.916, SE 0.04). This approach, while easy to implement by combing the literature, assumes that all data are created equal, but we know that each estimated slope, $\hat{\beta}$ comes with error, and some of the studies we included had relatively large standard errors (**Table 1**). Our data with fewer total species than most meta-analysis, but using individual data instead of species or trial means, proved to be sufficient to address the question concerning the universality of scaling relationship between fish body mass and metabolic rate.

The evidentialist approach is useful in addressing long-standing scientific debates (such as universal scaling relationships of metabolism), consistent with the practice of applied scientists, and relatively easy to implement using existing evidence functions and programing packages. It provides path forward for dismissing models (hypotheses) with little to no support, identifying and retaining hypotheses needing further evaluation, and provides a philosophy that emphasizes accumulation of evidence, through additional data and confronting that data with more complex models of how the nature works. We look forward to further refinement of the approach not only through philosophical insights and mathematical rigor, but through application of the approach to long-standing, pressing ecological and environmental science problems.

DATA AVAILABILITY

The data sets analyzed for this study, with exception of the measurement error data for lobster, are peer-reviewed and published (see **Table 1** for citations). Data sets are available from the originating author(s). The lobster data are available in the **Supplementary Material**.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of the manuscript. MT, CJ, EE, and KK conceptualized the project. CJ and MT conducted the analyses. EE and KK created the database. SC and KK collected the lobster data.

FUNDING

CJ was partially funded by UCSB's Office of Research. EE was partially funded by a UCSB Faculty Research Grant.

ACKNOWLEDGMENTS

We thank Tristan McArley, Javed Khan, Tommy Norin, Martin Boldsen, and Yangfan Zhang, from whom we personally received datasets, and the authors of all other studies who made their data available online. Jose Ponciano and Subhash Lele provided helpful comments and guidance regarding the application and explanation of evidence functions. Our reviewers provided insightful comments which led to demonstrable improvement.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2019.01166/full#supplementary-material

REFERENCES

Agutter, P. S., and Wheatley, D. N. (2004). Metabolic scaling: consensus or controversy? *Theor. Biol. Med. Model.* 1:13.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 60, 255–265. doi: 10.1093/biomet/60.2.255

Akaike, H. (1981). Likelihood of a model and information criteria. *J. Econom.* 16, 3–14. doi: 10.1016/0304-4076(81)90071-3

Auer, S. K., Dick, C. A., Metcalfe, N. B., and Reznick, D. N. (2018). Metabolic rate evolves rapidly and in parallel with the pace of life history. *Nat. Commun.* 9:14. doi: 10.1038/s41467-017-02514-z

Auer, S. K., Salin, K., Rudolf, A. M., Anderson, G. J., and Metcalfe, N. B. (2015). The optimal combination of standard metabolic rate and aerobic scope for somatic growth depends on food availability. *Funct. Ecol.* 29, 479–486. doi: 10.1111/1365-2435.12396

Bandyopadhyay, P. S., Brittan, G., and Taper, M. L. (2016). *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference*. Berlin: Springer Briefs in Philosophy of Science.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48.

Behrens, J. W., van Deurs, M., and Christensen, E. A. (2017). Evaluating dispersal potential of an invasive fish by the use of aerobic scope and osmoregulation capacity. *PloS One* 12:e0176038. doi: 10.1371/journal.pone.0176038

Bokma, F. (2004). Evidence against universal metabolic allometry. *Funct. Ecol.* 18, 184–187. doi: 10.1111/j.0269-8463.2004.00817.x

Boldsen, M. M., Norin, T., and Malte, H. (2013). Temporal repeatability of metabolic rate and the effect of organ mass and enzyme activity on metabolism in European eel (*Anguilla anguilla*). *Comp. Biochem. Phys. Part A Mol. Integr. Physiol.* 165, 22–29. doi: 10.1016/j.cbpa.2013.01.027

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135. doi: 10.1016/j.tree.2008.10.008

Brett, J. R., and Glass, N. R. (1973). Metabolic rates and critical swimming speeds of sockeye salmon (*Oncorhynchus nerka*) in relation to size and temperature. *J. Fish. Board Can.* 30, 379–387. doi: 10.1139/f73-068

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., and West, G. B. (2004). Toward a metabolic theory of ecology. *Ecology* 85, 1771–1789. doi: 10.1890/03-9000

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection, and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer Science, and Business Media.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304. doi: 10.1177/0049124104268644

Burnham, K. P., Anderson, D. R., and Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav. Ecol. Sociobiol.* 65, 23–35. doi: 10.1007/s00265-010-1084-z

Burton, T., Killen, S. S., Armstrong, J. D., and Metcalfe, N. B. (2011). What causes intraspecific variation in resting metabolic rate and what are its ecological consequences? *Proc. R. Soc. B Biol. Sci.* 278, 3465–3473. doi: 10.1098/rspb.2011.1778

Caselle, J. E., Davis, K., and Marks, L. M. (2018). Marine management affects the invasion success of a non-native species in a temperate reef system in California, USA. *Ecol. Lett.* 21, 43–53. doi: 10.1111/ele.12869

Chabot, D., Steffensen, J. F., and Farrell, A. P. (2016). The determination of standard metabolic rate in fishes. *J. Fish Biol.* 88, 81–121. doi: 10.1111/jfb.12845

Cheng, C.-L., and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*, First Edn. London: Arnold.

Cheung, W. W., Close, C., Lam, V., Watson, R., and Pauly, D. (2008). Application of macroecological theory to predict effects of climate change on global fisheries potential. *Mar. Ecol. Prog. Ser.* 365, 187–197. doi: 10.3354/meps07414

Clarke, A., and Johnston, N. M. (1999). Scaling of metabolic rate with body mass and temperature in teleost fish. *J. Anim. Ecol.* 68, 893–905. doi: 10.1046/j.1365-2656.1999.00337.x

Collins, G. M., Clark, T. D., and Carton, A. G. (2016). Physiological plasticity v. inter-population variability: understanding drivers of hypoxia tolerance in a tropical estuarine fish. *Mar. Freshwater Res.* 67, 1575–1582.

Collins, G. M., Clark, T. D., Rummer, J. L., and Carton, A. G. (2013). Hypoxia tolerance is conserved across genetically distinct sub-populations of an iconic, tropical Australian teleost (*Lates calcarifer*). *Conserv. Physiol.* 1:cot029. doi: 10.1093/conphys/cot029

Cooper, B., Adriaenssens, B., and Killen, S. S. (2018). Individual variation in the compromise between social group membership and exposure to preferred temperatures. *Proc. R. Soc. Lond. B Biol. Sci.* 285:20180884. doi: 10.1098/rspb.2018.0884

Dennis, B., Ponciano, J. M., Taper, M. L., and Lele, S. R. (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Evol.*

Dunn, R. P., Baskett, M. L., and Hovel, K. A. (2017). Interactive effects of predator and prey harvest on ecological resilience of rocky reefs. *Ecol. Appl.* 27, 1718–1730. doi: 10.1002/eap.1581

Eliason, E. J., Clark, T. D., Hague, M. J., Hanson, L. M., Gallagher, Z. S., Jeffries, K. M., et al. (2011). Differences in thermal tolerance among sockeye salmon populations. *Science* 332, 109–112. doi: 10.1126/science.1199158

Eliason, E. J., Higgs, D. A., and Farrell, A. P. (2007). Effect of isoenergetic diets with different protein and lipid content on the growth performance and heat increment of rainbow trout. *Aquaculture* 272, 723–736. doi: 10.1016/j.aquaculture.2007.09.006

Elith, J., Burgman, M. A., and Regan, H. M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Model.* 157, 313–329. doi: 10.1016/s0304-3800(02)00202-8

Enquist, B. J., and Niklas, K. J. (2001). Invariant scaling relations across tree-dominated communities. *Nature* 410, 655–660. doi: 10.1038/35070500

Farrell-Gray, C. C., and Gotelli, N. J. (2005). Allometric exponents support a 3/4-powerscaling law. *Ecology* 86, 2083–2087. doi: 10.1890/04-1618

Fry, F. E. J. (1947). *Effects of the Environment on Animal Activity*. Toronto, ON: University of Toronto Press, 1–60.

Gillies, C. S., Hebblewhite, M., Nielsen, S. E., Krawchuk, M. A., Aldridge, C. L., Frair, J. L., et al. (2006). Application of random effects to the study of resource selection by animals. *J. Anim. Ecol.* 75, 887–898. doi: 10.1111/j.1365-2656.2006.01106.x

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M., and Charnov, E. L. (2001). Effects of size and temperature on metabolic rate. *Science* 293, 2248–2251. doi: 10.1126/science.1061967

Glazier, D. S. (2005). Beyond the '3/4-power law': variation in the intra-and interspecific scaling of metabolic rate in animals. *Biol. Rev.* 80, 611–662.

Glazier, D. S. (2008). Effects of metabolic level on the body size scaling of metabolic rate in birds and mammals. *Proc. R. Soc. Lond. B Biol. Sci.* 275, 1405–1410. doi: 10.1098/rspb.2008.0118

Glazier, D. S. (2018). Effects of contingency versus constraints on the body-mass scaling of metabolic rate. *Challenges* 9:4. doi: 10.3390/challe9010004

Hubbard, R., and Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am. Stat.* 57, 171–178. doi: 10.1198/0003130031856

Isaac, N. J., and Carbone, C. (2010). Why are metabolic scaling exponents so controversial? Quantifying variance and testing hypotheses. *Ecol. Lett.* 13, 728–735. doi: 10.1111/j.1461-0248.2010.01461.x

Jennings, S., and Blanchard, J. L. (2004). Fish abundance with no fishing: predictions based on macroecological theory. *J. Anim. Ecol.* 73, 632–642. doi: 10.1111/j.0021-8790.2004.00839.x

Johnston, I. A., and Dunn, J. (1987). Temperature acclimation and metabolism in ectotherms with particular reference to teleost fish. *Symp. Soc. Exp. Biol.* 41, 67–93.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* 30, 3050–3056. doi: 10.1002/sim.4323

Khan, J. R., Johansen, D., and Skov, P. V. (2018a). The effects of acute and long-term exposure to CO2 on the respiratory physiology and production performance of Atlantic salmon (*Salmo salar*) in freshwater. *Aquaculture* 491, 20–27. doi: 10.1016/j.aquaculture.2018.03.010

Khan, J. R., Lazado, C. C., Methling, C., and Skov, P. V. (2018b). Short-term feed and light deprivation reduces voluntary activity but improves swimming performance in rainbow trout *Oncorhynchus mykiss*. *Fish Physiol. Biochem.* 44, 329–341. doi: 10.1007/s10695-017-0438-0

Khan, J. R., Pether, S., Bruce, M., Walker, S. P., and Herbert, N. A. (2014). Optimum temperatures for growth and feed conversion in cultured hapuku (*Polyprion oxygeneios*)—is there a link to aerobic metabolic scope and final temperature preference? *Aquaculture* 430, 107–113. doi: 10.1016/j.aquaculture.2014.03.046

Khan, J. R., Pether, S., Bruce, M., Walker, S. P., and Herbert, N. A. (2015). The effect of temperature and ration size on specific dynamic action and production performance in juvenile hapuku (*Polyprion oxygeneios*). *Aquaculture* 437, 67–74. doi: 10.1016/j.aquaculture.2014.11.024

Killen, S. S. (2014). Growth trajectory influences temperature preference in fish through an effect on metabolic rate. *J. Anim. Ecol.* 83, 1513–1522. doi: 10.1111/1365-2656.12244

Killen, S. S., Atkinson, D., and Glazier, D. S. (2010). The intraspecific scaling of metabolic rate with body mass in fishes depends on lifestyle and temperature. *Ecol. Lett.* 13, 184–193. doi: 10.1111/j.1461-0248.2009.01415.x

Killen, S. S., Glazier, D. S., Rezende, E. L., Clark, T. D., Atkinson, D., Willener, A. S., et al. (2016). Ecological influences and morphological correlates of resting and maximal metabolic rates across teleost fish species. *Am. Nat.* 187, 592–606. doi: 10.1086/685893

Kleiber, M. (1932). Body size and metabolism. *Hilgardia* 6, 315–353. doi: 10.3733/hilg.v06n11p315

Kooijman, S. A. L. M. (1993). *Dynamic Energy Budgets in Biological Systems*. Cambridge: Cambridge University Press.

Kunz, K. L., Frickenhaus, S., Hardenberg, S., Johansen, T., Leo, E., Pörtner, H. O., et al. (2016). New encounters in Arctic waters: a comparison of metabolism and performance of polar cod (*Boreogadus saida*) and Atlantic cod (*Gadus morhua*) under ocean acidification and warming. *Polar Biol.* 39, 1137–1153. doi: 10.1007/s00300-016-1932-z

Lele, S. (2004). "Evidence functions, and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper, and S. R. Lele, (Chicago, IL: The University of Chicago Press), 191–216. doi: 10.7208/chicago/9780226789583.003.0007

Lighton, J. R. (2018). *Measuring Metabolic Rates: A Manual for Scientists*. Oxford: Oxford University Press.

Lindmark, M., Huss, M., Ohlberger, J., and Gårdmark, A. (2018). Temperature-dependent body size effects determine population responses to climate warming. *Ecol. Lett.* 21, 181–189. doi: 10.1111/ele.12880

Lorah, J., and Womack, A. (2019). Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behav. Res. Methods* 51, 440–450. doi: 10.3758/s13428-018-1188-3

Maino, J. L., Kearney, M. R., Nisbet, R. M., and Kooijman, S. A. (2014). Reconciling theories for metabolic scaling. *J. Anim. Ecol.* 83, 20–29. doi: 10.1111/1365-2656.12085

McArley, T. J., Hickey, A. J., and Herbert, N. A. (2017). Chronic warm exposure impairs growth performance and reduces thermal safety margins in the common triplefin fish (*Forsterygion lapillum*). *J. Exp. Biol.* 120(Pt 19), 3527–3535. doi: 10.1242/jeb.162099

McArley, T. J., Hickey, A. J., and Herbert, N. A. (2018). Hyperoxia increases maximum oxygen consumption and aerobic scope of intertidal fish facing acutely high temperatures. *J. Exp. Biol.* 221:jeb189993. doi: 10.1242/jeb.189993

McLean, S., Persson, A., Norin, T., and Killen, S. S. (2018). Metabolic costs of feeding predictively alter the spatial distribution of individuals in fish schools. *Curr. Biol.* 28, 1144–1149. doi: 10.1016/j.cub.2018.02.043

Metcalfe, N. B., Van Leeuwen, T. E., and Killen, S. S. (2016). Does individual variation in metabolic phenotype predict fish behaviour and performance? *J. Fish Biol.* 88, 298–321. doi: 10.1111/jfb.12699

Moses, M. E., Hou, C., Woodruff, W. H., West, G. B., Nekola, J. C., Zuo, W., et al. (2008). Revisiting a model of ontogenetic growth: estimating model parameters from theory and data. *Am. Nat.* 171, 632–645. doi: 10.1086/587073

Murtaugh, P. A. (2014). In defense of P values. *Ecology* 95, 611–617.

Nadler, L. E., Killen, S. S., McClure, E. C., Munday, P. L., and McCormick, M. I. (2016). Shoaling reduces metabolic rate in a gregarious coral reef fish species. *J. Exp. Biol.* 219, 2802–2805. doi: 10.1242/jeb.139493

Nelson, J. A. (2016). Oxygen consumption rate v. rate of energy utilization of fishes: a comparison and brief history of the two measurements. *J. Fish Biol.* 88, 10–25. doi: 10.1111/jfb.12824

Nishii, R. (1988). Maximum-likelihood principle and model selection when the true model is unspecified. *J. Multivar. Anal.* 27, 392–403. doi: 10.1016/b978-0-12-580205-5.50032-x

Norin, T., and Clark, T. D. (2017). Fish face a trade-off between 'eating big' for growth efficiency and 'eating small' to retain aerobic capacity. *Biol. Lett.* 13:20170298. doi: 10.1098/rsbl.2017.0298

Norin, T., and Gamperl, A. K. (2018). Metabolic scaling of individuals vs. populations: Evidence for variation in scaling exponents at different hierarchical levels. *Funct. Ecol.* 32, 379–388. doi: 10.1111/1365-2435.12996

Norin, T., and Malte, H. (2011). Repeatability of standard metabolic rate, active metabolic rate and aerobic scope in young brown trout during a period of moderate food availability. *J. Exp. Biol.* 214, 1668–1675. doi: 10.1242/jeb.054205

Norin, T., and Malte, H. (2012). Intraspecific variation in aerobic metabolic rate of fish: relations with organ size and enzyme activity in brown trout. *Physiol. Biochem. Zool.* 85, 645–656. doi: 10.1086/665982

Norin, T., Malte, H., and Clark, T. D. (2016). Differential plasticity of metabolic rate phenotypes in a tropical fish facing environmental change. *Funct. Ecol.* 30, 369–378. doi: 10.1111/1365-2435.12503

Ponciano, J. M., Burleigh, G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using Data Cloning. *Syst. Biol.* 61, 955–972. doi: 10.1093/sysbio/sys055

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Searle, S. R. (1971). *Linear models* (No. 04; QA279, S4.). New York, NY: John Wiley & Sons, Inc.

Sunday, J. M., Bates, A. E., and Dulvy, N. K. (2010). Global analysis of thermal tolerance and latitude in ectotherms. *Proc. R. Soc. B Biol. Sci.* 278, 1823–1830. doi: 10.1098/rspb.2010.1295

Taper, M. L., and Lele, S. R. (2011). Evidence, evidence functions, and error probabilities. *Philos. Stat.* 7, 513–532. doi: 10.1016/b978-0-444-51862-0.50015-0

Taper, M. L., and Marquet, P. A. (1996). How do species really divide resources? *Am. Nat.* 147, 1072–1086. doi: 10.1086/285893

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

West, G. B., Brown, J. H., and Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. *Science* 276, 122–126. doi: 10.1126/science.276.5309.122

White, C. R., and Kearney, M. R. (2013). Determinants of inter-specific variation in basal metabolic rate. *J. Comp. Physiol. B* 183, 1–26. doi: 10.1007/s00360-012-0676-5

White, C. R., and Kearney, M. R. (2014). Metabolic scaling in animals: methods, empirical results, and theoretical explanations. *Compr. Physiol.* 4, 231–256. doi: 10.1002/cphy.c110049

White, C. R., Phillips, N. F., and Seymour, R. S. (2005). The scaling and temperature dependence of vertebrate metabolism. *Biol. Lett.* 2, 125–127. doi: 10.1098/rsbl.2005.0378

White, C. R., and Seymour, R. S. (2003). Mammalian basal metabolic rate is proportional to body mass2/3. *Proc. Natl. Acad. Sci.* 100, 4046–4049. doi: 10.1073/pnas.0436428100

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.

Zhang, Y., Mauduit, F., Farrell, A. P., Chabot, D., Ollivier, H., Rio-Cabello, A., et al. (2017). Exposure of European sea bass (Dicentrarchus labrax) to chemically dispersed oil has a chronic residual effect on hypoxia tolerance but not aerobic scope. *Aquat. Toxicol.* 191, 95–104. doi: 10.1016/j.aquatox.2017.07.020

Zhang, Y., Timmerhaus, G., Anttila, K., Mauduit, F., Jørgensen, S. M., Kristensen, T., et al. (2016). Domestication compromises athleticism and respiratory plasticity in response to aerobic exercise training in Atlantic salmon (*Salmo salar*). *Aquaculture* 463, 79–88. doi: 10.1016/j.aquaculture.2016.05.015

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 21 October 2019
doi: 10.3389/fevo.2019.00372

Errors in Statistical Inference Under Model Misspecification: Evidence, Hypothesis Testing, and AIC

Brian Dennis[1]*, *José Miguel Ponciano*[2], *Mark L. Taper*[2,3] *and Subhash R. Lele*[4]

[1] Department of Fish and Wildlife Sciences and Department of Statistical Science, University of Idaho, Moscow, ID, United States, [2] Biology Department, University of Florida, Gainesville, FL, United States, [3] Department of Ecology, Montana State University, Bozeman, MT, United States, [4] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

The methods for making statistical inferences in scientific analysis have diversified even within the frequentist branch of statistics, but comparison has been elusive. We approximate analytically and numerically the performance of Neyman-Pearson hypothesis testing, Fisher significance testing, information criteria, and evidential statistics (Royall, 1997). This last approach is implemented in the form of evidence functions: statistics for comparing two models by estimating, based on data, their relative distance to the generating process (i.e., truth) (Lele, 2004). A consequence of this definition is the salient property that the probabilities of misleading or weak evidence, error probabilities analogous to Type 1 and Type 2 errors in hypothesis testing, all approach 0 as sample size increases. Our comparison of these approaches focuses primarily on the frequency with which errors are made, both when models are correctly specified, and when they are misspecified, but also considers ease of interpretation. The error rates in evidential analysis all decrease to 0 as sample size increases even under model misspecification. Neyman-Pearson testing on the other hand, exhibits great difficulties under misspecification. The real Type 1 and Type 2 error rates can be less, equal to, or greater than the nominal rates depending on the nature of model misspecification. Under some reasonable circumstances, the probability of Type 1 error is an increasing function of sample size that can even approach 1! In contrast, under model misspecification an evidential analysis retains the desirable properties of always having a greater probability of selecting the best model over an inferior one and of having the probability of selecting the best model increase monotonically with sample size. We show that the evidence function concept fulfills the seeming objectives of model selection in ecology, both in a statistical as well as scientific sense, and that evidence functions are intuitive and easily grasped. We find that consistent information criteria are evidence functions but the MSE minimizing (or efficient) information criteria (e.g., AIC, AICc, TIC) are not. The error properties of the MSE minimizing criteria switch between those of evidence functions and those of Neyman-Pearson tests depending on models being compared.

Keywords: model misspecification, evidential statistics, evidence, error rates in model selection, Kullback-Leibler divergence, hypothesis testing, Akaike's information criterion, model selection

OPEN ACCESS

Edited by:
Peng Liu,
Institute of Remote Sensing and Digital Earth (CAS), China

Reviewed by:
Venkata Krishna Jandhyala,
University of Western Ontario, Canada
Haruhiko Ogasawara,
Otaru University of Commerce, Japan

****Correspondence:***
Brian Dennis
brian@uidaho.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *07 March 2019*
Accepted: *17 September 2019*
Published: *21 October 2019*

Citation:
Dennis B, Ponciano JM, Taper ML and Lele SR (2019) Errors in Statistical Inference Under Model Misspecification: Evidence, Hypothesis Testing, and AIC. Front. Ecol. Evol. 7:372. doi: 10.3389/fevo.2019.00372

1. INTRODUCTION

1.1. Background

In the twentieth century, the bulk of scientific statistical inference was conducted with Neyman-Pearson hypothesis tests, a term which we broadly take to encompass significance testing, *P*-values, generalized likelihood ratio, and other special cases, adaptations, or generalizations. The central difficulty with interpreting NP tests is that the Type 1 error probability (usually denoted α) remains fixed regardless of sample size, rendering problematic the question of what constitutes evidence *for* the model serving as the null hypothesis (Aho et al., 2014; Murtaugh, 2014; Spanos, 2014). The fixed null error rate of hypothesis testing lies at the core of why model selection procedures based on hypothesis testing (such as stepwise regression and multiple comparisons) have always had the reputation of being jury-rigged contraptions that have never been fully satisfactory (Gelman et al., 2012). An additional problem with hypothesis tests arises from the "Type 3" error of model misspecification, in which neither the null nor the alternative hypothesis model adequately describes the data (Mosteller, 1948). The influence of model misspecification on all types of inference is under appreciated.

A substantial advance in late 20th century statistical practice was the development of information-theoretic indexes for model selection, namely the Akaike information criterion (AIC) and its variants (Akaike, 1973, 1974; Sakamoto et al., 1986; Bozdogan, 1987). The model selection criteria were slow in coming to ecology (Kemp and Dennis, 1991; Lebreton et al., 1992; Anderson et al., 1994; Strong et al., 1999) but have rapidly proliferated in the past 20 years, aided by a popular book (Burnham and Anderson, 2002) and journal reviews (Anderson et al., 2000; Johnson and Omland, 2004; Ward, 2008; Grueber et al., 2011; Symonds and Moussalli, 2011). Ecological practice has been indelibly shaped by the use of AIC and similar indexes (Guthery et al., 2005; Barker and Link, 2015). Notwithstanding, ecologists, traditionally introspective about and scrutinizing of statistical practices (Strong, 1980; Quinn and Dunham, 1983; Loehle, 1987; Yoccoz, 1991; Johnson, 1999; Hurlbert and Lombardi, 2009; Gerrodette, 2011), have generated much critique and discussion of the appropriate uses of the information criteria (Guthery et al., 2005; Richards, 2005; Arnold, 2010; Barker and Link, 2015; Cade, 2015). Topics of discussions have focused on the contrast of information-theoretic methods with frequentist hypothesis testing methods (Anderson et al., 2000; Stephens et al., 2005; Murtaugh, 2009) and with Bayesian statistical approaches (Link and Barker, 2006; Barker and Link, 2015).

In an apparently separate statistical development, the concept of statistical evidence was refined in light of the shortcomings of using as evidence quantities such as *P*-values that emerge from frequentist hypothesis testing (Royall, 1997, 2000; Taper and Lele, 2004, 2011; Taper and Ponciano, 2016). Crucial to the evidence concept was the idea of an evidence function (Lele, 2004). An evidence function is a statistic for comparing two models that has a suite of statistical properties, among them two critical properties: (a) both error probabilities (analogous to Type 1 and Type 2 error probabilities in hypothesis testing) approach zero asymptotically as the sample size increases, and (b) when the models are misspecified and the concept of "error" is generalized to be the selection of the model "farthest" from the true data-generating process, the two error probabilities still approach zero as sample size increases.

Despite widespread current usage of AIC-type indexes in ecology, the inferential basis and implications of the use of information criteria are not fully developed, and what is developed is commonly misunderstood (see the forum edited by Ellison et al., 2014). AIC-type indexes are used for different purposes: in some contexts in place of hypothesis testing, in some as evidence for model identification, in some as estimates of pseudo-Bayesian model probabilities, and in some purely as criteria for prediction (Anderson et al., 2001). Of concern is that few ecologists can explain the inferences they are conducting with AIC, as Akaike's (Akaike, 1973, 1974) mathematical argument is not an easy one, and more recent accounts (Bozdogan, 1987; Burnham and Anderson, 2002; Claeskens and Hjort, 2008) are heavily mathematical as well. A clear and accessible inferential concept is needed to promote confidence in and appropriate uses of the information-theoretic criteria. We believe that the concept of statistical evidence can serve well as the inferential basis for the uses of and distinctions among the AIC-type indexes.

This paper contrasts the concept of evidence with classical statistical hypothesis testing and demonstrates that many information-based indexes for model selection can be recast and interpreted as evidence functions. We show that the evidence function concept fulfills many seeming objectives of model selection in ecology, both in a statistical as well as scientific sense, and that evidence functions are intuitive and easily grasped. Specifically, the difference of two values of an information-theoretic index for a pair of models possesses in whole or in part the properties of an evidence function and thereby grants to the resulting inference a scientific warrant of considerable novelty in ecological practice.

Of particular importance is the desirable behavior of evidence functions under model misspecification, behavior which, as we shall show, departs sharply from that of statistical hypothesis testing. As ecologists grapple increasingly with issues related to multiple quantitative hypotheses for how data arose, the evidence function concept can serve as a scientifically satisfying basis for model comparison in observational and experimental studies.

1.2. Method of Analysis and Notation

For convenience we label as Neyman-Pearson (NP) hypothesis tests a broad collection of interrelated statistical inference techniques, including *P*-values for likelihood ratios, confidence intervals, and generalized likelihood ratio tests, that are connected to Neyman and Pearson's original work (Neyman and Pearson, 1933) and that form the core of modern applied statistics. We distinguish Fisher's use of *P*-values as a measure of the adequacy of the null hypothesis from the use of *P*-values in likelihood ratio hypothesis tests.

NP hypothesis tests and evidential comparisons are conducted in very different fashions and operate under different warrants. Thus, comparison is difficult. However, they both make inferences. One fundamental metric by which they can be compared is the frequency that inferences are made in error.

In this paper we seek to illuminate how the frequency of errors made by these methods is influenced by sample size, the differences among models being compared, and also the differences between candidate models and the true data generating process. Both of these inferential approaches can be, and generally are, constructed around a base of the likelihood ratio (LR). By studying the statistical behavior of the LR, we can answer our questions regarding frequency of error in all approaches considered.

Throughout this discussion, one observation (datum) is represented using the random variable X with $g(x)$ being the probability density function representing the true, data-generating process and $f(x)$ being the probability density function of an approximating model. If the observed process is discrete, $g(x)$ and $f(x)$ will represent probability mass functions. For simplicity we refer to these functions in both the discrete and continuous cases as pdf's, thinking of the abbreviation as "probability distribution function." The likelihood function under the true model, for n independent and identically distributed (*iid*) observations $x_1, x_2, \ldots x_n$ is written as

$$L_g = g(x_1)g(x_2)\ldots g(x_n), \tag{1}$$

whereas under the approximating model it is

$$L_f = f(x_1)f(x_2)\ldots f(x_n). \tag{2}$$

In cases where there are two candidate models $f_1(x)$ and $f_2(x)$, we write the respective likelihoods as L_1 and L_2 to avoid double subscript levels.

We make much use of the Kullback-Leibler (KL) divergence, one of the most commonly used measures of the difference between two distributions. The KL divergence of $f(x)$ from $g(x)$, denoted $K(g,f)$, is defined as the expected value of the log-likelihood ratio of g and f (for one observation) given that the observation came from the process represented by $g(x)$:

$$K\left(g,f\right) \equiv \mathrm{E}_g\left[\log\left(\frac{g\left(X\right)}{f\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int g\left(x\right)\log\left(\frac{g\left(x\right)}{f\left(x\right)}\right). \tag{3}$$

Here E_g denotes expectation with respect to the distribution represented by $g(x)$. The expectation is a sum or integral (or both) over the entire range of the random variable X, depending on whether the probability distributions represented by $g(x)$ and $f(x)$ are discrete or continuous (or both, such as for a zero-inflated continuous distribution). The functions must give positive probability to the same sets (along with other technical mathematical requirements which are usually met by the common models of ecological statistics).

The KL divergence is interpreted as the amount of information lost when using model $f(x)$ to approximate the data generating process $g(x)$ (Burnham and Anderson, 2001). Its publication (Kullback and Leibler, 1951) was a highpoint in the golden age of the study of "information theory." The KL divergence is always positive if $g(x)$ and $f(x)$ represent different distributions and is zero if the distributions are identical ("identical" in the mathematical sense that the distributions give the same probabilities for all events in the sample space). The KL divergence is not a mathematical distance measure in that $K(g,f)$ is not in general equal to $K(f,g)$.

The relevant KL divergences under correct model specification are for $f_1(x)$ and $f_2(x)$ with respect to each other:

$$K_{12} \equiv K\left(f_1,f_2\right) = \mathrm{E}_1\left[\log\left(\frac{f_1\left(X\right)}{f_2\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int f_1\left(x\right)\log\left(\frac{f_1\left(x\right)}{f_2\left(x\right)}\right), \tag{4}$$

$$K_{21} \equiv K\left(f_2,f_1\right) = \mathrm{E}_2\left[\log\left(\frac{f_2\left(X\right)}{f_1\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int f_2\left(x\right)\log\left(\frac{f_2\left(x\right)}{f_1\left(x\right)}\right). \tag{5}$$

By reversing numerator and denominator in the log function in Equation (5), one finds that

$$\mathrm{E}_2\left[\log\left(\frac{f_1\left(X\right)}{f_2\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int f_2\left(x\right)\log\left(\frac{f_1\left(x\right)}{f_2\left(x\right)}\right) = -K_{21}. \tag{6}$$

The convention for which subscript is placed first varies among references; we put the subscript of the reference distribution first as it is easy to remember.

The likelihood ratio (LR) and its logarithm figure prominently in statistical hypothesis testing as well as in evidential statistics. The LR is

$$\frac{L_1}{L_2} = \frac{f_1\left(x_1\right)f_1\left(x_2\right)\cdots f_1\left(x_n\right)}{f_2\left(x_1\right)f_2\left(x_2\right)\cdots f_2\left(x_n\right)}, \tag{7}$$

and the log-LR is

$$\log\left(\frac{L_1}{L_2}\right) = \sum_{i=1}^{n}\log\left(\frac{f_1\left(x_i\right)}{f_2\left(x_i\right)}\right). \tag{8}$$

In particular, the log-LR considered as a random variable is a sum of iid random variables, and its essential statistical properties can be approximated using the central limit theorem (CLT). The CLT (**Box 1**) provides an approximate normal distribution for a sum of iid random variables and requires the expected value (mean) and the variance of one of the variables. Under correct model specification, the observations came from either $f_1(x)$ or $f_2(x)$, and Equations (4)–(6) above give the expected value of one of the random variables in the sum as K_{12} or $-K_{21}$, depending on which model generated the data. Let σ_1^2 and σ_2^2 be the variances of $\log\left[f_1\left(X\right)/f_2\left(X\right)\right]$ with respect to each model:

$$\sigma_1{}^2 = \mathrm{V}_1\left[\log\left(\frac{f_1\left(X\right)}{f_2\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int f_1\left(x\right)\left(\log\left(\frac{f_1\left(x\right)}{f_2\left(x\right)}\right)\right)^2 - K_{12}{}^2. \tag{9}$$

$$\sigma_2{}^2 = \mathrm{V}_2\left[\log\left(\frac{f_1\left(X\right)}{f_2\left(X\right)}\right)\right] = \sum\!\!\!\!\!\!\int f_2\left(x\right)\left(\log\left(\frac{f_1\left(x\right)}{f_2\left(x\right)}\right)\right)^2 - K_{21}{}^2. \tag{10}$$

One can envision cases in which these variances might not exist, but we do not consider such cases here. The CLT, which requires that the variances be finite, provides the following approximations. If the data arise from f_1:

$$\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(nK_{12}, n{\sigma_1}^2), \tag{11}$$

$$\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(K_{12}, {\sigma_1}^2/n), \tag{12}$$

$$\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - K_{12}\right] \dot{\sim} \text{normal}(0,1). \tag{13}$$

Here, $\dot{\sim}$ means "is approximately distributed as." If the data arise from f_2:

$$\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(-nK_{21}, n{\sigma_2}^2), \tag{14}$$

$$\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(-K_{21}, {\sigma_2}^2/n), \tag{15}$$

$$\frac{\sqrt{n}}{\sigma_2}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) + K_{21}\right] \dot{\sim} \text{normal}(0,1)\,. \tag{16}$$

The device of using the CLT to study properties of the likelihood ratio is old and venerable and figures prominently in the theory of sequential statistical analysis (Wald, 1945).

A model, f, can be said to be misspecified if the distribution of data implied by the model (under best possible parameterization) differs in any way from the distribution of data under the true generating process. In the Kullback-Leibler divergence setting within which we are working, f is misspecified if $K(g,f) > 0$. A model set can be said to be misspecified if all of its member models are misspecified. Misspecification can have a host of causes, including omission of real covariates, inclusion of spurious covariates, incorrect specification of functional form, incorrect specification of process error structure, and incorrect specification of measurement error structure.

The approximate behavior of the LR under misspecification can also be represented with the CLT. To our two model candidates $f_1(x)$ and $f_2(x)$, we add the pdf $g(x)$ defined as the best possible mathematical representation of the distribution of data stemming from the actual stochastic mechanism generating the data, the unknown "truth" sought by scientists. We denote by ΔK the difference of the KL divergences of $f_1(x)$ or $f_2(x)$, from $g(x)$:

$$\Delta K = K\left(g,f_2\right) - K\left(g,f_1\right). \tag{17}$$

We note that ΔK could be positive, negative, or zero: if ΔK is positive, then f_1 is "closer" to truth, if ΔK is negative, then f_2 is closer to truth, and if ΔK is zero, then both models are equally distant from truth. To deploy the CLT, we need the mean and variance of the single-observation LR under misspecification. For the mean we have

$$\text{E}_g\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] = \sum\!\!\!\!\!\!\int g(x)\log\left(\frac{f_1(x)}{f_2(x)}\right) = \Delta K \tag{18}$$

The rightmost equality is established by adding and subtracting $\text{E}_g\left[\log\left(g(X)\right)\right]$. We denote the variance by ${\sigma_g}^2$ which becomes

$$\text{V}_g\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] \equiv {\sigma_g}^2 = \sum\!\!\!\!\!\!\int g(x)\left(\log\left(\frac{f_1(x)}{f_2(x)}\right)\right)^2 - (\Delta K)^2. \tag{19}$$

And now by the CLT, if the data did not arise from $f_1(x)$ or $f_2(x)$, but rather from some other pdf $g(x)$, we have:

$$\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(n\Delta K, n{\sigma_g}^2), \tag{20}$$

$$\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) \dot{\sim} \text{normal}(\Delta K, {\sigma_g}^2/n), \tag{21}$$

$$\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - \Delta K\right] \dot{\sim} \text{normal}(0,1). \tag{22}$$

Critical to the understanding, both mathematical and intuitive, of inference on models is an understanding of the topology of models. Once one has a concept of distances between models, a topology is implied. A model with one or more unknown parameters represents a whole family or set of models, with each parameter value giving a completely specified model. At times we might refer to a model set as a model if there is no risk of confusion. Two model sets can be only be arranged as nested, overlapping, or non-overlapping. A set of models can be correctly specified or misspecified depending on whether or not the generating process can be exactly represented by a model in the model set. Thus, there are only six topologies relating two model sets to the generating process (**Figures 1**, **2**).

2. EVIDENCE, NEYMAN-PEARSON TESTING, AND FISHER SIGNIFICANCE

2.1. Correctly Specified Models

In the canon of traditional statistical practices for comparing two candidate models, $f_1(x)$ and $f_2(x)$ say, with or without unknown parameters involved, the assumption that the data arose from either $f_1(x)$ or $f_2(x)$ is paramount. In this section we adopt this assumption of correctly specified models and compare the properties of statistical hypothesis testing with those of the evidence approach. The correct model assumption is the home turf, so to speak, of hypothesis testing, and so the comparison should by rights highlight the strengths of traditional statistical practice. To focus the issues with clarity we concentrate on the case in which $f_1(x)$ and $f_2(x)$ are statistically simple hypotheses (a.k.a. completely specified models, not to be confused with correctly specified models). In other words, we assume for now there are no unknown parameters in either model, deferring until later in this paper a discussion of unknown parameters.

2.1.1. Neyman-Pearson Statistical Hypothesis Tests

Neyman and Pearson (1933) proved in a famous theorem (the "Neyman-Pearson Lemma") that basing a decision between two completely specified hypotheses (H_1: the data arise from model

BOX 1 | The Central Limit Theorem (CLT)

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with common finite mean denoted $\mu = E(X_i)$, and finite variance denoted $\sigma^2 = E\left[(X_i - \mu)^2\right]$. Let

$$S_n = X_1 + X_2 + \ldots + X_n$$

be the sum of the X_is. Let $\mathrm{P}\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le s\right) = F_n(s)$ be the cumulative distribution function (CDF) for S_n standardized with its mean $n\mu$ and its variance $n\sigma^2$, equivalently written as $\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, where $\bar{X}_n = \frac{1}{n}S_n$. Then as $n \to \infty$, $F_n(s)$ converges to the cdf of a normal distribution with mean of 0 and variance of 1. We say that $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ converges in distribution to a random variable with a normal (0, 1) distribution, and we write

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \text{normal}\,(0, 1)\,.$$

From the CLT one can obtain normally distributed approximations for various quantities of interest:

$$S_n \,\dot{\sim}\, \text{normal}\left(n\mu, n\sigma^2\right),$$

$$\bar{X}_n = \frac{1}{n}S_n \,\dot{\sim}\, \text{normal}\left(\mu, \frac{\sigma^2}{n}\right).$$

Here, $\dot{\sim}$ means “approximately distributed as.” A general proof of the CLT as presented in advanced mathematical statistics texts typically uses the theory of characteristic functions (Rao, 1973).

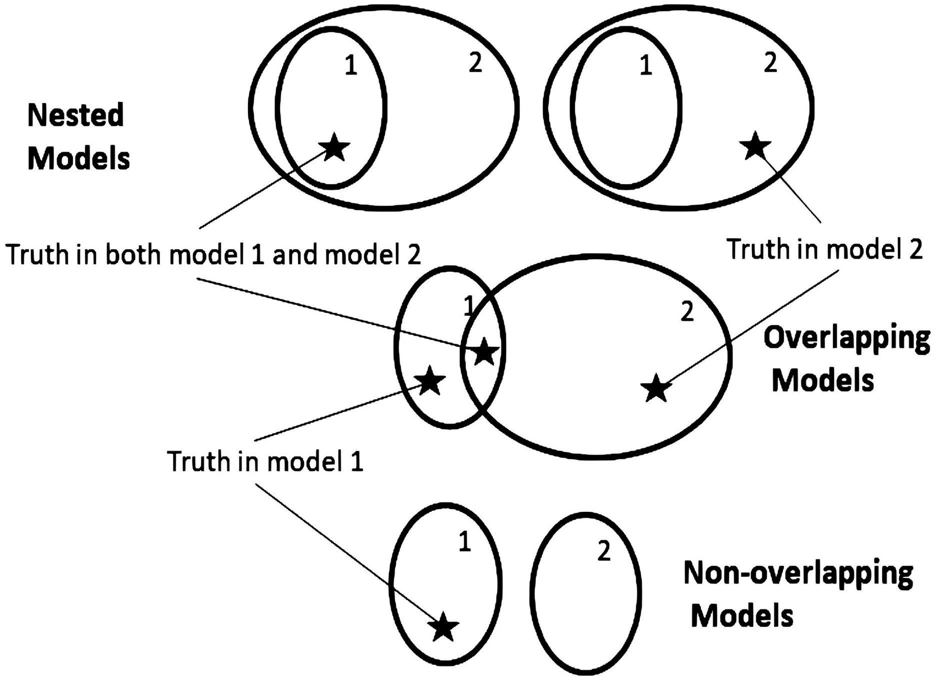


FIGURE 1 | Model topologies when models are correctly specified. Regions represent parameter spaces. Star represents the true parameter value corresponding to the model that generated the data. **Top**: a nested configuration would occur, for example, in the case of two regression models if the first model had predictor variables R_1 and R_2 while the second had predictor variables R_1, R_2, and R_3. **Middle**: an overlapping configuration would occur if the first model had predictor variables R_1 and R_2 while the second had predictor variables R_2 and R_3. Three locations of truth are possible: truth in model 1, truth in model 2, and truth in both models 1 and 2. **Bottom**: an example of a non-overlapping configuration is when the first model has predictor variables R_1 and R_2 while the second model has predictor variables R_3 and R_4.

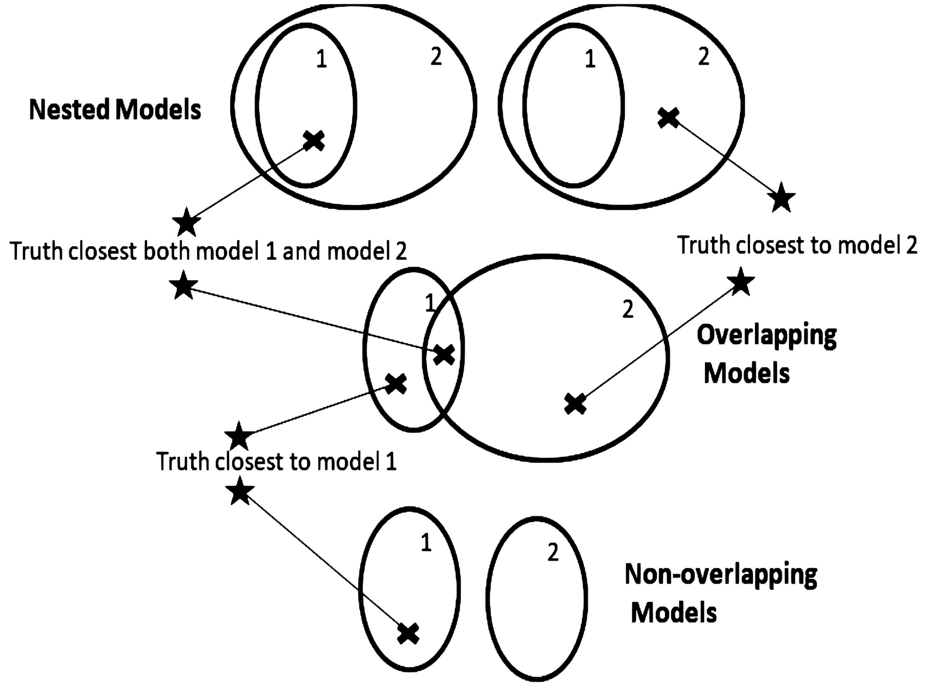


FIGURE 2 | Model topologies when models are misspecified. Regions represent parameter spaces. Star represents the true model that generated the data. Exes represent the point in the parameter space covered by the model set closest to the true generating process.

$f_1(x)$, and H_2: the data arise from model $f_2(x)$) on the likelihood ratio had certain optimal properties. Neyman and Pearson's LR decision rule has the following structure:

$$\begin{array}{ll}\text{decide on } H_1 \text{ if} & L_1/L_2 > c, \\ \text{decide on } H_2 \text{ if} & L_1/L_2 \leq c.\end{array} \tag{23}$$

Here the cutoff quantity (or critical value) c is determined by setting an error probability equal to a known constant (usually small), denoted α. Specifically, the conditional probability of wrongly deciding on H_2 given that H_1 is true is the "Type 1 error probability" and is denoted as α.

$$P\left(L_1/L_2 \leq c \mid H_1\right) = \alpha. \tag{24}$$

Often for notational convenience in lieu of the statement "H_i is true" we will simply write "H_i." Now, such a data-driven decision with fixed Type 1 error probability is the traditional form of a statistical hypothesis test. A test with a Type 1 error probability of α is said to be a size α test. The other error probability ("Type 2"), the conditional probability of wrongly deciding on H_1 given H_2, is usually denoted β:

$$P\left(L_1/L_2 > c \mid H_2\right) = \beta \tag{25}$$

The power of the test is defined as the quantity $1 - \beta$. Neyman and Pearson's theorem, stating that no other test of size α or less has power that can exceed the power of the likelihood ratio test, is a cornerstone of most contemporary introductions to mathematical statistics (Rice, 2007; Samaniego, 2014).

With the central limit theorem results (Equations 11–16), the error properties of the NP test can be approximated. To find the critical value c, we have under H_1:

$$\frac{L_1}{L_2} \leq c \Rightarrow \frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - K_{12}\right] \leq \frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log(c) - K_{12}\right], \tag{26}$$

and so the CLT tells us that

$$\alpha = P\left(\frac{L_1}{L_2} \leq c \mid H_1\right) \approx \Phi\left(\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log(c) - K_{12}\right]\right), \tag{27}$$

where $\Phi(z)$ is the cumulative distribution function (cdf) of the standard normal distribution. The approximate critical value c required for a size α test is then found by solving Equation (27) for c:

$$\begin{array}{l}\Phi\left(\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log(c) - K_{12}\right]\right) = \alpha \\ \Rightarrow \frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log(c) - K_{12}\right] = \Phi^{-1}(\alpha) = -z_\alpha \\ \Rightarrow c = \exp\left[\sqrt{n}\left(\sqrt{n}K_{12} - \sigma_1 z_\alpha\right)\right].\end{array} \tag{28}$$

Here $z_\alpha = \Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$ is the value of the $1 - \alpha$ quantile of the standard normal distribution. Thus, for error rate α to be fixed, the critical value as a function of n is seen to be a rapidly moving target.

The error probability β is approximated in similar fashion. We have, under H_2,

$$\begin{aligned}&\tfrac{L_1}{L_2} > c \Rightarrow \tfrac{\sqrt{n}}{\sigma_2}\left[\tfrac{1}{n}\log\left(\tfrac{L_1}{L_2}\right) + K_{21}\right] > \tfrac{\sqrt{n}}{\sigma_2}\left[\tfrac{1}{n}\log(c) + K_{21}\right]\\ &\Rightarrow \tfrac{\sqrt{n}}{\sigma_2}\left[\tfrac{1}{n}\log\left(\tfrac{L_1}{L_2}\right) + K_{21}\right] > \tfrac{\sqrt{n}}{\sigma_2}\left[\tfrac{1}{n}\log(c) + K_{21}\right],\end{aligned} \quad (29)$$

so that, after substituting for c,

$$\begin{aligned}\beta &= \mathrm{P}\left(\frac{L_1}{L_2} > c \mid \mathrm{H}_2\right) \approx 1 - \Phi\left(\frac{\sqrt{n}}{\sigma_2}(K_{12} + K_{21}) - \frac{\sigma_1}{\sigma_2}z_\alpha\right)\\ &= \Phi\left(\frac{\sigma_1}{\sigma_2}z_\alpha - \frac{\sqrt{n}}{\sigma_2}(K_{12} + K_{21})\right).\end{aligned} \quad (30)$$

It is seen that $\beta \rightarrow 0$ as sample size n becomes large. Here $K_{12} + K_{21}$ is an actual distance measure between $f_1(x)$ and $f_2(x)$ (Kullback and Leibler (1951); sometimes referred to as the "symmetric" KL distance) and can be regarded as the "effect size" as used in statistical power calculations.

Five important points about the Neyman-Pearson Lemma are pertinent here. First, the theorem itself is just a mathematical result and leaves unclear how it is to be used in scientific applications. The prevailing interpretation that emerged in the course of 20^{th} century science was that one of the hypotheses, H_1, would be accorded a special status ("the null hypothesis"), having its error probability α fixed at a known (usually small) value by the investigator. The other hypothesis, H_2, would be set up by experiment or survey design to be the only reasonable alternative to H_1. The other error probability, β, would be managed by study design characteristics (especially sample size), but would remain unknown and could at best only be estimated when the model contained parameters with unknown values. The hypothesis H_1 would typically play the role of the skeptic's hypothesis, as in the absence of an effect (absence of a difference in means, absence of influence of a predictor variable, absence of dependence of two categorical variables, etc.) under study. The other hypothesis, H_2, contains the effect under study and serves as the hypothesis of the researcher, who has the scientific charge of convincing a reasoned skeptic to abandon H_1 in favor of H_2.

Second, the theorem in its original form does not apply to models with unknown parameters. Various extensions were made during the ensuing decades, among them Wilks' (Wilks, 1938) and Wald's (Wald, 1943) theorems. The Wilks-Wald extension allows the test of two composite models (models with one or more unknown parameters) in which one model, taken as the null hypothesis, is formed from the other model (the alternative) by placing one or more constraints on the parameters. An example is a normal (μ, σ^2) distribution with both mean μ and variance σ^2 unknown as the model for the alternative hypothesis H_2, within which the null hypothesis model f_1 constrains the mean to be a fixed known constant: $\mu = \mu_1$. In such scenarios, the null model is "nested" within the alternative model, that is, the null is a special version of the alternative in which the parameters are restricted to a subset of the parameter space (set of all possible parameter values). Wilks' (Wilks, 1938) and Wald's (Wald, 1943) theorems together provide the asymptotic distribution of a function of the likelihood ratio under both the null and alternative hypotheses, with estimated parameters taken into account. The function is the familiar "generalized likelihood ratio statistic," usually denoted G^2, given by

$$G^2 = -2\log\left(\hat{L}_1/\hat{L}_2\right), \quad (31)$$

where $\hat{L}_1$ and $\hat{L}_2$ are the likelihood functions, respectively for models f_1 and f_2, with each likelihood maximized over all the unrestricted parameters in that model. The resulting parameter estimates, known as the maximum likelihood (ML) estimates, form a prominent part of frequentist statistics theory (Pawitan, 2001). Let θ be the vector of unknown parameters in model f_2 formed by stacking subvectors θ_{21} and θ_{22}. Likewise, let θ under the restricted model f_1 be formed by stacking the subvectors θ_{11} and θ_{12}, where θ_{11} is a vector of fixed, known constants (i.e., all values in θ_{21} are fixed) and θ_{12} is a vector of unknown parameters. Wald's (1943) theorem (after some mathematical housekeeping: Stroud, 1972) gives the asymptotic distribution of G^2 as a non-central chisquare(ν, λ) distribution, with degrees of freedom ν equal to the difference between the number of estimated parameters in f_2 and the number of estimated parameters in f_1, and non-centrality parameter λ being a statistical (Mahalanobis) distance between the true parameter values under H_2 and their restricted versions under H_1:

$$\lambda = n(\theta_{21} - \theta_{11})'\Sigma^{-1}(\theta_{21} - \theta_{11}). \quad (32)$$

Here Σ is a matrix of expected log-likelihood derivatives (details in Severini, 2000). Technically the true values θ_{21} must be local to the restricted values θ_{11}; the important aspects for the present are that λ increases with n as well as with the effect size represented by the distance $(\theta_{21} - \theta_{11})'\Sigma^{-1}(\theta_{21} - \theta_{11})$. With the true parameters equal to their restricted values, that is with H_1 governing the data production, the non-centrality parameter becomes zero, and Wald's theorem collapses to Wilks' theorem, which gives the asymptotic distribution of G^2 under H_1 to be an ordinary chisquare(ν) distribution. For linear statistical models in the normal distribution family (regression, analysis of variance, etc.), G^2 boils down algebraically into monotone functions of statistics with exact (non-central and central) t- or F-distributions, and so the various statistical hypothesis tests can take advantage of exact distributions instead of asymptotic approximations.

The concept of a confidence interval or region for one or more unknown parameters follows from Neyman-Pearson hypothesis testing in the form of a region of parameter values for which hypothesis H_1 would not be rejected at fixed error rate α. We remark further that although a vast amount of every day science relies on the Wilks-Wald extension of Neyman-Pearson testing (and confidence intervals), frequentist statistics theory prior to the 1970s had not provided much advice on what to do when the two models are not nested.

Certainly nowadays one could setup a model $f_1(x)$ as H_1 in a hypothesis test against a non-overlapping model $f_2(x)$ taken as H_2 and obtain the distributions of the generalized likelihood ratio under both models with simulation/bootstrapping.

Third, the Neyman-Pearson Lemma provides no guidance in the event of model misspecification. The theorem assumes that the data was generated under either H_1 or H_2. However, the "Type 3" error of basing inferences on an inadequate model family is widely acknowledged to be a serious (if not fatal) scientific drawback of the Neyman-Pearson framework (and parametric modeling in general, see Chatfield, 1995). Modern applied statistics rightly stresses rigorous checking of model adequacy with various diagnostic procedures, such as the standard battery of residual analyses in regression models. Deciding between two models based on diagnostic qualities has been a standard workaround in the situation mentioned above for which the two models are not nested. For instance, one might choose the model with the most homoscedastic residuals.

Fourth, the asymmetry of the error structure has led to difficulties in scientific interpretation of Neyman-Pearson hypothesis testing results. The difficulties stem from α being a fixed constant. A decision to prefer hypothesis H_2 over H_1 because the LR (Equation 23) is smaller than c is not so controversial. The H_2 over H_1 decision has some intuitively desirable statistical properties. For example, the error rate β asymptotically approaches 0 as the sample size n grows larger. Further, β asymptotically approaches 0 as model f_2 becomes "farther" from f_1 (in the sense of the symmetric KL distance $K_{12} + K_{21}$ as seen in Equation 30). Mired in controversy and confusion, however, is the decision to prefer H_1 over H_2 when the LR is larger than c. The value of c is set by the chosen value of the error rate α, using the probabilistic properties of model f_1. If a larger sample size is used, the LR has more terms, and the value of c necessary to attain the desired value of α changes. In other words, c depends on sample size n and moves in such a way as to keep α fixed (at 0.05 or whatever other value is used; Equation 28). The net effect is to leave the Neyman-Pearson framework without a mechanism to assess evidence *for* H_1, for no matter how far apart the models are or how large a sample size is collected, the probability of wrongly choosing H_2 when H_1 is true remains stuck at α.

Fifth, scientific practice rarely stops with just two models. In an analysis of variance, after an overall test of whether the means are different, one usually needs to sort out just who is bigger than whom. In a multiple regression, one is typically interested in which subset of predictor variables provide the best model for predicting the response variable. In a categorical data analysis of a multiway contingency table, one is often seeking to identify which combination of categorical variables and lower and higher order interactions best account for the survey counts. For many years (through the 1980s at least), standard statistical practice called for multiple models to be sieved through some (often long) sequence of Neyman-Pearson tests, through processes such as multiple pairwise comparisons, stepwise regression, and so on. It has long been recognized, however, that selecting among multiple models with Neyman-Pearson tests plays havoc with error rates, and that a pairwise decision tree of "yes-no's" might not lead to the best model among multiple models (Whittingham et al., 2006 and references therein). Using Neyman-Pearson tests for selection among multiple models was (admittedly among statisticians) a kludge to be used only until something better was developed.

2.1.2. Fisher Significance Analysis

R. A. Fisher never fully bought into the Neyman-Pearson framework, although generations of readers have debated about what exactly Fisher was arguing for, due to the difficulty of his writing style and opacity of his mathematics. Fisher rejected the scientific usefulness of the alternative hypothesis (likely in part because of the lurking problem of misspecification) and chose to focus on single-model decisions (resulting in lifelong battles with Neyman; see the biography by Box, 1978). Yea or nay, is model f_1 an adequate representation of the data? As in the Neyman-Pearson framework, Fisher typically cast the null hypothesis H_1 in the role of a skeptic's hypothesis (the lady cannot tell whether the milk or the tea was poured first). It was scientifically sufficient in this approach for the researcher to develop evidence to dissuade the skeptic of the adequacy of the null model. The inferential ambitions here are necessarily more limited, in that no alternative model is enlisted to contribute more insights for understanding the phenomenon under study, such as an estimate of effect size. As well, Fisher's null hypothesis approach preserves the Neyman-Pearson incapacitation when the null model is not contradicted by data, in that at best, one will only be able to say that the data are a plausible realization of observations that could be generated under H_1.

Fisher's principal tool for the inference was the *P*-value. For Fisher's preferred statistical distribution models, the data enter into the maximum likelihood estimate of a parameter in the form of a statistic, such as the sample mean. The implication is that such a statistic carries all the inferential information about the parameter; knowing the statistic's value is the same (for purposes of inference about the parameter) as knowing the values of all the individual observations. Fisher coined the term "sufficient statistic" for such a quantity. The null model in Fisher significance analysis is formed by constraining a parameter to a pre-specified value. In the tea testing example, the probability of correct identification is constrained to one half. Fisher's *P*-value is the probability that data drawn from the model H_1 yield a sufficient statistic as extreme or more extreme than the sufficient statistic calculated from the real data.

In absence of an alternative model, Fisher's strict *P*-value accomplishes an inference similar to what is called a goodness of fit test (or model adequacy test) in contemporary practice, as the inference seeks to establish whether or not the data plausibly could have arisen from model f_1. Accordingly, just about any statistic (besides a sufficient statistic) can be used to obtain a *P*-value, provided the distribution of the statistic can be derived or approximated under the model f_1. Goodness of fit tests therefore tend to multiply, as witnessed by the plethora of tests available for the normal distribution. To sort out the qualities of different goodness of fit tests, one usually has to revert to a Neyman-Pearson two-model framework to establish for what types of alternative models a particular test is powerful.

2.1.3. Neyman-Pearson Testing With P-values

P-values are, of course, routinely used in Neyman-Pearson hypothesis testing, but the inference is different from that made with Fisher significance. A *P*-value in the context of the generalized LR test above (Equation 31) is defined as the probability that, if the data generation process were to be repeated, the new value of the LR would exceed the one already observed, provided that the data were generated under H_1. Hinkley (1987) interprets the *P*-value as the Type 1 error rate that an ensemble of hypothetical experiments would have if their critical level c was set to the observation of this experiment. In the generalized LR test, the approximate *P*-value would simply be the area to the right of the observed value of G^2 under the chisquare pdf applicable for H_1-generated data. For Fisher's preferred statistical distributions (those with sufficient statistics, nowadays called exponential family distributions), the generalized LR statistic G^2 algebraically reduces to a monotone function of one or more sufficient statistics for the parameter or parameters under constraint in the model f_1. In the generalized likelihood ratio framework, the hypothesis test decision between H_1 and H_2 can be made by comparing the *P*-value to the fixed value of α, rejecting H_1 as a plausible origin of the data if the *P*-value is $\leq \alpha$.

In both Neyman-Pearson hypothesis testing and Fisher significance analysis, the *P*-value provides no evidence for model H_1. The *P*-value in the two-model framework has been thought of as an inverse measure of the "evidence" for H_2, as the distribution of the *P*-value under data generated by H_2 becomes more and more concentrated near zero as sample size becomes large or as model f_2 becomes "farther" from f_1. In the Fisher one-model framework an alternative model is unspecified. Consequently, a low *P*-value has been interpreted as "evidence" against H_1. However, the *P*-value under data generated by H_1 has a uniform distribution (because a continuous random variable transformed by its own cumulative distribution function has a uniform distribution) no matter what the sample size is or how far away the true data generating process is. Hence, as with NP tests, Fisher's *P*-value has no evidential value toward f_1, as any *P*-value is equally likely under H_1.

Ecologists use and discuss hypothesis testing in both the Fisher sense and the Neyman-Pearson sense, sometimes referring to both enterprises as "null hypothesis testing." The use of *P*-values, strongly argued for by some (Hurlburt and Lombardi 2009), does not in and of itself distinguish the two approaches. Rather, a specific alternative hypothesis, an estimable effect size, and (most controversially) a decision rule fixing a Type 1 error rate (i.e., comparing a *P*-value to α) identifies the analysis as more Neyman-Pearsonian than Fisherian. While Fisher himself originated the $P \leq 0.05$ tradition for judging whether a deviation is significant [... "it is convenient to draw the line at about the level at which we can say: 'Either there is something in the treatment, or a coincidence has occurred, such as does not occur more than once in twenty trials.'" Fisher (1926)], he was mostly casual about the cutoff and viewed *P*-values more as evidence against the null hypothesis in question. In ecology, null hypotheses in the Fisherian sense are seen, for instance, in analyses of species assembly patterns in ecological communities, such as in testing whether bird species groups on offshore islands could be modeled as randomly drawn from the mainland (Connor and Simberloff, 1979). By contrast, a field experiment aimed at demonstrating the existence of competition and estimating an effect size (Underwood, 1986) would take on a Neyman-Pearsonian flavor.

2.1.4. Equivalence Testing and Severity

Attempts have been made to modify the Neyman-Pearson framework to accommodate the concept of evidence for H_1. In some applied scientific fields, for example in pharmacokinetics and environmental science, the regulatory practice has created a burden of proof around models normally regarded as null hypothesis models: the new drug has an effect equal to the standard drug, the density of a native plant has been restored to equal its previous level (Anderson and Hauck, 1983; McDonald and Erickson, 1994; Dixon, 1998). Equivalence testing and non-inferiority testing (e.g., Wellek, 2010) are statistical methods designed to address the problem that "absence of a significant effect" is not the same as "an effect is significantly absent." In practice, the equivalence testing methods reverse the role of null and alternative hypotheses by specifying a parameter region that constitutes an acceptably small departure from the parameter's constrained value and then casting the region as the alternative hypothesis. Typically, two statistical hypothesis tests are required to conclude that the parameter is within the small region containing the constraint, such as two one-sided *t*-tests (to show that the parameter is bounded by each end of the region).

Another proposed solution for the evidence-for-the-null-hypothesis problem is the concept of severity (Mayo, 1996, 2018; Mayo and Spanos, 2006) and the closely related method of reverse testing (Parkhurst, 2001). Severity is a sort of *P*-value under a specified (or possibly estimated) version of the alternative hypothesis. It is the probability that a test statistic more extreme than the one observed would be obtained if the experiment were to be repeated, if the data were arising from model f_2 (with the particular effect size specified). In the generalized likelihood ratio framework, the severity would be calculated as the area to the right of the observed value of G^2 under the non-central chisquare pdf applicable for data generated under model f_2, with the non-centrality parameter set at a specified value. Thus, severity is a kind of attained power for a particular effect size. Also, severity is mostly discussed in connection with one-sided hypotheses, so that its calculation under the two-sided generalized likelihood ratio statistic is at best an approximation. However, if the effect size is substantial, the probability contribution from the "other side" is low, and the approximation is likely to be fine. In general, the severity of the test is related to the size of the effect, so care needs to be taken in the interpretation of the test.

For a given value of the LR, if the effect size is high, the probability of obtaining stronger evidence against H_1 is high, and the severity of the test against H_1 is high. "A claim is severely tested to the extent that it has been subjected to and passes a test that probably would have found flaws, were they present" (Mayo, 2018).

For both equivalence testing and severity, we are given procedures in which consideration of evidence requires two statistics and analyses. In the case of equivalence testing, we have a statistical test for each side of the statistical model specified by H_1, and for severity we have a statistic for H_2 and a statistic for H_1. Indeed, Thompson (2007), section 11.2, considers that for *P*-values to be used as evidence for one model over another, these must be used in pairs. There is evidence for H_1 relative to H_2 if the first *P*-value, say P_1, is large and the second *P*-value, say P_2, is small. The requirement for two analyses and two interpretations seems a disadvantageous burden for applications. More importantly, the equivalence testing and severity concepts do not yet accommodate the problems of multiple models or non-nested models.

2.1.5. Royall's Concept of Evidence

The LR statistic (Equation 7), as discussed by Hacking (1965) and Edwards (1972), can be regarded as a measure of *evidence* for H_1 and against H_2 (Edwards 1972 termed it *support*, but the word has a different technical meaning in probability and is better avoided here), or equivalently, an inverse measure of evidence for H_2 and against H_1. The evidence concept here is post-data in that the realized value of the LR itself, and not a probability calculated over hypothetical experiment repetitions, conveys the magnitude of the empirical scientific case for H_1 or H_2. However, restricting attention to just the LR itself leaves the prospect of committing an error unanalyzed; while scientists want to search for truth, they strongly want (for reasons partly sociological) to avoid being wrong.

Royall (1997, 2000) argued forcefully for greater use of evidence-based inferences in statistics, and to Hacking's and Edwards' frameworks he added formal procedures and consideration of errors. Royall's basic setup uses completely specified models as in Neyman-Pearson, but the conclusion about which model is favored by the data is based on fixed thresholds for the LR value, not thresholds determined by any error rate. The idea is to conclude there is strong evidence in favor of model H_1 when L_1 is k times L_2 and strong evidence in favor of H_2 when L_2 is k times L_1. Royall's conclusion structure in terms of the LR then has a trichotomy of outcomes:

$$\begin{array}{ll} L_1/L_2 \geq k: & \text{Strong evidence for } \mathrm{H}_1. \\ 1/k < L_1/L_2 < k: & \text{Weak or inconclusive evidence.} \\ L_1/L_2 \leq 1/k: & \text{Strong evidence for } \mathrm{H}_2. \end{array} \tag{33}$$

For k, values of 8, 20, or 32 are mentioned. The k value is chosen by the investigator, but unlike α in the Neyman-Pearson framework, k is not dependent on sample size. Viewed as evidence, LR is a post-data measure. The inference does not make appeals to hypothetical repeated sampling.

Royall (1997, 2000) moreover defines pre-data error rates which are potentially useful in experimental design and serve as reassurance that the evidential approach will not lead investigators astray too often. Suppose the data were generated by model f_1. It is possible that the LR could take a wayward value, leading to one of two possible errors in conclusion that could occur: (1) the LR could take a value corresponding to weak or inconclusive evidence (the error of weak evidence), or (2) the LR could take a value corresponding to strong evidence for H_2 (the error of misleading evidence). Given the data are generated by model f_1, the probabilities of the two possible errors are defined as follows:

$$\mathrm{P}\left(\text{weak evidence} \mid \mathrm{H}_1\right) = \mathrm{P}\left(1/k < L_1/L_2 < k \mid \mathrm{H}_1\right) = W_1 \tag{34}$$

$$\mathrm{P}\left(\text{misleading evidence} \mid \mathrm{H}_1\right) = \mathrm{P}\left(L_1/L_2 \leq 1/k \mid \mathrm{H}_1\right) = M_1. \tag{35}$$

Similarly, given the data are generated under H_2,

$$\mathrm{P}\left(\text{weak evidence} \mid \mathrm{H}_2\right) = \mathrm{P}\left(1/k < L_1/L_2 < k \mid \mathrm{H}_2\right) = W_2, \tag{36}$$

$$\mathrm{P}\left(\text{misleading evidence} \mid \mathrm{H}_2\right) = \mathrm{P}\left(L_1/L_2 \geq k \mid \mathrm{H}_2\right) = M_2. \tag{37}$$

The error probabilities M_1, M_2, W_1, and W_2 can be approximated with the CLT results for L_1/L_2 (Equations 11–16). Proceeding as before with the Neyman-Pearson error rates, we find that

$$M_1 \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(k\right) + K_{12}\right]\right), \tag{38}$$

$$M_2 \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_2}\left[\frac{1}{n}\log\left(k\right) + K_{21}\right]\right), \tag{39}$$

$$\begin{aligned} W_1 \approx\ & \Phi\left(\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(k\right) - K_{12}\right]\right) \\ & - \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(k\right) + K_{12}\right]\right), \end{aligned} \tag{40}$$

$$\begin{aligned} W_2 \approx\ & \Phi\left(\frac{\sqrt{n}}{\sigma_2}\left[\frac{1}{n}\log\left(k\right) - K_{21}\right]\right) \\ & - \Phi\left(-\frac{\sqrt{n}}{\sigma_2}\left[\frac{1}{n}\log\left(k\right) + K_{21}\right]\right). \end{aligned} \tag{41}$$

The error probabilities M_1, M_2, W_1, and W_2 depend on the models being compared, but it is easy to show that all four probabilities, as approximated by Equations (38–41), converge to zero as sample size n becomes large. For either hypothesis H_i $(i = 1, 2)$, the total error probability given by $M_i + W_i$ is additionally a monotone decreasing function of n, as for instance

$$M_1 + W_1 = \Phi\left(\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(k\right) - K_{12}\right]\right), \tag{42}$$

in which the argument of the cdf $\Phi\left(\cdot\right)$ is seen (by ordinary differentiation, assuming $k > 1$) to be monotone decreasing in n (the expression for $M_2 + W_2$ would have σ_2 and K_{21} in place of σ_1 and K_{12}).

The probability V_1 of strong evidence for model $f_1\left(x\right)$, given the data are indeed generated by model $f_1\left(x\right)$, becomes

$$V_1 = 1 - M_1 - W_1, \tag{43}$$

with $V_2 = 1 - M_2 - W_2$ defined in kind. Here V stands for veracity or veridicality (because of context, there should be no confusion with the variance operator). It follows from the monotone property of $M_i + W_i$ that V_i is a monotone increasing function of n. Furthermore, it is straightforward to show that $V_i > M_i$, $i = 1, 2$.

Note that V_1, M_1, and W_1 are not in general equal to their counterparts V_2, M_2, and W_2, nor should we expect them to be; frequencies of errors will depend on the details of the model generating the data. One model distribution with, say, a heavy tail could produce errors at a greater rate than a light-tailed model. The asymmetry of errors suggests possibilities of pre-data design to control errors. For instance, instead of LR cutoff points $1/k$ and k, one could find and use cutoff values k_1 and k_2 that render M_1 and M_2 nearly equal for a particular sample size and particular values of σ_1, K_{12}, σ_2, and K_{21}. Such design, however, will induce an asymmetry in the error rates (defined below) for misspecified models.

Interestingly, as a function of n, M_i $(i = 1, 2)$ increases at first, rising to a maximum value before decreasing asymptotically to zero. The value $\tilde{n}_1$ at which M_1 is maximized is found by maximizing the argument of the normal cdf in Equation (38):

$$\tilde{n}_1 = \frac{\log(k)}{K_{12}}, \tag{44}$$

with the corresponding maximum value of M_1 being

$$\tilde{M}_1 = \Phi\left(-\frac{2\sqrt{K_{12}\log(k)}}{\sigma_1}\right). \tag{45}$$

Expressions for $\tilde{n}_2$ and $\tilde{M}_2$ are similar and substitute K_{21} and σ_2 in place of the H_1 quantities. That the M_i functions would increase with n initially is counterintuitive at first glance. With just a few observations, the variability of the likelihood ratio is not big enough to provide much chance of misleading evidence, although the chance of weak evidence is high. As the sample size increases, the chance of misleading evidence grows at first, replacing some of the chance of weak evidence, before decreasing. It is the overall probability of either weak or misleading evidence, $W_i + M_i$, that decreases monotonically with sample size.

2.1.6. Illustration of the Concept of Evidence

We illustrate the error properties of evidence under correct model specification with an example. Suppose the values $x_1, x_2, \ldots, x_n$ are zeros and ones that arose as iid observations from a Bernoulli distribution with $P(X = 1) = p$. The pdf is $f(x) = p^x(1-p)^{1-x}$, where x is 0 or 1. The sum of the observations of course has a binomial distribution. We wish to compare hypothesis $\mathrm{H}_1\colon p = p_1$ with $\mathrm{H}_2\colon p = p_2$, where p_1 and p_2 are specified values. The log-likelihood ratio is

$$\log\left(\frac{L_1}{L_2}\right) = \left(\sum_{i=1}^{n} x_i\right)\log\left(\frac{p_1}{p_2}\right) + \left(n - \sum_{i=1}^{n} x_i\right)\log\left(\frac{1-p_1}{1-p_2}\right). \tag{46}$$

From Equations (4) and (9) we find that

$$K_{12} = p_1\log\left(\frac{p_1}{p_2}\right) + (1-p_1)\log\left(\frac{1-p_1}{1-p_2}\right), \tag{47}$$

$$\sigma_1{}^2 = p_1\left[\log\left(\frac{p_1}{p_2}\right)\right]^2 + (1-p_1)\left[\log\left(\frac{1-p_1}{1-p_2}\right)\right]^2 - K_{12}{}^2. \tag{48}$$

In the top panel of **Figure 3**, simulated values of the probability of strong evidence for model H_1, given by $V_1 = 1 - M_1 - W_1$, are compared with the values as approximated with the CLT (Equations 38, 40). The simulated values create a jagged curve due to the discrete nature of the Bernoulli distribution but are well-characterized by the CLT approximation. The lower panel of **Figure 3** portrays the probability of misleading evidence given by M_1 as a function of n. The discrete serrations are even more pronounced in the simulated values of M_1, and the CLT approximation (Equation 38) follows only the lower edges; the approximation could likely be improved (i.e., set toward the middle of the serrated highs and lows) with a continuity correction. The CLT nonetheless picks up the qualitative behavior of the functional form of M_1.

2.1.7. P-values, Severity, and Evidence

The concept of evidence allows re-interpretation of P-values in a clarifying manner. Suppose we denote by l_1/l_2 the realized (i.e., post-data) value of the LR, the lower case signaling the actual outcome rather than the random variable (pre-data) version of the LR denoted by L_1/L_2. The classical P-value is the probability, given the data arise from model H_1, that a repeat of the experiment would yield a LR value more extreme than the value l_1/l_2 that was observed. In our CLT setup, we can write

$$P = \mathrm{P}\left(\frac{L_1}{L_2} \leq \frac{l_1}{l_2} \mid \mathrm{H}_1\right) \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(\frac{l_2}{l_1}\right) + K_{12}\right]\right). \tag{49}$$

Comparing P with the expression for M_1 (Equation 38), we find that P is the probability of misleading evidence under model f_1 if the experiment were repeated and the value of k were taken as l_2/l_1.

If the value of l_1/l_2 is considered to be the evidence provided by the experiment, the value of P is a monotone function of l_1/l_2 and thereby might be considered to be an evidence measure on another scale. P however is seen to depend on other quantities as well: for a given value of l_1/l_2, P could be greater or less depending on the quantities n, K_{12}, and σ_1. Furthermore, K_{21} and σ_2 are left out of the value of P, giving undue influence to model f_1 in the determination of amount of evidence, a finger on the scale so to speak. The evidential framework therefore argues for the following distinction in the interpretation of P: the *evidence* is l_1/l_2, while P, like M_1, is a probability of misleading evidence, except that P is defined post-data.

In fairness to both models, we can define two P-values based on the extremeness of evidence under model f_1 and under model f_2:

$$P_1 = P\left(\frac{L_1}{L_2} \leq \frac{l_1}{l_2} \mid H_1\right) \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_1}\left[\frac{1}{n}\log\left(\frac{l_2}{l_1}\right) + K_{12}\right]\right), \tag{50}$$

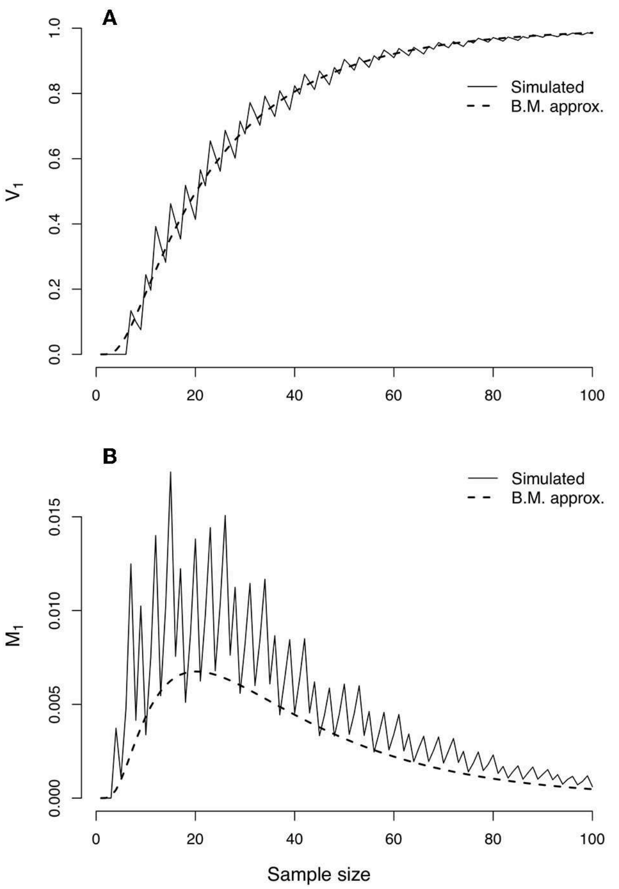


FIGURE 3 | Evidence error probabilities for comparing two Bernoulli(p) distributions, with $p_1 = 0.75$ and $p_2 = 0.50$. **(A)** Simulated values (jagged curve) and values approximated under the Central Limit Theorem of the probability of strong evidence for model H_1, $V_1 = 1 - M_1 - W_1$. **(B)** Simulated values (jagged curve) and approximated values for the probability of misleading evidence M_1. Note that the scale of the bottom graph is one fifth of that of the top graph.

$$P_2 = P\left(\frac{L_1}{L_2} \le \frac{l_2}{l_1} \mid H_2\right) \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_2}\left[\frac{1}{n}\log\left(\frac{l_2}{l_1}\right) + K_{21}\right]\right). \tag{51}$$

These are interpreted as the probabilities of misleading evidence under models 1 and 2, respectively if the value of k were taken to be l_2/l_1. The quantity $1 - P_2$ in this context is the severity as defined by Mayo (1996, 2018) and Mayo and Spanos (2006). Taper and Lele (2011) termed P_1 or P_2 as a local probability of misleading evidence (M_{L} in their notation), as opposed to a global, pre-data probability of misleading evidence (M_{G} in their notation; M_1 and M_2 here) characterizing the long-range reliability of the design of the data-generating process.

2.2. Misspecified Models

George Box's (Box, 1979) oft-quoted aphorism that "all models are wrong, but some are useful" becomes pressing in ecology, a science in which daily work and journal articles are filled with statistical and mathematical representations. Ecologists must assume in abundance that Type 3 errors are prevalent, even routine, in their work. Here we compare Neyman-Pearson hypothesis testing with evidential statistics to try to understand how analyses can go wrong, and how analyses can be made better, in ecological statistics. For a statistical method of choosing between $f_1(x)$ or $f_2(x)$, we now ask how well the method performs toward choosing the model closest to the true model $g(x)$ when both candidate models are misspecified.

2.2.1. Neyman-Pearson Hypothesis Testing Under Misspecification

Statisticians have long cautioned about the prospect that both models f_1 and f_2 in the Neyman-Pearson framework, broadly interpreted to include testing composite models with generalized likelihood ratio and other approaches, could be misspecified, and as a result that the advertised error rates (or by extension the coverage rates for confidence intervals) would become distorted in unknown ways (for instance, Chatfield, 1995). The approximate behavior of the LR under the CLT under misspecification (Equations 20–22) allows us to view directly how the error probabilities α and β can be affected in Neyman-Pearson testing when the models are misspecified.

The critical value c (Equation 28) is chosen as before, under the assumption that the observations are generated from model f_1. We ask the following question: "Suppose the real Type 1 error is defined as picking model f_2 when the model f_1 is actually closest to the true pdf $g(x)$ (that is, when $\Delta K > 0$). What is the probability, let us say α', of this Type 1 error, given that f_1 is the better model?" We now have

$$\begin{aligned}\frac{L_1}{L_2} \le c \Rightarrow \frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - \Delta K\right] &\le \frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log(c) - \Delta K\right] \\ &= \frac{\sqrt{n}}{\sigma_g}(K_{12} - \Delta K) - \frac{\sigma_1}{\sigma_g}z_\alpha\end{aligned} \tag{52}$$

after substituting for c (Equation 28), and so the CLT (Equation 22) tells us that

$$\begin{aligned}\alpha' &= \text{P}\left(\frac{L_1}{L_2} \le c \mid \Delta K > 0\right) \approx \Phi\left(\frac{\sqrt{n}}{\sigma_g}(K_{12} - \Delta K) - \frac{\sigma_1}{\sigma_g}z_\alpha\right) \\ &\neq \Phi(-z_\alpha) = \alpha.\end{aligned} \tag{53}$$

In words, the Type 1 error realized under model misspecification is generally not equal to the specified test size. Note that Equation (53) collapses to Equation (28), as desired, if $f_1 = g$.

Whether the actual Type 1 error probability α' is greater than, equal to, or less than the advertised level α depends on the various quantities arising from the configuration of $f_1(x)$, $f_2(x)$, and $g(x)$ in model space. Because the standard normal cdf $\Phi(\bullet)$ is a monotone increasing function, we have

$$\alpha' > \alpha \Rightarrow \frac{\sqrt{n}}{\sigma_g}(K_{12} - \Delta K) - \frac{\sigma_1}{\sigma_g}z_\alpha > -z_\alpha. \tag{54}$$

The inequality reduces to three cases, depending on whether $\sigma_1 - \sigma_g$ is positive, zero, or negative:

$$\alpha' > \alpha \;\Rightarrow$$

$$\sqrt{n}\frac{(K_{12} - \Delta K)}{\left(\sigma_1 - \sigma_g\right)} > z_\alpha, \text{ if } \sigma_1 - \sigma_g > 0, \tag{55}$$

$$K_{12} - \Delta K > 0, \text{ if } \sigma_1 - \sigma_g = 0, \tag{56}$$

$$\sqrt{n}\frac{(K_{12} - \Delta K)}{\left(\sigma_1 - \sigma_g\right)} < z_\alpha, \text{ if } \sigma_1 - \sigma_g < 0. \tag{57}$$

The ratio $(K_{12} - \Delta K)/\left(\sigma_1 - \sigma_g\right)$ compares the difference between what we assumed about the LR mean (K_{12}) and what is the actual mean (ΔK) with the difference between what we assumed about the LR variability (σ_1) with what is the actual variability (σ_g). The left-hand inequalities for each case are reversed if $\alpha' < \alpha$.

The persuasive strength of Neyman-Pearson testing always revolved around the error rate α being known and small, and the P-value, if used, being an accurate reflection of the probability of more extreme data under H$_1$. When $L_1/L_2 \leq c$ in the Neyman-Pearson framework with correctly specified models, the reasoned observer is forced to abandon H$_1$ as untenable. However, in the presence of misspecification, the real error rate α' is unknown, as is a real P-value for a generalized likelihood ratio test. Furthermore, α' is seen in Equation (53) to be an increasing function of n if $K_{12} > \Delta K$ (remember that for a generalized LR test the Type 1 error is predicated on $\Delta K > 0$), *with 1 as an upper asymptote*. If model f_2 is very different from model f_1 (K_{12} large) but is almost as close to truth as f_1 (ΔK small), then Type 1 errors will be rampant, more so with increasing sample size.

That greater sample size would make error more likely seems counterintuitive, but it can be understood from the CLT results for the average log-LR given by $(1/n)\log\left(L_1/L_2\right)$ (Equations 12, 21). If the observations arise from $f_1(x)$ (correct specification), the average log-LR has a mean of K_{12} and its distribution becomes more and more concentrated around K_{12} as n becomes large. If however the observations arise from $g(x)$ (misspecification), the average log-LR has a mean of ΔK and its distribution becomes more and more concentrated around ΔK as n becomes large. A Neyman-Pearson test based on a statistic that has a null hypothesis mean of K_{12} will become more and more certain to reject the null hypothesis when the true mean is ΔK. Thus, the Neyman-Pearson framework can be a highly unreliable approach for picking the best model in the presence of misspecification.

The error probability β' is defined and approximated in similar fashion. If model f_2 is closer to truth, we have $\Delta K < 0$, and from Equations (28–30) we now have

$$\frac{L_1}{L_2} > c \Rightarrow \frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - \Delta K\right] > \frac{\sqrt{n}}{\sigma_g}\left(K_{12} - \Delta K\right) - \frac{\sigma_1}{\sigma_g}z_\alpha. \tag{58}$$

The CLT then gives

$$\begin{aligned}\beta' &= \mathrm{P}\left(\tfrac{L_1}{L_2} > c \mid \Delta K < 0\right)\\ &\approx 1 - \Phi\left(\tfrac{\sqrt{n}}{\sigma_g}\left(K_{12} - \Delta K\right) - \tfrac{\sigma_1}{\sigma_g}z_\alpha\right)\\ &\neq 1 - \Phi\left(\tfrac{\sqrt{n}}{\sigma_2}\left(K_{12} + K_{21}\right) - \tfrac{\sigma_1}{\sigma_2}z_\alpha\right) = \beta.\end{aligned} \tag{59}$$

As a function of n, β' goes to zero as n becomes large, preserving that desirable property of β from Neyman-Pearson testing under correct specification. However, if the experiment or survey is being planned around the value of β, under misspecification the actual value as defined by β' could be quite different. In particular, if $\beta' > \beta$, we must have

$$\frac{\sqrt{n}}{\sigma_g}\left(K_{12} - \Delta K\right) - \frac{\sigma_1}{\sigma_g}z_\alpha < \frac{\sqrt{n}}{\sigma_2}\left(K_{12} + K_{21}\right) - \frac{\sigma_1}{\sigma_2}z_\alpha. \tag{60}$$

The inequality reduces to three cases depending on whether $\sigma_2 - \sigma_g$ is positive, zero, or negative:

$$\beta' > \beta \;\Rightarrow$$

$$\frac{\sqrt{n}}{\sigma_1}\left[\frac{\sigma_2\left(K_{12} - \Delta K\right) - \sigma_g\left(K_{12} + K_{21}\right)}{\sigma_2 - \sigma_g}\right] < z_\alpha, \text{ if } \sigma_2 - \sigma_g > 0, \tag{61}$$

$$(-\Delta K) - K_{21} < 0, \text{ if } \sigma_2 - \sigma_g = 0, \tag{62}$$

$$\frac{\sqrt{n}}{\sigma_1}\left[\frac{\sigma_2\left(K_{12} - \Delta K\right) - \sigma_g\left(K_{12} + K_{21}\right)}{\sigma_2 - \sigma_g}\right] > z_\alpha, \text{ if } \sigma_2 - \sigma_g < 0. \tag{63}$$

The left inequalities for the three cases are reversed for $\beta' < \beta$. The degree to which β' departs from β is seen to depend on a tangled bank of quantities arising from the configuration of $f_1(x)$, $f_2(x)$, and $g(x)$ in model space.

2.2.2. P-values, Equivalence Testing, and Severity Under Misspecification

The problems with α and β, and with P-values as defined for the generalized LR setting in Equations (50) and (51), under misspecification highlight problems that might arise in significance testing, equivalence testing or severity analysis. With misspecification, the true P-value (P' say) can differ greatly from the P-value (Equation 49) calculated under H$_1$ and thereby could promote misleading conclusions (P' is formed from Equation (49) by substituting σ_g for σ_1 and $-\Delta K$ for K_{12}). Equivalence testing, being retargeted hypothesis testing, will take on all the problems of hypothesis testing under misspecification. Severity is $1 - P_2$ as defined by Equation (51), but with misspecification the true value of P_2 is Equation (51) with σ_g substituted for σ_2 and $-\Delta K$ substituted for K_{21}. With misspecification, the true severity could differ greatly from the severity calculated under H$_2$. One might reject H$_1$ falsely, or one might fail to reject H$_1$ falsely, or one might fail to reject H$_1$ and falsely deem it to be severely tested. Certainly, in equivalence testing and severity analysis, the problem of model misspecification is acknowledged as important (for instance, Mayo and Spanos, 2006; Spanos, 2010) and is addressed with model evaluation techniques, such as residual analysis and goodness of fit testing.

2.2.3. Evidence Under Misspecification

To study the properties of evidence statistics under model misspecification, we redefine the probabilities of weak evidence and misleading evidence in a manner similar to how the error probabilities were handled above in the Neyman-Pearson formulation. We take $W_1{}'$ and $M_1{}'$ to be the probabilities of weak and misleading evidence, given that model f_1 is closer to truth, that is, given that $\Delta\ K > 0$:

$$\begin{aligned} \mathrm{P}\left(\text{weak evidence} \mid \Delta K > 0\right) &= \mathrm{P}\left(1/k < L_1/L_2 < k \mid \Delta K > 0\right) \\ &= W_1{}', \qquad (64) \\ \mathrm{P}\left(\text{misleading evidence} \mid \Delta K > 0\right) &= \mathrm{P}\left(L_1/L_2 \leq 1/k \mid \Delta K > 0\right) \\ &= M_1{}'. \qquad (65) \end{aligned}$$

Similarly, given model f_2 is closer to truth,

$$\mathrm{P}\left(\text{weak evidence} \mid \Delta K < 0\right) = \mathrm{P}\left(1/k < L_1/L_2 < k \mid \Delta K < 0\right) = W_2{}', \qquad (66)$$

$$\mathrm{P}\left(\text{misleading evidence} \mid \Delta K < 0\right) = \mathrm{P}\left(L_1/L_2 \geq k \mid \Delta K < 0\right) = M_2{}'. \qquad (67)$$

The error probabilities $M_1{}'$, $M_2{}'$, $W_1{}'$, and $W_2{}'$ can be approximated with the CLT results for L_1/L_2 (Equations 20–22) under misspecification. For example, to approximate $M_1{}'$ we note that

$$\begin{aligned} &\frac{L_1}{L_2} \leq \frac{1}{k} \Rightarrow \frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(\frac{L_1}{L_2}\right) - \Delta K\right] \leq \frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(\frac{1}{k}\right) - \Delta K\right] \\ &= -\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) + \Delta K\right]. \qquad (68) \end{aligned}$$

We thus obtain

$$M_1{}' \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) + \Delta K\right]\right). \qquad (69)$$

The other error probability under misspecification, with $\Delta K < 0$, likewise becomes

$$M_2{}' \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) + |\Delta K|\right]\right). \qquad (70)$$

The expression is identical to Equation (69) where $\Delta K > 0$ and so we may write

$$M_i{}' \approx \Phi\left(-\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) + |\Delta K|\right]\right), \quad i = 1, 2. \qquad (71)$$

In words, for models with no unknown parameters under misspecification, the error probabilities $M_1{}'$ and $M_2{}'$ are identical. Using different LR cutoff points k_1 and k_2 to control error probabilities M_1 and M_2 under correct specification would break the symmetry of errors under misspecification. The consideration of evidential error probabilities in study design forces the investigator to focus on what types of errors and possible model misspecifications are most important to the study.

The symmetry of error rates is preserved for weak evidence, for which we obtain

$$\begin{aligned} W_i{}' \approx\ &\Phi\left(\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) - |\Delta K|\right]\right) \\ &-\Phi\left(-\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) + |\Delta K|\right]\right), \quad i = 1, 2. \qquad (72) \end{aligned}$$

The formulae for α' (Equation 53), β' (Equation 59), and M_i', W_i', $i = 1, 2$ (Equations 71, 72) allow the investigation of how these error rates change as a function of the sample size n. However, given that these formulae also involve ΔK, K_{12}, and K_{21}, multiple configurations should be explored in model space. **Figure 4** illustrates how changing parameters can change KL divergences. For instance, the generating process and the approximating models could be aligned in space (see **Figure 4A**) or not (**Figure 4B**). Other configurations are explored in **Figures 4C,D**. The error rates for each one of these configurations are shown in **Figure 5**.

Four properties of the error probabilities under misspecification are noteworthy. First, $M_1{}'$, $M_2{}'$, $W_1{}'$, and $W_2{}'$ all asymptotically approach zero as n becomes large provided $\Delta K \neq 0$ (that is, provided one of the models is measurably better than the other), consistent with their behavior under correct specification. Second, for a given value of $|\Delta K|$, that is, for a given difference in the qualities of models H_1 and H_2 in representing truth, $M_1{}'$ is equal to $M_2{}'$, and $W_1{}'$ is equal to $W_2{}'$. Thus, neither model has special standing. Third, $M_1{}'$ and $W_1{}'$ asymptotically approach M_1 and W_1 as model f_1 becomes better at representing truth (i.e., as $K\left(g,f_1\right) \rightarrow 0$), and likewise $M_2{}'$ and $W_2{}'$ approach M_2 and W_2 as f_2 becomes better. Fourth, if $\Delta K = 0$, that is, if both models are equal in quality, then $M_1{}'$ and $M_2{}'$ each approach $1/2$, and $W_1{}'$ and $W_2{}'$ each approach zero, as n becomes large. The above four properties are intuitive and sensible.

The total error probability under misspecification given by $M_i{}' + W_i{}'$ ($i = 1, 2$) is identical for both models and remains a monotone decreasing function of n:

$$M_i{}' + W_i{}' \approx \Phi\left(\frac{\sqrt{n}}{\sigma_g}\left[\frac{1}{n}\log\left(k\right) - |\Delta K|\right]\right). \qquad (73)$$

The probability of strong evidence for model f_i if f_i is closer to g is given by $V_i{}' = 1 - M_i{}' - W_i{}'$ thus remains a monotone increasing function of n with an asymptote of 1. As was the case for correctly specified models, $V_i{}' > M_i{}'$. Also, $M_i{}'$ increases at first as a function of n, rising to a maximum value before decreasing asymptotically to zero. The value $\tilde{n}_i'$ at which $M_i{}'$ is maximized is given by

$$\tilde{n}_i' = \frac{\log\left(k\right)}{|\Delta K|}, \qquad (74)$$

with the corresponding maximum value of $\tilde{M}_i'$ being

$$\tilde{M}_i' = \Phi\left(-\frac{2\sqrt{|\Delta K| \cdot \log\left(k\right)}}{\sigma_g}\right). \qquad (75)$$

The expressions for $\tilde{n}_i'$ and $\tilde{M}_i'$ revert to their counterparts $\tilde{n}_i$ and $\tilde{M}_i$ when one of the models is correctly specified. If both models are of equal quality, that is, $\Delta K = 0$, then the probabilities $M_i{}'$ can be considered as probabilities of evidence

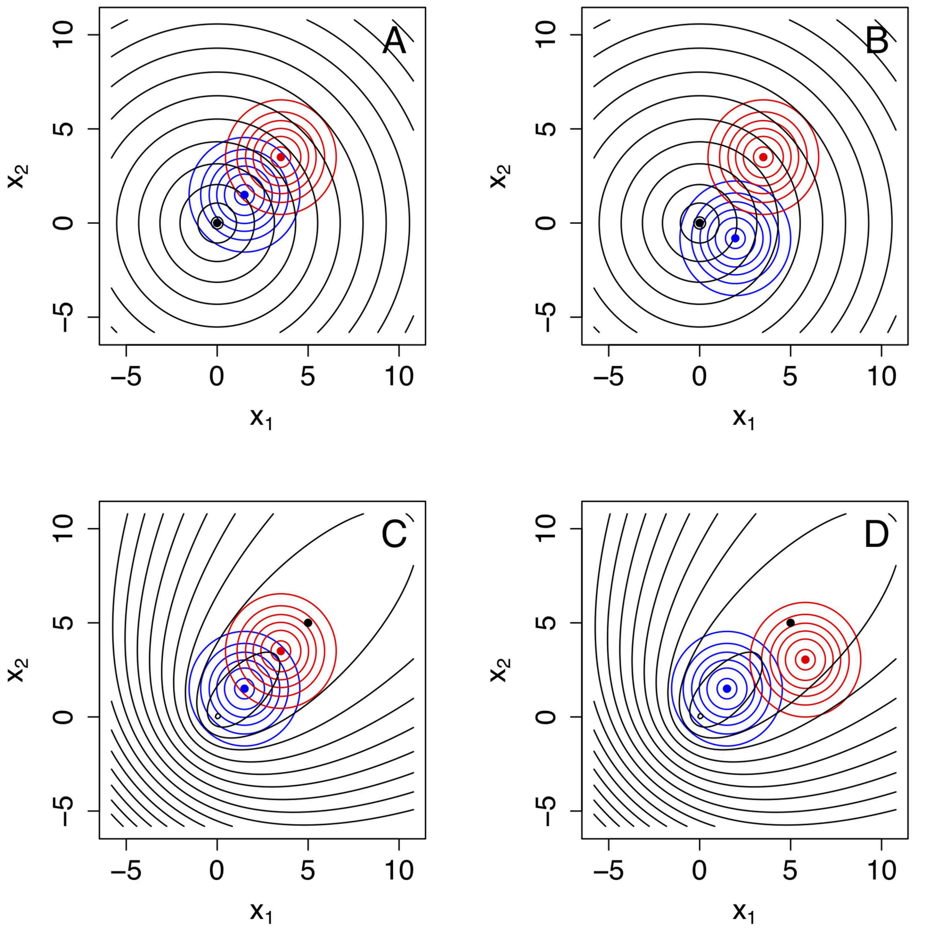


FIGURE 4 | Four model configurations involving a bivariate generating process $g(x_1, x_2)$ (in black), and two approximating models $f_1(x_1, x_2)$ (in blue) and $f_2(x_1, x_2)$ (in red). In all cases the approximating models are bivariate normal distributions whereas the generating process is a bivariate Laplace distribution. These model configurations are useful to explore changes in α' (Equation 53), β' (Equation 59) and $M_i', W_{i'}, i = 1, 2$ (Equations 71, 72) as a function of sample size, as plotted in **Figure 6**. **(A)** $g(x_1, x_2)$ is a bivariate Laplace distribution centered at 0 with high variance. All three models have means aligned along the $1:1$ line and marked with a black, blue, and red filled circle, respectively. Model $f_1(x_1, x_2)$ is closest to the generating process. **(B)** Model $f_1(x_1, x_2)$ is still the model closest to the generating process, at exactly the same distance as in **(A)** but misaligned from the $1:1$ line. **(C)** Here all three models are again aligned, but the generating process $g(x_1, x_2)$ is an asymmetric bivariate Laplace that has a large mode at 0, 0 and smaller mode around the mean, marked with a black dot. In this case, the generating model is closer to model $f_2(x_1, x_2)$ (in red). **(D)** Same as in **(C)**, except model $f_2(x_1, x_2)$ (in blue) is now misaligned, but still the closest model to the generating process.

favoring (wrongly, as the models are a tossup in quality) one or the other models. When $\Delta K = 0$, $M_i{}'$ as a function of n has no local maximum and asymptotically approaches 1/2 as sample size increases. The possibility that $M_i{}'$ might be as great as 1/2 seems distressing, but this only occurs when the two models become equally good (not necessarily identical) approximations of the generating process.

2.2.4. Illustration of Neyman-Pearson Testing and Evidence Under Misspecification

An extension of the Bernoulli example from **Figure 3** serves to sharply contrast the error properties of NP testing and evidence analysis. We construct as before two candidate Bernoulli models with respective success probabilities p_1 and p_2. Suppose however that the data actually arise from a Bernoulli distribution with success probability p_g. From Equation (17), the value of ΔK becomes

$$\begin{aligned}\Delta K &= p_g \log\left(\frac{p_g}{p_2}\right) + (1-p_g)\log\left(\frac{1-p_g}{1-p_2}\right) - p_g \log\left(\frac{p_g}{p_1}\right) \\ &\quad - (1-p_g)\log\left(\frac{1-p_g}{1-p_1}\right) \\ &= \log\left(\frac{1-p_1}{1-p_2}\right) + p_g \log\left[\frac{p_1(1-p_2)}{(1-p_1)p_2}\right] \qquad (76)\end{aligned}$$

Note that ΔK is here a simple linear function of p_g. In the **Figure 3** example, $p_1 = 0.75$ and $p_2 = 0.50$. If we take $p_g = 0.65$, we have a situation in which model 1 is slightly closer to the true model than model 2. As well, we readily calculate that $K_{12} = 0.130812$ and $\Delta K =$

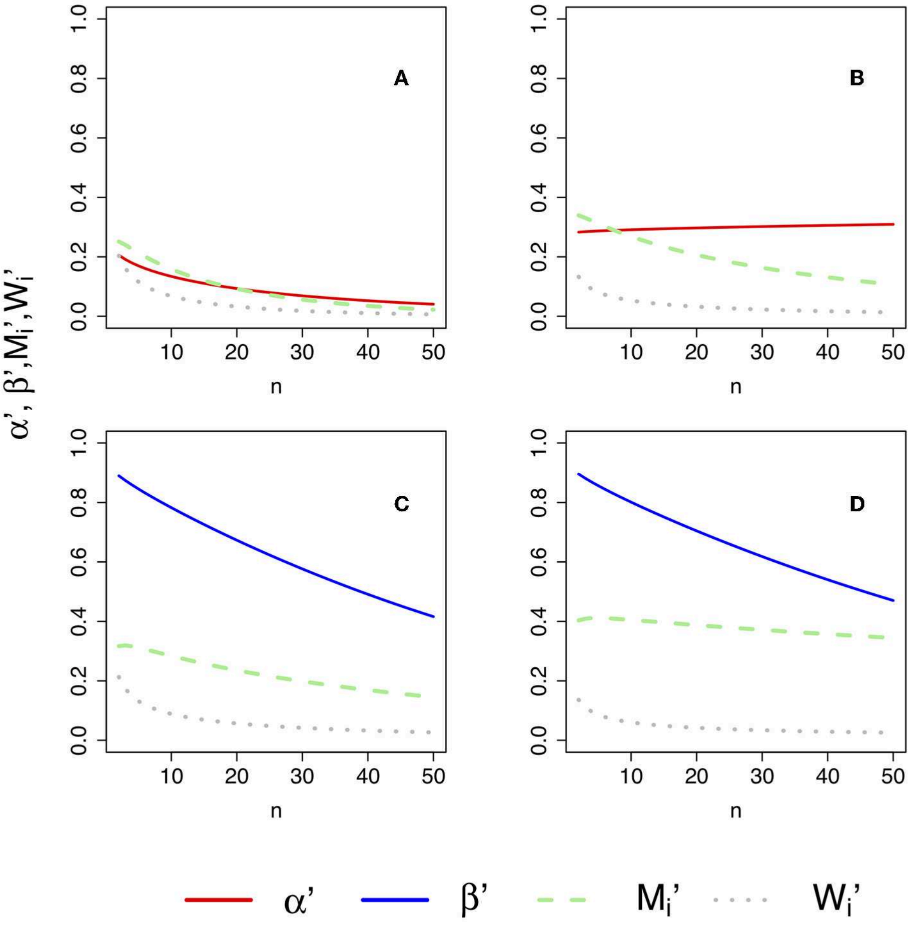


FIGURE 5 | Changes in α' (Equation 53), β' (Equation 59) and $M_i', W_{i'}, i = 1, 2$ (Equations 71, 72) as a function of sample size. The plot in **(A–D)** were computed under each of the geometries plotted in **Figures 4A–D**. **(A)** α', M_1', and W_1' for the models geometry in **Figure 4A**, where all models are aligned and model f_1 is closest to the generating process. **(B)** Same as in **(A)** but model f_1 is misaligned. **C** β', M_2', and W_2' for model geometry in **Figure 4C**, where model f_2 is closer to the generating process and all models are aligned. **D**: β', M_2', and W_2' for model geometry in **Figure 4D**, where model f_2 is closer to the generating process but model f_2 is misaligned.

0.02095081, so that $K_{12} > \Delta K$, a situation in which we expect α' to be an increasing function of n (as dictated by Equation 53).

The top panel of **Figure 6** should give pause to all science. Shown is the probability (α') of wrongly rejecting the null hypothesis of model 1 in favor of the alternative hypothesis of model 2 with Neyman-Pearson testing, under the example scenario of model misspecification in which model 1 is closer to truth. Both simulated values and the CLT approximation (Equation 53) are plotted as a function of sample size. The nominal value of α for setting the critical value (c) was taken to be 0.05. The curves rapidly approach an asymptote of 1 as sample size increases. With NP testing under model misspecification, picking the wrong model can become a near certainty.

In the bottom panel of **Figure 6**, the probability of misleading evidence for model 2 (M_2'), that is, of picking the model farther from truth, increases at first but eventually decreases to zero (**Figure 6**, bottom panel shows simulated values as well as CLT approximation given by Equation 70). Under evidence analysis, the probability of wrongly picking the model farthest from truth converges to 0 as sample size increases.

The example illustrates directly the potential effect of misspecification on the results of the Neyman-Pearson Lemma. The lemma is of course limited in scope, and we should in all fairness note that a classical extension of the lemma to one-sided hypotheses seemingly ameliorates the problem in this particular example. Suppose the two models are expanded: model 1 is the Bernoulli distribution with $p \geq 0.75$, with model 2 becoming the Bernoulli with $p < .75$. Then, the "Karlin-Rubin Theorem" (Karlin and Rubin, 1956) finds the LR test to be uniformly most powerful size α (or less) test of model 1 vs. model 2. Three key ideas enter the proof of the theorem. First, for any particular value p_2 such that $p_2 < p_1$, the Neyman-Pearson Lemma gives the LR test as most powerful. Second, the cutoff point c for the Neyman-Pearson LR test does not depend on the value of p_2. Third, the LR is a monotone function of a sufficient test statistic given by $(x_1+x_2+\ldots+x_n)/n$. The upshot is that α would remain constant in the expanded scenario, and β would decrease toward zero as advertised.

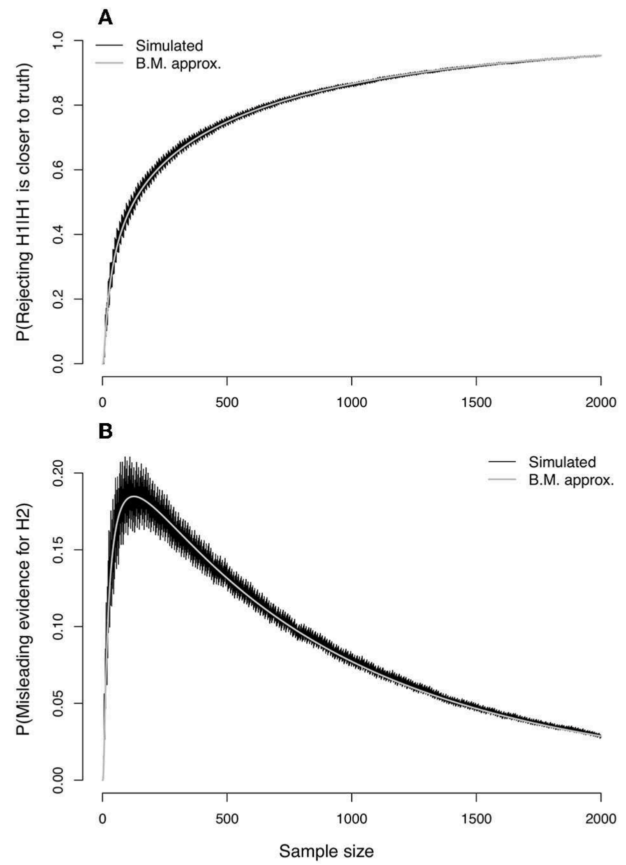


FIGURE 6 | Evidence error probabilities for comparing two Bernoulli(p) distributions, with $p_1 = 0.75$ and $p_2 = 0.50$, when the true data-generating model is Bernoulli with $p = 0.65$. **(A)** Simulated values (jagged curve) and values approximated under the Central Limit Theorem of the probability (α') of rejecting model H_1 when it is closer than H_2 to the true model. **(B)** Simulated values (jagged curve) and approximated values for the probability (M'_1) of misleading evidence for model H_2 when model H_1 is closer to the true data-generating process.

However, the one-sided extension of our Bernoulli example expands the model space to eliminate the model misspecification problem. We regard the hypotheses $H_1: p \geq 0.75$ and $H_2: p < 0.75$ to be a case of two non-overlapping models (**Figures 1**, **2**, bottom) which may or may not be correctly specified. The Karlin-Rubin Theorem would govern if the models are correctly specified. Misspecification in this one-sided context would be exemplified, for instance, by data arising from some other distribution family besides the Bernoulli(p), such as an overdispersed family like a beta-Bernoulli (Johnson et al., 2005). Under misspecification, Karlin-Rubin lacks jurisdiction.

2.2.5. Evidence Functions

Lele (2004) took Royall's (Royall, 1997) approach to using the LR for model comparison and generalized it into the concept of evidence functions. Evidence functions are developed mathematically from a set of desiderata that effective measures of evidence intuitively should satisfy (see Taper and Ponciano, 2016).

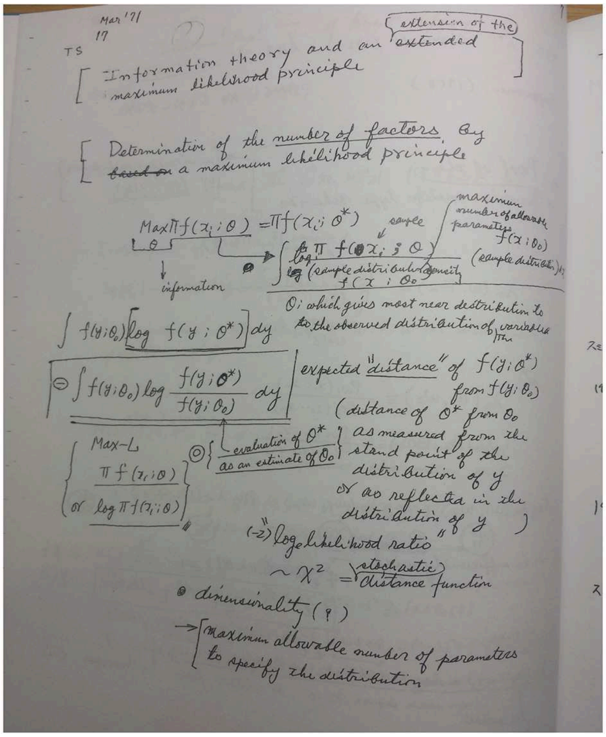
The basic insight is that the log-LR emerges as the function to use for model comparison when the discrepancy between models is measured by the KL divergence (Equation 3). The reason is that $(1/n)\log(L_1/L_2)$ is a natural estimate of ΔK, the *difference of divergences of* $f_1(x)$ *and* $f_2(x)$ *from truth* $g(x)$. However, numerous other measures of divergence or distance between statistical distributions have been proposed (see Lindsay, 2004; Pardo, 2005; Basu et al., 2011), the KL divergence merely being the most well-known. Each measure of divergence or distance would give rise to its own evidence function. Lele (2004) defines an evidence function for a given divergence measure as a data-based estimate of the difference of divergences of two approximating models from the underlying process that generated the data. The motivating idea is to use the data to estimate which of two models is "closer" in some sense to the data generating process. The evidence function concept requires a measure of divergence of a model $f(x)$ from the true data generating process $g(x)$ and a statistic, the evidence function, for estimating the difference of divergences from truth of two models $f_1(x)$ and $f_2(x)$. Important among the desiderata for evidence functions (Taper and Ponciano, 2016) is that the probabilities of strong evidence *as defined under misspecification* should asymptotically approach 1 as sample size increases (and so the error probabilities as embodied in $M_1{}'$, $M_2{}'$, $W_1{}'$, and $W_2{}'$ would approach zero). It is noteworthy that the prospect of model misspecification is baked into the very definition of an evidence function.

Lele (2004) further proved an optimality property of the LR as evidence function similar to the optimality of the LR in the Neyman-Pearson Lemma. Lele's Lemma states that, out of all evidence functions, asymptotically, that is for large sample sizes, the probability of strong evidence is maximized by the LR. The result combines the Neyman-Pearson Lemma of hypothesis tests with Fisher's lower bound for the variance of estimators (see Rice, 2007), extending both. Thus, the information in the data toward quantifying evidence is captured the most by the LR statistic or, equivalently, KL divergence. Other divergence measures, however, have desirable properties, such as robustness against outliers. Modified profile likelihood and conditional likelihood also lead to desirable evidence functions that can account for nuisance parameters, although these modifications to the original LR statistics still are unexplored in terms of their optimality.

3. EVIDENCE FUNCTIONS FOR MODELS WITH UNKNOWN PARAMETERS

3.1. Information-Theoretic Model Selection Criteria

The latter part of the 20th Century saw some statistical developments that made inroads into the problems of models with unknown parameters (composite models), multiple models, model misspecification and non-nested models, among the more widely adapted of which were the model selection indexes based on information criteria. The work of Akaike (Akaike, 1973, 1974,



Mar '71
17

TS

[Information theory and an {extension of the} extended maximum likelihood principle]

[Determination of the number of factors by ~~based on~~ a maximum likelihood principle

$\mathrm{Max}_{\theta} \prod f(x_i;\theta) = \prod f(x_i;\theta^*)$ sample

maximum number of allowable parameters

information

$\int \log \frac{\prod f(x_i;\theta)}{(\text{sample distribution density } f(x;\theta_0))}$ $f(x;\theta_0)$ (sample distribution)

θ; which gives most near distribution to the observed distribution of the variables

$\int f(y;\theta_0)\left[\log f(y;\theta^*)\right]dy$

$\ominus \int f(y;\theta_0) \log \frac{f(y;\theta^*)}{f(y;\theta_0)}\,dy$ expected "distance" of $f(y;\theta^*)$ from $f(y;\theta_0)$

(distance of θ^* from θ_0 as measured from the stand point of the distribution of y or as reflected in the distribution of y)

Max-L $\prod f(x_i;\theta)$ or $\log \prod f(x_i;\theta)$

$\circledcirc$ {evaluation of θ^* as an estimate of θ_0}

(-2) "log likelihood ratio" $\sim \chi^2$ = stochastic distance function

● dimensionality (?)

→ [maximum allowable number of parameters to specify the distribution

FIGURE 7 | Moment of discovery: page from Professor H. Akaike's research notebook, written while he was commuting on the train in March 1971. Photocopy kindly provided by the Institute for Statistical Mathematics, Tachikawa, Japan.

Figure 7) revealed a novel way of formulating the model selection problem and ignited a new statistics research area. Akaike's ideas found immediate use in the time series models of econometrics (Judge et al., 1985), were studied and disseminated for statistics in general by Sakamoto et al. (1986) and Bozdogan (1987) and popularized, especially in biology, by Burnham and Anderson (2002).

The information criteria are model selection indexes, the most widely used of which is the AIC (originally, "an information criterion," Akaike, 1981; now universally "Akaike information criterion"). The AIC is minus two times the maximized log-likelihood for a model, the maximization taken across unknown parameters, with a penalty for the number of unknown parameters added in: $\mathrm{AIC}_i = -2\log\left(\hat{L}_i\right) + 2r_i$, where $\hat{L}_i$ is the maximized likelihood for model H_i, and r_i is the number of unknown parameters in model H_i that were estimated through the maximization of L_i. We are now explicitly considering the prospect of more than two candidate models, although each evidential comparison will be for a pair of models.

Akaike's fundamental intuition was that it would be desirable to select models with the smallest "distance" to the generating process. The distance measure he adopted is the KL divergence. The log-likelihood is an estimate of this distance (up to a constant that is identical for all candidate models). Unfortunately, when parameters are estimated, the maximized log-likelihood as an estimate of the KL divergence is biased low. The AIC is an approximate bias-corrected estimate of an expected value related to the distance to the generating process. The AIC is an index where goodness of fit as represented by maximized log-likelihood is penalized by the number of parameters estimated. Penalizing likelihood for parameters is a natural idea for attempting to balance goodness of fit with usefulness of a model for statistical prediction (which starts to break down when estimating superfluous parameters). To practitioners, AIC is attractive in that one calculates the index for every model under consideration and selects the model with the lowest AIC value, putting all models on a level playing field so to speak.

Akaike's inferential concept underlying the AIC represented a breakthrough in statistical thinking. The idea is that in comparing model H_i with model H_j using an information criterion, both models are assumed to be misspecified to some degree. The actual data generating mechanism cannot be represented exactly by any statistical model or even family of statistical models. Rather, the modeling process seeks to build approximations useful for the purpose at hand, with the left-out details deemed negligible by scientific argument and empirical testing.

Although AIC is used widely, the exact statistical inference presently embodied by AIC is not widely understood by practitioners. What Akaike showed is that under certain conditions $-\mathrm{AIC}_i/(2n)$ is (up to an unknown constant) an approximately unbiased estimator of $\mathrm{E}_g\left\{K\left[g(x), f_i\left(x, \hat{\theta}_i\right)\right]\right\}$, where θ_i is a vector of unknown parameters and $\hat{\theta}_i$ is its ML estimate, the parameter penalty in AIC being the approximate bias correction. The expectation has two variability components, (1) the distribution of $f_i\left(X, \hat{\theta}_i\right)$ given the ML estimate value, and (2) the distribution of the ML estimate, both expectations with respect to truth $g(x)$ (In Akaike's formulation, truth was a model $f(\cdot)$ with some high-dimensional unknown parameter, while all the candidate models are also in the same form $f(\cdot)$ except with the parameter vector constrained to a lower-dimensional subset of parameter space. Truth in Akaike's approach is as unattainable as $g(x)$). The double expectation is termed the "mean expected log-likelihood." The difference $\mathrm{AIC}_i - \mathrm{AIC}_j$ then is a *point estimate* of which model is closer on average to truth, in the sense estimating $(-2n)$ times the difference of mean expected log-likelihoods. The approximate bias correction incorporated in AIC is technically correct only if $f_i\left(x, \hat{\theta}_i\right)$ is rather "close" to $g(x)$; Takeuchi (1976) subsequently provided a mathematically improved (but statistically more difficult to estimate) approximation. "Information theoretic" indexes for model selection have proliferated since, with different indexes refined to perform well for different sub-purposes (Claeskens and Hjort, 2008).

In practice, the AIC-type inference represents a relative comparison of two models, not necessarily nested or even in the same model family, requiring only the same data and the same response variable to implement. The inference is post-data, in that there are (as yet) no appeals to hypothetical repeated sampling and error rates. All candidate models, or rather, all pairs of models, can be inspected simultaneously simply by obtaining the AIC value for each model. But, as is the case with all point estimates, without some knowledge of sampling variability and error rates we lack assurance that the comparisons are informative.

3.2. Differences of Model Selection Indexes as Evidence Functions

We propose that information-based model selection indexes can be considered as generalizations of LR evidence to models with unknown parameters, for model families obeying the usual regularity conditions for ML estimation. The evidence function concept clarifies and makes accessible the nature of the statistical inference involved in model selection. Like LR evidence, one would use information indexes to select from a pair of models, say $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$, where θ_1 and θ_2 are vectors of unknown parameters. Like LR evidence, the selection is a post-data inference. Like LR evidence, the prospect of model misspecification is an important component of the inference. And critically, like LR evidence, the error probabilities W_i and M_i $(i = 1, 2)$ can be defined for the information indexes and can in principal be calculated (or simulated) as discussed below. Additionally, as discussed below, many of the existing information indexes retain the desirable error properties of evidence functions. Oddly, the AIC itself does not.

3.3. Nested Models, Correctly Specified

As noted earlier, the generalized LR framework of two nested models under correct model specification is a workhorse of scientific practice and a prominent part of applied statistics texts. It is worthwhile then in studying evidence functions to start with the generalized LR framework, in that the model selection indexes are intended in part to replace the hierarchical sequences of generalized LR hypothesis testing (stepwise regression, multiple comparisons, etc.) for finding the best submodel within a large model family.

The model relationships diagrammed in the top portion of **Figure 1** depict the two cases. In case 1 (top left), a parameter vector in model f_1 identifies the true model giving rise to the data. Technically the parameter vector is contained in model f_2 as well, but the scientific interest focuses on whether the additional parameters in the unconstrained parameter space of f_2 can be usefully ignored. Case 2 (top right) portrays the situation in which the true parameter vector is in the unconstrained parameter space of model f_2; model f_1 is too simple to be useful.

Suppose we decide to use $\Delta\mathrm{AIC}_{12} = \mathrm{AIC}_1 - \mathrm{AIC}_2$ as an evidence function. For convenience, we have defined this AIC-based evidence function to vary in the same direction as G^2 (Equation 31) in NP hypothesis testing, so that large values of $\Delta\mathrm{AIC}$ correspond to large evidence for f_2 (opposite to the direction for the ordinary LR-evidence function given by Equation 33). For instance, the early rule of thumb in the AIC literature was to favor model f_1 when $\Delta\mathrm{AIC}_{12} \leq -2$ and to favor model f_2 when $\Delta\mathrm{AIC}_{12} \geq 2$. Note that

$$\Delta\mathrm{AIC}_{12} = G^2 - 2\nu, \tag{77}$$

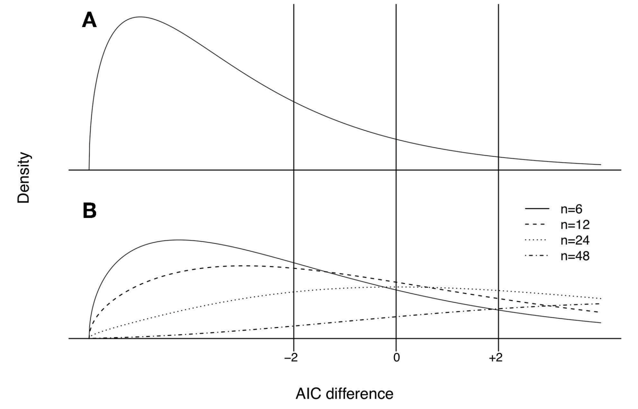


FIGURE 8 | (A) Location-shifted chisquare distribution of the difference of AIC values, when data arise from model 1 nested within model 2. In this plot, the degrees of freedom for this distribution are equal to $\nu = 3$, and the shift to the left of 0 is equal $2\nu = 6$ (see Equation 77 and text below it). This chisquare distribution is invariant to sample size. As a result, the areas under this distribution in the intervals $(-2, +2)$ and $(+2, \infty)$ corresponding to W_1 and M_1, respectively, are invariant to sample size. **(B)** Non-central chisquare distribution of the difference of AIC values, when data arise from model 2 (but not model 1), plotted for different sample sizes. This distribution is also location-shifted but its non-centrality parameter λ, which determines both its mean and variance, is proportional to sample size. In this illustration, $\lambda = n(1/4)$. As a result, the areas under the intervals $(-2\nu, -2)$ and $(-2, +2)$ corresponding to the error probabilities M_2 and W_2 decrease as the sample size increases.

where $\nu = r_2 - r_1$, the difference of the numbers of unknown parameters in the two models. The behavior of our candidate evidence function $\Delta\mathrm{AIC}_{12}$ can be studied using the Wilks/Wald results for the asymptotic distribution of G^2. Under case 1, $\Delta\mathrm{AIC}_{12}$ has (approximately) a chisquare(ν) distribution that has been location-shifted to begin at -2ν instead of at 0 (top of **Figure 8**). Under case 2, $\Delta\mathrm{AIC}_{12}$ has (approximately) a non-central chisquare(ν, λ) distribution with the same -2ν location shift (bottom of **Figure 8**). The areas under the shifted chisquare pdf in the intervals $(-2, +2)$ and $(+2, \infty)$ are respectively the generalized error probabilities W_1 and M_1 (**Figure 8**, top). Likewise, the areas under the shifted non-central pdf in the intervals $(-2\nu, -2)$ and $(-2, +2)$ are respectively the generalized error probabilities M_2 and W_2 (**Figure 8**, bottom).

As sample size increases, the error probabilities W_1 and M_1 for the AIC-based evidence function do not go to zero but rather remain positive (**Figure 8**, top). The value of n appears nowhere in the location-shifted chisquare pdf for $\Delta\mathrm{AIC}_{12}$, and so the error probabilities W_1 and M_1 remain static. Thus, for the AIC, the probabilities of weak and misleading evidence given model f_1 generates the data both behave like the Type 1 error probability α in Neyman-Pearson testing. The simulation results of Aho et al. (2014) showing a Type-1-like behavior of the AIC with increasing sample size for particular statistical models are thereby explained (see also Taper and Ponciano, 2016).

As sample size increases, the error probabilities W_2 and M_2 for the AIC-based evidence function do go to zero (**Figure 8**, bottom). The non-centrality parameter λ in the location-shifted non-central chisquare pdf for $\Delta\mathrm{AIC}_{12}$ is proportional to the value of n, and the mean $(\nu + \lambda)$ of the non-central distribution increases faster than the standard deviation $([2(\nu + 2\lambda)]^{1/2})$, driving the error probabilities W_2 and M_2 to zero. Thus, for the AIC, the probabilities of weak and misleading evidence given model f_2 generates the data both behave like the Type 2 error probability β in Neyman-Pearson testing.

Thus, within the generalized likelihood ratio framework, the AIC appears to bring no particular improvement in the sense of evidence to ordinary Neyman-Pearson testing using G^2. Indeed, at least in the Neyman-Pearson approach, the value of α is fixed by the investigator and is therefore *known* if the models are correctly specified. The error probabilities attending the use of AIC however are unknown, as they generally are in evidence functions, although they in principle can be estimated with simulation. AIC-based model selection does not have the error properties of an evidence function within the classical milieu of nested statistical models.

Other information-theoretic indexes used for model selection, however, do have performance characteristics of evidence functions. Consider the Schwarz information criterion (SIC; also known as Bayesian information criterion or BIC) given by

$$\text{SIC}_i = -2\log\left(\hat{L}_i\right) + r_i\log(n).$$

The index originally had a Bayesian-based derivation (Schwarz, 1978), but its frequentist error properties when employed as an evidence function become apparent with the methods used above for the AIC. As with the AIC, the evidence function version of the SIC would use the difference of SIC values:

$$\Delta\text{SIC}_{12} = \text{SIC}_1 - \text{SIC}_2 = G^2 - \nu\log(n).$$

As with the AIC also, the asymptotic distributions of the SIC evidence function under model f_1 and model f_2 are respectively, location-shifted chisquare and non-central chisquare distributions. For the SIC though, the location of the lower bound of the two distributions at $-\nu\log(n)$ decreases as sample size increases (**Figure 9**, top). If the data arise from model f_1, the chisquare distribution is pulled to the left, and the areas under the pdf corresponding to and eventually decrease asymptotically to zero. If the data arise from model f_2, although the non-central chisquare distribution is also pulled to the left at a rate proportional to $\log(n)$, the mean is pulled to the right at a rate proportional to n, and the coefficient of variation around the mean goes to zero at a rate $1/\sqrt{n}$. The areas under the pdf corresponding to W_2 and M_2 eventually decrease asymptotically to zero (**Figure 9**, bottom). Thus, unlike the AIC, for nested, correctly specified models the SIC possesses a key quality of an evidence function: all the probabilities of weak and misleading evidence eventually decrease asymptotically to zero.

3.4. Misspecified Models

To be fair, AIC as well as evidence functions were forged in the fiery world of misspecified models. Does the AIC difference gain the properties of an evidence function when neither f_1 nor f_2 give rise to the data?

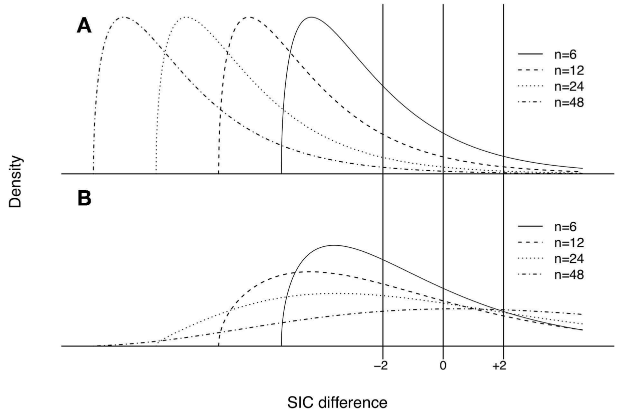


FIGURE 9 | (A) Chisquare distribution of the difference of SIC values, when data arise from model 1 nested within model 2. The chisquare distribution is shifted left as sample size increases. **(B)** Non-central chisquare distribution of the difference of SIC values, when data arise from model 2 (but not model 1), plotted for increasing sample sizes.

If the models are nested or overlapping, the answer is no. To understand this, we must appeal to modern statistical advances in the theory of maximum likelihood estimation and generalized likelihood ratio testing when models are misspecified. The relevant and general theory can be found in White (1982), Nishii (1988), Vuong (1989), and references therein.

Suppose a model with pdf $f(x, \theta)$ is fitted using ML estimation to observations that came from a distribution with pdf $g(x)$. Under a variety of regularity conditions on the pdfs, the ML estimate has an asymptotic multivariate normal distribution centered on a value θ^*, where θ^* is the value of θ that minimizes $K\left(g(x), f(x, \theta)\right)$ (White, 1982). The multivariate normal distribution furthermore concentrates around θ^* as n becomes large, reflecting the fact that the ML estimate under misspecification is a statistically consistent estimate of (converges in probability to) θ^*.

Now, any two models $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$ being compared will be in one of nested, overlapping, or non-overlapping configurations (see **Figure 2**). Under misspecification in each case, the truth $g(x)$ is out there, somewhere. We now ask of an evidence function: "Which model contains a parameter set that brings it closer to truth? Is $K\left(g(x), f_1(x, \theta_1{}^*)\right)$ smaller than $K\left(g(x), f_2(x, \theta_2{}^*)\right)$ or vice versa?"

The question needs modification in the nested and overlapping cases. If f_1 is nested within f_2, $K\left(g(x), f_1(x, \theta_1{}^*)\right)$ cannot be smaller than $K\left(g(x), f_2(x, \theta_2{}^*)\right)$. The modified question becomes "Is $f_1(x, \theta_1{}^*)$ as close to truth as $f_2(x, \theta_2{}^*)$?" The question in the nested case is a natural extension of the question asked under correct specification. In the nested case, $K\left(g(x), f_1(x, \theta_1{}^*)\right)$ being the same as $K\left(g(x), f_2(x, \theta_2{}^*)\right)$ signifies that $f_1(x, \theta_1{}^*)$ and $f_2(x, \theta_2{}^*)$ are the same model. If f_1 overlaps f_2, the model closest to truth could be in the overlapping region, $K\left(g(x), f_1(x, \theta_1{}^*)\right)$ would be the same as $K\left(g(x), f_2(x, \theta_2{}^*)\right)$, and $f_1(x, \theta_1{}^*)$ and $f_2(x, \theta_2{}^*)$ would be the same model. However, in the overlapping case,

$K\left(g\left(x\right), f_1\left(x, \theta_1{}^*\right)\right)$ being the same as $K\left(g\left(x\right), f_2\left(x, \theta_2{}^*\right)\right)$ does not necessarily signify that $f_1\left(x, \theta_1{}^*\right)$ and $f_2\left(x, \theta_2{}^*\right)$ are the same model. The question in the overlapping case becomes "Is the best model in the overlapping region?"

Vuong (1989) derived the asymptotic distributions of G^2 under the nested, overlapping, and non-overlapping cases in the presence of misspecification. His main results relevant here are the following, presented in our notation:

A. When $f_1\left(x, \theta_1{}^*\right)$ and $f_2\left(x, \theta_2{}^*\right)$ are the same model (either f_1 is nested within f_2, or f_1 overlaps f_2, and the best model is in the nested or overlapping region), then the asymptotic distribution of G^2 is a "weighted sum of chisquares" in the form $\sum a_j Z_j^2$, in which the Z_j are independent, standard normal random variables (each Z_j^2 being chisquare with 1 df) and the a_j values are eigenvalues of a square matrix ($r_1 \times r_2$ rows) of expected values of various derivatives of the two log-pdfs with respect to the parameters (generalization of the Fisher information matrix). The point is, the asymptotic distribution of G^2 does not depend on n. ΔAIC_{12} and ΔSIC_{12}, along with evidence functions formed from other information indexes, then have location-shifted versions of the weighted sum of chisquares distribution. The error probabilities $M_1{}'$ and $W_1{}'$ defined for AIC become static and do not decrease to zero as n becomes large. The error probabilities $M_1{}'$ and $W_1{}'$ defined for SIC do decrease to zero, because the location quantity decreases as becomes large, pulling the weighted sum of chisquares pdf to the left (similar to the chisquare distribution in **Figure 9**). This scenario is simulated and then plotted in **Figure 10A**.

B. Suppose the models are nested, overlapping, or non-overlapping, but a non-overlapping part of f_1 or f_2 is closer to truth, that is, when $f_1\left(x, \theta_1{}^*\right)$ and $f_2\left(x, \theta_2{}^*\right)$ are not the same model as in **Figure 2**. Then G^2 has an asymptotic normal distribution with mean $2n\Delta K^*$ and variance $4n\sigma_g{}^2*$, where

$$\Delta K^* = K\left(g\left(x\right), f_2\left(x, \theta_2{}^*\right)\right) - K\left(g\left(x\right), f_1\left(x, \theta_1{}^*\right)\right), \quad (78)$$

and

$$\sigma_g{}^2* = \text{V}_g\left\{\log\left[\frac{f_1\left(X, \theta_1{}^*\right)}{f_2\left(X, \theta_2{}^*\right)}\right]\right\}. \quad (79)$$

The result parallels the CLT results (Equations 20–22) for completely specified models, with the added condition that each candidate model is evaluated at its "best" set of parameters. In this situation, the mean of G^2 increases or decreases in proportion to n, while the standard deviation increases only in proportion to $\sqrt{n}$. All of the error probabilities, $M_1{}'$, $M_2{}'$, $W_1{}'$ and $W_2{}'$ defined for ΔAIC_{12} as well as for ΔSIC_{12} do decrease to zero as n becomes large. This scenario is simulated and plotted in **Figure 10B**.

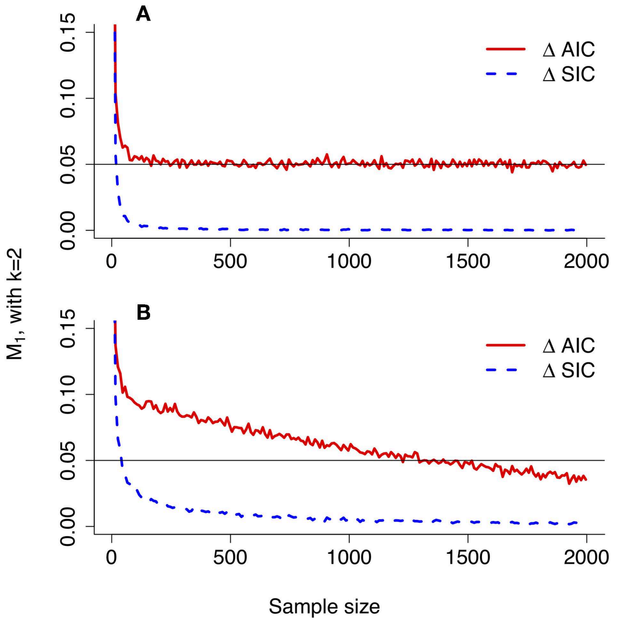


FIGURE 10 | Simulation of Vuong (1989) results for misspecified models. **(A)** When $f_1\left(x, \theta_1{}^*\right)$ and $f_2\left(x, \theta_2{}^*\right)$ are the same model (either f_1 is nested within f_2, or f_1 overlaps f_2, and the best model is in the nested or overlapping region), then the asymptotic distribution of G^2 is a "weighted sum of chisquares" that does not depend on n. The error probabilities M_1 and W_1 do not decrease to 0 for ΔAIC_{12} but do decrease for ΔSIC_{12}. **(B)** When the models are nested, overlapping, or non-overlapping, but a non-overlapping part of f_1 or f_2 is closer to truth, then G^2 has an asymptotic normal distribution with mean and variance that depend on the sample size, and the error probabilities M_1 and W_1 decrease to 0 for both ΔAIC_{12} and ΔSIC_{12}. Details of these two settings in **(A,B)** are found in a fully commented R code.

We must point out that a generalized Neyman-Pearson test (via simulation/bootstrap) of two non-overlapping models with misspecification can suffer the same fate as the completely specified models in the Neyman-Pearson Lemma. The large sample distribution of G^2, assuming model 1 generates the data, would have a mean involving K_{12} (evaluated at true parameter value in model 1 and best parameter value in model 2); the cutoff point c and other test characteristics would be obtained from this distribution. Under misspecification, the true asymptotic distribution of G^2 has a mean involving ΔK^* (Equation 78). As was the case for the two models in the Neyman-Pearson Lemma (**Figure 6**), discrepancy between K_{12} and ΔK^* can cause the generalized Neyman-Pearson test to pick the wrong model with Type 1 error probability approaching 1. The Karlin-Rubin Theorem and the forceful language of uniformly most powerful tests does not rescue Neyman-Pearson testing from derailment when inadequate models are deployed.

Error probabilities going to zero can alternatively be derived as a consequence of the (weak or strong) "consistency" of the model selection index. Consistency here means that the index asymptotically picks the model closest to truth as sample size becomes large. Nishii (1988) studied information indexes in the form $-2\log\left(\hat{L}_i\right) - c_n r_i$, where the parameter penalty coefficient c_n is a possible function of n. The parameter penalty determines whether an information-theoretic index behaves like an evidence function. If c_n grows at a rate $< n$ but $> \log\left(\log\left(n\right)\right)$ then an information-theoretic index will asymptotically pick the model closest to truth Nishii (1988).

The difference of such indexes will therefore behave as an evidence function, as the probabilities of picking any of the contending models go to zero. If, however, the penalty term is constant or asymptotically constant, and the model closest to truth is in a parameter region common to two or more models, then the probabilities of weak and misleading evidence are or become constant. The problematic error properties of Neyman-Pearson testing from the standpoint of evidence are thereby preserved in such model selection indexes. For instance, the AIC-corrected index is (Hurvich and Tsai, 1989).

$$\text{AICc}_i = \text{AIC}_i + 2r_i\,(r_i + 1)\,/\,(n - r_i - 1,)$$

in which the correction term is designed to improve the behavior of the index under small sample sizes. However, the correction term asymptotically approaches zero as n becomes large, and so AICc reverts to AIC, with all its asymptotic error properties, for large samples.

Thus, for either correctly specified or misspecified models in which the best model is in a region of model space that does not overlap any other model under consideration, ΔAIC_{12} indeed behaves like an evidence function. However, many model selection problems, such as in multiple regression, involve collections of models in which model pairs can be nested or overlapping as well as non-overlapping. ΔAIC_{12} will behave more like Neyman-Pearson hypothesis testing for models within overlapping regions and therefore will not possess evidence function properties. Differences of information indexes that adjust G^2 with a constant or asymptotically constant location shift, such as the TIC and AICc will share the Neyman-Pearson properties of ΔAIC_{12} and cannot be regarded as evidence functions. Differences of those information indexes, such as SIC that produce a location shift that decreases to $-\infty$ as n increases (provided that rate is within the Nishii (1988) bounds) will have the error properties of evidence functions.

4. DISCUSSION

4.1. Comparing Approaches to Statistical Inference

We have shown that key inferential characteristics for Fisher significance analysis, Neyman-Pearson hypothesis testing, and evidential comparison differ substantially. Evidence has inferential qualities that match or surpass Fisher significance and Neyman-Pearson tests (see **Table 1**):

TABLE 1 | A comparison of inferential characteristics between Fisherian significance testing (*P*-values *sensu stricto*), Neyman-Pearson hypothesis tests (including *P*-values for likelihood ratios) and evidential statistics.

Inferential characteristic	*P*-value	NP-test	Evidence
Equal status for null and alternatives	NA	No	Yes
Allows evidence for Null	No	No	Yes
Accommodates multiple models	No	Awkward	Yes
All error rates go to zero as sample size increases	No	No	Yes
Total error rate always decreases with increasing sample size	No	No	Yes
Can be used with non-nested models	NA	Not Standard	Yes
Evidence and error rates distinguished	No	No	Yes
Robust to model misspecification	Yes	No	Yes
Promotes exploration of new models	Yes	No	Yes

- *Equal status for both models.* In Fisher significance analysis, there is only one model under consideration. Neyman-Pearson testing compares two models but one of them is accorded special status as the null model and endowed with a fixed error rate (α). Evidence analysis compares two models without giving either model special status.
- *Evidence for the null.* Neither Fisher significance analysis nor the conventional form of Neyman-Pearson testing provides evidence for the null hypothesis. Extra analyses (equivalence testing, severity) have been proposed to quantify evidence for the null hypothesis, but such approaches reverse model roles and give special status to the alternative hypothesis. In evidence analysis, one statistic called an evidence function quantifies the evidence for one model and against each of the models in the model set.
- *Accommodates multiple models.* Under Fisher significance analysis, the *P*-values for different models are based on different sufficient statistics and are not strictly comparable. One could compare multiple *P*-values using a shared goodness of fit statistic (not necessarily sufficient), such as the Kolmogorov-Smirnoff. However, pure goodness of fit favors overparameterization (overfitting). Neyman-Pearson testing has been jury-rigged in various forms (stepwise regression, multiple comparisons) to sort through multiple models, but the results at best have only had fair statistical properties. With evidence analysis, all pairs of candidate models can be compared, and thereby all candidate models can be ranked.
- *All error rates go to zero.* Neyman-Pearson testing fixes the Type 1 error probability to be constant, thereby structuring the error rate to be constant regardless of sample size. Fisher significance analysis acquires such a constant error rate when the decision to reject a model is based on a threshold for the *P*-value. Under evidence analysis all error rates approach zero asymptotically with increasing sample size.
- *Total error monotonically decreasing.* In evidence analysis, the total error under each model (1 minus the probability of strong evidence under the model) decreases monotonically and asymptotically to zero with increasing sample size. Because of the special status of the null hypothesis in Neyman-Pearson testing, the total error rate is the Type 1 error rate which remains constant. Fisher significance analysis dons the Type 1 error properties of Neyman-Pearson testing if the decision to reject the model is based on a *P*-value threshold.
- *Non-nested models.* Fisher significance analysis deals with one model at a time, so the idea of comparing two non-nested models is not applicable. The standard extensions (such as

generalized likelihood ratio) of the original Neyman-Pearson framework to models with unknown parameters assume that one of the models is nested within the other. Evidence analysis compares two models regardless of their nested or non-nested configuration.

- *Evidence and errors rates distinguished.* The interpretation of a *P*-value has long been a source of confusion among scientists. Because the *P*-value is calculated under the properties of just one model, it is not satisfactory as a measure of evidence for one model over another (Royall, 1986, 1997). Evidence analysis regards error rates and evidence as separate concepts. The evidence approach clarifies *P*-values as error rates defined post-data (see section 2.1.7).
- *Robustness to model misspecification.* Evidence functions are defined in terms of the misspecification of two candidate models. Evidence functions are statistical estimates of which of two models is closer to the true data-generating process. The error rates of evidence analysis, defined robustly as the probabilities of wrong conclusions about which model is closer, go to zero as sample size increases, even under model misspecification. Under model misspecification, Neyman-Pearson testing can fail spectacularly: the Type 1 error rate, defined as the probability of wrongly picking the alternative hypothesis model when the null hypothesis model is just as close to truth, can approach 1 asymptotically as sample size increases. Fisher significance analysis, being in essence a test of whether a given model is misspecified, can be considered to be defined under a presumption of misspecification.
- *Promotes exploration of new models.* Perhaps the most important property of evidence analysis in scientific endeavors is that it explicitly encourages discovery of new models that are closer to truth than models already analyzed. An evidence analysis leaves "room at the top," or the possibility that a new approach could yield a much better model for the data. In the scientific world, the daily *t*-tests and regressions under Neyman-Pearson testing produces an inertia, a perfunctory routine in statistical analysis often characterized by working scientists as "cookbook" in nature. Barnard's (1949) observation had Bayesian statistics as its target, but his excruciating words apply to any kind of modeling: "To speak of the probability of a hypothesis implies the possibility of an exhaustive enumeration of all possible hypotheses, which implies a degree of rigidity foreign to the true scientific spirit. We should always admit the possibility that our experimental results may be best accounted for by a hypothesis which never entered our own heads."

4.2. Prediction-Efficient vs. Consistent Criteria

4.2.1. Prediction-Efficiency

AIC and its asymptotic relatives like AICc are built around statistical prediction. The difference of mean expected log-likelihoods is different from what we have defined above as ΔK^*. The mean expected log-likelihood has a second, predictive layer of expectation in its definition, the idea being to identify the model that could best predict a new observation from $g(x)$, taking into account the uncertainty in the estimation of unknown parameters. For this reason these criteria have been termed the efficient, asymptotically efficient, or prediction-efficient criteria (Shibata, 1980; Hurvich and Tsai, 1990).

The tendency for AIC related criteria to over fit is a natural consequence of their design goal of prediction mean square error (MSE) minimization. When parameters are estimated, the increase in prediction MSE due to adding a spurious covariate is generally less than the reduction in prediction MSE caused by including a relevant covariate.

The tendency of stepwise regression to overfit using Neyman-Pearson testing has long been noted (Wilkinson and Dallal, 1981; Hurvich and Tsai, 1990; Harrell, 2001; Rao et al., 2001; Blanchet et al., 2008; Mundry and Nunn, 2008). The fixed Type 1 error rate as a criterion for entry (or exit) of a variable is at the heart of the overfitting problem, and methods for altering the Type 1 error rate based on the number of model parameters have been proposed (e.g., Foster and George, 1994). Such interventions without sample size in the recipe do not produce error rates that universally converge to zero as sample size becomes large.

Model selection with AIC or AICc improves somewhat on the Neyman-Pearson overfitting problem in that the misleading error probabilities both go to zero as sample size increases when two non-overlapping models are being compared. However, overlapping models, in which AIC and AICc are prone to overfit, are typically a substantial subset of the models in contention in multiple regression. The AIC and AICc indexes will tend to include spurious variables too often and thus represent only a partial improvement over stepwise regression.

4.2.2. Identifying Causal Structure

Scientific prediction, however, can be broader than pure statistical prediction. The scientist often desires to predict the outcome of a system manipulation: what will happen if harvest rate is increased, or if habitat extent is halved? Modeling such manipulation might translate as a structural change in a statistical model of the system. The predictive quality of the model then lies more in getting mechanisms in the model as right as possible.

The consistent criteria will asymptotically select the generating process if it is in the model set. If the generating process is not in the model set, the consistent criteria will asymptotically select the model in the set that under best possible parameterization is closest (in the KL sense) to the generating process. The estimation of ΔK^* by the difference of SIC values represents a quest for a different kind of prediction that might come from a structural understanding of the major forces influencing the system under study. The tendency of the prediction efficient criteria to include spurious covariates promotes a mis-understanding of the generating mechanism (Taper, 2004).

Certainly, the finite-sample properties of SIC and other consistent indexes require substantial further study, but the property that more data should be able to distinguish among candidate models with fewer errors seems an important property to preserve.

The scientific allure of information-theoretic indexes resided in the idea that all models were evaluated on a level playing field. One would calculate the index for each model and select the model with the best index, a procedure which promised considerably more clarity over hierarchical sequences of Neyman-Pearson tests, such as stepwise regression.

4.3. Uncertainty in Evidence

AIC and its descendants were originally built around concepts of statistical point estimation. The statistical inference represented by AIC is that of an approximately unbiased point estimate of the mean expected log-likelihood. The statistical concepts of errors and variability in information indexes have by contrast not often been emphasized. Partly as a result, model selection with information indexes has been somewhat of a black box for investigators, as achieving a good understanding of the inferences represented by model selection analyses is a mathematical challenge (see Taper and Ponciano, 2016).

4.3.1. Evaluating Model Adequacy

We have illustrated that, unlike the error rates in Neyman-Pearson hypothesis testing, all of the error rates of evidence analysis converge to zero as sample size increases. However, the errors we have discussed deal only with the determination which of two models is closer to truth; the error rates do not shed light on whether either model is close enough to truth to be scientifically or managerially valuable. This question is the realm of model adequacy analysis.

Whether the statistical inference is a hypothesis test, equivalence analysis, severity analysis, or evidence analysis, whether for a pair of models or multiple pairs of models, a follow-up evaluation of model adequacy looms ever more important as a crucial step (Mayo and Spanos, 2004; Spanos, 2010). Lindsay (2004) and Markatou and Sofikitou (in review) discuss ideas about the statistical evaluation of model adequacy. Mac Nally et al. (2018) give an impassioned editorial plea for routine model adequacy evaluation in scientific model selection. Ponciano and Taper (2019) show how to directly incorporate model adequacy evaluation into information criterion based model selection.

Considering the likely prevalence of model misspecification in ecological statistics, analysts will need to consider how a candidate model could be misspecified as well as the effects of such misspecification on the intended uses of the model. Practically, the analyst can introduce models formulated in diverse fashions and let the model identification process itself reduce model misspecification. Further experimental or observational tests of model predictions (e.g., Costantino et al., 2005) and their associated error rates are necessary to map the conditions under which a given model is reliable.

The error properties of evidence analysis are more difficult to calculate than classical NP tests because model misspecification is involved. But once calculated, the rates are likely to be more accurate than classical tests that pretend misspecification does not exist.

4.3.2. Approaches to Estimating Post-data Error Rates

Error rates are different pre and post-data. W, M and α are pre-data error rates calculated under a model that is assumed to be true. The P-value is a post-data error rate. The pre-data error rates are useful for experimental design, but should be viewed with suspicion as a post-data inference tool because as we have shown these error rates are only accurate if the generating process is the assumed model. Little work has been performed on evidential error rates under the realistic assumption of model misspecification (but see Royall and Tsou, 2003). This area is an important field for future work.

Non-parametric bootstrapping shows great promise for calculating evidential error rates, for data structures that allow bootstrapping. In work in preparation, we (Taper, Lele, Ponciano, and Dennis) show that bootstrapping greatly aids in the interpretation of evidential results. **Figures 4**, **5** indicate that evidential error rates depend on the structure of the model space. Taper and Ponciano (2016) and Ponciano and Taper (2019) show that given data and a set of models, estimation of the model space structure including the location of the unknown generating process is feasible. This gives a direct measure of model adequacy. Future extensions of this work may allow for the direct estimation of realistic error rates as well.

4.4. How Should One Use Evidential Statistics in Practice?

A basic recommendation is to stop using NP tests for inference and be cautious about using the AIC family of information criteria for model selection. These are known as the “efficient” or “MSE minimizing” criteria and include the AIC, the AICc, the TIC, many forms of ICOMP and the EIC. These criteria are recognized by a complexity penalty whose expectation is asymptotically constant. Asymptotically equivalent to the AIC is the use of leave-one-out cross-validation (Stone, 1977); cross-validation will have model selection properties similar to AIC but has the advantage that it can be calculated in the absence of a likelihood function.

There is no reason that the multiple comparisons inference from traditional ANOVAs cannot be made using information criteria (e.g., Kemp et al., 2004; Jerde et al., 2019).

Classical methods will work well for state description and less well for process identification. Unbiased scientific inferences of process are better made using consistent information criteria (see Jerde et al., 2019; Lorah and Womack, 2019 for examples). Analysts have a convenient spectrum of choices for many standard modeling situations in a suite of consistent information criteria: The HQIC (also known as the HQC, Hannan and Quinn, 1979), the HIC (aka BIC$*$ and HBIC, Haughton, 1988), the SIC (aka BIC and SBC, Schwarz, 1978), and the CAIC (Bozdogan, 1987). The analyst can opt for a criterion that matches her goals. The sample size multiplier in the HQIC grows at the minimal rate to generate a consistent form. As a consequence the HQIC will behave very much like the AIC, selecting models with low MSE of prediction by capturing real but small effects at the cost of including spurious covariates. The HIC tends to balance underfitting and overfitting errors. The SIC and CAIC both favor compact models, with all the included components well-supported, and both tend to underfit. The CAIC has the strongest complexity penalty and thus makes the most underfitting errors and the fewest overfitting errors.

Besides being influenced by inferential goals, the choice of evidence function should depend on the modeling framework. Information criteria had their beginnings as a tool for variable selection in linear regression with independent observations. In such situations, as derived by Akaike, the number of parameters is a good first order bias correction to the observed likelihood. But, statistics is a world of special cases. The dizzying diversity of information criteria in the literature produces the desire to optimize the bias correction under different modeling frameworks. For instance, in mixed models, even the meaning of the number of parameters or the number of observations becomes ambiguous due to the dependence structure of mixed models. Information criteria have been developed using estimates of the effective number of parameters (e.g., Vaida and Blanchard, 2005; You et al., 2016). Similarly, information criteria have been constructed using estimates of the effective number of observations (e.g., Jones, 2011; Berger et al., 2014).

If the generating process is in the model set, or in flat model spaces, such as those in linear regression, the ΔAIC is an unbiased estimate of $2n\Delta K$ regardless of how near or far each of the approximating models is to the generating process (Burnham and Anderson, 2002; Choi and Kiefer, 2011). In curved model spaces (as in Efron, 1975), ΔAIC is not unbiased, and the estimation is only good if both approximating models are close to the generating process. The Takeuchi's information criterion, the TIC (Takeuchi, 1976; Shibata, 1989), is nearly unbiased even for curved models at great distances from the generating process (Burnham and Anderson, 2002; Choi and Kiefer, 2011). Optimal multiplicative coefficients of bias adjustment for the AIC and TIC have been given (Ogasawara, 2016). Also, Ogasawara showed that when the penalty term in TIC (a random variable, not a constant) is negatively correlated with the main term, the higher-order asymptotic variance of the TIC becomes smaller than that common to the AIC and BIC. Unfortunately, the complexity penalty for the TIC must be estimated from data and cannot be specified a priori, as with the other criteria mentioned. The uncertainty in penalty estimation makes the use of the TIC impractical unless sample size is large. A second problem with the TIC is that like the AIC, it is not consistent, but any efficient information criterion can be made consistent either by multiplying the complexity penalty by a consistent multiplier (Nishii, 1988) or by averaging the penalty with a consistent penalty (Lorah and Womack, 2019). Lorah and Womack (2019) also report on testing a list of various model selection criteria. In a nutshell, model selection criteria made into evidence functions as a whole give reasonable and responsible results, with none of the criteria being universally best. Which evidence function is better depends on the nature of the problem at hand, that is, the characteristics of the model space being investigated. The technical difficulties of criterion selection aside, the most important aspect of applying evidential statistics is approaching problems evidentially.

5. CONCLUSION

Evidence is not so much a new statistical method for model selection as it is a new way of thinking about the inference involved with existing model selection methods. The evidential way of thinking has two main components: (1) A post-data trichotomy of outcomes (strong evidence for model f_i, weak or inconclusive evidence, strong evidence for model f_j). (2) A framework of pre-data error probabilities, which are assured to go to zero as sample size increases. The evidential approach invites exploration of the error probabilities, usually via simulation, to aid in study design, the selection of evidence thresholds, the effects of different types of misspecification, and the interpretation of study results.

We have proposed here a different way of thinking about statistical analyses and model selection, based on the concept of evidence functions. Evidence is an intuitive way to decide between two models that avoids the famously upside-down logic that accompanies Neyman-Pearson testing. Evidential thinking has helped us reveal the shortcomings of Fisher significance analysis and Neyman-Pearson testing. The errors that can arise in evidence analysis are straightforward to explain, and the frequentist properties of such errors as functions of sample size and effect size are easy to understand and highly compelling in a scientific chain of argument. The information indexes, when differenced, represent a collection of potential evidence functions that extend the evidence ideas to models with unknown parameters. The desirable error properties are preserved in the presence of model misspecification, when the model choice is generalized to be an inference about which model is closer to the stochastic process that generated the data. The error properties of AIC and AICc are similar to those of Neyman-Pearson testing when the candidate models are nested or overlapping and so the AIC-type indexes are not satisfactory evidence functions in those common circumstances. The indexes like SIC in which the parameter penalty is an increasing function of sample size retain the frequentist error properties of evidence functions for all model pairs.

Evidence works well for science in part because its explicit conditioning on the model set invites thinking about new models. Evidence has inferential qualities that match or surpass Fisher significance analysis and Neyman-Pearson tests. Evidence represents a compelling scientific warrant for formulating statistical analyses as model selection problems.

DATA AVAILABILITY STATEMENT

This paper's code is available at https://github.com/jmponciano/ModelsProjection.

AUTHOR CONTRIBUTIONS

BD wrote the initial draft of the manuscript. BD, JP, and MT jointly derived the mathematical statistics results in Idaho, the summers of 2015, 2016, and 2017, and JP wrote an initial draft of these results. Figures were drawn by JP, BD, and MT. SL and MT contributed with critical insights, discussion, re-organization of the manuscript as well as editing. The ideas reflected in this work started to coalesce when all authors organized courses and lectures at the Center for Research in Mathematics, CIMAT A.C. in 2008, 2009, and 2010, and was reprised when MT and JP organized the International Symposium What is a good model?: Evidential statistics, information criterion and model evaluation held January 2016 at the Institute of Statistical Mathematics in Tokyo, Japan, in which all authors participated.

FUNDING

All authors were partially funded via a grant from JSPS KAKENHI Grant Number 25340110 (to Shimatani Kenichiro) and from the Risk Analysis Center in the Institute of Statistical Mathematics, Japan. JP was partially supported by the US National Institute of Health grant number R01 GM103604 to him.

ACKNOWLEDGMENTS

The authors appreciate the feedback and insights from the audience at the International Symposium What is a good model?: Evidential statistics, information criterion and model evaluation. We also appreciate the insightful comments from the audiences from Virginia Tech's Statistics Department Seminar and University of Idaho's Applied Statistics Seminar, specially to Drs. Leah Thompson and Bill Price.

REFERENCES

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95, 631–636. doi: 10.1890/13-1452.1

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. Petrov, and F. Csaki (Budapest: Akademiai Kiado), 267–281.

Akaike, H. (1974). A new look at statistical-model identification. *IEEE Trans. Autom. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1981). Likelihood of a model and information criteria. *J. Econ.* 16, 3–14. doi: 10.1016/0304-4076(81)90071-3

Anderson, D., Burnham, K., and Thompson, W. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manag.* 64, 912–923. doi: 10.2307/3803199

Anderson, D., Burnham, K., and White, G. (1994). Aic model selection in overdispersed capture-recapture data. *Ecology* 75, 1780–1793. doi: 10.2307/1939637

Anderson, D. R., Burnham, K. P., Gould, W. R., and Cherry, S. (2001). Concerns about finding effects that are actually spurious. *Wildl. Soc. Bull.* 29, 311–316. Available online at: http://www.jstor.org/stable/3784014

Anderson, S., and Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun. Stat. Theory Methods* 12, 2663–2692.

Arnold, T. W. (2010). Uninformative parameters and model selection using akaike's information criterion. *J. Wildl. Manag.* 74, 1175–1178. doi: 10.1111/j.1937-2817.2010.tb01236.x

Barker, R. J., and Link, W. A. (2015). Truth, models, model sets, aic, and multimodel inference: a Bayesian perspective. *J. Wildl. Manag.* 79, 730–738. doi: 10.1002/jwmg.890

Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. New York, NY: Chapman and Hall; CRC.

Berger, J., Bayarri, M., and Pericchi, L. (2014). The effective sample size. *Econ. Rev.* 33, 197–217. doi: 10.1080/07474938.2013.807157

Blanchet, F. G., Legendre, P., and Borcard, D. (2008). Forward selection of explanatory variables. *Ecology* 89, 2623–2632. doi: 10.1890/07-0986.1

Box, G. E. (1979). "Robustness in the strategy of scientific model building," in *Robustness in Statistics*, ed G. Wilkinson (New York, NY: Academic Press), 201–236.

Box, J. (1978). *RA Fisher: The Life of a Scientist*. New York, NY: John Wiley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Burnham, K. P., and Anderson, D. R. (2001). Kullback-leibler information as a basis for strong inference in ecological studies. *Wildl. Res.* 28, 111–119. doi: 10.1071/WR99107

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology* 96, 2370–2382. doi: 10.1890/14-1639.1

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. A* 158, 419–444. doi: 10.2307/2983440

Choi, H.-S., and Kiefer, N. M. (2011). Geometry of the log-likelihood ratio statistic in misspecified models. *J. Stat. Plan. Infer.* 141, 2091–2099. doi: 10.1016/j.jspi.2010.12.019

Claeskens, G., and Hjort, N. L. (2008). *Model Selection and Model Averaging*. New York, NY: Cambridge Books.

Connor, E. F., and Simberloff, D. (1979). The assembly of species communities: chance or competition? *Ecology* 60, 1132–1140.

Costantino, R. F., Desharnais, R. A., Cushing, J. M., Dennis, B., Henson, S. M., and King, A. A. (2005). Nonlinear stochastic population dynamics: the flour beetle tribolium as an effective tool of discovery. *Adv. Ecol. Res.* 37, 101–141. doi: 10.2307/1936961

Dixon, P. M. (1998). "12. Assessing effect and no effect with equivalence tests," in *Risk Assessment: Logic and Measurement*, eds N. M.C., and C. Strojan (Chelsea, MI: Ann Arbor Press), 275–301.

Edwards, A. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.* 3, 1189–1242.

Ellison, A. M., Gotelli, N. J., Inouye, B. D., and Strong, D. R. (2014). *P* values, hypothesis testing, and model selection: it's déjà vu all over again 1. *Ecology* 95, 609–610. doi: 10.1890/13-1911.1

Fisher, R. A. (1926). The arrangement of field experiments. *J. Ministry Agriculture* 33, 503–513. Available online at: https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15191/1/48.pdf

Foster, D. P., and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* 22, 1947–1975.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* 5, 189–211. doi: 10.1080/19345747.2011.618213

Gerrodette, T. (2011). Inference without significance: measuring support for hypotheses rather than rejecting them. *Mar. Ecol.* 32, 404–418. doi: 10.1111/j.1439-0485.2011.00466.x

Grueber, C., Nakagawa, S., Laws, R., and Jamieson, I. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *J. Evol. Biol.* 24, 699–711. doi: 10.1111/j.1420-9101.2010.02210.x

Guthery, F. S., Brennan, L. A., Peterson, M. J., and Lusk, J. J. (2005). Information theory in wildlife science: critique and viewpoint. *J. Wildl. Manag.* 69, 457–465. doi: 10.2193/0022-541X(2005)069[0457:ITIWSC]2.0.CO;2

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. B Methodol.* 41, 190–195.

Harrell, F. E. Jr. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York, NY: Springer.

Haughton, D. M. (1988). On the choice of a model to fit data from an exponential family. *Ann. Stat.* 16, 342–355.

Hurlbert, S. H., and Lombardi, C. M. (2009). Final collapse of the neyman-pearson decision theoretic framework and rise of the neofisherian. *Ann. Zool. Fennici* 46, 311–349. doi: 10.5735/086.046.0501

Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307.

Hurvich, C. M., and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *Am. Stat.* 44, 214–217.

Jerde, C. L., Kraskura, K., Eliason, E. J., Csik, S., Stier, A. C., and Taper, M. L. (2019). Strong evidence for an intraspecific metabolic scaling coefficient near 0.89 in fish. *Front. Physiol.* 10:1166. doi: 10.3389/fphys.2019.01166

Johnson, D. H. (1999). The insignificance of statistical significance testing. *J. Wildl. Manag.* 63, 763–772.

Johnson, J. B., and Omland, K. S. (2004). Model selection in ecology and evolution. *Trends. Ecol. Evol.* 19, 101–108. doi: 10.1016/j.tree.2003.10.013

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions, 3rd Edn.* Hoboken, NJ: John Wiley & Sons.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* 30, 3050–3056. doi: 10.1002/sim.4323

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*. New York, NY: Wiley.

Karlin, S., and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Stat.* 27, 272–299.

Kemp, W., Bosch, J., and Dennis, B. (2004). Oxygen consumption during the life cycles of the prepupa-wintering bee megachile rotundata and the adult-wintering bee osmia lignaria (hymenoptera: Megachilidae). *Ann. Entomol. Soc. Am.* 97, 161–170. doi: 10.1603/0013-8746(2004)097[0161:OCDTLC]2.0.CO;2

Kemp, W. P., and Dennis, B. (1991). Toward a general model of rangeland grasshopper (orthoptera: Acrididae) phenology in the steppe region of montana. *Environ. Entomol.* 20, 1504–1515.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Monogr.* 62, 67–118.

Lele, S. (2004). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, eds M. Taper, and S. Lele (Chicago, IL: The University of Chicago Press), 191–216.

Lindsay, B. G. (2004). "Statistical distances as loss functions in assessing model adequacy," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. Taper, and S. Lele (Chicago, IL: The University of Chicago Press), 439–488.

Link, W. A., and Barker, R. J. (2006). Model weights and the foundations of multimodel inference. *Ecology* 87, 2626–2635. doi: 10.1890/0012-9658(2006)87[2626:MWATFO]2.0.CO;2

Loehle, C. (1987). Hypothesis testing in ecology: psychological aspects and the importance of theory maturation. *Q. Rev. Biol.* 62, 397–409. doi: 10.1086/415619

Lorah, J., and Womack, A. (2019). Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behav. Res. Methods* 51, 440–450. doi: 10.3758/s13428-018-1188-3

Mac Nally, R., Duncan, R. P., Thomson, J. R., and Yen, J. D. (2018). Model selection using information criteria, but is the "best" model any good? *J. Appl. Ecol.* 55, 1441–1444. doi: 10.1111/1365-2664.13060

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mayo, D. G., and Spanos, A. (2004). Methodology in practice: statistical misspecification testing. *Philos. Sci.* 71, 1007–1025. doi: 10.1086/425064

Mayo, D. G., and Spanos, A. (2006). Severe testing as a basic concept in a neyman–Pearson philosophy of induction. *Br. J. Philos. Sci.* 57, 323–357. doi: 10.1093/bjps/axl003

McDonald, L., and Erickson, W. (1994). "Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed," in *Statistics in Ecology and Environmental Monitoring*, eds D. Fletcher, and B. Manly (Dunedin: University of Otago Press), 183–197.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *Ann. Math. Stat.* 19, 58–65.

Mundry, R., and Nunn, C. L. (2008). Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am. Nat.* 173, 119–123. doi: 10.1086/593303

Murtaugh, P. A. (2009). Performance of several variable-selection methods applied to real ecological data. *Ecol. Lett.* 12, 1061–1068. doi: 10.1111/j.1461-0248.2009.01361.x

Murtaugh, P. A. (2014). In defense of *p* values. *Ecology* 95, 611–617. doi: 10.1890/13-0590.1

Neyman, J., and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. A* 231, 289–337.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivar. Anal.* 27, 392–403.

Ogasawara, H. (2016). Optimal information criteria minimizing their asymptotic mean square errors. *Sankhya B* 78, 152–182. doi: 10.1007/s13571-016-0115-9

Pardo, L. (2005). *Statistical Inference Based on Divergence Measures*. Boca Raton, FL: Chapman and Hall; CRC.

Parkhurst, D. F. (2001). Statistical significance tests: equivalence and reverse tests should reduce misinterpretation: equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. *Bioscience* 51, 1051–1057. doi: 10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Ponciano, J. M., and Taper, M. L. (2019). Model projections in model space: a geometric interpretation of the AIC allows estimating the distance between truth and approximating models. *Front. Ecol. Evol.* doi: 10.3389/fevo.2019.00413

Quinn, J. F., and Dunham, A. E. (1983). On hypothesis testing in ecology and evolution. *Am. Nat.* 122, 602–617.

Rao, C., Wu, Y., Konishi, S., and Mukerjee, R. (2001). On model selection. *Lect. Notes Monogr. Ser.* 38, 1–64. doi: 10.1214/lnms/1215540960

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Vol. 2. New York, NY: Wiley.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Belmont, CA: Brooks; Cole.

Richards, S. A. (2005). Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology* 86, 2805–2814. doi: 10.1890/05-0074

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. London, UK: Chapman & Hall.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *J. Am. Stat. Assoc.* 95, 760–768. doi: 10.1080/01621459.2000.10474264

Royall, R., and Tsou, T.-S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *J. R. Stat. Soc. B Stat. Methodol.* 65, 391–404. doi: 10.1111/1467-9868.00392

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *Am. Stat.* 40, 313–315.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. New York, NY: D. Reidel.

Samaniego, F. J. (2014). *Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists*. Boca Raton, FL: CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Stat.* 8, 147–164.

Shibata, R. (1989). "Statistical aspects of model selection," in *From Data to Model* (London: Springer), 215–240.

Spanos, A. (2010). Akaike-type criteria and the reliability of inference: model selection versus statistical model specification. *J. Econ.* 158, 204–220. doi: 10.1016/j.jeconom.2010.01.011

Spanos, A. (2014). Recurring controversies about p values and confidence intervals revisited. *Ecology* 95, 645–651. doi: 10.1890/13-1291.1

Stephens, P. A., Buskirk, S. W., Hayward, G. D., and Del Rio, C. M. (2005). Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* 42, 4–12. doi: 10.1111/j.1365-2664.2005.01002.x

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. R. Stat. Soc. B Methodol.* 39, 44–47. doi: 10.1111/j.2517-6161.1977.tb01603.x

Strong, D., Whipple, A., Child, A., and Dennis, B. (1999). Model selection for a subterranean trophic cascade: root-feeding caterpillars and entomopathogenic nematodes. *Ecology* 80, 2750–2761.

Strong, D. R. (1980). Null hypotheses in ecology. *Synthese* 43, 271–285.

Stroud, T. (1972). Fixed alternatives and Wald's formulation of the noncentral asymptotic behavior of the likelihood ratio statistic. *Ann. Math. Stat.* 43, 447–454. doi: 10.1214/aoms/1177692625

Symonds, M. R., and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behav. Ecol. Sociobiol.* 65, 13–21. doi: 10.1007/s00265-010-1037-6

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Math. Sci.* 153, 12–18.

Taper, M. (2004). "Model identification from many candidates," in *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, eds M. Taper, and S. R. Lele (Chicago, IL: The University of Chicago Press), 448–524.

Taper, M., and Lele, S. (2011). "Evidence, evidence functions, and error probabilities," in *Handbook of the Philosophy of Science, Volume 7: Philosophy of Statistics*, eds P. Bandyopadhyay, and M. Forster (London: Elsevier), 439–488.

Taper, M. L., and Lele, S. R. (2004). *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago, IL: The University of Chicago Press.

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Pop. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

Thompson, B. (2007). *The Nature of Statistical Evidence*. New York, NY: Springer.

Underwood, T. (1986). "Analysis of competition by field experiments," in *Community Ecology: Pattern and Process*, ed J. Kikkawa, and D. J. Anderson (London, UK: Blackwell), 240–268.

Vaida, F., and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika* 92, 351–370. doi: 10.1093/biomet/92.2.351

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* 54, 426–482.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* 16, 117–186.

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecol. Modell.* 211, 1–10.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Boca Raton, FL: Chapman and Hall; CRC.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.* 75, 1182–1189. doi: 10.1111/j.1365-2656.2006.01141.x

Wilkinson, L., and Dallal, G. E. (1981). Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics* 23, 377–380.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* 9, 60–62.

Yoccoz, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull. Ecol. Soc. Am.* 72, 106–111.

You, C., Müller, S., and Ormerod, J. T. (2016). On generalized degrees of freedom with application in linear mixed models selection. *Stat. Comput.* 26, 199–210. doi: 10.1007/s11222-014-9488-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

HYPOTHESIS AND THEORY
published: 22 October 2019
doi: 10.3389/fevo.2019.00399

Ecology, Evidence, and Objectivity: In Search of a Bias-Free Methodology

*Gordon Brittan Jr. and Prasanta Sankar Bandyopadhyay**

Department of History and Philosophy, Montana State University, Bozeman, MT, United States

For at least the past 25 years or so, there has been a twofold sense of "crisis" in ecology. One indication of this is the spate of articles and books calling for a reformation of the discipline and bearing such titles as "The New Ecology." On the part of practitioners, the unease concerns its theories, concepts, and methods. On the part of the general public, the unease concerns the perceived "bias" of its results. This paper is an attempt by two philosophers of science to clarify one critical methodological issue—hypothesis/model testing—and in the process to identify ways to gird the objectivity of ecological claims. What is significant about our approach is a distinction between the tasks appropriate to Bayesian Inference and Evidential Statistics—confirming hypotheses on the one hand and measuring evidence for models on the other. These two inferential paradigms are contrasted with the testing methods long-dominant in the discipline—Fisher-Neyman-Pearson Significance Testing and Popper Falsificationism—and a case made for a much greater use of Bayesian and Evidentialist Methods. In particular, it is argued that Evidential Statistics, here in the form of the likelihood ratios of competing predictive and explanatory multiple models avoids the main forms of otherwise unsettling cognitive bias. It also provides a Darwinian alternative to the "convergence" accounts of objectivity associated with the development of physics which is more appropriate to ecology.

Keywords: bayesian inference, evidential statistics, significance testing, falsificationism, hypothetico-deductivism

OPEN ACCESS

Edited by:
Yukihiko Toquenaga, Graduate School of Life and Environmental Sciences, University of Tsukuba, Japan

Reviewed by:
Jonah N. Schupbach, The University of Utah, United States
Christopher L. Jerde, University of California, Santa Barbara, United States

****Correspondence:***
Prasanta Sankar Bandyopadhyay
psb@montana.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *19 March 2019*
Accepted: *08 October 2019*
Published: *22 October 2019*

Citation:
Brittan G Jr and Bandyopadhyay PS (2019) Ecology, Evidence, and Objectivity: In Search of a Bias-Free Methodology. Front. Ecol. Evol. 7:399. doi: 10.3389/fevo.2019.00399

INTRODUCTION

Twenty-five years ago, Shrader-Frechette and McCoy (1993) wrote that

> On the whole, general ecological theory has, so far, been able to provide neither the largely descriptive, scientific conclusions often necessary for conservation decisions, nor the normative basis for policy.

Judging by the titles of more recent textbooks, and despite an immense amount of very interesting ecological research and theorizing carried out in the meantime, the situation appears basically unchanged. These books bear such titles as *Scientific Method for Ecological Research* (Ford, 2000), *Ecological Understanding: The Nature of Theory and the Theory of Nature* (Pickett et al., 2007) and *The New Ecology: Re-Thinking A Science for the Anthropocene* (Schmitz, 2017). All are premised on the complex claim that there is as yet little consensus on either the correct theoretical structures or the proper experimental/inferential methods of the subject; the result is that ecological science has not yet had the desired and necessary influence on policy formation and implementation.

Ford, for example, begins the final chapter of his book with a list of criticisms that he takes seriously. After all, they provide the motives for developing what he takes to be a new and improved approach.

i) There has been a lack of progress in ecology.
ii) No general theory has emerged.
iii) Ecological concepts are inadequate.
iv) Ecologists fail to test their theories.

Picket, Kolesa, and Jones echo the discontent. In their view, at least part of the problem stems from the fact that the great growth of ecological information has occurred in evermore Balkanized sub-disciplines, each with its own assumptions, concepts, methods, and hypotheses. Hence, the progress made has been (in their word) "narrow," focusing on specific scales and levels of organization, and making communication between sub-disciplinarians, not to mention with the general educated public, increasingly difficult. There is no larger and consistent picture on which to get a grip, no uniform set of methods to employ, and (although they do not put it this way) no firm basis on which to formulate, much less implement, coherent public policies–in particular regarding the multi-scale impacts of human actions on specific plant and animal populations. As with these other authors, Schmitz tries to provide a new and more implementable picture.

One source of the discontent both with and within ecology is the relative absence of understanding the role and scope of the methods used to test ecological hypotheses and models. The authors of this paper have been invited to expand this understanding by placing it in a larger philosophical perspective.

THE DEFORESTATION CONTROVERSY: HYPOTHESIS, POLICY, AND LACK OF TRUST

On October 3, 2018, the environment, development, and agricultural heads of the United Nations issued a joint statement declaring that

> Forests are a major, requisite front of action in the global fight against climate change – thanks to their unparalleled capacity to absorb and store carbon. Stopping deforestation and restoring damaged forests could provide up to 30% of the climate solution (Da Silva et al., 2018).

All well and good. On the assumption (on which more later) that climate change is (to a significant degree) human-induced, and given that we have every reason to resist it, we need to stop deforestation and restore damaged forests. The rational place to begin is with a factual assessment of the situation. The immediate problem is that "there are two main data sources for tree loss, and they are increasingly contradictory" (Pearce, 2008). One source is the Global Forest Watch (GFW). Its data are compiled from satellite images by the World Resources Institute. These data indicate a decline in tree cover in 2017 of 72.6 million acres, almost 50% more than in 2015. The other source of deforestation data is the Global Forces Resource Assessment (FRA), which is based on government inventories compiled by the UN Food and Agricultural Organization. It estimates the annual loss at just 8.2 million acres, and adds that deforestation rates have declined by more than 50% since 2008. In individual countries the data-inconsistency is even more dramatic, the FRA showing forest gains in the US, for instance, while the GFW indicates big losses.

In this case, the data-inconsistency can be explained in terms of the types of data gathered—Landsat tree-cover images as against government-designated land uses—employed by the two organizations. More inclusive and sophisticated models are being developed[1]. But it is not at all clear whether they will reinforce the on-going attempt to protect intact forests or put the emphasis on re-growing temporarily degraded areas. The correct policy perspective depends, at least to an important extent, on the time-scale chosen. Once-deforested areas in New England are now overgrown with trees.

Even when there is a clear consensus among scientists about both fact and policy, the general public is often slow to follow. *Yale Environment 360*[2] ran the headline, "Americans Who Accept Climate Change Outnumber Those Who Don't 5-1" on April 4, 2018, but a closer look at the survey numbers indicates that no more than 58% believe that global warming is mostly caused by human factors, and no more than 49% (2% *less* than in 2008) are "extremely" or "very" sure that it is really happening. Again according to the Yale survey, only 6% of the population believes that anything much can be done to slow or reverse it.

ENTER PHILOSOPHY OF SCIENCE

There are, of course, many reasons for the discrepancy between expert and popular opinion. Some of them are familiar—politics, economics, spatial and temporal scales. But what runs through all of them is distrust, sometimes of "science" generally, on religious or other cultural grounds, more often of ecology or similarly policy-connected disciplines. The main brief against them is that their research is often "biased," aligned in one way or another with "liberal" or "environmentalist" agendas. In one word, ecology and its brethren are not "objective," and for that reason not to be taken seriously. This is disturbing not only from a policy perspective, but also because a good percentage of ecological research is government-funded and depends on broad political support.

It is to be expected, then, that ecology textbooks would concentrate as they tend to do on questions concerning *objectivity*—how it is to be understood and obtained. Since the hallmark of and the means by which it is ensured, at least in our culture for the past several hundred years, has been the "scientific method," much of the discussion in these books quickly focuses on it. The discussion of method, in turn, is deeply informed by the philosophy of science[3].

But can philosophical reflection aid ecologists in either their methodology or their communication with the public? Our aim is to answer the question affirmatively by focusing on the *objectivity*

[1] It is worth noting that neither the GFW nor FRA models is to this point sensitive to changes in biodiversity or carbon uptake in the forests modeled, although both factors enter into cause-of-warming considerations.

[2] Yale Environment 360 (2018, April 14). Americans Who Accept Climate Change Outnumber Those Who Don't 5 to 1."

[3] If we can take the textbooks by Pickett et al., and Ford as representative. Of course, a great many books on general ecology do not focus on methodological issues, although they do underline the necessity of re-conceptualizing the subject.

of the claims that ecologists make. Objectivity in turn has to do with the methods by which these claims are tested. This is the nub of the controversies surrounding ecology as regards both its scientific status and reliable source of informed public policy. It is also the way in which the indispensability of philosophical reflection can best be demonstrated. A brief review of the testing methods already in widespread practice should provide context, and a distinction between the concepts of confirmation and evidence add clarity.

HYPOTHESIS-TESTING METHODS IN ECOLOGY

Hypothetico-Deductive Testing

At present, inferential methods are routinely characterized within one or another statistical framework. It was not always thus. Discussions of theory-testing, at least among philosophers interested in the subject, were dominated in the middle years of the twentieth century by the so-called Hypothetico-Deductive or H-D model. On this model, to test a hypothesis is, schematically, to derive a statement, via initial and boundary conditions[4], describing an observation. If the derivation is carried out before the observation is made, it is predicted; if detected or measured, the hypothesis is confirmed. If the derivation is carried out after the observation has been made, the hypothesis retrodicts and explains it. The underlying point remains the same: to test a theory is to derive statements describing observations or, ideally, experimental results. If verified, belief that the theory is true has to some indeterminate degree been justified. There were several variations on the H-D model, for example Hempel's view (1965) that not the observational consequences but the instantiations of empirical hypotheses justified or confirmed belief in them, but the leading theme remained untouched, that the credibility of scientific claims rested on successful prediction/explanation and that prediction/explanation in turn could be characterized by a simple deductive relationship between hypotheses and observations or (to use a more bracing and embracing term) *data*[5].

A relatively early and interesting case for the H-D model as a reliable way of testing wildlife, and by extension ecological, hypotheses generally, was made by Romesburg (1981), although in doing so he departed from the Positivist original in a significant way. On his account, wildlife science was dominated into the 1980's, although in his view wrongly, by the methods of "induction" and "retroduction." On Romesburg's somewhat non-standard use of the terms, the former involves correlating variables, the amount of edge vegetation in fields, say, with an index of game abundance; the greater the degree of observed correlation, the more reliable the hypothesis linking the variables. The latter (retroductive) method involves providing an explanation of the observed linkages simply by providing a generalization from which all of them can be derived.

According to Romesburg, the major difficulty with both methods is that they are used to *generate* rather than to *test* hypotheses. In his view, a subsidiary difficulty with the inductive method is that it wrongly assimilates correlation to causation; that two variables are usually, if not also invariably, conjoined does not by itself demonstrate a *direct* (or *directional*) causal connection between them[6]. A reliable hypothesis must in one way or another *explain* the connection, it must provide a reason for and not simply "fit" it. A subsidiary difficulty with the retroductive method is that it is tied closely to the facts that it is invoked to explain; it doesn't provide a way of ruling out incompatible hypotheses that explain all the same facts. Although Romesburg doesn't put it this way, one might say that the inductive method leads to predictive but not explanatory hypotheses, the retroductive method to explanatory but not predictive hypotheses and that any adequate ("reliable") hypothesis must be both explanatory and predictive. It is only if they satisfy both criteria that hypotheses are testable. In a Positivist vocabulary, induction and retroduction are methods of "discovery," not "justification," and discovery is methodologically moot; for the most part, adequate hypotheses are *invented*, products of insight and imagination. Justification alone has its own logic[7].

The distinction between *discovery* and *justification* is classically Positivist, Romesburg's distinction between *prediction* and *explanation*[8] is not. On the original H-D model, prediction and explanation are asymmetrical only with respect to the time, before or after the fact, when the derivation of an observational

[4] And if necessary, rules by which to translate theoretical terms in the hypotheses so that they had observational content and application, usually in the form of measurable quantities. These rules were often referred to as "operational definitions." That said, there is no commonly-accepted way in which to characterize such "definitions." Perhaps most often it is to provide quantitative indices for the application of theoretical terms, means by which they may be measured and thereby applied to observational or experimental data. It has proven to be particularly difficult to operationalize theoretical terms in ecology—think "ecosystem," "niche," and "diversity" (all of which have come to have normative dimensions). One virtue of testing mathematical models is that they postpone the problem; to test the model is simply to measure the quantities that it contains and verify the data-distributions in which it issues. It can later be decided how the model should, if desired, be integrated into a more explanatory and policy-guiding theory.

[5] Elaboration of the H-D model included attempts to characterize "data" more precisely as well, including the methods of their measurement and the errors to which it would inevitably be subject, but nothing in what follows turns on these attempts.

[6] They may be linked by a common cause or confounded with another variable, for example.

[7] Saint-Mont (2018) has recently made an up-dated and well-informed case for the "inductive" (data-first) approach to testing. On the assumption that samples test generalizations about populations, the law of large numbers guarantees that the distance between them shrinks quickly as the sample increases in size, and "the true distribution comes into focus almost inevitably" (p. 686). Saint-Mont's perspective contrasts sharply with the hypothesis/model-first approach of the other accounts of testing we will consider (although he includes elements of these accounts in his own; the implication is correct, both models and data are involved inferentially, in this respect like the analysis of ecosystems, trophic cascade from the top down, nutrient supply from the bottom up). Although it is in certain respects problematic, he implicitly blurs Romesburg's line between causation/explanation and correlation, ignores problems associated with (random) sampling, and shares the questionable "true-model" aim of testing with other statistical paradigms. For all of its sophistication, Saint-Mont's view of testing represents a return to a form of Positivism on which the role of theoretical concepts in science/ecology is at best unclear and predictive success is the sole criterion of evaluation.

[8] Or his corollary distinction between correlation and causation.

consequence is carried out[9]. Romesburg's case for restricting the model to explanatory, which is to say causal, hypotheses (although he does not frame it as a restriction) rests on the close connection he posited between wildlife science and public policy; it is only when "cause-effect relationships among variables are found [that] control [of outcomes] is possible" (p. 304).

The difficulties with the structural identity of prediction and explanation aside, a number of criticisms were later made of the H-D model (or better: account) of theory-testing. Three of these criticisms proved to be of special significance, not only because they undermined the H-D account, but more importantly because they led to alternative and very fruitful testing accounts. The first criticism was that the H-D account is no more than qualitative. It provides necessary and sufficient conditions for the truth of "*D* confirms *H*," but without a rule for determining the *degree* to which it is able to do so. This is troubling. An adequate account of confirmation should capture the universally-held belief that some hypotheses are (perhaps much) better supported by the available data than others, and be able to measure the difference. The second criticism of the H-D account was that while it indicates a logical relationship, usually entailment, between hypothesis and data, it does not specify an inverse relationship, neither entailment nor any other, between data and hypothesis, no way, so to speak, to retrace the bottoms-up route. The third main criticism was that the H-D account is, without further modification and amendment, restricted to non-statistical hypotheses, typically illustrated by universal generalizations of the form "All A are B." That is, from a simple statistical hypothesis of the form "$\Pr(B|A) = r$" it does not follow logically that a description of *A* entails a description of *B*. In fact, it doesn't follow that the probability of an instance of *B* given an instance of *A* is equal to *r*. In such cases, the relationship between hypothesis and data must be inductive and characterized in probabilistic terms.

One way to lump all three of these criticisms together is to say that the hypothetico-deductive account had some serious *gaps* in it. The option was to fill the gaps and in the process reconfigure the structure of scientific testing. Several alternative and gap-filling accounts have been proposed. The first is deductive in character, the three others are statistical.

Falsification and Corroboration

The first alternative was set out by Popper in his classic *The Logic of Scientific Discovery* (Popper, 1934/1959). His approach was striking both in its ease of application and intuitive appeal[10]: Re-construe the H-D account in such a way that there are *no* gaps to fill. In Popper's view, this is fortunate since there is no way in which they can be filled coherently in any case. His point of departure was the fact that while a hypothesis can never be "confirmed," it can be *falsified*. The point is purely logical. No number of confirming instances, no matter how great, can ever guarantee that a universal generalization is true. Yet a single disconfirming consequence will show, other things being equal, that the generalization is false in a deductively straightforward way. It doesn't follow from the fact that any number of swans are white that all swans are, but it does follow from the fact that there is a black swan[11] that the generalization concerning them is false. Moreover, in the case of falsification there are no gaps to fill, no new relationship between data and hypothesis to be discovered or invented, no need to add probabilistic operators and rules governing them to our traditional methods. The rule of *modus ponens*—if *p* then *q* and $\sim q$, therefore $\sim p$–by itself suffices as the "logic," not so much of justification (for there is no such thing according to Popper) as of scientific discovery.

Popper reinforces his proposal by way of a reflection on actual scientific practice. Scientists do not keep repeating the same experiments in the attempt to pile up confirming consequences of a hypothesis (although they do attempt to diversify the conditions with respect to which these consequences are derived). Once an experiment has been performed, and replicated by others, they move on to other ways in which to *test* the hypothesis. But, Popper contends, to (really) test a hypothesis is to find new ways to falsify it, other kinds of data. Since no hypothesis can ever be *established* as true, the best one can say of a particular hypothesis is that it has survived a number of tests, the more varied and severe the better. A hypothesis which has so survived is said to be *corroborated*, i.e., has not been shown to be false. In science as in the biotic community, the fittest survive. The idea that biotic communities are self-regulating, that there is "a balance in nature," is an old, indeed ancient, ecological truism. Yet it has been shown over the last 30 years or so that the assumptions on which such equilibrium rests do not hold generally[12]. This is, at least according to the conventional wisdom, what characterizes the scientific mind: never to accept some truth as given, but to question it constantly.

[9] Romesburg's article, though written almost 40 years ago, still makes good reading, not only because of an extended (and mathematically-sophisticated) description of how Errington's constant threshold-of-security hypothesis ("For a given area and species, the number of animals surviving fall to spring can be no greater than a threshold value. This threshold accounts for all forms of natural mortality, barring catastrophic weather events, and is constant from year to year") is to be reconstructed/tested on the H-D model, but also because of his careful attention to the details of evaluating the observational consequences (for the most part statistical) of the hypothesis, the vagaries of "general-purpose data" not collected under controlled conditions, and the necessity of cost/benefit analyses of experiments before they are actually initiated. For Errington's classic study (later modified to include a variable threshold), see Errington (1945).

[10] Indeed, it is difficult to overstate the impact of Popper's account on the methodology of practicing scientists, among which ecologists. Thus, the bio-scientists Cassey and Blackburn (2006): "It is widely agreed that modern scientific inference relies on the vulnerability to refutation of its general theories, which have the characteristic quality of being both general and falsifiable." Indeed, there are more references in the index to Ford's book to Popper than to anyone else, philosopher or scientist. Neither Ford nor Picket et al., discuss either Bayesianism or Evidentialism, although Pickett et al.'s, discussion of "pairwise alternative hypothesis testing" and the reference in it to Platt (1964) include elements of the latter.

[11] The stunning *Cygnus atratus* discovered in 1790 by Latham.

[12] See Botkin (1990). For Schmitz (2017), the "New Ecology" rests principally on a rejection of the twin classic theses that ecosystems are (relatively) self-regulating and isolated (from each other and, as objects of study, from human intervention).

Popper is right to stress the "testing" intuition[13]. But whatever logical advantage a program of principled falsification enjoys is no more than apparent. The French physicist, philosopher, and historian of science, Duhem (1962) was perhaps the first to emphasize that hypotheses are never tested in isolation, but only in conjunction with other hypotheses and appropriate initial and boundary conditions[14]. A negative result does not by itself show which of these hypotheses or conditions is false. To put it another way, the logical asymmetry to which Popper draws attention is matched by another: a confirming prediction confirms *all* of the hypotheses and conditions from which it follows; a falsifying observation does not similarly falsify all of the hypotheses and conditions from which *it* follows[15].

Finally, the Popperian methodology shares an important difficulty with H-D accounts generally. Both are premised on the assumption that hypotheses take the form of universal conditionals. But it is often the case, perhaps almost always in ecology, that hypotheses have a probabilistic or statistical form. We have already referred to the difficulty in deducing observational consequences from such hypotheses. The falsifiability criterion is similarly tailored to "All *A* are *B*" examples. It cannot deal effectively with the multi-factor multi-causal hypotheses typical of ecology. All of this said, it must be added at once that Popper's methodology has not itself been "falsified." A great deal of valuable research has been carried out by ecologists in attempts to follow Popper's guidelines (albeit substituting "rejection" for falsification properly so-called, as is necessarily the case when hypotheses do not take the form of universal conditionals)[16].

In brief, problems with Popper do not show that all of the research done in his name is either misguided or without value. They do prompt us to look for other accounts of hypothesis testing that avoid failures in Popper's own. It is in any case a mistake to fix on one method as uniquely satisfactory. Different testing methods are appropriate as different types of research questions are asked.

Error-Statistical and Significance Testing

Significance or error-statistical testing in fact pre-dated the H-D model, both as regards its initial formalization and its widespread acceptance among ecologists. The latter undoubtedly had to do with the fact that ecological generalizations, even those taken as lawlike, are for the most part statistical in character. It involves a procedure not unlike Popper's. That is, it provides a way of rejecting (not falsifying) hypotheses and at least indirectly provides support for their alternatives. Variants on this testing theme are associated with Fisher, Neyman, and Pearson. Since it is so well-known among ecologists, to the extent that significance testing is virtually synonymous with "statistical testing," and even "testing" *tout court*, there is no need for much detail. It suffices to point out in a very broad way why it is inadequate, and then to discuss briefly its recent redeployment by the philosopher Mayo (1996; 2018) and Mayo and Spanos (2010).

On its Fisher variant, the viability of a hypothesis is probed by comparing an observed result with the distribution of results predicted by the hypothesis. That is, any hypothesis (typically described as "null") is rejected if an observed result (and results more deviant) would be predicted by the hypothesis with a low probability (*P*-value). Commonly, a result is judged "*significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator*" (Fisher, 1929, p. 191), viz., no more than 5% of the time. On the other hand, if results as or more deviant than the observed results would be predicted more than 5% of the time, the proper "Fisherian" conclusion is not to accept the hypothesis, but to recognize a failure to reject the hypothesis. The obvious problem is that any number of otherwise incompatible hypotheses in the same area of research could predict the results and in this very general sense be confirmed[17]. The Fisher singular hypothesis account is too weak to discriminate them.

On the Neyman-Pearson variant[18], "the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the observed result with a greater degree of probability" (Pearson, 1938). One of the hypotheses compared is invariably in practice if not also in theory "null," and the commonly accepted significance level continues to be conventionally set at 0.05, however arbitrary the number. In essence, an NP test is a Fisherian test of the null hypothesis using a test statistic designed to maximally differentiate between the two hypotheses. Usually, this statistic is the likelihood ratio for the two models or its logarithm.

The NP approach differs from Fisher's view in a second respect as well. The Fisherian test has an inductive and rather open-ended character. The Fisherian *P*-value is just something for the scientist to think about when trying to come to grips with nature. The NP test on the other hand is set up to be a clear-cut decision procedure. A critical level (designated α) for the *P*-value of the null hypothesis is set *a priori,* with the result that you either accept the null hypothesis or accept the alternative. An artifact of the black or white nature of NP testing is that small differences in the data can make large differences in inference. While in a properly interpreted Fisherian test, the difference between a *P*-value of 0.051 and 0.049 makes very little difference, in a properly interpreted NP test, if the critical level has been set to 0.05, this small difference makes a great deal of inferential difference.

[13] Up to a point. There are notable examples of non-falsifiable zero-force principles that play an indispensable role, the First Law of Motion in Newtonian mechanics, the Hardy-Weinberg Law in ecology.

[14] See Houston (2014) for a case study in ecology of the ways in which "the logic of every hypothesis is based on the underlying assumptions."

[15] Popper (1974, p. 1035) recognized the difficulty. Yet he does not resolve it beyond leaving it to "the scientific instinct of the investigator," as did Duhem himself. See also *The Logic of Scientific Discovery*, p. 76n. It is also always possible in principle to re-interpret the allegedly falsifying data. See Kidwell and Holland (2002) for a taphonomic/stratigraphic re-interpretation of the fossil record on which it is consistent with classical evolutionary theory (and not, as Darwin himself was worried, a straightforward falsification of it).

[16] An especially interesting ecological example is the study of individualist and community-unit concepts carried out by Shipley and Keddy (1987).

[17] See Anderson et al. (2000). See also Läärä (2009).

[18] Which might more accurately be called Neyman-Pearson or NP *hypothesis testing.*

NP analysts realize that their procedure makes mistakes. Neymann and Pearson distinguished two types of errors: Type I (rejecting a true null hypothesis) and Type II (accepting a false null hypothesis). They console themselves with the belief that they can both measure and control the rate of errors. In fact the magnitude of those error rates is the sole measure of the validity of NP test inferences.

A cryptic consequence of NP test construction now emerges. The calculation of error rates is tightly bound to the assumption that one or the other alternative is true. If the data are generated by some process other than one of the two alternative hypotheses, the calculation of error rates may be deeply disrupted (see Dennis et al., 2019, for a detailed analysis of this problem).

Thus, while the inference from Fisherian tests may be too weak, the inference from NP tests may be too strong. As pointed out by Chatfield (1995), analyzing the wrong models is likely to be the greatest source of error in statistical analysis. Further errors are often made, in part because many ecological hypotheses lack measurable power and precision, in part because of the number and complexity of the variables to be taken into account in the case of field observations. The result, or so outsiders often agree, is a widespread lack of confidence in significance testing generally[19]. On both variants, it is too easy to attribute biological to mere statistical significance.

Mayo attempts to bolster confidence in error-statistical testing by imposing Popper-like severity constraints on it. However, there are at least two ways her account differs from Popper's. First, hers is probabilistic, his deductive. Second, she wants to "go smaller" and focus on testing individual statistical hypotheses; his focus is on testing "global theories" like Newton's and Einstein's. On Mayo's account, an adequately stringent test combines weak and strong severity principles. The weak principle has two key features. One is that a severe test is such that the probability is low that the test procedure would pass a hypothesis subjected to it if the hypothesis were false. The other feature is that the probability that the data agree with the alternative hypothesis is very low. On the strong severity principle, data provide good evidence for a hypothesis if it passes the severe test procedure, that is, is in agreement with the data. Like Popper, Mayo emphasizes that the more severe the test, the greater its probative value. She also shares with him the assumption that hypotheses may be tested individually, in a non-comparative context (or rather, that the test is always with respect to a hypothesis and its negation). But this assumption introduces the potential for bias, not simply by way of adding auxiliaries to it so as to square the hypothesis with the data once errors have been detected in it, but also by leaving out of account that *other* hypotheses might be better supported (more severely tested) by the same data. To alleviate this problem, Mayo and Spanos (2004) advocate "misspecification testing," but this only helps for misspecifications that can be conceived of.

Bayesian Inference

A third and increasingly influential option to the H-D model has been to fill the gaps in it by providing an inverse characterization of the way in which data directly confirm or otherwise support hypotheses. It does so by supplementing deductive logic with the full resources of probability theory and is known as Bayesian Inference.

The first gap in the H-D model is the absence of any way of determining both the means by which and the extent to which data confirm a hypothesis. The gap is filled by Bayes Theorem, so-named after its eighteenth century originator. The Theorem is easily derived from the axioms of probability theory together with a definition of conditional probability. The probabilities within it are interpreted as measures of belief. It states that if the probability of the data is not equal to zero, the posterior probability of the hypothesis is equal to its prior probability, i.e., the willingness of particular agents to bet that it is true, before new or additional data bearing on it have been gathered, multiplied by the probability of the data given the hypothesis divided by the "expectedness" of the data, i.e., the marginal probability of the data averaged over the hypothesis and its alternatives. More compactly,

- $\Pr(H|D) = [\Pr(H) \times \Pr(D|H)]/\Pr(D)$.

On the Bayesian account, to confirm (disconfirm) a hypothesis is just to raise (lower) its prior probability, viz.,

- D confirm H just in case $\Pr(H|D) > \Pr(H)$[20].

This measure of confirmation is qualitative. There are alternative ways to measure the *degree to* which a hypothesis is confirmed, but a common metric is in terms of the difference between the prior and posterior probabilities, $\Pr(H|D) - \Pr(H)$. Whether the degree is "high" or "low" depends on the particular confirmation measure chosen, the implicit standards of disciplinary scientific communities, and the research purposes of the investigator.

It follows as an immediate corollary of the Bayesian account of confirmation that it applies to probabilistic or statistical hypotheses as much as it does to universal conditionals, thus filling a second gap left open by the H-D and falsification accounts.

The third "gap," if such it can be called, left open by the H-D and its falsification variant is that they provide no way on the basis of which to choose hypotheses to put to experimental test. For traditional H-D theorists, they have no particular advance rationale, for Popper they are merely "conjectures" on an individual scientist's part, the bolder the better. But as Aaron Ellison, one of a rather small number of ecologists in the 1980's to urge adoption of Bayesian methods, puts it (Ellison, 1986),

> We rarely, if ever, test all possible hypotheses, and most of us use substantial prior knowledge about the behavior

[19] A lack exacerbated by widespread inability to replicate results published in peer-reviewed articles. It is troubling without further explanation that (a) there is a growing number of P values per ecology article published (since "the more P values, the higher the odds that any given result will be significant even if it's just the result of chance") and (b) "the reported value of the coefficient of determination, R^2, has been falling steadily (suggesting a decrease in the marginal explanatory power of ecology)." See Low-Décarie et al. (2014), and for the first embedded quotation (Stokstad, 2014). Murtaugh (2014), among others, defends the traditional use of P values by ecologists, but on mathematical grounds.

[20] Which might also be put in terms of the relevance of the data to the belief that H is true. I.e., If $\Pr(H|D) = \Pr(H)$ then the data D are irrelevant re belief adjustment.

> of a system in designing our experiments. Unlike classical frequentist statistical practice, Bayesian inference requires the investigator to state assumptions explicitly and use pre-existing information quantitatively to define the prior distribution or hypothesis (p. 1041).

In what are sometimes referred to as "empirical" or "standard" applications of Bayes Theorem, the prior probability distributions are *estimated* on the basis of observed relative frequencies in the data. In non-standard cases, the distributions are a function of the ecologist's previously acquired beliefs (including hunches and intuitions) about the object of investigation.

The fourth and final gap, very much underscored by Ellison, is that Bayesian inference lends itself in a uniquely transparent way to adaptive management and environmental decision-making. On the one hand, just as Bayesian agents begin, as most of us in fact (and rather unconsciously) do, with an initial probability distribution over plausible hypotheses and expected outcomes, up-dating and re-adjusting the distribution as data accumulate, continually learning from experience[21], so too (ideally) adaptive land and wildlife managers treat decisions as hypotheses to be tested, choosing them where possible on the basis of past experience, and modifying them as necessary in the light of the observed outcomes to which they lead. To manage adaptively is to learn from experience, to acknowledge the inevitability of uncertainty is to be open to policy changes as additional data are brought to bear on policies already in place. That the degree of uncertainty with which initial decisions are made can be measured and then re-evaluated as time goes by, moreover, reassures the public that policy shifts are never arbitrary or capricious, and nearly always open to revision.

On the other hand, the usual decision-protocols routinely use Bayes Theorem to calculate optimal courses of action on the basis of the probability of outcomes and their respective utilities. A rational agent—manager, politician, or citizen—chooses the course of action that maximizes the product of the (posterior) probability of its outcome and its expected utility. This is as should be expected. We act rationally in such a way as to maximize our desires (utilities) given that we have particular beliefs (probabilities) concerning the future, at least insofar as our actions are intentional[22].

CONFIRMATION AND EVIDENCE

Ellison admits that

> Not all ecologists … appreciate the philosophical underpinnings of Bayesian inference. In particular, Bayesians and frequentists differ in their definition of probability and in their treatment of model parameters as random variables or estimates of true values. These assumptions must be addressed explicitly before deciding whether or not to analyze ecological data (Ellison, 2004, p. 509).

Agreed: the assumptions must be addressed. In brief, (a) the decision whether or not to use Bayesian methods depends on the *type* of research question being asked, (b) there are several clear differences between these types, and (c) an unaided used of Bayesian methods does not ensure the *objectivity* rightly held to follow from an appropriate use of "scientific method(s)."

There are various types of research question[23]. One is: given a datum, what should I *believe*, and to what degree? This question has to do with the *confirmation* of my beliefs. A second question is: what kind of *evidence* does the datum provide for one hypothesis as against another, and how strong is the evidence? Admittedly, "confirmation" and "evidence" are used interchangeably; D is often taken as evidence for H just in case D confirms H. But they should be distinguished rather sharply. Their conflation is the source of a great deal of error in philosophy, statistics, and perhaps also in the practice of science. Intuitively, *confirmation* is agent-dependent in the sense that a hypothesis is confirmed if and only if the agent's degree of belief in it is raised. Incorporating as it does an agent's *belief*, it in this same sense subjective. *Evidence* in the narrow technical sense used here, a relation between the likelihoods of data/models, however, is agent-independent; it has to do not with raising agents' prior degrees of belief in a hypothesis on the basis of the data subsequently collected, but in assessing the relative probability of the data under one hypothesis as opposed to under another. It is in this sense, objective, incorporating a logical and belief-free relation between data and hypothesis. It is also intuitively comparative. Evidence consists of data more probable

[21] If learning from experience is to be possible, then it is reasonable to insist that learners should not have an *a priori*, and hence prior, beliefs that an empirical hypothesis is true to degree 1, i.e., could not possibly be false, or false to degree 0, i.e., could not possibly be true. Empirical hypotheses are never more than merely probable, which is to say that our beliefs concerning their truth are always uncertain to one degree or another. Nevertheless, some hypotheses are much better confirmed than others, and provide a more secure basis for action. It is the task of an adequate theory of confirmation, or so the Bayesian argues, to make clear the grounds of the difference. Although uncertainty can never be eliminated, it can be brought to heel.

[22] Of course, this is no more than an idealization. In practice, policy decisions involve reconciling a number of different, often-conflicting, objectives, and there is no algorithm by means of which all can be optimized. The relatively recent discipline of multiple-criteria decision-making seeks to optimize, if not the criteria, then the trade-offs between them (see Deb, 2013).

[23] Royall (1997) was perhaps the first statistician to distinguish sharply between confirmation and evidence questions in just the way that we do here. One of the anonymous referees of this paper has reminded us that the confirmation question is normative – what *should* an agent believe? – while the second, evidential, question asks the merely *descriptive* question – in what conditions do data provide evidence for a hypothesis. This distinction is important; what "should" be believed brings with it the presupposition that the agent is *rational*, and this presupposition, in turn, constrains the limits of belief, imposing a measure of "objectivity" on them. It is certainly more plausible to contend that D provide evidence for H just in case they bolster *rational* belief that it is true. The immediate difficulty with this sort of attempt to square confirmation with evidence is that it eventually requires imposing such strong constraints that all fully rational agents will assign the same prior probabilities at the outset of their inferences given that they share the same background information (see Williamson, 2005, pp. 11–12). There are several problems with the "unique probability constraint" (see Bandyopadhyay and Brittan, 2010), perhaps the most important of which is that "sharing the same background information" is vague if not also question-begging, an unhelpful proxy for objectivity. In part for this reason, traditional Bayesians make their case for objectivity not on the constraining of priors but the convergence of posterior probabilities. That convergence is not a sufficient condition of unbiased objectivity will be demonstrated later.

on one hypothesis than on another. The greater the likelihood ratio, the stronger the evidence. In contrast, hypotheses are confirmed one at a time as the probability of (belief in) their truth is raised (strengthened).

The same idea, that "confirmation" and "evidence" vary conceptually, is perhaps best illustrated by "crucial experiments." Such experiments discriminate one equally-well confirmed hypothesis from another and at the same time provide *evidence* for one as against the other. Although Darwin's explanation of evolution by way of natural selection had been generally accepted by that time, it remained an open question in the early 1940's whether mutations among bacteria occur as either an adaptive response to an environmental stimulus (an instance of the Lamarckian theory of the heritability of acquired characteristics) or randomly (in which case they are transmitted to the next generation as a function of reproductive fitness). Both theories had their defenders. In what is arguably the most famous single experiment in the history of ecology/evolutionary biology, Luria and Delbruck devised a way to test the two hypotheses (Luria and Delbruck, 1943). They exposed a number of parallel cultures to viruses known as phages, "subcellular parasites that infest, multiply within, and kill bacteria." On the Lamarckian theory, bacteria adapt to their phage environment; hence, the number of mutations that occur should be both relatively small and constant across bacterial cultures. The Darwinian theory, that phage-resistant mutations occur randomly and prior to exposure entails that the number of phage-resistant mutations should vary dramatically from one culture to the next, and since available earlier in the process, the mutations accumulate much more rapidly where they occur in lines before phage exposure.

Slightly more precisely, Delbruck and Luria hypothesized that if phage-resistant mutations occurred after exposure, the number of survivors would approximate a Poisson distribution, on which the mean would equal the variance. What they found was that the variance was much greater than the mean. They then drew the Darwinian conclusion (as did the rest of the biological community) that bacterial mutations are indeed random, as are macro-organism mutations, rather than post-adaptive or "directed."

EVIDENTIAL STATISTICS[24]

This is what evidence does, allows us to discriminate in a straightforwardly objective way between hypotheses that may be otherwise equally well-confirmed. As we have just seen, there are cases in which there is strong evidence for one of a pair of equally well-confirmed hypotheses. There are also cases in which there is no such evidence[25], cases in which the evidence is strong and the degree of confirmation low, and so on.

We can make this account of evidence more precise. It involves the comparison of the merits of two models, M_1 and M_2 (possibly, but not generally $\sim M_1$) relative to the data D and background information B.

- D is evidence for M_1as against M_2just in case $\Pr(D| M_1) > \Pr(D|M_2)$[26].

This is often called the Likelihoodist (LR) account of evidence. It follows at once that "data" are to be distinguished from "evidence." Data constitute evidence only with respect to models in a well-defined comparative context[27]. To put this more precisely, evidence is a data-based estimate of the relative discrepancy of two models to the generating process. In the case of simple models, the log-likelihood happens to be an estimate (up to a constant) of the Kullback-Leibler discrepancy between the generating process and a model. As with the original confirmation account, this formulation is qualitative. A commonly-used measure of the degree of evidence vis-à-vis a model comparison is the numerical ratio of the likelihoods[28]. Note in this connection that if $1 < \text{LR} \leq 8$, then D is often held to provide "weak" evidence for M_1as against M_2, while when LR > 8, D provides "strong" evidence for M_1as against M_2. Note also the shift from talking about "hypotheses" and "theories" to talking about "models." *Hypotheses* are often formulated in verbal rather than mathematical terms and they rarely provide potential and predictive data-distributions. They can be either true or false. *Models* are mathematically-formulated and idealized data-generating mechanisms. They can generate data-distributions near or far from what might be termed the "naturally" generated distributions. Observed data or experimental results support model$_1$ over model$_2$ if the potential/predictive data generated by the first are by some agreed-upon measure closer to the observed data than the potential data generated by the second. Differences of information criteria are estimates of KL discrepancy differences that adjust for biases caused by estimation.

Although the terms have been used rather carelessly to this point, "hypothesis" is more helpfully associated with Bayesian inference, "model" with Evidential testing, leaving to the side any questions concerning how either relates to "theory" (which intuitively includes both hypotheses and models as components, links them by way of explanatory principles and basic concepts, and affords successful prediction of data in a wide variety of sub-disciplines; *force* is such a basic concept and the laws of motion the explanatory principles in classical physics, *natural selection*,

[24] We use "the likelihood-ratio account of evidence, "evidential statistics," and "Likelihoodism" somewhat interchangeably. It is important to note that the likelihood ratio is only an important special case of a more general class of measures that constitute the core of evidential statistics. See Lele (2004) on the "efficiency" of this particular evidential function.

[25] See Rosenzweig's (1936).

[26] Equivalently, the ratio of the two likelihoods is >1.

[27] The quantity $\Pr(D|M)$ is usually referred to in the philosophical literature as a "likelihood." But while numerically the likelihood of the model given the data is proportional to the probability of the data given the model, likelihood and probability differ conceptually; the likelihood is considered a function of the model, whereas the probability is considered a function of the data. The common philosophical notation of $\Pr(D|M)$ rather than the common statistical notation of $L(M;D)$ is adopted here, but is not meant to imply that the model M needs to be considered a random variable.

[28] Or the logarithm of the likelihood ratio. Nothing in the present discussion turns on the difference; the respective ordinal structures remain the same.

and *adaptation* the basic concepts and explanatory principles in Darwinian biology)[29].

Evidence and confirmation thus characterized differ in a number of important respects, some of which have already been mentioned. Data can provide a high or low degree of confirmation to a hypothesis while at the same time providing weak or strong evidence for it in a comparative context. This is to say that while D confirms H if and only if D constitutes evidence for M with respect to $\sim M$[30], there is no linear relationship between their respective degrees[31]. Moreover, given the way in which they are quantified, degrees of confirmation can vary between 0 and 1 exclusive, while numerical values of evidence on a likelihood ratio can range from 0 to ∞ inclusive (or -∞ to ∞ in the case of the log-likelihood ratio).

CONFIRMATION, EVIDENCE, AND THE ANTHROPOGENIC CLIMATE-CHANGE HYPOTHESIS

The distinction between confirmation and evidence is indispensable for both theory and practice, and not often made. But it also bears directly on the public understanding of the way in which science informs policy formation. Simply put, controversy with sweeping social, political, and economic consequences sometimes arises, to one extent or another, from failure to draw it.

A sample controversy has to do with the anthropogenic "global warming" hypothesis, that is, the hypothesis that present warming trends are human-induced. If it does not raise questions concerning foundational physical theories, it does with respect to their application[32].

A wide spectrum of data raises the posterior probability of the hypothesis, in which case they confirm it. Indeed, in the view of most climatologists, this probability is very high. The Intergovernmental Panel on Climate Change contends that most of the observed temperature increase since the middle of the twentieth century has been caused by increasing concentrations of greenhouse gases resulting from human activity such as fossil fuel burning and deforestation. In part this is because the reasonable prior probability that global warming is human-induced is very high. It is assigned not on the basis of relative frequencies so much as on the explanatory power of the models linking human activity to the "greenhouse effect," and thence to rising temperatures. In part, the posterior probability of the hypothesis is even higher because there are so many strong correlations in the data. Not only is there a strong hypothesized mechanism for relating greenhouse gases to global warming, this mechanism has been validated in detail by physical chemistry experiments on a micro scale, and as already indicated there is a manifold correlation history between estimated CO_2 levels and estimated global temperatures. Of course, some climate skeptics emphasize how difficult it is to get standardized and reliable data for such a long period of time and from so many different places, others point out that it has not always been true that changes in CO_2 levels precede changes in temperature. But the main skeptical lines of argument are that (a) the likelihood of the data on the alternative default (certainly simpler) hypothesis, that past and present warming is part of an otherwise "natural" and long-term trend, and therefore not "anthropogenic" is just as great, (b) that the data are at least as likely on other, very different hypotheses, among which solar radiation and volcanic eruption, (c) that not enough alternative hypotheses have been considered to account for the data. That is, among credible climate skeptics there is some willingness to concede that burning fossil fuels leads to CO_2 accumulation in the atmosphere and that carbon dioxide is a greenhouse gas that traps heat before it can escape into the atmosphere, and that there are some data correlating a rise in surface temperatures with CO_2 accumulation. But, the skeptics continue, these correlations do not "support," let alone "prove," the anthropogenic hypothesis because they can be equally well accounted for on the default, "natural variation" hypothesis or by some specific alternative. Since there is very little evidence for the hypothesis, it is not, the skeptics conclude, very well confirmed (and for this and other reasons massive efforts to reduce carbon emissions are a costly mistake). But this conclusion rests on a conflation of evidence with confirmation, and provides a striking reason why it is necessary to distinguish the two.

Data are evidentially relevant only if they discriminate hypotheses, and such data in the case of human-induced warming have been difficult to come by. That fact has premised at least part of the skeptics' argument that the rise in atmospheric CO_2 comes from, e.g., the ocean, and is therefore "natural," at the very least as likely a cause of the greenhouse gases responsible for temperature rise as the human-induced explanation. Such data have, however, been identified increasingly[33]. For example, most carbon atoms have an atomic mass of 12, but about 1% has an atomic mass of 13. Both kinds can form CO_2 molecules, $^{12}CO_2$ and $^{13}CO_2$, distinguishable in the laboratory. To put a complex story very simply, it can be shown that if the CO_2 atmosphere comes from the surface (and not the depths) of the ocean, then $^{13}CO_2$ will increase over time. If the CO_2 comes from fossil

[29] Some readers of this paper may be disappointed by the lack of precision in this definition of "theory." They should look at Marquet et al. (2014): "In ecology, there is generally no consensus regarding the definition, role, and generality of theories….A summary of the ecological literature finds reference to 78 theories."

[30] I.e., it is provable that $\Pr(M|D) > \Pr(M)$ just in case $[\Pr(D|M)/\Pr(D| \sim M)] > 1$, i.e., when the two models are mutually exclusive and jointly exhaustive. Models, unlike hypotheses, don't often have negations, only alternatives. In this respect it differs from Bayesian-testing, which presupposes an implicit comparison between a hypothesis and its negation only. It is in this sense that evidence-testing allows for genuinely multiple models, Bayesian-testing does not.

[31] One of the reviewers has corrected the original formulation of this claim, and has also urged us to make clear that the claim presupposes the difference measure of confirmation we have taken as our model. It should be added that while numerical similarities/dissimilarities between the proposed measures of confirmation and evidence vary with the way in which each is characterized, the ways in which the probability operators in each are interpreted—in terms of beliefs or bets in the case of confirmation, in terms of formal relations in the case of evidence—force a conceptual distinction between them, as does the ability to unravel such heretofore intractable problems in the foundations of statistics as the notorious "paradoxes of confirmation" (see Bandyopadhyay et al., 2016, Chapter 9) or to clarify one source of public policy controversies.

[32] The following two paragraphs are drawn from Bandyopadhyay et al. (2016, pp. 40–44). References documenting the empirical claims made can be found there.

[33] What follows draws on the very accessible overview by Farley (2008).

fuel burning, then the relative abundance of $^{13}CO_2$ to $^{12}CO_2$ will decrease. Experimental results show that the $^{13}CO_2/^{12}CO_2$ ratio is decreasing, evidence for the hypothesis that fossil fuels rather than surface water is mainly responsible for rising levels of CO_2 in the atmosphere, and hence (on the assumption that rising levels of CO_2 are a cause of rising temperatures) for the anthropogenic hypothesis.

BAYESIAN OBJECTIVITY

Two crucial differences between confirmation and evidence have been alluded to but must be underlined. First, confirmation is psychological in character, involving as it does changes in an agent's personal degree of belief that a hypothesis is true. Evidence is logical in character, an agent-independent relationship between models and data[34]. It follows not only that confirmation is "kinematic," beliefs re-adjusted over time as data are accumulated, evidence "static," an atemporal as well as impersonal relationship between models and data, but also that the probability operators in their respective accounts are not to be interpreted in the same way. Confirmation tracks changes in belief and thus degrees of uncertainty in an agent's mind. Evidence has to do, rather, and as already noted, with a logical relation. The former probabilities are psychological and in this sense "subjective," the latter formal and for this reason "objective."

It would seem to follow that since the credibility of their claims depends on the extent to which they are objective, the method of model statistics should be preferred by practicing ecologists. But this is not the end of the matter. On the one hand, traditional, i.e., self-described "subjective Bayesians" make a case for the objectivity of their method of testing hypotheses. On the other, so-called "objective Bayesians" both curb the source of subjectivity in applications of Bayes Theorem and play down if not also discount completely the subjective/objective distinction as an unwanted philosophical distraction. Since both approaches are increasingly popular, each must be examined.

Confirmation and Convergence

Bayesian inferential techniques inform decision-making processes. They do so by way of the fact that decisions are to be explained in part in terms of their beliefs and desires. This is to imply that whether the decisions themselves are good or bad is agent-dependent. It is but a short natural if not also logical step to conclude that they are all, even research decisions, "biased." One much-discussed example is "confirmation bias," focusing one's efforts on finding data that confirm one's beliefs and thus potentially misrepresenting what is in fact the case[35]. Meta-studies of reported ecological claims provide some support for this conclusion[36], and of course it is a source of at least some of the public's resistance to take them seriously. In the case of Bayesian inference, the charge stems directly from the role that prior probabilities play. Such probabilities are "subjective" in the straightforward sense already indicated.

Traditional Bayesians contend that it is demonstrable (Walker, 1969) that the influence of priors "wash out" over time, in which case the inference is ultimately "objective." So long as certain conditions (event-exchangeability and the like) hold and assumptions (concerning parameter identifiability and the omission of idiosyncratic priors) are made, the beliefs of different agents, no matter how unlike at the outset, will eventually converge to the maximum likelihood solution as data accumulate. What is not often appreciated is that the rate of convergence (or whether it occurs at all) depends both on the nature of the data and of the models. Unfortunately, the real world of science is not always asymptotic. Data cost money and take time to acquire. So while Bayesian inference and maximum likelihood may often agree, sometimes in real world analyses they do not – with practical consequences (Lele, under review; see also the further discussion of this point below).

Further, the idea that Bayesian convergence is tantamount to objectivity in an adequately strong sense of the word is misleading. On the one hand, "objectivity" is here equated with "inter- subjective agreement," which is to say that for all of the consensus involved, the probability is not agent-independent[37]. Invariance is not to be confused with independence. However, much the prior probabilities might "wash out" numerically in the calculation of posterior probabilities via Bayes Theorem, they must still include reference to them in principle. The reference in principle is crucial not because it influences the calculation, but because it embeds stochasticity in the head as "uncertainty," and not in the world. A fully objective scientific inference draws conclusions about the way the world is, and not about the way in which consensus, however general, has been reached.

On the other hand, the asymptotic intuition embedded in the notion of Bayesian convergence again leads naturally if not also logically to the conclusion that common agreement about the way things stand in the world is tantamount to truth. But this optimistic suggestion is hostage to the history of science. Commitment to the belief that in their inter-dependency, self-regulation, and complexity, undisturbed biotic communities evolve in the direction of greater complexity was "settled science" among ecologists for well-over a 100 years. Only relatively recently has it been more and more challenged. Convergence of belief doesn't entail its truth. Confidence may be raised, even to the point of near-certainty, when it is in hindsight

[34] To avoid misunderstanding, the choice of models to test is not agent-independent, only the formal relationship between the models tested and the data-distributions in which they issue. Both Bayesianism and Evidential Statistics are "rationalist" or "top-down" in that they begin with hypotheses and models and then proceed to gather data, not the other way around. In this respect, both are to be sharply distinguished from frequentist approaches which begin with correlations in the data gathered. In that Mayo begins with simple statistical hypotheses, her approach (Mayo, 2018, p. 85) is in a related sense "bottoms-up."

[35] See Kahneman (2011, p. 81): "Contrary to the rules of philosophers [or at least of Karl Popper], who advise testing hypotheses by trying to refute them, people (and scientists quite often) seek data that are likely to be compatible with the beliefs they currently hold."

[36] For documentation of such bias see Fanelli (2010) and Holman et al. (2015).

[37] Nor, for that matter, independent of the many pressures brought to bear on the up-dating of beliefs by disciplinary communities (in the person of editorial staffs and funding agencies).

unwarranted. This is why the rote response to those who question anthropogenic global warming–"it is the consensus of experts"—is far from conclusive and to much of the public unconvincing.

Non-informative Priors and Invariance

There is a 2-fold option for the increasing number of ecologists who find it more computationally convenient to use Bayesian up-dating techniques to analyze multi-layered/factor hierarchical or space-state model of complex data, but who are uneasy about the apparent subjectivity of prior probabilities in the inferences they make. This option is set out in a very lucid and thought-provoking way by Clark in his widely-cited paper, "Why environmental scientists are becoming Bayesians" (Clark, 2005). It consists of placing a constraint on allowable priors and easing the tension sometimes induced when metaphysics and method are mixed.

A variety of constraints on priors have been proposed. Most of them are epistemic in character. They range from total knowledge on the part of the up-dating agent to total ignorance, some version of applying the Principle of Indifference to the choice of priors. Although both have a long history, it is not entirely clear how each is to be made precise (see Bandyopadhyay and Brittan, 2010). Clark opts for the latter—a flat or non-informational constraint. It is mathematically-convenient to do so. Moreover, it ensures agent-independency in this sense, that the agent is assumed to know nothing about the hypothesis at hand at the outset of his or her inferential activities; no prior beliefs are presupposed (in which case, at least in principle, "the data are allowed to speak for themselves"). But it also harbors problems, several of which are set out by Lele (*Frontiers of Ecology and Evolution*, this issue), and illustrated by case studies of the survival of the kit fox and declines in amphibian populations.

Since it is immediately available to the reader, there is little point in rehashing the rather technical paper here. Suffice it to say that Lele draws several unintended but important consequences from the long-known fact (see Fisher, 1930) that flat parameters are not invariant under transformation. For our purposes, two are particularly important. The first is that in a sample viability analysis, the population prediction interval (PPI) obtained by maximum likelihood ratios (MLR) under two parameterizations of the data are similar, while those obtained by non-informative priors differ from each other and from the MLR PPI. Despite what Clark says (Clark, 2005, pp. 3 and 5), Bayesian inferences based on flat priors do not lead to the same (numerical) conclusions as likelihood-based inferences on the same data.

The second consequence (Clark, 2005, lines 258–259) is that "different versions of the non-informative priors on the natural parameters induce different priors (and hence biases) on the induced parameters of scientific interest." Simply put, the fact that flat priors are not invariant under transformation can be used to demonstrate that while on occasion Bayesian inferences resemble likelihood-based inferences and appear bias-free, on closer examination and other occasions this is not the case.

Although Clark admits (Clark, 2005, p. 4) that "the importance of philosophy should not be understated," the "focus" of his paper is that "the emergence of modern [viz., objective hierarchical] Bayes has little to do with philosophy, but rather comes from pragmatism." But as Lele makes clear in some detail, Clark's failure to take philosophical questions concerning the concept of objectivity more seriously led him to ignore the problematic character of his answers to them, and opens up legal and legislative challenges to flat-prior ecological inferences which lack the requisite invariance under transformation and parameterization.

Computation and Cloning[38]

Modern Bayesianism is ostensibly (but problematically) superior to its unacceptably "subjective" original by way of restricting allowable priors. It is often held to be similarly superior to Likelihoodism in its apparently unique ability to compute the likelihood function in complex statistical inferences from and to hierarchical models. These models are very useful, indeed indispensable, in understanding the processes underlying complex ecological data. As Ponciano et al. (2009, p. 356) put it, "computing the likelihood function needed for such inferences requires an intractable, high-dimensional integral. [But] inferences using computer intensive Bayesian methods sidestep this difficulty by simulating observations from a prior distribution using one of the various Markov chain Monte Carlo algorithms." This surmounting of very genuine computational problems is undoubtedly an important factor in the growing popularity of these Bayesian methods.

Lele et al. (2007, 2010) recognized that the Bayesian computational methods could be coopted to calculate fully frequentist maximum likelihood estimates and their standard errors using an approach called data cloning. Ponciano et al. (2009), developed an extension to data cloning (the data cloned likelihood ratio or DCLR) that in a similar way affords the calculation of likelihood ratios or the differences of information criterion values. These are the fundamental tools of evidence, and hence of evidence comparing hierarchical models.

Thus, the computational advantage enjoyed by Bayesian methods is no more than apparent. If one assumes that statistical paradigms should be (mainly) compared computationally and conceptually, and if (at least in the wake of Ponciano et al., 2009 and also Lele et al., 2007) there is nothing (basically) to choose between the Bayesian and Likelihood paradigms computationally, then the difference is conceptual, and in this sense "philosophical." The announcement of philosophy's irrelevancy by Clark and others was premature.

COGNITIVE BIASES AND THE METHOD OF MULTIPLE MODELS

It needs to be made clear that convergence *per se* is a demonstrable consequence of the Likelihood account of evidence. Indeed on any adequate statistical paradigm, inferences should improve as more model-relevant data are analyzed[39].

[38] Another referee helpfully asked for a brief comment on cloning.

[39] Because model misspecification is allowed in an evidential framework, data-model consistency is not identical with classical consistency. Of course, if the generating process is actually a model in the model set, it should be asymptotically identified. Under model misspecification, however, misleading and weak evidence still both need to go to zero as sample size increases to infinity. Asymptotically the model selected should be the model in the model set closest to the generating process.

But there is an underlying problem confronting Bayesian convergence. It has been called "availability bias." Bayesian model identification converges to truth only if the "true" model is in the set of hypotheses under consideration[40]. As the statistician Barnard (1949) once wrote:

> To speak of the probability of a hypothesis implies the possibility of an exhaustive enumeration of all possible hypotheses, which implies a degree of rigidity foreign to the true scientific spirit. We should always admit the possibility that the experimental results may best be accounted for by a hypothesis which never entered our heads.

In fact, there are two problems here. The more general is that Bayesian convergence assumes that all investigators start out with the same model set, however at variance their initial degrees of uncertainty with respect to its members' truth. The more specifically Bayesian problem is that it makes little sense to assign prior probabilities to members of the model set unless that set is assumed closed. Both problems result from the "availability bias."

But this bias, as also the "confirmation bias" mentioned earlier, is eliminated with the introduction of multiple models required by the Likelihood account of evidence. First, models on this account are pairwise compared, without assigning prior probabilities to any of them, i.e., without incorporating a subjective bias in the testing procedure. Second, data constitute evidence only in a contextual way, conditional on the two models compared. The most one can say is that one or the other is better supported, not that it is closer to the truth. The challenge is to find other models and new data against which models to compare it. Bayesian convergence does not allow for the heretofore unimagined, either with respect to the initial model set or heretofore unrealized conditions. As ecologists know perhaps too well, new models at different levels of organization are introduced all the time as explanatory insights emerge and ecosystems change[41].

A striking example is provided by research on stress-induced mutation. As Foster (2007) puts it,

> …after 20 years of research, evidence now suggests that various types of stresses induce responses that have mutagenic consequences, and that sometimes this essentially random process can appear to be directed…

Change, not stasis, is the rule. In this case what has emerged is a model on which mutations are generated even prior to the operation of Darwinian and Lamarckian selection pressures and appears consistent with both. This is to say that what we early took as an exemplary "crucial experiment," viz., Delbruck and Luria's analysis of phage-resistant bacteria cultures, was not[42]. As Turkey said in a memorable paper (Tukey, 1960, p. 425), "Conclusions are established with careful regard to the evidence…[but] accepted subject to future *rejection*.…[They are] taken to be of lasting value, but not everlasting value."

Darwinian Objectivity

Given that in its present conceptual-methodological state ecology generally considered appears so unsettled[43], it might be asked whether it really is a *science*, and not an area studies program, grouping together a number of rather different investigative activities under the general heading of "organisms and the environment." The emphasis on "integration" or "synthesis" in some of the textbooks mentioned suggests an urgent need to find, or impose, a common core. The traditional way of understanding the question, "is it really a science and, if so, in what respects?" viz., "how closely does it resemble classical physics in its general aims and methods?" has rightly been rejected. No one any longer thinks that philosophers can determine in a more or less *a priori* fashion what the "right" concept of science is, or pretend that there is one (and only one) method of implementing it, or that all scientific *laws* must take the form of universal conditionals. But there is more to be said.

In a broad perspective, theoretical and conceptual clarification in ecology continue; their integration remains a somewhat distant goal. The possibility of methodological progress is nearer at hand. We have illustrated such progress in the case of hypothesis and model testing. There are at present two particularly plausible accounts, Bayesian confirmation and the Likelihoodist account of evidence. Their integration depends less on their unification than on assigning them their proper roles. Choosing ecological models to test and formulating/implementing environmental policies are (like betting) actions; they are to be explained or justified (as a normative Bayesian does) in terms of a (rational) agent's beliefs and desires. Beliefs and desires in turn are traditionally and, we think, most plausibly to be understood in personal probabilistic terms; some beliefs are more certain than others, some desires stronger. It follows that one should use Bayesian methods when the question is: what should I *do*? The resulting answer concerns the extent to which particular beliefs have been fortified in the process of up-dating (up-data-ing) them. Likelihood ratios, on the other hand, are agent independent. The probabilities embedded in them have nothing to do with either beliefs or desires, but with logical relations between sentences describing data and articulating models. Such ratios answer the question: which model (among those compared) is *better supported* by the data?

[40] That the "true model" is assumed to be in the model set follows from the fact that the prior probabilities of the hypotheses considered must sum to 1. See Lindley (2001) for an attempt to avoid the problem by pointing out that Bayesian inference is always conditional on a set of models and "convergence" understood as relative to it. To relativize convergence in this way, however, is to relativize "true model," and with it the "objectivity" that Bayesian convergence is intended to ensure.

[41] Chamberlain (1897), Platt (1964), and Burnham and Anderson (2002) among others have understood the virtues of multiple models. On the LR account of testing, evidence has real bite only when it serves to distinguish between them. Human-caused and ocean-temperature caused global warming are not simply mutually-exclusive and (we assume for present purposes) jointly-exhaustive alternatives; stronger or weaker evidence for and against them can be gathered in a genuinely comparative context. Apparently such a context has not yet been developed for the deforestation hypotheses mentioned above.

[42] See Cairns et al. (1988) for the initial clue re stress-induced mutation, and Houston (2014) for a multi-model approach to re-thinking the rejection of at least two classic population equilibrium hypotheses.

[43] Some would say it is in a state of crisis, despite all of the enormously illuminating work done over the last several generations in its many sub-disciplines, island-biogeography among them.

Insofar as they measure uncertainty, Bayesian inferences lead to irreducibly personal conclusions, however great the agreement respecting these conclusions proves to be over time among different people. The LR account of evidence compares models with their alternatives, and each accumulates support or not as the predictions to which they lead are verified. Since ecological models are for the most part stochastic, so too are the events and processes that, in a clear sense, they objectively represent. Greater methodological self-consciousness about the methods they use to test hypotheses/models should provide helpful guidance to ecological scientists and, in identifying (at least in general terms) the sources of their objectivity, make the policy recommendations of individual wildlife and wildland managers more credible with the general public[44].

Our emphasis on the L-R account of evidence is not new, but it has yet to gain much ground in ecology. As a recent paper by Betini et al. (2017) discovered, "only 21 of 100 randomly selected studies from the ecological and evolutionary literature tested more than one hypothesis, only eight tested more than two hypotheses." Yet as we have argued, it is only insofar as multiple models are pairwise compared vis-à-vis the data and in this way tested that the main forms of cognitive bias can be ruled out.

Two final notes. One is that both Bayesian confirmation and Likelihood evidence rely importantly, although not completely[45], on the predictions to which hypotheses and models lead. Prediction in ecology is very difficult[46]. Humans are an integral part of many if not all of the ecological systems they study and their research and policy interventions alter them in the process of such study. As a result, a majority of ecologists still fall back on falsification and simplified versions of significance testing. It very much needs more methodological attention.

The other note is this. The Likelihood account leads to convergence, as do all good parameter estimators, in the sense that as favorable data continue to be accumulated, the evidence for particular models becomes stronger and stronger. However, there is not convergence toward a "final theory," or even assumed that the "true model" is among those already under consideration. New explanations of events and instruments of their prediction may from 1 min to the next be discovered or invented. As in the case of Darwin's theory of natural selection, progress is measured in terms of the survival of mutations in the face of environmental pressures. From this point of view, testing methods do not result in approximations to some stipulated goal but are measures of survival value. What we term "Darwinian objectivity" presupposes competition between accounts, whether of ecological phenomena or appropriate methodologies.

In this deep way, the Likelihood account is well-suited to ecology. On it, models are never more or less true but epistemic stages in the course of evolution, contingent on conditions which are themselves subject to continuing change, and on the intervention in the biotic and abiotic environments of human beings whose behavior is itself conditioned on the success or failure of the models they test. Stochasticity and survival are fundamental dimensions of natural processes. So too are they features of any adaptive, objective and self-conscious account of model testing, and therefore of scientific method generally.

AUTHOR CONTRIBUTIONS

GB and PB originally conceived the main the theme of the paper. GB completed the first draft. Both authors contributed to the revisions of the manuscript and approved the final version for submission.

ACKNOWLEDGMENTS

Two of the editors of this issue of *Frontiers in Ecology and Evolution*, Mark L. Taper and José M. Ponciano, have made some very useful suggestions regarding the paper. The three referees for this issue of *Frontiers in Ecology and Evolution* made a number of helpful comments on our original submission, all of which we have tried to address in our revisions.

[44] See Maunder and Piner (2015): "Bayesian analysis accommodates the use of prior information in integrated assessments, allowing sharing of information from other species. It also allows for the representation of uncertainty in a probabilistic context, which is ideal for decision analysis." In this way Maunder and Piner take it as supplementing Likelihood testing which is widely used in fisheries management.

[45] The extent to which they explain and aid understanding of ecological events and processes also factors into their evaluation.

[46] See Dietze (2017) and Maris (2018).

REFERENCES

Anderson, D., Burnham, K., and Thompson, W. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manag.* 64, 912–923. doi: 10.2307/3803199

Bandyopadhyay, P., and Brittan, G. (2010). Two dogmas of strong objective Bayesianism. *Int. Stud. Philos. Sci.* 24, 45–65. doi: 10.1080/02698590903467119

Bandyopadhyay, P., Brittan, G., and Taper, M. (2016). *Belief, Evidence, and Uncertainty*. New York, NY: Springer.

Barnard, G. (1949). Statistical inference. *J. R. Stat. Soc. Ser. B* 11, 115–149. doi: 10.1111/j.2517-6161.1949.tb00028.x

Betini, G., Avgar, T., and Fryxell, J. (2017). Why are we not evaluating multiple competing hypotheses in ecology and evolution? *R. Soc. Open Sci*. 4:160756. doi: 10.1098/rsos.160756

Botkin, D. (1990). *Discordant Harmonies: A New Ecology for the Twenty-First Century*. New York, NY: Oxford University Press.

Burnham, K., and Anderson, D. (2002). *Model Selection and Multi-Model Information: A Practical Information-Theoretic Approach, 2nd Edn*. New York, NY: Springer.

Cairns, J., Overbaugh, J., and Miller, S. (1988). The origin of mutants. *Nature (London)* 335, 142–145. doi: 10.1038/335142a0

Cassey, P., and Blackburn, T. (2006). Reproducibility and repeatability in ecology. *Bioscience* 56, 958–959. doi: 10.1641/0006-3568(2006)56[958:RARIE]2.0.CO;2

Chamberlain, T. (1897). The method of multiple working hypotheses. *J. Geol.* 5, 837–848. doi: 10.1086/607980

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. Ser A* 158, 419–466. doi: 10.2307/2983440

Clark, J. (2005). Why environmental scientists are becoming Bayesians. *Ecol. Lett.* 8, 2–14. doi: 10.1111/j.1461-0248.2004.00702.x

Da Silva, J., Steiner, A., and Schreiner, E. (2018, October 3). Forests: a natural solution to climate change, crucial for a sustainable future. *United Nations Development Programme*.

Deb, K. (2013). An evolutionary based Bayesian design optimization approach under incomplete information. *Eng. Optim.* 45, 151–165. doi: 10.1080/0305215X.2012.661730

Dennis, B., Poinciano, J., Taper, M., and Lele, S (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Evol.* doi: 10.3389/fevo.2019.00372

Dietze, M. (2017). Prediction in ecology: a first principles framework. *Ecol. Appl.* 27, 2048–2070. doi: 10.1002/eap.1589

Duhem, P. (1962). *The Aim and Structure of Physical Theory*, ed P. Weiner. New York, NY: Atheneum.

Ellison, A. (1986). An introduction to Bayesian inference for ecological research and decision-making. *Ecol. Appl.* 64, 1036–1046.

Ellison, A. (2004). Bayesian inference in ecology. *Ecol. Lett.* 7, 509–520. doi: 10.1111/j.1461-0248.2004.00603.x

Errington, P. (1945). Some contributions of a fifteen-year local study of the northern bob-white to a knowledge of population phenomena. *Ecol. Monogr.* 15, 1–34. doi: 10.2307/1943293

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE* 5:e10271. doi: 10.1371/journal.pone.0010271

Farley, J. (2008). The scientific case for modern anthropogenic global warming. *Monthly Rev.* 60. doi: 10.14452/MR-060-03-2008-07_5

Fisher, R. (1929). The statistical method in psychical research. *Proc. Soc. Psych. Res.* 39, 189–192.

Fisher, R. (1930). Inverse probability. *Proc. Camb. Philos. Soc.* 26, 528–535.

Ford, D. (2000). *Scientific Method for Ecological Research*. Cambridge: Cambridge University Press.

Foster, P. (2007). Stress-induced mutagenesis in bacteria. *Crit. Rev. Mol. Biol.* 42, 373–397. doi: 10.1080/10409230701648494

Hempel, C. (1965). *Studies in the Logic of Confirmation. Aspects of Scientific Explanation*. New York, NY: Free Press.

Holman, L., Head, M., Lanfear, R., and Jennions, M. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol.* 13:e1002190. doi: 10.1371/journal.pbio.1002190

Houston, M. (2014). Disturbance, productivity, and species diversity: empiricism vs. logic in ecology theory. *Ecology* 9, 2382–2396. doi: 10.1890/13-1397.1

Kahneman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Strauss, and Giroux.

Kidwell, S., and Holland, S. (2002). The quality of the fossil record: implications for evolutionary analyses. *Annu. Rev. Ecol. Syst.* 33, 561–588. doi: 10.1146/annurev.ecolsys.33.030602.152151

Läärä, E. (2009). Statistics: reasoning on uncertainty, and the insignificance of testing null. *Ann. Zool. Fennici* 46, 138–157. doi: 10.5735/086.046.0206

Lele, S. (2004). "Evidence function and the optimality of the law of likelihood," in *The Nature of Scientific Evidence*, eds M. Taper and S. Lele. Chicago, IL: University of Chicago Press.

Lele, S., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov Chain Mlonte Carlo methods. *Ecol. Lett.* 10, 551–563. doi: 10.1111/j.1461-0248.2007.01047.x

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *J. Am. Stat. Assoc.* 105, 1617–1625. doi: 10.1198/jasa.2010.tm09757

Lindley, D. (2001). The philosophy of statistics. *J. R. Stat. Soc.* 49, 293–337. doi: 10.1111/1467-9884.00238

Low-Décarie, E., Chivers, C., and Grenados, M. (2014). Rising complexity and falling explanatory power in ecology. *Front. Ecol. Environ.* 12, 412–418. doi: 10.1890/130230

Luria, S., and Delbruck, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491.

Maris, V. (2018). Prediction in ecology: promises, obstacles, and clarifications. *Oikos* 127, 171–183. doi: 10.1111/oik.04655

Marquet, P., Allen, A. P., Brown, J. M., Dunne, J. A., Enquist, B. J., Gillooly, J. M., et al. (2014). On theory in ecology. *Bioscience* 64, 701–710. doi: 10.1093/biosci/biu098

Maunder, M., and Piner, K. (2015). Contemporary fisheries stock assessment: many issues still remain. *ICES J. Mar. Sci.* 72, 7–18. doi: 10.1093/icesjms/fsu015

Mayo, D. (1996). *Error and the Growth of Knowledge*. Chicago, IL: University of Chicago Press.

Mayo, D. (2018). *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

Mayo, D., and Spanos, A. (2004). Methodology in practice: statistical misspecification testing. *Philos. Sci.* 71, 1007–1025.

Mayo, D., and Spanos, A. (2010). *Error and Inference*. Cambridge: University of Cambridge Press.

Murtaugh, P. (2014). In defense of *P* values. *Ecology* 95, 611–617. doi: 10.1890/13-0590.1

Pearce, F. (2008, October 9). Conflicting data: how fast is the world losing its forests? *Yale Environment 360*.

Pearson, E. (1938). Student vs. statistician. *Biometrica* 30, 210–250.

Pickett, S., Kolesa, J., and Jones, C. (2007). *Ecological Understanding: The Nature of Theory and the Theory of Nature, 2nd Edn*. Amsterdam: Elsevier.

Platt, J. (1964). Strong inference. *Science* 146, 347–353. doi: 10.1126/science.146.3642.347

Ponciano, J., Taper, M., Dennis, B., and Lele, S. (2009). Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology* 90, 356–362. doi: 10.1890/08-0967.1

Popper, K. (1934/1959). *Logik der Forschung. The Logic of Scientific Discovery.* London: Hutchinson.

Popper, K. (1974). *Intellectual Autobiography*. *The Philosophy of Karl Popper*, ed P. Schilpp. LaSalle, IL: Open Court.

Romesburg, C. (1981). Wildlife science: gaining reliable knowledge. *J. Wildl. Manag.* 45, 293–313. doi: 10.2307/3807913

Rosenzweig, S. (1936). Some implicit factors in diverse methods of psychotherapy. *Am. J. Orthopsychiatry* 6, 412–415. doi: 10.1111/j.1939-0025.1936.tb05248.x

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall.

Saint-Mont, U. (2018). Where Fisher, Neyman and Pearson went astray: on the logic (plus some history and philosophy) of statistical tests. *Adv. Soc. Sci. Res.* 5, 672–691. doi: 10.14738/assrj.58.4867

Schmitz, O. (2017). *The New Ecology: Rethinking a Science for the Anthropocene*. Princeton, NJ: Princeton University Press.

Shipley, B., and Keddy, P. (1987). The individualistic and community-unit concepts as falsifiable hypotheses. *Vegetatio* 69, 47–55. doi: 10.1007/BF00038686

Shrader-Frechette, K., and McCoy, E. (1993). *Method in Ecology: Strategies for Conservation*. Cambridge: Cambridge University Press.

Stokstad, E. (2014, August 25). Is ecology explaining less and less? *Science*.

Tukey, J. (1960). Conclusions vs. decisions. *Technometrics* 2, 423–433. doi: 10.1080/00401706.1960.10489909

Walker, A. M. (1969). On the asumptotic behavior of posterior distributions. *J. R. Stat. Soc. Ser. B* 31, 423–433.

Williamson, J. (2005). Bayesian nets and causality. Oxford: Oxford University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/fevo.2019.00413

Model Projections in Model Space: A Geometric Interpretation of the AIC Allows Estimating the Distance Between Truth and Approximating Models

José Miguel Ponciano[1]* *and Mark L. Taper*[1,2]

[1] *Biology Department, University of Florida, Gainesville, FL, United States,* [2] *Department of Ecology, Montana State University, Bozeman, MT, United States*

Information criteria have had a profound impact on modern ecological science. They allow researchers to estimate which probabilistic approximating models are closest to the generating process. Unfortunately, information criterion comparison does not tell how good the best model is. In this work, we show that this shortcoming can be resolved by extending the geometric interpretation of Hirotugu Akaike's original work. Standard information criterion analysis considers only the divergences of each model from the generating process. It is ignored that there are also estimable divergence relationships amongst all of the approximating models. We then show that using both sets of divergences and an estimator of the negative self entropy, a model space can be constructed that includes an estimated location for the generating process. Thus, not only can an analyst determine which model is closest to the generating process, she/he can also determine how close to the generating process the best approximating model is. Properties of the generating process estimated from these projections are more accurate than those estimated by model averaging. We illustrate in detail our findings and our methods with two ecological examples for which we use and test two different neg-selfentropy estimators. The applications of our proposed model projection in model space extend to all areas of science where model selection through information criteria is done.

Keywords: error rates in model selection, Kullback-Leibler divergence, model projections, model averaging, Akaike's Information Criterion

Edited by:
Hidetoshi Shimodaira,
Kyoto University, Japan

Reviewed by:
Ichiro Ken Shimatani,
Institute of Statistical Mathematics (ISM), Japan
Hirohisa Kishino,
The University of Tokyo, Japan

***Correspondence:**
José Miguel Ponciano
josemi@ufl.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *18 March 2019*
Accepted: *17 October 2019*
Published: *08 November 2019*

Citation:
Ponciano JM and Taper ML (2019) Model Projections in Model Space: A Geometric Interpretation of the AIC Allows Estimating the Distance Between Truth and Approximating Models. Front. Ecol. Evol. 7:413. doi: 10.3389/fevo.2019.00413

1. INTRODUCTION

Recent decades have witnessed a remarkable growth of statistical ecology as a discipline, and today, stochastic models of complex ecological processes are the hallmark of the most salient publications in ecology (e.g., Leibold et al., 2004; Gravel et al., 2016; Zeng and Rodrigo, 2018). Entropy and the Kullback-Liebler divergence as instruments of scientific inquiry are now at the forefront of the toolbox of quantitative ecologists, and many exciting new opportunities for their use are constantly being proposed (e.g., Casquilho and Rego, 2017; Fan et al., 2017; Kuricheva et al., 2017; Milne and Gupta, 2017; Roach et al., 2017; Cushman, 2018). One of the most important, but under explored, applications of the

Kullback-Liebler divergence remains the study or characterization of the error rates incurred while making model selection according to information criteria (Taper and Ponciano, 2016b). This research is particularly relevant when, as it almost always happens in science, none of the candidate models exactly corresponds to the chance mechanism generating the data.

Understanding the impact of misspecification of statistical models constitutes a key knowledge gap in statistical ecology, and many other areas of biological research for that matter (e.g., Yang and Zhu, 2018). Research by us and many others (see citations in Taper and Ponciano, 2016b and in Dennis et al., 2019) has led to detailed characterizations of how the probability of making the wrong model choice using any given information criterion, not only may depend on the amount of information (i.e., sample size) available, but also on the degree of model misspecification.

Consequently, in order to estimate the error rates of model selection according to any information criterion, practitioners are left with the apparent paradox ("catch-22") of being able to estimate how likely it is to erroneously deem as best that model which is furthest apart from the generating model, only after having accomplished the unsolved task of estimating the location of the candidate models relative to the generating process and to each other.

In this paper, we propose a solution to this problem. Our solution was motivated by the conceptualization of models as objects in a multi-dimensional space as well as an extension of the geometrical thinking that Akaike used so brilliantly in his 1973 paper introducing the AIC. Starting from Akaike's geometry, we show how to construct a model space that includes not only the set of candidate models but also an estimated location for the generating process. Now, not only can an analyst determine which model is closest to the generating process, she/he can also determine the (hyper)spatial relationships of all models and how close to the generating process the best model is.

In 1973, Hirotugu Akaike wrote a truly seminal paper presenting what came to be known as the AIC. Akaike initially called the statistic "An Information Criterion," but soon after its publication it came to be known as "Akaike's Information Criterion." Various technical accounts deriving the AIC exist (e.g., Burnham and Anderson, 2004, Chapter 7), but few explain in detail every single step of the mathematics of Akaike's derivation (but see De Leeuw, 1992). Although focusing on the measure-theoretic details, deLeeuw's account makes it clear that Akaike's paper was a paper about ideas, more than a paper about a particular technique. Years of research on this project has led us to understand that only after articulating Akaike's ideas, the direction of a natural extension of his work is easily revealed and understood. Although thinking of models and the generating mechanism as objects with a specific location in space is mathematically challenging, this exercise may also prove to be of use to study the adequacy of another common statistical practice in multi-model inference: model averaging.

Intuitively, if one thinks of the candidate models as a cloud of points in a Euclidean space, then it would only make sense to "average" the model predictions if the best approximation of the generating chance mechanism in that space is located somewhere inside the cloud of models. If however the generating model is located outside such cloud, then performing model average will only at best, worsen the predictions of the closest models to the generating mechanism. The question then is, can this idea of thinking about models as points in a given space be mathematically formalized? Can the structure and location of the candidate models and the generating mechanism be somehow estimated and placed in a space? If so, then the answer to both questions above (i.e., the error rates of multi-model selection under misspecification and when should an analyst perform model averaging) could be readily explored. These questions are the main motivation behind the work presented here.

2. THE AIC AND A NATURAL GEOMETRIC EXTENSION: MODEL PROJECTIONS IN MODEL SPACE

In his introduction to Akaike (1973)'s original paper, De Leeuw (1992) insisted on making sure it was understood that Akaike's contribution was much more valuable for its ideas than for its technical mathematical developments: "...This is an 'ideas' paper,' promoting a new approach to statistics, not a mathematics paper concerned with the detailed properties of a particular technique..." After this explanation, De Leeuw undertakes the difficult labor of teasing Akaike's thought process from the measure-theoretic techniques. In so doing, the author manages to present a clear and concise account clarifying both, Akaike's mathematical approach and his ideas. De Leeuw was keenly aware of the difficulty of trying to separate the ideas from the mathematical aspects of the paper: in introducing the key section in Akaike's paper, he describes it as "a section not particularly easy to read, that does not have the usual proof/theorem format, expansions are given without precise regularity conditions, exact and asymptotic identities are freely mixed, stochastic and deterministic expressions are not clearly distinguished and there are some unfortunate notational... typesetting choices" (De Leeuw, 1992). To us, however, the importance of De Leeuw's account stems from the fact that it truly brings home the crucial point that at the very heart of Akaike's derivation there was a geometrical use of Pythagoras' theorem (see Equation 1, page 604 in De Leeuw, 1992). The modern literature has been able to reduce Akaike's derivation to just a few lines (see Davison, 2003). However, such condensed proofs conceal the original geometric underpinnings of Akaike's thinking, which De Leeuw exposed. Our contribution for this special issue consists of taking Akaike's derivation one step further by using Pythagoras' theorem again to attain not a relative, but an absolute measure of how close each model in a model set is from the generating process.

Akaike's (1973) paper is difficult and technical but at the same time, it is a delightful reading because he managed to present his information criterion as the natural consequence of a logical narrative. That logical narrative consisted of six key insights that we strung together to arrive at what we believe is a second natural consequence of Akaike's foundational thoughts: our model projections proposal. After introducing our notation following Akaike's, we summarize those six key insights. We stress that these insights and the accompanying key figure we present below are none other than a simple geometric representation of De Leeuw's measure-theoretic re-writing of Akaike's proof. We encourage readers with a strong probability background to read De Leeuw's account. We then present our main model projections proposal and contribution and support it with a fully illustrated example.

2.1. Theoretical Insights From Akaike (1973)

Akaike's quest was motivated by a central goal of modern scientific practice: obtaining a comparison measure between many approximating models and the data-generating process. Akaike began thinking about how to characterize the discrepancy between any given approximating model and the generating process. He denoted the probability densities of the generating process and of the approximating model as $f(x, \theta_0)$ and $f(x, \theta)$, respectively, where θ_0 denoted the column vector of dimension L of true parameter values. Although he started by characterizing the discrepancy between the true model and the approximating model, his objective was to come up with an estimate of such discrepancy that somehow was free of the need of knowing either the dimension or the model form of $f(x, \theta_0)$. The fact that he was able to come up with an answer to such problem is not only outstanding, but the reason why the usage of the AIC has become ubiquitous in science. Akaike's series of arguments arriving to the AIC can be summarized by stringing together these six key insights:

2.1.1. Insight 1: Discrepancy From the Generating Process (Truth) Can Be Measured by the Average of Some Function of the Likelihood Ratio

Akaike's first important insight follows from two observations. First, under the parametric setting defined above, a direct comparison between an approximating model and the true, generating stochastic process can be achieved *via* the likelihood ratio, or some function of the likelihood ratio. Second, because the data X are random, the expected discrepancy (average over all possible realizations of the data) would be written as

$$\begin{aligned}\mathcal{D}(\theta, \theta_0;\, \Phi) &= \int f(x;\, \theta_0)\Phi(\tau(x, \theta, \theta_0))dx \\ &= \mathbb{E}_X\left[\Phi(\tau(X, \theta, \theta_0))\right],\end{aligned}$$

where the expectation is, of course, taken with respect to the generating stochastic process X. We denote the likelihood ratio as $\tau(x, \theta, \theta_0) = \frac{f(x;\theta)}{f(x;\theta_0)}$ and a twice differentiable function of it as $\Phi(\tau(x, \theta, \theta_0))$.

Akaike then proposed to study under a general framework how sensitive this average discrepancy would be to the deviation of θ from the truth, θ_0.

2.1.2. Insight 2: $\mathcal{D}(\theta, \theta_0;\, \Phi)$ Is Scaled by Fisher's Information Matrix

Akaike thought of expanding the average discrepancy $\mathcal{D}(\theta, \theta_0;\, \Phi)$ using a second order series approximation around θ_0. Akaike's second insight then consisted of noting the strong link between such approximation and the theory of Maximum Likelihood (ML).

For a univariate θ, the Taylor series approximation of the average function Φ of the likelihood ratio is written as

$$\begin{aligned}\mathcal{D}(\theta, \theta_0;\, \Phi) \approx\ & \mathcal{D}(\theta_0, \theta_0;\, \Phi) + (\theta - \theta_0)\left.\frac{\partial \mathcal{D}(\theta, \theta_0;\, \Phi)}{\partial \theta}\right|_{\theta\ =\theta_0} \\ & + \frac{(\theta - \theta_0)^2}{2!}\left.\frac{\partial^2 \mathcal{D}(\theta, \theta_0;\, \Phi)}{\partial \theta^2}\right|_{\theta=\theta_0} + \dots \qquad (1)\end{aligned}$$

To find an interpretable form of this approximation, just like Akaike did following Kullback and Leibler (Kullback and Leibler, 1951; Akaike, 1973), we use two facts: first, by definition $\tau(x, \theta, \theta_0)|_{\theta=\theta_0} = 1$ and second, that $\int f(x;\, \theta)dx = 1$ because f is a probability density function. Together with the well-known regularity conditions used in mathematical statistics that allow differentiation under the integral sign (Pawitan, 2001), these two facts give us the following: first, $\int \frac{\partial f(x;\theta)}{\partial \theta}dx = \int \frac{\partial^2 f(x;\theta)}{\partial \theta^2}dx = 0$. Hence, $\left.\frac{\partial \mathcal{D}(\theta,\theta_0;\Phi)}{\partial \theta}\right|_{\theta=\theta_0} = 0$. This result then allows writing the second derivative of the approximation as

$$\begin{aligned}\left.\frac{\partial^2 \mathcal{D}(\theta, \theta_0;\, \Phi)}{\partial \theta^2}\right|_{\theta=\theta_0} &= \int \frac{\partial}{\partial \theta}\left(\frac{\partial \Phi(\tau)}{\partial \tau}\frac{\partial \tau}{\partial \theta}\right) f(x;\, \theta_0)dx\bigg|_{\theta=\theta_0} \\ &= \int \frac{\partial^2 \Phi(\tau)}{\partial \tau^2}\left(\frac{\partial \tau}{\partial \theta}\right)^2 f(x;\, \theta_0)dx\bigg|_{\theta=\theta_0} \\ &\quad + \int \frac{\partial^2 \tau}{\partial \theta^2}\frac{\partial \Phi(\tau)}{\partial \tau} f(x;\, \theta_0)dx\bigg|_{\theta=\theta_0} \\ &= \Phi''(1)\int \left(\frac{1}{f(x;\, \theta_0)}\frac{\partial f(x;\, \theta)}{\partial \theta}\right)^2 f(x;\, \theta_0)dx\, |S_{\theta=\theta_0} \\ &= \Phi''(1)\int \left(\frac{\partial \log f(x;\, \theta)}{\partial \theta}\right)^2 f(x;\, \theta)dx\, |_{\theta=\theta_0} \\ &= \Phi''(1)\mathcal{I}(\theta_0),\end{aligned}$$

where $\mathcal{I}(\theta_0)$ is Fisher's information. To move from the first line of the above calculation to the second line we used a combination of the product rule and of the chain rule. To go from the second to the third line, note that because the first derivative is equal to 0 as shown immediately above of this equation, the integral in the right hand is null.

Hence, in this univariate case, the second order approximation is given by $\mathcal{D}(\theta, \theta_0) \approx \Phi(1) + \frac{1}{2}\Phi''(1)(\theta - \theta_0)^2\mathcal{I}(\theta_0)$, where $\mathcal{I}(\theta_0)$ is Fisher's information. Thus, the average discrepancy between an approximating and a generating model is scaled by the inverse of the theoretical variance of the Maximum Likelihood estimator, regardless of the form of the function $\Phi()$.

2.1.3. Insight 3: Setting $\Phi(t) = -2\log t$ Connects $\mathcal{D}(\theta, \theta_0;\, \Phi)$ With Entropy and Information Theory

Akaike proceeded to arbitrarily set the function $\Phi(t)$ to $\Phi(t) = -2\log t$. Using this function not only furthered the connection with ML theory, but also introduced the connection of his thinking with Information Theory. By using this arbitrary function, the average discrepancy becomes a divergence because

$\mathcal{D}(\theta_0, \theta_0) = \Phi(1) = 0$ and the approximation of the average discrepancy, heretofore denoted as $\mathcal{W}(\theta, \theta_0)$, is modulated by Fisher's information, the variance of the Maximum Likelihood estimator: $\mathcal{D}(\theta, \theta_0) \approx \mathcal{W}(\theta, \theta_0) = (\theta - \theta_0)^2 \mathcal{I}(\theta_0)$. For a multivariate θ_0 we get then that $\mathcal{W}(\theta, \theta_0) = (\theta - \theta_0)' \mathcal{I}(\theta_0)(\theta - \theta_0)$ where $\mathcal{I}(\theta_0)$ is Fisher's Information matrix (Pawitan, 2001). Conveniently then, the arbitrary factor of 2 gave his general average discrepancy function the familiar "neg-entropy" or Kullback-Leibler (KL) divergence form

$$\begin{aligned}
\mathcal{D}(\theta, \theta_0) &= -2 \int f(x; \theta_0) \log\left(\frac{f(x;\theta)}{f(x;\theta_0)}\right) dx \\
&= -2\mathbb{E}_X\left[\log \frac{f(X;\theta)}{f(X;\theta_0)}\right] \\
&= -2\left[\mathbb{E}_X\left(\log f(X;\theta)\right) - \mathbb{E}_X\left(\log f(X;\theta_0)\right)\right] \\
&= 2\mathbb{E}_X\left(\log f(X;\theta_0)\right) - 2\mathbb{E}_X\left(\log f(X;\theta)\right) \\
&= 2KL(\theta, \theta_0) \qquad (2)
\end{aligned}$$

thus bringing together concepts in ML estimation with a wealth of results in Information Theory. The two expectations (integrals) in the last line of the above equation were often succinctly denoted by Akaike as Sgg and Sgf, respectively: these are the neg-selfentropy and the neg-crossentropy terms. Thus, he would write that last line as $2KL(\theta, \theta_0) = 2[Sgg - Sgf]$. Note that for consistency with Akaike (1973) we have retained his notation and in particular, the order of arguments in the KL function, as opposed to the notation we use in Dennis et al. (2019).

2.1.4. Insight 4: $\mathcal{D}(\theta, \theta_0)$ Is Minimized at the ML Estimate of θ

Aikaike's fourth critical insight was to note that a Law of Large Numbers (LLN) approximation of the Kullback-Leibler divergence between the true, generating stochastic process and a statistical model is minimized by evaluating the candidate model at its maximum likelihood estimates. Such conclusion can be arrived at even if the generating stochastic model is not known. Indeed, given a sample of size n, $X_1, X_2, \ldots, X_n$ from the generating model, from the LLN we have that

$$\hat{\mathcal{D}}_n(\hat{\theta}, \theta_0) = -2 \times \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i; \hat{\theta})}{f(x_i; \theta_0)},$$

which is minimized at the ML estimate $\hat{\theta}$. Akaike actually thought that this observation could be used as a *justification for the maximum likelihood principle*: "Though it has been said that the maximum likelihood principle is not based on any clearly defined optimum consideration, our present observation has made it clear that it is essentially designed to keep minimum the estimated loss function which is very naturally defined as the mean information for discrimination between the estimated and the true distributions" Akaike (1973).

2.1.5. Insight 5: Minimizing $\mathcal{D}(\theta, \theta_0)$ Is an Average Approximation Problem

Akaike's fifth insight was to recognize the need to account for the randomness in the ML estimator. Because multiple realizations of a sample $X_1, X_2, \ldots, X_n$ each results in different estimates of θ, the average discrepancy should be considered a random variable. The randomness hence, is with respect to distribution of the maximum likelihood estimator $\hat{\theta}$. Let $\mathcal{R}(\theta_0) = \mathbb{E}_{\hat{\theta}}\left[\mathcal{D}(\hat{\theta}, \theta_0)\right]$ denote our target average over the distribution of $\hat{\theta}$. Then, the problem of minimizing the Kullback Leibler divergence can be conceived as an approximation problem where the target is the average:

$$\begin{aligned}
\mathcal{R}(\theta_0) = \mathbb{E}_{\hat{\theta}}\mathcal{D}(\hat{\theta}, \theta_0) &= 2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f(X;\theta_0)\right) \right. \\
&\qquad \left. - \mathbb{E}_X\left(\log f(X;\hat{\theta})|\hat{\theta}\right)\right] \\
&= 2\mathbb{E}_X\left(\log f(X;\theta_0)\right) \\
&\qquad - 2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f(X;\hat{\theta})|\hat{\theta}\right)\right].
\end{aligned}$$

In the final expression of the equation above, the first term is an unknown constant. The second term on the other hand, is the expected value of a conditional expectation.

2.1.6. Insight 6: $\mathcal{D}(\theta, \theta_0)$ Can Be Approximated Geometrically Using Pythagoras' Theorem

Instead of estimating the expectations above, Akaike thought of substituting the probabilistic entropy $\mathcal{D}(\hat{\theta}, \theta_0)$ with its Taylor Series approximation $\mathcal{W}(\hat{\theta}, \theta_0) = (\hat{\theta} - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)$, which can then be interpreted as a squared statistical distance. This approximation is indeed the square of a statistical distance wherein the divergence between any two points $\hat{\theta}$ and θ_0 is weighted by their dispersion in multivariate space, measured by the eigenvalues of the positive definite matrix $\mathcal{I}(\theta_0)$. This sixth insight led him straight into the path to learning about the KL divergence between a generating process and a set of proposed probabilistic mechanisms/models. By viewing this quadratic form as a statistical distance, Akaike was able to use a battery of clear measure-theoretic arguments relying on various convergence proofs to derive the AIC.

Interestingly, and although he doesn't explicitly mentions it in his paper, his entire argument can be phrased geometrically: if the average discrepancy that he was after could be approximated with the square of a statistical distance, its decomposition using Pythagoras theorem was the natural thing to do. By doing such decomposition, one can immediately visualize the ideas in his proof with a simple sketch. We present such sketch in **Figure 1**. In that figure, the key triangle with a right angle has as vertices the truth θ_0 of unknown dimension L, the ML estimator $\hat{\theta}$ of dimension $k \leq L$, denoted $\hat{\theta}_k$ and finally, θ_{0k}. This quantity represents the orthogonal projection of the truth in the plane where all estimators of dimension k lie, which is in turn denoted as Θ_k (**Figure 1A**). **Figure 1B** shows a fourth crucial point in this geometrical interpretation: it is the estimator of θ_0 from the data using a model with the same model form as the generating model, but with parameters estimated from the data. To distinguish it from $\hat{\theta}_k$ we denote this estimator $\hat{\theta}_0$. Because it has the same dimensions than the generating model, $\hat{\theta}_0$ can be thought of as being located in the same model surface as the generating model θ_0. Akaike's LLN approximation of the KL

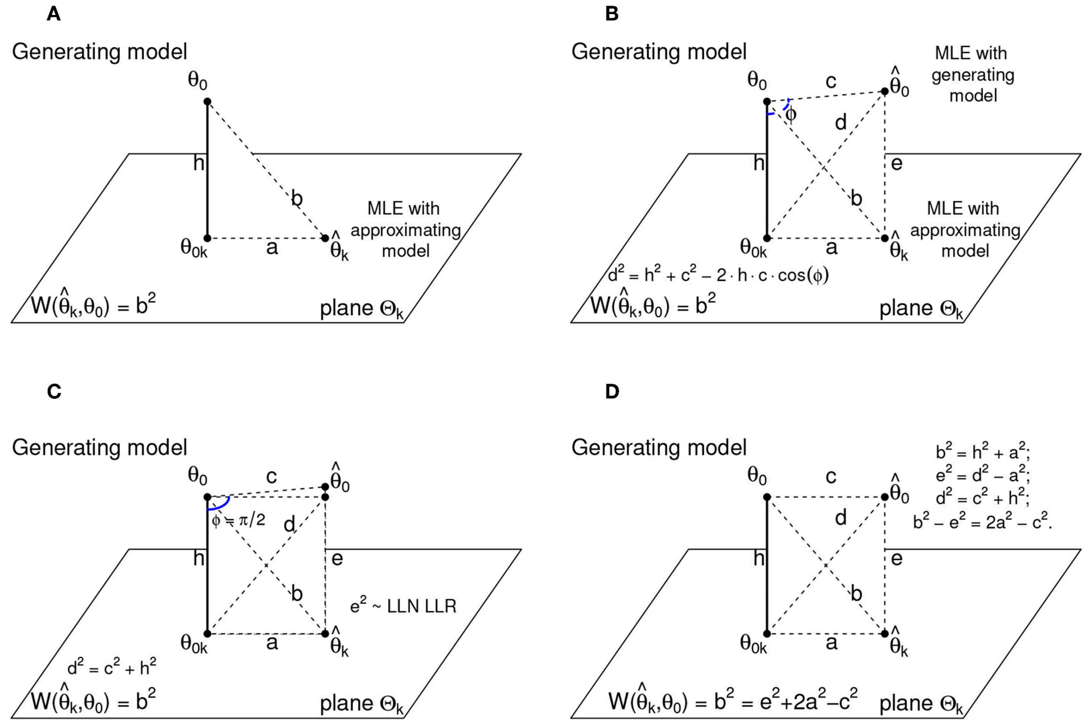


FIGURE 1 | The geometry of Akaike's Information Criterion. **(A)** Shows θ_0, which is the generating model and θ_{0k} which is the orthogonal projection of the generating model into the space Θ_k of dimension k. $\hat{\theta}_k$ is the ML estimate (MLE) of an approximating model of dimension k given a data set of size n. Akaike's objective was to solve for b^2, which represents in this geometry $\mathcal{W}(\hat{\theta}, \theta_0)$, the quadratic form approximation of the divergence between the generating and the approximating models. Akaike showed that $\hat{\theta}_k$ can be thought of as the orthogonal projection of the MLE of $\hat{\theta}_0$ **(B)**. This last quantity $\hat{\theta}_0$ represents the MLE of θ_0 with a finite sample of size n and assuming that the correct model form is known. The angle ϕ is not necessarily a right angle, but Akaike used $\phi \approx \pi/2$ so that the generalized Pythagoras theorem [equation on the lower left side of **(B)**] could be approximated with the simple version of Pythagoras [equation on the lower left side of **(C)**] when the edge h is not too long. When implemented, this Pythagoras equation can be used in conjunction with the other Pythagorean triangles in the geometry to solve for the squared edge b. The equations leading to such solution are shown in **(D)**.

divergence as an average of log-likelihood ratios $\hat{\mathcal{D}}_n(\hat{\theta}, \theta_0) = -2 \times \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\hat{\theta})}{f(x_i;\theta_0)}$ comes to play in this geometric derivation as the edge labeled e^2 in **Figure 1B** that traces the link between $\hat{\theta}_0$ and the ML estimator $\hat{\theta}_k$. Following Akaike's derivation then, the ML estimator $\hat{\theta}_k$ can be thought as the orthogonal projection of $\hat{\theta}_0$ onto the plane Θ_k.

Before continuing with our geometric interpretation, we alert the reader that in **Figure 1** all the edges are labeled with a lowercase letter with the purpose of facilitating this geometric visualization. The necessary calculations to understand Akaike's results are presented as simplified algebraic calculations but the reader however, is warned that these edges or lower case letters denote for the most part random variables. We leave these simple letters here because in Akaike's original derivations, the technical measure-theoretic operations may end up distracting the reader from a natural geometric understanding of the AIC.

In simple terms then, the objective of this geometric representation is to see that obtaining an estimate of the discrepancy between the approximating model and the generating process amounts to solving for the square of the edge length b, which is in fact the KL divergence quadratic form approximation. That is, $b^2 = \mathcal{W}(\hat{\theta}, \theta_0)$. Proceeding with our geometric interpretation, note that the angle ϕ between edges h and c in **Figure 1B** is not by necessity a right angle, and that the generalized Pythagoras Theorem to find the edge length d applies. Akaike then noted that provided that the approximating model is in the vicinity of the generating mechanism, the third term of the generalized Pythagoras form of the squared distance $d^2 = c^2 + h^2 - 2ch \cos \phi$ was negligible when compared with c^2 and h^2 [see Akaike, 1973, his Equation (4.15) and his comment about that term in the paragraph above his Equation (4.19). See also De Leeuw 1992, text under his Equation (4)], and so he proceeded to simply use only the first two terms, c^2 and h^2 (see **Figure 1C**). The immense success of the AIC in a wide array of scientific settings to date shows that this approximation, as rough as it may seem, is in fact quite reliable. This approximation allowed him to write the squared distance d^2 in two different ways: as $d^2 \approx c^2 + h^2$ and as $d^2 = a^2 + e^2$. Because by construction, we have that $b^2 = h^2 + a^2$, one can immediately write the difference $b^2 - e^2$ as

$$\begin{aligned} b^2 - e^2 &= h^2 + a^2 - d^2 + a^2 \\ &= h^2 + a^2 - c^2 - h^2 + a^2, \end{aligned}$$

and then solve for b^2 (see **Figure 1D**):

$$b^2 = e^2 + 2a^2 - c^2. \tag{3}$$

Using asymptotic expansions of these squared terms, the observed Fisher's information and using known convergence in probability results, Akaike showed when multiplied by the sample size *n*, the difference of squares $c^2 - a^2$ was approximately chi-squared distributed with degrees of freedom $L - k$ and that $na^2 \sim \chi^2_k$. Then, multiplying equation (3) by *n* gives

$$nb^2 = n\mathcal{W}(\hat{\theta}_k, \theta_0) \approx \underbrace{n\mathcal{D}_n(\hat{\theta}_k, \hat{\theta}_0)}_{=2\times\text{log-likelihood ratio}} + \underbrace{na^2}_{\sim\chi^2_k} - \underbrace{n(c^2 - a^2)}_{\sim\chi^2_{L-k}}.$$

Finally, one may arrive at the original expected value of the conditional expectation shown above by replacing the chi-squares with their expected values, which are given by their degrees of freedom. Hence,

$$\begin{aligned} n\mathbb{E}_{\hat{\theta}_k}\left[\mathcal{W}(\hat{\theta}_k, \theta_0)\right] &\approx n\mathcal{D}_n(\hat{\theta}_k, \hat{\theta}_0) + 2k - L, \text{ or} \\ \mathbb{E}_{\hat{\theta}_k}\left[\mathcal{W}(\hat{\theta}_k, \theta_0)\right] &\approx \frac{-2}{n}\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_k) + \frac{2k}{n} - \frac{L}{n} \\ &\quad + \frac{2}{n}\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_0). \end{aligned} \tag{4}$$

The first two terms in the above expression, $-2\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_k) + 2k$, constitute what came to be known as the AIC. These terms correspond respectively to twice the negative log-likelihood evaluated at the MLE and twice the number of parameters estimated in the approximating model. To achieve multi-model comparison (see **Figure 2**), Akaike swiftly pointed out that in fact, only these first two terms are needed because the true model dimension L and the term $\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_0)$ both terms (1) remain the same across models, as long as the same data set is used and (2) *cannot be known* because they refer to the true model dimension. Akaike rightly noted that if one were to compute Equation (4) for a suite of approximating models, these two terms would remain the same across all models and hence, could in practice be ignored for comparison purposes: these unknowns then act as constants of proportionality that are invariant to model choice. Therefore, in order to compare the value of this estimated average discrepancy across a suite of models, the user only needs to calculate the AIC score $-2\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_k) + 2k$ for each model and deem as best that model for which the outcome of this calculation is the smallest. The logic embedded in Akaike's reasoning is represented graphically in **Figure 2** (redrawn from Burnham et al., 2011). This reasoning kickstarted the practice, still followed in science 46 years later, to disavow the absolute truth in favor of a careful examination of multiple, if not many, models.

Finally, the reader should recall that what Equation (4) is in fact approximating is

$$\begin{aligned} \mathcal{R}(\theta_0) = \mathbb{E}_{\hat{\theta}}\mathcal{D}(\hat{\theta}_k, \theta_0) &= -2\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f(X; \hat{\theta}_k)|\hat{\theta}_k\right)\right] \\ &\quad + 2\mathbb{E}_X\left(\log f(X; \theta_0)\right). \end{aligned} \tag{5}$$

and that this last expression is in fact the expectation with respect to $\hat{\theta}_k$ of

$$\begin{aligned} -2\int f(x; \theta_0)\log\frac{f(x; \hat{\theta}_k)}{f(x; \theta_0)}dx &= -2\int f(x; \theta_0)\log f(x; \hat{\theta}_k)dx \\ &\quad + 2\int f(x; \theta_0)\log f(x; \theta_0)dx. \end{aligned} \tag{6}$$

Later, Akaike (1974) referred to the integral $\int f(x; \theta_0)\log f(x; \hat{\theta}_k)dx$ as *Sgf* and to $\int f(x; \theta_0)\log f(x; \theta_0)dx$ as *Sgg*, which are names easy to remember because it's almost as if the *S* in *Sgf* and *Sgg* represent the integral sign and *g* and *f* are a short hand representation of the probability density function of the generating stochastic process and of the approximating model, respectively.

One of our central motivations to write this paper is the following: by essentially ignoring the remainder terms in Equation (4), since 1973 practitioners have been almost invariably selecting the "least worst" model among a set of models (but see Spanos, 2010). In other words, we as a scientific community, have largely disregarded the question of how far, *in absolute terms not relative*, is the generating process from the best approximating model. Suppose the generating model is in fact very far from all the models in a set of models currently being examined. Then, the last term in Equation (4) will be very large with respect to the first two terms for all the models in a model set that is being examined, and essentially any differences between the terms $-2\sum_{i=1}^{n}\log f(x_i; \hat{\theta}_k) + 2k$ for every model will be meaningless.

2.2. The Problem of Multiple Models

Akaike's realization that "truth" did not need to be known in order to select from a suite of models which one was closest to truth shaped the following four and a half decades of scientific undertaking of model-centered science. Scientists were then naturally pushed toward the confrontation of not one or two, but multiple models with their experimental and observational data. Such approach soon led to the realization that basing the totality of the inferences on the single best model was not adequate because it was often the case that a small set of models would appear indistinguishable from each other when compared (Taper and Ponciano, 2016b).

Model averaging is by far, the most common approach used today to make inferences and predictions following an evaluation of multiple models *via* the AIC. Multiple options to do model averaging exist but in all cases, this procedure is an implicit Bayesian methodology that results in a set of posterior probabilities for each model. These posterior probabilities are called the "Akaike weights." For the i^{th} model in a set of candidate models, this weight is computed as

$$w_i = \frac{e^{(-\Delta_i/2)}}{\sum_{r=1}^{R} e^{(-\Delta_r/2)}}.$$

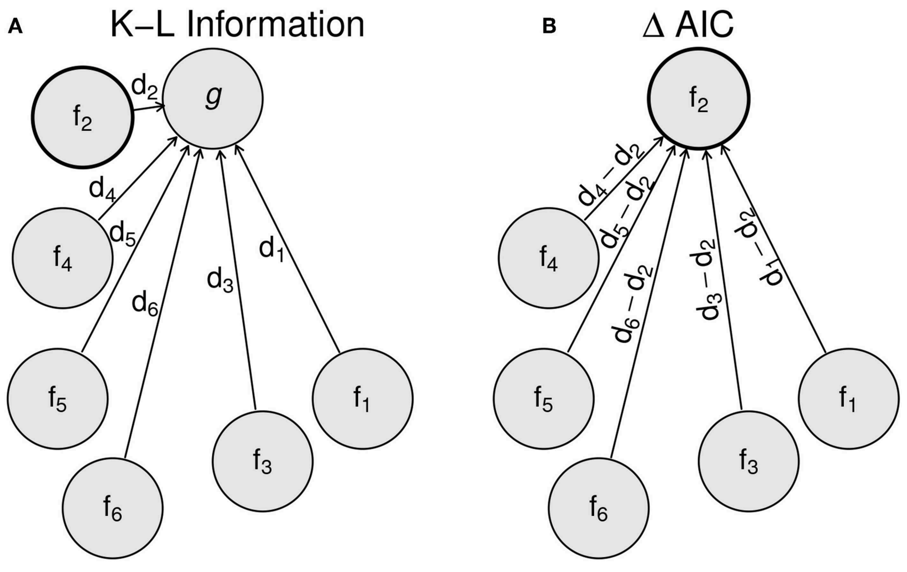


FIGURE 2 | Schematic representation of the logic of multi-model selection using the AIC. *g* represents the generating model and f_i the i^{th} approximating model. The Kullback-Leibler information discrepancies (d_i) are shown on the left **(A)** as the distance between approximating models and the generating model. The ΔAICs shown on the right **(B)** measures the distance from approximating models to the best approximating model. All distances are on the information scale.

In this expression, Δ_i is the i^{th} difference between the AIC value and the best (i.e., the lowest) AIC score in the set of R candidate models. Although this definition is very well-known, cited and used (Taper and Ponciano, 2016b), it is seldom acknowledged that because these weights are in fact posterior probabilities, they must result from adopting a specific set of subjective model priors. Burnham et al. (2011) actually show that the weights shown above result from adopting the following subjective priors q_i:

$$q_i = C \cdot \exp\left(\frac{1}{2}k_i \log(n) - k_i\right), \tag{7}$$

where C is a normalization constant, k_i is the model dimension (the estimated number of parameters) of model i and n denotes the total sample size. Note that with sample sizes above 7, those weights increase with the number of parameters, thus favoring parameter rich models. The use of these priors makes model averaging a confirmation approach (Bandyopadhyay et al., 2016).

For someone using evidential statistics, adopting the model averaging practice outline above presents two important problems: first, the weights are based on prior beliefs that favor more parameter rich models and are not based on actual evidence (data). Second, and much more practically, model averaging appears to artificially favor redundancy of model specification: the more models that are developed in any given region of model space, the stronger this particular region gets weighted during the model averaging process. To counter these two problems, here we propose alternatively to estimate (1) the properties of a hyper-plane containing the model set, (2) the location in such plane of the best projection of the generating process and (3) an overall general discrepancy between each of the models in the model set and the generating process or truth. We achieve these goals by using the estimated KL divergences amongst all estimated models, that is, the estimated Sf_if_j for all models i and j in the candidate set. This is information that is typically ignored. Here again, we use Akaike's mnemonic notation where g denotes the generating model and f the approximating model. Then the so called neg-crossentropy and neg-selfentropy are written as

$$\begin{aligned} Sgf &= \int f(x;\theta_0)\log f(x;\hat{\theta}_k)dx \quad \text{and} \\ Sgg &= \int f(x;\theta_0)\log f(x;\theta_0)dx, \quad \text{respectively.} \end{aligned}$$

In his 1974 paper, Akaike observed that the neg-crossentropy could be estimated with

$$\widehat{Sgf} = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\hat{\theta}_k) - \frac{k}{n} = -\frac{AIC}{2n}. \tag{8}$$

We wish to point out that in the "popular" statistical literature within the Wildlife Ecology sciences (e.g., Burnham and Anderson, 2004; Burnham et al., 2011), it is often repeated that an estimator of $\mathbb{E}_{\hat{\theta}}\left[\mathbb{E}_X\left(\log f(X;\hat{\theta}_k)|\hat{\theta}_k\right)\right]$ is given by $-AIC/2$. In fact, Akaike (1974) shows that the correct estimator is given by Equation (8). This distinction, albeit subtle, marks a difference when the analyst wishes to compare not only which model best approximates the generating process, but also the strength of the evidence for one or the other model choice.

In what follows, we extend Akaike's geometric derivation to make inferences regarding the spatial configuration of the ensemble of models being considered as approximations to the generating process. As we show with an ecological example, unlike model averaging this natural geometric extension of the

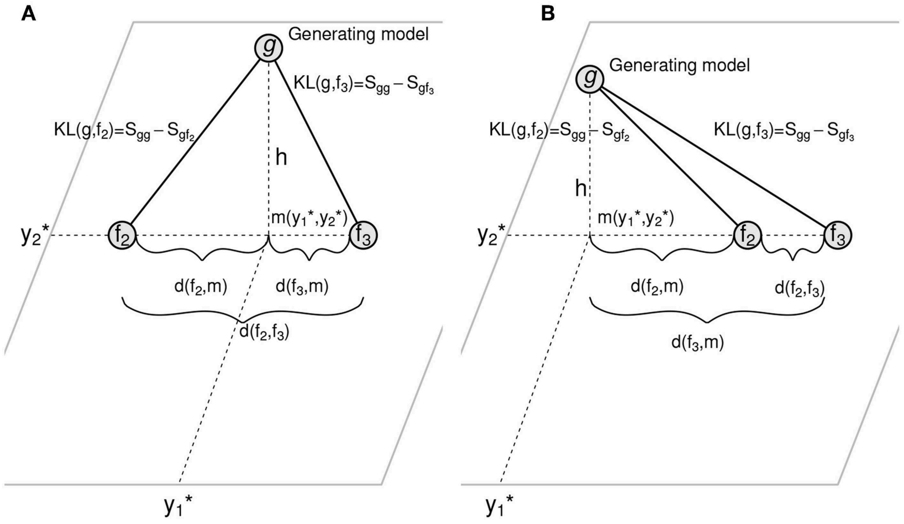


FIGURE 3 | The geometry of model space. In this figure, f_2 and f_3 are approximating models residing in a (hyper)plane. g is the generating model. m is the projection of g onto the (hyper)plane. $d(;)$ are distances between models in the plane. $d(f_2, f_3) \approx KL(f_2, f_3)$ with deviations due to the dimension reduction in NMDS and non-Euclidian behavior of KL divergences. As KL divergences decrease, they become increasingly Euclidian. **(A)** Shows a projection when m is within the convex hull of the approximating models, and **(B)** shows a projection when m is outside of the convex hull. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

AIC is fairly robust to the specification of models around the same region of model space and is actually aided, not hampered, by proposing a large set of candidate models.

2.3. A Geometrical Extension of Akaike's Extension to the Principle of Maximum Likelihood

As modelers, scientists are naturally drawn to visualize a suite of candidate models as entities in a (hyper)plane. By so doing, the geometric proximities between these entities are then intuitively understood as similarities amongst models. The key questions we answer in this paper are whether it is possible to estimate the architecture of such model space, locate a suite of approximating models within such space as well as estimating the location of the projection of truth onto that plane. All of this while not having to formulate an explicit model for the generating model. The estimation of the location of the truth projection in that plane would open the door to a formulation of an overall goodness of fit measure qualifying every single one of the AIC scores computed for a set of candidate models. Additionally, answering these questions automatically provides valuable insights to intuitively understand why or why not model averaging may be an appropriate course of action. As we show below, these questions are answerable precisely because any given set of models has a set of relationships which are typically ignored but that can be translated directly to a set of geometrical relationships that carry all the needed information and evidence.

One of the key observations of this contribution is the fact that while at the time of Akaike's publication his approach could not be extended due to mathematical intractabilities, nowadays computer intensive methods allow the design of a straightforward algorithm to solve the model projection problem outlined above. These computational tools basically involve two methodologies: first, a numerical estimation of Kullback-Leibler (KL) divergences between arbitrary distributions and second, parallel processing to carry a Non-Metric Multidimensional (NMDS) space scaling algorithm. With the help of a NMDS, a matrix of amongst-candidate models estimated KL divergences can be transformed into an approximated Euclidean representation of models in a (hyper)plane. The coordinates of each model in that plane, that we heretofore denote $\left(y_1, y_2, ...\right)$ are used to solve the model projection problem. The algorithm presented here is not necessarily restricted to a two-dimensional representation of model space, but for the sake of visualization we present our development in $\mathcal{R}^2$.

Consider the sketch in **Figure 3**. There, to begin with we have drawn only two approximating models f_2 and f_3 on a Euclidean space, along with a depiction of the location of the generating process g outside that plane. Such representation immediately leads to the definition of a point m in that plane that correspond to the orthogonal projection of the generating process onto the plane. The location of such point is denoted as $(y_1^\star, y_2^\star)$. The length h in that sketch represents the deviation of the generating process from the plane of approximating models as a line from g to the plane that crosses such plane perpendicularly. Note also that every one of the approximating models f_i in that plane is situated at a distance $d(f_i, m)$ from the orthogonal projection m. In reality, both the edges as well as the points in this plane are random variables associated with a sampling error. But we ask the reader's indulgence for the sake of the argument, just as we did above when we explained Akaike's results, and think of these simply as points and fixed lengths. Doing so, one may also indulge, as Akaike did, in using the right-angle, simple version of the Pythagoras theorem, and assume that all the amongst-models KL divergences have a corresponding squared Euclidean distance in that representation. Then, the following equations hold

$$\begin{cases} KL(g,f_1) = d(f_1,m)^2 + {h_1}^2 \\ KL(g,f_2) = d(f_2,m)^2 + {h_2}^2 \\ \quad\quad\vdots \end{cases}$$

where necessarily $h_1 = h_2 = h_i = \ldots = h$. Recalling Equation (8) we note that every one of the divergences between the approximating models and g can be expressed as a sum of an estimable term and a fixed, unknown term. These terms are Sgf_i and Sgg, respectively. Writing such decomposition of the KL divergences for all the equations above, and explicitly incorporating the coordinates of m then results in this system of equations

$$\begin{cases} Sgg - \widehat{Sgf_1} - d(f_1, m(y_1^\star, y_2^\star))^2 = h_1^2, \\ Sgg - \widehat{Sgf_2} - d(f_2, m(y_1^\star, y_2^\star))^2 = h_2^2, \\ \quad\quad\vdots \quad\quad\quad\quad\quad\quad\quad \vdots \;\; \vdots \end{cases} \tag{9}$$

which can be solved and optimized computationally by constructing an objective function that, for any given set of values of $Sgg, y_1^\star, y_2^\star$ in the left hand of these equations returns the sum of squared differences between all the h_i. Because by necessity (see **Figure 3**) $h^2 = h_i^2$ for all i, a routine minimization of this sum of squared differences can be used as the target to obtain optimal values of the unknown quantities of interest and obtain the model-projection representation shown in **Figure 5**. Although previously unrecognized by Taper and Ponciano (2016a), in these equations the terms Sgg and h^2 appear always as a difference, and hence are not separable. Fortunately, a non-parametric, multivariate estimate of Sgg can be readily computed. We use the estimator proposed by Berrett et al. (2019), a multivariate extension of the well-known univariate estimator by Kozachenko and Leonenko (1987). Other non-parametric entropy estimators could be used if they prove to be more appropriate. For instance, the Berrett et al. (2019) estimator assumes that the data are iid. This restricts the class of problems for which we are able to separate Sgg and h^2. An estimator for Sgg for dependent data would expand the class.

3. EXAMPLES

In what follows we illustrate our ideas and methodology with two ecological examples. The first example is an animal behavior study aiming to understand the mechanism shaping patterns of animal aggregations. The second one is an ecosystems ecology example, where the aim was to try to understand the biotic and abiotic factors that shape the species diversity and composition of a shrubland ecosystem in California.

3.1. An Application in Animal Behavior

The phenomenon of animal aggregations has long been the focus of interest for evolutionary biologists studying behavior (Brockmann, 1990). In some animal species, males form groups surrounding females, seeking breeding opportunities. Often, these mating groups vary substantially in size, even during the same breeding season and breeding occasion. This is particularly true in some species with external fertilization where females spawn the eggs and one or more males may fertilize them. The females of the American horseshoe crab, *Limulus polyphemus* leave sea "en masse" to spawn at the beach during high tide, 1–4 times a year. As females enter the beach and find a place to spawn, males land in groups and begin to surround the females. Nesting typically occur in pairs, but some females attract additional males, called satellites, and spawn in groups. As a result, when surveys of the mating group size are done, one may encounter horseshoe crab pairs with $0, 1, 2, 3, \ldots$ satellite males. That variation in the number of satellite males is at the root of the difficulty in characterizing the exact make-up of the crab population. Hence, for years during spawning events, Brockmann (1990) focused on recording not only the total number of spawning females in a beach in Seahorse Key (an island along Florida's northern west coast) but also the number of satellite males surrounding each encountered pair. Those data have long been the focus of attempts at a probabilistic description of the distribution of the number of satellite males surrounding a pair of horseshoe crabs using standard distribution models (e.g., Poisson, zero inflated Poisson, negative binomial, zero inflated negative binomial, hurdle-negative binomial distributions).

When one of us (JMP) met H. J. Brockmann in 2010, she asked the following: "how will fitting different discrete probability distributions to my data help me understand the biological mechanisms underlying group formation in this species?" After years of occasional one-on one meetings and back and forth discussions, we put together a detailed study (Brockmann et al., 2018) in which we compared the observed distribution of the number of satellites surrounding a female to the same distribution resulting from a complex, individual-based model simulation program. Importantly, this individual-based model allowed us to translate different hypotheses regarding the influence of different factors, like female density or male density around a female, into the decision by a new satellite male of joining a mating group or continuing the search.

The comparison between the real data and the simulated data *via* discrete probability distributions then allowed these authors to identify the biological settings that resulted in *in silico* distributions of satellites that most resembled the real, observed distributions of satellite males. To do that comparison, Brockmann et al. (2018) first fitted a handful of discrete probability models to the counts of the number of satellites surrounding each pair from each one of $N = 339$ tides, and proceeded to find the standard probability model that best described the data. These authors then fitted the same models to the simulated data sets under different biological scenarios and found the simulation setting that yielded the highest resemblance between the real data and the digital data. Finally Brockmann et al. (2018) discuss the implications of the results.

One of the most relevant conclusions of these authors was that their comparative approach was useful as a hypothesis generator. Indeed, by finding via trial and error which biological processes gave rise in the individual-based simulations to distributions of satellites that most resembled the real distributions, the researchers basically came up with a system to elicit viable biological explanations for the mechanisms shaping the distribution of the number of satellite males surrounding a pair. This approach was an attempt to answer Brockmann's initial question to JMP.

Here, we used the simulation setting of Brockmann et al. (2018) to exemplify how our Model Projections in Model Space (MPMS) approach can further our understanding of what are the model attributes that make a model a good model to better understand the underlying mechanisms generating the data. By having a complex simulation program, we can describe exactly the probability distribution of the data-generating process and we can validate our MPMS approach.

In what follows we first explain how we fitted our proposed models to the tides' count data, and then how we compute the quantities needed to generate an approximate representation of models in model space that includes the estimated projection of the true, data-generating process.

3.1.1. Likelihood Function for the Satellites Count Data

A handful of discrete probability models can be fit conveniently to the male satellites counts data using the same general likelihood functions by means of a reduced-parameter multinomial distribution model parameterization. As we will see below, this reduced-parameter multinomial likelihood formulation is instrumental to compute analytically the KL divergences between each one of the models as well as the neg-selfentropy. Many modern biological models, like phylogenetic Markov models, use this reduced-parameter formulation (Yang, 2000), and the example presented here can be readily used in many other settings in ecology and evolution (e.g., Rice, 1995).

In this example we adopt the following notation: the probability mass function of each discrete probability model i ($i = 1, 2, \ldots r$ where r is the number of models in the model set) is denoted as $f_i(x)$. Following Brockmann et al. (2018), we use $f_1(x)$ to denote the Poisson distribution (Poisson), $f_2(x)$ the negative binomial distribution (NegBin), $f_3(x)$ the zero inflated Poisson distribution (ZIP), $f_4(x)$ the zero inflated negative binomial distribution (ZINegBi), $f_5(x)$ a hurdle negative binomial distribution (HurdNBi), $f_6(x)$ a Poisson-negative binomial mixture (PoiNB), $f_7(x)$ a negative-binomial-Poisson mixture (NBPois), $f_8(x)$ a one-inflated Poisson distribution (OIPoiss), and $f_9(x)$ a one inflated negative-binomial distribution (OINegBi). In this example, $r = 9$.

We begin with the likelihood function for the counts for one tide, and extend it to the ensemble of counts for N tides Because for each tide $j, j = 1, 2, \ldots, N$ the data consisted of the number of 0's, 1's, etc..., the data can be represented as a multinomial sample with k categories and probabilities $\pi_1, \pi_2, \ldots, \pi_k$: Let Y_1 be the number of pairs with no satellites found at the beach in one tide, Y_2 the number of pairs with 1 satellite male in one tide, Y_3 the number of pairs with 2 satellite males in one tide, ..., Y_{k-1} the number of pairs with $k-2$ satellites in one tide and Y_k the number of pairs with $k-1$ or more satellites in one tide. Suppose for instance that we are to fit the Poisson distribution model with parameter λ to the counts of one tide. Then, the reduced parameter multinomial distribution arranged to fit the Poisson model would be parameterized using the following probabilities for each category:

$$\begin{aligned}
\pi_1 &= P(X=0) &&= f_1(0) &&= e^{-\lambda},\\
\pi_2 &= P(X=1) &&= f_1(1) &&= \lambda e^{-\lambda},\\
\pi_3 &= P(X=2) &&= f_1(2) &&= \tfrac{\lambda^2 e^{-\lambda}}{2!},\\
&\vdots\\
\pi_{k-1} &= P(X=k-2) &&= f_1(k-2) &&= \tfrac{\lambda^{k-2} e^{-\lambda}}{(k-2)!}\\
\pi_k &= P(X \geq k-1) &&= 1 - \textstyle\sum_{s=0}^{k-2} f_1(s) &&= 1 - \textstyle\sum_{s=0}^{k-2} \tfrac{\lambda^s e^{-\lambda}}{(s)!}.
\end{aligned} \tag{10}$$

It follows that if in a given tide j a total of n_j pairs are counted and $y_{j,1}$ is the number of females with no satellites, $y_{j,2}$ is the number with one satellites, etc., such that $\sum_{i=1}^{k} y_{j,k} = n_j$, the likelihood function needed to fit the Poisson probability model to the data of one tide is simply written as:

$$\begin{aligned}
L_j(\lambda) &= P(Y_{j,1} = y_{j,1}, Y_{j,2} = y_{j,2}, \ldots, Y_{j,k-1} = y_{j,k-1}, Y_{j,k} = y_{j,k})\\
&= \tfrac{n!}{y_{j,1}! y_{j,2}! y_{j,3}! \ldots y_{j,k}!} \pi_1^{y_{j,1}} \pi_1^{y_{j,2}} \ldots \pi_k^{y_{j,k}},
\end{aligned}$$

and the overall likelihood function for the N tides is simply

$$L(\lambda) = \prod_{i=1}^{N} L_j(\lambda).$$

Finally, note that for this reduced parameter multinomial model, the ML expected frequencies would simply be computed as $n_j \hat{\pi}_i$. For example, under the Poisson model, the expected number of 0's in a sample would be computed as $n_j \hat{\pi}_1 = \widehat{P(X=0)} = e^{-\hat{\lambda}}$, where $\hat{\lambda}$ denotes the ML estimate of λ.

The likelihood function and each of the predicted probabilities for every model were computed using the programs in the files `CrabsExampleTools.R` and `AbundanceToolkit2.0.R` downloadable from our github webpage, which works as follows. Suppose that for a single tide, the counts of the number of pairs with $0, 1, 2, 3, 4,$ and 5 or more satellites are $112, 96, 101, 48, 22, 16$, respectively. Then, the program `abund.fit` (found in the set of functions `AbundanceToolkit2.0.R`) takes those counts and returns, for every model in a pre-specified model set, the expected frequencies (from which the probabilities of every category in the reduced-parameter multinomial are retrievable), the ML estimates of each set of model parameters, the maximized log-likelihood and other statistics.

The processes of simulating any given number of tide counts according to Brockmann et al. (2018) and computing the ML estimates and other statistics for every model and every tide in a pre-specified model set are packaged within our function `short.sim()` whose output is (1) a matrix of simulated counts, with one row per tide. In each row the data for a single tide is displayed from left to right, showing the number of pairs with $0, 1, 2, 3, 4$, and 5 or more satellites. (2) a list with the statistics (ML estimates, maximized log-likelihood, predicted counts, etc...) for every model and every tide. (3) A matrix of information criteria values for every tide (row) and every model (column) in the set of tested models.

3.1.2. Calculation of Quantities Needed to Generate a MPMS

The generation of the MPMS necessitates solving the system of Equation (9). To solve that system of equations for any given dat set we need

1. A non-parametric estimate of the neg-selfentropy Sgg, $\widehat{Sgg}$. Berrett et al. (2019) recently proposed such an estimator. Their estimator is in essence a weighted (Kozachenko and Leonenko, 1987) estimator, and uses k nearest neighbors of each observation as follows:

$$H_n^w = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j \log \xi_{(j),i},$$

where $\xi_{(j),i} = (n-1)e^{-\psi(j)} V_q ||X_{(j),i} - X_i||^q$ with $X_{(j),i}$ indicating the j-th nearest neighbor from the i-th observation X_i. Also, in these equation n indicates the number of observations, $\psi(j)$ is the digamma function and $V_q = \pi^{q/2}/\Gamma(1+q/2)$ is the volume of the unit q-dimensional ball and q is the dimension of the multivariate observations.

The focus of Berrett et al. (2019) was writing a complete theoretical proof of the statistical properties of their estimator. Practical guidance as to how to find these weights is however lacking in their paper, but through personal communication with T. Berrett we learned that their weights w_j must only satisfy the constraints (see their Equation 2):

$$\sum_{j=1}^{k} w_j = 1 \quad \text{and} \quad \sum_{j=1}^{k} w_j \Gamma(j + 2l/q)/\Gamma(j) = 0 \quad \text{for} \\ l = 1, \ldots, \lfloor q/4 \rfloor,$$

where k is the number of observations that define a local neighborhood of observations around any given observation. Berrett (personal communication) recommends arbitrarily choosing k as the sample size to the power of a third. The other restrictions on Berrett et al. (2019)' theorem about the support of these weights were needed only for technical convenience for the proof. Berrett et al. (2019) also mentioned that for small sample sizes, the unweighted estimator may be preferable. For larger problems he recommended solving the above restrictions with a non-linear optimizator. We wrote such non-linear optimization routine to compute the weights w_j's and tested it extensively via simulations and embedded it into a function whose only argument is the data itself. Through extensive simulations we have verified that this routine works well for dimensions at least up to $q = 15$. We coded our optimization in R and is now part of a package of functions accompanying this paper. The function is found in the file `MPcalctools.R` and was named `Hse.wKL`. Finally, note that a typical data set for our crabs example is of dimension 6, so our routine is more than enough for a typical set of counts similar to the ones in this example. For instance, one set of counts of pairs with 0 satellites, 1 satellite, 2 satellites, ..., 5 or more satellite males for one tide is $y_1 = 112, y_2 = 96, y_3 = 101, y_4 = 48, y_5 = 22, y_6 = 16$.

2. A matrix of KL divergences between all models estimated in the model set being considered. If a total of r models are being considered, then the elements of this matrix are $\left\{KL(f_i, f_j)\right\}_{i,j}, i, j = 1, 2, \ldots, r$. Computing these divergences may seem like a daunting task, especially because these quantities are, in fact, different expectations (i.e., infinite sums) evaluated at the ML estimates for each model in the model set. However, those calculations are enormously simplified by adopting the general reduced-parameter likelihood approach because the neg-crossentropy $H(f_r, f_s)$ between two multinomial models f_r and f_s with a total sample size n can be computed exactly:

$$\begin{aligned} H(f_r, f_s) &= \sum\nolimits_{(y_1, y_2, \ldots, y_k) \geq 0, (\sum_k y_k) = n} \frac{n!}{y_1! \ldots y_k!} \pi_{1,r}^{y_1} \cdots \\ &\quad \pi_{k,r}^{y_k} \log \left[\frac{n!}{y_1! \ldots y_k!} \pi_{1,s}^{y_1} \ldots \pi_{k,s}^{y_k} \right] \\ &= \log n! + n \sum i = 1^k \pi_{i,r} \log \pi_{i,s} \\ &\quad - \sum_{i=1}^{k} \sum_{y_{i=0}}^{n} \binom{n}{y_i} \pi_{i,r}^{y_i} (1 - \pi_{i,r})^{(n-y_i)} \log \mathrm{y_i}!. \end{aligned} \quad (11)$$

Note that when $s = r$, then $H(f_r, f_s)$ becomes the neg-selfentropy. Because the KL divergence is the sum of a neg-selfentropy and a crossentropy, in practice, to compute the KL divergence between two count models for a single vector of counts for one tide we only needed to compute the probabilities in Equation (10) for every model using the ML estimates for each data set and use Equation (11) above. The function in `R` used to compute either the neg-crossentropies or the neg-selfentropies is named `H.multinom.loop()` and found in the file `MPcalctools.R`. Following simple rules of expected values, the overall KL divergence between two count models for a set of N vectors of tide counts, each drawn from the same true generating process (the individual-based model simulator program), was just computed as the sum of the divergences between the two models for each vector of counts. Note that the same simplification in Equation (11) applies to the computation of the neg-selfentropy for a multinomial distribution, a fact that we used to compute the true Sgg for our simulator algorithm, given that the individual-based model simulator of Brockmann et al. (2018) could be used to the estimate numerically true probabilities for 0,1,2,... satellites.

3. The estimates of the neg-crossentropies $\widehat{Sgf_i}$ and of $\widehat{Sf_ig}$ for $i = 1, 2, \ldots, r$. Although the first set of divergences, the $\widehat{Sgf_i}$, can be estimated either using the AIC and Equation (8), by definition of the KL divergence, the estimates $\widehat{Sf_ig}$ are in general not

equal to the estimates $\widehat{Sgf_i}$ and cannot be computed using the AIC and Equation (8). If however, h^2 is very small, then using the approximation $Sgf_i \approx \widehat{Sf_ig}$ works quite well as we show in example 3.2 and in Taper and Ponciano (2016a). Fortunately, using this approximation is not always necessary and does not have to be used for a large class of statistical problems. Indeed, for the example at hand where we are fitting multiple count models and for any other case where the likelihood function may be written by means of a reduced-parameter multinomial model (like the likelihood function for most phylogenetics models, for instance), both the $\widehat{Sgf_i}$ and the $\widehat{Sf_ig}$ can be computed using Equation (11) by using the ML estimates of the multinomial π's for each model and the ML estimates of the π's for the fully parameterized (i.e., the empirical model) in lieu of the π parameter values for g. We will denote these empirical estimates (i.e., the sample proportions) as $\bar{\pi}_i$. These estimates and Equation (11) can be used to compute $\widehat{Sgg}$. For a set of models and a data set including one or more tides, the estimates $\widehat{Sgf_i}$, $\widehat{Sf_ig}$, $\widehat{Sf_if_j}$ and $\widehat{Sgg}$ are computed using the function `entropies.matcalc()` found in the file `MPcalctools.R`.

For a simulated example where the data consisted of counts for 300 tides for which the first 5 tides were

```
> simdat[1:5,]
       0   1   2  3  4  5
[1,] 112  96 101 48 22 16
[2,] 135 125 108 44 19 12
[3,] 141 108  91 55 23 16
[4,] 119 117  99 60 18 10
[5,] 139 120 117 37 26 11
```

The estimated matrix of neg-crossentropies for these $N = 300$ tides was

```
$Sfifjs.hat
          Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
Poisson -4693.860 -6788.198 -7144.127 -7261.358 -7276.347 -7240.670 -7268.412 -4694.360 -5474.936
NegBin  -7140.595 -4801.609 -5748.393 -5402.616 -5393.795 -5141.644 -5417.371 -7142.133 -6271.978
ZIPoiss -7269.760 -5688.618 -4778.731 -4958.789 -4991.042 -5157.351 -4973.420 -7271.568 -6703.770
ZINegBi -7483.025 -5374.228 -4980.700 -4792.072 -4868.308 -5004.579 -4952.840 -7484.694 -6719.950
HurdNBi -7504.330 -5366.989 -5015.873 -4870.053 -4792.890 -4980.233 -4974.241 -7503.976 -6688.178
PoisNB  -7476.723 -5131.165 -5178.562 -5008.269 -4982.275 -4794.920 -4975.992 -7477.044 -6595.687
NBPois  -7453.540 -5389.715 -4984.999 -4949.951 -4969.751 -4970.038 -4790.289 -7455.437 -6702.133
OIPoiss -4694.328 -6789.643 -7145.910 -7262.925 -7275.909 -7240.883 -7270.251 -4693.826 -5474.358
OINegBi -5606.894 -6051.992 -6648.368 -6590.061 -6559.321 -6453.790 -6596.554 -5606.393 -4735.259
```

The true generating process neg-selfentropy, Sgg was -16.01199 and the estimated neg-selfentropy $\widehat{Sgg}$ was -15.96137. The real neg-crossentropies between the

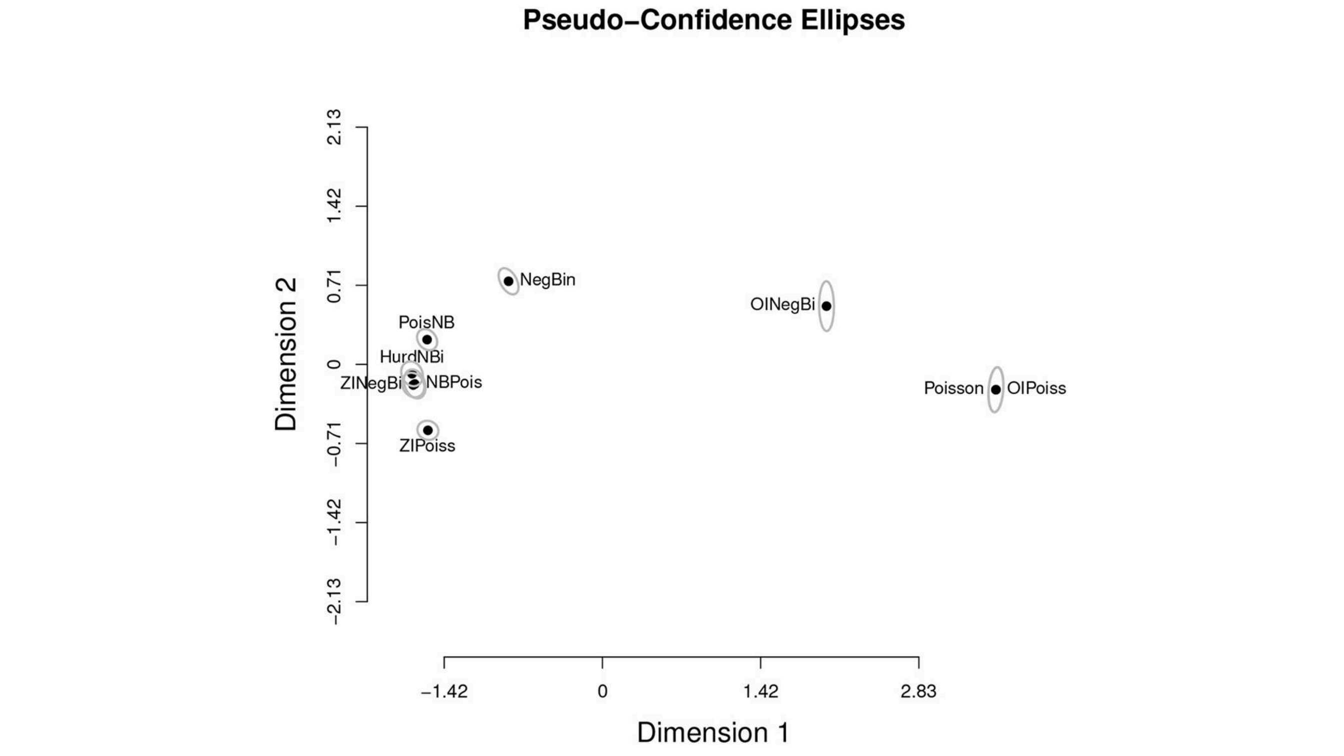


FIGURE 4 | Count models for the horseshoe crab example (section 3.1) in NMDS space, along with pseudo-confidence ellipses (95%). These ellipses are based on Stress derivatives (Mair et al., 2019) and indicate in this case that the NMDS is quite stable (overall stress is 0.029).

generating process and each of the models *Sgf_i*'s and *Sf_ig*'s were:

```
$Sgfis
  Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
-7601.606 -5603.842 -5485.607 -5432.819 -5457.657 -5450.109 -5463.949 -7606.407 -6832.673

$Sfisg
  Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
-7276.506 -5615.999 -5431.557 -5414.428 -5439.071 -5436.292 -5435.686 -7281.250 -6653.690
```

whereas the estimated neg-crossentropies were

```
$Sgfis.hat
  Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
-7891.890 -5652.049 -5403.773 -5200.907 -5179.331 -5323.294 -5253.094 -7891.705 -7051.753

$Sfisg.hat
  Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
-7638.034 -5682.275 -5396.418 -5213.190 -5193.804 -5339.449 -5264.447 -7637.716 -6906.558
```

4. The coordinates of every model in an NMDS space. Multidimensional scaling, MDS, is an established method (Borg et al., 2018) for representing the information in the $s \times s$ matrix D of distances/divergences among s objects as a set of coordinates for the objects in a $k-$dimensional euclidian space ($k \leq s$). If $k < s$, there may be some loss of information. MDS has two major varieties, metric multidimensional scaling, MMDS, in which D is assumed to be comprised of Euclidean distances, and non-metric multidimensional scaling, NMDS, in which D can be made up of divergences only monotonically related to distances. The MMDS projection can be made analytically, while the NMDS projection can only be found algorithmically by iteratively adjusting the configuration to minimize a statistic known as "Stress," which is a weighted average squared deviation of the distances between points (models in our case) calculated from the proposed configuration and the distances given in D.

   The matrix D required by NMDS should be symmetric. KL divergences are not, however, symmetric. The KL divergence can be reasonably symmetrized in a number of ways (Seghouane and Amari, 2007). We symmetrize using the arithmetic average of $KL(\theta_i, \theta_j)$ and $KL(\theta_j, \theta_i)$. As mentioned above in this problem we can directly calculate the symmetric KL. For other applications the symmetric KL can be estimated (up to the constant Sgg) using the KIC and its small sample version the KiCc (Cavanaugh, 1999, 2004). We follow Akaike in considering the KL divergence as a squared distance, and thus construct the matrix D from the square roots of the symmetrized KL divergence. We use the function smacofSym (De Leeuw and Mair, 2009) from the R package smacof (version 2.0, Mair et al., 2019) to calculate the NMDS. For the purposes of this paper we chose $k = 2$ so that we could have a graphical representation after augmenting the dimension to 3 to show the orthogonal distance from the generating process to its orthogonal projection M in the estimated plane of models. Nevertheless, the Stress of 0.029 indicates an excellent fit. Except for the very important aspect of visualization, dimension reduction is not an essential aspect of our method. Finally, the tight pseudo-confidence ellipses (95%) illustrated in **Figure 4**, based on Stress derivatives (Mair et al., 2019) indicate that this NMDS is quite stable.

Once all these components are computed, the system of Equation (9) can be solved with non-linear optimization. We coded such solution in the R function `MP.coords` found in the file *MPcalctools.R*. This function takes as input the estimated neg-crossentropies between all models, an estimate of *Sgg* or the neg-selfentropy of the generating process, and the vectors of estimated neg-crossentropies $\widehat{Sgf_i}$ and $\widehat{Sf_ig}$ to output the matrix of dimension $(r + 1) \times (r + 1)$ of symmetrized KL divergences, and the results of the NMDS with the coordinates of every model in a two-dimensional space, the estimated location of the orthogonal projection of g in such plane, M, and the estimate of h. Notably, this function works for any example for which these estimated quantities are available. Its output is taken by our function `plot.MP` to produce the three-dimensional representation of the Model Projection in Model Space shown in **Figure 5**. For this example, the estimated distances in the model projection space between all models, g and its projection M were

```
        Poisson  NegBin ZIPoiss ZINegBi HurdNBi  PoisNB  NBPois OIPoiss OINegBi       M
NegBin  4.46711
ZIPoiss 5.09358 1.51616
ZINegBi 5.21422 1.26069 0.42784
HurdNBi 5.22444 1.19699 0.52347 0.09773
PoisNB  5.10858 0.89561 0.81309 0.42542 0.33690
NBPois  5.19883 1.24348 0.43208 0.01798 0.09139 0.41228
OIPoiss 0.00181 4.46859 5.09530 5.21588 5.22609 5.11018 5.20050
OINegBi 1.69500 2.85332 3.73875 3.76818 3.75926 3.58746 3.75125 1.69618
M       4.51235 2.33280 1.26566 1.67451 1.76140 1.97444 1.67387 4.51416 3.50306
g       4.51235 2.33280 1.26566 1.67451 1.76140 1.97444 1.67387 4.51416 3.50306 0.00032
```

whereas the real distances (because we knew what the simulation setting was) were

```
> dist(true.MP$XYs.mat)
         Poisson  NegBin ZIPoiss ZINegBi HurdNBi  PoisNB  NBPois OIPoiss OINegBi       M
NegBin  4.46711
ZIPoiss 5.09358 1.51616
ZINegBi 5.21422 1.26069 0.42784
HurdNBi 5.22444 1.19699 0.52347 0.09773
PoisNB  5.10858 0.89561 0.81309 0.42542 0.33690
NBPois  5.19883 1.24348 0.43208 0.01798 0.09139 0.41228
OIPoiss 0.00181 4.46859 5.09530 5.21588 5.22609 5.11018 5.20050
OINegBi 1.69500 2.85332 3.73875 3.76818 3.75926 3.58746 3.75125 1.69618
M       4.14688 2.14942 1.34587 1.70959 1.78455 1.93970 1.70493 4.14868 3.12059
g       4.14688 2.14942 1.34587 1.70959 1.78455 1.93970 1.70493 4.14868 3.12059 0.00037
```

From these matrices, it is readily seen that the real value of *h* in the model projection space was 0.000372 whereas its corresponding estimated *h* value is 0.000323. A quick calculation yields the distances between the true location of the orthogonal projection *M*, its estimate, the true location of *g* and its estimate:

```
           hat.m    hat.g   true.m
hat.g  0.000323
true.m 0.383074 0.383074
true.g 0.383074 0.383074 0.000372
```

As expected, variation in the quality of these estimates and the difference with the true locations changes from simulated dat set

```
> AICs
  Poisson    NegBin   ZIPoiss   ZINegBi   HurdNBi    PoisNB    NBPois   OIPoiss   OINegBi
14301.321  9823.639  9327.086  8923.355  8880.203  9168.129  9027.727 14302.951 12625.047
> delta.is
   Poisson     NegBin    ZIPoiss    ZINegBi    HurdNBi     PoisNB     NBPois    OIPoiss    OINegBi
5421.11814  943.43554  446.88328   43.15146    0.00000  287.92594  147.52444 5422.74769 3744.84389
```

to simulated data set. Two questions are a direct consequence of this observation: first, the MPMS data representation in **Figure 5** could be more accurately depicted via bootstrap and confidence clouds or spheres for the location of each model in model space could be drawn. Such task would however involve entertaining the problem of the representation of multiple bootstrap NMDS runs in a single space, using the same rotation.

Classically, variation among NMDS object has been estimated only after Procrustes rotation has oriented the various coordinate systems for maximal similarity among the NMDS objects (see Mardia et al., 1979). A long series of articles involving authors, such as T. M. Cole, S. R. Lele, C. McCulloch, and J. Richtsmeir demonstrates that this approach is deeply flawed. This work is summarized in the monograph by Lele and Richtsmeier (2001). The problem is that the apparent variability among equivalent points in the multiple objects depends on distance from the center of rotation. Lele and Richtsmeir argue that inference is better made regarding variation in estimated distances between points than on the coordinates of points. A mean distance matrix can be estimated from a set of bootstrapped replicates, and it is almost certain that the mean distance matrix will be the most informative matrix both for inference and for graphical purposes as this mean corresponds to the expectation with respect to $\hat{\theta}$ in Akaike's 5th insight (see section 2.1.5). Further, variation and covariation in all estimated distances and contrasts of distances can be invariantly calculated and used for inference. Finally, extending our MPMS methodology to include confidence bounds for our estimates is a topic of current research in our collaboration and will be treated in a future manuscript because it necessitates the same degree of care used to generate confidence intervals for Model-Average inferences (see for instance Turek, 2013).

The second question has to do with how would our estimate of the location in model space of the orthogonal projection of the generating process compare to the location of the model-average. For our example at hand, the AIC values as well as the Δ AICs were:

To compare the estimated location of the model average with our estimated our model projection, we plotted both panels in **Figure 6** into a single, two-dimensional figure with: the location of every estimated model, the location of the model averaged coordinates using the AIC weights, the location of the estimated orthogonal projection of *g*, and the location of the true location of the orthogonal projection *g*. Such figure is presented in **Figure 6**. In this figure, the distance between the real projection *M* of *g* and our estimated projection is 0.383074 whereas the distance between the model-average and the real projection of *g* is 1.784555. A quick inspection of **Figure 6** shows that this case in fact, is a real-life illustration of the point brought up by **Figure 3B**. When the geometry of the model space is as in **Figures 3B**, **5**, **6**, model averaging may not be a suitable enterprise.

3.2. An Ecosystems Ecology Application

Here we discuss a worked example highlighting the strengths of the model projections approach to multi-model inference. This example was originally presented in Taper and Ponciano (2016a) which is freely downloadable from:

```
https://link.springer.com/book/10.1007/978-3-319-
27772-1
```

This example is an analysis of data simulated from a structural equation model (SEM) based on a study by Grace and Keeley (2006). Simulation from a known model in necessary to

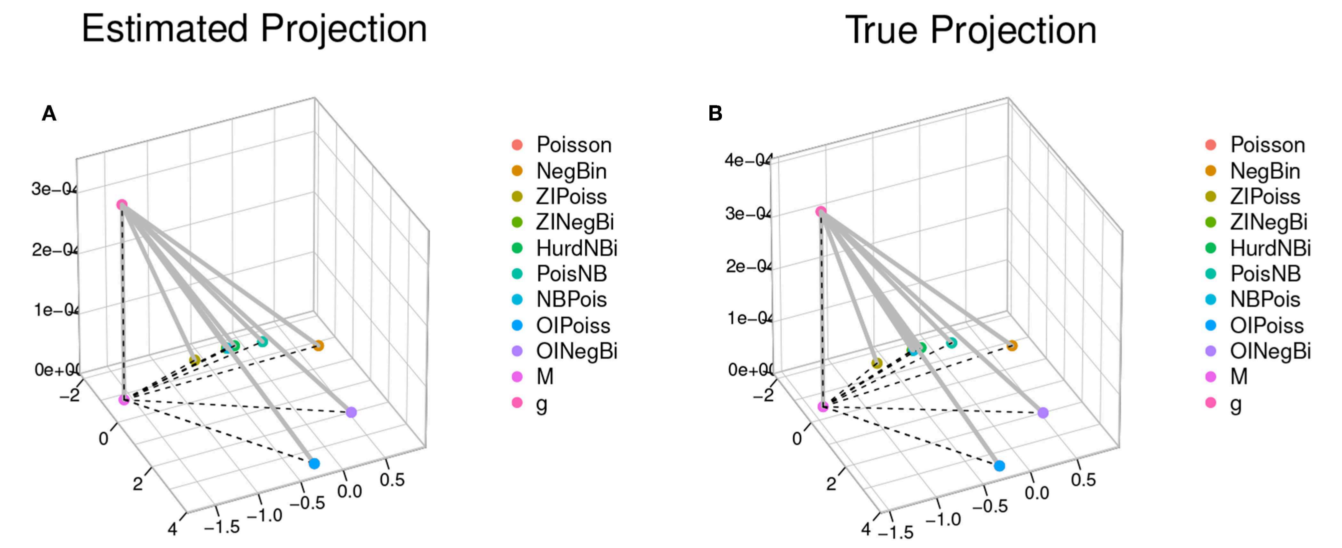


FIGURE 5 | The models of **Figure 2** visualized by our new methodology, and applied to our Horseshoe crab example (section 3.1). As before, *g* is the generating model and models $f_1, \ldots, f_9$, are the approximating models and named in the legend of each panel. **(A)** Shows the estimated model projection "M" and the estimated location of the true generating process whereas **(B)** shows the location of the true model projection "M" and of the true generating process. The dashed lines are KL distances between approximating models, which were calculated according to Equation 2. The solid gray lines are the KL distances from approximating models to the generating model. The vertical dotted line shows h, the discrepancy between the generating model and its best approximation in the NMDS plane, whereas all the other dotted lines mark the discrepancy between the approximating models and the model projection "M." A 2-dimensional representation of only the plane of models, the estimated *g* model projection and the true model projection of *g* onto that plane is shown in **Figure 6**.

understand how well our methods capture information about the generating process, while basing that model on published research guarantees that our test-bed is not a toy, but is a problem of scientific interest. SEM is a flexible statistical method that allows scientists to analyze the causal relationship among variables and even general theoretical constructs (Grace and Bollen, 2006, 2008; Grace, 2008; Grace et al., 2010). Grace and Keeley (2006) analyzed the development of plant diversity in California shrublands after natural fires. Structural equations models were used to make inferences as to the causal mechanisms influencing changes in diversity. Plant composition at 90 sites was followed for 5 years. The Grace and Keely final model is displayed in **Figure 7**. To summarize the causal influences, species richness is directly affected by heterogeneity, local abiotic conditions, and plant cover. Heterogeneity and local abiotic conditions are in turn affected by landscape position, but total cover is only directly affected by burn intensity. Burn intensity is in turn only affected by stand age, which itself depend on landscape position. Affects and their direction are shown as arrows in the figure. The strength of affects (i.e., the path coefficients) are shown both as numbers on the figure and as the thickness of the arrows).

Forty-one models were fit to our generated data. The models ranged from underfitted to overfitted approximations of the generating process. The actual generating model was not included in this model set. Using this set of fitted models, we estimated a 2-d Non-Metric Dimensional Scaling model space as discussed above. The calculated stress was tiny (0.006%) indicating almost all higher dimensional structure is captured by an $\mathcal{R}^2$ plane. A mapping of the estimated space analogous to our **Figure 6** is shown in their Figure 6 Taper and Ponciano (2016a). ΔAIC values are indicated by color. As in **Figure 6** of this paper, on this map of model space we also indicated: (1) The estimated projection (location) of the generating process to the 2-d NMDS space, (2) The Akaike weighted model averaged location and 3) The actual projection of the true generating process l onto the 2-d manifold (in this worked example this can be done because we have simulated from a known model).

Two important observations can be made based on the graph in **Figure 6** (both in this manuscript and in Taper and Ponciano, 2016a) : First while there is a rough agreement between proximity to the generating process and ΔAIC values, this relationship is not as tight as one might naively expect. The inter-model KL distances do have substantial impact on the map. Second, using our methods and just like in example 3.1 above, the estimated projection of the generating process is somewhat nearer to the actual projection of the generating process than the location produced by model-averaging (**Figure 6** in this manuscript).

Figure 8 demonstrates the sensitivities of both the estimated projection and model average of eliminating fitted models from the estimation of the NMDS space. We repeatedly eliminate the left-most model in the model set and reestimate the space after each cycle. With each model elimination, the model-averaged location moves toward the right. On the other hand, the estimated projection stays near its original location, even after all fitted models in that side of the map have been eliminated. Conversely, eliminating from the right, the model average shifts to the left as anticipated. Under right-side model elimination,

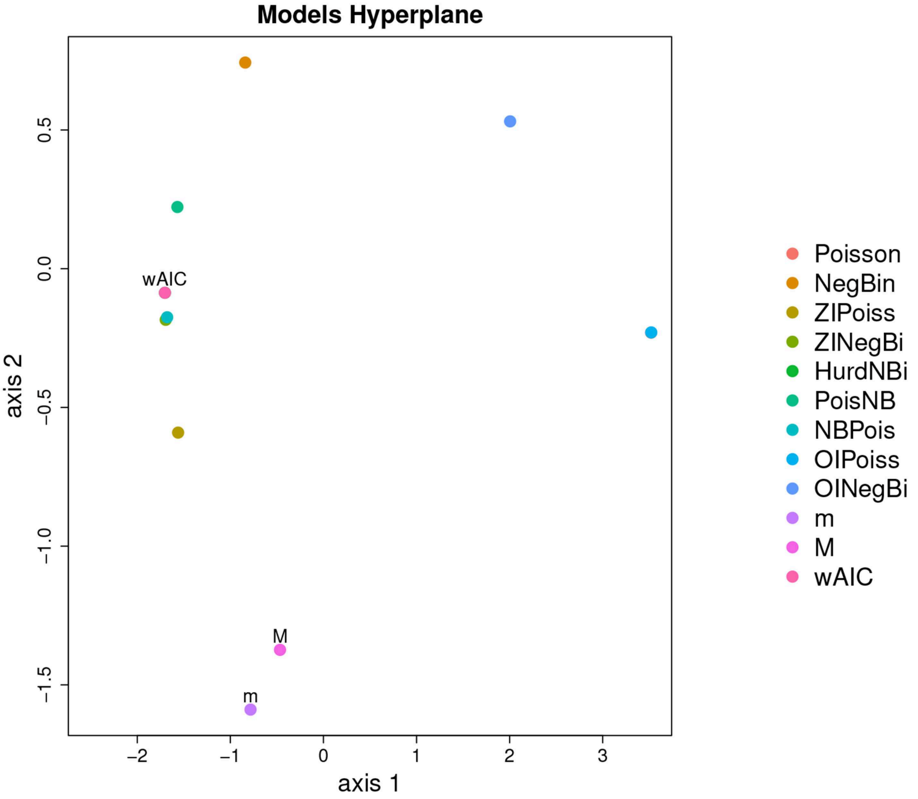


FIGURE 6 | NMDS space of nine models for the Horseshoe crab example (section 3.1). The true projection, *M*, of the generating model onto the NMDS plane is shown, along with the location of the estimated location of such projection, *m*, and of the model average, *wAIC*.

the model projection is somewhat more variable than under elimination from the other direction.

This model elimination example illuminates differences in the two kinds of estimates the generating process location. These differences follow directly from the geometric development of the AIC by Akaike, and from the mathematics of model averaging. (1) The model average must fall inside of the bounds of the fitted models. changing the model set will, except in contrived cases, change the model average. (2) Because it is a projection, our method's estimate of the generating process' location can fall outside the bounds of the model set. And (3), because of the nature of projection geometry, farther models can inform the estimated situation of the generating process in the NMDS map. Point (3) is demonstrated in the discrepancy in the stability of the model projection location under model elimination from the left and model elimination from the right. There are several models with high influence that are deleted quickly under model elimination from the right that stay in the model set much longer under elimination from the left.

Our approach calculates two important diagnostic statistics not even thought of in model averaging. The first is measure of the dispersion of the generating process. This is the negselfentropy or *Sgg*. In this example it is calculated to be −9.881, very close to the known magnitude of −9.877. The second statistic is an estimate of the perpendicular distance of the generating process to the NMDS manifold (h in Equation 9). This diagnostic is critical for proper interpretation of your model set. If the generating process is far from NMDS manifold, then any statistic based on models in the model set is likely to be inaccurate. Using our approach we calculate h to be 0.0002. The known h is $6e - 08$.

3.3. Testing the Non-parametric Estimation of *Sgg*

To exemplify the independent estimation of *Sgg* with a data set we simulated samples from a seven-dimensional multivariate normal distribution and compared the true value of *Sgg* with its non-parametric estimate according to Berrett et al. (2019). We chose to simulate data from a multivariate normal distribution because its *Sgg* value is known analytically. When the dimension of a multivariate normal distribution is p and is variance-covariance matrix is Σ, then

$$Sgg = -\frac{1}{2}\ln\left\{(2\pi e)^{p}\det(\Sigma)\right\}. \qquad (12)$$

To carry our test, we chose five testing sample sizes $10, 25, 50, 75, 150$, and for each sample size we simulated 2,000 data sets according to a multivariate normal distribution with $p = 7$ and $\Sigma = I$, and computed each time Berrett et al.'s non-parametric estimate. The resulting estimates, divided by the true value of 9.93257 are plotted as boxplots in **Figure 9**.

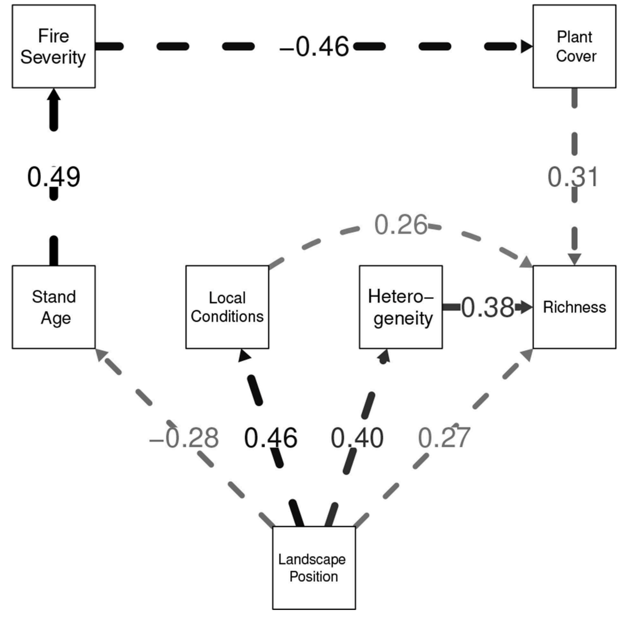


FIGURE 7 | The final, simplified model explaining plant diversity from Grace and Keeley (2006). Arrows indicate causal influences. The standardized coefficients are indicated by path labels and widths. See section 3.2 for details. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

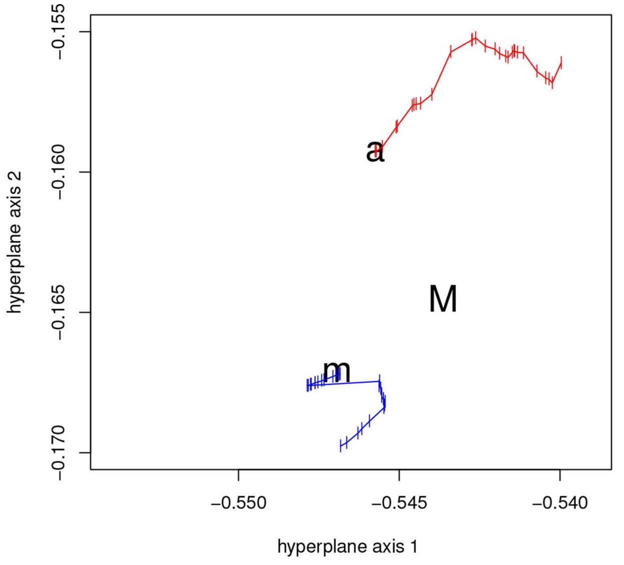


FIGURE 8 | Stability test of the displacement (trajectories) of the model prediction (in blue) and the model average (in red) under deletion of 1 − 30 models. *M* denotes the true location of the orthogonal projection of the generating model in the hyperplane. *m* and *a* mark the location of the model projection and the model average, respectively, when the 30 models are used. In both cases, as models are removed one by one from the candidate model set, the location of both *m* and *a* changes (little vertical lines). Note how the model projection estimate is more stable to changes in the model set than the model average. Prasanta S. Bandyopadhyay, Gordon Brittan Jr., Mark L. Taper, Belief, Evidence, and Uncertainty. Problems of Epistemic Inference, published 2016 Springer International Publisher, reproduced with permission of Springer Nature Customer Service Center.

4. DISCUSSION

We have constructed a novel approach to multi-model inference. Standard multi-model selection analyses only estimate the relative, not overall divergences of each model from the generating process. Typically, divergence relationships amongst all of the approximating models are also estimable (dashed lines in **Figure 5**). We have shown that using both sets of divergences, a model space can be constructed that includes an estimated location for the generating process (the point *g* in **Figure 5**). The construction of such model space stems directly from a geometrical interpretation of Akaike's original work.

The approach laid out here has clear and substantial advantages over standard model identification and Bayesian based model averaging. A heuristic approach aiding the development of novel models is now possible by simply being able to visualize a set of candidate models in an Euclidean space. Now the overall architecture of model space vis-a-vis the generating process is statistically estimable. Such architecture is composed of a critical set of quantities and relationships. Among these objects, we now include the estimated coordinates of the closest orthogonal generating model projection onto the manifold of candidate models (the point *M* in **Figure 5**). Second, the estimated magnitude of the total divergence between the truth and its orthogonal projection onto the manifold of models can give the analyst an indication of whether important model attributes have been overlooked.

In the information criterion literature and all scientific application, the neg-selfentropy *Sgg* of the generating process is simply treated as an unknown quantity. In fact, it can be estimated quite precisely as our example shows. *Sgg* is itself of great interest because with it the overall discrepancy to the generating process becomes estimable. Because this quantity is estimable, now the analyst can discern the overall quality and proximity of the model set under scrutiny. Thus, our approach solves a difficulty that has long been recognized (Spanos, 2010) but yet treated as an open problem.

Studying the model space architecture gives the information to correct for misleading evidence (the probability of observing data that fails to support the best model), accommodation (over-fitting), and cooking your models (Dennis et al., 2019). The scaffolding from which to project the location of the generating process is estimated can be rendered more robust simply by considering more models. This is an interesting result that we expect will later contribute to the discussion of data dredging. On the other hand, non-identifiability and weak estimability (Ponciano et al., 2012) are, of course, still a problem, but at least the model space approach will clearly indicate the difficulties.

As conceived here, model projection is an evidential alternative (Taper and Ponciano, 2016b) to model averaging

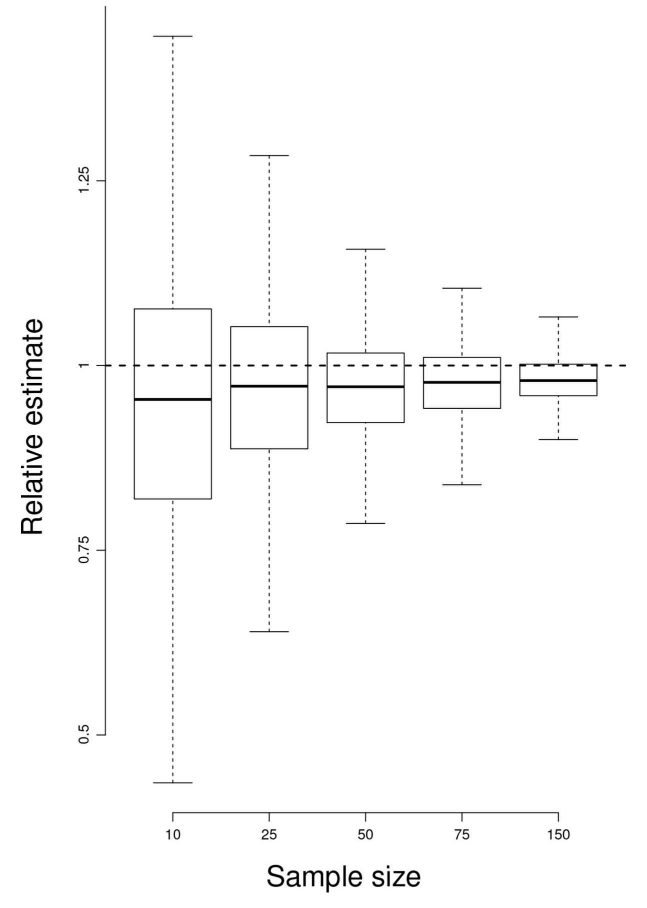


FIGURE 9 | Boxplots of sets of 2,000 non-parametric estimates of *Sgg* (from Berrett et al., 2019) relative to the true *Sgg* value of 9.93257, for different sample sizes. The simulated data comes from a seven-dimensional Multivariate Normal distribution with means equal to 10 and the identity matrix as a variance-covariance matrix. The dashed, horizontal line at 1 shows the zero-bias mark.

using Akaike weights (or other Bayesian alternatives) because it incorporates the available information estimated by many models without the redundancy inherent in model averaging. Through model projection the analyst can use more of the information available but usually ignored. Furthermore, our methodology provides new important diagnostic statistics previously not considered by model averaging: *Sgg* and *h*. As we showed in our results, model projection is not as sensitive as model average to the composition of the set of candidate models being investigated. Model averaging appears to artificially favor redundancy of model specification: the more models are developed in any given region of model space, the stronger this particular region gets weighted during the model averaging process. Finally, an emergent pattern in the analysis is that the optimization problem of our model projection methodology can be used to project outside the bounds of the available model set whereas the model averaging methodology, by definition, cannot.

As well as proposing solutions to existing problems, any new method also raises a variety of technical problems that need to be solved. This is certainly the case with the model projection approach presented here.

Our methodology bears a near-model limitation that, although important, is shared with the usage of Akaike's Information Criterion. Our exposition makes it clear that near model requirement is due to the imperfect yet useful approximation employed by Akaike while setting $\phi \approx \pi/2$ (see **Figure 1**). It was only thanks to this approximation that Akaike was able to solve for the estimable divergence contrasts between all approximating models and the generating process. This approximation breaks down in curved model spaces as the divergence from the generating process increases. Indeed, as the KL distance between approximating models and the generating model increases, $-AIC/2n$ becomes an increasingly biased and variable estimate of the Sgf component of the KL distance between the approximating model and the generating model. This effect is strong enough that sometimes very bad models can have very low ΔAIC scores, sometimes even as low as the minimum score. The TIC (Takeuchi, 1976) and the EIC2 (Konishi and Kitagawa, 2008; Kitagawa and Konishi, 2010) are model identification criteria designed to be robust to model misspecification. Substituting one of these information criteria for the AIC in constructing the matrix of inter-model divergences should allow the use of models more distant from truth than is acceptable using the AIC.

Our methodology focuses on estimation of the model space geometry but uncertainties around such estimation are not fully worked out as of yet. Work in progress by Taper, Lele, Ponciano and Dennis, the estimation of the uncertainties associated with doing inference with evidence functions, such as ΔSIC scores, can be assessed *via* non-parametric bootstrap techniques. We expect bootstrap to be also useful to reduce the variance of information criterion's bias correction (Kitagawa and Konishi, 2010).

We think that this model projection methodology should be the starting point to do a careful, science-based inquiry of what are the model attributes that make a model a good model. Knowing the location of the projected best model is an essential component of our multi-model development strategy because a response surface analysis can reveal what model attributes tend to be included near the location of the projected best model thus aiding in the construction of a model closer to the best projection.

DATA AVAILABILITY STATEMENT

This paper's code is available at https://github.com/jmponciano/ModelsProjection.

AUTHOR CONTRIBUTIONS

JP and MT contributed equally to the conceptualization, development, programing, analysis, writing, and figure creation.

FUNDING

Both authors were partially funded via a grant from JSPS KAKENHI Grant Number 25340110 (to Shimatani Kenichiro) and from the Risk Analysis Center in the Institute of Statistical Mathematics. JP was partially supported by two grants from the US National Institute of Health grant number to University of Florida NIH R01 GM103604 to JP and NIH 1R01GM117617 to Profs. Jason K. Blackburn and JP.

ACKNOWLEDGMENTS

The authors appreciate the feedback and insights from the audience at the International Symposium What is a good model?: Evidential statistics, information criterion and model evaluation held January 2016 at the Institute of Statistical Mathematics in Tokyo, Japan (the home institute of Hirotugu Akaike). We also appreciate the insightful comments from the audiences from Virginia Tech's Statistics Department Seminar and University of Idaho's Applied Statistics Seminar, specially to Drs. Leah Thompson and Bill Price.

REFERENCES

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. Petrov, and F. Csaki (Budapest: Akademiai Kiado), 267–281.

Akaike, H. (1974). A new look at statistical-model identification. *IEEE Trans. Autom. Control* 19, 716–723.

Bandyopadhyay, P. S., Brittan G. Jr., and Taper, M. L. (2016). *Belief, Evidence, and Uncertainty Problems of Epistemic Inference.* Basel: Springer International Publisher.

Berrett, T. B., Samworth, R. J., Yuan, M. (2019). Efficient multivariate entropy estimation via *k*-nearest neighbour distances. *Ann. Stat.* 47, 288–318. doi: 10.1214/18-AOS1688

Borg, I., Groenen, P. J., and Mair, P. (2018). *Applied Multidimensional Scaling and Unfolding*. Cham: Springer.

Brockmann, H. J. (1990). Mating behavior of horseshoe crabs, limulus polyphemus. *Behaviour* 114, 206–220.

Brockmann, H. J., St Mary, C. M., and Ponciano, J. M. (2018). Discovering structural complexity and its causes: breeding aggregations in horseshoe crabs. *Anim. Behav.* 143, 177–191. doi: 10.1016/j.anbehav.2017.10.020

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociol. Method Res.* 33, 261–304. doi: 10.1007/b97636

Burnham, K. P., Anderson, D. R., and Huyvaert, K. P. (2011). Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav. Ecol. Sociobiol.* 65, 23–35. doi: 10.1007/s00265-010-1029-6

Casquilho, J. P., and Rego, F. C. (2017). Discussing landscape compositional scenarios generated with maximization of non-expected utility decision models based on weighted entropies. *Entropy* 19:66. doi: 10.3390/e19020066

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Probab. Lett.* 42, 333–343.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on kullback's symmetric divergence. *Austr. N. Zeal. J. Stat.* 46, 257–274. doi: 10.1111/j.1467-842X.2004.00328.x

Cushman, S. A. (2018). Calculation of configurational entropy in complex landscapes. *Entropy* 20:298. doi: 10.3390/e20040298

Davison, A. C. (2003). *Statistical Models*, Vol. 11. Cambridge, UK: Cambridge University Press.

De Leeuw, J. (1992). "Introduction to akaike (1973) information theory and an extension of the maximum likelihood principle," in *Breakthroughs in Statistics*, eds S. Kotz, and N. L. Johnson (London: Springer), 599–609.

De Leeuw, J., and Mair, P. (2009). Multidimensional scaling using majorization: Smacof in R. *J. Stat. Softw.* 31, 1–30. doi: 10.18637/jss.v031.i03

Dennis, B., Ponciano, J., Taper, M., and Lele, S. (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Evol.* 7:372. doi: 10.3389/fevo.2019.00372

Fan, Y., Yu, G., He, Z., Yu, H., Bai, R., Yang, L., et al. (2017). Entropies of the chinese land use/cover change from 1990 to 2010 at a county level. *Entropy* 19:51. doi: 10.3390/e19020051

Grace, J. B. (2008). Structural equation modeling for observational studies. *J. Wildl. Manage.* 72, 14–22. doi: 10.2193/2007-307

Grace, J. B., Anderson, T. M., Olff, H., and Scheiner, S. M. (2010). On the specification of structural equation models for ecological systems. *Ecol. Monogr.* 80, 67–87. doi: 10.1890/09-0464.1

Grace, J. B., and Bollen, K. A. (2006). *The Interface Between Theory and Data in Structural Equation Models*. Reston, VA: US Geological Survey.

Grace, J. B., and Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: the role of composite variables. *Environ. Ecol. Stat.* 15, 191–213. doi: 10.1007/s10651-007-0047-7

Grace, J. B., and Keeley, J., E. (2006). A structural equation model analysis of postfire plant diversity in California shrublands. *Ecol. Appl.* 16, 503–514. doi: 10.1890/1051-0761(2006)016[0503:ASEMAO]2.0.CO;2

Gravel, D., Massol, F., and Leibold, M. A. (2016). Stability and complexity in model meta-ecosystems. *Nat. Commun.* 7:12457. doi: 10.1038/ncomms12457

Kitagawa, G., and Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Ann. Stat. Math.* 62:209. doi: 10.1007/s10463-009-0237-1

Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York, NY: Springer Science & Business Media.

Kozachenko, L., and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Pered. Inform.* 23, 9–16.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.

Kuricheva, O., Mamkin, V., Sandlersky, R., Puzachenko, J., Varlagin, A., and Kurbatova, J. (2017). Radiative entropy production along the paludification gradient in the southern taiga. *Entropy* 19:43. doi: 10.3390/e19010043

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., et al. (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett.* 7, 601–613. doi: 10.1111/j.1461-0248.2004.00608.x

Lele, S. R., and Richtsmeier, J. T. (2001). *An Invariant Approach to Statistical Analysis of Shapes*. Boca Raton, FL: Chapman and Hall/CRC.

Mair, P., Groenen, P., and De Leeuw, J. (2019). More on multidimensional scaling and unfolding in R: smacof version 2. *J. Stat. Softw.* [Epub ahead of print].

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Statistics*. San Diego, CA: Academic Press.

Milne, B. T., and Gupta, V. K. (2017). Horton ratios link self-similarity with maximum entropy of eco-geomorphological properties in stream networks. *Entropy* 19:249. doi: 10.3390/e19060249

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Syst. Biol.* 61, 955–972. doi: 10.1093/sysbio/sys055

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury advanced series. Belmont, CA: Duxbury Press.

Roach, T. N., Nulton, J., Sibani, P., Rohwer, F., and Salamon, P. (2017). Entropy in the tangled nature model of evolution. *Entropy* 19:192. doi: 10.3390/e19050192

Seghouane, A.-K., and Amari, S.-I. (2007). The aic criterion and symmetrizing the Kullback–Leibler divergence. *IEEE Trans. Neural Netw.* 18, 97–106. doi: 10.1109/TNN.2006.882813

Spanos, A. (2010). Akaike-type criteria and the reliability of inference: model selection versus statistical model specification. *J. Econometr.* 158, 204–220. doi: 10.1016/j.jeconom.2010.01.011

Takeuchi, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Math. Sci.* 153, 12–18.

Taper, M. L., and Ponciano, J. (2016a). "Book appendix. projections in model space: multi-model inference beyond model averaging," in *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference*, eds P. S. Bandyopadhyay, G. G. Brittan, and M. L. Taper (Springer), 157–173.

Taper, M. L., and Ponciano, J. M. (2016b). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

Turek, D. (2013). *Frequentist model-averaged confidence intervals* (Ph.D. thesis), University of Otago, Dunedin, New Zealand.

Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. B Biol. Sci.* 267, 109–116. doi: 10.1098/rspb.2000.0974

Yang, Z., and Zhu, T. (2018). Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1854–1859. doi: 10.1073/pnas.1712673115

Zeng, Q., and Rodrigo, A. (2018). Neutral models of short-term microbiome dynamics with host subpopulation structure and migration limitation. *Microbiome* 6:80. doi: 10.1186/s40168-018-0464-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 14 November 2019
doi: 10.3389/fevo.2019.00427

Incorporating Parameter Estimability Into Model Selection

Jake M. Ferguson [1]*, ***Mark L. Taper*** [2,3], ***Rosana Zenil-Ferguson*** [1], ***Marie Jasieniuk*** [4] **and** ***Bruce D. Maxwell*** [5]

[1] Department of Biology, University of Hawaii at Manoa, Honolulu, HI, United States, [2] Department of Ecology, Montana State University, Bozeman, MT, United States, [3] Department of Biology, University of Florida, Gainesville, FL, United States, [4] Department of Plant Sciences, University of California, Davis, Davis, CA, United States, [5] Land Resources and Environmental Sciences, Montana State University, Bozeman, MT, United States

We investigate a class of information criteria based on the informational complexity criterion (ICC), which penalizes model fit based on the degree of dependency among parameters. In addition to existing forms of ICC, we develop a new complexity measure that uses the coefficient of variation matrix, a measure of parameter estimability, and a novel compound criterion that accounts for both the number of parameters and their informational complexity. We compared the performance of ICC and these variants to more traditionally used information criteria (i.e., AIC, AICc, BIC) in three different simulation experiments: simple linear models, nonlinear population abundance growth models, and nonlinear plant biomass growth models. Criterion performance was evaluated using the frequency of selecting the generating model, the frequency of selecting the model with the best predictive ability, and the frequency of selecting the model with the minimum Kullback-Leibler divergence. We found that the relative performance of each criterion depended on the model set, process variance, and sample size used. However, one of the compound criteria performed best on average across all conditions at identifying both the model used to generate the data and at identifying the best predictive model. This result is an important step forward in developing information criterion that select parsimonious models with interpretable and tranferrable parameters.

Keywords: informational complexity, ICOMP, AIC, BIC, variable selection, covariance, coefficient of variation, prediction

Edited by:
Diego P. Ruiz,
University of Granada, Spain

Reviewed by:
Luis Gomez,
University of Las Palmas de Gran Canaria, Spain
Konstantinos Demertzis,
Democritus University of Thrace, Greece

Correspondence:
Jake M. Ferguson
jakeferg@hawaii.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *22 February 2019*
Accepted: *23 October 2019*
Published: *14 November 2019*

Citation:
Ferguson JM, Taper ML, Zenil-Ferguson R, Jasieniuk M and Maxwell BD (2019) Incorporating Parameter Estimability Into Model Selection. Front. Ecol. Evol. 7:427. doi: 10.3389/fevo.2019.00427

1. INTRODUCTION

It is through models that scientists continually refine their descriptions of nature (Giere, 2004; Taper, 2004; Pickett et al., 2010; Taper and Lele, 2011). Scientists interpret models as descriptions of observations, as representations of causal processes, or as predictions of future observations. Often scientists test a set of probabilistic models representing alternative hypotheses. A critical scientific goal is identifying reliable methods to determine the best predictive model, or set of models, among the candidates. Prediction has emerged as a primary goal for many ecological applications (Dietze et al., 2018) but commonly used information criterion have been shown to be inadequate in many ecological applications (Link and Sauer, 2016; Link et al., 2017).

To most profitably select among set of models we should be able to measure the evidence of each model relative to others (Lele, 2004; Taper and Lele, 2011). Model selection criteria do this by ranking a set of models based on their relative ability to achieve a specific goal. Two common goals of model selection are the minimization of the approximation error and the minimization of the prediction error (Taper, 2004), corresponding to two principal functions of modeling, explanation, and prediction (Cox, 1990; Lele and Taper, 2012).

Most commonly used model selection criteria apply asymptotic theory developed under the assumption of large sample sizes (Bozdogan, 1987; Cavanaugh, 1997; Burnham and Anderson, 2002). This has led to criteria that are easily calculated from standard regression output; however, the criterion's effectiveness may be limited when applied to sets of complex models with low sample sizes, such as those often encountered in ecological inference. Relatively few studies have tested how well selection criterion can deal with such scenarios, but work by Hooten (1995) and Ward (2008) have tested the ability of criteria to answer questions about nonlinear animal population dynamics, while Murtaugh (2009) looked at how different model selection techniques affected predictability across nine different ecological datasets.

The discrepancy between an estimated probability distribution and the true underlying distribution can be partitioned into two terms. The first, termed the model discrepancy, is due to limitations in model formulation while the second, termed the estimation discrepancy, arises due to difficulties in estimation (Bozdogan, 1987). The model discrepancy arises from how close the approximating model is to the data generating mechanism, given the best possible parameter values. The second quantity, called the estimation discrepancy, arises from the poor estimation of model parameters. An extreme example of poor estimability is parameter non-identifiability (e.g., when parameters only occur in fixed combinations, such as sums or products) leading to complete correlation or collinearity. Although this is an extreme example and not likely to appear in a well-considered model, there are various degrees of collinearity in models and not all strong collinearities are obvious (e.g., Polansky et al., 2009; Ponciano et al., 2012).

Collinear parameters will be unstable to small changes in the data (Schielzeth, 2010; Freckleton, 2011), thus affecting the interpretability of estimates. Collinearity also impacts to the generality of a model by affecting the ability to make reliable out-of-sample predictions (Brun et al., 2001; Dormann et al., 2013), the interpretability of model-averaged coefficents (Cade, 2015), and the ability to transferrable parameters estimated from one context to another (Yates et al., 2018). This final property is especially desirable for generating estimates that will be useful for fields that rely on parameterizing complex model using estimates pulled from the literature [e.g., food web ecology (Ferguson et al., 2012) and epidimiology (Ruktanonchai et al., 2016)]. Bozdogan and Haughton (1998) showed that the performance standard information criterion can significantly decline in the presence of collinearity.

We argue that when dealing with complex models, estimation accuracy should be considered in measures of model quality because accuracy is necessary to correctly interpret parameter estimates, make reliable predictions, and to use estimated parameters in new scientific settings—three common goals of scientific practice. Below, we discuss previous work that incorporates measures of parameter interdependency into model selection criterion. We use this to motivate a new class of information criteria that incorporates measures of interdependency into traditional forms of information criterion. We test the ability of new and existing information criteria over three model sets of increasing complexity, looking at selection behavior in each model set over different levels of process variability and sample size.

1.1. Introduction to Information Criterion

In ecology, primarily due to the influential work of Burnham and Anderson (2002), attention has focused on estimating the Kullback-Leibler divergence as a measure of model discrepancy. Akaike (1974) measured this discrepancy by minimizing the cross-entropy between the model distribution, $m(x)$, and the true distribution, $t(x)$. The difference between the entropy of a distribution and the cross-entropy is called the Kullback-Leibler (KL) divergence. This measures the amount of information lost about $t(x)$ when using $m(x)$ to approximate. The KL divergence is given by $D_{\mathrm{KL}}(t, m) = \mathrm{E}_{\mathrm{t(x)}}\left[\ln(t(x))\right] - \mathrm{E}_{\mathrm{t(x)}}\left[\ln(m(x))\right]$.

Increasing values of D_{KL} are interpreted as poorer approximations of the model $m(x)$ to $t(x)$ (Burnham and Anderson, 2002). In a typical application we don't know the true underlying distribution, $t(x)$. However, when making relative comparisons between two or more approximating models we do not need to consider the first term of the KL divergence, the entropy of the true distribution, as this is the same for all models in the comparison and is eliminated in the contrast between models. Differences between models therefore only depend only on the second term, the cross-entropy. Akaike (1974) showed that if the model is sufficiently close to the generating process, twice this cross-entropy term could be estimated in what has become called Akaike's Information Criterion:

$$\mathrm{AIC} = -2\ln(L(\hat{\theta})) + 2k. \tag{1}$$

Here, $L(\theta)$ is the likelihood function of the pdf $m(x)$, evaluated at the maximum likelihood parameter values, $\hat{\theta}$, and k is the number of parameters in the model (including estimates of variance parameters).

Given a set of AIC values, we declare the parameterized model with the lowest value to have the minimum estimated KL divergence from the generating process and therefore to be most similar to it. Because AIC values all lack the unknown self-entropy term in the KL divergence, they are often presented as a contrast between a given model and the best model in the set. This measure is often denoted as the ΔAIC value. Values of $\Delta\mathrm{AIC} > 2$ indicate there is some evidence for the model with the lower value relative to the other model, while models with of $\Delta\mathrm{AIC} < 2$ are considered to be indistinguishable (Taper, 2004) (see Jerde et al. in this issue for a more fine-grained discussion on the strength of evidence).

While the use of the AIC has flourished in ecological modeling, there are several important properties of the AIC that are not well known to ecologists. For example, Nishii (1984) and Dennis, Ponciano, Taper and Lele (submitted this issue) showed that in linear models the AIC has a finite probability of overfitting even when the sample size is large. Thus the AIC is not statistically consistent. However, the AIC does minimize the mean squared prediction error in linear models as sample size increases, making it asymptotically efficient for prediction (Shibata, 1981), a property that does not require the generating model to be in the set.

Many other criteria have been developed which are similar in form to the AIC. These criteria are composed of a goodness of fit term, based on the log-likelihood, and a penalty term, based on some measure of the model complexity. For the AIC in Equation (1) this penalty is the number of parameters. The AICc is a small sample bias correction to the AIC derived under the assumption of a standard regression model with the sampling distribution of the estimated parameters normally distributed around the true parameter values (Hurvich and Tsai, 1989). The AICc is given by $\mathrm{AICc} = -2\ln(L(\hat{\theta})) + 2k + \frac{2k(k+1)}{n-k-1}$. Here, n, is the sample size and k is the number of estimated parameters. Like the AIC, this criterion is not consistent but it is asymptotically efficient with linear models (Shibata, 1981; Hurvich and Tsai, 1989).

The Schwarz information criterion or BIC (Schwarz, 1978) (also sometimes called the SIC), is used to estimate the marginal likelihood of the generating model, a quantity often used in Bayesian model selection. Originally derived under a general class of priors the BIC is given by $\mathrm{BIC} = -2\ln(L(\hat{\theta})) + k\ln(n)$. The BIC is consistent, in that it will asymptotically choose the model closest to truth (in the Kullback-Leibler sense). However, the BIC is not asymptotically efficient, an important difference between it and the AIC and AICc (Aho et al., 2014). Finally, the BIC* (also sometimes called HBIC or the HIC) (Haughton, 1988) is an alternative derivation of the BIC and a slightly weaker penalty that may serve as a useful compromise between the AIC and BIC, $\mathrm{BIC}{*} = -2\ln(L(\hat{\theta})) + k\ln(n/2\pi)$. This criterion is thought to have greater efficiency than the BIC at higher sample sizes while still being consistent. This allows the criterion to balance underfitting and overfitting errors.

The informational complexity criterion, or ICC, developed by Bozdogan (2000) examines a different kind of complexity than the previously described methods. In the ICC the number of parameters, k, is not considered to be a full characterization of a models complexity. Instead, ICC seeks to capture dependencies among model parameters. The approach applies an information-based covariance complexity term (van Emden, 1969), in addition to the cross-entropy term used in the AIC. The ICC constructs its penalty term from the trace and the determinant of the parameter covariance matrix $\boldsymbol{\Sigma}$, characterizing complexity through measures of parameter redundancy and estimation instability. ICC is given by

$$\mathrm{ICC}(\boldsymbol{\Sigma}) = -2\ln(L(\hat{\theta})) + 2\mathrm{C}(\boldsymbol{\Sigma}), \tag{2}$$

where $\mathrm{C}(\boldsymbol{\Sigma})$ has replaced k, the number of parameters, in the AIC. The complexity penalty, $\mathrm{C}(\boldsymbol{\Sigma})$, takes into account not just the number of parameters but also the degree of interdependence among parameters, measured using the covariance matrix of the estimated parameters, $\boldsymbol{\Sigma}$.

1.2. Deeper Into C(Σ)

According to Bozdogan (2000), the "complexity of a system (of any type) is a measure of the degree of interdependency between the whole system and a simple enumerative composition of its subsystems or parts." Intuitively, this means that the more complex a system is, the more information is needed to reconstruct the whole from the constituent components. A mathematical realization of this definition can be realized by measuring the mutual information between the joint sampling distribution $(s(\theta_1\,\theta_2,\ldots,\theta_k))$ and the product of marginal sampling distributions $(s(\theta_1)s(\theta_2)\cdots s(\theta_k))$. The mutual information is

$$I(\theta_1\,\theta_2,\cdots,\theta_k) = \mathrm{E}\left[\ln\left(\frac{s(\theta_1\,\theta_2,\ldots,\theta_k)}{s(\theta_1)s(\theta_2)\cdots s(\theta_k)}\right)\right], \tag{3}$$

where the expectation is taken over the joint distribution.

Equation (3) is a measure of the information shared between the estimated parameters. It is zero, corresponding to no complexity penalty, when parameter estimates are all independently distributed and increases with increased covariation between parameters. Assuming the estimated parameters follow a multivariate normal distribution leads to a form of this mutual information that can be readily calculated. Because normality is an asymptotic property of maximum likelihood estimation, the assumption is valid in many settings. Equation (3) then simplifies to the van Emden complexity, given by $\mathrm{C}_{\mathrm{vE}}(\boldsymbol{\Sigma}) = \frac{1}{2}\sum_i^k \ln(\boldsymbol{\sigma}_i^2) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}|)$. Here, σ_i denotes the standard error of the i^{th} parameter estimate. diagonal elements of the estimated parameters covariance matrix, $\boldsymbol{\Sigma}$, for each of the k parameters. The determinant of this matrix is noted as $|\boldsymbol{\Sigma}|$. This quantity measures the amount of information lost when parameter estimates are assumed to be independent.

The van Emden complexity is not invariant to rotations of the parameter space; therefore Bozdogan maximized this quantity over all possible orthonormal parameter transformations (Bozdogan, 2000). The maximal complexity is $\mathrm{C}_{\max}(\boldsymbol{\Sigma}) = \frac{k}{2}\ln\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{k}\right) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}|)$. In this study we examined penalties based on both $\mathrm{C}_{\mathrm{vE}}(\boldsymbol{\Sigma})$ and $\mathrm{C}_{\max}(\boldsymbol{\Sigma})$ complexity terms as they behave differently and previous work has suggested that both may be useful (Clark and Troskie, 2008). We differentiate the ICC (Equation 2) that use these different complexity measures using the notation $\mathrm{ICC}_{\mathrm{vE}}(\boldsymbol{\Sigma})$ and $\mathrm{ICC}_{\max}(\boldsymbol{\Sigma})$.

An illustration of the complexity measures in **Figure 1** for a two-dimensional covariance matrix gives the qualitative behavior of both complexity terms. Both terms increase as the magnitude of the correlation increases, however, the increase in the van Emden complexity is independent of the variance while the maximal complexity is a non-monotonic function of the variance. The maximal complexity is minimized when the relative variance terms are equal, and increases when one variance term diverges from the other. Thus, the maximal complexity can actually increase with increases in precision of parameter estimates, a property that may not be desirable.

In order to apply the penalties $\mathrm{C}_{\mathrm{vE}}(\boldsymbol{\Sigma})$ and $\mathrm{C}_{\max}(\boldsymbol{\Sigma})$ to real data we use the estimated covariance matrix, $\hat{\boldsymbol{\Sigma}}$. The parameter covariance matrix is extractable from the output of virtually all estimation packages. If parameters are estimated through direct optimization, optimization routines typically report an approximate Hessian. The inverse of the Hessian matrix is an

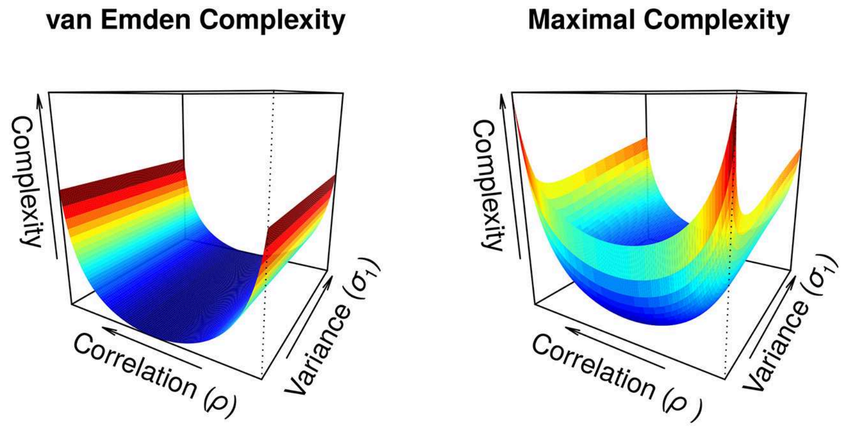


FIGURE 1 | The van Emden complexity, $C_{\mathrm{vE}}(\boldsymbol{\Sigma})$, and maximal complexity, $C_{\max}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$ and $\sigma_2 = 1$.

TABLE 1 | Invariance of complexity forms to different linear transformations of the covariates.

	Additive transforms	Multiplicative transforms	Rotational transforms
$C_{\mathrm{vE}}(\boldsymbol{\Sigma})$	✓	✓	
$C_{\max}(\boldsymbol{\Sigma})$	✓		✓
$C_{\mathrm{vE}}(\boldsymbol{\Psi})$		✓	
$C_{\max}(\boldsymbol{\Psi})$		✓	✓

approximation of the covariance matrix. Thus, $C_{\mathrm{vE}}(\hat{\boldsymbol{\Sigma}})$ and $C_{\max}(\hat{\boldsymbol{\Sigma}})$ can be easily calculated from output given by standard statistical packages such as R (R Core Team, 2015), which typically will report an approximate Hessian matrix that can be used to estimate the approximate covariance matrix by solving for the inverse of the matrix. For the small variance covariance matrices explored here solving for the inverse matrix is fast (less than 1 s) and the matrix would have to get quite large for the calculation time to be noticible, on the order of thousands of parameters. Other methods to estimate the covariance matrix such as using the least squares estimators or by bootstrapping estimates of the covariance matrix could also be applied.

ICC is not a scale-invariant penalty and transformations of the data may yield different model selections. Another form of ICC calculates the complexity penalty as a function of the correlation matrix, denoted as **R**, rather than of the covariance matrix (Bozdogan and Haughton, 1998). However, this quantity does not incorporate information about the precision of parameters estimates, as the variance terms are not present.

To overcome the limitations of the current form of scale-invariant ICC, we introduce a new variant based on a complexity measure that uses the coefficient of variation matrix (Boik and Shirvani, 2009). This matrix is independent of scale, but it retains information on the relative precision of the parameter estimates. The coefficient of variation matrix is defined as the covariance matrix scaled by the vector of parameter estimates in such a way that the diagonals are the squared coefficients of variation. This is a matrix with entries defined as $\Psi_{i,j} = \frac{\mathrm{Cov}(\theta_1,\theta_2)}{\theta_1\theta_2}$. Applying penalties of the form $C_{\mathrm{vE}}(\boldsymbol{\Psi})$ may be desirable because the matrix $\boldsymbol{\Psi}$ is invariant to multiplicatively rescaling the covariates but is still sensitive to the relative magnitude of coefficient uncertainty. Scale invariance means that going from one unit to another for a specific covariate, e.g., meters to kilometers, does not affect the inference. $C_{\mathrm{vE}}(\boldsymbol{\Psi})$ and $C_{\max}(\boldsymbol{\Psi})$ are sensitive to additive transformations, thus, shifting all measurement units by a constant factor will lead to different inferences. A table summarizing the properties of the different forms of informational complexity is given in **Table 1**.

2. METHODS

2.1. Incorporating Parameter Estimability Into Information Criterion

The standard ICC does not penalize increasing complexity in a manner that leads to asymptotically consistent model selections (Nishii, 1988). Bozdogan and Haughton (1998) proposed a consistent form of the ICC that scaled the complexity parameter, $C_{\max}(\Sigma)$ by the log of the sample size, however this criterion did not perform well in their simulation experiments. Therefore, we propose a new compound selection criterion that is the sum of two divergences in order to develop a consistent form of criteria. The first divergence is a model parsimony measure. The second divergence, $C_{-}(\cdot)$, gives a measure of parameter estimability, a useful property for model interpretability and prediction. This criterion is defined as $\mathrm{IC} + C_{-}(\cdot) \equiv -2\ln(L(\hat{\theta})) + kf(n) + 2C_{-}(\cdot)$. The first piece of this criterion measures the goodness of fit through the maximum log-likelihood, the second piece measures model complexity, where $f(n,k)$ is a function of the sample size and possibly of the number of parameters. The final piece, $C_{-}(\cdot)$, measures the parameter estimability. The exact form of this compound criterion depends on both the choice of the model parsimony criterion as well as the choice of the parameter complexity measure which regulates the strength of the penalty based on the complexity of the parameter.

Our motivation for using two divergences in this compound criterion is that we believe accounting for both goodness of fit and parameter estimability when using finite datasets will better reflect the underlying complexity and usefulness of the model. These compound criteria also deal with a critical issue in the ICC that yield a penalty of zero when parameters are orthogonal. Given that there are a number of measures of both goodness of fit and parameter complexity, we tested several different forms of the compound criterion. The forms we tested were $\mathrm{AIC}+2\mathrm{C}_{-}(\cdot)$, $\mathrm{AICc} + 2\mathrm{C}_{-}(\cdot)$, $\mathrm{BIC} + 2\mathrm{C}_{-}(\cdot)$, and $\mathrm{BIC}^{*} + 2\mathrm{C}_{-}(\cdot)$ where $\mathrm{C}_{-}(\cdot)$ can be $\mathrm{C}_{\mathrm{vE}}(\hat{\boldsymbol{\Sigma}})$, $\mathrm{C}_{\mathrm{vE}}(\hat{\boldsymbol{\Psi}})$, $\mathrm{C}_{\max}(\hat{\boldsymbol{\Sigma}})$, or $\mathrm{C}_{\max}(\hat{\boldsymbol{\Psi}})$.

2.2. Performance Comparisons With Simulation Studies

We conducted simulation studies that tested the capabilities of all 25 of the simple and compound information criteria discussed above under different conditions. We compared the behavior of model selection criteria using three attributes:

1. Selection: how frequently the criterion identifies the generating model.
2. Prediction: how well models selected by a criterion can predict new observations.
3. KL approximation: how well criterion values estimated KL divergence between model and truth.

The first two attributes reiterate the primary goals of model selection described in the introduction. The third attribute addresses the ability of a model to determine the relative KL divergence, a question of only collateral interest to practitioners interested in the application of model selection techniques to scientific problems. However, the estimated KL divergence may be a useful proxy for similarity to the generating model. In addition, much of the development and discussion of model identification criteria in ecology is framed around the estimation of the KL divergence as a metric between model and truth (for a review on other possible metrics see discussion in Lele, 2004).

We quantified attribute 1, the ability of a criterion to determine the generating model by counting the percentage of time that each criterion selected the generating model in our simulations. We measured attribute 2, a criterion's predictive ability, by its prediction sum of squares (PRESS) given by $\mathrm{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{-i}\right)^2$. Here $\hat{y}_{-i}$ is the predicted value at the i^{th} data point, which is omitted when fitting the model to the data, and y_i is the true, unobserved i^{th} value (Allen, 1974). A low PRESS value indicates that the criterion chooses a model that gives low prediction errors for in-sample prediction.

To calculate attribute 3, the frequency that criteria selected the minimum KL divergence between the j^{th} model and the generating distribution, we used the formula for the divergence between normal distributions (Bozdogan, 1987). We determined the agreement of each criterion with the true minimum KL divergence by calculating the frequency that the criterion selected the model with the minimum KL divergence.

To better understand the properties of these model selection criterion under a broad range of conditions we performed our simulation experiments with a set of linear models and two sets of nonlinear models of ecological interest. The linear model simulations varied the strength of correlation present in the design matrix as well as sample size and process variance. The first nonlinear model set examined time series models of population dynamics, while the second examined highly nonlinear models of barley yield. In both nonlinear model sets sample size and process variance were varied as well as the generating model. Using these different model sets, we sought to define criterion performance over the range of model complexity found in the ecological literature.

2.3. Linear Models

Our linear regression simulation experiment follows a design based on previous simulation studies by Bozdogan and Haughton (1998), Clark and Troskie (2006), and Yang and Bozdogan (2011). These studies explored the application of the ICC criterion under differing levels of correlation among explanatory variables. Correlation in explanatory variables is likely to be common in ecological covariates, causing the performance of the AIC to suffer (Bozdogan and Haughton, 1998).

We generated a 7 parameter design matrix by transforming 8 randomly drawn standard normal random variables, $Z \sim N(0, 1)$, using the relationships:

$$X_{i,j} = \begin{cases} \sqrt{1-\alpha_1^2}Z_{i,j} + \alpha_1 Z_{i,8} \text{ for } j = 1, 2, 3 \quad i = 1, 2, \ldots, n \\ \sqrt{1-\alpha_2^2}Z_{i,j} + \alpha_2 Z_{i,8} \text{ for } j = 4, 5, 6, 7 \quad i = 1, 2, \ldots, n. \end{cases}$$

Where $X_{i,j}$ is the ith entry of the jth covariate. The α_1 and α_2 values control the degree of correlation present among the elements of the design matrix. The covariates generated by using this procedure have a covariance between row j and row k given by,

$$\mathrm{Cov}(X_j, X_k) = \begin{cases} \alpha_1^2 & \text{for } j = 1, 2, 3, \quad k = 1, 2, 3 \\ \alpha_2^2 & \text{for } j = 4, 5, 6, 7, \quad k = 4, 5, 6, 7 \\ \alpha_1\alpha_2 & \text{for } j = 1, 2, 3, \quad k = 4, 5, 6, 7. \end{cases}$$

The covariate parameters, β, were generated from the maximum eigenvector of the matrix $\mathbf{X}'\mathbf{X}$ following Bozdogan and Haughton (1998).

We generated data at three different levels of collinearity, low ($\alpha_1 = 0.3$, $\alpha_2 = 0.7$), medium ($\alpha_1 = 0.9$, $\alpha_2 = 0.9$), and high ($\alpha_1 = 0.99$, $\alpha_2 = 0.99$); three levels of sample size, low ($n = 20$), medium ($n = 50$), and high ($n = 100$); and three levels of variability, low ($\sigma^2 = 0.25$), medium ($\sigma^2 = 1$), and high ($\sigma^2 = 2.5$). We simulated 100 datasets from the rank 5 model at each level of collinearity, sample size, and variance. We then repeated each set of 100 simulations 10 times to estimate a mean and standard error of each selection attribute.

We then fit all models of ranks (1–7) to the generated data using the BFGS algorithm in the optim routine in the R statistical language. Hessian matrices were calculated using the Hessian function in the numDeriv package (Gilbert and Varadhan, 2016).

We checked convergence of the optimization by checking that all eigenvalues of the Hessian matrix were positive. For each simulated dataset we calculate the information criteria of each fitted model and determined how well the criteria performed in our three selection attributes.

We determined performance at each level of collinearity, sample size, and variance by averaging the statistic of interest (e.g., the number of correct generating model selections) over all of the simulation study parameters except the level of interest. For example, to determine performance at the low collinearity level we averaged the performance statistic of interest over all sample sizes and variances that had design matrices with low collinearity.

TABLE 2 | Forms of density dependence used in the population dynamics study.

Model	**Functional form ($\boldsymbol{f(N_t)}$)**
Exponential	rN_t
Ricker	$rN_te^{bN_t}$
Theta-Ricker	$rN_te^{bN_t^{\theta}}$
Gompertz	$rN_te^{b\ln(N_t)}$
Beverton-Holt	$\frac{rN_t}{1+bN_t}$
Hassell	$\frac{rN_t}{(1+bN_t)^{\theta}}$

The intrinsic population growth rate is given by the parameter r, while b is the strength of density dependence. The degree of compensation in the Theta-Ricker and Hassell models is controlled by θ.

2.4. Population Dynamics Models

Dynamical time series models are a common applied modeling technique for forecasting future ecological conditions, a major goal of ecological modeling (Clark et al., 2001). Applications of time series models include forecasting fisheries stocks (Lindegren et al., 2010) and assessing extinction risk (Ferguson and Ponciano, 2014). Measuring the strength of evidence among a set of forecast models is critical for generating reliable predictions, but it's known that many nonlinear dynamical models yield correlated parameter estimates (Polansky et al., 2009). These correlations may impact the performance of traditional information criterion (Bozdogan, 1990). Here, we study the properties of information criterion in a set of nonlinear dynamical models.

The population dynamics simulation experiment used time series models to describe the projected population abundance in the next year given the abundance in the current year. In order to ensure that population dynamics were realistic, we generated data based on parameter estimates made from the Global Population Dynamics Database (GPDD) to simulate data (NERC, 2010). The GPDD contains approximately 5000 time series related to plant and animal index measurements. We used a subset of these studies, chosen for their length and the indicated data quality following the methods described in more detail in Ferguson and Ponciano (2015) and Ferguson et al. (2016). We only used time series with a length of at least 15 samples and a GPDD reliability rating of 3–5. The reliability rating is a qualitative measure of data quality made by the database authors. These quality standards left us with 391 time series to generate data from.

We examined six density-dependent models encompassing a wide range of functional forms. All models were of the form, $N_{t+1} = rN_tf(N_t)$ where $f(N_t)$ can take one of the commonly used functional forms of density dependence given in **Table 2**. These functional forms represent different hypotheses about the strength of density dependence. As in the linear model design, we examined a range of sample sizes ($n = 25$, $n = 50$, $n = 100$) and low, medium, and high process variances (see below for how we calculated these variances).

In order to determine realistic levels of variance to use in our simulations, we fit an additive, normally distributed environmental variance model to the population growth rate (pgr), where pgr $= \ln\left(\frac{N_{t+1}}{N_t}\right)$. To determine realistic levels of environmental variation, we fit the pgr to a linear model (corresponding to Gompertz density dependence in **Table 2**) for all 391 time series. Optimization and convergence checks were performed on the pgr using the same methods described in the linear model section. We then used the 10, 50, and 90% quartiles of the estimated environmental variance over all time series to determine the low, medium, and high variance levels used in the simulations.

To simulate data, we first fit each of the density dependence model to each of the 391 GPDD datasets. We then simulated a new dataset from each fitted model at each level of sample size and variance, repeating this process for all of the density dependence models in **Table 2**. We repeated this process for every possible generating model, sample size, and variance combination, repeating the whole procedure 10 times to obtain a standard error for the model selection attributes. We averaged criteria performance over sample size, variance level, and generating model to examine the average selection rate for a given factor of interest. We did not need to vary the correlations between parameters in this experiment as in the linear models because the nonlinear model structure induces correlations between parameters.

2.5. Barley Yield Models

Bioeconomic modeling is an increasingly important application of ecological modeling (Grafton et al., 2017). Here, we examined the selection properties information criterion applied to a set of crop-weed competition models. These models explain crop yield (Y) as a function of crop (D_c) and weed (D_w) density, as well as the relative difference in time to emergence (T) between crop and weed. Here we examined our ability to accurately select the correct barley yield model from a set of candidate models. The nine models considered for this simulation experiment are a subset from a previous study (Jasieniuk et al., 2008) that used the ICC. These models have more complex forms than the population dynamics models used above, as well as more parameters and more covariates. Thus, this model set is a step up in complexity from the population dynamics models explored in the previous section. The nine models used in this simulation study are defined in **Table 3**. We refer readers to the original study by Jasieniuk et al. (2008) for further motivation for these models.

TABLE 3 | Functional forms of the models used for the barley yield simulations.

Functional form	Fitted parameters	Observed variables
$Y = R_c D_c \left(1 - \frac{R_w D_w}{e^{-cT} + a_w D_w}\right)$	$R_c, R_w, a_w, c, \sigma^2$	D_c, D_w, T
$Y = \frac{R_c D_c}{1 + a_c D_c} \left(1 - \frac{R_w D_w}{1 + a_w D_w}\right)$	$R_c, R_w, a_c, a_w, \sigma^2$	D_c, D_w
$Y = R_c D_c \left(1 - \frac{R_w D_w}{1 + a_w D_w}\right)$	R_c, R_w, a_w, σ^2	D_c, D_w
$Y = R_c D_c$	R_c, σ^2	D_c
$Y = \frac{R_c D_c}{1 + a_c D_c + a_w D_w}$	R_c, a_c, a_w, σ^2	D_c, D_w
$Y = \frac{R_c D_c}{1 + a_w D_w}$	R_c, a_w, σ^2	D_c, D_w
$Y = \frac{R_c D_c}{1 + a_w D_w e^{-cT}}$	R_c, a_w, c, σ^2	D_c, D_w, T
$Y = \frac{R_c D_c}{1 + \frac{a_w D_w e^{-cT}}{1 + b D_w}}$	R_c, a_w, c, b, σ^2	D_c, D_w, T
$Y = R_c D_c e^{-i D_w e^{-cT}}$	R_c, i, c, σ^2	D_c, D_w, T

Y is the crop yield response. The covariates are, R_c, the observed crop density, R_w, the observed weed density, and T, the observed relative emergence time between the crop and weeds. Estimated parameters are, D_c, the slope of the increase in crop yield with increasing crop density below the asymptote, D_w, the slope of the proportional yield loss as weed density approaches 0, a_c, the maximum expected crop yield, a_w, the asymptotic maximum proportional yield loss at high weed densities, and c, the relative time of emergence between crop and weed is scaled.

We generated datasets by first fitting each of the models to the dataset from the Bozeman 1994 dataset reported in Jasieniuk et al. (2008). We simulated new datasets by adding a normal random noise term to the log of the empirically predicted response using data from Jasieniuk et al. (2008). We examined three sample size levels ($n = 25, n = 50, n = 125$) and three variance levels ($\sigma^2 = 0.5\hat{\sigma}^2, \sigma^2 = \hat{\sigma}^2, \sigma^2 = 4\hat{\sigma}^2$), where $\hat{\sigma}^2$ was the empirically estimated variance of the observed data under the given generating model. We generated 100 simulated datasets for each model in **Table 3** at each sample size and variance level. As before, we averaged over sample size, variance level, and generating model to examine the average selection rate for a given factor of interest. We repeated each set of simulations 10 times to order to estimate the mean and standard error of the selection statistics. We only performed the PRESS calculation on one set of simulations due to the length of time it took to do this calculation. Therefore, there is no standard error associated with prediction for these models.

Due to these models presenting a more difficult optimization problem than the other model sets, we modified our fitting procedure. From an initial set of parameters, we applied the Nelder-Mead optimization algorithm (also known as the downhill simplex method) followed by the BFGS method to maximize the log-likelihood function. The simplex method was run first because it is robust, although it converges slowly. This two-step process provided the initial parameter estimates for the quasi-Newton method, which converges relatively quickly near a maximum. We repeated this procedure for 100 random initial points and chose the parameters associated with the maximum likelihood value, and convergence was determined as previously described.

3. RESULTS

Here, we will focus on presenting the criteria that performed best under one or more of our experimental conditions. Figures of performance for all criterion under all experimental conditions are presented in the **Supplementary Material**.

3.1. Linear Models

We present the overall criterion performance averaged over all conditions, along with standard errors for the linear model simulations in **Figure 2**. The best criterion at selecting the generating model on average was the AICc (**Table 4**), the best at prediction was also the AICc (**Table 5**), and the best at selecting the minimum KL divergence was AICc+2C$_{\max}(\mathbf{\Sigma})$ (**Table 6**). The ICC tended to be the worst performers at all selection goals (**Figure 2**), however the ICC$_{\max}(\mathbf{\Psi})$ tended to behave similarly to the AIC and the BIC*. We also see in **Figure 2** that the average performance of the criterion for all selection goals was strongly correlated but the PRESS and KL minimum selection was nearly completely correlated. Several of the compound criteria performed well with AICc+2C$_{\max}(\mathbf{\Sigma})$, AICc+2C$_{\text{vE}}(\mathbf{\Sigma})$, and AICc+2C$_{\text{vE}}(\mathbf{\Psi})$ performing nearly as well as AICc for all performance attributes.

In general, criteria performed better as sample size increased and variance decreased as expected (**Supplementary Figures S1–S6**). In most trials some form of the compound criterion performed better than traditional criterion (**Figure 2**). However, performance differences among most of the criteria differed only by a few percentage points and the difference in top performers was within the range of the performance uncertainty (**Figure 2**).

3.2. Population Dynamics Models

We present the overall criterion performance for the population dynamics simulation experiments averaged over all conditions, along with standard errors, in **Figure 3**. While the class of ICC criteria performed poorly in the linear model selections, here they tended to perform as well as or better than the traditional criteria. While the performance of all selection goals in the linear models simulations were strongly correlated, here they differed. The variation in the performance of the ability to select the generating model was much greater than for the other selection goals, though the compound criteria did tend to perform better than both traditional criteria and the ICC.

Out of the ICC the ICC$_{\max}(\mathbf{\Psi})$ tended to perform as good as, or better than the other forms. The best criterion at selecting the generating model overall was the BIC+2C$_{\max}(\mathbf{\Sigma})$ (**Table 4**), the best at prediction was the BIC+2C$_{\max}(\mathbf{\Sigma})$ (**Table 5**), and the best at selecting the minimum KL divergence was the AICc (**Table 6**).

In general, criteria performed better at selecting the generating model and the KL minimum as sample size increased and variance decreased, as expected, however, the ability to select the minimum PRESS model actually declined in the traditional criteria with sample size (**Supplementary Figures S7–S12**). Additionally, we found that some form of the compound criterion tended to perform better than the traditional criterion for all selection goals with BIC+2C$_{\max}(\mathbf{\Sigma})$ performing best at selecting the generating model the best predictive model. However, the AIC and AICc tended to dominate the performance

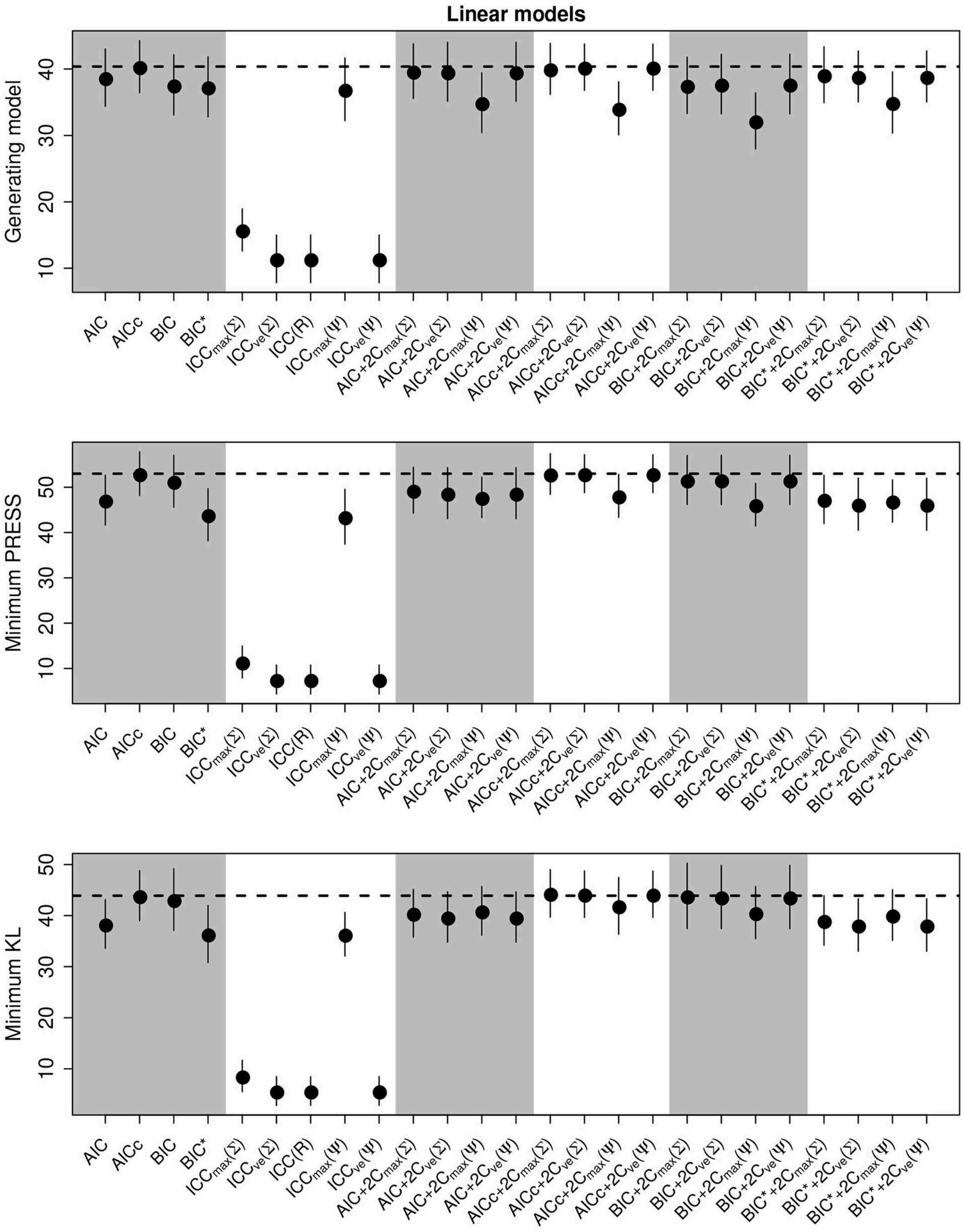


FIGURE 2 | Performance of the criterion for all selection goals for the linear model simulation experiment. Points are the average performance-level, bars give standard errors. The dashed horizontal line gives the performance of AICc for reference. The top panel gives the frequency that each criterion selects the model used to generate the data, the middle pannel gives the frequency of selecting the model that minimizes the predicted residual error sum of squares (PRESS), while the bottom panel gives the frequency of selection by the criterion of the model corresponding to the minimum Kullback-Leibler divergence (KL).

TABLE 4 | Best performing information criteria at selecting the generating model from the candidate set.

		Linear models	Population models	Barley yield models	Overall
Sample size	Low	AICc	BIC+2$C_{max}(\boldsymbol{\Psi})$	AIC+2$C_{max}(\boldsymbol{\Psi})$	AICc
	Medium	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	BIC+2$C_{max}(\boldsymbol{\Sigma})$	BIC	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
	High	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AIC+2$C_{max}(\boldsymbol{\Sigma})$	BIC*
Variance	Low	AICc+2$C_{max}(\boldsymbol{\Sigma})$	BIC*+2$C_{max}(\boldsymbol{\Sigma})$	BIC*	BIC
	Medium	AIC+2$C_{vE}(\boldsymbol{\Sigma})$	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	AICc
	High	AIC	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
Collinearity	Low	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
	Medium	AIC+2$C_{max}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{max}(\boldsymbol{\Sigma})$
	High	AIC+2$C_{vE}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
Overall		AICc	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$

TABLE 5 | Best performing information criteria at selecting the optimal predictive model from the candidate set.

		Linear models	Population models	Barley yield models	Overall
Sample size	Low	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	BIC	AICc+2$C_{max}(\boldsymbol{\Psi})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
	Medium	AICc+2$C_{max}(\boldsymbol{\Sigma})$	BIC+2$C_{max}(\boldsymbol{\Sigma})$	BIC*	BIC
	High	BIC*	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AIC + 2$C_{vE}(\boldsymbol{\Sigma})$	BIC*
Variance	Low	BIC+2$C_{max}(\boldsymbol{\Sigma})$	BIC	BIC*	BIC
	Medium	AICc	BIC*+2$C_{max}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	AICc
	High	AICc+2$C_{max}(\boldsymbol{\Sigma})$	BIC*+2$C_{max}(\boldsymbol{\Sigma})$	BIC+2$C_{vE}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
Collinearity	Low	BIC+2$C_{max}(\boldsymbol{\Sigma})$	NA	NA	BIC+2$C_{max}(\boldsymbol{\Sigma})$
	Medium	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
	High	AICc	NA	NA	AICc
Overall		AICc	BIC+2$C_{max}(\boldsymbol{\Sigma})$	BIC	AICc+2$C_{vE}(\boldsymbol{\Sigma})$

TABLE 6 | Best performing information criteria at selecting the minimum KL divergence from the candidate set.

		Linear models	Population models	Barley yield models	Overall
Sample size	Low	AICc+2$C_{max}(\boldsymbol{\Sigma})$	AICc	AICc+2$C_{max}(\boldsymbol{\Psi})$	AICc+2$C_{max}(\boldsymbol{\Sigma})$
	Medium	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AIC	BIC	AICc+2$C_{max}(\boldsymbol{\Psi})$
	High	BIC*	AIC	AIC	BIC
Variance	Low	BIC+2$C_{max}(\boldsymbol{\Sigma})$	AIC	BIC	BIC
	Medium	AICc	AIC	AICc+2$C_{vE}(\boldsymbol{\Sigma})$	AICc
	High	AICc+2$C_{max}(\boldsymbol{\Sigma})$	$ICC_{max}(\boldsymbol{\Sigma})$	BIC+2$C_{vE}(\boldsymbol{\Sigma})$	AICc+2$C_{vE}(\boldsymbol{\Sigma})$
Collinearity	Low	BIC+2$C_{max}(\boldsymbol{\Sigma})$	NA	NA	BIC+2$C_{max}(\boldsymbol{\Sigma})$
	Medium	AICc+2$C_{max}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{max}(\boldsymbol{\Sigma})$
	High	AICc+2$C_{max}(\boldsymbol{\Sigma})$	NA	NA	AICc+2$C_{max}(\boldsymbol{\Sigma})$
Overall		AICc +2$C_{max}(\boldsymbol{\Sigma})$	AICc	BIC	BIC

of the KL divergence selection. Performance differences among most of the criteria differed only by a few percentage points and the difference between top performers was within the range of the performance uncertainty (**Figure 3**).

3.3. Barley Yield Models

We present the overall criterion performance for the barley yield model simulation experiments, along with standard errors, in **Figure 4**. While the performance of criteria was strongly

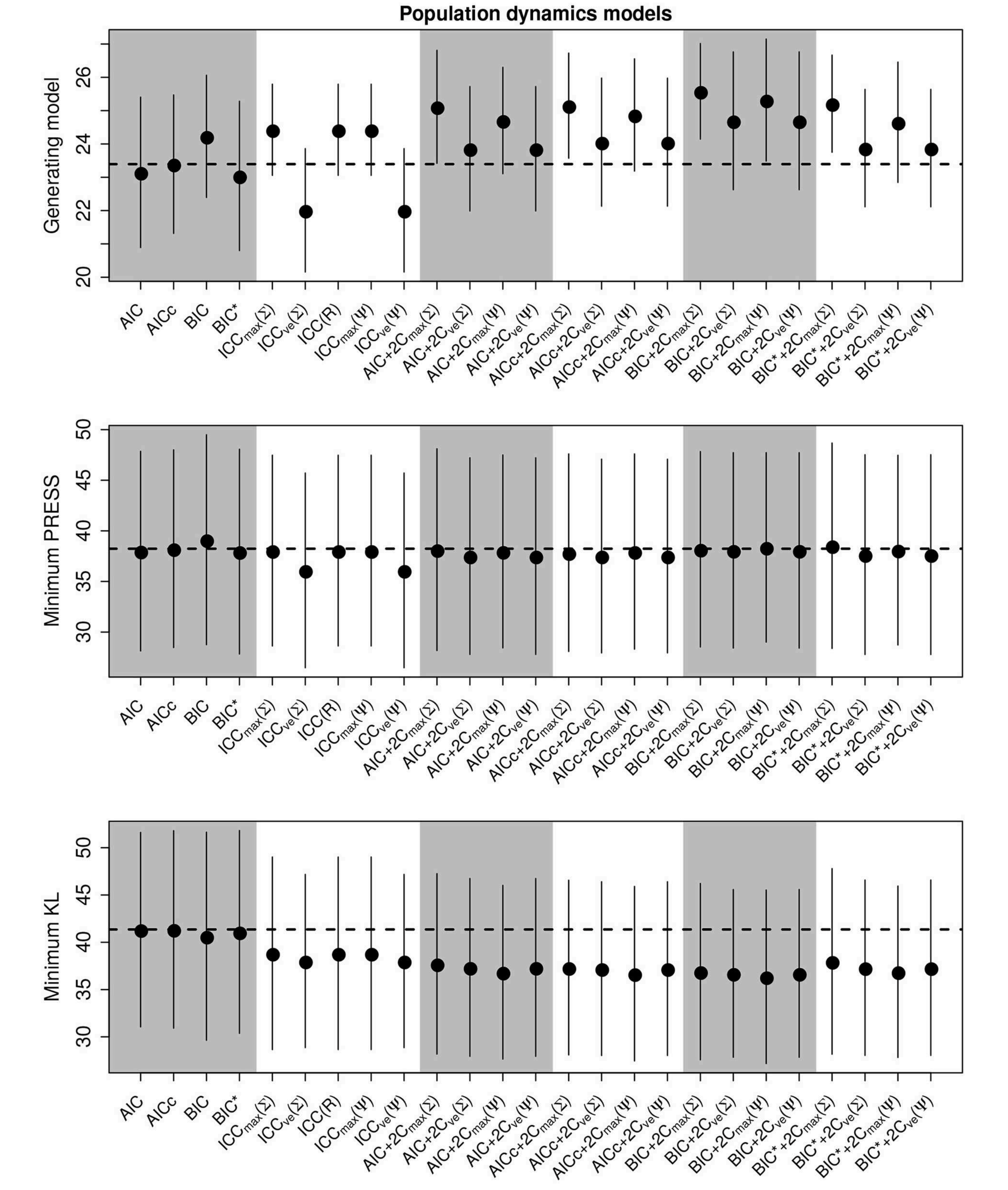


FIGURE 3 | Performance of the criterion for all selection goals for the nonlinear population dynamics model simulation experiment. Points are the average performance-level, bars give standard errors. The top panel gives the frequency that each criterion selects the model used to generate the data, the middle pannel gives the frequency of selecting the model that minimizes the predicted residual error sum of squares (PRESS), while the bottom panel gives the frequency of selection by the criterion of the model corresponding to the minimum Kullback-Leibler divergence (KL).

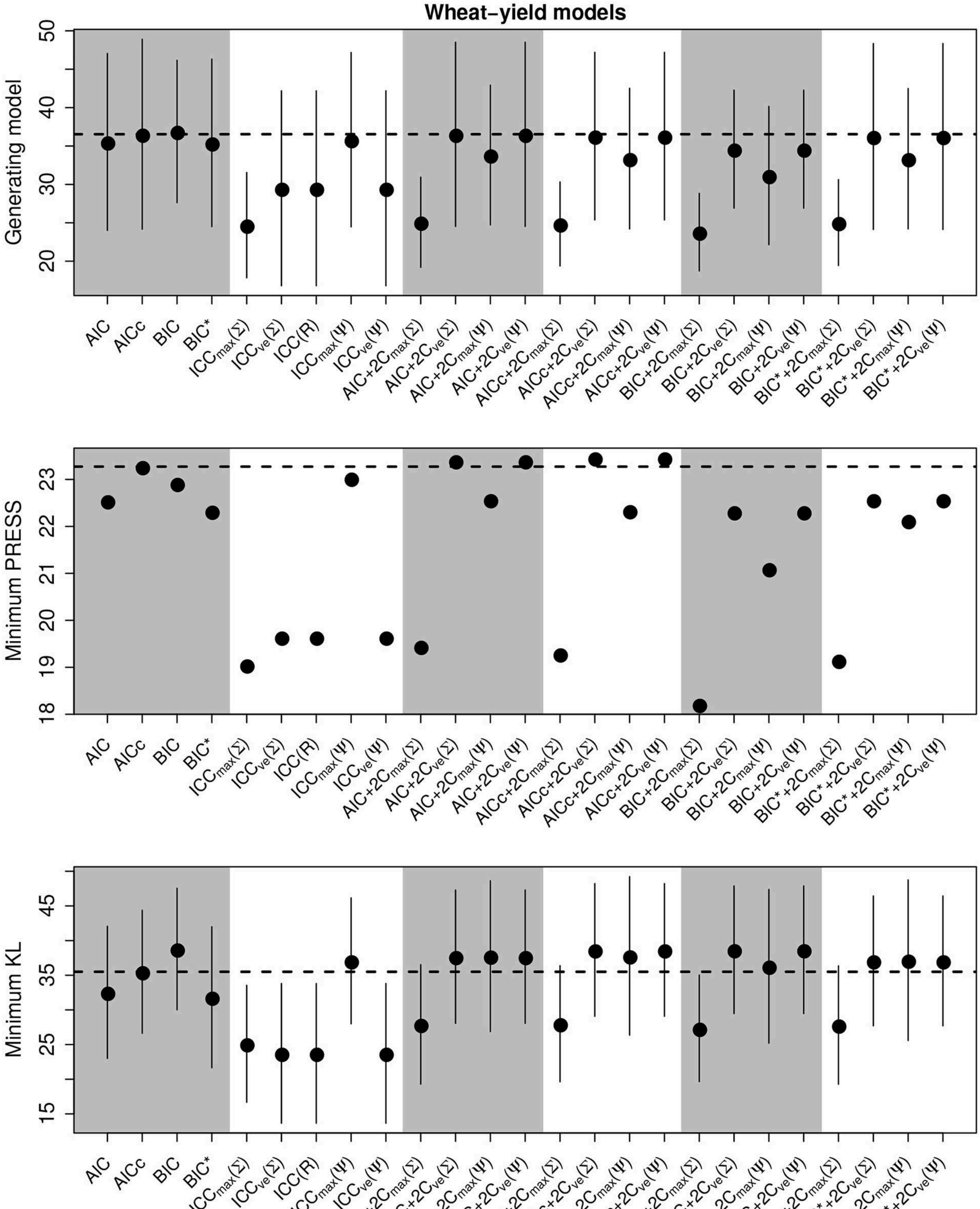


FIGURE 4 | Performance of the criterion for all selection goals for the nonlinear barley yield simulation experiment. Points are the average performance-level, bars give standard errors. The top panel gives the frequency that each criterion selects the model used to generate the data, the middle pannel gives the frequency of selecting the model that minimizes the predicted residual error sum of squares (PRESS), while the bottom panel gives the frequency of selection by the criterion of the model corresponding to the minimum Kullback-Leibler divergence (KL). No error bars are present in the PRESS results becauase we did not repeat these experiments.

correlated across all selection goals in the linear models, here performance was not correlated. While the selection of the minimum KL divergence was highly variable, similar to the population dynamics models, the PRESS performance was very consistent between criterion. Here, the class of ICC criteria tended to perform poorly though $\mathrm{ICC}_{\max}(\boldsymbol{\Psi})$ again tended to be consistent with the standard criterion and to perform better than the other forms of ICC (**Figure 3**). The compound criterion tended to perform better than the standard criterion but tended to perform worse at selecting the generating model.

Overall, the best criterion at selecting the generating model on average was the $\mathrm{AICc}+2C_{\mathrm{vE}}(\boldsymbol{\Sigma})$, while the best at prediction and at selecting the minimum KL divergence was the BIC (**Table 6**). In general, criteria performed better as sample size increased and variance decreased for selecting the generating model and the KL minimum, as expected (**Supplementary Figures S13–S18**). Performance differences among most of the criteria differed only by a few percentage points and the difference in top performers was within the range of performance uncertainty (**Figure 4**).

Finally, we found that overall performance across all simulations varied by the selection goals. The best at selecting both the generating model and the best predictive model overall was $\mathrm{AICc}+2C_{\mathrm{vE}}(\boldsymbol{\Sigma})$ (**Tables 4**, **5**). The criterion that performed best at selecting the KL minimum was BIC (**Table 6**).

4. DISCUSSION

The compound criterion $\mathrm{AICc}+2C_{\mathrm{vE}}(\boldsymbol{\Sigma})$ performed best on average at selecting both the generating model and the best predictive model, two important goals of ecological modeling. Surprisingly, the BIC performed best at selecting the model corresponding to the minimum KL divergence even though it is not meant to be an estimate of this quantity. Although the KL divergence is not a quantity that is itself of interest to scientists, it may be useful as a measure of the distance to truth. Despite the strong overall performance of the compound criteria, differences in performance between the top criteria were small. For example, while $\mathrm{AICc}+2C_{\mathrm{vE}}(\boldsymbol{\Sigma})$ performed best and selected the generating model 33.1% of the time across all experimental conditions, AICc selected the generating model 32.1% of the time and BIC selected the generating model 31.0% of the time.

Previous studies have looked at the performance of the ICC on linear regression models (Bozdogan, 1990; Bozdogan and Haughton, 1998; Clark and Troskie, 2006; Yang and Bozdogan, 2011), mixture models (Windham and Cutler, 1992; Bozdogan, 1993; Miloslavsky and Laan, 2003) and time series models (Bozdogan, 2000; Clark and Troskie, 2008). This past work has generally found much better performance of the ICC's than our study. For example, linear regression simulations suggest that the criteria may often outperform AIC and BIC, though limitations in study design are likely responsible for the different results. Two of these studies on linear regression (Bozdogan and Haughton, 1998; Clark and Troskie, 2006) did not allow for overfitting the generating model. While a third study (Yang and Bozdogan, 2011) did include the potential for overfitting, the variation in the extra model covariates were two orders of magnitude larger than the covariates of the generating model. This may not provide a realistic assessment of performance, as practitioners are often interested in distinguishing between effects that vary on the same scale. The results of the time series model application of ICC appeared more promising as ICC tended to do better than AIC or BIC most of the time when selecting among autoregressive moving average models (Clark and Troskie, 2008). Our population dynamics simulations also suggest that the ICC criteria perform better at selecting the generating model in nonlinear time series analysis than in linear regression, however we found that performance of the ICC criterion was rarely a significant improvement over AICc.

While many of the ICC performed poorly in our simulation experiments, the newly developed $\mathrm{ICC}_{\max}(\boldsymbol{\Psi})$ was comparable to the traditional criterion for all selection goals. $\mathrm{ICC}_{\max}(\boldsymbol{\Psi})$ uses the coefficient of variation matrix, accounting for uncertainty in parameter estimation. The compound criteria tended to provide superior performance over the other ICC measures. Even though the $\mathrm{ICC}_{\max}(\boldsymbol{\Psi})$ performed well as a criterion on its own, when incorporated as a compound criterion it tended to slightly underperform the best compound criteria. This is likely because the penalty term of the compound criteria ended up being too severe. Further work designed to optimize the weighting of the components might improve the performance of the compound criteria.

In the linear model simulation experiment, the AICc tended to do better than the BIC at selecting the generating model (**Figure S1**). In contrast, our population dynamics and barley yield simulation experiments found that BIC outperformed the AICc at selecting the generating model (**Figures S7, S13**). These results are broadly consistent with guidelines developed by Burnham and Anderson (2004) who outline how the BIC can be expected to outperform AIC when there are a few large effects. In systems with many small effects, such as the one used in our linear model experiments, the AIC will be expected to perform best. Further work by Brewer et al. (2016) has highlighted that the presence of multicollinearity can reverse these recommendations, with BIC generally selecting select better predictive models than the AIC. A previous study on population models by Corani and Gatto (2007) found that AICc outperformed BIC; however, this study was on nested models so the scenario more closely resembled our linear model simulation experiment. In a study design similar to our own, Hooten (1995) found that the BIC did better than either the AIC or the AICc at selecting the form of the generating model when selecting among density dependence forms, consistent with our results.

Averaging over all experimental factors provides a useful metric for assessing the general performance in complex ecological models. However, performance was highly variable on specific simulation experiments and even among experimental factors. We ascribe the differences between our results, which found only modest differences among criteria, and previous work to the broad array of simulation conditions. Averaging across these conditions provides a better guide to how criterion perform under a range of scenarios, though at the cost of providing less guidance for specific modeling scenarios. As Forster (2000) points out the performance of any criterion is context dependent and criteria will have a domain where they may be superior and where they may be inferior.

Designers and consumers of simulation validation studies need to carefully consider if performance is being assessed in a domain relevant to their modeling objectives. One potential approach to deal with the variability in performance is to conduct simulation experiments for every particular study to determine the optimal criterion. We would caution against this, besides performance being conditional on the particular model set, we expect this would lead to an anthology of idiosyncratic selection methodologies. Instead, we advise practitioners to rely on a criterion that has been shown to be consistent with their modeling goals and effective in a wide range of scenarios. Finally, there is no automated model selection approach that will substitute the clear-headed thinking that necessary to develop distinct, testable hypotheses that will answer the scientific question at hand. When this clarity is not possible, it may be preferable to develop a single, comprehensive model rather than performing model selection.

Our compound criteria are the sum of two estimated divergences. The first divergence attempts to measure the discrepancy between the model and truth. This model discrepancy can be estimated by AIC, AICc, BIC, BIC*, or one of the many other existing criteria. The second divergence estimates the distance between the joint sampling distribution of the parameters and the product of the marginal sampling distributions of the parameters. The motivation behind including this second divergence is to assess the estimability of parameters, a model quality that is often overlooked but has important implications when interpreting estimates, making out-of-sample predictions, and transferring parameters and models for use in other contexts. Thus, this divergence is a measure of a models usefulness. Our results suggest compound criterion that balance traditional measures of fit and complexity with an additional measure of usefulness can improve ecological inference. We found the AICc+2$C_{vE}(\mathbf{\Sigma})$ to be the best combination of these terms out of those considered here for both selecting the generating model and for prediction. AICc likely performed well because even the largest sample sizes explored here were relatively low, a common issue in many ecological datasets.

For the informational complexity we used a measure developed in past work based on the KL divergence between the joint and marginal sampling distributions of parameter estimates (van Emden, 1969; Bozdogan, 2000) (Equation 3). While the KL divergence has taken a primary role in ecological model selection, it is a divergence not a true distance. This means that the KL divergence between the distributions f and g is not necessarily equal to the KL divergence between g and f. In contrast, the Hellinger and Bhattacharyya distances are both true distance measures and have this symmetry property. Using an alternative measure may improve interpretability of the informational complexity, however it is not clear that these quantities have the same informational interpretation as the KL divergence, therefore it is not clear how to best combine these distance measures with information criterion.

Bozdogan and Haughton (1998) developed a consistent form of ICC by scaling the complexity measure, $C_{max}(\mathbf{\Sigma})$ by $\ln(n)$. While this does yield a consistent criterion, the performance of this *ad-hoc* approach was poor in their simulation studies. Our own preference is to use a compound criterion with a consistent form such as BIC. This study shows that BIC+2$C_{vE}(\mathbf{\Psi})$ achieves all measures of quality well under a broad range of modeling frameworks and it has the theoretical advantage of being scale invariant and consistent. Furthermore, the BIC is consistent at large sample size. At small sample size the BIC tends to choose compact model where all of the model components are well supported. Leading, we think, to a greater ease of interpretation (e.g., Arnold, 2010; Leroux, 2019).

While our analysis only considers a single best model, there are often likely to be several models that perform nearly as well due to the flexibility of the models in our simulation designs. Bayesian model averaging, and the complementary model averaging approach developed using AIC (Burnham and Anderson, 2002), is one common approach to account for uncertainty in model selection (but see Ponciano and Taper, submitted this issue). Model averaging can provide more precise parameter estimates (e.g., Vardanyan et al., 2011) and ensemble predictions can be more accurate than a single model (e.g., Martre et al., 2015). Given that our compound criterion performed slightly better than the standard information criterion for in-sample prediction and provides a measure of parameter dependence we expect that the compound criteria are suitable for model averaging and may directly address one major criticism of model averaging, the necessity of covariate independence (Cade, 2015).

We have assumed an equal weighting of the divergence between model and truth and the divergence measuring parameter complexity, though we could also choose to weight these contributions differently. One approach would be to calculate the optimal weights using simulation methods, while another approach is to allow the researcher to apply *a priori* weights based on the value a researcher places on model parsimony and estimability. It is these epistemic considerations that served as inspiration for developing these compound criteria so such a weighting would be consistent our original motivation.

This study provides evidence that developing information criterion based on measures other than the divergence between model and truth can yield improved model selection performance. However, we found that differences in performance between the best compound criterion and standard criteria were often small. This result aligns with previous work (Murtaugh, 2009) suggesting that standard methods tend to consistently produce models that are statistically and scientifically useful, though not necessarily optimal. Given that standard criteria are typically easy to calculate from regression output they provide useful and reliable tools for practicing ecologists. The compound criteria here can also be calculated from standard output suggesting that they could also be widely applied. Computational procedures such as regression trees (Murtaugh,

2009) or statistical learning methods (Corani and Gatto, 2006a,b, 2007) may also be useful tools under a wide variety of conditions, however these methods can be time demanding. The compound criteria examined here yield improved performance of model selection without dramatically increasing the amount of work needed to do inference.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.imperial.ac.uk/cpb/gpdd2/secure/login.aspx.

AUTHOR CONTRIBUTIONS

MT and JF conceived of the presented ideas. JF developed the theory and implemented the simulations. RZ-F, MT, MJ, and BM verified the analyses. MJ developed the suite of yield models. All authors discussed the results and contributed to the final manuscript.

FUNDING

MT has been supported in part by MTFWP contract #060327 while JF and BM were supported in part by USDA NRI Weedy and Invasive Plants Program Grant No. 00-35320-9464.

ACKNOWLEDGMENTS

We thank R. Boik, Mark Greenwood, Steve Cherry, Subhash Lele, Jose Ponciano, Taal Levi, Kristen Sauby, and Robert Holt for comments on earlier drafts of this work. MT's understanding of the problem of model identification has been greatly enhanced over the years by discussions with Subhash Lele, Brian Dennis, and Jose Ponciano.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00427/full#supplementary-material

REFERENCES

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95, 631–636. doi: 10.1890/13-1452.1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Allen, D. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* 16, 125–127. doi: 10.1080/00401706.1974.10489157

Arnold, T. W. (2010). Uninformative parameters and model selection using Akaike's information criterion. *J. Wildl. Manage* 74, 1175–1178. doi: 10.1111/j.1937-2817.2010.tb01236.x

Boik, R., and Shirvani, A. (2009). Principal components on coefficient of variation matrices. *Stat. Methodol.* 6, 21–46. doi: 10.1016/j.stamet.2008.02.006

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Commun. Stat. A Theor.* 19, 221–278. doi: 10.1080/03610929008830199

Bozdogan, H. (1993). "Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix," in *Information and Classification* (Berlin; Heidelberg: Springer), 40–54.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *J. Math. Psychol.* 44, 62–91. doi: 10.1006/jmps.1999.1277

Bozdogan, H., and Haughton, D. (1998). Informational complexity criteria for regression models. *Comput. Stat. Data Anal.* 28, 51–76. doi: 10.1016/S0167-9473(98)00025-5

Brewer, M. J., Butler, A., and Cooksley, S. L. (2016). The relative performance of AIC, AICCand BIC in the presence of unobserved heterogeneity. *Methods Ecol. Evol.* 7, 679–692. doi: 10.1111/2041-210X.12541

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resour. Res.* 37, 1015–1030. doi: 10.1029/2000WR900350

Burnham, K. K. P., and Anderson, D. D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Method Res.* 33, 261–304. doi: 10.1177/0049124104268644

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. New York, NY: Springer-Verlag.

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology* 96, 2370–2382. doi: 10.1890/14-1639.1

Cavanaugh, J. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probabil. Lett.* 33, 201–208. doi: 10.1016/S0167-7152(96)00128-9

Clark, A. E., and Troskie, C. G. (2006). Regression and ICOMP-A simulation study. *Commun. Stat. B Simul.* 35, 591–603. doi: 10.1080/03610910600716910

Clark, A. E., and Troskie, C. G. (2008). Time series and model selection. *Commun. Stat. B Simul.* 37, 766–771. doi: 10.1080/03610910701884153

Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J., et al. (2001). Ecological forecasts: an emerging imperative. *Science* 293, 657–660. doi: 10.1126/science.293.5530.657

Corani, G., and Gatto, M. (2006a). Model selection in demographic time series using VC-bounds. *Ecol. Model* 191, 186–195. doi: 10.1016/j.ecolmodel.2005.08.019

Corani, G., and Gatto, M. (2006b). VC-dimension and structural risk minimization for the analysis of nonlinear ecological models. *Apple Math. Comput.* 176, 166–176. doi: 10.1016/j.amc.2005.09.050

Corani, G., and Gatto, M. (2007). Erratum selection in demographic time series using VC-bounds. *Ecol. Model* 200, 273–274. doi: 10.1016/j.ecolmodel.2006.08.006

Cox, D. (1990). Role of models in statistical analysis. *Stat. Sci.* 5, 169–174. doi: 10.1214/ss/1177012165

Dietze, M. C., Fox, A., Beck-johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., et al. (2018). Iterative near-term ecological forecasting: needs, opportunities, and challenges. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1424–1432. doi: 10.1073/pnas.1710231115

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Ferguson, J. M., Carvalho, F., Murillo-García, O., Taper, M. L., and Ponciano, J. M. (2016). An updated perspective on the role of environmental autocorrelation in animal populations. *Theor. Ecol.* 9, 129–148. doi: 10.1007/s12080-015-0276-6

Ferguson, J. M., and Ponciano, J. M. (2014). Predicting the process of extinction in experimental microcosms and accounting for interspecific interactions in single-species time series. *Ecol. Lett.* 17, 251–259. doi: 10.1111/ele.12227

Ferguson, J. M., and Ponciano, J. M. (2015). Evidence and implications of higher-order scaling in the environmental variation of animal population growth. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2782–2787. doi: 10.1073/pnas.1416538112

Ferguson, J. M., Taper, M. L., Guy, C. S., and Syslo, J. M. (2012). Mechanisms of coexistence between native bull trout (*Salvelinus confluentus*) and non-native lake trout (*Salvelinus namaycush*): inferences from pattern-oriented modeling. *Can. J. Fish. Aquat. Sci.* 769, 755–769. doi: 10.1139/f2011-177

Forster, M. (2000). Key concepts in model selection: performance and generalizability. *J. Math. Psychol.* 44, 205–231. doi: 10.1006/jmps.1999.1284

Freckleton, R. P. (2011). Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behav. Ecol. Sociobiol.* 65, 91–101. doi: 10.1007/s00265-010-1045-6

Giere, R. N. (2004). How models are used to represent reality. *Philos. Sci.* 71, 742–752. doi: 10.1086/425063

Gilbert, P. and Varadhan, R. (2016). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1. Available online at: https://CRAN.R-project.org/package=numDeriv

Grafton, R. Q., Kirkley, J., and Squires, D. (2017). *Economics for Fisheries Management*. New York, NY: Routledge.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Stat.* 16, 342–355. doi: 10.1214/aos/1176350709

Hooten, M. (1995). *Distinguishing forms of statistical density dependence and independence in animal time series data using information criteria.* (Ph.D. thesis). Montana State University, Bozeman, MT, United States.

Hurvich, C. M., and Tsai, C.-L. L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307. doi: 10.1093/biomet/76.2.297

Jasieniuk, M., Taper, M. L., Wagner, N. C., Stougaard, R. N., Brelsford, M., and Maxwell, B. D. (2008). Selection of a barley yield model using information-theoretic criteria. *Weed Sci.* 56, 628–636. doi: 10.1614/WS-07-177.1

Lele, S. R. (2004). "Error functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 191–203.

Lele, S. R., and Taper, M. L. (2012). "Information criteria in ecology," in *Encyclopedia of Theoretical Ecology*, eds A. Hastings and L. Gross (Berkeley, CA: University of California Press), 371–376.

Leroux, S. J. (2019). On the prevalence of uninformative parameters in statistical models applying model selection in applied ecology. *PLoS ONE* 14:e0206711. doi: 10.1371/journal.pone.0206711

Lindegren, M., Mollmann, C., Nielsen, A., Brander, K., Mackenzie, B. R., and Stenseth, N. C. (2010). Ecological forecasting under climate change : the case of Baltic cod. *Proc. R. Soc. B Biol. Sci.* 277, 2121–2130. doi: 10.1098/rspb.2010.0353

Link, W. A., and Sauer, J. R. (2016). Bayesian cross-validation for model evaluation and selection, with application to the North American breeding survey. *Ecology* 97, 1746–1758. doi: 10.1890/15-1286.1

Link, W. A., Sauer, J. R., and Niven, D. K. (2017). Model selection for the North American Breeding Bird Survey: a comparison of methods. *Condor* 119, 546–556. doi: 10.1650/CONDOR-17-1.1

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., et al. (2015). Multimodel ensembles of wheat growth: Many models are better than one. *Glob. Change Biol.* 21, 911–925. doi: 10.1111/gcb.12768

Miloslavsky, M., and Laan, M. J. V. D. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Comput. Stat. Data Anal.* 41, 413–428. doi: 10.1016/S0167-9473(02)00166-4

Murtaugh, P. A. (2009). Performance of several variable-selection methods applied to real ecological data. *Ecol. Lett.* 12, 1061–1068. doi: 10.1111/j.1461-0248.2009.01361.x

NERC (2010) *The Global Population Dynamics Database Version 2.* Available online at: http://www.sw.ic.ac.uk/cpb/cpb/gpdd.html

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Stat.* 12, 758–765. doi: 10.1214/aos/1176346522

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivar. Anal.* 27, 392–403. doi: 10.1016/0047-259X(88)90137-6

Pickett, S. T., Kolasa, J., and Jones, C. G. (2010). *Ecological Understanding: The Nature of Theory and the Theory of Nature*. Burlington, MA: Academic Press.

Polansky, L., de Valpine, P., Lloyd-Smith, J. J. O., and Getz, W. W. M. (2009). Likelihood ridges and multimodality in population growth rate models. *Ecology* 90, 2313–2320. doi: 10.1890/08-1461.1

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Syst. Biol.* 61, 955–972. doi: 10.1093/sysbio/sys055

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team. Available online at: https://www.R-project.org/

Ruktanonchai, N. W., DeLeenheer, P., Tatem, A. J., Alegana, V. A., Caughlin, T. T., zu Erbach-Schoenberg, E., et al. (2016). Identifying malaria transmission foci for elimination using human mobility data. *PLoS Comput. Biol.* 12:e1004846. doi: 10.1371/journal.pcbi.1004846

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods Ecol. Evol.* 1, 103–113. doi: 10.1111/j.2041-210X.2010.00012.x

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* 68, 45–54. doi: 10.1093/biomet/68.1.45

Taper, M., and Lele, S. (2011). "Evidence, evidence functions, and error probabilities," in *Philosophy of Statistics*, eds M. Forster and P. Bandyophadhyay (Oxford, UK: North Holland), 1–31.

Taper, M. L. (2004). "Model identification from many candidates," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, chapter 15, eds M. L. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 488–524.

van Emden, M. (1969). *On the hierarchical decomposition of complexity* (Ph.D. thesis). Stichting Mathematisch Centrum, Amsterdam, Netherlands.

Vardanyan, M., Trotta, R., and Silk, J. (2011). Applications of Bayesian model averaging to the curvature and size of the universe. *Month. Notices R Astron. Soc.* 413, L91–L95. doi: 10.1111/j.1745-3933.2011.01040.x

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecol. Model.* 211, 1–10. doi: 10.1016/j.ecolmodel.2007.10.030

Windham, M., and Cutler, A. (1992). Information ratios for validating mixture analyses. *J. Am. Stat. Assoc.* 87, 1188–1192. doi: 10.1080/01621459.1992.10476277

Yang, H., and Bozdogan, H. (2011). Model selection with information complexity in multiple linear regression modeling. *Multiple Linear Regression Viewpoints* 37, 1–13.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33, 790–802. doi: 10.1016/j.tree.2018.08.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

METHODS
published: 14 November 2019
doi: 10.3389/fevo.2019.00415

Model Selection via Focused Information Criteria for Complex Data in Ecology and Evolution

Gerda Claeskens[1], *Céline Cunen*[2] *and Nils Lid Hjort*[2]*

[1] *ORSTAT and Leuven Statistics Research Centre, KU Leuven, Leuven, Belgium,* [2] *Department of Mathematics, University of Oslo, Oslo, Norway*

Datasets encountered when examining deeper issues in ecology and evolution are often complex. This calls for careful strategies for both model building, model selection, and model averaging. Our paper aims at motivating, exhibiting, and further developing focused model selection criteria. In contexts involving precisely formulated interest parameters, these versions of FIC, the focused information criterion, typically lead to better final precision for the most salient estimates, confidence intervals, etc. as compared to estimators obtained from other selection methods. Our methods are illustrated with real case studies in ecology; one related to bird species abundance and another to the decline in body condition for the Antarctic minke whale.

Keywords: bird species abundance, ecology, evolution, FIC and AFIC, focused model selection, linear mixed effects, minke whales

OPEN ACCESS

Edited by:
Jose Miguel Ponciano,
University of Florida, United States

Reviewed by:
Subhash Ramkrishna Lele,
University of Alberta, Canada
Christopher Jackson,
University of Cambridge,
United Kingdom

****Correspondence:***
Nils Lid Hjort
nils@math.uio.no

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *27 February 2019*
Accepted: *17 October 2019*
Published: *14 November 2019*

Citation:
Claeskens G, Cunen C and Hjort NL (2019) Model Selection via Focused Information Criteria for Complex Data in Ecology and Evolution. Front. Ecol. Evol. 7:415. doi: 10.3389/fevo.2019.00415

1. INTRODUCTION

Only rarely will initial modeling efforts lead to "one and only one model" for the data at hand. This simple empirical statement applies in particular to situations with complex data for complicated and not-yet-understood mechanisms underlying the phenomena being studied, in ecology and evolution, as well as other sciences. Thus, methods for model comparison, model selection, and model averaging are called for. Not surprisingly there must be several such methods, since the question "what is a good model for my data?" cannot be expected to have a simple and clear-cut answer.

There are indeed several model selection schemes in the statistics literature, with the more famous ones being the AIC (the Akaike Information Criterion) and the BIC (the Bayesian Information Criterion; see Claeskens and Hjort, 2008b) for a general overview. The AIC and BIC are able to compare and rank competing models for a given dataset, as long as they are all parametric. These and yet other methods work in an "overall modus," in appropriate senses comparing overall fit with overall complexity, but they do not take on board *the intended use of the fitted models*. This is where FIC (the Focused Information Criterion) comes in, along with certain relatives. The FIC aims at giving the most relevant model comparison and ranking, and hence also pointing to the best model, for *the given purpose*. What this given purpose is depends on the scientific context. Indeed, two research teams might ask different focused questions, with the same data and the same list of candidate models, and we judge it not to be a contradiction in terms that three focused questions might have three different best models.

The present article gives an account of FIC and its relatives, including also certain extensions of previously published methods. We do have models for ecology and evolution in mind, though it is clear that the view is broader: we wish to find good statistical models for complex data, and can do so, once crucial and context driven questions are translated to *focus parameters*. Our paper's contribution is 2-fold. (i) We aim at introducing the FIC methodology to researchers in ecology and evolution. We have therefore strived to include relevant examples, along with some R code. We also discuss various topics of interest to applied researchers, particularly in section 5. In this partly tutorial spirit, various technical details have been placed in the **Appendix**. (ii) Our article also serves as an outlet for a somewhat new FIC framework, termed the "fixed wide model framework," different from the "local asymptotics framework" used in the majority of previous publications. Details are in section 3, with material not been presented in this general form before. In particular, the extension of this framework to generalized linear models is novel.

To help fix ideas and some basic notation, we start with a concrete application. We use the dataset from Hand et al. (1994) regarding counts of the number of bird species on fourteen areas, vegetation islands, in the Andes mountains with páramo vegetation. In addition to the number of bird species y, there are four covariates recorded for each such vegetation island: x_1, the area of the vegetation island in thousands of square kilometers; x_2, the elevation in thousands of meters; x_3, the distance between the area and Ecuador in kilometers; and x_4, the distance from the nearest island in kilometers.

```
y        x1       x2        x3       x4
36       0.33     1.26        36     14
30       0.50     1.17       234     13
37       2.03     1.06       543     83
35       0.99     1.90       551     23
11       0.03     0.46       773     45
21       2.17     2.00       801     14
11       0.22     0.70       950     14
13       0.14     0.74       958      5
17       0.05     0.61       995     29
13       0.07     0.66      1065     55
29       1.80     1.50      1167     35
 4       0.17     0.75      1182     75
18       0.61     2.28      1238     75
15       0.07     0.55      1380     35
```

We model the number of bird species Y by a Poisson distribution with mean $\exp(x^{t}\beta)$, where x in the widest model consists of the constant 1 (modeling the intercept), all four covariates $x_1, \ldots, x_4$ as main effects, and all six pairwise interactions between these main effects. This amounts to a total of 11 parameters $\beta_0, \ldots, \beta_{10}$. We wish to include the intercept parameter β_0 in all candidate models, and hence take it as a "protected parameter," whereas the other parameters are "open," and can be pushed in and out of candidate models. For this application, all submodels of the largest 11-parameter model are considered, with the further restriction that interactions between two covariates can be included only if the two main effects are present. This results in a total of 113 models.

The main distinction between FIC and various other information criteria is the presence of a *focus*. This is a quantity of interest that depends on the model parameters and is estimable from the data. The generic notation for the focus used in our paper is μ. Its dependence on the model parameters might be indicated by writing $\mu(\beta)$.

In the bird species study, our first focus concerns one of the vegetation islands, Chiles. This area is the one among the fourteen that is closest to Ecuador, and has covariate values $x_1 = 0.33$, $x_2 = 1.26$, $x_3 = 36$, $x_4 = 14$. We wish to select a model that best estimates the expected number of bird species for this island, that is, $\mu(\beta) = \exp(x^{t}\beta)$ for the given covariate values for Chiles. In our model search problem there are 113 models and hence 113 estimators for μ. Each such estimator, say $\widehat{\mu}_M$ for a candidate model M, comes with its own bias and variance, say b_M and τ_M^2. Thus, for each candidate model there is a corresponding mean squared error (mse)

$$\mathrm{mse}_M = \tau_M^2 + b_M^2. \qquad (1)$$

The basic idea of the FIC is to estimate these mse values from the data, for the wide as well as for each candidate model, i.e., to construct

$$\mathrm{FIC}_M = \widehat{\mathrm{mse}}_M = \widehat{\tau}_M^2 + \widehat{\mathrm{bsq}}_M, \qquad (2)$$

with the second term indicating estimation of the squared bias $\mathrm{bsq}_M = b_M^2$. In the end one selects the model with the smallest estimated mse.

For the bird species application, we use FIC for finding the best model to estimate the expected number of bird species for Chiles. We use the R package `fic` with the following lines of R code, where we fit the wide model, specify the focus function, the covariate value in which to evaluate this focus, and the specific models that we wish to search through. In this example we restrict the built-in all subsets specification to only using models that obey the hierarchy principle (so out of the $2^{10} = 1024$ potential submodels, only the 113 pointed to above are included).

```
library(fic)
wide.birds = glm(y~.^2, data=birds,
 family=poisson)
focus1 = function(par, X) exp(X %*% par)
inds0 = c(1,rep(0,10)) # only the intercept
 is in the narrow model
A = all_inds(wide.birds, inds0)  # use all
 subsets of the wide model
#exclude models with interactions that do
 not have both main effects:
inds <- with(A,A[!(A[,2]==0 & (A[,6]==
1|A[,7]==1|A[,8]==1) |
   A[,3]==0 & (A[,6]==1|A[,9]==1
   |A[,10]==1) |
   A[,4]==0 & (A[,7]==1|A[,9]==1
   |A[,11]==1) |
   A[,5]==0 & (A[,8]==1|A[,10]==1|
   A[,11]==1)), ])
# specify the X used to evaluate the focus
 function:
XChiles=model.matrix(wide.birds)[1, ]
fic(wide=wide.birds, inds=inds, inds0=inds0,
 focus=focus1, X=XChiles)
```

For each of the 113 models we get via the output values of the focus estimate, the estimated bias, standard error, and actually two versions of the FIC of (2), corresponding to two related but different ways of estimating the b_M^2 part (for details, see section 2). For FIC tables and FIC plots we prefer working with the square-root of the FIC, i.e., estimates of the root-mse (rmse) rather than of the mse, as these are on the original scale of the focus and easier to interpret.

Table 1 is constructed from the output for a selection of models, including the narrow model (1) which has a relatively large (in absolute value) bias estimate of −19.035, a relatively small standard error of 2.247 and a focus estimate of 20.71; the wide model (113) with zero as the bias estimate though with a large standard error of 6.051. This is a typical output: the wide model contains 11 parameters to estimate which causes the standard error to be large, the narrow model only contains the intercept resulting in a small standard error. For the bias estimate the scenario is reversed: the wide model has the smallest bias, while the narrow model has a larger bias. The balancing act of the FIC via the mean squared error finds a compromise. The selected model (5) results in the smallest value of the square root of the estimated mean squared error (rmse). Its indicator sequence 10010,000000, with a one for β_0 and β_3, and zeroes for the interactions, points toward the selected focus $\mu(\beta) = \exp(\beta_0 + \beta_3 x_3)$ with corresponding estimated focus value 38.88. Using the wide model would have resulted in a close 38.27 though with a larger estimated root mean squared error. The wide model only ranks at the 73rd place according to estimated rmse. Model (20) is selected by the Bayesian information criterion BIC, it consists of the intercept, all four main effects and the interaction between x_1 and x_2. In the rmse ranking it comes at the 42nd place. Model (67) is the one selected by the Akaike information criterion, next to the intercept and all main effects it consists of the interactions x_1x_3, x_2x_3, x_2x_4. This models ranks 32nd.

The second focus concerns the probability of having more than 30 bird species, $P(Y > 30 \,|\, x)$. Now we do not specify a particular island but use the average FIC (see section 2.2), with equal weights for the fourteen vegetation islands (non-equal weights can easily be worked with too).

```
focus2 = function(par, X) 1-ppois(30,
lambda=exp(X %*% par))
Xall = model.matrix(wide.birds)
fic2 = fic(wide=wide.birds,inds=inds,
inds0=inds0,focus=focus2,X=Xall)
AVE = fic2[fic2$vals=="ave",]
which.min(AVE$rmse.adj)
```

The AFIC selects the following form for the mean: $\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_7 x_1 x_4)$. The averaged focus estimate of the probability of observing over 30 bird species in the selected model equals 15.73%, while the wide model's estimate is 21.83%, though with a substantial larger estimated mean squared error due to the estimation of 11 parameters instead of only 5 for the selected model. Of course, AIC and BIC ignore any information regarding the focus, and thus still recommend the very same models, model (67) for AIC, with estimate 21.15%, and model (20) for BIC, with estimate 21.59%. The AIC model ranks 16th, the BIC model is now at the third place.

Figure 1 displays for these two foci the root-FIC and root-AFIC values, as well as the estimated focus values, for all of the 113 models. The FIC or AFIC selected values, minimizing the respective criteria, are indicated in red, while the wide model's values are in blue.

Several traditional model selection criteria, such as the AIC and the BIC (see Claeskens and Hjort, 2008b, Chs. 2, 3) work in an overall modus, finding models that in a statistical sense are good on average, not taking on board the specific aims of a study. The FIC works explicitly with such specific aims, formalized via the focus parameters. Thus, FIC might find that one model works very well for covariates "in the middle," whereas another model could work rather better for covariates outside mainstream. Similarly, one model might work well for explaining means, and another for explaining variances. We stress that the FIC apparatus works for *any specified focus parameter*, and is not limited to e.g., regression coefficients and the customary selection of covariates from that perspective.

The generic FIC formula (2) cannot be immediately applied, as efforts are required to establish formulae for approximations to biases and variances, along with construction of estimators for these quantities. Thus, the FIC formula pans out differently in different situations, depending on the general framework, the complexity of models, and estimators of the focus parameters. A brief overview of general principles, leading to such approximations and estimators, is given in section 2. This also encompasses AFIC, ways of creating average-FIC scores in situations where more than one focus parameter is at stake.

In section 3 we provide the general FIC formulae in the so-called fixed wide model framework. The development of FIC formulae ingredients in a somewhat different framework, with local neighborhood models, is placed in **Appendix**. Generalized linear models are used as examples, encompassing linear regression, logistic and Poisson regression, etc. The more general

TABLE 1 | Bird species example.

Model	Coef. indicators	Focus	Bias	Se	$\sqrt{\text{FIC}}$	AIC	BIC
1	10000,000000	20.714	−19.035	2.247	19.167	143.26	143.90
5	10010,000000	38.882	0.000	4.383	4.383	112.65	113.93
20	11111,100000	33.718	−2.156	4.670	5.143	91.91	95.74
28	11101,001000	26.356	−11.0468	3.674	11.642	98.54	101.74
67	11111,010110	39.784	0.000	5.296	5.296	91.44	96.55
113	11111,111111	38.269	0.000	6.051	6.051	95.72	102.75

This table is constructed from output of the R function fic for six of the 113 models, together with the AIC and BIC values. FIC selection takes place via the square root of the estimated mean squared error of the focus estimator.

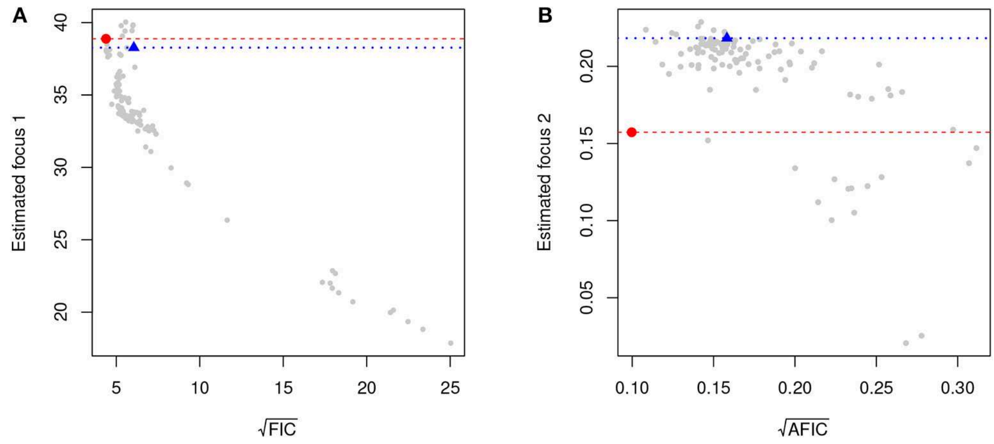


FIGURE 1 | The two plots give values for a total of 113 Poisson regression models, related to two different focused questions. **(A)** FIC plot for estimating the expected number of bird species for the Chiles region. **(B)** AFIC plot for estimating the probability of observing over 30 species, averaging over all 14 islands. The red dot and line indicate the selected value, the blue triangle and line are for the wide model.

class of linear mixed models has proven important for various applications to ecology, and in section 3.3 FIC formulae are reached for such. In section 4 we use linear mixed effects models with FIC for analyzing the body conditions of minke whales in the Antarctic, where one focus parameter is the yearly decline in energy storage. A general but brief discussion is then offered in section 5. Here we touch on aspects of performance, along with a few concluding remarks, some of which point to future research.

2. FOCUSED INFORMATION CRITERIA

The application concerning birds on vegetation islands in the previous section was meant to provide intuition for the use of FIC for model selection. Here we give a more formal, but brief, overview of the FIC and AFIC schemes.

2.1. General FIC Scheme

Suppose we have defined a *wide model* which is assumed to be the true data-generating mechanism. Estimating the focus parameter using the wide model leads to $\widehat{\mu}_{\text{wide}}$, which under broad regularity conditions will aim at μ_{true}, the unknown true value of the focus parameter. Estimation via fitting a candidate model M leads to $\widehat{\mu}_M$, say, aiming for some least false parameter $\mu_{0,M}$, typically different from μ_{true}, due to modeling bias. The least false parameter in question relates to the best approximation candidate model M can manage to be, to the true model. There is therefore an inherent bias, say

$$b_M = \mu_{0,M} - \mu_{\text{true}},$$

associated with using M. We saw estimates of this bias in the birds application above, where small models could have larger biases.

The estimators will have certain variances. In most frameworks, involving independent or weakly dependent data, these tend to zero with speed $1/n$, in terms of growing sample size n. It is therefore convenient and informative to write these variances as $\tau^2_{\text{wide}} = \sigma^2_{\text{wide}}/n$ and $\tau^2_M = \sigma^2_M/n$, where the mathematics and approximation theorems associated with different frameworks typically yield expressions for or approximations to the σ_{wide} and σ_M. The mse of the focus parameter estimators is the sum of the variance and the bias squared,

$$\text{mse}_{\text{wide}} = \sigma^2_{\text{wide}}/n + 0^2 \quad \text{and} \quad \text{mse}_M = \sigma^2_M/n + b^2_M. \tag{3}$$

These quantities are measures of the *risk*, in the statistical sense, associated with using each of the models for estimating μ. As explained in the introduction, the FIC scores of (2) are estimates of the mse of the focus parameter estimators, i.e., the $\widehat{\mu}_M$, for a specific dataset, for each of the models under consideration. Equation (3) is also an informative reminder that with more data, variances get small, but biases remain. So using a model which is not fully correct can still yield sharper estimators, as long as the bias is moderate or small: $|b_M| < (\sigma^2_{\text{wide}} - \sigma^2_M)^{1/2}/\sqrt{n}$. It is also clear that with steadily more data, steadily more sophisticated models can and indeed should be used. The FIC makes these ideas operative.

In various cases the variance terms σ^2_M/n are easier to estimate than the squared biases b^2_M. A starting point for the latter is $\widehat{b}_M = \widehat{\mu}_M - \widehat{\mu}_{\text{wide}}$, but the corresponding $\widehat{b}^2_M$ will overshoot b^2_M with about κ^2_M/n, which is the variance of $\widehat{b}_M$. With appropriately constructed estimators of the quantities σ_{wide}, σ_M, κ_M (with different recipes for different situations), this yields two natural ways of estimating the actual mse values:

$$\begin{aligned} \text{FIC}^u_{\text{wide}} &= \widehat{\sigma}^2_{\text{wide}}/n + 0^2 \quad \text{and} \quad \text{FIC}^u_M = \widehat{\sigma}^2_M/n + \widehat{b}^2_M - \widehat{\kappa}^2_M/n, \\ \text{FIC}_{\text{wide}} &= \widehat{\sigma}^2_{\text{wide}}/n + 0^2 \quad \text{and} \\ \text{FIC}_M &= \widehat{\sigma}^2_M/n + \max(\widehat{b}^2_M - \widehat{\kappa}^2_M/n, 0). \end{aligned} \tag{4}$$

The FIC^u scores are (approximately) unbiased estimates of the mse, since $\widehat{b}^2_M - \widehat{\kappa}^2_M/n$ is (approximately) unbiased for b^2_M, whereas the FIC scores are adjusted versions, by truncating any negative estimates of squared bias to zero, as we did in the first example. If the true bias in question is some distance away from zero, FIC^u_M will be equal to FIC$_M$. When faced with a specific application one should decide on one of these two FIC versions, and use the same choice for all models under consideration.

In order to turn the general scheme (4) into clear formulae, with consequent algorithms, we need expressions for or approximations to the population quantities σ_M, b_M, κ_M, followed by clear estimation strategies for these again. In most cases we need to rely on large-sample approximations. Arriving at clear formulae for σ_M etc. depends on the particularities of the wide model, the candidate models, and the focus parameter. We provide such FIC formulae, for *two different frameworks* or setups. *The first* involves local asymptotics, with candidate models being a local distance $O(1/\sqrt{n})$ away from the wide model. This derivation is placed in **Appendices A1** and **A2**. *The second* avoids such local asymptotics and works from a fixed wide model and a collection of candidate models (see section 3). It is not a contradiction in terms that these two frameworks lead to related but not identical FIC formulae, as different mathematical approximations are at work.

2.2. AFIC, the Averaged-Weighted Selection Scheme

The FIC apparatus above is tailored to one specific focus parameter at a time. In a regression context this applies e.g., to estimating the mean response function for one covariate vector at a time, say $\mu(\theta; x_0)$. Often there would be active interest in several parameters, however, as with such a $\mu(\theta; x_0)$ for all x_0 in a segment of covariates, or a probability $P(Y \geq y_0 \,|\, x_0)$ for a set of thresholds, as in the birds study.

Suppose in general that an ensemble of estimands is of interest, say $\mu(\theta; v)$ with $v \in V$, and that a measure of relative importance $\mathrm{d}W(v)$ is assigned to these. There could be only a few such estimands under scrutiny, say μ_j for $j = 1, \ldots, k$, along with weights of importance $w_1, \ldots, w_k$. Estimation involving all higher quantiles, or all covariates within a certain region, however, would constitute examples where we need the more general $v \in V$ notation. Here we sketch the AFIC approach, for estimating the relevant integrated weighted risk.

For each focus parameter in the ensemble of estimands there is an associated mse or risk, mse(v). The combined risk associated with using model M then becomes

$$r_n(M) = \int \mathrm{mse}(v)\,\mathrm{d}W(v) = \int \{\sigma_M(v)^2/n + b_M(v)^2\}\,\mathrm{d}W(v),$$

with the appropriate $\sigma_M(v)$ and $b_M(v) = \mu_{0,M,n}(v) - \mu_{\mathrm{true}}(v)$. An approximately unbiased estimate of this combined risk is

$$\mathrm{afic}^u(M) = \int \{\widehat{\sigma}_M(v)^2/n\}\,\mathrm{d}W(v) + \int \{\widehat{b}_M(v)^2 - \widehat{\kappa}_M(v)^2/n\}\,\mathrm{d}W(v).$$

This is the same as a direct weighted sum or integral of the individual FIC$^u(M, v)$ scores. The adjusted version, however, where a potentially negative value of the estimated integrated squared bias is being truncated to zero, is not identical to the integral of the FIC(M, v) scores. It is rather equal to

$$\begin{aligned}\mathrm{afic}(M) \;=\; & \int \{\widehat{\sigma}_M(v)^2/n\}\,\mathrm{d}W(v) \\ & + \;\max\Big[\int \{\widehat{b}_M(v)^2 - \widehat{\kappa}_M(v)^2/n\}\,\mathrm{d}W(v), 0\Big].\end{aligned}$$

As with FIC, there are two related, but not identical, approximation schemes, the fixed wide model setup and the local asymptotics, of respectively section 3.1 and **Appendix A1**, leading now to somewhat different AFIC formulae. For details and applications (see Claeskens and Hjort, 2008a,b, Ch. 6).

There is a connection between Akaike's information criterion AIC and AFIC with certain model dependent weights (see Claeskens and Hjort, 2008a, Sec. 6.2). Broadly speaking, the AIC turns out to be large-sample equivalent to cases with AFIC where "all things are equally important."

3. FIC WITHIN A FIXED WIDE MODEL FRAMEWORK

The FIC as used in the bird species example is the version as derived in Claeskens and Hjort (2003), see also Claeskens and Hjort (2008b, Ch. 6). For the estimation of bias and variance a local asymptotic framework is used in which the parameters of the true density of the data are assumed to be of the form $\gamma = \gamma_0 + \delta/\sqrt{n}$, with n the sample size, see **Appendix A1** for more explanation. This assumptions means in practice that we believe that all models are relative close to each other and to the truth. Moreover, all models are submodels of a wide model. Since the derivation of the FIC formulae is contained in the references above, we only place a summary in the **Appendix**.

In this section we present the "fixed wide model" framework, which is particularly useful if the set of candidate models are seen as not being in a reasonable vicinity of each other. This second framework allows candidate models of a different sort from the wide model; in particular, a candidate model does not have to be a clear submodel of the wide model. Keep in mind that the two different FIC frameworks have the same aims and motivation; the difference between them lies in the different mathematical tools for estimating the relevant mse quantities, which lead to different formulae. In the discussion section 5 we come back to some differences between the two frameworks. Here we start in section 3.1 by presenting the fixed wide model FIC in a general regression setup. Then in the two following subsections we deal with two specific model classes of general interest, generalized linear models and linear mixed models, in more detail.

3.1. General Regression Models

In this subsection we use the familiar (x_i, y_i) notation for the regression data, with x_i the covariate vector in question. The FIC machinery we develop here starts from the existence of a fixed wide model. The development represents an extension of earlier work of Jullum and Hjort (2017, 2019) for i.i.d. data and survival analysis, Ko et al. (2019) for copulae models, Cunen et al. (2019) for power-law distributions (with applications to war and conflict data) and Cunen et al. (in review)[1,2] for linear mixed effects models (with application to whale ecology).

[1] Cunen, C., Walløe, L., and Hjort, N. L. (2019). Focused model selection for linear mixed models, with an application to whale ecology. *Ann. Appl. Stat.*

[2] Cunen, C., Walløe, L., Konishi, K., and Hjort, N. L. (2019). Decline in energy storage for the Antarctic minke whale (*Balaenoptera bonaerensis*) in the Southern Ocean during the 1990s.

Since we wish to estimate the mse of the focus estimator in different models, we first consider the asymptotic distribution of the parameter estimator in the wide model and next in the other models of interest. The distributions are used to form the mse's of the focus estimators and finally we construct the fic as an estimated mse and select the model with the smallest fic value.

Suppose a wide model density is agreed upon, of the form $f(y_i \mid x_i, \theta)$, for a certain parameter vector θ, of length p. We consider this to be the true model. This θ would typically encompass both regression coefficients and parameters related to the spread and shape of error distributions. Define $u(y_i \mid x_i, \theta) = \partial \log f(y_i \mid x_i, \theta)/\partial\theta$ the score function, and $J_n = n^{-1} \sum_{i=1}^n \mathrm{Var}_{\mathrm{wide}}\, u(Y_i \mid x_i, \theta_{\mathrm{true}})$ the normalized Fisher information matrix at the true parameter. Under mild regularity conditions we have the following well-known result for the maximum likelihood estimator $\widehat\theta_{\mathrm{wide}}$,

$$\sqrt{n}(\widehat\theta_{\mathrm{wide}} - \theta_{\mathrm{true}}) \approx_d \mathrm{N}_p(0, J_n^{-1}). \tag{5}$$

The notation indicates approximate multinormality to the first order as the sample size grows, and can also be supplemented with a clear limit distribution statement, in that case involving a limit covariance matrix J for J_n. Consider now a candidate model M, different from the wide one, perhaps also in structure and form. With notation $f_M(y_i \mid x_i, \theta_M)$ for its density, and $u_M(y \mid x_i, \theta_M)$ for its score function, we have a maximum likelihood estimator $\widehat\theta_M$, of length p_M, maximizing the log-likelihood function $\ell_{n,M}(\theta_M) = \sum_{i=1}^n \log f_M(y_i \mid x_i, \theta_M)$. If the wide model is considered to be the truth, the estimator in model M does not necessarily aim at the true parameter, but at the least false parameter $\theta_{0,M,n}$, which is the minimizer of the Kullback–Leibler distance from the data-generating mechanism to the model; see details in **Appendix A3**. The estimator in the candidate model has a limiting multinormal distribution, with a sandwich type variance matrix,

$$\sqrt{n}(\widehat\theta_M - \theta_{0,M,n}) \approx_d \mathrm{N}_{p_M}(0, J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1}), \tag{6}$$

where

$$\begin{aligned} J_{M,n} &= -n^{-1} \sum_{i=1}^n \mathrm{E}_{\mathrm{wide}} \frac{\partial^2 \log f(Y_i \mid x_i, \theta_{0,M,n})}{\partial\theta_M\, \partial\theta_M^{\mathrm{t}}} \quad \text{and} \\ K_{M,n} &= n^{-1} \sum_{i=1}^n \mathrm{Var}_{\mathrm{wide}}\, u_M(Y_i \mid x_i, \theta_{0,M,n}). \end{aligned}$$

The variance matrices here are defined with respect to the wide model, at position θ_{true}.

From approximations (5–6) the delta method may be called upon to read off relevant expressions for the approximate distributions of the focus parameter estimators $\widehat\mu_{\mathrm{wide}} = \mu(\theta)$ and $\widehat\mu_M = \mu_M(\theta_M)$, where the latter is aiming for the least false parameter value $\mu_{0,M,n} = \mu_M(\theta_{0,M,n})$ associated with model M. Crucially, we also need a multinormal approximation to the *joint* distribution of $(\widehat\mu_{\mathrm{wide}}, \widehat\mu_M)$, in order to assess the distribution of the bias estimator $\widehat b_M = \widehat\mu_M - \widehat\mu_{\mathrm{wide}}$; without that part we can't build an appropriate estimator for b_M^2. In **Appendix A3**, we go through such arguments, and reach

$$\begin{pmatrix} \sqrt{n}(\widehat\mu_{\mathrm{wide}} - \mu_{\mathrm{true}}) \\ \sqrt{n}(\widehat\mu_M - \mu_{0,M,n}) \end{pmatrix} \approx_d \mathrm{N}_2(0, \Sigma_{M,n}). \tag{7}$$

Here the 2 × 2 matrix $\Sigma_{M,n}$ has diagonal terms $c^{\mathrm{t}} J_n^{-1} c$ and $c_{M,n}^{\mathrm{t}} J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} c_{M,n}$, with gradient vectors

$$c = \partial\mu(\theta_{\mathrm{true}})/\partial\theta \quad \text{and} \quad c_{M,n} = \partial\mu(\theta_{0,M,n})/\partial\theta_M$$

of lengths p and p_M. The off-diagonal term of $\Sigma_{M,n}$ takes the form $c^{\mathrm{t}} J_n^{-1} C_{M,n} J_{M,n}^{-1} c_{M,n}$, with a formula for the required covariance related term $C_{M,n}$ in the **Appendix**.

From (7) we can read off mse approximations,

$$\begin{aligned} \mathrm{mse}_{\mathrm{wide}} &\doteq c^{\mathrm{t}} J_n^{-1} c/n + 0^2 \quad \text{and} \\ \mathrm{mse}_M &\doteq c_{M,n}^{\mathrm{t}} J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} c_{M,n} + b_M^2, \end{aligned}$$

with bias $b_M = \mu_{0,M,n} - \mu_{\mathrm{true}}$. For the latter we use the estimator $\widehat b_M = \widehat\mu_M - \widehat\mu_{\mathrm{wide}}$, where the result above also leads to a clear approximation for the distribution of $\sqrt{n}(\widehat b_M - b_M)$. This leads to FIC formulae, unbiased and adjusted, as

$$\begin{aligned} \mathrm{FIC}^u_{\mathrm{wide}} &= \widehat c^{\mathrm{t}} \widehat J_n^{-1} \widehat c/n + 0^2 \quad \text{and} \\ \mathrm{FIC}^u_M &= \widehat c_M^{\mathrm{t}} \widehat J_M^{-1} \widehat K_M \widehat J_M^{-1} \widehat c_M/n + \widehat b_M^2 - \widehat\kappa_M^2/n, \\ \mathrm{FIC}_{\mathrm{wide}} &= \widehat c^{\mathrm{t}} \widehat J_n^{-1} \widehat c/n + 0^2 \quad \text{and} \\ \mathrm{FIC}_M &= \widehat c_M^{\mathrm{t}} \widehat J_M^{-1} \widehat K_M \widehat J_M^{-1} \widehat c_M/n + \max(\widehat b_M^2 - \widehat\kappa_M^2/n, 0). \end{aligned} \tag{8}$$

Here $\widehat c$ and $\widehat c_M$ emerge by computing gradients of $\mu(\theta)$ and $\mu_M(\theta_M)$ at their respective maximum likelihood positions, and $\widehat J_n, \widehat J_M$ are computed as normalized observed Fisher information matrices, for the wide and for the candidate model in question; specifically, $\widehat J_M$ is $1/n$ times minus the Hessian matrix from the log-likelihood, $-\partial^2 \ell_{n,M}(\widehat\theta_M)/(\partial\theta_M \partial\theta_M^{\mathrm{t}})$. Also, the $p_M \times p_M$ matrix $\widehat K_M$ is $n^{-1} \sum_{i=1}^n \widehat u_{M,i} \widehat u_{M,i}^{\mathrm{t}}$, with $\widehat u_{M,i} = u_M(y_i \mid x_i, \widehat\theta_M)$. Finally, the $\widehat\kappa_M^2/n$ estimates involves also the $p \times p_M$ matrix $\widehat C_M$, which is $n^{-1} \sum_{i=1}^n \widehat u_{\mathrm{wide},i} \widehat u_{M,i}^{\mathrm{t}}$. Model selection proceeds by computing FIC_M, the estimated mse of the focus estimator $\widehat\mu_M$, for all models M of interest, and then selecting that model for which this score is the lowest.

3.2. FIC for Generalized Linear Models, With a Fixed Wide Model

We illustrate this FIC machinery for one popular class of generalized linear models, namely the Poisson regression models. Generalizations to other generalized linear models are relatively immediate. Suppose therefore that we have count data y_i along with a covariate vector x_i of length p. For the fixed wide model we take the Poisson regression model with $y_i \sim \mathrm{Pois}(\xi_i)$, with $\xi_i = \exp(x_i^{\mathrm{t}}\beta)$ containing all covariate information; in particular, there is also a true parameter β_{true} there. Consider then an alternative candidate model M which instead takes the means to be $\xi_{M,i} = \exp(x_{M,i}^{\mathrm{t}}\beta_M)$, with $x_{M,i}$ of length p_M, perhaps a subset of the full x_i, or perhaps with some entirely other pieces of covariate information. Here the log-densities take the form $-\xi_i + y_i \log \xi_i - \log(y_i!)$, which means

$$\log f = -\exp(x_i^{\rm t}\beta) + y_i x_i^{\rm t}\beta - \log(y_i!) \quad \text{and}$$
$$\log f_M = -\exp(x_{M,i}^{\rm t}\beta_M) + y_i x_{M,i}^{\rm t}\beta_M - \log(y_i!),$$

for the wide model and the candidate model, along with score functions

$$u(y_i \,|\, x_i, \beta) = \{y_i - \exp(x_i^{\rm t}\beta)\}x_i \quad \text{and}$$
$$u_M(y_i \,|\, x_{M,i}, \beta_M) = \{y_i - \exp(x_{M,i}^{\rm t}\beta_M)\}x_{M,i}.$$

From this we deduce

$$J_n = n^{-1}\sum_{i=1}^n \exp(x_i^{\rm t}\beta_{\rm true})x_i x_i^{\rm t},$$
$$J_{M,n} = n^{-1}\sum_{i=1}^n \exp(x_{M,i}^{\rm t}\beta_{0,M,n})x_{M,i}x_{M,i}^{\rm t},$$
$$K_{M,n} = n^{-1}\sum_{i=1}^n \exp(x_i^{\rm t}\beta_{\rm true})x_{M,i}x_{M,i}^{\rm t},$$

along with the $p \times p_M$ covariance matrix $C_{M,n}$, defined as

$$n^{-1}\sum_{i=1}^n {\rm E}_{\rm wide}\,\{Y_i - \exp(x_i^{\rm t}\beta_{\rm true})\}x_i\{Y_i - \exp(x_{M,i}^{\rm t}\beta_{0,M,n})\}x_{M,i}^{\rm t}$$
$$= n^{-1}\sum_{i=1}^n \exp(x_i^{\rm t}\beta_{\rm true})x_i x_{M,i}^{\rm t}.$$

Consistent estimates of these population matrices are obtained by inserting $\widehat\beta_{\rm wide}$ for $\beta_{\rm true}$ and $\widehat\beta_M$ for $\beta_{0,M,n}$.

Notably, as long as there is a well-defined wide Poisson regression model, as assumed here, the framework is sufficiently flexible and broad to encompass also non-Poisson candidate models. Using the FIC apparatus involves working with log-likelihood functions and score functions for these alternative models, leading to different but workable expressions for the matrices $J_{M,n}$, $K_{M,n}$, $C_{M,n}$ above. The stretched Poisson models used in Schweder and Hjort (2016, Exercise 8.18) are a case in point; these allow both over- and underdispersion.

3.3. FIC for Linear Mixed Effects Models

Models with random effects, often called mixed effect models, are widely used in ecological applications. In Cunen et al. (in review)[1] FIC formulae have been developed for the class of *linear* mixed effect models (often abbreviated LME models). Here we will give a brief description of that approach, which also serves as a special case of the general FIC approach for a fixed wide model framework, see (8). Generalizations to classes of non-linear mixed effect models, and also to heteroscedastic situations where variance parameters depend on covariates, can be foreseen, following similar chains of arguments but involving more elaborations.

Suppose we have n observations of y_i, a vector of length m_i. The m_i datapoints within each y_i vector are assumed to be dependent, and will often correspond to data collected in the same space or time. Here we will refer to these data as belonging to the same *group*. Each y_i vector is associated with a regressor matrix X_i of dimension $m_i \times p$ for the fixed effects, and a design matrix Z_i of dimension $m_i \times k$ for the random effects. The linear mixed effects model takes the form

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with the $b_i \sim {\rm N}_k(0, D)$ independent of the errors $\varepsilon_i \sim {\rm N}_{m_i}(0, \sigma^2 I_{m_i})$. The model may also be represented as

$$Y_i \sim {\rm N}_{m_i}(X_i\beta, \sigma^2(I_{m_i} + Z_i D Z_i^{\rm t})), \tag{9}$$

and its parameters are $\theta = (\beta, \sigma, D)$. Note that the ordinary linear regression model is a special case, corresponding to $D = 0$. The log-likelihood contribution for this group of the data may be written

$$\ell_i(\theta) = -m_i \log\sigma - \tfrac{1}{2}\log|I_{m_i} + Z_i D Z_i^{\rm t}| - \tfrac{1}{2}(1/\sigma^2)(y_i - X_i\beta)^{\rm t}$$
$$\times (I_{m_i} + Z_i D Z_i^{\rm t})^{-1}(y_i - X_i\beta).$$

The combined log-likelihood $\sum_{i=1}^n \ell_i(\theta)$ leads to maximum likelihood estimators and hence also to $\widehat\mu_{\rm wide} = \mu(\widehat\beta_{\rm wide}, \widehat\sigma_{\rm wide}, \widehat D_{\rm wide})$ for any focus parameter $\mu = \mu(\beta, \sigma, D)$ of interest.

In applied situations we will spend efforts and call on biological knowledge to construct a well-motivated wide model, of the form (9). This wide model will typically be based on our knowledge of the system under study and, crucially, on how the data were collected. Quite often the resulting model could become *big*, in the sense that it includes a large number p of fixed effects and also a large number k of random effects. Assume, as we do throughout this paper, that our primary interest lies in the precise estimation of some focus parameter μ, which could be a function of the fixed effect coefficients β, and/or the variance components (σ, D). For such a $\mu = \mu(\beta, \sigma, D)$, can we find another model which offers more precise estimates of μ than $\widehat\mu_{\rm wide} = \mu(\widehat\beta_{\rm wide}, \widehat\sigma_{\rm wide}, \widehat D_{\rm wide})$ implied by the wide model?

FIC answers the question above; we can search among a set of candidate models for one giving more precise estimates of μ. In the simplest setting, the candidate model is defined with respect to the same n groups as in the wide model in (9), and we write

$$y_i \sim {\rm N}_{m_i}\big(X_{M,i}\beta_M, \sigma_M^2(I + Z_{M,i}D_M Z_{M,i}^{\rm t})\big).$$

This model has design matrices, $X_{M,i}$ and $Z_{M,i}$, potentially different from those of the wide model, and hence also a different set of parameters, say $\theta_M = (\beta_M, \sigma_M, D_M)$. Often, but not necessarily, the candidate model will involve subsets of the covariates (i.e., columns) included in X_i and Z_i, respectively. Let the covariate matrix $X_{M,i}$ have dimension $m_i \times p_M$, and $Z_{M,i}$ being $m_i \times k_M$. The focus parameter must then be represented properly inside the candidate model, as $\mu_M = \mu_M(\beta_M, \sigma_M, D_M)$, leading to the estimate $\widehat\mu_M = \mu_M(\widehat\beta_M, \widehat\sigma_M, \widehat D_M)$.

In order to work out FIC formulae, we first need to study the joint large-sample behavior of the estimator from the wide model

$\widehat{\mu}_{\rm wide}$ and the estimator from the candidate model $\widehat{\mu}_M$. This is as with Equation (7) in section 3.1, but the current framework is more complicated and needs further efforts. Such work is carried out in Cunen et al. (in review)[1], and lead to

$$\begin{pmatrix}\sqrt{n}(\widehat{\mu}_{\rm wide}-\mu_{\rm true})\\ \sqrt{n}(\widehat{\mu}_M-\mu_{0,M,n})\end{pmatrix}\approx_d \mathrm{N}_2(0,\Sigma_{M,n}),$$

with all quantities defined analogously to what is presented in section 3.1. These include matrices J_n, $J_{M,n}$, $K_{M,n}$, $C_{M,n}$ and gradient vectors c and $c_{M,n}$, defined similarly to those in section 3.1, but here involving more complicated details than for the plainer regression models worked with there.

This work then yields the same type of FIC formulae as for Equation (8), but with other recipes and formulae for the required estimators for the quantities mentioned. Regarding estimators for the matrices involved, we have three general possibilities: (i) working out explicit formulae and plug in the necessary parameter estimates; (ii) computing the matrices numerically, involving certain numerical integration details; (iii) via bootstrapping from the estimated wide model. In Cunen et al. (in review)[1] the first option is pursued, involving lengthy derivations of log-density derivatives and their means, variances, covariances, computed under the wide model. The resulting formulae are too long for this review, but are fast to compute. Options (ii) and (iii) have yet to be fully investigated, but will likely be fruitful when extending this FIC approach to the wider class of generalized linear mixed models (the so-called GLMMs).

The approach described here will be illustrated in section 4, but we first offer some comments of a more general nature. Readers familiar with linear mixed effects models will be aware that there are two different estimation schemes for models of this class, full maximum likelihood and so-called REML estimators, for restricted or residual maximum likelihood. The REML method takes the estimation of the fixed effects of the model into account when producing estimators of the variance parameters. For the computation of FIC scores the user might employ either maximum likelihood or residual maximum likelihood estimates, since these are large-sample equivalent; see for instance Demidenko (2013, Ch. 3). As with the general FIC formulae (8) there are two versions, the approximately unbiased estimates of risks and the adjusted ones. In Cunen et al. (in review)[1] it is argued that the unbiased version

$$\mathrm{FIC}^u_M = \widehat{c}^{\,t}_M \widehat{J}_M^{-1}\widehat{K}_M\widehat{J}_M^{-1}\widehat{c}_M/n + \widehat{b}^2_M - \widehat{\kappa}^2_M/n \qquad (10)$$

tends to work best for linear mixed effects models. The benefit of this version is that good candidate models with small biases earn more, compared to the wide model. Investigations show that the FIC formulae of (10) work well, in the sense that they accurately estimate the risk associated with the use of the different candidate models. The FIC formulae are based on large-sample arguments, which for the case of the linear mixed effects models involves approximations to normality when the number n of groups increases to infinity. These normal approximations work well as long as the full sample size $\sum_{i=1}^n m_i$ grows, particularly for functions of the linear mean parameters. More care is sometimes required when it comes to applications involving non-linear functions of both mean and variance parameters, as with estimates of probabilities $\mu = P(Y \geq y_0 \,|\, x_0, z_0)$.

4. APPLICATION: THE SLIMMING OF MINKE WHALES

Our second application story concerns the potential change in body condition of Antarctic minke whales over a period of 18 years. For a more thorough investigation consult Cunen et al. (in review)[2]. Questions treated there have been discussed in the Scientific committee of the International Whaling Commission (IWC) for a number of years, and a full consensus has not been reached. In the context of this review, therefore, the analysis below should be taken as an illustration, and not necessarily the last word on the topic of the decline in energy storage or body condition for the minke whales.

Using data from the Japanese Whale Research Program under Special Permit in the Antarctic (the so-called JARPA-1) we have studied the evolution of fat weight in Antarctic minke whales caught in 18 consecutive years, from 1988 and 2005. The main biological interest lies in whether or not the whales experienced a decline in body condition during the study period, and the dissected fat weight (in tons or kg) is taken to be a proxy for this body condition. Thus, there is a clear focus parameter in this application: the yearly decline in fat weight (which we will parametrize in a suitable fashion in the following).

The whales caught in each year are unevenly sampled with respect to a number of covariates, for instance sex, body length, age, and longitudinal region in the Antarctic ocean. Since all these covariates may influence body condition we need to include them in a model aiming at estimating the potential yearly decline in the response. Based on lengthy and detailed discussions in the Scientific Committee of the IWC, we have chosen a wide model within the class of linear mixed effect models, see section 3.3. In Cunen et al. (in review)[2] we have used considerable efforts to motivate the choice of covariates, interactions, and random effect terms in the wide model, but these arguments are outside the scope of the present article. In R-package-type notation, the wide model can be given as

$$\begin{aligned}\texttt{fatweight} \sim\ & \texttt{year} + \texttt{year}^2 + \texttt{bodylength} + \texttt{sex}\\ & + \texttt{diatom} + \texttt{date} + \texttt{date}^2 + \texttt{age}\\ & + \texttt{sex} * \texttt{diatom} + \texttt{diatom} * \texttt{date}\\ & + \texttt{diatom} * \texttt{date}^2 + \texttt{bodylength} * \texttt{sex}\\ & + \texttt{bodylength} * \texttt{date}\\ & + \texttt{bodylength} * \texttt{date}^2 + \texttt{sex} * \texttt{date}\\ & + \texttt{sex} * \texttt{date}^2\\ & + \texttt{bodylength} * \texttt{sex} * \texttt{date}\\ & + \texttt{bodylength} * \texttt{sex} * \texttt{date}^2\\ & + \texttt{age} * \texttt{sex}\\ & + \texttt{age} * \texttt{date} + \texttt{age} * \texttt{date}^2\\ & + \texttt{age} * \texttt{sex} * \texttt{date} + \texttt{age} * \texttt{sex} * \texttt{date}^2\\ & + \texttt{year} * \texttt{sex} + \texttt{year}^2 * \texttt{sex} + \texttt{region}\end{aligned}$$

$$\begin{aligned}&+\texttt{year}*\texttt{region}+\texttt{year}^2*\texttt{region}\\&+\texttt{sex}*\texttt{region}+\texttt{diatom}*\texttt{region}\\&+\texttt{region}*\texttt{date}\\&+\texttt{region}*\texttt{date}^2+(1+\texttt{date}\\&+\texttt{date}^2\,|\,\texttt{year}).\end{aligned}$$

The `region` covariate reflects three different geographical regions, associated with three regression coefficients summing to zero.

The model defined above has $p = 40$ fixed effect coefficients. The notation $(1 + \texttt{date} + \texttt{date}^2\,|\,\texttt{year})$ specifies the random effect structure; the groups are defined by a categorical version of the year variable (so $n = 18$), and the Z_i matrix has $k = 3$ columns (a column of ones for the intercept, date, and date squared). According to prior biological knowledge, `date` is assumed to be one of the most important effects governing the fat weight. The variable refers to the day of the season when each whale was caught, and since the whales are in the Antarctic to gain weight the coefficient related to date is expected to be large and positive. Also, the effect of date is expected to be different from year to year, possibly due to fluctuations in krill production. Hence, a random effect on `date` is included. We thus have a total of $40 + 1 + 6 = 47$ parameters to estimate. The total number of observations, i.e., $\sum_{i=1}^{n} n_i$, was 683.

As mentioned above the main interest, for discussions at several IWC meetings, has been the yearly decline in the `fatweight` outcome variable. Since we have a quadratic year term in our wide model, with that part taking the form $\beta_{\text{year}}x + \beta_{\text{year2}}x^2$ for year x, a natural definition of the yearly decline is $\mu = \beta_{\text{year}} + 2\beta_{\text{year2}}x_0$, with x_0 the mean year in the dataset. The focus parameter corresponds to the derivative of the mean response, with respect to year, and evaluated in this mean year time point. For candidate models with only a linear effect of year the parameter simplifies to β_{year} only. Furthermore, for those submodels where there is no year effect included, we have $\beta_{\text{year}} = 0$, a parameter value which then is estimated with zero variance but with potentially big bias. For this example, we have limited ourselves to investigating five candidate models only, in addition to using the wide model itself; see **Table 2**.

We do not actually expect the mean level of decline in energy storage to be either exactly linear or exactly quadratic over 18 years, but take this level of approximation to be adequate for the purpose, since the decline over time curve is not far from zero; also, our focus parameter is identical to the overall slope, the mean curve evaluated at the end point minus its value at the start point, divided by the length of time.

All the candidate models have a smaller number of fixed effects than the wide model. Note that the first candidate model M_1 has a more complex random effect structure than the wide model itself (with $k = 6$ giving a total of 21 random effect parameters). This choice also demonstrates that there is nothing in the formulae hindering us from having candidate models with more random effects (or also more fixed effects) than the wide model. When it comes to interpreting the results, it is usually more natural to choose the wide model to be the largest possible plausible model, however. The models M_2 and M_3 are very simple (with few fixed effects), and differ only in the their random effects. Model M_4 includes only the linear year effect in addition to a single random effect in the intercept. The last model, M_5, is the model without any year effect, so $\mu_{M_5} = 0$. With the present focus parameter, the FIC score of such a model will have zero variance and a bias which only depends on the estimated focus parameter in the wide model, and its estimated variance, so $\text{FIC}^u_{M_5} = (0 - \widehat{\mu}_{\text{wide}})^2 - \widehat{\kappa}^2_{\text{wide}}/n$, for the relevant κ^2_{wide}/n approximation to the variance of $\widehat{\mu}_{\text{wide}}$. Thus, further specification of M_5 is unnecessary; it includes all possible LME models without any year effect. As the candidate models worked with are not close enough to each other to warrant the use of the local neighborhoods framework, we use the “fixed wide model” approach.

TABLE 2 | Brief descriptions of the wide model and the five additional candidate models, with the number of fixed effects, the number of random effects, and the total number of parameters to be estimated, for each model.

	Description	*p*	*k*	*d*
M_0	Wide model	40	3	47
M_1	Less interactions, quadratic year effect	9	6	31
M_2	Very simple, linear year effect	5	2	9
M_3	Very simple, linear year effect	5	1	7
M_4	Only linear year effect	2	1	4
M_5	Like the wide, but without year effect	32	3	39

After carefully constructing our wide model, and checked that it passes various diagnostic tests, we can proceed to model selection with the FIC. The results are given in the form of a FIC-plot in **Figure 2**. We see that M_2 gets the lowest FIC score, with $\widehat{\mu} = -7.76$. The models M_1 and M_3 are close to the winning one, both in terms of their FIC scores and their estimates of the focus parameter. Model M_5, without the year effect had a considerably larger FIC score than any of the other models (which can be seen as an implicit test for the the null hypothesis of there being no year effect). From the plot we can conclude that our best estimate of the focus parameter is around −8 kg, or 80 kg loss of fat over a decade. Furthermore, since the root-FIC values are about 1.50, confidence intervals associated with these best point estimates will clearly fall to the left of zero. A natural interpretation of the FIC plot is therefore that the body condition decline, for the Antarctic minke whales, has been *negative and significant* over the study period.

To demonstrate the versatility of our approach, we have investigated the same six models with respect to another focus parameter, the probability of observing a whale with more than a certain amount of fat, say 1.5 tons (1,500 kg), given some covariate values: $\mu_2 = P(Y \geq 1.5\,|\,x_0, z_0)$. Here we chose to look at a 20 year old male whale, caught in 1991 in the eastern region, of approximately mean length (8 m), and which is caught toward the end of the season. Over the full dataset, the average fat weight of a whale is close to 1.5 tons. The FIC scores and estimates are given in **Figure 2**. We observe that the models give widely different estimates, ranging from around 0.50 to 0.90, and that the ranking of the models is very different from the ranking when the focus was the yearly decline in fat weight. The smallest model M_4 is considered the best for estimating the probability of

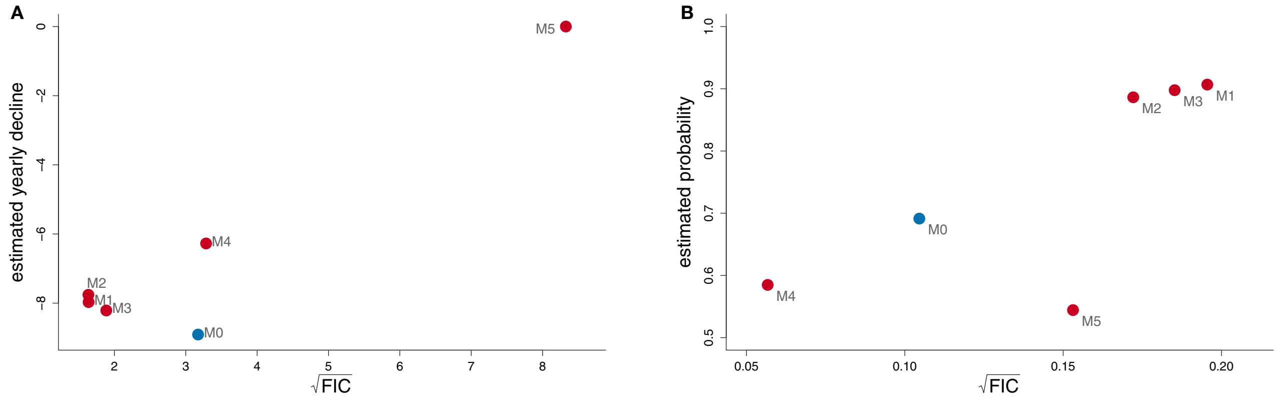


FIGURE 2 | (A) Estimates of the yearly decline in fat weight, for the Antarctic minke whale population (vertical axis), along with root-FIC scores (horizontal axis), for the wide model M_0, marked in blue, and five additional candidate models $M_1, \ldots, M_5$. The scale is in kilograms of fat. **(B)** Root-FIC scores and estimates of the probability of observing a whale with more than 1.5 tons of fat for the wide model (marked in blue) and the five candidate models.

observing a "medium fat" whale. Here, we see the typical bias-variance trade-off at work: using M_4 clearly gives an estimate with some bias compared to the wide model (estimate of 0.60 instead of around 0.70), but the bias is compensated for by a strong decrease in variance.

5. DISCUSSION

Our article has motivated, exhibited, developed, and extended the machinery of Focused Information Criteria for model selection and model ranking, with a few illustrations for ecological data. Here we offer some general remarks.

1. The role of the wide model. The FIC idea is to examine how different candidate models work regarding what they actually deliver, in terms of point estimates for the most crucial parameters of interest. This examination involves approximations to and estimates of the risks, which for the usual squared error loss function means mean squared error. Quantifying the implied variances and biases relies on the notion of a clearly defined (though unknown) data generating mechanism. This is one of the roles of our *wide model*. In the local asymptotics framework of **Appendix A1** this is the full model $f(y_i \,|\, x_i, \theta, \gamma)$ of (12), with $p + q$ parameters; in the alternative framework of section 3.1 it is what we term the fixed wide model. Such a wide model needs to be well argued, as being sufficiently rich to encompass the anticipated salient features of the phenomena studied. Since quantification and consequent estimation of variances and biases rest on the wide model being adequate it ought also to be given a goodness-of-fit verification, involving diagnostic checks etc.

One might inquire how sensitive the FIC scores are to the choice of the wide model. In connection with the application described in section 4 we have conducted some sensitivity checks and found that moderate changes to the wide model had little effect on the ranking of the different candidate models. Also, for the wide models we have investigated, the estimate of the focus parameter in the selected models was reasonably stable. More radical changes to the wide model should be expected to have greater effect, but we have not fully investigated this issue. Fully guarding against all misspecification of the wide model is unattainable, but extending our approach to even wider and more flexible wide models may lead to some improvements.

2. When should you use FIC? Practitioners may be interested in model selection for different, overlapping reasons. On one hand the goal might be to select the candidate model which in a relevant sense is the closest to the true data generating mechanism. Criteria based on model fit and some penalization for complexity aim at this goal, for instance the well-known AIC and BIC (see Claeskens and Hjort, 2008b) for a general discussion. On the other hand, practitioners often seek a small model offering precise estimates of the quantities they are interested in. It is important to keep in mind that FIC specifically aims at the second goal, and is not necessarily suitable for the first goal. FIC offers a principled way to *simplify* a large, realistic model which the user assumes to hold (i.e., to be realistically and adequately close to the complicated truth). The goal of the simplification is to obtain more precise estimates of quantities of interest, say $\widehat{\mu}$ for an underlying focus parameter μ. This also includes producing predictions for not yet seen outcomes of random variables, like the abundance of a certain species over the coming twenty years. Here simplification must be understood in a wide sense, as the candidate models do not necessarily need to be nested within the wide model, as we have seen. The two different motivations for model selection alluded to above partly relate to the two goals for statistical modeling: to explain or to predict, i.e., the "two cultures of statistics" (see Breiman, 2001; Shmueli, 2010). For yet further perspectives on model selection with focused views, coupled with model structure adequacy analysis (see Taper et al., 2008).

Once a practitioner has decided to use FIC, she then has to make a choice between the two FIC frameworks we have discussed, using local asymptotics or a fixed wide model. As a tentative guiding rule we advocate turning to the "fixed wide model" setup if the set of candidate models are seen as not being in a reasonable vicinity of each other. Also, we have seen that this framework allows candidate models of a different sort from the wide model; in particular, a candidate model does not have to be a clear submodel of the wide model. As stated before, the two frameworks aim at the same quantities, and the choice may thus also be guided by convenience. Note also that in many situations the two frameworks may give similar results. For the special case of linear regression models with focus parameters being linear functions of the coefficients, the formulae turn out to be identical. Also, for the classical generalized linear models, including logistic and Poisson regressions, the formulae yield highly correlated scores, as long as the focus parameters under study are functions of such linear combinations $x_0^{\rm t}\beta + z_0^{\rm t}\gamma$. For more complicated focus parameters, like probabilities for crossing thresholds, the answers are not necessarily close, and will depend on both the sample size and the degree to which the candidate models are not close.

3. Model averaging. Model averaging is sometimes used as an alternative to model selection to avoid the perhaps brutal throwing away of all but one model. With model averaging one computes the estimate of the focus quantity in all of the models separately and then forms a weighted average which is used as the final "model averaged" estimate of the focus. See for example the overview paper about model averaging in ecology by Dormann et al. (2018). Averaging estimates has as the advantage that all models are used. The flexibility of choosing the weights allows to give a larger weight to the estimate of a model that one prefers most. Weights could be set in a deterministic way, such as giving equal weights to all estimates, or could be data-driven. It makes sense to use values of information criteria to set the weights. Especially AIC has been popular (see Burnham and Anderson, 2002) for examples of the use of "Akaike weights." Also FIC could be used to form weights that are proportional to $\exp(-\lambda\,{\rm FIC}_M/{\rm FIC}_{\rm wide})$ for a user-chosen value of λ. One could also try to set the weights such that the mean squared error of the weighted estimator is as small as possible (Liang et al., 2011). Such theoretically optimal weights need to be estimated for practical use, which induces again estimation variability, and might lead to a more variable weighted estimator as when simple equal weights would have been used (Claeskens et al., 2016).

Model averaging with data-driven weights has consequences for inference similar to the post-selection inference (see below). Indeed, model selection may be seen as a form of model averaging, with all but one of the weights equal to zero and the remaining weight equal to one. Correct frequentist inference for model averaged estimators needs to take the correlations between the separate estimators into account, as well as the randomness of the weights in case of data-driven weights.

4. Post-selection issues. Model selection by the use of an information criterion (such as FIC, or AIC) comes with several advantages as compared to contrasting models two by two via hypothesis testing. With model selection there is no need to single out one model that would be placed in a null hypothesis. All models are treated equally. Multiple testing issues do not occur because no testing takes place. The set of models that is searched over can be large. The ease of calculating such information criteria makes it fast and allows to include many models in the search. However, there is a price to pay when one puts the next step to perform inference using the selected model. Simply ignoring that a model is arrived at via a selection procedure results in p-values that are too small and confidence intervals that are too narrow.

With a replicated study resulting in a dataset similar to but independent of the current one, it might happen that a different model gets selected, all the rest left unchanged. This illustrates that variability is involved in the process of model selection. One way to address such variability is via model averaging (see e.g., Hjort and Claeskens, 2003, Claeskens and Hjort, 2008b, Ch. 7, Efron, 2014). Berk et al. (2013) develop an approach for the construction of confidence intervals for parameters in a linear regression model that uses a selected model. Their approach is conservative, in the sense that the intervals tend to be wide and sometimes have a coverage that is quite a bit larger than the nominal value. Other approaches to take the uncertainty induced by the selection procedure into account is via selective inference leading to so-called "valid" inference. See, for example, Tibshirani et al. (2016, 2018). By using information about the specifics of the selection method such inference methods result in narrower confidence intervals as compared to the Berk et al. (2013) method. The effect of increasing the number of models results in getting larger confidence intervals (see Charkhi and Claeskens, 2018). Valid inference after selection is currently investigated for several model selection methods. It is to be expected that more results will become available in the future that guarantee that working with a selected model happens in a honest way that takes all variability into account.

It is well-known that estimators computed under a given model become approximately normal, under mild regularity conditions. It is however clear from the brief discussion above that post-selection and more general model-average estimators have more complicated distributions, as they often are non-linear mixtures of approximately normal distributions, with different biases, variances, and correlations. Clear descriptions of large-sample behavior, for even complex model-selection and model-average schemes, can be given inside the local asymptotics $O(1/\sqrt{n})$ framework of **Appendix A1**, as shown in Hjort and Claeskens (2003), Claeskens and Hjort (2008b, Ch. 7), with further generalizations in Hjort (2014). Inside the general framework of (12), with estimators $\widehat\mu_M$ as in (13), consider the combined or post-selection estimator

$$\widehat\mu^* = \sum_M \widehat w(M)\widehat\mu_M,$$

with data-driven weights $\widehat w(M)$ summing to one. If these are weights take the form $w(M\,|\,D_n)$, with $D_n = \sqrt{n}(\widehat\gamma_{\rm wide} - \gamma_0)$ as in (15), there is a very clear limit distribution,

$$\sqrt{n}(\widehat\mu^* - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widehat\delta(D)\}, \quad \text{where} \quad \widehat\delta(D) = \sum_M w(M\,|\,D) G_M D. \qquad (11)$$

This extends the master theorem result (17), to allow even for very complicated post-selection and model averaging schemes.

The $q \times q$ matrices G_M in this orthogonal decomposition are as in (16). The result remains true also for schemes based on weights involving AIC or FIC weights, as the appropriate weights can be shown to be close enough to the relevant $w(M \mid D_n)$. These limiting distributions can be simulated, at any position in the δ domain. Yet further efforts are required to turn such into valid post-selection or post-averaging confidence intervals, however see Claeskens and Hjort (2008b, Ch. 7) for one particular general (conservative) recipe, and for further discussion of these issues.

5. Performance. It is beyond the scope of this article to go into the relevant aspects of statistical performance of the FIC methods. One may indeed study both the accuracy of the final post-selection or post-averaged estimator, say for the $\widehat{\mu}^*$ above, and the probabilities for selecting the best models. Such questions are to some extent discussed in Hjort and Claeskens (2003) and Claeskens and Hjort (2008b, Ch. 7); broadly speaking, the FIC outperforms the AIC in large parts of the parameter space, but not uniformly. There are also several advantages with FIC, when compared with the BIC, regarding precision of the finally evaluated estimators. Notably, all of these questions can be studied accurately in the limit experiment alluded to above, where all limit distributions can be given in terms of the orthogonal decomposition $\Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\}$ of (11).

6. FIC for high-dimensional data. When models contain a large number of parameters, perhaps even larger than the sample size, maximum likelihood estimation might no longer be appropriate. The use of regularized estimators, such as ridge regression, lasso, scad, etc. requires adjustment to the FIC formulae. Even when the regularization takes automatic care of selection, Claeskens (2012) showed that selection via FIC is advantageous to get better estimators of the focus. Pircalabelu et al. (2016) used FIC for high-dimensional graphical models. For models with a diverging number of parameters FIC formulae using a so-called desparsified estimator have been obtained by Gueuning and Claeskens (2018). FIC may also be used to select tuning parameters for ridge regression. The focused ridge procedure of Hellton and Hjort (2018) is applicable to both the low and high-dimensional case and has been illustrated in linear and logistic regression models.

7. Extensions to yet other models. The methods exposited in section 3.1, yielding FIC machinery under a fixed wide model, can be extended to other important classes of models. The essential assumptions are those related to smooth log-likelihood functions and approximate normality for maximum likelihood estimators for the candidate models. Sometimes developing such FIC methods would take considerable extra efforts, though, as exemplified by our treatment in section 3.3 of linear mixed effects models. In particular, the methodology extends to models with dependence, as for time series and Markov chains with covariates (see Haug, 2019). This involves certain lengthier efforts regarding deriving expressions and estimation methods for the $K_{M,n}$ and $C_{M,n}$ matrices of (6, 7). Analogous FIC methods for time series are shown at work in Hermansen et al. (2016) for certain applications in fisheries sciences. Similar remarks also apply to the advanced Ornstein–Uhlenbeck process models used in Reitan et al. (2012) for modeling complex layered long-term evolutionary data. Specifically, these authors studied cell size evolution over 57 million years, and entertained 710 candidate models of this sort. An extension of our paper's FIC methods to their process models is possible and would lead to additional insights in their data.

A challenge of a different sort is to develop FIC methods also when the models used are too complicated for log-likelihood analyses, but where different estimation methods may be used. A case in point are models used in Dennis and Taper (1994), for dynamically evolving times series models of the form $y_{t+1} = y_t + a + b\exp(y_t) + \sigma z_t$, met in density dependence analyses for ecology. These models do not have stationary distributions and special estimation methods are needed to analyse the candidate models.

DATA AVAILABILITY STATEMENT

One of the datasets in this manuscript is not publicly available. The access to the minke whale dataset is controlled by the scientific committee of the International Whaling Commission. Requests to access the datasets should be directed to IWC Scientific Committee's Data Availability Group (DAG).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

Some of our FIC and AFIC calculations have been carried out with the help of the R library `fic`, developed by Christopher Jackson, see `github chjackson/fic`, also available on CRAN; some of our algorithms are extensions of his. CC and NH thank Kenji Konishi and the other scientists at the Institute of Cetacean Research for obtaining the body condition data for the Antarctic minke whale, the IWC Scientific Committee's Data Availability Group (DAG) for facilitating the access to these data, and Lars Walløe for valuable discussions regarding the modeling. GC acknowledges support of KU Leuven grant GOA/12/14, and CC and NH acknowledge partial support from the Norwegian Research Council through the FocuStat project 235116 at the Department of Mathematics, University of Oslo. Finally and crucially, the authors express gratitude to three reviewers, for their detailed suggestions, which led to a better and more clearly structured article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00415/full#supplementary-material

REFERENCES

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Stat.* 41, 802–837. doi: 10.1214/12-AOS1077

Breiman, L. (2001). Statistical modeling: the two cultures [with discussion contributions and a rejoinder]. *Stat. Sci.* 16, 199–215. doi: 10.1214/ss/1009213726

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edn*. New York, NY: Springer-Verlag.

Charkhi, A., and Claeskens, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* 105, 645–664. doi: 10.1093/biomet/asy018

Claeskens, G. (2012). Focused estimation and model averaging with penalization methods: an overview. *Stat. Neerland.* 66, 272–287. doi: 10.1111/j.1467-9574.2012.00514.x

Claeskens, G., and Hjort, N. L. (2003). The focused information criterion [with discussion contributions and a rejoinder]. *J. Am. Stat. Assoc.* 98, 900–916. doi: 10.1198/016214503000000819

Claeskens, G., and Hjort, N. L. (2008a). Minimizing average risk in regression. *Econometr. Theor.* 24, 493–527. doi: 10.1017/S0266466608080201

Claeskens, G., and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: a simple theoretical explanation. *Int. J. Forecast.* 32, 754–762. doi: 10.1016/j.ijforecast.2015.12.005

Cunen, C., Hjort, N. L., and Nygard, H. M. (2019). Statistical sightings of better angels: analysing the distribution of battle deaths in interstate conflict over time. *J. Peace Res.* (accepted).

Demidenko, E. (2013). *Mixed Models: Theory and Applications With R*. New York, NY: Wiley.

Dennis, B., and Taper, M. L. (1994). Density dependence in time series observations of natural populations: estimation and testing. *Ecol. Monogr.* 64, 205–224. doi: 10.2307/2937041

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., et al. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* 88, 485–504. doi: 10.1002/ecm.1309

Efron, B. (2014). Estimation and accuracy after model selection [with discussion contributions and a rejoinder]. *J. Am. Stat. Assoc.* 110, 991–1007. doi: 10.1080/01621459.2013.823775

Gueuning, T., and Claeskens, G. (2018). A high-dimensional focused information criterion. *Scand. J. Stat.* 45, 34–61. doi: 10.1111/sjos.12285

Hand, D. J., Daly, F., Lunn, A., McConway, K. J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Haug, J. (2019). *Focused model selection criteria for Markov chain models, with applications to armed conflict data* (Master Thesis). Technical report, Department of Mathematics, University of Oslo, Oslo, Norway.

Hellton, K. H., and Hjort, N. L. (2018). Fridge: focused fine-tuning of ridge regression for personalized predictions. *Stat. Med.* 37, 1290–1303. doi: 10.1002/sim.7576

Hermansen, G. H., Hjort, N. L., and Kjesbu, O. S. (2016). Recent advances in statistical methodology applied to the Hjort liver index time series (1859-2012) and associated influential factors. *Can. J. Fish. Aquat. Sci.* 73, 279–295. doi: 10.1139/cjfas-2015-0086

Hjort, N. L. (2014). Discussion of Efron's "Estimation and accuracy after model selection." *J. Am. Stat. Assoc.* 110, 1017–1020. doi: 10.1080/01621459.2014.923315

Hjort, N. L., and Claeskens, G. (2003). Frequentist model average estimators [with discussion and a rejoinder]. *J. Am. Stat. Assoc.* 98, 879–899. doi: 10.1198/016214503000000828

Jullum, M., and Hjort, N. L. (2017). Parametric of nonparametric: the FIC approach. *Stat. Sin.* 27, 951–981. doi: 10.5705/ss.202015.0364

Jullum, M., and Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Anal.* 25, 406–438. doi: 10.1007/s10985-018-9450-7

Ko, V., Hjort, N. L., and Hobæk Haff, I. (2019). Focused information criteria for copulae. *Scand. J. Stat.* doi: 10.1111/sjos.12387

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *J. Am. Stat. Assoc.* 106, 1053–1066. doi: 10.1198/jasa.2011.tm09478

Pircalabelu, E., Claeskens, G., Jahfari, S., and Waldorp, L. (2016). A focused information criterion for graphical models in fMRI connectivity with high-dimensional data. *Ann. Appl. Stat.* 9, 2179–2214. doi: 10.1214/15-AOAS882

Reitan, T., Schweder, T., and Hendericks, J. (2012). Phenotypic evolution studied by layered stochastic differential equations. *Ann. Appl. Stat.* 6, 1531–1551. doi: 10.1214/12-AOAS559

Schweder, T., and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge: Cambridge University Press.

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25, 289–310. doi: 10.1214/10-STS330

Taper, M. L., Staples, D. F., and Shepard, B. S. (2008). Model structure adequacy analysis: selecting models on the basis of their ability to answer scientific questions. *Synthese* 163, 357–370. doi: 10.1007/s11229-007-9299-x

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Stat.* 46, 1255–1287. doi: 10.1214/17-AOS1584

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Stat. Assoc.* 111, 600–620. doi: 10.1080/01621459.2015.1108848

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, CJ, declared a past collaboration with one of the authors, GC, to the handling editor.

frontiers in Ecology and Evolution

HYPOTHESIS AND THEORY
published: 22 November 2019
doi: 10.3389/fevo.2019.00447

Statistical Distances and the Construction of Evidence Functions for Model Adequacy

Marianthi Markatou and Elisavet M. Sofikitou*

Department of Biostatistics, SPHHP, University at Buffalo, Buffalo, NY, United States

Over the past years, distances and divergences have been extensively used not only in the statistical literature or in probability and information theory, but also in other scientific areas such as engineering, machine learning, biomedical sciences, as well as ecology. Statistical distances, viewed either as building blocks of evidence generation or as evidence generation vehicles in themselves, provide a natural way to create a global framework for inference in parametric and semiparametric models. More precisely, quadratic distance measures play an important role in goodness-of-fit tests, estimation, prediction or model selection. Provided that specific properties are fulfilled, alternative statistical distances (or divergences) can effectively be used to construct evidence functions. In the present article, we discuss an intrinsic approach to the notion of evidence and present a brief literature review related to its interpretation. We examine several statistical distances, both quadratic and non-quadratic, and their properties in relation to important aspects of evidence generation. We provide an extensive description of their role in model identification and model assessment. Further, we introduce an explanatory plot that is based on quadratic distances to visualize the strength of evidence provided by the ratio of standardized quadratic distances and exemplify its use. In this setting, emphasis is placed on determining the sense in which we can provide meaningful interpretations of the distances as measures of statistical loss. We conclude by summarizing the main contributions of this work.

Keywords: evidence functions, inference, kernels, model selection, quadratic and non-quadratic distances, statistical distances, statistical loss measures

OPEN ACCESS

Edited by:
Jose Miguel Ponciano,
University of Florida, United States

Reviewed by:
Nils Lid Hjort,
University of Oslo, Norway
Subhash Ramkrishna Lele,
University of Alberta, Canada

***Correspondence:**
Marianthi Markatou
markatou@buffalo.edu

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *17 February 2019*
Accepted: *06 November 2019*
Published: *22 November 2019*

Citation:
Markatou M and Sofikitou EM (2019)
Statistical Distances and the
Construction of Evidence Functions
for Model Adequacy.
Front. Ecol. Evol. 7:447.
doi: 10.3389/fevo.2019.00447

1. INTRODUCTION

What is evidence? The Oxford dictionary defines *evidence* as "the available body of facts or information indicating whether a belief or proposition is true or valid." The fundamental knowledge of a science or an art, which at the same time embeds basic philosophical principles, can also be characterized as evidence.

In the scientific world the concept of evidence is crucial as it accumulates all the pieces/sources of information one has at hand and can assess in a variety of ways to judge whether something is true or not. The term *statistical evidence* (Royall, 2004) refers to observations interpreted under a probability model. To reject or support a hypothesis we use data obtained from the phenomena that occur in the natural world or we perform experiments and combine/match with some background information, resources and scientific tools such as theories, tests and models.

How do we measure the strength of evidence? In statistics, different strategies have been suggested to measure the strength of evidence. *Fisher's method* (Fisher, 1935) uses extreme value probabilities known as *p-values* from several independent tests which consider the same null hypothesis. *Fisher's p-value tests* may provide a measure of evidence (Cox, 1977); however, only a single hypothesis is taken into consideration and no reference to any alternative hypothesis is provided. On the contrary, in *Neyman-Pearson tests* the decision rule is based on two competing hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1. This approach divides all the possible outcomes of the sample space into two distinct regions, the acceptance and the rejection region. The specific data values that lead to the rejection of H_0 form the rejection region. The aim is to define the best significant level a, that is the probability of rejecting the null hypothesis when in fact it is true. According to Lewin-Koh et al. (2004), *Neyman-Pearson tests* may not provide an appropriate measure of evidence, in the sense that a decision should be made between two hypotheses of which one is accepted and the other is rejected. As a result, minor data changes could alter the final decision making (Taper and Lele, 2004).

Under the *Bayesian framework*, the decision is made based on some prior probabilities which try to quantify the scientist's belief about the competing hypotheses. The stronger the scientist's belief is that a hypothesis is true, the higher probability this hypothesis is given. The use of Bayesian tests to measure the strength of evidence has raised questions as the priors' choice may not be objective (Lewin-Koh et al., 2004). Bayesianism as well as likelihoodism are both based on the same principle, the *law of likelihood* (Sober, 2008). *Likelihood* and, by extension, *likelihood ratio* are basic statistical tools used for the quantification of the strength of evidence. For instance, consider the case where there are two hypotheses $H_0 : \tau = m_{\theta_0}$ vs. $H_1 : \tau = m_{\theta_1}$; then, the likelihood ratio is defined by $L(\theta_0; x)/L(\theta_1; x)$. The likelihood ratio of H_0 vs. H_1 measures the strength of evidence for the first hypothesis H_0 vs. the second hypothesis H_1. A likelihood ratio takes values that are greater than or equal to zero; a value of one indicates that the evidence does not support one hypothesis over the other. On the other hand, a value of the likelihood ratio substantially greater than 1, indicates support of H_1 vs. H_0.

The evidential paradigm uses likelihood ratios as measures of statistical evidence for or against hypotheses of interest. Royall (1997, 2000, 2004) suggests that the use of likelihood ratio to quantify strength of evidence of one model over another. Although likelihood ratio is a useful measure of strength of evidence, it has some practical limitations. More precisely, it is sensitive to outliers and it requires the specification of a complete statistical model (Lele, 2004).

However, all the basic theories of inference and evidence described above have disadvantages. To overcome their drawbacks, these techniques have been extended to address the problem of multiple comparisons and composite hypotheses testing, as well as to deal with situations where nuisance parameters are present. In particular, Royall (2000) suggests the use of profile monitoring for evidential inference purposes when one has to cope with nuisance parameters and composite hypotheses. A further, though quite challenging, generalization would be the case of unequal nuisance parameter number between the compared models (Taper and Lele, 2004). Moreover, the idea of evidence and its measurement has been extended to model adequacy and selection problems.

Fundamental to scientific work is the use of models. In analyzing and interpreting data, the use of models, explicit or implicit, is unavoidable. Models are used to summarize statistical properties of data, to identify parameters, and to evaluate different policies. Where do models come from? The literature provides very little help on answering the question of model formulation, yet this is arguably the most difficult aspect of model building. Cox (1990) and Lehmann (1990) discuss this question and offer various classifications of statistical models. Following Cox (1990), we define a statistical model as:

(1) A specification of a joint probability distribution of a single random variable or a vector of random variables,
(2) A definition of a vector of parameters of interest, ideally such that each component of the vector has a subject-matter interpretation as representing some understandable stable property of the system under study, and
(3) At least an indication of or a link with the process that could have generated the data.

In this paper, we are not concerned with the origins of models. We take as given that a class of models $\mathscr{M}$ is under consideration and we are concerned with methods of obtaining evidence characterizing the quality of an aspect of model assessment, that is the adequacy of a model in answering questions of interest and/or our ability to perform model selection. Measuring model adequacy centers on measuring the model misspecification cost. Lindsay (2004) discusses a distance-based framework for assessing model adequacy, a fundamental tenet of which is that one is able to carry out a model-based scientific inquiry without assuming that the model is true and without assuming that "truth" belongs in the model class under investigation. However, we make the assumption that the "truth" exists and it is knowable given the presence of data. The evidence for the adequacy of the model is measured via the concept of a statistical distance.

We discuss therefore statistical distances as evidence functions in the context of model assessment. We show that statistical distances that can be interpreted as loss functions can be used as evidence functions. We discuss in some detail a specific class of statistical distances, called *quadratic distances*, and illustrate their use in applications. For ease of presentation, we only use simple hypotheses, however these measures can handle both simple and composite hypotheses. Methods based on distances compare models by estimating from data the relative distance of hypothesized models to "truth," and transform composite hypotheses into a model selection problem. Furthermore, if multiple models are available, all models are compared on the basis of the value of the distance to the truth (selecting the model with the lowest distance as the best supported by the evidence model). Section 2 presents the idea of an evidence function as introduced by Lele (2004) and Lindsay (2004). Section 3 illustrates the statistical properties of various statistical distances and discusses their potential in the context of model adequacy. Section 4 compares, theoretically, some of the presented distances, while section 5 provides illustrations and examples of use of a specific class of distances, the quadratic distances. Finally, section 6 offers discussion and conclusions.

2. EVIDENCE FUNCTIONS AND STATISTICAL DISTANCES

A generalization of the idea of the likelihood ratio as a measure of strength of evidence to the idea of comparing two different competing models by comparing the difference in disparities between the data and each competing model is discussed in Lele (2004). The author formulates a class of functions, called *evidence functions*, which can be exploited not only to characterize but also to measure the strength of evidence. It should be mentioned that the term evidence functions may have been introduced by Lele, but as we shall see later on the concept of such functions is not new. Incidentally, Royall (2000, pp. 8) defines implicitly the concept of evidence function. Lele (2004) made an attempt to provide a formal definition of evidence functions by describing in detail several intuitive conditions that such a function should satisfy. We briefly present these conditions below.

Let us denote by Θ the parameter space and by X the sample space. Provided that an evidence function measures the strength of evidence by comparing two parameter values (hypotheses) that are based on the observed data, the domain of the evidence function is $X \times \Theta \times \Theta$. A real-valued function of the form $h_n : X \times \Theta \times \Theta \to \mathbb{R}$ will be called evidence function. As an example of an evidence function, we offer the likelihood function which is a special case of the class of general statistical distances. Given an evidence function, one could have strong evidence of θ_1 over θ_2 if $h_n(X, \theta_1, \theta_2) < -K$, for some fixed $K > 0$. Alternatively, one could have strong evidence of θ_2 compared to θ_1 if $h_n(X, \theta_1, \theta_2) > K$, for some fixed $K > 0$ and weak evidence if $-K < h_n(X, \theta_1, \theta_2) < K$. Lele (2004) characterizes this as *indifference zone*. An evidence function should at the same time satisfy the following conditions:

C1. *Translation Invariance*
C2. *Scale Invariance*
C3. *Reparameterization Invariance*
C4. *Invariance Under Data Transformation*

The first condition is very important as it does not allow the practitioner to change the strength of evidence by adding a constant to the evidence function. The *translation invariance* of the evidence function as well as $h_n(X, \theta_1, \theta_1) = 0$ are implied due to the antisymmetric condition $h_n(X, \theta_1, \theta_2) = -h_n(X, \theta_2, \theta_1)$. Without the second condition, one can change the strength of evidence by simply multiplying an evidence function by a constant. The *scale invariance* property is ensured by the use of "standardized evidence functions" defined as $\tilde{h}_n(X, \theta_1, \theta_2) = h_n(X, \theta_1, \theta_2)/[I^{1/2}(\theta_1)I^{1/2}(\theta_2)]$, where the function $I(\theta_1)$ is assumed to be continuously differentiable up to second order and $0 < I(\theta_1) < \infty$, $I(\theta_1)$ is defined in **R5** below. The *reparameterization invariance* condition reassures that, given a function ψ (where $\psi : \Theta \to \Psi$ is a one-to-one mapping of the parameter space), the comparison between (θ_1, θ_2) and between the corresponding points in the transformed space (ψ_1, ψ_2) is identical. In simple words, the quantification of the strength of evidence cannot change by stretching the coordinate system. Finally, the fourth condition implies that if $g : X \to Y$ is a one-one onto transformation of the data and $\bar{g}(\cdot)$ is the corresponding transformation in the parameter, the evidence function satisfies the property $h_n(X, \theta_1, \theta_2) = h_n(Y, \bar{g}(\theta_1), \bar{g}(\theta_2))$. As a result, the comparison of evidence is not affected by changes in the measuring units.

Lele (2004) states that in order to obtain a reasonable evidence function, the probability of strong evidence in favor of the true hypothesis has to converge to 1 as the sample size increases. Consequently, he presents the following additional regularity conditions:

R1. $E_{\theta_1}(h_n(X, \theta_1, \theta_2)) < 0$ for all $\theta_1 \neq \theta_2$.

R2. $n^{-1}(h_n(X, \theta_1, \theta_2) - E_{\theta_1}(h_n(X, \theta_1, \theta_2))) \xrightarrow{P} 0$, given that θ_1 is the true value or the best approximating model.

R3. The evidence functions $h_n(X, \theta_1, \theta_2)$ are twice continuously differentiable and the Taylor series approximation is valid in the vicinity of the true value θ_1.

R4. The central limit theorem is applicable; this implies that there exists a function $J(\theta_1)$ such that $0 < J(\theta_1) < \infty$ and $n^{-1/2}\left(\frac{d}{d\theta} h_n(X, \theta_1, \theta)\big|_{\theta_1}\right) \xrightarrow{D} N(0, J(\theta_1))$.

R5. The weak law of large numbers is applicable and as a result $n^{-1}\left(\frac{d^2}{d\theta^2} h_n(X, \theta_1, \theta)\big|_{\theta_1}\right) \xrightarrow{p} -I(\theta_1)$, where $0 < I(\theta_1) < \infty$ and the function $I(\theta_1)$ is assumed to be continuously differentiable up to second order.

The first regularity condition implies that evidence for the true parameter is maximized on average at the true parameter only and not at any other parameter. The first and the second conditions impose that the probability of strong evidence in favor of the true parameter compared to any other parameter converges to 1 as the sample size increases $\big(P_{\theta_1}[h_n(X, \theta_1, \theta_2) < -K] \to 1$, for any fixed $K > 0\big)$; while the last three regularity conditions are just provided for facilitating analytical and asymptotic calculations.

Different evidence functions have been proposed in the literature that satisfy conditions **R1** and **R2**. For instance, the *log-likelihood-ratio evidence functions* which are sensitive to outliers. Additionally, *disparity-based evidence functions* such as functions based on the Kullback-Leibler disparity measure, functions based on Jeffreys's disparity measure (Royall, 1983) or functions based on Hellinger's distance satisfy the first two regularity conditions. The later functions are robust to outliers and they do not fail to maintain their optimality property (Lindsay, 1994). Evidence functions that overcome the problem of complete model specification are *the log-quasi-likelihood-ratio functions*. Indeed, as underlined by Lele (2004), additional evidence functions can be constructed based on *composite likelihood* (Lindsay, 1988), *profile likelihood* (Royall, 1997), *potential function* (Li and McCullogh, 1994) and *quadratic inference functions* (Lindsay and

Qu, 2000). Therefore, Lele (2004) uses statistical distances or divergencies as building blocks in the construction of evidence functions to carry out model selection. Lele (2004) compares evidence for two models by comparing the disparities between the data and the two models under investigation. We note here that, for simplicity reasons, we stated conditions **R1**–**R5** for the uni-dimensional parameter θ. However, the restriction to a uni-dimensional parameter is unnecessary −the d-dimenional case can be treated analogously.

Disparities or statistical distances (defined formally in section 3) can be used as evidence functions to study model assessment, that is, model adequacy and model selection problems if they can be interpreted as measures of risk. In this context, understanding the properties of the distance provides for understanding the magnitude of the incurred statistical risk when a model is used. Two components of error are important in this setting. One is due to model misspecification −this is the intrinsic error made because the model we use can never be true. The second is the parameter estimation error (Lindsay, 2004). Within this framework, we discuss in the next section the statistical properties of several statistical distances as measures of model adequacy.

In evidential statistics three quantities are of primary interest; the strength of evidence, expressed in terms of likelihood ratios of two hypotheses H_1 and H_2, the probability of observing misleading evidence, and the probability of weak evidence. The probability of observing misleading evidence is denoted by M and it is defined as the probability of the likelihood of H_2 over H_1 being greater than a threshold k, where the probability is calculated under H_1. The constant k is the lower limit of strong evidence. In other words, misleading evidence is strong evidence for a hypothesis that is not true. We would then like to have the probability of misleading evidence as small as possible. An additional measure introduced by Royall (1997, 2004) is the probability of weak evidence, defined as the probability that an experiment will not produce strong evidence for either hypothesis relative to the other.

In suggesting the use of statistical distances as evidence functions, we propose, in connection with the use of quadratic distances, a quantity analogous to the likelihood ratio. This quantity is the standardized ratio of the quadratic distance of the hypothesis H_2 over the quadratic distance of hypothesis H_1. The squared root of this quantity can be interpreted as measuring the strength of evidence against the hypothesis H_2 and can be used as a general strength of evidence function. Although we propose an exploratory device to visually depict the strength of evidence based on the aforementioned quantity, it may be of interest to study the behavior of error probabilities analogous to the probability of misleading evidence and weak evidence, associated with statistical distances. We conjecture that, under appropriate conditions, it is possible to calculate the probability of misleading evidence for at least evidence functions of the form suggested by Lele (2004, p. 198). These functions, using our notation, have the form $n[\rho(P_\tau, M_{\theta_1}) - \rho(P_\tau, M_{\theta_2})]$, where n is the sample size, P_τ is the true probability model and M_{θ_1}, M_{θ_2} are the models under the two hypotheses H_1 and H_2. Our current work consists of establishing conditions to carefully study these probabilities. Alternatively, one may be able to construct a confidence interval for the model misspecification cost along the lines suggested by Lindsay (2004).

3. STATISTICAL DISTANCES AS EVIDENCE FUNCTIONS IN MEASURING MODEL ADEQUACY

In this section, we examine several classes of statistical distances in terms of their suitability as evidence functions. After presenting preliminaries on models and model adequacy, we discuss statistical distances as evidence functions. Specifically, we present the class of chi-squared distances and their extension the class of quadratic distances, the class of probability integral transform based distances and the class of non-convex distances.

3.1. Preliminaries

We construct a probability-based framework that mimics the data generation process and it is reasonable in light of the collected data. Our goal for this framework is to allow one to incorporate all aspects of uncertainty into the assessment of scientific data. We call this framework *the approximation framework* (Lindsay, 2004) and offer a brief description of it below. The interesting reader can find details in Lindsay (2004) and Lindsay and Markatou (2002).

Our basic modeling assumption is that the experimental data constitute a realization from a random process that has probability distribution P_τ, where τ stands for "true." That is, the data generated from such a probability mechanism mimics closely the properties of data generated from an actual scientific experiment. We treat this modeling assumption as correct, hence there exists a $P_\tau \in \mathscr{P}$, where $\mathscr{P}$ is the class of all distributions consistent with the basic assumptions. Through a set of additional secondary assumptions, we arrive at a class of models $\mathscr{M} = \{M_\theta : \theta \in \Theta\} \subset \mathscr{P}$. The individual distributions, denoted as M_θ, are the model elements. Following Lindsay (2004), we resist the temptation of assuming that the true probability model P_τ belongs in $\mathscr{M}$. Instead, we take the point of view that P_τ does not necessarily belong in the model class under consideration. Therefore, there is a permanent model misspecification error present. Statistical distances can be used to measure the model misspecification error; they can reconcile the use of $\mathscr{M}$ while not believing it to be true, by allowing one to carry out statistical analysis using models only as approximations to P_τ. An important conceptual issue that is raised by the approximation framework relates to the question of the existence of a "true" distribution. Lindsay (2004) has addressed this issue and we are in agreement, thus we do not address this point here. However, it is important to address, albeit briefly, the use of parametric models since it is possible to carry out statistical analyses completely nonparametrically, without the use of any model.

Models and modeling constitute a fundamental part of scientific work. Models (deterministic or stochastic) are used in almost every field of scientific investigation. A very general statement is that we need models in order to structure our ideas and conclusions. Lindsay (2004) discusses the question of why we need to use models when we know that they can only provide approximate validity by offering examples where the use of models provides insights into the scientific problem under study. In general, we would like our models to offer parsimonious descriptions of the systematic variation, concise summary of the statistical (random) variation and point toward meaningful interpretation of the data. We continue to use models because we think, in some sense, that models are still informative if they approximate the data generating mechanism in a reasonable fashion. We take this as being a general justification for continuing to use concise models. But the word "approximation" needs a more formal examination. To do so, we use statistical distances as evidence measures that allow formal examination of the adequacy of a model.

3.1.1. Model Assessment

There are two aspects to the problem of model assessment. The first aspect corresponds to treating the scientific problem from the point of view of one fixed model $\mathscr{M}$. For example, $\mathscr{M}$ might be the family of binomial distributions, or the family of multivariate normal distributions that is used to model the experimental data of interest. In this case, model misspecification error occurs when we assume that P_τ, the true distribution, belongs to $\mathscr{M}$ when it does not. Our goal then is to measure the cost in uncertainty due to specification of a restricted statistical model $\mathscr{M}$ relative to the unrestricted global model. We call this *the model adequacy problem*.

A different type of problem occurs when there are multiple models of interest, indexed by a, say $\mathscr{M}_a$, and one is interested in selecting one or more models that are most descriptive for the process at hand. In this case, we are interested in minimizing the model misspecification error, and less interested in assessing the model misspecification error for the sake of determining overall statistical error. This problem is called *the model selection problem*.

Both model selection and model adequacy problems are closely linked because we are interested, in both cases, in assessing the magnitude of the model misspecification error. In this paper, we will focus on the model adequacy problem.

3.1.2. The Approximation Framework

The approximation framework (Lindsay and Markatou, 2002; Lindsay, 2004) is a statistical distance-based framework that allows one to carry out model-based inference in the presence of model misspecification error. This involves the construction of a loss function that measures both within model and outside model errors. The construction of this loss function (or statistical distance) is discussed in Lindsay (2004). In the model adequacy problem, we will need to define a loss function $\rho(P_\tau, M_\theta)$ that describes the loss incurred when the true distribution is P_τ but instead M_θ is used. Such a loss function will, in principle, tell us how far apart, in an inferential sense, the two distributions are.

If we adopt the usual convention that loss functions are nonnegative in their arguments, are zero if the correct model is used, and are taking larger values when the distributions are dissimilar, then $\rho(P_\tau, M_\theta)$ can be viewed as a distance between the two distributions. Generally, if F, G are two distributions such that $\rho(F, G) \geq 0$ and $\rho(F, F) = 0$, we will call ρ a *statistical distance*. As an example of a statistical distance, we mention the familiar likelihood concept. An extensively used distance in statistics is the Kullback-Leibler distance. The celebrated AIC model selection procedure is based on the Kullback-Leibler distance. Other examples include Neyman's chi-squared, Pearson's chi-squared, L_1 and L_2 distances, and Hellinger distance. Furthermore, additional examples of statistical distances can be found in Lindsay (1994), Cressie and Read (1984), and Pardo (2006). Note that we only require that the distance is non-negative. We do not require symmetry in the arguments because the roles of P_τ and M_θ (or generally $\mathscr{M}$) are different. Neither do we require the distance to satisfy the triangle inequality. Thus, our measures are not distances in the formal mathematical sense.

As a historical note, we mention that statistical distances or divergences have a large history and are defined in a variety of ways, by comparing distribution functions, density functions or characteristic and moment generating functions.

3.1.3. Model Misspecification and Decomposition of Model Fitting Error

Given a statistical distance between probability distributions represented by P_τ and M, we can define the distance from the model class $\mathscr{M}$ to the true distribution P_τ by

$$\rho(P_\tau, \mathscr{M}) = \inf_{M \in \mathscr{M}} \rho(P_\tau, M).$$

Therefore, the distance from a class of models $\mathscr{M}$ to the true distribution P_τ equals the smallest distance generated by an element of the model class $\mathscr{M}$. This is called *the model misspecification cost*. It corresponds to finding the minimal model misspecification cost associated with the elements in the model class $\mathscr{M}$. If the true model P_τ belongs in the model class $\mathscr{M}$, say it is equal to M_{θ_0}, and M has density M_θ, then $\rho(P_\tau, M)$ induces a loss function on the parameter space via the relation $L(\theta_0, \theta) \overset{def}{=} \rho(M_{\theta_0}, M_\theta)$. Therefore, if the true model belongs to the model class $\mathscr{M}$, the losses are strictly parametric (Lindsay, 2004).

However, if P_τ does not belong in the model class $\mathscr{M}$, the overall cost can be broken into two parts, as follows

$$\rho(P_\tau, M_{\hat{\theta}}) = \rho(P_\tau, M_{\theta_\tau}) + [\rho(P_\tau, M_{\hat{\theta}}) - \rho(P_\tau, M_{\theta_\tau})],$$

where $\rho(P_\tau, M_{\theta_\tau}) = \inf_\theta \rho(P_\tau, M_\theta)$, that is M_{θ_τ} defines the best model element in $\mathscr{M}$ that is closest to P_τ in the given distance. Furthermore, $\hat{\theta}$ is the estimator of θ representing the particular method of estimation used to obtain it.

The first term in the decomposition of the overall misspecification cost is an unavoidable error that arises from using $\mathscr{M}$. This is the model misspecification cost. The second term is nonnegative and represents the error made due to point estimation. This is *the parameter estimation cost* (Lindsay, 2004). The way to balance these two costs depends on the basic modeling goals.

In the problem of model adequacy discussed here, one has a fixed model and interest centers in measuring the quality of the approximation offered by the model. In this case, it makes sense to perform post-data inference on the magnitude of the statistical distance to see if the approximation of the model to the "true" distribution is "adequate" relative to some standard. In what follows, we present specific classes of statistical distances (or loss functions) that can be used to measure model adequacy, and hence they can be used (potentially) as evidence functions.

3.2. Statistical Distances as Evidence Functions for Model Adequacy

In this section, we study the characteristics, that is, the mathematical properties of statistical distances to assess their suitability as evidence functions for model adequacy. Our point of view is that the choice of an appropriate statistical distance to use as an evidence function for evaluating model adequacy will depend on the aspects of model fit that a researcher is most interested in and the ability of the statistical distance to have a clear interpretation as a measure of risk. We note that Lele (2004) constructed evidence functions of the form $h_n(x;\theta_1,\theta_2) = n\{\rho(p_n,p_{\theta_1}) - \rho(p_n,p_{\theta_2})\}$, where $\rho(\cdot;\cdot)$ is a disparity or statistical distance, $p_{\theta_1}, p_{\theta_2}$ are two discrete probability models indexed by the parameters θ_1, θ_2 and p_n is the empirical probability mass function. In this way, Lele generalizes the likelihood paradigm and argues that the disparity-based evidence functions, under appropriate conditions, satisfy the property of strong evidence. We now examine three broad classes of statistical distances with respect to their suitability as evidence functions for model adequacy. To indicate the versatility of the methods, we work with both, continuous and discrete distributions and denote by $\mathscr{X}$ the associated sample space.

3.2.1. The Class of Chi-Squared Distances

Define $\tau(t)$ to be the "true" distribution and $\mathscr{M} = \{m_\theta(t) : \theta \in \Theta\}$ be a model class such as $\tau \notin \mathscr{M}$, Θ is the parameter space such that $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$. If $\tau(t), m(t)$ are two discrete probability distributions the generalized chi-squared distances are defined as

$$\sum_t \frac{\left(\tau(t) - m(t)\right)^2}{a(t)},$$

where $a(t)$ is a suitable probability mass function (see Lindsay, 1994; Markatou et al., 2017). For example, when $a(t) = m(t)$ and $\tau(t) = d(t)$ the proportion of observations in the sample with value t, we obtain Pearson's chi-squared distance. Other choices of $a(t)$ result in different members of the chi-squared family.

The family of chi-squared distances has a very clear interpretation as a risk measure (Lindsay, 2004; Markatou et al., 2017). First, the chi-squared distance is obtained as the solution of an optimization problem with interpretable constraints. This result helps the interpretation of the chi-squared distance measures as well as our understanding of their robustness properties. To exemplify, note that Pearson's chi-squared can be obtained as

$$\sum_t \frac{\left(d(t) - m(t)\right)^2}{m(t)} = \sup_h \frac{[E_d h(X) - E_m h(X)]^2}{Var_m(h(X))}, \tag{1}$$

where $h(\cdot)$ is a function that has finite second moments. Furthermore, relationship (1) gives

$$\sum_t \frac{\left(d(t) - m(t)\right)^2}{m(t)} = \sup_h \frac{\left(\frac{1}{n}\sum h(X_i) - E_m h(X)\right)^2}{Var_m(h(X))} = \frac{1}{n}\sup_h Z_h^2,$$

that is, Pearson's chi-squared is the supremum of squared Z statistics. As such, Pearson's chi-squared cannot possibly be robust. On the other hand, Neyman's chi-squared distance given as $\sum_t (d(t) - m(t))^2/d(t)$ equals $(1/n)\sup_h t_h^2$, the supremum of squared t statistics and hence is more robust. In general, the chi-squared distances are affected by outliers. However, a member of this class, the symmetric chi-squared distance is obtained if we use in place of $a(t)$ the mixture $0.5m(t) + 0.5d(t)$, and provides estimators that are unaffected by outliers (see Markatou et al., 1998; Markatou et al., 2017; Markatou and Chen, 2018). An attractive characteristic of the symmetric chi-squared distance is that it admits a testing interpretation. For details, see Markatou et al. (2017).

The fact that it is possible to obtain the chi-squared distances as solutions of a certain optimization problem with interpretable as a variance constraint, allows us by analogy to the construction of Scheffé's confidence intervals for parameter contrasts, to interpret chi-squared distances as tools that permit the construction of "Scheffé-type" confidence intervals for models. Therefore, the assessment of the adequacy of a model is done via the construction of a confidence interval for the model.

In contrast with the class of chi-squared distances, distance measures that are used frequently in practice do not arise as solutions to optimization problems with interpretable as variance constraints. For example, the Kullback-Leibler distance or the Hellinger distance can be obtained as solutions of similar optimization problems but with constraints that are not interpretable as suitable variance functions (see Markatou et al., 2017). As such, their interpretation as measures of risk, as well as their suitability in constructing confidence intervals for models is unclear. However, we note here that there is a near equivalence between the Hellinger distance and chi-squared distance, therefore justifying the use of Hellinger distance as a measure for model adequacy.

A classical distance for continuous probability models that is very popular is the L_2 distance (Ahmad and Cerrito, 1993; Tenreiro, 2009) defined as

$$L_2^2(\tau, m) = \int [\tau(x) - m(x)]^2 dx. \tag{2}$$

While the L_2 distance is location invariant, it is not invariant under monotone transformations. Moreover, scale changes appear as a constant factor multiplying the L_2 distance. However, other features of L_2^2 may not be invariant.

3.2.2. General Quadratic Distances

Lindsay et al. (2008) introduce the concept of quadratic distance defined as

$$\rho_K(F, M) = \int\int K_M(x,y)d(F-M)(x)d(F-M)(y), \tag{3}$$

where $K_M(x, y)$ is a nonnegative definite kernel function that possibly depends on the model M and F corresponds to the distribution function of the unknown "true" model. An example of a kernel function that is quite popular as a smoothing kernel in density estimation is the normal kernel with smoothing parameter h. We note that quadratic distances are defined for both, discrete and continuous probability models. To calculate $\rho_K(F, M)$ we write it as

$$\rho_K(F, M) = K(F, F) - K(F, M) - K(M, F) + K(M, M), \quad (4)$$

where $K(A, B) = \int\int K_M(x, y)dA(x)dA(y)$. Since the true distribution F is unknown, a nonparametric estimator of F, $\hat{F}$, can be used. We call $\rho_K(\hat{F}, M)$ the *empirical distance* between $\hat{F}$ and M.

An example of a quadratic distance is Pearson's chi-squared distance. The kernel of this distance is given as

$$K(x, y) = \sum_{i=1}^{m} \frac{I(x \in A_i)I(y \in A_i)}{M(A_i)}. \quad (5)$$

Here, $I(\cdot)$ is the indicator function and $A_1, A_2, \ldots, A_m$ represent the partitioning of the sample space into m bins. The empirical distance is then given by

$$\sum_{i=1}^{m} \frac{(\hat{F}(A_i) - M(A_i))^2}{M(A_i)}, \quad (6)$$

where $M(A_i)$ indicates the probability of the $i-$th partition under the model M and $\hat{F}(A_i)$ is the corresponding empirical probability.

Lindsay et al. (2008) showed that in order to obtain the correct asymptotic distribution of the quadratic distance, the kernel K needs to be modified. This means that the kernel needs to be centered with respect to model M. Centering is also necessitated by the need to obtain, for a given kernel, uniquely defined distances. We define the *centered kernel* with respect to model element M by

$$K_{cen}(x, y) = K(x, y) - K(x, M) - K(M, y) + K(M, M), \quad (7)$$

where $K(x, M) = \int K(x, y)dM(x)$ and the remaining terms are defined analogously.

The centering of the kernel has the additional benefit to allow one to write the quadratic distance as

$$\rho_K(F, M) = \int\int K_{cen(M)}(x, y)dF(x)dF(y). \quad (8)$$

The two relations above [that is (7) and (8)] guarantee that the expectation of the centered kernel with respect to the true model is the same with $\rho_K(P_\tau, M)$, the distance between the true distribution P_τ and the model M. Furthermore, relationship (8) shows that, for a fixed model M, the empirical distance $\rho_K(\hat{F}, M) = K_{cen(M)}(\hat{F}, \hat{F})$ equals to

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{cen(M)}(\mathbf{x}_i, \mathbf{x}_j), \quad (9)$$

and hence it is easily computable. It can be calculated in a matrix form as $(\mathbf{1}^T \mathbb{K}_{cen(M)} \mathbf{1})/n^2$, where $\mathbf{1}^T = (1, 1, \ldots, 1)$ and $\mathbb{K}_{cen(M)}$ is a matrix with $ij-$th elements being equal to $K_{cen(M)}(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, 2, \ldots, n$.

We can also estimate unbiasedly the quadratic distance using the formula

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i=1}^{n} K_{cen(M)}(\mathbf{x}_i, \mathbf{x}_j), \quad (10)$$

where the notation $K_{cen(M)}$ indicates the centered, with respect to the model M, kernel. The fundamental distinction between V_n and U_n is the inclusion (in V_n) of the diagonal terms $K_{cen(M)}(\mathbf{x}_i, \mathbf{x}_i)$.

Fundamental aspects of the construction of quadratic distances are the kernel selection and the selection of the kernel's tuning parameter. This parameter in fact determines the sensitivity of the quadratic distance in identifying departures between the adopted model and the true model. Lindsay et al. (2014) offer a partial solution to the issue of kernel selection and an algorithm of selecting the tuning parameter h in the context of testing goodness-of-fit of the model M.

In section 5, we illustrate the use of quadratic distances in the model adequacy problem, through the use of an explanatory analysis device, which we call *the ratio of the standardized distances plot*. This plot is based on the idea that when the true model is not in the model class under consideration, the standardized quadratic distance distribution can be proved to be normal with mean zero and standard deviation $\sigma_h(F)$. One can then construct the quantities $\rho_K(F, M)/\sigma_h(F)$, where h is a tuning parameter of the kernel. If a variety models M_i are under consideration, one can compute an estimate of the quantity $\rho_K(F, M_i)/\sigma_h(F)$ for each M_i.

To estimate $\rho_K(F, M)/\sigma_h(F)$, we use the ratio $U_n(M)/\hat{\sigma}_h(U_n(M))$, where $\hat{\sigma}_h(U_n(M))$ is the exact variance of U_n under the true distribution F (estimated by $\hat{F}$, the empirical cumulative distribution function). The quantity $U_n(M_\ell)/\hat{\sigma}_h(U_n(M_\ell))$, $\ell = 1, 2, \ldots, L$ is computed for each of the L model elements under consideration. This quantity is the *standardized distance* corresponding to each model element M_ℓ.

The ratio of the standardized distances plot is a plot where the $x-$axis depicts different models M_ℓ, $\ell = 1, 2, \ldots, L$ and the $y-$axis depicts the squared root of the ratios

$$\frac{U_n(M_\ell)/\hat{\sigma}_h(U_n(M_\ell))}{U_n(M_k)/\hat{\sigma}_h(U_n(M_k))}, \quad \ell, k = 1, 2, \ldots, L, \quad \ell \neq k.$$

This plot is analogous to the likelihood ratio plot that we define as the plot of the standardized, by their maximum value, likelihood functions $L(H_i)$ vs. H_i. For more information about the use of standardized likelihood functions, we refer the interested reader to Blume (2002). We further discuss and interpret the introduced distance plot in section 5. It graphically presents the strength of evidence for the model M_k or the strength of evidence against the model M_ℓ, $\ell \neq k$.

When the ratio of the standardized distances is approximately 1, then both models M_k, M_ℓ fit the data equally well. A ratio greater than 1 indicates that the standardized distance in the denominator is smaller than the standardized distance of the numerator. Depending on the magnitude of this ratio, it indicates that model M_k provides a better fit than the model M_ℓ. The greater this ratio is, the stronger the evidence against model M_ℓ.

We close this section by noting that quadratic distances, as defined above, can be thought of as extensions of the class of chi-squared distances. They can be interpreted as risk measures, and certain distances exhibit robustness properties. Additionally, they are locally equivalent to Fisher's information. As such, they can be used as evidence functions.

3.2.3. Non-convex Statistical Distances and Probability Integral Transformation Distances

Prominent among the non-convex distance functions is the total variation distance defined as $V(\tau, m) = (1/2)\sum_t |\tau(t) - m(t)|$ when the probability distributions are discrete or $V(\tau, m) = (1/2)\int |\tau(t) - m(t)|dt$ when the probability distributions are continuous. An alternative representation of the total variation distance allows us to interpret it as a measure of risk and hence as a measure for model adequacy. A statistically useful interpretation of the total variation is that it can be thought of as the worst error we can commit in probability when we use the model m instead of τ. This error has maximum value of 1 that occurs when τ, m are mutually singular. Although the total variation distance can be interpreted as a risk measure assessing the overall risk of using a model m instead of the true but unknown model τ, it has several disadvantages including the fact that if $V(d, m_\theta)$ is used as an inference function it yields estimators of the parameter θ that are not normal when the model $\mathscr{M}$ is true. This is related to the pathologies of the variation distance described by Donoho and Liu (1988). On the other hand, of note here is that the total variation distance is locally equivalent to the Fisher information number, and it is invariant under monotone data transformations. Both of these are desirable properties for evidence functions. Further discussion of the properties of total variation can be found in Markatou and Chen (2018).

The mixture index of fit distance is a nonconvex distance defined as $\pi^*(\tau, \mathscr{M}) = \inf_{m \in \mathscr{M}} \pi^*(\tau, m)$, where $\mathscr{M}$ is a model class or model, and $\pi^*(\tau, m)$ is the mixture index of fit that is defined as the smallest proportion π for which we can express the model $\tau(t)$ as follows: $\tau(t) = (1 - \pi)m_\theta(t) + \pi e(t)$, where $m_\theta(t) \in \mathscr{M}$ and $e(t)$ is an arbitrary distribution. The mixing proportion π is interpreted as the proportion of the data that is outside the model $\mathscr{M}$. The mixture index of fit distance is closely related to total variation, and for small values of the total variation distance the mixture index of fit and the total variation distance are nearly equal. See Markatou and Chen (2018) for a mathematical derivation of the aforementioned result. The mixture index of fit has an attractive interpretation as the fraction of the population intrinsically outside the model $\mathscr{M}$, that is, the proportion of outliers in the sample. However, despite this attractive interpretation, the mixture index of fit does not provide asymptotically normal estimators in the case of $\mathscr{M}$ being the true model, hence exhibits the same behavior with the total variation when used as an inference function. This behavior makes it less attractive for use as an evidence function.

Many invariant distances are based on the probability integral transformation, which says that if X is a random variable that follows a continuous distribution function F, then $F(X) = U$ is a uniform random variable on (0,1). Thus, it allows a simple analysis by reducing our probabilistic investigations to the uniform random variables. One distance that is used extensively in statistics and can be treated using the probability integral transformation is the Kolmogorov-Smirnov distance, that is defined as

$$\rho_{KS}(P_\tau, M) = \sup_x |P_\tau(x) - M(x)|, \tag{11}$$

where P_τ, M are two probability models, with P_τ indicating the true model distribution and M indicating a model element. This distance can be thought of as the total variation analog on the real line and hence it can be interpreted as a risk measure.

Markatou and Chen (2018) show that the Kolmogorov-Smirnov distance is invariant under monotone transformations, and that it can be interpreted as the test function that maximizes the difference between the power and size when testing the null hypothesis of the true distribution $P_\tau = F$ vs. the alternative $P_\tau = M$. A fundamental drawback however of the Kolmogorov-Smirnov distance is that there is no obvious extension of the distance and methods based on it to the multivariate case. Attempts to extend the Kolmogorov-Smirnov test to two and higher dimensions exist in the literature (Peacock, 1983; Fasano and Franceschini, 1987; Justel et al., 1997), but the test based on the Kolmogorov-Smirnov distance is not very sensitive in, generally, establishing differences between two distributions unless these differ in a global fashion near their centers. Since, there is not a direct interpretation of these distances as risks measures when the model is incorrect, they are not attractive for use as evidence functions.

4. THEORETICAL COMPARISONS

We begin with some comparisons between different statistical distances. We choose to compare the quadratic distance with L_2-distance and the total variation (or L_1-distance). This choice is based on the popularity of L_1 and L_2-distances, as well as on the fact that L_2 is a special case of the quadratic distance.

To better understand how these distances behave, and before we apply those to data for judging the evidence for or against hypotheses of interest, we present explicit theoretical computations that aim to elucidate their performance as functions of various aspects of interest, such as mean and/or variability of distributions. To make the comparisons as clear as possible, we concentrate in the uni-dimensional case.

Assume that we are interested in choosing between two normal models for describing our data, and suppose those are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with respective cumulative distribution functions F_1 and F_2. We use two different scenarios, the case of equal variances: $\sigma_1^2 = \sigma_2^2 = 1$ and in the case of unequal variances: $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, for different values of the tuning parameter h ($h \in [0.5, 2]$) and the mean difference $\mu_1 - \mu_2$ ($\mu_1 - \mu_2 \in [0, 5]$). To compute the quadratic distance between the two normal models, we use a normal kernel with tuning parameter h^2. Therefore, the kernel is expressed as $K(x, y) = (1/2\pi h) \cdot \exp\big[-(x-y)^2/(2h^2)\big]$. This produces the quadratic distance between the two aforementioned normal distributions given by

$$\rho_K(F_1, F_2) = \frac{1}{\sqrt{2\pi}} \cdot \left\{ \frac{1}{\sqrt{2\sigma_1^2 + h^2}} + \frac{1}{\sqrt{2\sigma_2^2 + h^2}} - 2 \cdot \frac{\exp\Big[-\frac{1}{2} \cdot \frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2+h^2)}\Big]}{\sqrt{\sigma_1^2 + \sigma_2^2 + h^2}} \right\}.$$

Lele (2004) lists as one of the desirable properties of an evidence function the property of scale invariance. Quadratic distances can be made scale invariant, and the scale-invariant quadratic distance of the aforementioned two normal distributions is given by

$$\rho_K^{(inv)}(F_1, F_2) = 1 - 2 \cdot \frac{\frac{\exp\left[-\frac{1}{2}\cdot\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2+h^2)}\right]}{\sqrt{\sigma_1^2+\sigma_2^2+h^2}}}{\frac{1}{\sqrt{2\sigma_1^2+h^2}} + \frac{1}{\sqrt{2\sigma_2^2+h^2}}}. \qquad (12)$$

Notice that, when the two distributions are equal, the two means and variances are equal to the distance is 0. Furthermore, for fixed variances σ_1^2, σ_2^2, the quadratic distance between the two normal populations is an increasing function of the distance between their respective means. For completeness of the discussion, we also note that the L_2 distance is a special case of the quadratic distance when $h = 0$.

The aforementioned distances are presented in $3D-$plots as functions of h and $\mu_1 - \mu_2$ in **Figures 1**, **2**. The blue color indicates small values of the distances and the mean difference, while graduate changes in the color indicate larger values. The various values of h provide different levels of smoothness. In practice, the selection of this parameter is connected to the specific data analytic goals under consideration. For example, Lindsay et al. (2014) select h such that the power of the goodness-of-fit test is maximized (for details see Lindsay et al., 2014).

In the sequel, we take into account only the case of equal variances: $\sigma = \sigma_1^2 = \sigma_2^2 = 1$, and we plot three distances as a function of the mean difference $\mu_1 - \mu_2$ ($\mu_1 - \mu_2 \in [0, 5]$). The solid lines in **Figure 3** illustrate the scale-quadratic distance for two different values of the tuning parameter ($h = 0.5$ and $h = 1$). When the variances are equal, Equation (12) reduces to

$$\rho_K^{(inv)}(F_1, F_2) = 1 - \exp\left[-\frac{1}{2} \cdot \frac{(\mu_1 - \mu_2)^2}{(2\sigma^2 + h^2)}\right].$$

The non-solid lines illustrated in **Figure 3** represent the L_1 distance (also known as total variation), which is given by the formula

$$TV(F_1, F_2) = 1 - 2\Phi\left(-\frac{|\mu_1 - \mu_2|}{2\sigma}\right),$$

and the scaled L_2 distance, which, as mentioned before, can be derived from Equation (12) by setting $h = 0$.

The graphs illustrated in **Figures 1–3** were created using the *Wolfram Mathematica* 11.1 program. For this purpose and in order to calculate the values of the depicted points, code was written in *Wolfram Language* by exploiting the formulae presented above.

In summary, the quadratic distance between two normal populations is an increasing function of the difference between the two parameter means when the two normal populations have equal variability. The shape of the distance does depend on the smoothing parameter that is selected by the user and provides different levels of smoothing, with higher values of h to correspond to greater smoothing. On the other hand, smaller values of h produce quadratic distances that are closer (in shape) to the L_2 distance, for which $h = 0$.

The results presented in this section provide guidance on the performance of these distances in practical applications. The following section presents data examples with the purpose of illustrating these distances as evidence functions for model adequacy.

5. ILLUSTRATIONS AND EXAMPLES

In this section, we present different examples using both simulated and real-world data with two or six dimensions. Our aim is to provide illustrations related to distances computed under different models so that the interested reader will get a better understanding on how the evidence functions and distances work in practice. **Figures 4–6** and the numbers described in **Table 1** were generated using the *Wolfram Mathematica* 11.1 program exploiting multivariate formulae analogous to the (univariate) ones presented in section 4.

5.1. Example # 1

The purpose of this illustration is to understand the behavior of quadratic distances as measures of evidence for model adequacy in various data structures arising when experimental data are generated.

We generate a single sample of size $n = 400$ from a mixture of two bivariate normal distributions as follows; 200 data points follow a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$ (abbrev. $MVN_2(\mathbf{0}, \mathbf{I})$). Another 200 data points are generated form a bivariate normal with the same covariance matrix $\mathbf{I}$ and mean $\mu^T = (6, 8)$. The different hypotheses postulate that the data are from models M_i, $i = 1, 2, 3, 4, 5$, where M_1 corresponds to a bivariate normal with mean $\mu_1^T = (0, 0)$ and covariance matrix $\mathbf{I}$ and the remaining models are bivariate normal with covariance matrix $\mathbf{I}$ and corresponding means $\mu_2^T = (-1, -2)$, $\mu_3^T = (3, 4)$, $\mu_4^T = (6, 8)$ and $\mu_5^T = (10, 20)$. For each case, we compute an estimate of the distance $\rho_K(\hat{F}, M_i)$, $i = 1, 2, 3, 4, 5$ denoted by $U_n(M_i)$ and its

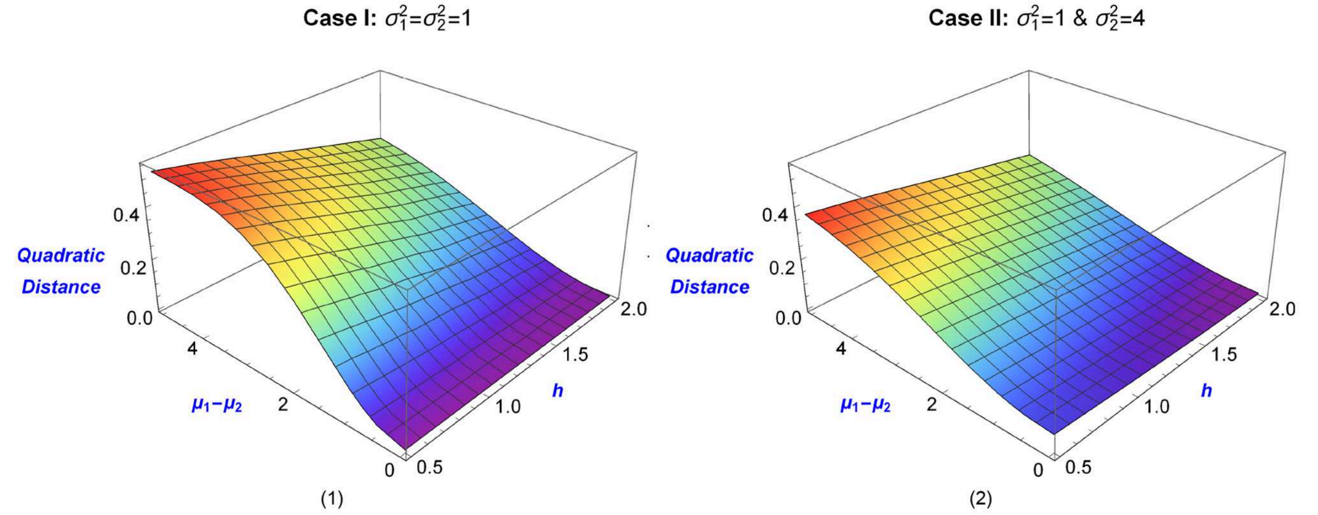


FIGURE 1 | *Quadratic Distances* of two univariate normal distributions, as a function of the tuning parameter *h* and the mean difference $\mu_1 - \mu_2$. Graph (1) shows the distance between $N(\mu_1, 1)$ & $N(\mu_2, 1)$, while Graph (2) shows the distance between $N(\mu_1, 1)$ & $N(\mu_2, 4)$. Blue color corresponds to small distances (as measured by the magnitude of $\mu_1 - \mu_2$), with the change of color indicating a larger difference in means.

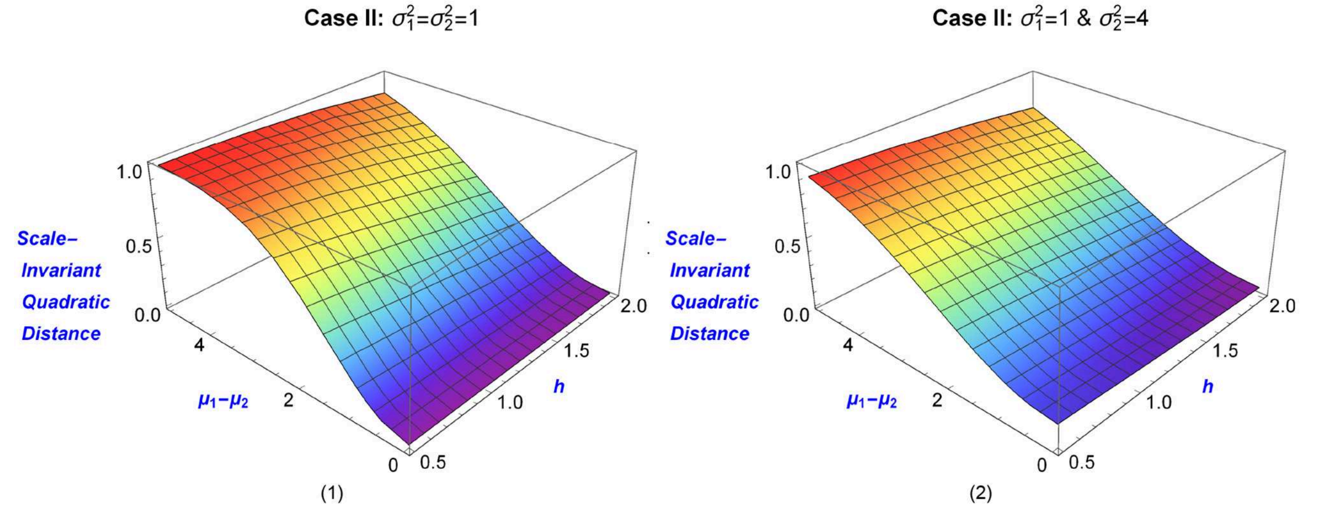


FIGURE 2 | *Scale-Invariant Quadratic Distances* of two univariate normal distributions, as a function of the tuning parameter *h* and the mean difference $\mu_1 - \mu_2$. Graph (1) shows the distance between $N(\mu_1, 1)$ & $N(\mu_2, 1)$, while Graph (2) shows the distance between $N(\mu_1, 1)$ & $N(\mu_2, 4)$. Blue color corresponds to small distances (as measured by the magnitude of $\mu_1 - \mu_2$), with the change of color indicating a larger difference in means.

associated variance. The kernel used to carry out the computation is the density of a multivariate normal with mean the observation $\mathbf{x_j}, j = 1, 2, \ldots, n$ and covariance matrix $h \cdot \mathbf{I}$. We use $h^2 = 0.5$.

Figure 4 plots the squared root of the standardized estimates of the quadratic distances between data expressed as $\hat{F}$ and the various fitted models. The plot indicates that models M_1 and M_4 provide an equally good fit to the data (the corresponding standardized distances are equal to 0.032), with the other models providing a worse fit to the data. This is actually expected because 50% of the sample comes from a bivariate normal with mean vector $\mathbf{0}$ (model M_1) and 50% of the sample comes from a bivariate normal with mean $\boldsymbol{\mu}_4$ (model M_4). The quadratic distance, interpreted as an evidence function, provides evidence that supports equally well the use of models M_1 and M_4.

Figure 5 presents a plot of the squared root of the standardized distance ($\sqrt{SD_{H_i}/SD_{H_1}}$) vs. the models fitted. This second plot is analogous to the log-likelihood plot of a hypothesis of interest vs. other competing hypothesis. The likelihood function is graphed to provide visual impression of the evidence over the parameter space. In analogy, we plot the ratio of the square root of the standardized distance for the different hypothesis over the standardized distance of the hypotheses of

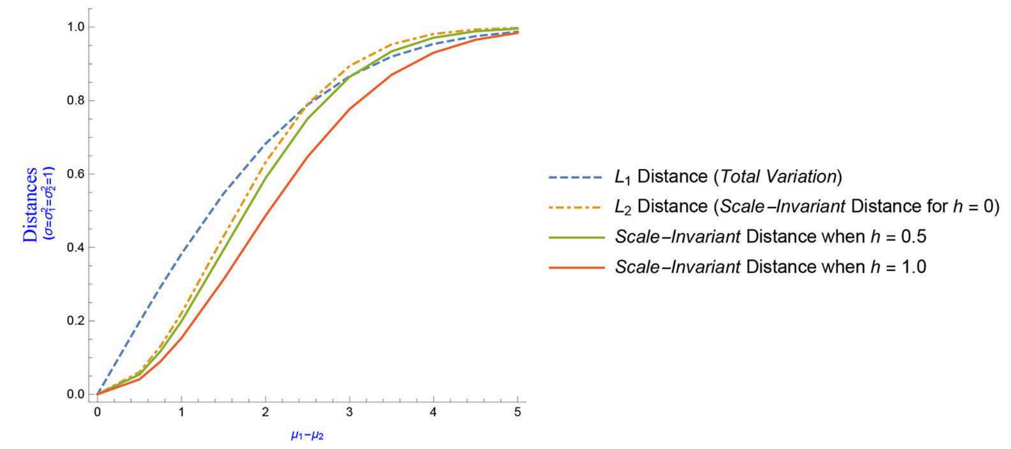


FIGURE 3 | L_1, L_2 and *Scale-Invariant Quadratic Distances* between two univariate normal distributions $N(\mu_1, 1)$ & $N(\mu_2, 1)$, as a function of the mean difference $\mu_1 - \mu_2$.

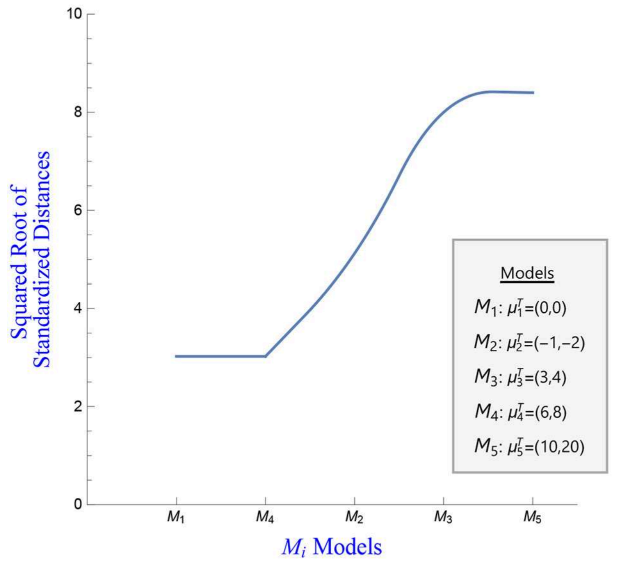


FIGURE 4 | Squared root of the standardized estimates of the distances between model $\hat{F}$ and the various fitted models.

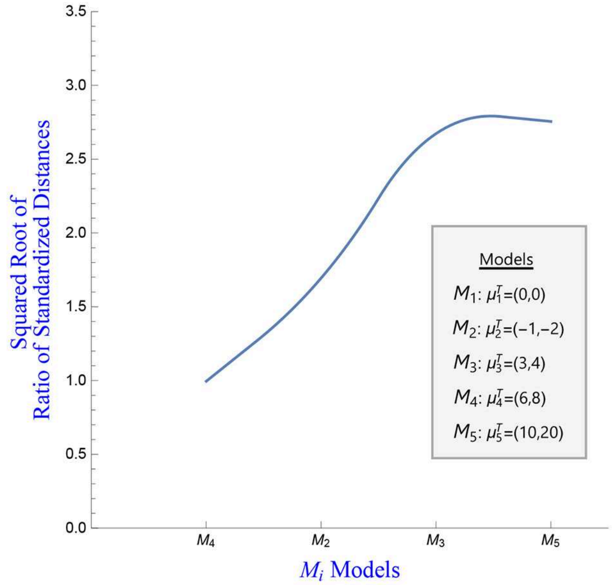


FIGURE 5 | Squared root of the ratio of the standardized distance vs. the various fitted models.

interest. Given that a small distance provides evidence of the model fit, the greater the value of the aforementioned ratio the stronger the evidence against the hypotheses H_i, $i = 2, \ldots, 5$. A ratio of approximately 1 indicates that both models are almost equally supported by the data, hence both hypotheses are approximately equally supported by the data, so the evidence does not indicate any preference for one hypothesis over the other. For example, the squared root of the standardized distance of model M_4 vs. model M_1 equals 0.995107, indicating that both models M_4 and M_1 are equally supported by the data. This is indeed the case since, by design, 50% of the data points come from $MVN_2(\mathbf{0}, \mathbf{I})$ with the remaining 50% of the data coming from a $MVN_2(\boldsymbol{\mu}, \mathbf{I})$, where $\boldsymbol{\mu}^T = (6, 8)$.

5.2. Example # 2

A second illustration of quadratic distances as evidence functions for model adequacy is provided below. We generate a single sample of size 250 from a $MVN_2(\mathbf{0}, I)$. We use this single sample as our baseline data and fit ten different models to obtain estimates of the standardized distances. The fitted models have a covariance matrix $\mathbf{I}$ and corresponding means as follows: $\boldsymbol{\mu}_0^T = (0, 0)$, $\boldsymbol{\mu}_1^T = (0.3, 0)$, $\boldsymbol{\mu}_2^T = (0.5, 0)$, $\boldsymbol{\mu}_3^T = (-3, 1)$, $\boldsymbol{\mu}_4^T = (1, 3)$, $\boldsymbol{\mu}_5^T = (3, 1)$, $\boldsymbol{\mu}_6^T = (-3, -2)$, $\boldsymbol{\mu}_7^T = (5, 4)$, $\boldsymbol{\mu}_8^T = (-5, -5)$ and $\boldsymbol{\mu}_9^T = (6, 9)$. We use $h^2 = 0.5$ and a normal kernel as before. **Table 1** presents the estimates of the distances for the different models and their associated standard deviations.

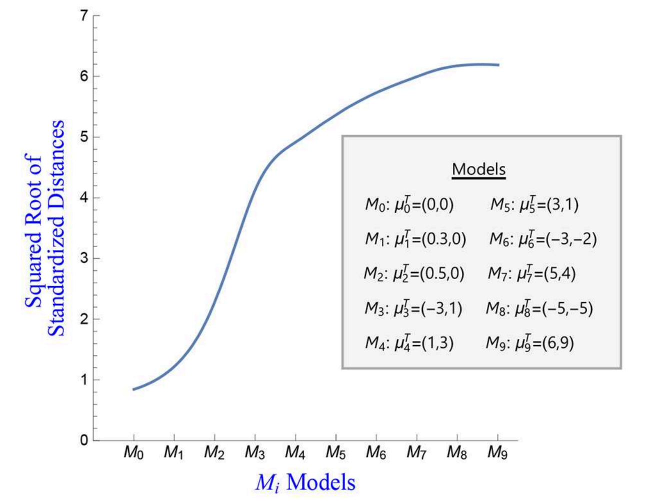


FIGURE 6 | Squared root of the standardized estimates of the distances between model $\hat{F}$ and the various fitted models.

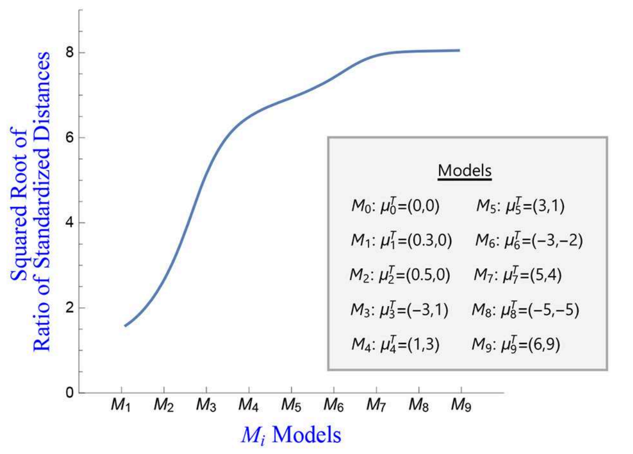


FIGURE 7 | Squared root of the ratio of the standardized distance vs. the various fitted models.

TABLE 1 | Estimates of the distances for ten different models and their associated standard deviations.

Models	Distances $U_n(M_i)$	Standard Deviations $\sigma_n(U_n(M_i))$
$M_0 : \boldsymbol{\mu}_0^T = \ (0,0)$	0.03772	0.06354
$M_1 : \boldsymbol{\mu}_1^T = (0.3,0)$	0.15479	0.11774
$M_2 : \boldsymbol{\mu}_2^T = (0.5,0)$	0.52944	0.18271
$M_3 : \boldsymbol{\mu}_3^T = (-3,1)$	10.95950	0.43770
$M_4 : \boldsymbol{\mu}_4^T = \ (1,3)$	11.10480	0.42908
$M_5 : \boldsymbol{\mu}_5^T = \ (3,1)$	11.16740	0.40285
$M_6 : \boldsymbol{\mu}_6^T = (-3,-2)$	11.94250	0.37171
$M_7 : \boldsymbol{\mu}_7^T = \ (5,4)$	12.78660	0.33466
$M_8 : \boldsymbol{\mu}_8^T = (-5,-5)$	12.79010	0.33421
$M_9 : \boldsymbol{\mu}_9^T = \ (6,9)$	12.79020	0.33422

A single sample of size 250 was used as the baseline sample coming from a $MVN_2(\mathbf{0}, \mathbf{I})$. Distances and their standard deviations are multiplied by 100.

Notice that when the mean of the fitted model is $\boldsymbol{\mu}_0^T = (0,0)$ the estimate of the distance is close to 0 with a very small standard deviation. Further, the more different the means are, the bigger the value of the distance estimate. **Figures 6**, **7** plot the squared root of the standardized distance estimates and the standardized ratio distance estimates. Interpretation of these plots is similar to the ones presented before.

5.3. Example # 3

This example uses a real experimental data set and illustrates that the quadratic distance evidence functions can be easily computed in higher than two dimensions and offer meaningful results. We use a multivariate data set introduced by Lubischew (1962). This data set contains three classes of *Chaetocnema*, a genus of flea beetles. Each class refers to a different type of species: *Chaetocnema Concinna Marsh*, *Chaetocnema Heikertingeri Lubisch*, and *Chaetocnema Heptapotamica Lubisch* of $n_1 = 21$, $n_2 = 31$ and $n_3 = 23$ instances each. Six features/characteristics were measured from each species: the width of the first and the second joint of the first tarsus in microns (the sum of measurements for both tarsi), the maximal width of the aedeagus in the fore-part (in microns), the front angle of the aedeagus (1 unit = 7.5°), the maximal width of the head between the external edges of the eyes (in 0.01 mm), the aedeagus width from the side (in microns).

In this example, we take two of the chaetocnema species, *Chaetocnema Heikertingeri Lubisch* and *Chaetocnema Heptapotamica Lubisch*. Measurements are taken on six dimensions. There are 31 observations in the first group of species and 22 observations in the second group (*Heptapotamica*), in total 53 observations. To estimate the mean vector μ_i and the covariance matrix Σ_i for each group we use the maximum likelihood. Each group, therefore, is described by a six-dimensional normal distribution with corresponding means given as $\boldsymbol{\mu}_{Hr}^T = (201, 119, 49, 125, 14, 81)$ for the *Heikertingeri* species and $\boldsymbol{\mu}_{Hp}^T = (138, 125, 52, 138, 10, 107)$ for the *Heptapotamica* species with their associated covariance matrices. In this case, we use the models $MVN_6(\mathbf{0}, \mathbf{I})$, $MVN_6(\mu_{Hr}, \Sigma_{Hr})$ and $MVN_6(\mu_{Hp}, \Sigma_{Hp})$ and computed their distance from the data set of 53 observations. Again, we used the multivariate normal kernel with $h = 0.1$. Notice that the standard multivariate normal model is also used in order to clearly indicate the difference in the values of the distance calculations. The fitting of the $MVN_6(\mu_{Hr}, \Sigma_{Hr})$ and $MVN_6(\mu_{Hp}, \Sigma_{Hp})$ offers estimators of the distance of 3.52×10^{-8} and 8.63×10^{-8}, while the fitting of $MVN_6(\mathbf{0}, \mathbf{I})$ gives a distance of 0.0005, indicating an estimate several orders of magnitude greater than the one obtained from the previous two cases. That is, the largest quadratic distance observed corresponds to the six-dimensional multivariate standard normal model. Furthermore, the squared root of the ratio of the standardized distances between the fitted *Heptapotamica* normal model (numerator) and the *Heikertingeri* normal model equals 1.57, implying that the evidence is inconclusive as to what model is supported. On the other hand, the corresponding quantities when the multivariate normal $MVN_6(\boldsymbol{0}, \boldsymbol{I})$ model is used in the numerator and the *Heptapotamica* model is used in the denominator is 76.11, and when the *Heikertingeri* is used the corresponding ratio is 119.18, clearly indicating that the data does not support the $MVN_6(\boldsymbol{0}, \boldsymbol{I})$ model.

6. DISCUSSION AND CONCLUSIONS

In this paper, we discuss the role of statistical distances as evidence functions. We review two definitions of evidence functions, one proposed by Lele (2004) and a second proposed by Lindsay (2004). We then examine the mathematical properties of some commonly used statistical distances and their suitability as evidence functions for model adequacy. Our investigation indicates that the class of the chi-squared distances and their extension, the class of quadratic distances introduced by Lindsay et al. (2008) and Lindsay et al. (2014) can be used as evidence functions for measuring model adequacy. This is because they can be interpreted as measures of risk, certain members of each class exhibit robustness properties, and if used as inference functions produce estimators that are asymptotically normal.

We propose also an explanatory analysis tool, namely the standardized distance ratio plot, that can be used to visualize the strength of evidence provided for, or against, hypotheses of interest and illustrate its use on experimental and simulated data. Our results indicate that quadratic distances perform well as evidence functions for measuring model adequacy. Furthermore, quadratic distances are of interest for a variety of reasons including the fact that several important distances are quadratic or they can be shown to be distributionally equivalent to a quadratic distance.

One of the reviewers raised the question of error probabilities associated with the use of statistical distances. Specifically, the reviewer asked whether the probabilities of misleading evidence and weak evidence are relevant in our context. We believe that measurement of model misspecification is an important step toward clarifying the suitability of a model class to explain the experimental data. However, we also think that a careful study of the behavior of these probabilities may shine additional light on distinguishing between different distances. The careful study of these questions is the topic of a future paper.

A second reviewer raised the question of potential connections of our work with work on the Focused Information Criterion (FIC) (Jullum and Hjort, 2017). The focus of our paper is on articulating the properties and illustrating, via data examples, the potential of statistical distances in assessing model adequacy. Connections with other model selection methods such as FIC will constitute the topic of future work. Finally, we would like to mention here that statistical distance concepts and ideas can be adapted to address model adequacy and model selection problems in many settings including linear, nonlinear and mixed effects models. Dimova et al. (2018) discuss in detail the case of linear regression and show that AIC and BIC are special cases of a general information criterion, the Quadratic Information Criterion (QIC).

Model assessment, that is, model adequacy and model selection is a fundamental and very important stage of any statistical analysis. Different techniques of model selection have been proposed in the literature describing how one could choose the best model among a spectrum of other competing models which best captures reality. However, provided that data were generated according to that specific model, the next logical step of a statistical analysis is to make statements about the study population. This implies making statistical inferences about the parameters of the chosen (data-dependent) model. Indeed, model selection strategies may have a significant effect or impact on inference of estimated parameters. Consequently, it is also crucial attention to be given to inference after model selection. For more information on estimation and inference after model selection, the interested reader is referred to Shen et al. (2004), Efron (2014), Fithian et al. (2017) and Claeskens and Hjort (2008, Chapters 6,7).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.jstor.org/stable/2527894?seq=1#metadata_info_tab_contents.

AUTHOR CONTRIBUTIONS

MM developed the structure of the paper and contributed to writing of the paper. ES searched the literature and contributed to writing of the paper.

FUNDING

The authors acknowledge financial support from the Troup Fund, KALEIDA Health Foundation, award number 82114, in the form of a grant award to MM.

REFERENCES

Ahmad, I. A., and Cerrito, P. B. (1993). Goodness of fit tests based on the l2-norm of multivariate probability density functions. *J. Nonparametr. Stat.* 2, 169–181. doi: 10.1080/10485259308832550

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Stat. Med.* 21, 2563–2599. doi: 10.1002/sim.1216

Claeskens, G., and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511790485

Cox, D. R. (1977). The role of significance tests. *Scand. J. Stat.* 4, 49–70.

Cox, D. R. (1990). Role of models in statistical analysis. *Stat. Sci.* 5, 169–174. doi: 10.1214/ss/1177012165

Cressie, N., and Read, T. R. (1984). Multinomial goodness-of-fit tests. *J. R. Stat. Soc. Ser. B (Methodol.)* 46, 440–464. doi: 10.1111/j.2517-6161.1984.tb01318.x

Dimova, R., Markatou, M., and Afendras, G. (2018). *Model Selection Based on the Relative Quadratic Risk*. Technical report, Department of Biostatistics, University at Buffalo, Buffalo, NY, United States.

Donoho, D. L., and Liu, R. C. (1988). Pathologies of some minimum distance estimators. *Ann. Stat.* 16, 587–608. doi: 10.1214/aos/1176350821

Efron, B. (2014). Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* 109, 991–1007. doi: 10.1080/01621459.2013.823775

Fasano, G., and Franceschini, A. (1987). A multidimensional version of the KolmogorovSmirnov test. *Month. Notices R. Astron. Soc.* 225, 155–170. doi: 10.1093/mnras/225.1.155

Fisher, R. A. (1935). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. *arXiv:1410.2597v4 [math.ST]*.

Jullum, M., and Hjort, N. L. (2017). Parametric or nonparametric: the fic approach. *Stat. Sin.* 27, 951–981. doi: 10.5705/ss.202015.0364

Justel, A., Peña, D., and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test for goodness of fit. *Stat. Probabil. Lett.* 35, 251–259. doi: 10.1016/S0167-7152(97)00020-5

Lehmann, E. L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Stat. Sci.* 5, 160–168. doi: 10.1214/ss/1177012164

Lele, S. R. (2004). "Evidence functions and the optimality of the law of likelihood (with comments and rejoinder by the author)," in *The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations*, eds M. P. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 191–216.

Lewin-Koh, N., Taper, M. L., and Lele, S. R. (2004). "A brief tour of statistical concepts," in *The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations*, eds M. P. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 3–16.

Li, B., and McCullogh, P. (1994). Potential functions and conservative estimating equations. *Ann. Stat.* 22, 340-356. doi: 10.1214/aos/1176325372

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* 80, 221–239. doi: 10.1090/conm/080/999014

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.* 22, 1081-1114. doi: 10.1214/aos/1176325512

Lindsay, B. G. (2004). "Statistical distances as loss functions in assessing model adequacy," in *The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations*, eds M. P. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 439–487.

Lindsay, B. G., and Markatou, M. (2002). *Statistical Distances: A Global Framework to Inference*. New York, NY: Springer.

Lindsay, B. G., Markatou, M., and Ray, S. (2014). Kernels, degrees of freedom, and power properties of quadratic distance goodness-of-fit tests. *J. Am. Stat. Assoc.* 109, 395–410. doi: 10.1080/01621459.2013.836972

Lindsay, B. G., Markatou, M., Ray, S., Yang, K., and Chen, S. C. (2008). Quadratic distances on probabilities: a unified foundation. *Ann. Stat.* 36, 983–1006. doi: 10.1214/009053607000000956

Lindsay, B. G., and Qu, A. (2000). *Quadratic Inference Functions*. Technical report, Pennsylvania State University.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics* 18, 455–477. doi: 10.2307/2527894

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *J. Am. Stat. Assoc.* 93, 740–750. doi: 10.1080/01621459.1998.10473726

Markatou, M., and Chen, Y. (2018). Non-quadratic distances in model assessment. *Entropy* 20:464. doi: 10.3390/e20060464

Markatou, M., Chen, Y., Afendras, G., and Lindsay, B. G. (2017). "Statistical distances and their role in robustness," in *New Advances in Statistics and Data Science*, eds D. G. Chen, Z. Jin, G. Li, Y. Li, A. Liu, and Y. Zhao (New York, NY: Springer), 3–26.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. London: Chapman and Hall/CRC.

Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Month. Notices R. Astron. Soc.* 202, 615–627. doi: 10.1093/mnras/202.3.615

Royall, R. M. (1983). *Theory of Probability*. New York, NY: Oxford University Press.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall.

Royall, R. M. (2000). On the probability of observing misleading statistical evidence (with discussion). *J. Am. Stat. Assoc.* 94, 760–780. doi: 10.1080/01621459.2000.10474264

Royall, R. M. (2004). "The likelihood paradigm for statistical evidence," in *The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations*, eds M. P. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 119–152.

Shen, X., Huang, H.-C., and Ye, J. (2004). Inference after model selection. *J. Am. Stat. Assoc.* 99, 751–762. doi: 10.1198/016214504000001097

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Taper, M. L., and Lele, S. R. (2004). "The nature of scientific evidence: a forward-looking synthesis," in *The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations*, eds M. P. Taper and S. R. Lele (Chicago, IL: The University of Chicago Press), 527–551.

Tenreiro, C. (2009). On the choice of the smoothing parameter for the bhep goodness-of-fit test. *Comput. Stat. Data Anal.* 53, 1038–1053. doi: 10.1016/j.csda.2008.09.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fevo.2019.00486

Second-Generation *P*-Values, Shrinkage, and Regularized Models

Thomas G. Stewart† *and Jeffrey D. Blume**†

Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, United States

OPEN ACCESS

Edited by:
Mark Louis Taper,
Montana State University System,
United States

Reviewed by:
Tobias Roth,
University of Basel, Switzerland
Ben Bolker,
McMaster University, Canada

****Correspondence:***
Jeffrey D. Blume
j.blume@vanderbilt.edu

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *15 March 2019*
Accepted: *29 November 2019*
Published: *17 December 2019*

Citation:
Stewart TG and Blume JD (2019) Second-Generation P-Values, Shrinkage, and Regularized Models. Front. Ecol. Evol. 7:486. doi: 10.3389/fevo.2019.00486

Second-generation *p*-values (SGPVs) are a novel and intuitive extension of classical *p*-values that better summarize the degree to which data support scientific hypotheses. SGPVs measure the overlap between an uncertainty interval for the parameter of interest and an interval null hypothesis that represents the set of null and practically null hypotheses. Although SGPVs are always in the unit interval, they are not formal probabilities. Rather, SGPVs are summary measures of when the data are compatible with null hypotheses (SGPV $= 1$), compatible with alternative hypotheses (SGPV $= 0$), or inconclusive ($0 <$ SGPV < 1). Because second-generation *p*-values differentiate between inconclusive and null results, their Type I Error rate converges to zero along with the Type II Error rate. The SGPV approach is also inferentially agnostic: it can be applied to any uncertainty interval about a parameter of interest such as confidence intervals, likelihood support intervals, and Bayesian highest posterior density intervals. This paper revisits the motivation for using SGPVs and explores their long-run behavior under regularized models that provide shrinkage on point estimates. While shrinkage often results in a more desirable bias-variance trade-off, the impact of shrinkage on the error rates of SGPVs is not well-understood. Through extensive simulations, we find that SPGVs based on shrunken estimates retain the desirable error rate behavior of SGPVs that we observe in classical models—albeit with a minor loss of power—while also retaining the benefits of bias-variance tradeoff.

Keywords: *p*-value, inference, bayes, shrinkage, regularization, second-generation *p*-value

INTRODUCTION

Despite decades of controversy, *p*-values remain a popular tool for assessing when observed data are incompatible with the null hypothesis. While *p*-values are widely recognized as imperfect, they continue to flourish in the scientific literature even when their shortcomings have real consequences. This reluctance to change occurs, in large part, because *p*-values are being used as quick-and-dirty summary assessments of the underlying data (instead of as a perfectly precise measure of evidence for a statistical hypothesis). In some cases, *p*-values are undeniably misused, abused and selectively misinterpreted. However, most researchers look to the *p*-value for an objective assessment of when the data are worthy of further detailed inspection. Given the large amount of information published on a daily basis, there is a critical role for a summary statistic to do just that. Blume et al. (2018, 2019) proposed the second-generation *p*-value (SGPV) as an improved *p*-value as used in practice. The SGPV is intended to serve as a summary measure of the data at hand, regardless of the statistical approach.

The SGPV is a formalization of today's best practices for interpreting data. According to the American Statistical Association (Wasserstein and Lazar, 2016), "best practice" amounts to de-emphasizing the magnitude of the *p*-value and inspecting the associated uncertainty interval (typically a confidence interval) to see if contains only scientifically meaningful effects. That is, researchers are supposed to check to see if the uncertainty interval rules out the null hypothesis and all other trivial, scientifically uninteresting effects. The problem with this approach is that it is *post-hoc*; the assessment of scientific meaningfulness is conducted after examining the data. As a result, the researcher's *post-hoc* assessments are influenced by results at hand, and this leads to the embellishment of effectively inconclusive data that supports practically null effects simply because the classical *p*-value is small. Given this, it should be no surprise that many "findings" fail to replicate; those "findings" were often mischaracterized in the first place.

A straightforward remedy for this is to require researchers to specify interesting and uninteresting effect sizes before the data are collected. This is routinely done in clinical trials, for example. The observed results can then be contrasted against initial benchmarks, uncorrupted by the observed data. Findings that fail to meet those benchmarks should still be reported, of course. But now they will be correctly reported as exploratory results and not as confirmatory ones. This is a critical step toward reproducibility: requiring the experimenter to define what constitutes a "successful experiment" before data are collected and interpreted.

The second-generation *p*-value (SGPV) is an improved *p*-value that has been adapted to this new level of exactness. It depends on the researcher's a priori definition of what constitutes an interesting or uninteresting effect and it indicates when the experiment has met that pre-specified benchmark. Blume et al. (2018, 2019) showed that this formalization leads to improved statistical properties in terms of a reduced Type I Error rate (it converges to zero as the sample size grows, much like the Type II error rate) and reduced false discovery rates.

The SGPV also depends on an uncertainty interval that characterizes the effect sizes that are supported by the data. Blume et al. (2018) shows how the SGPV's frequency properties are derived from the uncertainty interval. Blume et al. (2018, 2019) show than if a $(1-\alpha)100\%$ confidence interval or properly calibrated likelihood support interval is used, then the SGPV has desirable error rate behavior, with a Type I Error rate that remains bounded by α. The SGPV can just as easily be based on a Bayesian credible interval. The ability to incorporate an uncertainty interval from any of the three inferential schools of thought is why the SGPV is "method-agnostic." This also highlights the SGPVs role as a global indicator of when the study has collected sufficient data to draw conclusions, regardless of the underlying inferential approach used in the analysis.

In this paper, we examine what happens when the uncertainty interval upon which the SGPV is based comes from a model that is regularized. This is most easily thought of as using a Bayes or credible interval with a pre-specified prior. The Bayes approach provides shrinkage, which often results in reduced mean square error because the added bias is offset by a larger reduction of variance (confidence intervals, on the other hand, are routinely based on unbiased estimates). The question of interest is what happens to the frequency properties of the SGPV when uncertainty intervals are derived from a procedure that adds bias to reduce the variance. We investigate this by examining the behavior of SGPVs based on Bayes uncertainty intervals in a variety of simulations. We find that the SGPV easily incorporates these intervals while maintaining the improved Type I/Type II Error rate tradeoff. That is, the Type I error rate still converges to zero and the associated reduction of power tends to be minor. As a result, SGPVs based on Bayes intervals are similarly reliable inferential tools.

BACKGROUND: THE SECOND-GENERATION *P*-VALUE

Blume et al. (2018) present **Figure 1** (below) to illustrate the SGPV. The top diagram depicts the typical scenario: an estimated effect (denoted by $\hat{H}$), the traditional null hypothesis (denoted by H_0) and a confidence interval (CI) for the uncertainty interval. Here we take the uncertainty interval to be a collection of hypotheses, or effect sizes, that are supported by the data by some criteria (in this case at the 95% level). Classical hypothesis testing follows by simply checking if H_0 is in the CI or not.

There will always be a set of distinct hypotheses that are close to the null hypothesis but are scientifically inconsequential. This group represents null effects and practically null results, which we sometimes call trivial hypotheses, so it makes sense to group them together. This collection of hypotheses represents an indifference zone or interval null hypothesis. The bottom diagram depicts what happens when the null hypothesis is an interval instead of a single point. The null zone contains effect sizes that are indistinguishable from the null hypothesis, due to limited precision or practicality.

An interval null always exists, even if it is narrow, which is why the inspection of a CI for scientific relevance is essential and considered best practice. It is not sufficient to simply rule out the mathematically exact null; one must also rule similarly inconsequential scientific hypotheses/models. At its core, the problem of statistical significance not implying clinical significance boils down to this very issue. It is a matter of scale; the SGPV forces the experimenter to anchor that scale. As we will see, the reward for doing this is a substantially reduced false discovery rate.

Note that the experimental precision, which is finite, can serve as a minimum set for the interval null hypothesis. Finite experimental precision means there is some resolution along the x-axis (**Figure 1**) within which it is impossible to distinguish between hypotheses. This is a constraint of the experimental design, not the statistical methods. For example, when measuring income, hypotheses differing by <1 cent cannot be compared because the data on income are only measured to within 1 cent. Hypotheses differing by <1 cent are within the fundamental measurement error of the experiment. Typically, however, we are interested in hypotheses that are less precise than the experimental precision, e.g., income differences at the level of 1

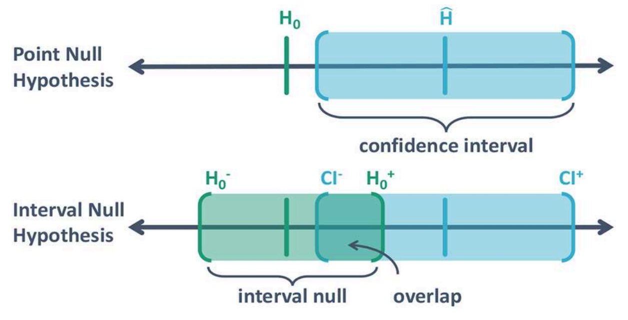


FIGURE 1 | Illustration of a point null hypothesis, H_0; the estimated effect that is the best supported hypothesis, $\hat{H}$; the a confidence interval (CI) for the estimated effect $[CI^-, CI^+]$; and the interval null hypothesis $\left[H_0^- \; H_0^+\right]$.

dollar. It is this scientific determination that sets the indifference zone around the null interval.

The SGPV is a scaled measurement of the overlap between the two intervals. If there is no overlap, the SGPV is zero. The data only support meaningful non-null hypotheses. If the overlap is partial, so that some of hypotheses supported by the data are in the interval null and some are out, we say the data are inconclusive. The degree of inconclusivity is directly related to the degree of overlap. But the general message is clear: more data are required for a definitive result. If the uncertainly interval is completely contained within the null zone—so the SGPV is 1—then the data support only null or scientifically trivial effects. This is how the SGPV indicates support for alternative hypotheses or null hypotheses, or indicates the data are inconclusive.

An important side note is that a SGPV of 0 or 1 is an endpoint in the sense that the study has completed its primary objective. It has collected sufficient information to screen out/in the null hypothesis. This does not imply that the data have achieved sufficient precision for policy implementation; the resulting uncertainty intervals can still be wide.

Formally, let interval I represent an uncertainty interval, e.g., a 95% CI or 95% credible interval, and let H_0 represent the interval null hypothesis. If $I = \left[a, b\right]$ where $a < b$ are real numbers, then its length is $|I| = b - a$. The second-generation p-value, denoted by p_δ, is defined as

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2\,|H_0|}, 1\right\} \qquad (1)$$

where $I \cap H_0$ is the overlap between intervals I and H_0. The subscript δ signals the reliance of the second-generation p-value on an interval null. Often δ represents the half-width of the interval null hypothesis. The value of δ is driven by scientific context and should be specified prior to conducting the experiment. The SGPV is often referred to as "p-delta."

The first term in Equation (1) is the fraction of best supported hypotheses that are also null hypotheses. The second term is a small-sample correction factor, which forces the second-generation p-value to indicate inconclusiveness when the observed precision is insufficient to permit valid scientific inferences about the null hypotheses. The second term applies whenever the uncertainty interval is more than twice as long as the null interval. It is this device that allows the SGPV to distinguish inconclusive results from those that support the null premise. See Blume et al. (2018) for a discussion of the correction factor. When the uncertainty interval is a traditional confidence interval, it is straightforward to determine the error rates and subsequent false discovery rates. Blume et al. (2018, 2019) provide these computations. Here we consider what happens when one uses an uncertainty interval from a regularized model, or a Bayes interval, for the basis of the SGPV and how that affects the statistical properties of the SGPV.

The use of an interval null hypothesis is not new in statistics. It is featured in equivalence testing (Schuirmann, 1987), non-inferiority testing (Wang and Blume, 2011), and the Bayesian Region of Practical Equivalence procedure [ROPE; Kruschke, 2014, chapter 12; Kruschke and Liddell, 2017]. Despite 30+ years of existence, equivalence tests have not garnered a large following in the statistical community. Factors contributing to this are the equivalence test's general behavior and non-optimality (Perlman and Wu, 1999) and a well-respected paper calling for the abandonment of a popular variant of equivalence tests—the two one-sided tests (Berger and Hsu, 1996). Of course, equivalence and non-inferiority tests are classical hypothesis tests. As a result, they inherit the shortcomings of the general approach. Flipping the null and alternative hypotheses does not alleviate the ills of hypothesis testing. For example, a p-value cannot measure the evidence for a null hypothesis; flipping the null and alternative hypotheses does not solve this problem, as support for the new null (the old alternative) can no longer be assessed. On the other hand, the SGPV is something different; it is not rooted in classical testing. The similarity between equivalence testing and SGPVs begins and ends in the mathematical formalization of the hypotheses. To the point, the SGPV easily indicates when the data support the null or alternative hypotheses, or when the data are inconclusive; there is no need to flip the hypotheses.

BACKGROUND: REGULARIZED MODELS

Regularized models are commonly used in quantitative research. These models can be generated using a wide variety of methods. Some common examples are LASSO (Tibshirani, 1996), elastic-net (Zou and Hastie, 2005), support vector machines (SVM) (Cortes and Vapnik, 1995), Bayesian regression models (Gelman et al., 2013), and even simple continuity corrections of 2 × 2 tables. Because regularized models are now ubiquitous, it is important to know how the SGPV performs when calculated with an uncertainty interval generated from a regularized model.

Broadly speaking, regularization is the practice of incorporating additional structure to a model beyond the typical likelihood or loss function. The additional structure is often incorporated into the model via (a) constraints on the model parameters (LASSO, elastic-net), (b) direct addition of model complexity terms to the loss function (SVM), (c) prior distributions of the model parameters or Bayesian models, or (d) augmented data. Operationally, the contribution of the additional structure—the regularization—relative to the typical likelihood or loss function is controlled by tuning parameters e.g., the severity of the constraint, the scale of the complexity penalty, the variation in the prior distributions, or the amount of augmented data. These tuning parameters are commonly set by cross-validation, although this is not the only approach. Such regularization helps to combat over-optimistic parameter estimation in models that have sparse information relative to the (number of) parameters of interest.

Consider, for example, the classical Bayesian model. The impact of the prior distribution can be minimal if the variation of the prior distribution is large enough that the prior distribution looks essentially flat relative to the likelihood function. When this happens the (flat) prior adds no additional structure to the model. In these cases, the posterior distribution looks very similar to the likelihood function. Conversely, the prior's impact can be substantial if the variation in the prior distribution is small and discordant with the likelihood. Such a prior adds considerable structure to the model; the resulting posterior is a weighted average of the likelihood and that prior.

To illustrate, consider the comparison of two group means using a Bayesian regression model. A detailed description of each regularized model is beyond the scope of this paper, but a simple summary is that the prior and likelihood are combined to yield uncertainty intervals from the posterior (regularized credible intervals). In this example, let $\beta = \mu_1 - \mu_0$ denote the difference in means between the two groups. In **Figure 2**, data collected from two groups is displayed as overlapping histograms, and the observed sample means are shown as $X_1$and X_0. In the bottom of **Figure 2**, the impact of the prior on the posterior is evident. Note that the 95% credible interval and posterior point estimate of β (displayed as a blue line and point overlaid on the posterior) are pulled toward zero. The data are not changing in this example; the different credible intervals are the result of changing the prior distribution.

The phenomenon evident in **Figure 2**, where posterior point estimates are pulled toward to the mean of the prior (usually 0), is called shrinkage. Shrinkage is natural a consequence of adding structure or information to the model. Notice also that the interval widths become narrower as the degree of regularization becomes larger. The shrunken point estimates are statistically biased but the standard errors of the estimates are smaller. The bias typically vanishes as the sample size grows if the added structure does not change as data accumulate (e.g., the prior is prespecified and remains fixed). Shrinkage often reduces the mean squared error (MSE, i.e., $bias^2 + variance$), which is why regularized methods are typically used on prediction models. The trade-off of bias and variance is an important one; smaller MSE is often a desirable operating characteristic.

However, there is no guarantee that regularization will generate smaller MSE. **Figure 3** shows the impact on MSE as outcome variation increases under various degrees of regularization. (The operational definition of the degree of regularization is described in section Methods: Simulation Setup.) For a given level of regularization, MSE is improved if the standard deviation of the outcome somewhat exceeds β (in standard deviation units) as depicted in **Figure 3**. However, regularization tends to increase MSE when the standard deviation of the outcome was comparable to, or less than, the effect size. This phenomenon becomes exaggerated as the degree of regularization increases.

The take home message from **Figure 3** is that regularization works well when the magnitude of the noise is substantially larger than the signal strength. But when the signal is larger than the noise, regularization can be counterproductive and increase the mean squared error (reduce predictive ability). It should also be said that some models cannot be estimated uniquely without regularization. That is, often the data do not provide enough information to identify a model by themselves. In such cases, adding structure to the model allows the enhanced model to be fit with the data. For example, when the number of predictor variables exceeds the number of observations, regularization can add sufficient structure to permit unique model estimation. LASSO regression and ridge regression are often used in these settings.

Because the SGPV is predicated upon the concept of interval estimates and interval nulls, the SGPV can be immediately applied to parameter estimates and uncertainty intervals constructed from regularized models. In the sections that follow, we examine how the SGPV performs when applied to a Bayesian regression model that estimates the difference in means between two groups. The Bayesian setting is quite flexible and generalizable, as virtually all popular regularization techniques can be re-written as a Bayesian model (albeit sometimes with empirical or specialized prior). Lasso, Ridge Regression, and the James-Stein estimator are some prominent examples. Other penalized likelihood formulations can be framed in a Bayesian context, although the corresponding prior may not be proper, smooth, or as well-behaved as the Lasso, Ridge, and JS estimation.

AN INTRODUCTORY EXAMPLE: LOGGERHEAD SHRIKE AND HORNED LIZARD

Data presented in Young (2004) and made public as part of the textbook Analysis of Biological Data (Whitlock and Schluter, 2015) compared the horn length of 30 dead and 154 alive lizards, *Phrynosoma mcalli.* Researchers hypothesized that larger horn length might be protective against the attack of the loggerhead shrike, *Lanius ludovicianus*. Here we present a reanalysis of the data in the context of regularization and SGPVs.

The model for the difference in mean horn length can be parameterized with β in the following linear model. In this model, $I(Alive)$ is an indicator variable which is equal to 1 if the lizard was alive at the time of measurement and 0 otherwise.

$$E\left[Horn\ Length \middle| Alive\right] = \alpha + \beta\ I\left(Alive\right)$$
$$V\left[Y|G\right] = \sigma^2$$

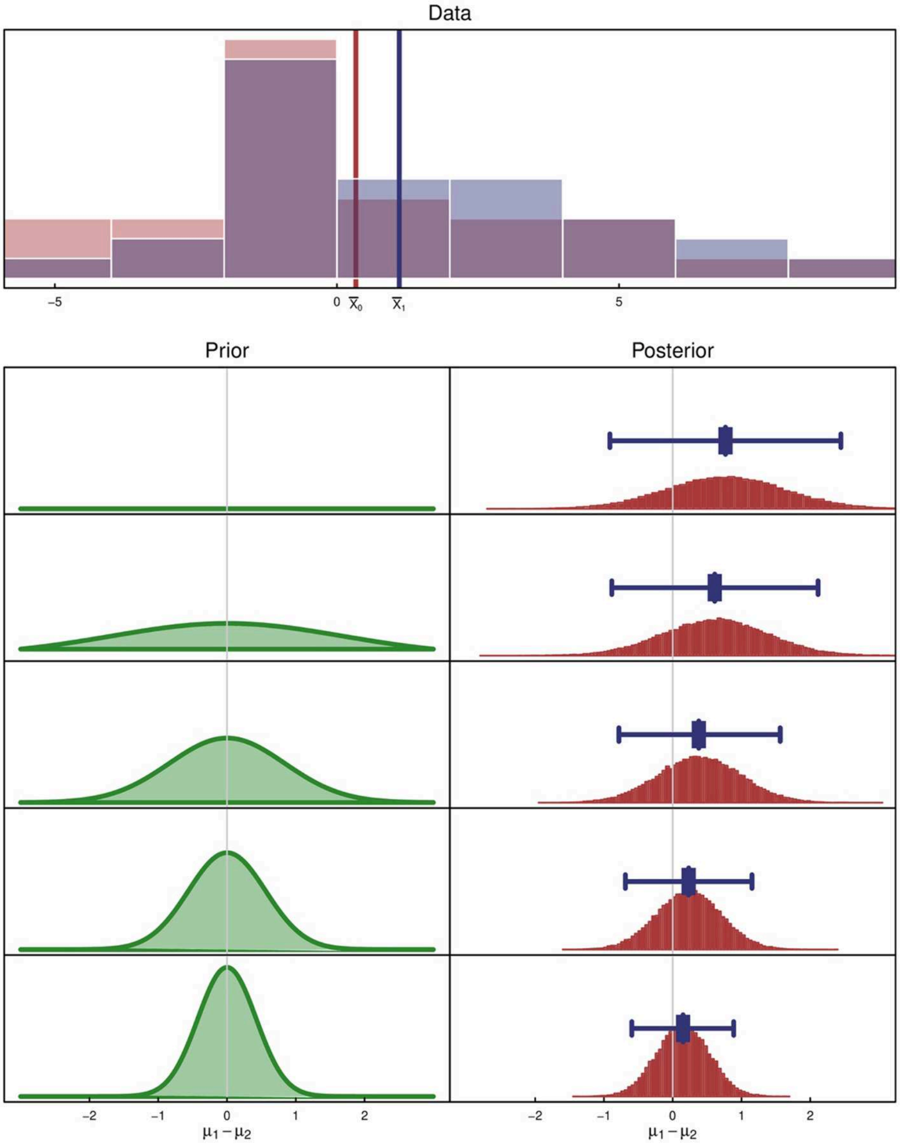


FIGURE 2 | An illustration of shrinkage when additional structure is incorporated into a model by regularization. In this illustration, regularization is achieved with a Bayesian prior, which ranges from flat to peaked. The top figure is the hypothetical data collected from two groups. N (per group) is 30. The observed sample means are shown as X_1 and X_0. The bottom panel shows the posterior distributions (right column) resulting from the choice of prior for the difference in means (left column). As variation in the prior decreases, the resulting interval estimate and point estimate (shown as a blue line and point overlaid on the posterior) are shrunk toward 0.

We set the Bayesian priors as follows:

$$\begin{aligned}\beta &\sim N(0,\ 4.25)\\ \alpha &\sim \textit{ Improper Flat Prior}\end{aligned}$$

Prior to the analysis, we set the null region from −0.5 to 0.5 mm, indicating that a mean difference <0.5 mm is scientifically equivalent to no difference. The null region should be based on researcher expertise. It is not a quantity driven by data; rather it is driven by scientific understanding of the subject matter. In this example, it is quite possible that different researchers will arrive at different null regions. The variance of the prior for beta was selected to be wide enough to be non-informative, but not so wide to allow implausible values of the treatment effect. There are different approaches to selecting a prior in a Bayesian analysis (Gelman et al., 2013). The approach used will impact the degree of regularization, and it is not a primary concern in this investigation because our focus is on the SGPV's behavior after regularization. However, in our experience, using an empirically derived prior, as we done here, often provides sensible shrinkage behavior.

In **Figure 4**, we show the prior, the likelihood for the observed data, and the resulting posterior and 95% credible interval for three different version of this analysis. Credible intervals were

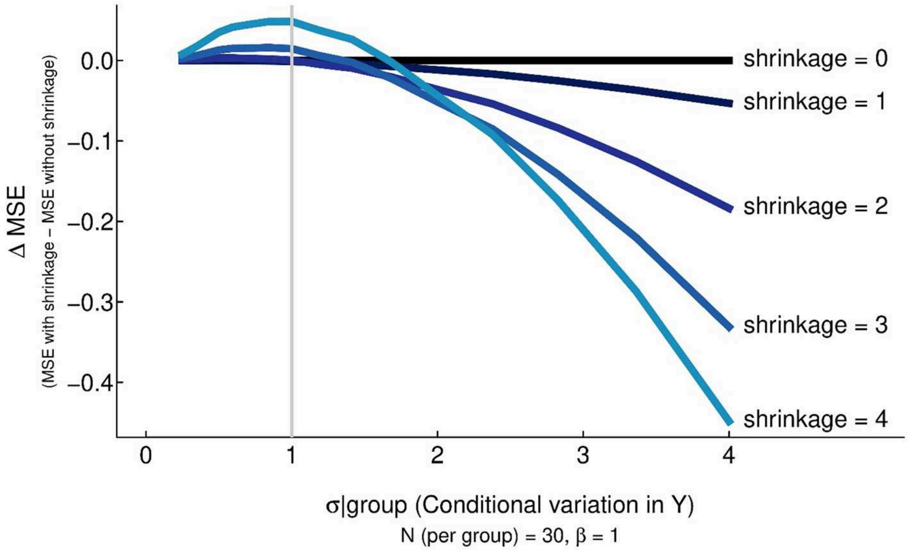


FIGURE 3 | An illustration of the relationship between the conditional variance, the degree of regularization/shrinkage, and the change in MSE when estimating the difference in means. The gray vertical line represents the effect size of the difference in means (1 SD). Shrinkage improves MSE when the conditional variance is large relative to the effect size, but it may increase MSE when the conditional variance is relatively comparable or smaller to the effect size.

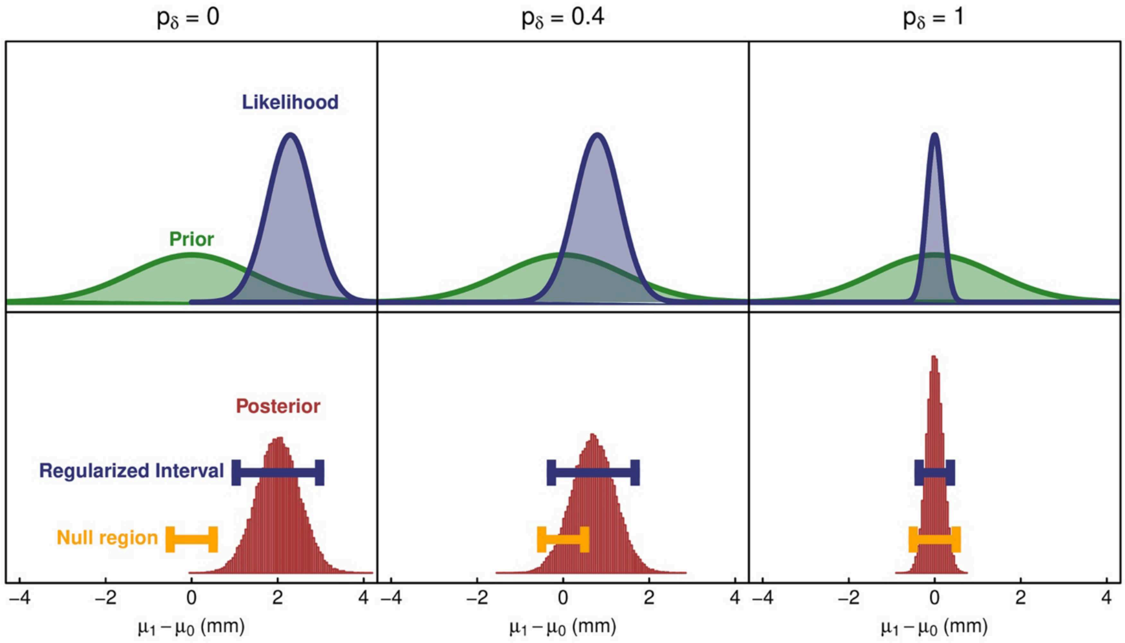


FIGURE 4 | A demonstration of three possible study results in the context of the horned lizard data. The left column shows an interval estimate that does not overlap with the null region, resulting in a second-generation p-value of 0. The middle column shows an interval estimate that straddles the null region, resulting in a second-generation p-value of 0.4. The right column shows an interval estimate that falls entirely within the null region, so the second-generation p-value is 1. The left column results from the unaltered data; whereas data for the middle and right column have been altered for demonstration purposes.

generated from 50,000 draws from the posterior distribution and an empirical calculation of the 95% highest posterior density. The null region is also shown. In the first panel (column), the null region and the credible interval do not overlap, so the SGPV is 0. A larger null region from −1 to 1 mm is needed to have any chance of overlapping.

To demonstrate how the analysis might proceed for different effect sizes, we artificially shifted the outcome values for the living lizards by 1.5 mm so that differences in mean horn length are much smaller than the original data. After shifting the data, we see a different result, which is shown in the middle panel (column). The regularized interval straddles the boundary of the null region, so the analysis of this data generates an inconclusive result. The data support both null and meaningful effect sizes, and the second-generation *p*-value is 0.4.

We also artificially altered the dataset to demonstrate an analysis that results in a conclusive similarity between groups (last panel/column, **Figure 5**). First, we shifted the horn length for living lizards to match the mean horn length of dead lizards. Second, we increased the sample size of the dataset by a factor of 8 by resampling rows. As one would expect, the likelihood and posterior terms are tighter because of the increased sample size. The resulting interval calculated from the posterior is shorter and falls completely within the null region. The second-generation *p*-value is 1 in this case, which is an indication of a conclusive similarity.

We now understand how to compute and use SGPVs. The question that remains is whether the SGPV is reliable in a repeated sampling sense. In the following sections, we simulate similar types of data and perform similar analyses under a wide variety of settings in order to understand the operating characteristics of the SGPV with (smooth) regularization.

METHODS: SIMULATION SETUP

In order to better understand the properties of SGPV, we generated Gaussian outcome data for two groups of size N. The difference in means between the groups was β, and the conditional standard deviation was σ. Depending on the simulation, we varied N, β, and σ. In mathematical notation, the data generation procedure was:

$$\begin{aligned} For\ i &= 1, \ldots, N, \ldots, 2N \\ Let\ G_i &= \begin{cases} 0, & i \leq N \\ 1, & i > N \end{cases} \\ Draw\ \epsilon_i &\sim N(0,\ \sigma) \\ Calculate\ Y_i &= \beta G_i + \epsilon. \end{aligned}$$

G_i is the group indicator and the linear regression model for estimating β was

$$\begin{aligned} E[Y|G] &= \alpha + \beta G \\ V[Y|G] &= \sigma^2. \end{aligned}$$

When fitting a Bayesian regression model, the prior for the two mean parameters (α, β) was

$$\begin{aligned} \beta &\sim N\left(0,\ \frac{3\hat{\sigma}}{1.96} \times \frac{1}{shrinkage}\right) \\ \alpha &\sim Improper\ Flat\ Prior. \end{aligned}$$

where *shrinkage* is a variable set for each simulation. Setting shrinkage to 0 is equivalent to ordinary least squares. The prior for β is calibrated so that 95% of its probability mass is within $\pm\frac{3\hat{\sigma}}{shrinkage}$ (Gelman et al., 2008). The value $\hat{\sigma}$ is the unconditional standard deviation of the outcome. In a typical analysis setting, the prior for the treatment effect coefficient would be driven by expert opinion. In the simulation setting, we resort to an empirically driven prior. The resulting prior without shrinkage is

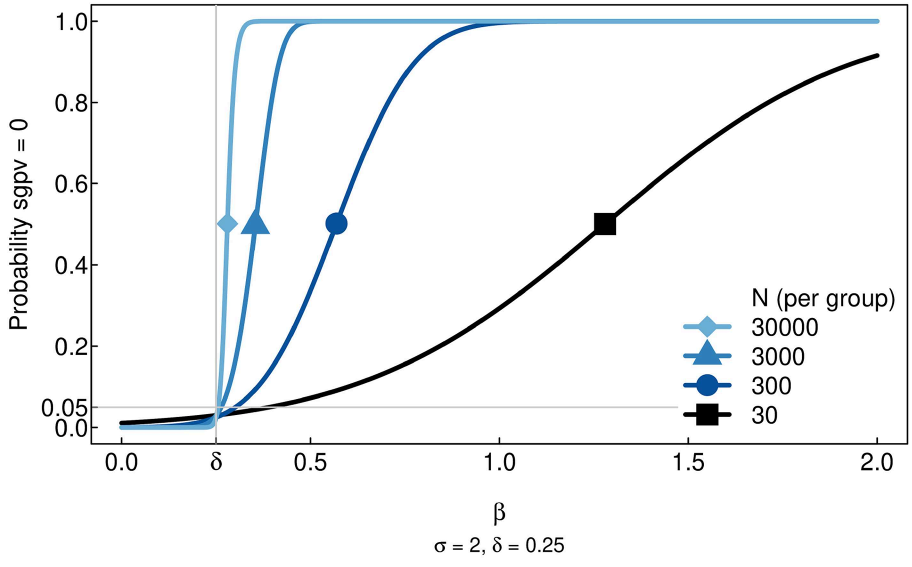


FIGURE 5 | Simulation results showing the Type I and Type II Errors rates for the SGPV as the sample size (N) increases. Within the null region (i.e., all values of beta less than delta), the probability that the SGPV = 0 gets increasingly small as N gets larger. In contrast, for beta values beyond the null region, the probability that the SGPV = 0 goes to 1 as N gets larger. At delta, the boundary of the null region, the probability that SGPV = 0 is controlled at α = 0.05 or less. Hence, the Type I Error rate goes to 0 as N increases. For non-zero values of beta within the null region, the Type II Error rate goes to 1.

non-informative without admitting implausible values (Gelman et al., 2008). As *shrinkage* increases, the probability mass becomes more concentrated around 0. We varied the shrinkage parameter from 0 to 9 in our simulations. For computing the SGPV, we used a null interval of $[-0.25, 0.25]$ or equivalently $\delta = 0.25$ (we used a relatively narrower null interval than in the example to account for possibly strong shrinkage in the simulations).

For each combination of simulation parameters, 5,000 replicate datasets were generated and analyzed. Credible intervals in each analysis were generated from 1,000 draws from the posterior distribution and an empirical calculation of the 95% highest posterior density. Power was calculated as the proportion of replicates where the SGPV equaled zero. If the interval null had been specified as a point, then this procedure would be equivalent to a standard two-sided t-test. MSE was calculated as the mean squared error of the difference between the known β (set by simulation) and estimated $\hat{\beta}$ (from simulated replicates).

RESULTS

Simulation 1: Power of SGPV as *N* Increases

As a starting point, we consider the traditional case of least squares to demonstrate the default trade-off of Type I and Type II/power rates for the SGPV. Data were generated under a range of effect sizes, β values, with an increasing number of observations in each group. The conditional standard deviation and null interval were held constant as indicated above.

The results are reported in **Figure 5**. The most noticeable feature of the figure is that errors within the null region tend toward 0, especially as N increases. Rather than an error rate of 5% at $\beta = 0$, there is an approximate error rate of 5% at $\beta = \delta$, the boundary of the null region. Consequently, power for values of β outside the null region is less than what would be observed with the traditional t-test. This agrees with the results in Blume et al. (2018). The SGPV's reduction in power is traded for a similar reduction in the Type I Error rate for clinically meaningless effect sizes. The reduction in power here is not substantial, but it might be larger in other cases. This is should be checked when planning studies.

Simulation 2: Power, Interval Null, Shrinkage

At the heart of this simulation study is the question of how SGPVs generated with intervals from regularized models compare to SGPVs generated without regularization. To that end, we simulated power curves for all four possible combinations of interval types and degree of shrinkage. In the top panel of **Figure 6** we see that mild shrinkage has negligible impact on the Type I and Type II error rates. The primary feature in the top panel is that SGPVs with an interval null spend power to reduce the Type I error rate. In the bottom panel of the same figure, the degree of shrinkage is exceedingly large, much larger than one would typically use in an actual analysis. Interestingly, even in this case, there is a real separation of the power curves when comparing regularized and non-regularized approaches. Given the extreme nature of the shrinkage, it is surprising that the differences are not larger.

Simulation 3: MSE, Interval Null, and Shrinkage

This final simulation reinforces that the MSE benefits of regularization are retained when SGPV is used as a summary measure. Because MSE is a function of the estimated and true β–values which are not altered when calculating or interpreting the SGPV—MSE will not change when a null interval is used for inference. In the simulation results below (**Figure 7**), the red line represents the difference between the MSE of a regularized model with an interval null compared to the same regularized model with a point null. As is clear from the plot, this difference in MSE is 0. As a point of reference, the black line shows that in this simulation setting, regularization does in fact lower MSE. This shows that using an interval null also yields the typical benefits seen with standard shrinkage estimators of improved prediction via lower MSE.

DISCUSSION

The SGPV promotes good scientific practice by encouraging researchers to *a priori* establish what are, and what are not, scientifically meaningful effects. By establishing the null interval at the start of the analysis, the SGPV can provide a summary of how consistent the data are with the null hypothesis or how consistent the data are with meaningful effects. Better Type I Error rates are achieved at the expense of power in the region outside the null interval. Because regularized models are now widely used, it is important to understand how the SGPV operates when applied to intervals impacted by shrinkage. Based on our simulations, SPGVs based on credible intervals retain their desirable Type I/Type II error rate tradeoff at a modest cost in power. Likewise, the same gains in MSE observed with Bayesian estimation are observed when a null interval and the SGPV are used. Consequently, SPGVs may be applied to Bayesian analyses (where classical p-values are currently not available) and to regularized models that exhibit some degree of natural and smooth shrinkage.

The simulation study in this manuscript focuses on shrinkage intervals constructed using a prior and a Bayesian posterior distribution. Because regularized likelihood and regularized machine learning methods can often be couched as special cases of Bayesian modeling, the focus on shrinkage-by-prior is a natural place to start. We note, however, that in focusing on the Bayesian approach in the simulations, we are really focusing on the subset of priors that induce shrinkage in the natural way; overly informative priors or priors which lead to pathological shrinkage are outside the scope of our investigation.

Statements regarding Type I or Type II error rates and Bayesian credible intervals may seem odd to some readers because some authors do not consider p-values or null hypothesis testing to fit within the Bayesian paradigm. Likewise, it can seem odd to estimate a posterior distribution in order to calculate a second-generation p-value. However, this highlights

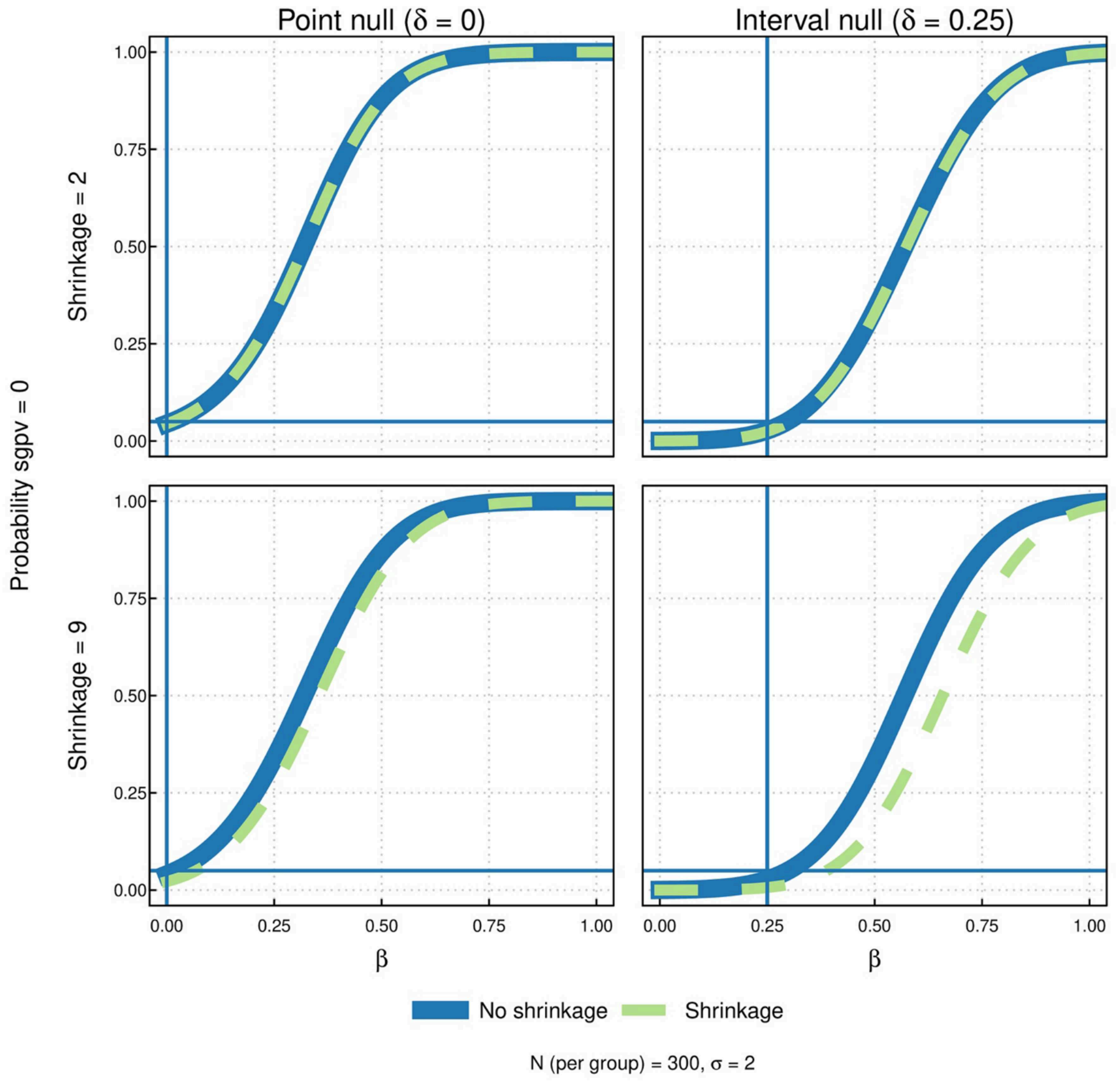


FIGURE 6 | Results of the simulation study with the following factors: degree of shrinkage [mild vs. none (row 1) and extreme vs. none (row 2)] and type of null [point (column 1) and interval (column 2)]. In each of the four graphs, the probability that the SGPV = 0 is plotted as a function of beta for both the shrinkage and no shrinkage intervals. Comparing column 1 to column 2, the estimated probability curves cross 0.05 at the boundary of the respective nulls (0 for the point null and 0.25 for the interval null). For mild shrinkage (row 1), there is little noticeable difference between probability curve estimates with or without shrinkage (solid blue vs. dashed green). More noticeable differences between shrinkage and no shrinkage occur with more extreme shrinkage (row 2).

an important point: the SGPV is not a probability. It is a summary measure—applicable to any inferential framework—for indicating the degree of conclusiveness of the analysis. An SGPV of 0 indicates a conclusive difference, a value near 1 indicates a conclusive similarity, and values between 0 and 1 indicate differing degrees of inconclusive results with a value at 0.5 indicating a maximum degree of inconclusiveness.

One might wonder why this summary measure is called a second-generation *p*-value if it is not a probability. It is our contention that the practical, every-day use of traditional *p*-values is as a marker for results deserving of increased scrutiny. That is, the traditional *p*-value and 0.05 threshold is used to answer the question: "Should I dive deeper into this hypothesis?" As many have noted, the traditional *p*-value is not a good filter for this in practice. The SGPV, in contrast, is designed to filter results that deserve greater attention vs. results that need more data and are currently inconclusive. So, the SGPV is a second generation of the *p*-value as it is used in practice; it is not an extension of the probability calculation for a null hypothesis test.

Evaluating the operating characteristics of the SGPV is routine step that is intended to be paradigm-agnostic. It is common these days to see a statistical approach, regardless of paradigm of origin, evaluated in this long-run sense. The Food and Drug Administration (FDA) which approves drugs

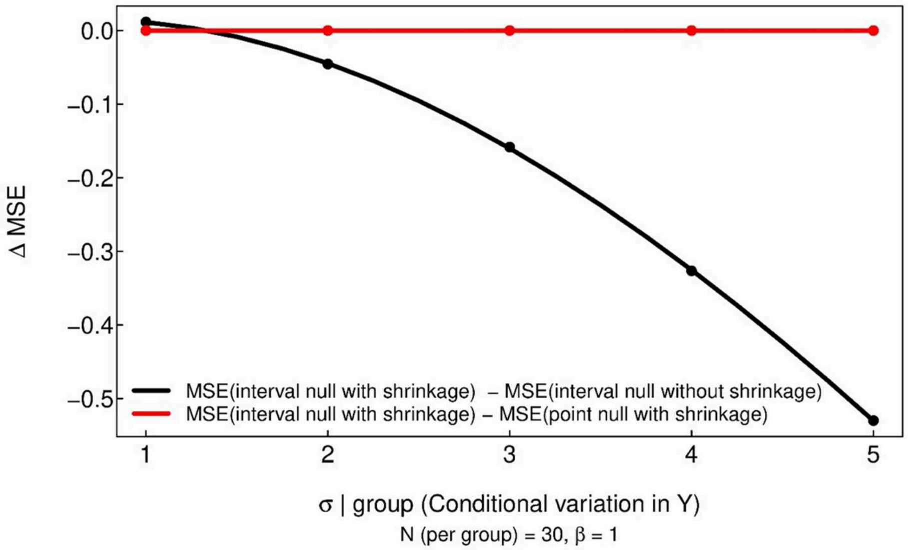


FIGURE 7 | Simulation results showing that MSE is not altered when using a point null or a region null (red line). The black line is a reference to show the change in MSE when incorporating shrinkage.

and medical devices for commercial use in the United States requires evaluation of operating characteristics as part of drug and device applications even when the submitted data analysis is a set of posterior probabilities from a Bayesian analysis (or set of likelihood ratios or p-values). In "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials[1]" (Young, 2004), the FDA recommends and provides guidance for computing Type I and Type II Error rates regardless of the analysis paradigm. Note that even when prominent Bayesian statisticians propose a new Bayesian clinical trial design, as with the "Bayesian decision-theoretic group sequential clinical trial design" (Lewis et al., 2007) or with "Phase II oncology clinical trials" (Berry et al., 2013), the analysis is calibrated so that the Type I Error rate is controlled. The SGPV is a tool for deciding when an analysis shows a conclusive difference, a conclusive similarity, or is not conclusive. As such, it is appropriate to explore the operating characteristics of this tool even if the interval is calculated from a Bayesian posterior or from a statistical learning method with regularization.

The calculation of the SGPV is simple and intuitive. Specifying the null region, however, is challenging. Ideally, the null region should reflect subject matter expertise of effect sizes that are not meaningful. Agreement among subject-matter experts on what the null region should be is potentially hard to achieve. Further, some research areas may be so new, there is no prior information to guide the decision. Some users may want to punt on deciding what the null region should be and may seek a "data-driven null region." Unfortunately, there is no such thing, at least not with data from the same dataset that one intends to analyze. The challenge of specifying a null region is, in our opinion, the biggest obstacle and limitation of the SGPV. However, it is the step that anchors the statistical analysis to the scientific context; it is the step that pushes that research team to decide what it means to be similar and what it means to be different for their particular research question, all prior to the analysis. These questions are exactly where discussion should focus; and they are precisely the questions that subject matter experts are best equipped to debate. Specifying the null region is a challenging task, but it is a scientific one worth the effort it requires.

There is still a lot to learn about the SGPV and a number of potentially fruitful areas of investigation or expansion. One outstanding question, for example, is what impact cross-validation of shrinkage parameters may have on the operating characteristics of the SGPV when used with intervals constructed with machine learning methods. Dezeure et al. (2015) show that this impact can be real. Another possible extension of particular interest to those analysts that use Bayesian methods is to expand the meaning of the null region. The null region as currently used with the SGPV treats all values in the region as equally unimportant. One may want to incorporate the idea that some values in the null region are more null than others. One approach would be to represent the relative "nullness" of the values by borrowing structure from mathematical distributions, similar in spirit to the likelihood. For example, the simple null region of the current SGPV can be described as a uniform null region in reference to the uniform distribution. A null region in which the relative "nullness" is maximized at zero but then fades to the interval endpoints might be represented with a beta distribution. This is an intriguing next step of research.

The SGPV is intended to be a method-agnostic indicator of when a prespecified evidential benchmark is achieved. Assessing

[1] *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials.* Available online at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials#7

the overlap of the null and uncertainty interval is easily mapped back to classical measures of statistical evidence like the likelihood ratio. For example, a SGPV that is based on a *1/k* likelihood support interval is set to zero whenever the likelihood ratio for the MLE vs. the nearest hypothesis in the null interval is $>k$ [most 95% CIs can be mapped to a 1/6.83 SI, see (Blume, 2002)]. A similar condition can be formulated for Bayes factors when SGPVs are based on credible intervals. In this sense, the SGPV just indicates when the observed evidence is sufficiently strong against the hypotheses in the interval null hypothesis.

CONCLUSIONS

The second-generation *p*-value is an intuitive summary of analysis results that is based on an uncertainty interval about the parameter of interest and a pre-specified null region. Previous publications on SGPVs focused on 95% confidence intervals and 1/8 likelihood support intervals. In the current manuscript, we explored the performance of SGPVs based on uncertainty intervals from a regularized model, specifically Bayesian credible intervals. While we considered intervals generated with Bayes regression, this framework is readily generalizable to many different types of regularization schemes. We saw nearly the same trade-off of Type I and Type II Error rates in SGPVs based on Bayesian credible intervals as SGPVs based on classical confidence intervals. Our results indicate that SPGVs based on regularized intervals retain this desirable error rate trade-off, at a slight loss in power, while benefiting from the bias-variance tradeoff imparted by regularization. Consequently, the SPGV is a meaningful summary of study results, even when applied in a Bayesian framework or other contexts that incorporate regularization.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://whitlockschluter.zoology.ubc.ca/wp-content/data/chapter12/chap12e3HornedLizards.csv.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Berger, R. L., and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.* 11, 283–319. doi: 10.1214/ss/1032280304

Berry, S. M., Broglio, K. R., Groshen, S, and Berry, DA. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. *Clin. Trials* 10, 720–734. doi: 10.1177/1740774513497539

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Stat. Med.* 21, 2563–2599. doi: 10.1002/sim.1216

Blume, J. D., D'Agostino McGowan, L, Dupont, W. D., and Greevy, R. A. Jr. (2018). Second-generation *p*-values: improved rigor, reproducibility, & transparency in statistical analyses. *PLoS ONE* 13:e0188299. doi: 10.1371/journal.pone.0188299

Blume, J. D., Greevy, R. A. Jr., Welty, V. F., Smith, J. R., and Dupont, W. D. (2019). An introduction to second-generation *p*-values. *Am. Stat.* 73, 157–167. doi: 10.1080/00031305.2018.1537893

Cortes, C., and Vapnik, V. N. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software Hdi. *Stat. Sci.* 30, 533–558. doi: 10.1214/15-STS527

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, 3rd Edn*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* 2, 1360–1383. doi: 10.1214/08-AOAS191

Kruschke, J. (2014). *Doing Bayesian Data Analysis, Second Edition: A Tutorial With R, JAGS, and Stan, 2nd Edn*. Boston, MA: Academic Press.

Kruschke, J., and Liddell, T. M. (2017). The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* 25, 178–206. doi: 10.3758/s13423-016-1221-4

Lewis, R. J., Lipsky, A. M., and Berry, D. A. (2007). Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: a frequentist evaluation. *Clin. Trials* 4, 5–14. doi: 10.1177/1740774506075764

Perlman, M. D., and Wu, L. (1999). The Emperor's new tests. *Stat. Sci.* 14, 355–369. doi: 10.1214/ss/1009212517

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharm. Biopharm.* 15, 657–680. doi: 10.1007/BF01068419

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Wang, S. J., and Blume, J. D. (2011). An evidential approach to noninferiority clinical trials. *Pharm. Stat.* 10, 440–447. doi: 10.1002/pst.513

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's Statement on *p*-values: context, process, and purpose. *Am. Stat.* 70, 129–133. doi: 10.1080/00031305.2016.1154108

Whitlock, M. C., and Schluter, D. (2015). *The Analysis of Biological Data*. New York, NY: W.H. Freeman and Company.

Young, K. V. (2004). How the horned lizard got its horns. *Science* 304, 65–65. doi: 10.1126/science.1094790

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 21 January 2020
doi: 10.3389/fevo.2019.00501

Consequences of Lack of Parameterization Invariance of Non-informative Bayesian Analysis for Wildlife Management: Survival of San Joaquin Kit Fox and Declines in Amphibian Populations

*Subhash R. Lele**

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

OPEN ACCESS

Edited by:
Yukihiko Toquenaga,
University of Tsukuba, Japan

Reviewed by:
Ichiro Ken Shimatani,
Institute of Statistical Mathematics (ISM), Japan
Brian Gray,
Upper Midwest Environmental Sciences Center (USGS), United States

***Correspondence:**
Subhash R. Lele
slele@ualberta.ca

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *12 March 2019*
Accepted: *05 December 2019*
Published: *21 January 2020*

Citation:
Lele SR (2020) Consequences of Lack of Parameterization Invariance of Non-informative Bayesian Analysis for Wildlife Management: Survival of San Joaquin Kit Fox and Declines in Amphibian Populations. Front. Ecol. Evol. 7:501. doi: 10.3389/fevo.2019.00501

Computational convenience has led to widespread use of Bayesian inference with vague or flat priors to analyze statistical models in ecology. Vague priors are claimed to be objective and to "let the data speak." However, statisticians have long disputed these claims and have criticized the use of vague priors from philosophical to computational to pragmatic reasons. One of the major criticisms is that the inferences based on non-informative priors are generally dependent on the parameterization of the models. Ecologists, unfortunately, often dismiss such criticisms as having no practical implications. One argument is that for large sample sizes, the priors do not matter. The problem with this argument is that, in practice, one does not know whether or not the *observed* sample size is sufficiently large for the effect of the prior to vanish. It intricately depends on the complexity of the model and the strength of the prior. We study the consequences of parameterization dependence of the non-informative Bayesian analysis in the context of population viability analysis and occupancy models and at the commonly obtained sample sizes. We show that they can have significant impact on the analysis, in particular on prediction, and can lead to strikingly different managerial decisions. In general terms, the consequences are: (1) All subjective Bayesian inferences can be masqueraded as objective (flat prior) Bayesian inferences, (2) Induced priors on functions of parameters are not flat, thus leading to cryptic biases in scientific inferences, (3) Unrealistic independent priors for multiparameter models lead to unrealistic priors on induced parameters, (4) Bayesian prediction intervals may not have correct coverage, thus leading to errors in decision making, (5) Reparameterization to facilitate MCMC convergence may influence scientific inference. Given the wide spread applicability of the hierarchical models and uncritical use of non-informative Bayesian analysis in ecology, researchers should be cautious about using vague priors as a default choice in practical situations.

Keywords: Bayesian analysis, flat priors, non-informative priors, occupancy models, parameterization invariance, population viability analysis, population prediction intervals, prior sensitivity

1. INTRODUCTION

Hierarchical models, also known as state-space models, mixed effects models or mixture models, have proved to be extremely useful for modeling and analyzing ecological data (e.g., Bolker, 2008; Kery and Schaub, 2011). Although these models can be analyzed using the likelihood methods (Lele et al., 2007, 2010), the Bayesian approach is the most advocated approach for such models. Many researchers even name hierarchical models as "Bayesian models" (Parent and Rivot, 2013). Of course, there are no Bayesian models or frequentist models. There are only statistical models that we fit to the data using either the Bayesian approach or the frequentist approach. The subjectivity of the Bayesian approach is bothersome to many scientists (Efron, 1986; Dennis, 1996) and hence the trend is to use the non-informative, also called vague or objective, priors instead of the subjective priors provided by an expert. These non-informative priors purportedly "let the data speak" and do not bias the conclusions with the subjectivity inherent in the subjective priors. It has been claimed that Bayesian inferences based on non-informative priors are similar to likelihood inference (e.g., Clark, 2005, p. 3, 5) although such a result has never been rigorously established.

The fact is that it is not even clear what a non-informative prior really means. There are many different ways to construct non-informative priors (Press, 2003, Chapter 5). The most commonly used non-informative priors are either the uniform priors or the priors with very large variances spreading the probability mass almost uniformly over the entire parameter space. These priors have been criticized on computational grounds (e.g., Natarajan and McCulloch, 1998) because they can inadvertently lead to improper posterior distributions. Link (2013) shows similar problems with using uniform prior on the population size when fitting capture-recapture models. However, how does one explain that a uniform prior on probability of success in a Binomial experiment represents non-information but a uniform prior on the population size does not? Gelman (2006) discusses similar computational problems associated with the non-informative priors for variance components and concludes uniform prior is, in fact, a good choice and not the commonly used inverse Gamma prior (e.g., King et al., 2009). The issue of choice of default priors and its impact on statistical inference has also been observed in genomics (Rannala et al., 2012).

More fundamentally, one of the founders of modern statistics, R.A. Fisher, objected to the use of flat priors because of their lack of invariance under transformation (De Valpine, 2009; Lele and Dennis, 2009). In Fisher's words (Fisher, 1930, p. 528), use of flat priors to represent ignorance is "fundamentally false and devoid of foundation." An excellent review of the problems with various kinds of non-informative or objective priors is available in Ronneberg (2017). For example, a uniform prior on (0,1) for the probability of success in a Binomial model turns into a non-uniform prior on the logit scale (see **Figure 1**). If a uniform prior is supposed to express complete ignorance about different parameter values, then this says that if one is ignorant about p, one is quite informative about $log(p/(1-p))$. Similarly a normal prior with large variance on the logit scale, that presumably represents complete ignorance, transforms into a non-uniform, informative prior on the probability scale (see **Figure 1**).

This makes no sense because they are one-one transformations of each other; if we are ignorant about one, we should be equally ignorant about the other.

Fisher's criticism was potent enough that it needed addressing. Harold Jeffreys tried to construct priors that yield parameterization invariant conclusions. They are now known as Jeffreys priors. A full description of these priors and how to construct them is beyond the scope of this paper (See Press, 2003 or Ronneberg, 2017 for easily accessible details). However, it suffices to say that they are proportional to the determinant of the inverse of the expected Fisher information matrix.

Despite its theoretical properties, there are practical issues that hinder the use of Jeffrey's priors. For example, in order to construct them, one needs to know the likelihood function and the exact analytic expression for the expected Fisher information matrix. Because it is nearly impossible to write the likelihood function explicitly for hierarchical models, computing the expected Fisher information matrix is seldom possible for hierarchical models. This makes the specification of Jeffrey's prior for a given hierarchical model difficult.

Let us look at Jeffreys prior for a simple, non-hierarchical model where $Y \sim Bernoulli(p)$. The Jeffreys prior for the probability of success p is the $Beta(0.5, 05)$ distribution. The density function of this random variable is U-shaped that is highly concentrated near 0 and 1 with very small weight in the middle (see **Figure 2**). It looks similar to the distribution in the lower right hand panel. Even when Jeffreys prior can be computed, it will be difficult to sell it as an objective prior to the jurors or the senators on the committee.

Construction of Jeffreys and other objective priors for multi-parameter models poses substantial mathematical difficulties (Ronneberg, 2017). A commonly proposed solution is to put independent Jeffreys or other non-informative prior on each of the parameter separately. Why such prior knowledge of independence of the parameters be considered "non-informative" is unclear. Assuming two quantities are independent of each other is considered to be a very strong assumption in practice. The assumption of a priori independence between parameters is more a matter of convenience than a matter of principle and is not justifiable.

Press (2003, Chapter 5) provides an excellent review of various problems associated with the definitions and use of non-informative priors along with interesting historical notes. It is clear that non-informative priors are chosen more for their mathematical or computational convenience than for their representation of no information or because they "let the data speak." Unfortunately, ecologists and practitioners tend to dismiss these criticisms; considering them to be of no practical relevance (e.g., Clark, 2005).

There are two prevalent notions, both false, about the non-informative Bayesian analysis. The first false notion is that there is no difference between non-informative Bayesian inference and likelihood-based inference and the second false notion is that the philosophical underpinnings of statistical inference are irrelevant to practice. Some researchers (e.g., Kery and Royle, 2016) even

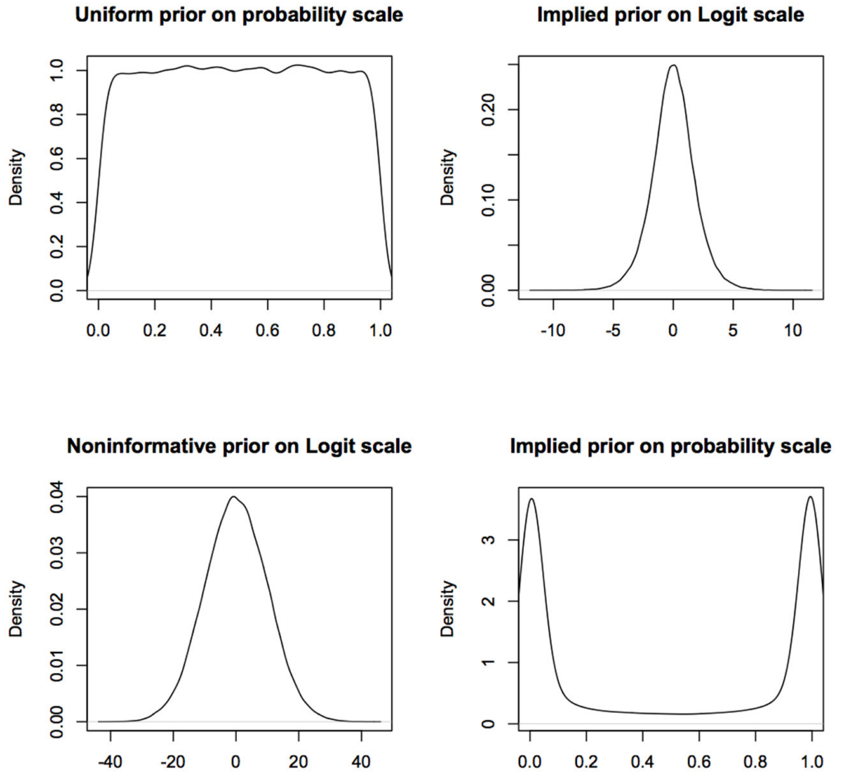


FIGURE 1 | Non-informative prior on one scale is informative on a different scale. What is considered non-informative on the logit scale will be considered quite informative on the probability scale and what is considered non-informative on the probability scale will be considered informative on the logit scale. For computational convenience, the figures are density plots of the random numbers generated from the corresponding distributions, instead of using the analytic expressions.

claim that Bayesian inference is "valid" for *all* sample sizes, but, unfortunately, without specifying the "validity" criterion. Of course, as the information in the sample increases, effects of the prior and consequences of lack of parameterization invariance become negligible. Although, caveat of large sample size is mathematically correct, whether or not the observed sample size is large depends on the complexity of the model and the strength of the prior (e.g., Dennis, 2004) and cannot be judged in practice. To illustrate the falsity of these notions for sample sizes observed in practice, we consider two important ecological problems: Population monitoring and population viability analysis. We show that, due to lack of invariance, analysis of the same data under the same statistical model can lead to substantially different conclusions under a non-informative Bayesian framework. This is disturbing because common sense dictates that same data analyzed using the same model should lead to the same scientific conclusions. The problem with the non-informative priors is that they do not "let the data speak"; contrary to what is commonly claimed, (absent large sample size) they bring in their own biases in the analysis. We do not suggest that likelihood analysis, which is parameterization invariant, is the only right way to do the data analysis in applied ecology. That debate is subtle and potentially unresolvable. Only goal of this paper is to show that implications of the lack of invariance of non-informative priors are of practical significance to wildlife managers.

2. POPULATION VIABILITY ANALYSIS (PVA) FOR SAN JOAQUIN KIT FOX

Let us consider the San Joaquin kit fox data set originally analyzed by Dennis and Otten (2000). This kit fox population inhabits a study area of size 135 km^2 on the Naval Petroleum Reserves in California (NPRC). The abundance time-series for the years 1983–1995 was obtained to conduct an extensive population dynamics study as part of the NPRC Endangered Species and Cultural Resources Program. The annual abundance estimates were obtained from capture-recapture histories generated by trapping adult and yearling foxes each winter between 1983 and 1995. We refer the reader to Dennis and Otten (2000) for further details on these data and abundance estimation technique.

Dennis and Otten (2000) analyzed these data using the Ricker model. The deterministic version of the Ricker model can be written in two different but mathematically equivalent forms. It may be written in terms of the growth parameter a and density dependence parameter b as $logN_{t+1} - logN_t = a - bN_t$ where $a > 0, b > 0$ or in terms of growth parameter a and carrying capacity parameter K as $logN_{t+1} - logN_t = a\left(1 - \frac{N_t}{K}\right)$ where $a > 0, K > 0$. We also know that $K = a/b$ and $b = a/K$. It is reasonable to expect that the conclusions about the survival of the San Joaquin kit fox population would remain

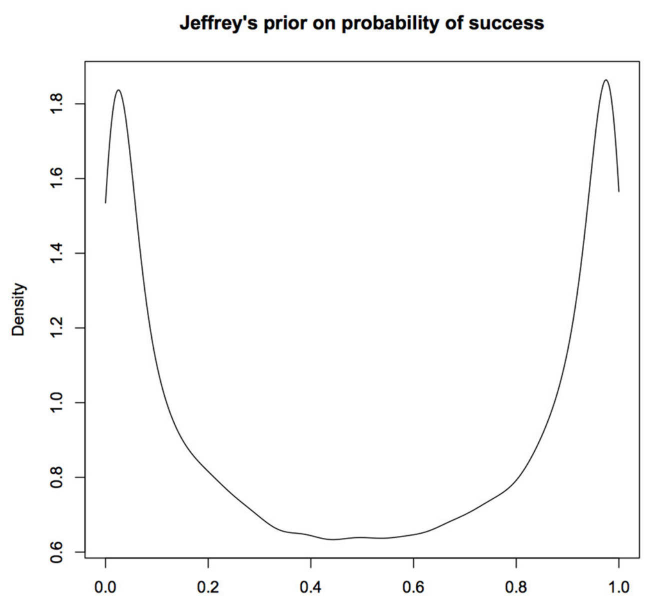


FIGURE 2 | Jeffreys prior for probability of success in a Binomial experiment; This prior concentrates the probability mass near 0 and 1. It is difficult to justify this as a prior that would be considered non-informative. Multivariate extensions of Jeffreys priors can lead to inconsistent estimators and hence seldom used in practice. For computational convenience, the figures are density plots of the random numbers generated from the Uniform(0.5,0.5) distribution, instead of using the analytic expression.

the same whether one uses the (a, b) formulation or the (a, K) formulation. In statistical jargon, we call this change in the form of the model "reparameterization" and we will use this term, instead of the term "different formulation," in the rest of the paper. Following Dennis and Otten (2000), we use a stochastic version of the Ricker model where the parameter a, instead of being fixed, varies randomly from year to year. Furthermore, the abundance values are themselves an estimate of the true abundances and hence we consider the sampling variability in the model as well. The variances for the abundance estimates were nearly proportional to the abundance estimates and hence the Poisson sampling distribution makes reasonable sense. The full model can be written as a state-space model as follows. In the following, N_t denotes the true abundance, Y_t denotes the estimated abundance and $X_t = logN_t$.

We call the following form of the model the (a, b) parameterization.

- Process model: $X_{t+1}|X_t \sim Normal(X_t + a - b * exp(X_t), \sigma^2)$
- Observation model: $Y_t|X_t \sim Poisson(exp(X_t))$

One can write this model in an alternative form that we call the (a, K) parameterization.

- Process model: $X_{t+1}|X_t \sim Normal(X_t + a\left(1 - \frac{exp(X_t)}{K}\right), \sigma^2)$ where K is the carrying capacity.
- Observation model: $Y_t|X_t \sim Poisson(exp(X_t))$

These two models are mathematically identical to each other. Our goal is to fit these models to the observed data and conduct population viability analysis using the population prediction intervals (PPIs) (Saether et al., 2000). To compute the one sided PPIs, that is usually of interest to managers, we predict the future values of the time series and connect the lower 10% values for each year to get a curve, as a function of time, indicating the lower envelope to the future population sizes. This lower envelope helps guide the management decisions. Common sense dictates that because the data are the same and the models are mathematically equivalent to each other, the PPIs computed under the two parameterizations should also be identical to each other.

We use Bayesian inference using non-informative priors to compute PPIs under these two forms. For Bayesian inference, we use the following non-informative priors for the parameters in the respective parameterization.

- Priors for the (a, b) parameterization: $a \sim LN(0, 10), b \sim U(0, 1), \sigma^2 \sim LogNormal(0, 10)$
- Priors for the (a, K) parameterization: $a \sim LN(0, 10), K \sim Gamma(100, 100), \sigma^2 \sim LogNormal(0, 10)$

These are some of the commonly used distributions for representing non-information on the appropriate ranges of the parameters (e.g., Kery and Schaub, 2011). Although note that there is no general agreement on what is a non-informative prior distribution. A reader who wants to use different non-informative priors can easily repeat the experiment by modifying the R code (see link in the data availability statement) appropriately. The qualitative conclusions will remain the same. For comparison, we use the data cloning algorithm (Lele et al., 2007, 2010) to compute the maximum likelihood estimators (MLE) based frequentist predictions to obtain PPIs under these two parameterizations. The analysis was conducted using the package "dclone" (Solymos, 2010; R Development Core Team, 2011), that is based on commonly used JAGS software (Plummer, 2003), within the R software. The data and the R program to conduct this analysis are available in the link provided in the data availability statement. Both Bayesian analysis and data cloning based maximum likelihood analysis are based on the Markov Chain Monte Carlo (MCMC) algorithms. Convergence diagnostics were based on the Gelman-Rubin Rhat statistics and the trace plots. For all cases, the Rhat statistics was very close to 1 and the trace plots showed good mixing of the chains (see link in the data availability statement). The resultant parameter estimates are given in the table below. To make the comparison easy to interpret, we report the estimates for (a, K, σ) under the two parameterizations.

Notice that (**Table 1**) the Bayesian parameter estimates for the two parameterizations, especially the estimates of the carrying capacity K, are quite a bit different. On the other hand, the data cloned maximum likelihood estimates (MLE) are nearly identical to each other under both parameterizations, as they should be. The small differences are due to the Monte Carlo error. In **Figure 3**, we show the PPIs obtained under the likelihood and the non-informative Bayesian approaches.

One can make two important observations:

1. The PPIs obtained under the (a, b) parameterization and the PPIs obtained under the (a, K) parameterization, under

TABLE 1 | Parameter estimates for the kit fox data using different parameterizations and non-informative priors and maximum likelihood.

Parameter	Bayes (a, b)	Bayes (a, K)	MLE (a, b)	MLE (a, K)
a	0.7542	0.4812	0.7404	0.7322
K	159.6425	141.39	160.1643	159.7164
σ	0.4916	0.5053	0.4360	0.4358

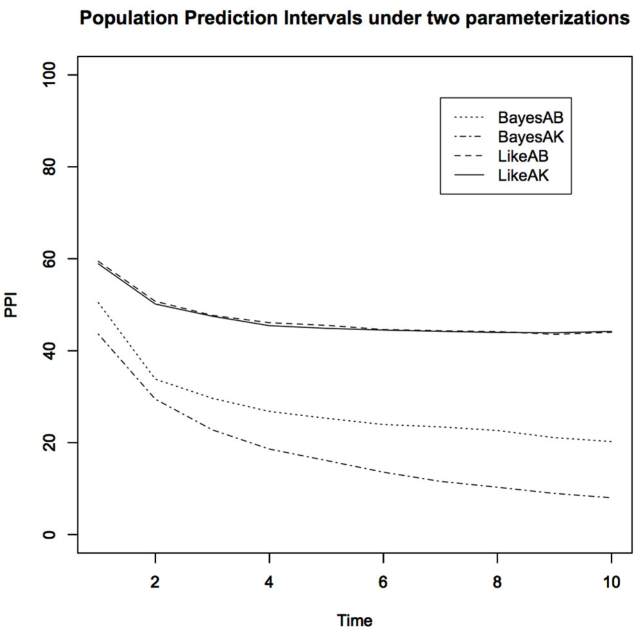


FIGURE 3 | Lower 10% Population Prediction Intervals (PPI) for the kit fox data using non-informative Bayesian analysis under two different parameterizations and the maximum likelihood analysis. Notice that non-informative Bayesian analysis does not approximate the maximum likelihood analysis and depends on the specific parameterization.

purportedly non-informative priors, are quite different. Depending on which parameterization the researcher happens to use, the scientific conclusions could be quite different. For the non-informative Bayesian analysis, instead of the Gamma distribution, we also used a uniform distribution prior for the carrying capacity parameter. The results were not only different from these but also very sensitive to the choice of the upper bound of the uniform distribution. This is at least disturbing, if not totally unacceptable. As we said earlier, analyzing the same data with the same model should lead to the same conclusions. However, non-informative Bayesian analysis does not satisfy this common sense requirement.

2. The MLE based PPIs are quite different than the non-informative prior based PPI. Contrary to what is commonly claimed, the non-informative priors do not lead to inferences that are similar to the likelihood inferences.

3. OCCUPANCY MODELS AND DECLINE OF AMPHIBIANS

One of the central tasks that applied ecologists are entrusted with is monitoring existing populations. These population monitoring data are the inputs to many further ecological analyses. We consider the following simple model that is commonly used in analyzing occupancy data with replicate visits (MacKenzie et al., 2002). We denote probability of occupancy by ψ and probability of detection, given that the site is occupied, by p. For simplicity (and, to emphasize that these results happen even for simple models), we assume that these do not depend on covariates. We assume that there are n sites and each site is visited k times. Other assumptions about close population and independence of the surveys are similar to the ones described in MacKenzie et al. (2002). In the following, Y_i indicates the true state of the ith location, occupied (1) or unoccupied (0). This is a latent, or unobservable, variable. Observations are denoted by O_{ij}. These are either 0 or 1, depending on the observed status of the location at the time of jth visit to the ith location. These can be different from the true state Y_i because of detection error. The replicate visit model can be written as follows.

- Hierarchy 1: $Y_i \sim Bernoulli(\psi)$ for $i = 1, 2, ..., n$
- Hierarchy 2: $O_{ij}|Y_i = 1 \sim Bernoulli(p)$ where $j = 1, 2, ..., k$

We assume that if $Y_i = 0$, then $O_{ij} = 0$ with probability 1 for $j = 1, 2, ..., k$. That is, there are no false detections. This model can also be written in terms of logit parameters as follows:

- Hierarchy 1: $Y_i \sim Bernoulli(\frac{exp(\beta)}{1+exp(\beta)})$ for $i = 1, 2, ..., n$ where $\beta = log(\psi/(1-\psi))$
- Hierarchy 2: $O_{ij}|Y_i = 1 \sim Bernoulli(\frac{exp(\delta)}{1+exp(\delta)})$ for $i = 1, 2, ..., n$ where $\delta = log(p/(1-p))$

The second parameterization is commonly used when there are covariates and the logit link is used to model the dependence of the occupancy and detection probabilities on the covariates. Notice that $p = \frac{exp(\delta)}{1+exp(\delta)}$ and $\psi = \frac{exp(\beta)}{1+exp(\beta)}$. If there are covariates that affect the occupancy probability, the familiar Logistic regression corresponds to $\psi(X_i) = \frac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)}$.

We use the following non-informative priors for the two parameterizations.

- The (ψ, p) parameterization: $\psi \sim Uniform(0, 1)$ and $p \sim Uniform(0, 1)$
- The (β, δ) parameterization: $\beta \sim N(0, 1000)$ and $\delta \sim N(0, 1000)$

These are commonly used non-informative priors on the respective scales (e.g., Kery and Schaub, 2011). One of the important goals of occupancy studies is to compute the probability that a site is, in fact, occupied when it is observed to be unoccupied on all visits. This is different than ψ. To compute this, we need to compute the probability that a site that is observed to be unoccupied is, in fact, occupied. We can compute it by using standard conditional probability arguments as: $P(Y_i = 1|O_{ij} = 0, j = 1, 2, .., k) = \frac{(1-p)^k\psi}{(1-p)^k\psi+(1-\psi)}$.

TABLE 2 | Simulation results for $n = 30$ and $k = 2$.

	$p = 0.3, \psi = 0.3$		$p = 0.8, \psi = 0.3$		$p = 0.3, \psi = 0.8$	
	Prob	**Logit**	**Prob**	**Logit**	**Prob**	**Logit**
p	0.3079	0.1864	0.7394	0.7855	0.3648	0.2950
	(0.3295)	(0.2391)	(0.733)	(0.7725)	(0.3865)	(0.3051)
ψ	0.4168	0.7786	0.3438	0.3240	0.6904	0.9174
	(0.4085)	(0.6865)	(0.3356)	(0.3297)	(0.6696)	(0.8666)
Occupancy	0.2535	0.7054	0.0324	0.0196	0.4581	0.8567
	(0.2574)	(0.5919)	(0.04142)	(0.0399)	(0.4537)	(0.7739)

Parameter estimates as well as predicted probability under probability scale and Logit scale are quite different. Numbers in the parentheses are the standard errors.

TABLE 3 | Parameter estimates for the American Toad occupancy data using non-informative Bayesian under different parameterization.

Parameter	**Bayes probability**	**Bayes Logit**
p	0.3245	0.2314
ψ	0.5770	0.8183
$P(Y_i = 1\|O_{ij} = 0, j = 1, 2, 3)$	0.2960	0.6715
Total occupancy	0.5568	0.7932

We first present a simulation study where we show the differences in the non-informative Bayesian inferences between the two parameterizations. The R program used to conduct these simulations is available in the link provided in the data availability statement. We present the simulation results for the case of 30 sites and two visits to each site. We consider three different combinations of probability of detection and probability of occupancy; both detection and occupancy small, occupancy large but detection small and occupancy small and detection large.

Table 2 shows that the Bayesian inferences about point estimates of the probability of occupancy and detection and more importantly about the probability that a site is, in fact, occupied when it is observed to be unoccupied on both visits (denoted by "occupancy" in the table for brevity) are dependent on the parameterization. This has significant practical implications: The total occupancy rate (defined explicitly in the next paragraph), that is often needed by the managers, depends on $P(Y_i = 1|O_{ij} = 0, j = 1, 2, .., k)$, will be quite different depending on which parameterization is used. The biases observed here, although somewhat reduced, persisted as the sample size was increased to 50 and 100 but with only two visits per site. Notice also that variation under probability scale and logit scale are quite different. For more detailed simulations on the effect of prior distributions on the parameter estimation (but not the prediction) when covariates are involved, see Northrup and Gerber (2018) and comments that follow the paper.

How does this work out in real life situation? Let us reanalyze the data presented in MacKenzie et al. (2002). We consider a subset of the occupancy data for American Toad (*Bufo americanus*) where we only consider the first three visits. The data and the R program for this analysis are provided in the link provided in the data availability statement. As in the previous example, we conducted standard diagnostic tests such as the value of Gelman-Rubin statistics and trace plots to judge the convergence of the MCMC algorithm. In all case, we had excellent convergence. There are 27 sites that have at least three visits. Number of sites that were observed to be occupied at least once during the three visits was 10. Hence, the raw occupancy rate, the proportion of sites occupied at least once in three visits, was 0.37. We fit the constant occupancy and constant detection probability model using the two different parameterizations described above. We report, in **Table 3**, the Bayesian point estimates of: Probability of detection (p), probability of occupancy (ψ), probability of occupancy when the site was never observed to be occupied during the three visits, namely, $P(Y_i = 1|O_{ij} = 0, j = 1, 2, .., k)$ under two different parameterizations. The total occupancy rate is computed by adding the number of sites that were observed to be occupied at least once during the surveys (these are the sites that are definitely occupied) to the probability of occupancy for those sites that were never observed to be occupied during the surveys (these sites might have been occupied but were not observed to have been occupied due to detection error), namely $P(Y_i = 1|O_{ij} = 0, j = 1, 2, .., k)$ and dividing the number by the total number of sites. Total occupancy rate is often used to make management decisions.

The differences in the two analyses are striking. According to one analysis, we will declare an (observed to be) unoccupied site to have probability of being occupied as 0.296 where as the other analysis it is 0.672, more than double the first analysis. Given the data, after adjusting for detection error, we will declare the study area to have occupancy rate to be 0.56 under one analysis but under the other analysis, we will declare it to be 0.79. Both of these Bayesian estimates also differ from the ML estimate of 0.55 (This is slightly different than the one reported, 0.49, in MacKenzie et al., 2002 because, unlike the original analysis, we have considered a subset of sites that were visited exactly three times for ease of computation). The ML estimate is close to the Bayesian estimate with a flat prior on the probability scale but not to the one obtained by non-informative prior on the Logit scale.

In **Figure 4**, we show the posterior distributions for the total occupancy rate under the two parameterizations.

The difference between the two posterior distributions is shocking. Such posterior distributions form the basis for deciding the status of the species. It is obvious that the decisions based on these two posterior distributions are likely to be very different. Now imagine facing a lawyer in the court of law or a politician who is challenging the results of the wildlife manager who is testifying that the occupancy rates are too low (or, too high for invasive species). All they have to do, while still claiming to do a legitimate non-informative analysis, is use a parameterization that gives different results to raise the doubt in the minds of the jurors or the senators on the committee. This is not a desirable situation.

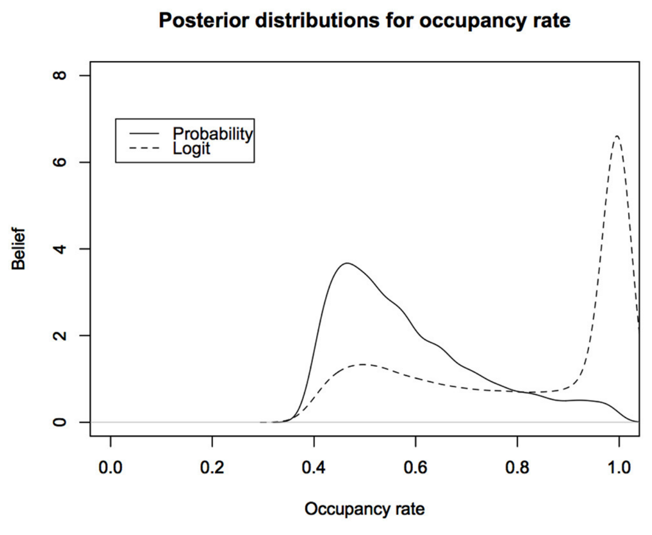


FIGURE 4 | Posterior predictive distributions for the occupancy rates of American Toad under different parameterizations: The logit scale leads to a distribution that is highly skewed toward probability of occupancy close to 1. Probability of occupancy often depends on habitat covariates and is modeled with Logistic regression. This figure indicates that we might be biasing the inferences about probability of occupancy under the non-informative Bayesian analysis.

4. UNINTENDED CONSEQUENCES OF OBJECTIVE PRIORS

Scientific and statistical inference is not limited to inference about the parameters of the generating mechanism as it is formulated. Inference also extends to inference on functions of the parameters, including predictions. So far, we have studied in concrete terms the consequences of the lack of parameterization invariance in important ecological problems at commonly observed sample sizes, especially in wildlife management. These consequences, of course, vanish as the information in the data increases. Unfortunately, whether or not the observed sample size is large depends on the complexity of the model. In this section, we provide general arguments against subjective and objective priors in scientific inference in general.

Consequence 1: All subjective Bayesian inferences can be masqueraded as objective (flat prior) Bayesian inferences.

This result is simply a converse of Fisher's result that all flat priors on one scale are not flat on any other scale. Let Y be a random variable such that $Y \sim f(.;\theta)$. Let $\theta \subset \Theta$ be continuous and Θ be a compact subset of the real line. Let $\pi(\theta)$ denote the prior distribution. A basic probability result on transformation (e.g., Casella and Berger, 2002) is the following: If $\theta \sim \pi(\theta)$, then the probability density function of $g(\theta)$, a one-one, differentiable transformation of θ is given by $\pi(g^{-1}(\theta))|\frac{dg^{-1}(\theta)}{d\theta}|$.

Thus, if θ has uniform distribution on Θ, any one-one, differentiable transformation of it has a density function that is proportional to $|\frac{dg^{-1}(\theta)}{d\theta}|$ which is not a uniform distribution. This is the basis of Fisher's criticism of the flat priors: What is "non-informative" on one scale is "informative" on any transformed scale.

The converse of the result, not noted in the literature to the best of our knowledge, is equally devastating. Suppose a researcher has a subjective prior in mind, say $\pi(\theta)$ that is not a uniform distribution on Θ. Let $G(\theta)$ denote the cumulative distribution function corresponding to this density. The researcher may have this particular prior in mind because he truly believes it but he realizes that he may face the criticism of being biased with an agenda to prove. To avoid the criticism, he can easily rewrite his model in terms of $\varphi = G^{-1}(\theta)$. This transformation is also known as the probability transform and is used to generate random numbers from univariate, continuous random variables (Gentle, 2004). It is well known that φ has a Uniform distribution on $(0, 1)$. When presenting the results of his analysis, the researcher simply presents his model in terms of φ and a Uniform prior on $(0, 1)$. Many Bayesian analysts would consider this as an "objective" Bayesian analysis that is not tainted by subjective priors and that it has "let the data speak." This is patently a false statement: The researcher started with a subjective prior but was able to masquerade it as an "objective" analysis.

Consequence 2: Induced priors on functions of parameters are not flat, thus leading to cryptic biases in scientific inference.

Scientific inference is usually not limited to the natural parameters of the generating mechanism but may be based on functions of parameters. Often these functions of the natural parameters are really the parameters of scientific interest. For example, in the PVA example that we studied, the natural parameters of the Ricker model were (a, b) or (a, K). But the analysis was not limited to conducting inference about these parameters alone. We are interested in computing the probability of extinction or the time to extinction or the PPI (e.g., Dennis et al., 1991). These are usually *functions* of the natural parameters. Similarly for the occupancy model, the natural parameters are (ψ, p) but quantities of interest are predicted probability of occupancy when the site was never observed to be occupied. As we saw earlier, this is also a *function* of the natural parameters. When one specifies a prior distribution on the natural parameters, it induces a prior distribution on all transformations of the natural parameters including such functions of the parameters.

In PVA, one of the quantities of interest is the probability of (quasi)extinction, that is, the probability that the population will dip below a threshold. For the stochastic versions of the continuous time exponential growth models, Dennis et al. (1991) compute this explicitly. The basic model (for discrete time case) may be written as: $X_{t+1}|X_t \sim Normal(X_t + \mu, \sigma^2)$. Let x_e be the log-threshold population size and x_0 be the current log-population. Let $x_d = x_0 - x_e$. The probability of (quasi)extinction is given by $\pi(x_d, \mu, \sigma^2) = exp(-2\mu x_d/\sigma^2)$ for $\mu > 0$. If $\mu < 0$, the population goes to extinction with certainty and hence that case is not of interest. In **Figure 6**, we show the priors induced on this quantity under different non-informative priors on μ and σ^2 (without changing the parameterization). The solid curve corresponds to using $\mu \sim logNormal(0, 10)$ and $\sigma \sim logNormal(0, 10)$ and the dotted curve corresponds to using $\mu \sim Uniform(0, 10)$ and $\sigma \sim Uniform(0, 10)$.

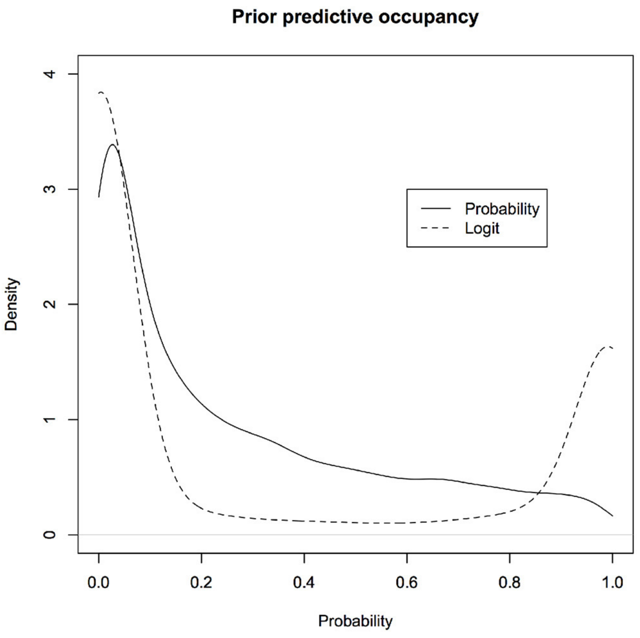


FIGURE 5 | Induced priors on the quasi extinction probability: Different non-informative priors on the parameters of a stochastic Exponential growth model lead to different induced priors on the probability of quasi-extinction. We are biasing the result even before any data are conducted. The induced prior on the probability of quasi-extinction, the parameter of interest, is not uniform. One induced prior (red) is implicitly assuming that probability of extinction is highly likely to be zero whereas another induced prior (blue) implicitly assumes the probability of extinction is mostly between 0 to 0.2 or 0.8 to 1 but not much in between.

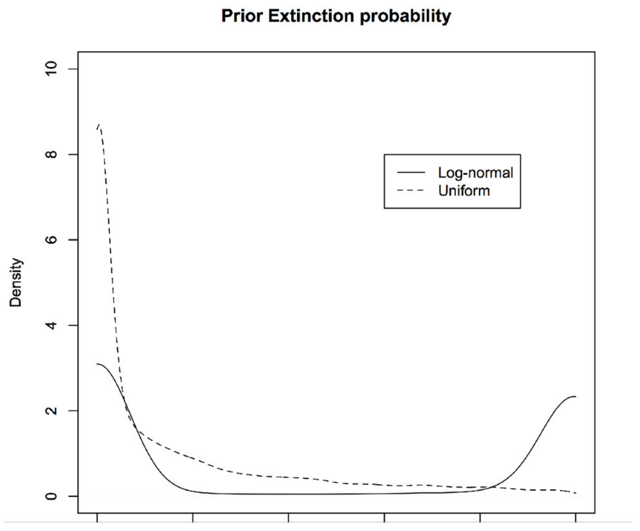


FIGURE 6 | Induced priors on the occupancy probability: Different non-informative priors induce different priors on the parameter of interest, namely, probability that a site is occupied given that we have not observed it to be occupied while surveying due to detection error. We are biasing the results of the survey even before conducting the survey. One induced prior (black) is implicitly assuming that probability of occupancy is more likely to be zero whereas another induced prior (red) implicitly assumes the probability of occupancy is near 0 or 1 but not much in between.

Let us look at the induced prior distribution for $P(Y_i = 1|O_{ij} = 0, j = 1, 2, .., k) = \frac{(1-p)^k\psi}{(1-p)^k\psi+(1-\psi)}$, the predicted occupancy, under different non-informative priors of p and ψ for $k = 2$. The solid curve in **Figure 5** corresponds to the induced prior on the predictive occupancy using $\psi \sim Uniform(0, 1)$ and $p \sim Uniform(0, 1)$ and the dotted curve corresponds to using non-informative prior commonly used on the (β, δ) scale as described previously in the discussion of the occupancy problem, namely, $\beta \sim Normal(0, 10)$ and $\delta \sim Normal(0, 10)$.

It is clear that different versions of the non-informative priors on the natural parameters induce different priors (and, hence biases) on the induced parameters that are of scientific interest. In Lele (2004) and Lele and Allen (2006), it was argued that even if one can elicit priors from the experts on the natural parameters, expert may not be aware of, and in fact, may not even agree with the prior distributions induced by his own priors on the natural parameters. In a recent paper, Seaman et al. (2012) point out the same issues but in the context of flat priors, extending Fisher's criticism of the flat priors.

There is a hidden danger of using flat priors uncritically. Unwittingly the researcher might be biasing the conclusion about the interesting functions of the parameters while falsely claiming the mantle of "objectivity." Even when we have flat priors on the natural parameters, the induced priors on the quantities of inferential interest are extremely likely to be biased to one conclusion or the other.

Consequence 3: The assumption of independent parameters, although convenient for MCMC calculations, creates unrealistic priors.

Most ecological models involve multiple parameters with complex parameter spaces. Because of the interdependencies between these parameters, the valid parameter values are dependent on each other. It is usually quite difficult to specify flat priors or almost flat priors as in priors with large variability on such non-trivial parameter spaces. Even Jeffreys priors (e.g., Ronneberg, 2017) that address Fisher's objection and are invariant to parameterization, or Bernardo's reference priors (e.g., Ronneberg, 2017) are extremely difficult to construct for multiparameter situations. In practice, most of the Bayesian analysis for multiparameter models is conducted with priors that assume that parameters are *a priori* independent. For example, in simple Capture-Recapture models, it is often assumed (e.g., Parent and Rivot, 2013) that probability of recapture and population size are independent of each other. Similarly in regression analysis, it is assumed that the regression parameters are independent of the each other *a priori*. In fact, as is clear from any analysis of Capture-Recapture experiments that the parameters are intricately related to each other. The assumption of prior independence of parameters is seldom justified but is taken as a convenient assumption. Questions one must ask are: What are the consequences on the distribution of the functions of parameters that are of real interest? What effect would this have if we reparameterize the model where new parameters are functions of the original parameters? For example, if we assume (a, b) are independent of each other, it is clear that (a, K) are bound to be correlated parameters because K, the carrying capacity, is a function of both a, growth parameter and b, the density dependence parameter. Is this correlation sensible *a priori*?

Ecologists are justifiably skeptical of the assumption that the data are independent of each other and are well aware of the consequences of such assumption; the famous pseudo replication problem in ecology (Hurlbert, 1984). However, they seem to accept, somewhat uncritically, the assumption of *a priori* independence between the parameters. Computational convenience should not be the driving force behind choosing prior distributions. Prior distributions have consequences; sometimes they are intended but most of the times they are unintended and not understood explicitly.

Consequence 4: Bayesian prediction intervals may not have correct coverage

Management decisions are based not only on the parameter estimates but also, and perhaps more importantly, on prediction of future events. In an important recent paper, Shen et al. (2018) consider the problem of prediction and predictive densities from the Classical and Bayesian perspective. They define a predictive density in a general form as: $f^P(y) = \int f(y;\theta)dQ(\theta)$. They show that such predictive density will lead to correct predictive coverage provided $Q(\theta)$ is a valid confidence distribution that has correct frequentist coverage properties. As a consequence, the Bayesian predictive density that uses posterior distribution as $Q(\theta)$, will lead to valid predictive coverage only if the posterior distribution has correct frequentist properties. Unfortunately, posterior distributions do not always have, in fact seldom have, correct frequentist coverage properties unless the information in the data is substantial. Of course, in that case, there remains no difference between a Bayesian and a frequentist inference.

Whether information in the data is substantial or not is not a simple function of the sample size; it also depends on the complexity of the model. The more complex the model is, the larger is the sample size required (e.g., Dennis, 2004). As a consequence, the objective, flat prior based analyses may not even lead to predictions that are valid. Why should we expect them to lead to management decisions that are sensible in practice?

Consequence 5: Reparameterization to facilitate MCMC convergence may influence scientific inference.

Markov Chain Monte Carlo algorithms have made it feasible to analyze highly complex, hierarchical models. One of the major difficulties in the application of the MCMC algorithms is the convergence of the underlying Markov chain to stationarity. When the parameters in the model are highly correlated (also, termed weakly estimable) or if the parameters are non-identifiable or non-estimable (See Ponciano et al., 2012; Campbell and Lele, 2014), it is difficult to obtain convergence and good mixing of the MCMC chains.

To alleviate this problem, one needs to reparameterize the model so that the parameters are orthogonal or weakly correlated with each other. Such reparameterization of the model will have no consequences if the inferences are based on the likelihood function which is invariant to such reparameterization but, as shown above, can have serious consequences for a non-informative Bayesian inference.

As an aside, such orthogonalization of the parameters is feasible only if the parameters are identifiable. Diagnostics for non-estimability and non-identifiability of the parameters is automatic under the data cloning based likelihood estimation (Lele et al., 2010; Ponciano et al., 2012; Campbell and Lele, 2014), however such diagnostics is not possible under the Bayesian approach (Lele, 2010).

5. DISCUSSION

Using different parameterizations of a statistical model depending on the purpose of the analysis is not uncommon. For example, in survival analysis the exponential distribution is written using the hazard function or the mean survival function depending on the goal of the study. They are simply reciprocals of each other. Similarly Gamma distribution is often written in terms of rate and shape parameter or in terms of mean and variance that is suitable for regression models. Beta regression is presented in two different forms: regression models for the two shape parameters or a regression model for the mean keeping variance parameter constant (Ferrari and Cribari-Neto, 2004). All these situations present a problem for flat and other non-informative priors because same data and same model can lead to different conclusions depending on which parameterization is used. One can possibly construct similar examples in the Mark-Capture-Recapture methods where different parameterizations are commonly used.

Indeed, as the sample size increases, effect of the prior diminishes and Bayesian and likelihood inferences become similar. However, in practice, hierarchical models are fairly complex and involve substantially more parameters than in the models considered in this paper. Dennis (2004) illustrates that as number of parameters increases, effects of the choice of a prior linger even for large samples.

Hierarchical models in ecology tend to be complex and can easily lead to non-identifiable parameters (Lele et al., 2010; Ponciano et al., 2012; Campbell and Lele, 2014). If there are non-identifiable parameters, effect of the prior *never* vanishes. Owhadi et al. (2015) explore effect of the priors on Bayesian inferences in a mathematically rigorous fashion and conclude that the Bayesian inference is very brittle. Hence, the results presented here are likely to be far more common in practice than may be imagined.

To summarize, we have shown that non-informative priors neither "let the data speak" nor does the analysis based on them correspond, even roughly, to likelihood analysis for the sample sizes feasible in ecological studies. Non-informative priors add their own cryptic biases to the scientific conclusions. Just because the terms objective priors, non-informative priors or objective Bayesian analysis are used, it does not mean that the analyses are not subjective. A truly subjective prior based on expert opinion is, perhaps, preferable to the non-informative priors because in the former case the subjectivity is clear and well quantified, and, may even be justified, whereas in the latter the subjectivity is hidden and not quantified.

Hierarchical models in themselves are extremely useful to model complex ecological phenomena. The many successes of the so called "Bayesian" approach are actually attributable to sensible uses of hierarchical models for pooling information, e.g across different studies or resolutions. These success stories have nothing to do with the use of the Bayesian philosophy or use of priors.

Many applied ecologists are using the non-informative Bayesian approach as a panacea to deal with hierarchical models, erroneously believing that they are presenting objective, unbiased results and that there are no alternative approaches. Hierarchical models can be and are analyzed using the likelihood and frequentist methods. Given the complexity of these models, the number of parameters involved and the different ways the same model potentially can be formulated; the resultant analysis, because of the lack of invariance to parameterization, may have unstated and unqualified biases. Hence it may be easily challenged in the legislature and in the court of law.

DATA AVAILABILITY STATEMENT

This R code and data are available at: https://github.com/jmponciano/LELE_NoninformativeBayes.

AUTHOR CONTRIBUTIONS

SL conceived of the project, conducted the analysis and wrote the paper.

FUNDING

This work was partially supported by a grant from National Science Engineering Research Council of Canada to SL.

ACKNOWLEDGMENTS

This work was initiated while the author was visiting the Center for Biodiversity Dynamics, NTNU, Norway and the Department of Biology, Sun Yat Sen University, Guangzhou in 2014. Two anonymous referees along with H. Beyer, B. Dennis, C. McCulloch, J. Ponciano, P. Solymos and especially M. Taper provided many useful comments that have improved the manuscript significantly. B. Dennis provided the kit fox data. This paper is an extended version of an arXiv paper (Lele, 2014).

REFERENCES

Bolker, B. M. (2008). *Ecological Models and Data in R.* Princeton, NJ: Princeton University Press.

Campbell, D., and Lele, S. R. (2014). An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Comput. Stat. Data Anal.* 70, 257–267. doi: 10.1016/j.csda.2013.09.013

Casella, G., and Berger, R. L. (2002). *Statistical Inference*, Vol. 2. Pacific Grove, CA: Duxbury Press.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecol. Lett.* 8, 2–14. doi: 10.1111/j.1461-0248.2004.00702.x

De Valpine, P. (2009). Shared challenges and common ground for Bayesian and classical analysis of hierarchical statistical models. *Ecol. Appl.* 19, 584–588. doi: 10.1890/08-0562.1

Dennis, B. (1996). Discussion: should ecologists become Bayesians? *Ecol. Appl.* 6, 1095–1103. doi: 10.2307/2269594

Dennis, B. (2004). "Statistics and the scientific method in ecology," in *The Nature of Scientific Evidence,* eds M. L. Taper and S. R. Lele (Chicago, IL: University of Chicago Press), 327–378.

Dennis, B., Munholland, P. L., and Scott, J. M. (1991). Estimation of growth and extinction parameters for endangered species. *Ecol. Monogr.* 61, 115–143. doi: 10.2307/1943004

Dennis, B., and Otten, M. R. (2000). Joint effects of density dependence and rainfall on abundance of San Joaquin kit fox. *J. Wildlife Manage.* 64, 388–400. doi: 10.2307/3803237

Efron, B. (1986). Why isn't everyone a Bayesian? *Am. Stat.* 40, 1–5. doi: 10.1080/00031305.1986.10475342

Ferrari, S. L. P., and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.* 31, 799–815. doi: 10.1080/0266476042000214501

Fisher, R. A. (1930). Inverse probability. *Math. Proc. Cambridge Philos. Soc.* 26, 528–535. doi: 10.1017/S0305004100016297

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-BA117A

Gentle, J. E. (2004). *Random Number Generation and Monte Carlo Methods*. New York, NY: Springer.

Hurlbert, S. H. (1984). Pseudo-replication and the design of ecological field experiments. *Ecol. Monogr.* 54, 187–211. doi: 10.2307/1942661

Kery, M., and Royle, A. (2016). *Applied Hierarchical Modelling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS.* New York, NY: Elsevier.

Kery, M., and Schaub, M. (2011). *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective.* New York, NY: Academic Press.

King, R., Morgan, B., Gimenez, O., and Brooks, S. (2009). *Bayesian Analysis for Population Ecology.* Boca Raton, FL: CRC Press.

Lele, S. R. (2004). "Elicit data, not prior: on using expert opinion in ecological studies," inn *The Nature of Scientific Evidence,* eds M. L. Taper and S. R. Lele (Chicago, IL: University of Chicago Press), 410–436.

Lele, S. R. (2010). Model complexity and information in the data should match: could it be a house built on sand? *Ecology* 91, 3493–3496. doi: 10.1890/10-0099.1

Lele, S. R. (2014). Is non-informative Bayesian analysis appropriate for wildlife management: survival of San Joaquin Kit fox and declines in amphibian populations. *arXiv:1502.00483*.

Lele, S. R., and Allen, K. L. (2006). On using expert opinion in ecological analyses: a frequentist approach. *Environmetrics* 17, 683–704. doi: 10.1002/env.786

Lele, S. R., and Dennis, B. (2009). Bayesian methods for hierarchical models: are ecologists making a Faustian bargain. *Ecol. Appl.* 19, 581–584. doi: 10.1890/08-0549.1

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* 10, 551–563. doi: 10.1111/j.1461-0248.2007.01047.x

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *J. Am. Stat. Assoc.* 105, 1617–1625. doi: 10.1198/jasa.2010.tm09757

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial N. *Ecology* 94, 2173–2179. doi: 10.1890/13-0176.1

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255. doi: 10.1890/0012-9658(2002)0832248:ESORWD2.0.CO;2

Natarajan, R., and McCulloch, C. E. (1998). Gibbs sampling with diffuse priors: a valid approach to data driven inference? *J. Comput. Graph. Stat.* 7, 267–277. doi: 10.1080/10618600.1998.10474776

Northrup, J. M., and Gerber, B. D. (2018). A comment on priors for Bayesian occupancy models. *PLoS ONE* 13:e0192819. doi: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

Owhadi, H., Scovel, C., and Sullivan, T. (2015). Brittleness of Bayesian inference under finite information in a continuous world. *Electr. J. Stat.* 9, 1–79. doi: 10.1214/15-EJS989

Parent, E., and Rivot, E. (2013). *Introduction to Hierarchical Bayesian Modelling of Ecological Data.* London: Chapman and Hall/CRC.

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 10.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Syst. Biol.* 61, 955–972. doi: 10.1093/sysbio/sys055

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models and Applications, 2nd Edn.* New York, NY: Wiley Interscience.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rannala, B., Zhu, T., and Yang, Z. (2012). Tail paradox, partial identifiabilty and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.* 29, 325–335. doi: 10.1093/molbev/msr210

Ronneberg, L. T. S. (2017). *Fiducial and objective inference: history, theory and comparisons* (Master's thesis). Department of Mathematics, University of Oslo, Oslo, Norway.

Saether, B. E., Engen, S., Lande, R., Arcese, P., and Smith, J. N. (2000). Estimating the time to extinction in an island population of song sparrows. *Proc. R. Soc. B Biol. Sci.* 267:621. doi: 10.1098/rspb.2000.1047

Seaman, J. W. III., Seaman, J. W. Jr., and Stamey, J. D. (2012). Hidden dangers of specifying non-informative priors. *Am. Stat.* 66, 77–84. doi: 10.1080/00031305.2012.695938

Shen, J., Liu, R. L., and Xie, M. (2018). Prediction with confidence? A general framework for predictive inference. *J. Stat. Plan. Inference* 195, 126–140. doi: 10.1016/j.jspi.2017.09.012

Solymos, P. (2010). dclone: data cloning in R. *R J.* 2, 29–37. doi: 10.32614/RJ-2010-011

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 13 March 2020
doi: 10.3389/fevo.2020.00035

How Should We Quantify Uncertainty in Statistical Inference?

*Subhash R. Lele**

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

An inferential statement is any statement about the parameters, form of the underlying process or future outcomes. An inferential statement, that provides an approximation to the truth, becomes "statistical" only when there is a measure of uncertainty associated with it. The uncertainty of an inferential statement is generally quantified in terms of probability of the strength of approximation to the truth. This is what we term "inferential uncertainty." Answer to this question has significant implications in statistical decision making where inferential uncertainty is combined with loss functions for predicted outcomes to compute the risk associated with the decision. The Classical and the Evidential paradigms use aleatory (frequency based) probability for quantifying uncertainty whereas the Bayesian approach utilizes epistemic (belief based) probability. To compute aleatory uncertainty, one needs to answer the question: *which experiment is being repeated, hypothetically or otherwise?* whereas computing epistemic uncertainty requires: *What is the prior belief?* Deciding which type of uncertainty is appropriate for scientific inference has been a contentious issue and without proper resolution because it has been commonly formulated in terms of statements about parameters, that are statistical constructs, not observables. Common to these approaches is the desire to understand the data generating mechanism. Whether one follows the Frequentist or the Bayesian approach inferential statements concerning prediction are aleatory in nature and are practically ascertainable. We consider the desirable characteristics for quantification of uncertainty as: (1) Parameterization and data transformation invariance, (2) correct predictive coverage, (3) uncertainty that depends only on the data at hand and the hypothesized data generating mechanism, and (4) diagnostics for model misspecification and guidance for correction. We examine the Classical, Bayesian and Evidential approaches in the light of these characteristics. Unfortunately, none of these inferential approaches possesses all of our desiderata although the Evidential approach seems to come closest. Choosing an inferential approach, thus, involves choosing between either specifying the hypothetical experiment that will be repeated or equivalently a sampling distribution of the estimator or a prior distribution on the model space or an evidence function.

Keywords: aleatory probability, conditional inference, empirical validation, epistemic probability, parameterization invariance, prediction, predictive densities, statistical paradigms

Edited by:
Yukihiko Toquenaga,
University of Tsukuba, Japan

Reviewed by:
Nathan Lemoine,
Colorado State University,
United States
Tore Schweder,
University of Oslo, Norway

Correspondence:
Subhash R. Lele
slele@ualberta.ca

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *12 March 2019*
Accepted: *06 February 2020*
Published: *13 March 2020*

Citation:
Lele SR (2020) How Should We Quantify Uncertainty in Statistical Inference? Front. Ecol. Evol. 8:35.
doi: 10.3389/fevo.2020.00035

1. INTRODUCTION

It is indisputable that statistical reasoning has become an essential component of modern scientific thinking (Taper and Ponciano, 2016). However, discussions on the philosophical foundations of statistical methods are often regarded as esoteric and of little practical importance to the scientific practitioners (e.g., Clark, 2005). It is commonly claimed that pragmatic scientists somehow know which method is appropriate for their own problem and they do not need to worry about the differences in the philosophies of statistics that underlie such methods. That such differences are too subtle to be of any practical relevance (e.g., Kery and Royle, 2016). One possible reason scientists feel this way is because they often make decisions solely on the basis of the estimated effect size while paying only a lip service to the magnitude and nature of the associated uncertainty, in spite of the repeated protestations by the statisticians that "effect size estimate without the associated uncertainty" is useless for decision making. Understanding the meaning and the quantification of uncertainty is a major hurdle, both in practical applications of statistics and in understanding the arguments for and against different paradigms in statistics.

Why is uncertainty quantification a critical endeavor for science and scientific decisions? Decisions are ultimately based on the predictions of the future outcomes of a statistical experiment. These predictions are uncertain and hence we need to quantify their uncertainty. Prediction uncertainty has several components. First component is the process variation. It exists even if all the parameters of the model are known. This variation can be reduced to some extent by appropriate use of covariates and auxiliary information in the process modeling. Second component is the estimation error. This occurs because parameters in the process model are generally unknown and not directly observable. These parameters need to be estimated using the observed data. Different methods of estimation lead to different estimation errors. Both these components assume that the form of the model used for prediction correctly represents the true underlying process. The third component to prediction uncertainty is the uncertainty about the form of the process model. This uncertainty can be controlled to some extent by appropriate model selection and model diagnostics. Prediction uncertainty is a combination of these three components. Given the prediction uncertainty, we can combine it with the loss function that quantifies the consequences of different decisions that are based on the uncertain predictions. The combination of the loss function and the three types of uncertainties leads to the quantification of risk. A rational decision maker is presumed to choose a course that minimizes the risk. Thus, if one wants to make rational decisions, one needs a verifiable quantification of the uncertainty in prediction. In this paper, we discuss the quantification of the prediction uncertainty when there is no model form uncertainty. Dennis et al. (2019) discuss the effect of model mis-specification on the quantification of uncertainty.

The goal of this paper is to discuss various ways statisticians quantify uncertainty in statistical inferential statements about the parameters of the model and the observables. Here observables refer to both observed data and future data that are potentially observable. Parameters of the model, although statistical constructs and not always useful for prediction in specific circumstances, are important for developing scientific understanding (e.g., Jerde et al., 2019). However, uncertainty statements about the parameter estimates are difficult to directly verify in practice. On the other hand, statements about the observables are aleatory or frequentist in nature and hence are directly ascertainable in practice. Predictive accuracy has been at the center of much of the development in the statistical learning literature (e.g., Hastie et al., 2009) and has also been suggested as the appropriate approach to statistical thinking (Billheimer, 2019). We emphasize, however, that it is not sufficient to compare predictive abilities of different procedures. Ability to diagnose and pinpoint errors in modeling and being able to learn from errors is an essential component when comparing the desirability of various inferential procedures (e.g., Dennis, 1996; Lele and Dennis, 2009).

Although many of the discussions in the literature often concentrate on estimation and testing of the parameters of the model, the scope of statistical inference is wider than that. For example, scientists want to be able to forecast future outcomes under different "what if" scenarios or they may be interested in studying derived quantities, such as probability of extinction or time to extinction of a species. Model choice, estimation and prediction are three important components of any scientific enquiry. In the next section, we discuss desiderata for uncertainty quantification in the context of this general scope. In section 3, we will discuss the basics of the Classical paradigm to quantify uncertainty. We emphasize the difference between pre-data and post-data measures of uncertainty and difficulties faced by the Classical approach. This will lead us to the discussion of conditional inference, relevant subsets and ancillary statistics. We discuss the quantification of uncertainty in the context of prediction. This discussion will clarify the importance of conditioning, not just on intuitive grounds, but in practical terms. In section 4, we will review the basics of (subjective) Bayesian inference, from estimation to prediction. We will briefly discuss the effect of the choice of the prior distribution. But the main emphasis will be on discussing the meaning of the uncertainty in the Bayesian context, namely the epistemic probability and its interpretation. Determination of the prior distribution along with the lack of ability to pinpoint errors in modeling are the main stumbling blocks in the Bayesian approach. In section 5, we will discuss the solution offered by the Evidential paradigm to the problem of prediction. In particular, we use normalized predictive likelihood to obtain evidential predictive density and study its performance. Section 6 summarizes the results and offers general conclusions. Throughout this paper, we assume that the reader is familiar with the basic concepts in statistical inference, such as different probability distributions, maximum likelihood estimation, confidence intervals etc. See, for example, any introductory level textbook on statistical inference, such as Ramsey and Schafer (2002) or a mathematical text, such as Casella and Berger (2002). Some of the topics, however, may need a somewhat more advanced mathematical understanding, although we have tried to make it accessible by providing simple examples and intuition where possible.

2. DESIDERATA FOR UNCERTAINTY QUANTIFICATION

Before we can compare different approaches to quantify uncertainty in statistical inference, we need to have a list of desirable characteristics that such quantification will possess in an ideal world. The following characteristics are generally agreed upon as desirable in the statistical literature, although not all in one place.

1. Uncertainty quantification should be invariant to both data transformation and parameterization of the model.
2. Uncertainty quantification should reflect the informativeness of the observed data for the underlying process.
3. Uncertainty quantification should be amenable to be probed empirically for possible violations. This is also sometimes described as "being ascertainable in practice."
4. If an uncertainty quantification is not sufficiently accurate, it should be possible to diagnose potential problems in the model and ways to correct them.

We will examine uncertainty quantifications in three inferential paradigms in the light of these desiderata.

Before we proceed further, we discuss the first desideratum that can be potentially confusing for a non-statistician. Let us consider the problem of prediction of amount of biomass of a grass species in a typical plot or a quadrat. Suppose we measure the biomass in the units of kilograms. We may report a 90% prediction interval as, say (2.3, 3.5). This says, that if we randomly select say 1,000 quadrats and measure their biomass in kilograms, then ∼90% of the quadrats will have biomass between 2.3 and 3.5 kg. Someone else, who happens to measure the biomass in the units of pounds, the corresponding 90% prediction interval would have been (5.06, 7.7). The equivalent prediction interval has different end points depending on the unit but the uncertainty, namely the probability content of the interval, 90%, does not depend on the unit of measurement or data transformation. Similarly, suppose we report a 90% confidence intervals for probability of occupancy of a plot by a species as, say (0.2, 0.8). The corresponding 90% confidence interval for the log-odds of occupancy will be, approximately (−1.38, 1.38). These intervals clearly look different with different widths but their coverage probabilities are identical, namely, 90%. The desideratum says that these *coverage probabilities*, that are a measure of uncertainty, should not change as a consequence of data transformation or a particular choice of parameterization.

In the following, we will be using two different notions of probability. Fox and Ulkumen (2011) give the following characteristics of the two kinds of probabilities or uncertainties:

Pure epistemic uncertainty:

- is represented in terms of a single case,
- is focused on the extent to which an event is or will be true or false,
- is naturally measured by confidence in one's knowledge or model of the causal system determining the outcome, and
- is attributed to missing information or expertise.

Pure aleatory uncertainty, in contrast:

- is represented in relation to a class of possible outcomes,
- is focused on assessing an event's propensity,
- is naturally measured by relative frequency, and
- is attributed to stochastic behavior.

They define the two concepts as follows.

- **Aleatory probability:** An aleatory conception of uncertainty involves unknown outcomes that can differ each time one runs an experiment under similar conditions.
- **Epistemic probability:** An epistemic conception of uncertainty involves missing knowledge concerning a fact that either is or is not true.

Fox and Ulkumen (2011) claim that disagreement concerning the nature of uncertainty persists to this day in the two dominant schools of probability theorizing, with frequentists treating probability as long-run stable frequencies of events, and Bayesians treating probability as a measure of subjective degree of belief.

3. HOW FREQUENTLY WOULD WE BE CONTRADICTED? ALEATORY PROBABILITY FOR UNCERTAINTY QUANTIFICATION

Let us consider one of the most common problems in ecology: prediction of the total biomass of a species in a study area. Let us assume that the study area can be divided in N management quadrats of equal area. For the time being, we will consider estimating the mean biomass in a typical management quadrat. Suppose we take a sample of n quadrats and measure the biomass in each of them. How can we use this information to infer about the mean biomass in a typical quadrat? Furthermore, how can we use this information to predict biomass in the unsampled quadrats? To be able to go from what we observe (biomass in the sampled quadrats) to what we have not observed (biomass in the unsampled quadrats), we need to make some assumptions. For the sake of simplicity, let us assume that the quadrats are similar to each other in terms of habitat covariates that may affect the amount of biomass and that amount of biomass in one quadrat does not affect the amount of biomass in other quadrats. Furthermore, we assume that the quadrats chosen for measurement were chosen randomly. If N is substantially larger than n, we can ignore the subtle differences between "with replacement" and "without replacement" sampling. Also, for the simplicity of notation, we will say that the sampled quadrats were the first n of the N quadrats.

In mathematical notation, the amount of biomass in the N quadrats, $Y_1, Y_2, ..., Y_N$, are assumed to be independent, identically distributed random variables. The sampled observations are the biomasses at the sampled quadrats, namely, $y_1, y_2, ..., y_n$. Let us further assume that $Y_i \sim N(\mu, \sigma)$ where μ indicates the mean biomass in a quadrat and σ indicates the natural variation. We use the standard deviation (sd) σ, instead of the commonly used parameterization σ^2, because it has the same unit as the mean. Let us look at a simple implication of this assumption. Suppose the mean biomass in a quadrat is 10 kg and sd is 1. Then, the distributional assumption implies that probability that Y, the biomass at any quadrat, is in the interval $(10 - 1, 10 + 1)$ is ∼0.68. What do we mean by this statement? To most scientists, this means that about 68% of the quadrats will have biomass between 9 and 11 kg. This is an aleatory probability. In statistical literature we call this the "frequentist" definition of probability. It is the proportion of times an event is observed in infinite replications of the experiment. The N quadrats are independent replications of the experiment and we expect about 68% of them to have biomass between 9 and 11 kg. If, in practice, the observed proportion turns out to be substantially different than 0.68, we know that our statistical model is inappropriate. An important characteristic of aleatory probability statements is that they are ascertainable in practice. Thus, they are *probeable* statements and we can also diagnose problems with the data generating mechanism if the statements are refuted in practice.

There are a few unknowns in our situation: (1) value of the parameters (μ, σ), and (2) appropriateness of the probability density function, namely the Normal density function to model the underlying process. Statistics, often, is considered the epistemology of science. We want to learn from the data about these unknowns. For the time being, let us assume that the Normality assumption is appropriate and also that $\sigma = 1$ is known. The maximum likelihood estimator (MLE) of the parameter μ is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$. Notice that $\hat{\mu}$ is a random variable and the corresponding estimate (the value obtained for a particular sample), with some abuse of notation, is given by $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$. This is simply a number. This number is an inferential statement about the mean biomass in a management quadrat, namely μ. Thus, after sampling, one may say that "mean biomass of a species in a management quadrat is 8.3 kg." We can also make statements such that if we sample a new management quadrat, assuming we know the true parameters, the probability that it will have biomass between 3 and 5 kg is about 0.68. Both these statements are "inferential statements" but are quite different in their nature. First statement is about a parameter, a statistical construct, whereas the second statement is about an observable. Given such statements, a natural question to ask is: How certain (or, uncertain) are we about these statements? This corresponds to determining the probability of the strength of approximation to the truth. Answering such questions is the crux of statistical inference.

3.1. Sampling Distribution and Confidence Intervals

We will start with discussing uncertainty in the parameter estimation. Later we will discuss inferential statements about observables. Neyman (1937) proposed to quantify uncertainty in the parameter estimation by answering the question: If there were another scientist who had sampled n quadrats, albeit different than the one we sampled, how different would be their estimate of μ? The distribution of the estimates obtained by infinitely many scientists repeating the experiment is called the *sampling distribution*. Sampling distribution quantifies uncertainty in the Classical statistical inference.

Let us continue with the biomass survey example. Suppose the true mean biomass in any quadrat is 10 kg and known true sd is 1. Suppose the sample size is 20. Then to obtain the true sampling distribution of the estimator of μ, namely $\hat{\mu} = \overline{Y}$, we generate 20 random numbers from $N(10, 1)$ and compute the sample mean. If we repeat this process, say 1,000 times, we will obtain 1,000 sample means (equivalent to estimates from 1,000 independent surveys). Histogram of these 1,000 means represents the *true* sampling distribution (strictly speaking, simulation based estimate of the true sampling distribution). It shows, if we repeat the study, how different the estimates will be, namely, probability of the strength of approximation. **Figure 1** (black curve) illustrates an example of the *true* sampling distribution. In reality, we cannot compute the *true* sampling distribution because we do not have data from replications of the experiment. Fortunately, given the data at hand, one can *estimate* the sampling distribution. In **Figure 1** (dotted curve), we illustrate a parametric bootstrap estimate of the sampling distribution given data in hand. For this, given the results of our one survey, we compute the sample mean. Then generate 20 random numbers from $N(\bar{y}, 1)$ and compute the sample mean. If we repeat this process, say 1,000 times. We will obtain 1,000 sample means (equivalent to estimates from 1,000 independent surveys). Histogram of these 1,000 means represents the parametric bootstrap estimate of the sampling distribution. Notice that we have replaced the true mean 10 by its estimate $\bar{y}$. Naturally the true and estimated sampling distributions are slightly different from each other but this is what one can do in practice because true mean is not known. For each data set in hand, because the sample means are different for different data sets, the bootstrap estimate of the sampling distribution is different.

Sampling distributions can be estimated using various other techniques, such as using pivotal statistics, asymptotic normal approximation, inversion of the likelihood ratio or by non-parametric bootstrapping (Casella and Berger, 2002). As an aside, the last two techniques are considered preferable because they lead to confidence intervals that are parameterization equivariant. That is, one can transform the confidence interval for μ to $log(\mu)$ by simply log-transforming the endpoints of the first interval. Although their lengths and end points will change, their coverage properties remain invariant. Thus, likelihood ratio based or bootstrap based confidence intervals satisfy desiderata 1 but confidence intervals based on other methods may not. We will discuss implications to other desiderata in the next section.

Let us look at how one can use the (true and estimated) sampling distribution for quantifying uncertainty about the inferential statements.

3.1.1. Confidence Intervals and Coverage

It is easy to see that we can use the *true* sampling distribution to compute an interval that indicates the range of estimates that one would obtain in replicated experiments with specific probability. For example, using the true sampling distribution which, in this case, can be analytically shown to be $N(\mu, \sigma/\sqrt{n})$,

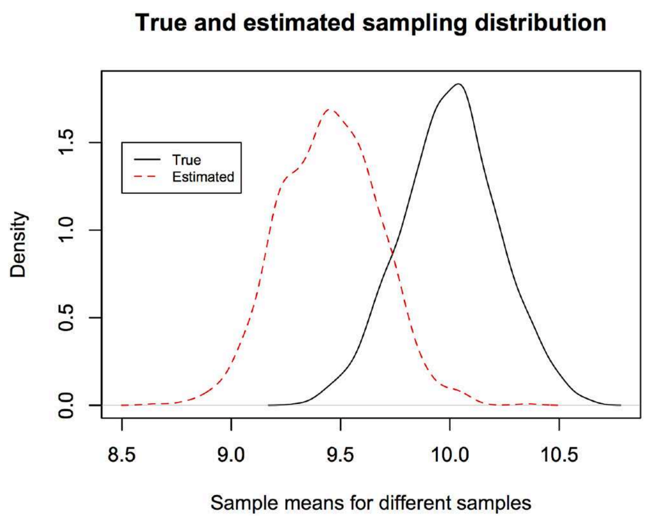


FIGURE 1 | The estimated sampling distribution depends on the observed data and is different from the true sampling distribution. Hence the parameter estimate of a new study may lie outside the confidence interval reported in an earlier study more often than the nominal error rate. The new estimate is occurring from the true sampling distribution and the previous confidence interval is based on the estimated sampling distribution. It is approximately the area outside the reported confidence interval under the true sampling distribution.

we can give 90% confidence interval as $(10 - 1.68\sigma/\sqrt{n}, 10 + 1.68\sigma/\sqrt{n})$ where n denotes the sample size and $\sigma = 1$. The confidence interval shrinks as we increase the sample size. As we noted before, it is impossible to compute this interval in practice because the true parameter values are unknown. The true 90% confidence interval for a sample size 20 is given by (9.624341,10.37566). A corresponding estimated 90% confidence interval based on the *estimated* sampling distribution, for a specific sample, turns out to be (9.716948,10.460803). This is different from the true confidence interval because we replace true mean by the estimated mean. For different samples, one would get different confidence intervals because each sample leads to a different estimate of the mean. The reader can use the R program in the **Supplementary Material** to see how parametric bootstrap sampling distribution and associated confidence interval varies depending on the sample in hand. Note that each run of the program will lead to different confidence intervals than reported above.

It is clear what information the true 90% confidence interval provides. It says that if you repeat the experiment, your estimate will lie inside the true confidence interval 90% of times. Hence your result will contradict the original result only 10% times. But what information does the *estimated* confidence interval provide about the true value of μ? We can make the following statement about the value of μ: If we replicate the experiment 100 times and calculate the estimated 90% confidence interval for each replication, then ∼90% of the intervals will cover the true value (that is, the true value will belong to the interval). Of course, any *particular* interval obtained from a single experiment may or may not contain the true value. This is the property of the procedure and not of the outcome of a single experiment. The interpretations of the true confidence interval (that can never be computed) and the estimated confidence interval are different.

Thus, we have answered the question, how often (in replicated experiments) would our interval cover the true parameter value of μ? This is called the coverage probability. Is this useful? We contend that this is the kind of probability we use in practice. For example, probability of an airplane crash on a take-off is say 1 in 10,000. This tells us nothing with certainty about what will happen on a particular flight; it may crash or it may not crash. However, we intuitively understand this uncertainty statement and are able to make decisions. It helps us behave in a rational manner. This is what Neyman called "inductive behavior" (Lehmann, 1995), behavior informed by the data.

Replicability of the conclusions: Another question explicitly addressed by the sampling distribution is: How replicable is our study? How likely is it that we would be contradicted by someone conducting similar experiment? This is sometimes crudely put as "Cover Your Ass" (CYA) statements. For example, suppose the first sampler publishes a confidence interval for the mean biomass in a given size quadrat. Then we can use the *true* sampling distribution to compute the probability that subsequent sampling of the biomass will yield a mean biomass estimate that will not belong to the first sampler's confidence interval and hence the first sampler's conclusions will be contradicted by the subsequent study. This probability is not the same as the coverage probability which is the property of the *estimated* confidence interval. For example, for the estimated sampling distribution in **Figure 1** (dotted curve), the probability that a new sampler will get an estimate outside the *estimated* confidence interval (μ_L, μ_U), namely, $P(\hat{\mu}_{new} \notin (\mu_L, \mu_U))$, turns out to be, on an average, 0.24. This is *larger* than the nominal 10% excesses under the true sampling distribution. Of course, as the sample size increases, this problem goes away. We conjecture that this is one of the reasons of the replicability crisis in science (e.g., Ioannidis, 2012), namely incorrect interpretation of the confidence interval; the other, perhaps far more important, being model misspecification or the model from one study not being applicable to other studies.

Replicability of the conclusions is an essential component of the scientific validity of the conclusions. Aleatory probability based quantification of uncertainty clearly tries to address this concern. Not everyone, however, agrees that classical quantification of uncertainty is useful. It is claimed that not all experiments can (will) be replicated. For example, the critics ask: How do we quantify uncertainty of the event of a nuclear war? How do we replicate a time series of populations? We find this objection fundamentally vacuous because, by its very nature, modeling of a natural phenomenon using a statistical model assumes the possibility of replication of the experiment. If replication of an experiment is impossible, statistical modeling of such an experiment is also impossible, nay meaningless. Unfortunately, even if we accept the Classical approach to quantification of uncertainty in principle, there are problems when applied to inferential statements.

3.2. Conditional Inference and Post-data Uncertainty

Let us continue with the question of estimating the mean biomass in each management quadrat. Previously, we assumed that all quadrats were identical to each other. It is reasonable to think that each quadrat has different mean biomass that depends on the habitat covariates of that quadrat. Let us assume that $Y_i \sim N(X_i\beta, \sigma)$. This is a simple linear regression through the origin model with a single habitat covariate and constant standard deviation.

Given the data, that now consist of (y_i, x_i) where $i = 1, 2, ..., n$, the MLE of β is given by $\hat{\beta} = \sum x_i y_i / \sum x_i^2$ where the summation runs over $i = 1, 2, ..., n$. Suppose, again unrealistically, that the standard deviation is known. The question now is: What is the uncertainty associated with the estimator of the slope β? Because of the Normality assumption, we can represent the uncertainty using the variance of the estimator. Surprisingly, there are two possible answers to this question.

1. Conditional variance: The standard answer in regression analysis, e.g., Ramsey and Schafer (2002), is $Var(\hat{\beta}|(x_i, i = 1, 2, ..., n)) = \sigma^2 / \sum x_i^2$. Notice that the variance of $\hat{\beta}$ depends on σ but more importantly also on the observed values of the covariates $x_1, x_2, ..., x_n$. If the observed set of covariates are widely dispersed, the variance of $\hat{\beta}$ is small whereas if the observed set of covariates are not dispersed, the variance is large. This is why, in planning ecological studies or constructing sampling designs, we aim to have high dispersion in the covariate values. To most researchers, this makes intuitive sense. With this, the true sampling distribution of $\hat{\beta}$ is given by:

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum x_i^2})$$

   This measure of uncertainty assumes that the replicated experiments are such that the covariate values are identical to the ones in the original experiment, namely, $x_i, i = 1, 2, ..., n$. The only difference between the replicate experiments is in the values of the responses Y_i, conditional on the original covariate values. This is why it is called "conditional variance."
2. Unconditional variance: On the other hand, one can argue that because our study is an observational study, if we replicate the experiment the specific covariate values that different experimenters would observe are likely to be different. Thus, an argument can be made that when characterizing uncertainty we should account for the possible variation in the covariates as well. Let us assume that the covariate values arise from $N(0, 1)$. That is, if we plot a histogram of the covariate values from all the management quadrats, it will have a bell shape. Under this assumption, it can be shown that, $Var(\hat{\beta}) = \sigma^2/(n-2)$. This is the variation in $\hat{\beta}$ that we will observe if we replicate the experiment where the covariate values are not fixed. This variation does not depend on the covariate values because their values across the replications are different and hence are averaged over. Because we do not condition on the covariate values, this is called "unconditional variance." In this case, the true sampling distribution is (now, approximately) given by:

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{(n-2)})$$

It is obvious that the length of the true confidence interval is constant in the unconditional case whereas it depends on the particular covariate composition in the conditional case. Using the distribution of $\sum x_i^2$, we can find that, for smallish sample sizes, about 60% of the conditional confidence intervals will be shorter than the unconditional intervals and as the sample size increases 50% of the conditional confidence intervals are shorter and 50% are longer than the unconditional confidence intervals.

These conditional and unconditional confidence intervals can be obtained in practice by using bootstrapping (Wu, 1986; Efron and Tibshirani, 1993). There are two different ways to conduct bootstrapping for regression. One is called pairwise bootstrap where we resample with replacement from the pairs (x_i, y_i). This leads to unconditional confidence interval. On the other hand, one can resample with replacement from the residuals $r_i = y_i - \hat{\beta}x_i$ denoted by r_i^* and then generate the bootstrap samples using $y_i^* = \hat{\beta} * x_i + r_i^*$. Notice that in this bootstrap, covariate values are identical throughout the boostrapping procedure. This conditional (also called, residual) bootstrap leads to conditional confidence intervals. Notice that residual bootstrap procedure assumes that the linear regression model is the true model whereas the pairwise bootstrap procedure does not assume the correctness of the linear regression. Thus, pairwise bootstrap is model robust.

Both conditional and unconditional answers are *mathematically* correct (that is, they have correct coverage under the appropriate replication, conditional or unconditional) but which one is *scientifically* appropriate? It makes sense to use the conditional variance if we want to report uncertainty about the estimate that we obtained based on our *own particular data*. For example, if we happen to get a really good sample, that is, observed sample covariate values are highly dispersed, we should be fairly confident that our particular estimated slope is pretty close to the true slope. On the other hand, if we were unlucky and got a sample such that the covariate values were not very dispersed, we should not be too confident about the slope estimate being close to the true slope. The unconditional variance, on the other hand, seems to penalize a lucky experimenter and award an unlucky experimenter by averaging over their performances. But if we want to protect against possible contradiction by other experimenters, who will get different covariate values than what we observed, reporting the unconditional variance makes sense. The answer seems to be "it depends on the scope of the inference."

This has puzzled, stumped and bothered the frequentist statisticians for a very long time (e.g., Fisher, 1955; Cox, 1958; Buehler, 1959; Royall and Cumberland, 1985; Casella and Goustis, 1995; among many other papers). We will let the reader read through these papers to see the full technical and scientific discussion. The ambiguity of when and how to condition has led to the study of relevant subsets, subsets of the sample space over which replication should be considered, along with conditioning on appropriate ancillary statistics and more. Much of this discussion revolves around uncertainty in the parameter estimates. These are statistical constructs. Although, intuition suggests that conditional inference is both mathematically correct and scientifically appropriate, there is no direct, operational way to justify the quantification of uncertainty about a statistical construct. Suppose we can relate the discussion to uncertainty about the observables then may be we can make such statements ascertainable in practice. Would the prediction accuracy help us decide if the conditional inference is "scientifically appropriate" without resorting to intuition alone?

3.3. Prediction and Prediction Intervals

Let us first look at how we can solve the problem of prediction and its uncertainty using the Classical approach (e.g., Lejeune and Faulkenberry, 1982). Let θ_T denote the true value of the parameter and let us assume that the model is correctly specified. The goal is, given the sampled data, to predict the new observation and associated prediction uncertainty. This could be equivalently translated into estimating either the density function $f(y;\theta_T)$, the corresponding cumulative distribution function (CDF) $F(y;\theta_T)$ or, more directly the inverse of the cumulative distribution function, the quantile function, $F^{-1}(\alpha;\theta_T)$. Let us look at the estimation of the density function.

3.3.1. Estimated Predictive Density

Given the data, we can simply replace the true, but unknown, parameter θ_T by its estimated value $\hat{\theta}$ and use $f^p_{est}(y) = f(y;\hat{\theta})$ to obtain prediction intervals for a new observation.

Here superscript p indicates predictive and subscript *est* indicates estimated predictive density approach. This is certainly parameterization invariant (at least when MLE is used to estimate the parameter), as it should be, but depends on the transformation of the observable. These properties can be proved quite easily.

1. Let us reparameterize the density using $\psi = g(\theta)$ where $g(.)$ is a one-to-one function. Then, we can write $\theta = g^{-1}(\psi)$ where $g^{-1}(.)$ is the inverse function of $g(.)$. The density is only a function of y and hence it follows that $f^p_{est}(y) = f(y;\hat{\psi}) = f(y;g(\hat{\theta}))$.
2. Let us do a data transformation where $z = h(y)$. In this case, we have to use the Jacobian of the transformation (Casella and Berger, 2002) to get the density in terms of z. The density in terms of z is given by $f^p_{est}(z) = f(z;\hat{\theta})|dh^{-1}(z)/dz|$. The density in terms of z looks quite different. However, if $z_1 = h(y_1)$ and $z_2 = h(y_2)$, then $P(Z \in (z_1,z_2)) = P(Y \in (y_1,y_2))$. The prediction intervals are different but the probability content is the same.

This makes perfect sense: If we measure the variable on a different scale, the prediction interval should depend on that scale. For example, suppose population abundances are modeled as Log-normal distributions. Then, log-abundances are distributed as a Normal distribution. One can obtain prediction intervals for the log-abundances using Normal distribution properties and simply transform the end points using the exponential transform to get the prediction intervals for the abundances. Both these intervals, although numerically quite different, have exactly the same probability content under the respective distributions. The coverage probability of the prediction interval, the uncertainty quantification, remains invariant to the choice of the data transformation as well as the choice of the parameterization.

The major problem with the estimated predictive density is that it tends to be too optimistic in the sense that it gives prediction intervals that are too short and that do not have appropriate coverage properties. Notice here that the predictive error statement is aleatory and probeable (Taper et al., 2019), either by using cross validation or by independent experiments. One reason for bad coverage property of the estimated predictive density is that it does not take into account the estimation error in $\hat{\theta}$ (e.g., Aitchison, 1975; Cox, 1975). There are many different approaches to account for the estimation error (e.g., Smith, 1998) each with its own pros and cons. One of the straightforward approaches (e.g., Hamilton, 1986) is based on accounting for estimation error by using the following.

3.3.2. Classical Predictive Density

$$f^p_C(y) = \int f(y;\theta)\phi(\theta;\hat{\theta},I^{-1}(\hat{\theta}))d\theta$$

Where $\phi(\theta;\hat{\theta},I^{-1}(\hat{\theta}))$ is the asymptotically Normal sampling distribution of the estimator and $I(\hat{\theta})$ is the usual estimated Fisher Information matrix (e.g., Casella and Berger, 2002; Ramsey and Schafer, 2002).

Notice that the integration is with respect to θ and not $\hat{\theta}$, which makes a clean, philosophically sound justification for this approach awkward. The estimated Fisher Information matrix can be replaced by the observed Fisher Information matrix (e.g., Efron and Hinkley, 1978). The above definition of predictive density, of course, assumes that the sampling distribution of $\hat{\theta}$ can be well-approximated by the specified Normal distribution. One can, naturally, replace the asymptotic approximation of the sampling distribution by the bootstrap estimate of the sampling distribution (Harris, 1989). In the context of the linear regression problem discussed above, this immediately raises the question: "which sampling distribution" should we use for integration, conditional or unconditional? For example, a pairwise bootstrap for regression (Efron and Tibshirani, 1993) will lead to different predictive density than using the residual bootstrap (e.g., Wu, 1986). The first one leads to unconditional whereas the second one leads to conditional sampling distribution but assumes that the regression model is appropriate. A conditionally appropriate solution to this problem was provided by Vidoni (1995) where he uses the $p^{\star}$-approximation to the distribution of the MLE as suggested by Barndorff-Nielsen (1983). He also uses the Laplace approximation (Tierney and Kadane, 1986) to avoid the integration altogether. What properties are satisfied by the Classical predictive density?

Shen et al. (2018) (see also Lawless and Fredette, 2005; Schweder and Hjort, 2016) consider the prediction problem from the frequentist perspective in detail. They consider a general form of the predictive density, namely $f^p_Q(y) = \int f(y;\theta)dQ(\theta) = \int f(y;\theta)q(\theta)d\theta$. where $Q(\theta)$ is any distribution on the parameter values of θ. The different predictive densities described above are particular cases of this general form with different $Q(\theta)$. For example, when we use the Classical predictive density following Hamilton (1986), we use $Q(\theta) = Normal(\theta, I^{-1}(\hat{\theta}))$. An important result they prove is that the Classical predictive density has correct coverage probabilities only if the estimated sampling distribution of $\hat{\theta}$ has correct frequentist coverage (Shen et al., 2018, p. 130). They show that the predictive densities in the form similar to the ones defined above are superior to the estimated predictive density (which is nothing but using a degenerate $Q(\theta)$, degenerate at $\hat{\theta}$) in terms of average Kullback-Leibler divergence and in terms of prediction error. They study parameterization invariance of the coverage in some cases. The conclusion is that it does not hold in general. The error probabilities (coverage properties) of these inferential statements are generally not parameterization invariant for small samples but they are parameterization invariant for large samples. This is because most estimators have sampling distributions that are asymptotically normal. If an estimator does not have asymptotically normal distribution, it is not clear if the parameterization invariance will hold true in such cases.

The predictive density for the linear regression through origin (also considered by Shen et al., 2018), using the conditional variance, is easy to derive and to justify by noting that:

$$Y_{new} - X_{new}\hat{\beta} \sim N(0, \sigma^2 + \sigma^2 \frac{X_{new}^2}{\sum X_i^2})$$

where the second component in the variance is due to the estimation error of $\hat{\beta}$. This is how, generally, one obtains the prediction interval for linear regression (e.g., Ramsey and Schafer, 2002).

One can obtain an approximate predictive density based on the unconditional variance as:

$$Y_{new} - X_{new}\hat{\beta} \sim N(0, \sigma^2 + \sigma^2 \frac{X_{new}^2}{n-2})$$

An obvious comparison would be to see which density comes closest to the true density

$$Y_{new} - X_{new}\beta \sim N(0, \sigma^2)$$

See **Figure 2** for a visual comparison between estimated, conditional and unconditional predictive densities (for a particular observed sample) along with the true predictive density. In the figure, we illustrate four different samples to show that sometimes estimated predictive density comes closer to the true density and sometimes it can be quite different, depending on how close the estimated parameters are to the true parameters. The general predictive density $f_Q^p(y)$ averages these different estimated predictive densities to get, on an average, better performance.

Shen et al. (2018) compare the prediction coverage performance of the estimated, exact conditional and using the conditional bootstrap sampling distribution. In the **Supplementary Material**, we provide an R code that confirms that both conditional and unconditional predictive densities lead to correct predictive coverage of a future observation but conditional prediction intervals are shorter than the unconditional intervals when $\sum X_i^2 > (n-2)$ and longer otherwise. An immediate implication is that because conditional prediction intervals have correct coverage, when the unconditional prediction interval is shorter than the conditional prediction interval, it will have less than nominal coverage for those covariate configurations and when unconditional interval is longer than the conditional interval, it will have larger than nominal coverage for other covariate configurations. This implies that unconditional intervals are either unnecessarily conservative or incorrectly optimistic, but never correct conditionally (although correct on an average). This justifies the use of conditional variance in practical terms instead of "intuition." See Royall and Cumberland (1985) for a similar argument in the context of finite population sampling. The differences between conditional and unconditional prediction intervals can be substantial when there are large number of covariates that leads to more variation in the covariate configurations.

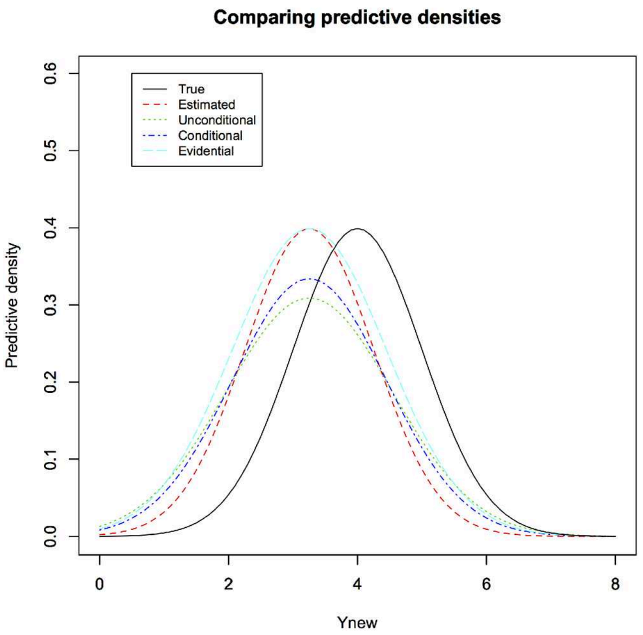


FIGURE 2 | The true density for the new observation under the linear regression through origin is different than the estimated predictive density based on the observed data. Classical conditional, Classical unconditional have slightly fatter tails than the estimated predictive densities. This leads to somewhat better coverage properties by accounting for the sampling variability. Evidential predictive density also has fatter tails than estimated predictive density. Its calculation, however, does not need sampling distribution and hence specification of the experiment to be repeated. It reflects the information in the observed data appropriately.

3.4. What Should We Do?

It is clear that reporting the uncertainty in inferential statements about the *parameters* is tightly related to the question of "which experiment do we replicate?" Reporting the uncertainty about the parameters leads to the difficulties of "unconditional" vs. "conditional" (sometimes also termed pre-data and post-data) uncertainty. Because models and parameters are purely a statistical construct, the uncertainty statements related to their values are not justifiable directly and in practical terms. On the other hand, the observations have real world meaning. Reporting the uncertainty in statistical inference procedure in terms of its predictive accuracy is unambiguous. Thus, we can compare and contrast different uncertainty quantifications in terms of their predictive accuracy. For example, looking at the predictive accuracy, we can conclude that conditional predictive uncertainty is not only scientifically appropriate but also practically correct and better than the unconditional predictive uncertainty. Let us summarize what we can say about the Classical predictive density in the light of the desiderata from section 2.

1. The Classical predictive density is not parameterization invariant unless the sampling distribution is completely known, that is, it is a pivotal statistics (Shen et al., 2018). Sampling distribution based on the asymptotic normal approximation or the inversion of the Likelihood ratio test based on the asymptotic Chi-square approximation or bootstrapping leads to parameterization invariance of the predictive density. Thus, parameterization invariance is achieved only when valid bootstrapping of the data is possible or when the sample size is sufficiently large. However, bootstrapping time series or spatial data is not possible without some, possibly strong, additional assumptions.
2. Most of the results regarding the predictive density are proved under the assumption that the estimators are consistent and have asymptotically normal (CAN) distribution. However, in many complex ecological models, the conditions for CAN estimation may not be satisfied. For example, estimation of the boundary parameter commonly leads to estimators that are not CAN estimators. Such models may require non-standard asymptotics where the estimators approach the true value of the parameter at a rate different than $\sqrt{n}$ or the asymptotic distribution may be different than Normal. It is unclear which of the above results hold true in such a situation.
3. The Classical predictive density does not automatically reflect how informative the observed data are. Unfortunately there is no general recipe to construct correct conditional or post-data sampling distribution for small samples. If one uses observed Fisher information (Efron and Hinkley, 1978) for the computation of predictive density, it appears to use the correct conditioning. See also Vidoni (1995) for appropriate conditioning in predictive density for small samples.
4. The Classical predictive density leads to correct predictive coverage only if the sampling distribution of $\hat{\theta}$ has correct frequentist coverage properties. In general, the validity of the confidence intervals or prediction intervals can be rigorously proved only for large samples. Unfortunately, what is a large sample and if one has it in practice is never known. Whether or not a sample size is large, depends on the complexity of the model (e.g., Dennis, 2004).
5. Of course, even with proper conditioning under the presumed model, if the true regression model in the above example were non-linear or if the variance depended on the habitat covariates, the prediction intervals would have incorrect coverage.
6. Ideas, such as cross validation can be used to test the validity of the predictive density. Thus, these inferential statements are fully probeable.
7. Model estimation and model selection using cross validation, one of the most commonly used approach in much of machine learning literature, is based on computing the mean prediction squared error or some modification of it (e.g., Hastie et al., 2009). It is important to note that the method of cross validation, as is commonly used, is based on minimizing the *unconditional* prediction error as described earlier. This is troublesome. Furthermore, cross validation based model selection and Akaike Information Criterion (AIC) are closely connected to each other (Stone, 1977). However, Dennis et al. (2019) show that, according to the Evidential paradigm, use of AIC for model selection is problematic because the probability of misleading evidence does not converge to zero as the sample size increases.
8. Instead minimizing the MPSE, we suggest that one should check if the predictive density leads to appropriate prediction coverage. One could compare the predictive density with a non-parametric estimate (if such an estimation is possible) of the data generating mechanism, e.g., a non-parametric density estimate in the case of independent, identically distributed random variables. Any differences not only indicate that the model is incorrect but also can lead to model diagnostics and model correction.

In summary, the Classical approach satisfies some of the desiderata for the quantification of uncertainty. However, in order to get the sampling distribution, we have to address the crucial question of "which experiment do we repeat" and the answer is not straight forward.

4. UNCERTAINTY IS ALL IN YOUR MIND: EPISTEMIC PROBABILITY FOR QUANTIFICATION OF UNCERTAINTY

Classical uncertainty quantification is based on the properties of the procedures over replications of a specified experiment. Implicitly what is being claimed is that if the procedure is good on an average, the specific inferences are good as well. Of course, a good cook does not guarantee that a specific meal would be good; by chance, although rarely, you might get a bad meal. Is a *specific* inferential statement based on more accurate procedure better than one based on a less accurate procedure? For example, suppose we get exactly the same blood pressure (BP) measurement based on a drug store machine vs. in a doctor's office, should we take both of them equally at face value? Intuitively most would say no. However, not all statisticians agree with the quantification of uncertainty in terms of the accuracy of the procedure. They claim, because accuracy of the procedure is no guarantee that a particular inferential statement is good or bad, we cannot use it as a measure of uncertainty. They do not think it is epistemically correct to average over samples that we could have, but did not, observe. So how should we approach the question of quantification of uncertainty of a statistical inferential statement that reflect the lucky (or unlucky) observed data appropriately?

Bayesian approach assumes that, even before collecting the data, the experimenter is able to quantify their uncertainty about the value of the parameter. This may be based on prior experience about a similar situation, e.g., measurement error of the BP machine, distribution of the BP measurements in the population, prevalence of a disease, previous surveys in that study area, related surveys elsewhere or basic natural history of the species. Suppose one can quantify such prior belief in terms of a proper statistical distribution (that means it should be positive, countably additive, integrate or sum to 1 etc.). Such a distribution is called a "prior distribution." This distribution describes the prior (to data) uncertainty about the parameter as quantified by the particular researcher. This is an epistemic uncertainty. This cannot be challenged nor can it necessarily be probed empirically. In this context, now we ask the question: In the light of the data, how do we change our prior beliefs (distribution)? Standard conditional probability calculation can be used to answer this question.

There are three components to every Bayesian analysis.

1. Prior distribution: Let θ denote the parameter of the model. This could be a vector indicating multiple parameters (as in multiple regression). Let Θ denote the parameter space, the set of values that the parameter can potentially take. We will generically denote the prior distribution by $\pi(\theta)$. This is assumed to be a proper statistical distribution. Thus, $\pi(\theta) > 0$ for all $\theta \in \Theta$ and $\int \pi(\theta)d\theta = 1$.
2. Data generation model: This is the process model that postulates how the data are generated in nature. This is a statistical distribution on the observables. It varies for different values of the parameter. We will generically denote this by $f(y_{(n)}|\theta)$ where $y_{(n)} = \{y_1, y_2, ..., y_n\}$ is the data vector.
3. Posterior distribution: The conditional probability distribution of the parameters given the data is called the posterior distribution. It is given by

$$\pi(\theta|y_{(n)}) = \frac{f(y_{(n)}|\theta)\pi(\theta)}{\int f(y_{(n)}|\theta)\pi(\theta)d\theta}$$

We want to emphasize that, under the Bayesian framework, the model, as indexed by the parameter value, itself is a random variable. The prior distribution represents the researcher's belief about how probable a particular model is to represent the underlying process. This is an epistemic probability.

The posterior distribution completely quantifies the researcher's belief about the appropriateness of the model in representing the underlying process, *after or in the light of* the observed data $y_{(n)}$. Although the process model component is an aleatory probability, the posterior distribution, that combines epistemic and aleatory probabilities, is an epistemic probability.

In the Bayesian paradigm, the posterior distribution plays the same role that sampling distribution played in the Classical paradigm. Using the sampling distribution, we obtained confidence intervals that represented the range of estimated values that one may obtain if we replicate the experiment. Using the posterior distribution, one computes an interval that represents the experimenter's *belief* about the range of values that the true parameter could take. This is called a "credible interval." There are no replicate experiments. Only one experiment was conducted and it resulted in the observed data. What changed, in the light of the data, are the prior probabilities about different parameter values. Just as the prior uncertainty was all in the mind of the experimenter, posterior uncertainty also is in the mind of the experimenter. See Brittan and Bandyopadhyay (2019) for a philosophical discussion on this point.

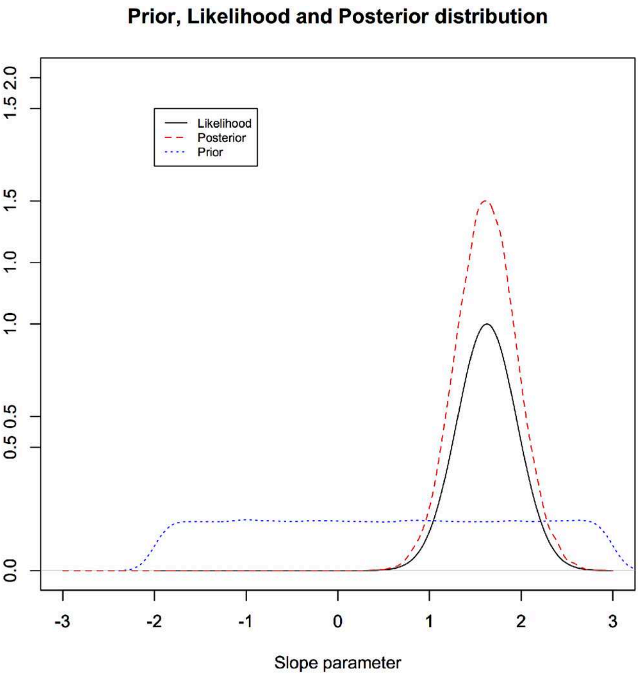


FIGURE 3 | Illustration of a prior distribution, likelihood function, and posterior distribution for the linear regression through origin: notice how much the data can change the prior beliefs. Highly informative data change the prior substantially and vice versa.

In **Figure 3**, we illustrate these three components for the linear regression through the origin example of section 3.2. We note that one can use credible interval to address the replicability of the inferential statement: How often do we believe we would be contradicted if someone replicates the experiment? The answer varies depending on the prior distribution. A credible interval does not have the interpretation of "how often would we cover the true parameter value if we repeat the experiment?" The uncertainty here is epistemic and is not testable.

Effect of the choice of the prior on the posterior distribution: It is obvious that if one has different prior beliefs, the posterior beliefs will be different even if the observed data are identical. In **Figure 4**, we illustrate how the posterior distribution changes with two different priors for the same observed data for the linear regression model considered earlier.

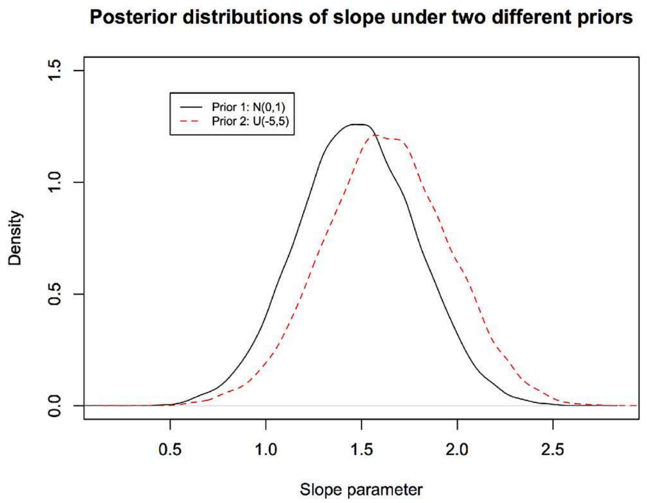


FIGURE 4 | Different prior distributions lead to different posterior distributions. They both cannot possibly have correct frequentist coverage. Their validity is epistemic and is not testable in any practical fashion.

We invite the reader to play with the R code provided in the **Supplementary Material** to see how choice of the prior affects the posterior distribution.

We emphasize again that the posterior uncertainty does not reflect simply what the data says but reflects a combined effect of the prior beliefs and the information in the data. The probability statement reflected in the credible interval has no aleatory meaning. The uncertainty here is epistemic; it is neither testable nor verifiable in any fashion.

Note: A referee raised the possibility of checking the frequentist validity of the Bayesian credible intervals using the replicate experiments. Various researchers (e.g., Datta and Ghosh, 1995) have tried to study the frequentist validity of the Bayesian credible intervals. There are two problems with this comment.

- First problem is that the Bayesian credible intervals depend on the choice of the prior. This implies that not all priors can lead to credible intervals with good frequentist properties. We do not know if our particular choice of the prior will lead to good frequentist coverage. The research related to constructing priors that lead to correct frequentist coverage, called Probability Matching Priors (e.g., Datta and Ghosh, 1995), shows that it is extremely difficult to construct such priors, even for simple models and single parameter situation.
- Second problem is that if we are using the frequentist validity as a criterion for justifying Bayesian inference, we again face the difficulty of answering the question: which experiment do we repeat, conditional or unconditional? Would we be reporting proper post-data uncertainty? This justification violates the strong likelihood principle (e.g., Berger and Wolpert, 1988), that says that uncertainty should depend only on the data at hand and not on what other data one could have observed had the experiment been repeated, that Bayesian approach considers sacrosanct.

4.1. Bayesian Prediction and Prediction Uncertainty

As we did previously, it seems reasonable to relate the uncertainty statements to observables rather than the parameters facilitating testing and falsification in practice. We will describe the ideas under the assumption that the data are independent and identically distributed but they are easily extended to non-identically distributed or dependent data, such as space-time series of population abundances.

1. Prior predictive density: We can obtain Bayesian predictions even before obtaining any data. This is called a "prior predictive distribution."

$$f(y) = \int f(y|\theta)\pi(\theta)d\theta$$

2. Bayesian predictive density: In the light of the data, the prior predictive distribution changes to posterior predictive distribution and is given by

$$f_B^p(y|y_{(n)}) = \int f(y|\theta)\pi(\theta|y_{(n)})d\theta$$

where $y_{(n)}$ denotes the data vector of length n.

4.1.1. Parameterization and Bayesian Predictive Density

According to desiderata 1, uncertainty about prediction of the future observation should not depend on the parameterization used in the modeling. To our surprise, unless we are misunderstanding, Bjornstad (1990) seems to claim that the Bayesian predictive density is not generally parameterization invariant. Suppose the prior distribution is uniform distribution on the parameter space. Then, using Laplace approximation (Tierney and Kadane, 1986), one can write the Bayesian predictive density approximately as (Leonard, 1982):

$$f_B^p(y_{n+1}|y_{(n)}) \doteq \left|I(\tilde{\theta})\right|^{0.5}\left|I(\hat{\theta})\right|^{-0.5}\frac{L(\tilde{\theta};y_{(n+1)})}{L(\hat{\theta};y_{(n)})}$$

where $y_{(n)} = y_1, y_2, ..., y_n$, $y_{(n+1)} = y_1, y_2, ..., y_n, y_{n+1}$, $\hat{\theta}$ is the MLE based on the data $y_{(n)}$ and $\tilde{\theta}$ is the MLE based on $y_{(n+1)}$. The matrix $I(.)$ is the Fisher Information matrix. The non-invariance of the Information matrix to parameterization seems to make the Bayesian predictive density non-invariant to parameterization.

4.1.2. Sensitivity to the Choice of the Prior Distribution

It is obvious that as different priors lead to different posterior distributions, they also lead to different post-predictive densities. In **Figure 5**, we first depict the prior predictive densities induced by different priors along with the true density of the new observation for the linear regression model. It is clear that prior predictive densities or equivalently, induced priors on the observations (Lele, 2020) can be quite different from each other and the true density,

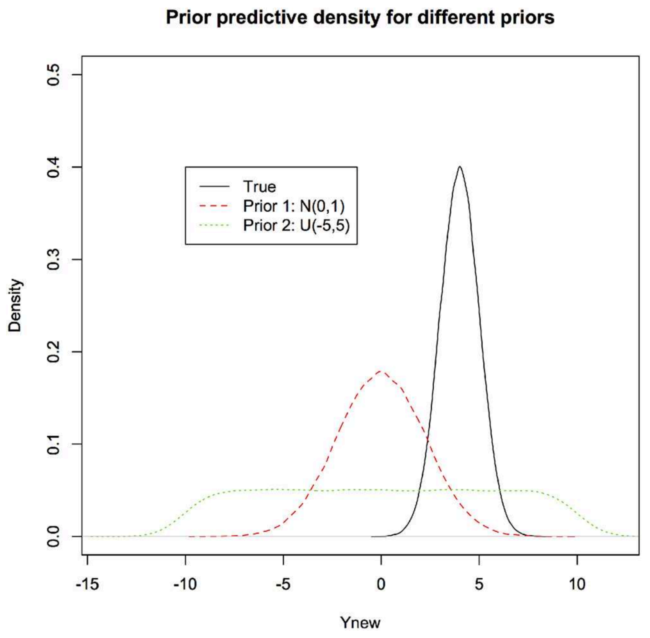


FIGURE 5 | Prior predictive densities for the linear regression through origin example under two different priors. These represent the prior beliefs about the observation to be predicted. The true density (black) is presented for comparison. Prior predictive densities could be close to the true density if the prior distribution on the parameters is "good" and they can be very far if the prior distribution on the parameters is "inappropriate." These are induced priors on the quantities of interest, namely values of the future data.

In **Figure 6**, we depict the Bayesian predictive densities corresponding to different prior distributions. The R code to produce these figures (with some Monte Carlo variation because the random numbers for each run are bound to be different) is provided in the **Supplementary Material**.

The predictive coverage for these two Bayesian predictive densities corresponds to their overlap with the true density of the new observation. Given that Bayesian predictive distributions are sensitive to the choice of the prior distribution, they all cannot possibly have correct predictive coverage.

Shen et al. (2018) show that this predictive density will lead to good coverage only if the posterior distribution is also a valid frequentist sampling distribution. Given this result, it is obvious that Bayesian predictive density is unlikely to have correct coverage properties except in special circumstances or if the sample size is large enough to use the asymptotic Normal distribution approximation. Lawless and Fredette (2005) pointed out that objective Bayesian methods do not have clear probability interpretations in finite samples, and subjective Bayesian predictions have a clear personal probability interpretation but it is not generally clear how this should be applied to non-personal predictions or decisions. Similar objections were raised by many authors, e.g., Lele and Dennis (2009), Bandyopadhyay et al. (2016), Taper and Ponciano (2016), and Brittan and Bandyopadhyay (2019).

In statistical ecological literature (e.g., Royle and Dorazio, 2008; Kery and Royle, 2016) claims are made that Bayesian procedures are valid for all sample sizes without clear specification of the criterion for validity. It is clear that Bayesian prediction intervals do not have proper coverage as they should, at least in the aleatory sense. Perhaps the validity of the Bayesian procedures is also in the minds of the researchers.

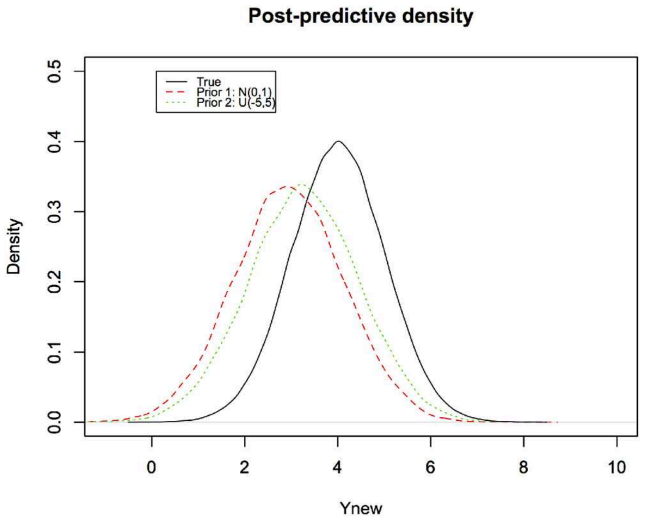


FIGURE 6 | Bayesian predictive densities representing the post-data belief about the observation to be predicted. Notice how the effect of different priors has been reduced by the data. They are much closer to the true density (black curve).

4.1.3. Using Prior Data to Construct Prior Distributions

It may be tempting to think that using past data to construct prior distributions would be a way out of the subjectivity inherent in specifying a prior distribution. There are several problems with this approach. First, using past data implies that the past experiments are identical to the present experiment. If they are not, the estimates from the prior data cannot simply be put together in a histogram and use it to construct a prior distribution. This assumption may be satisfied in a few instances but not always. Suppose it is satisfied. In that case, a question one should ask: Is this the optimal way to utilize the past data? There is an alternative approach to utilizing the past data using the so called (ironically, indeed) "Empirical Bayes approach" or "Hierarchical models" or "Meta analysis" that does not involve constructing prior distributions from the results of the past experiments. We simply combine the likelihood functions of the past data with the likelihood function of the present data, under the assumption that the parameters of these different experiments are identical to each other or somewhat related to each other. This is likely to be statistically more efficient than reducing the past data to a prior distribution.

4.1.4. Model Checking

Model diagnostics is an essential component of any statistical analysis. Bayesian model diagnostics is usually based on the Bayesian predictive density. If the data are consistent with the Bayesian predictive (commonly called, post-predictive) density, it is taken as an indication that the model structure is appropriate. However, if the data are inconsistent with the Bayesian predictive density, a natural question to ask is: What part of the model is possibly incorrect? How should we modify it? Notice that the Bayesian predictive distributions (post- or pre-data) are mixture distributions (Lindsay, 1995). It is well-known (e.g., Teicher, 1961; Lindsay, 1995) that given observations from the predictive (mixture) density, one cannot uniquely determine the data generating (mixture components) distribution and the prior (mixture weights) distribution. Hence bad post-predictive fit does not tell us whether our prior distribution that is incorrect or the data generating mechanism that is incorrect and in what fashion. Even when the Bayesian predictive density fits the observed data well, it could very well be the case that both the prior distribution and the data generating mechanism are wrong but they compensate each other's mistakes to produce the correct Bayesian predictive distribution. Hence these post-predictive checks and model diagnostics are more ambiguous and less useful for scientific analyses than one would like them to be.

4.2. What Should We Do?

As long as one is willing to provide the prior distribution, the Bayesian approach to uncertainty quantification simply follows the laws of probability to obtain posterior beliefs about the parameters and predictive distributions. This appears to be a simple, elegant and logically coherent solution to the problem of uncertainty quantification.

An oft quoted, important result related to the Bayesian paradigm, is called the Complete Class theorem (e.g., Robert, 1994). In statistical decision theory, an admissible decision rule is a rule for making a decision such that there is no other rule that is always "better" than it, where the definition of "better" depends on the loss function. According to the complete class theorems, under mild conditions every admissible rule is a (generalized) Bayes rule (with respect to some prior distribution). Conversely, while Bayes rules with respect to proper priors are virtually always admissible, generalized Bayes rules corresponding to improper priors need not yield admissible procedures. Stein's example is one such famous situation (e.g., Robert, 1994). The main caveat that is, conveniently, not stated in the quantitative ecological literature, is that Complete Class Theorem is only an existence theorem and it does not instruct us which prior leads to the admissible estimator or how to construct such a prior. If your prior happens to be different than this optimal prior distribution, your results are likely to be suboptimal, if not downright misleading.

Let us look at the Bayesian prediction in the light of the desiderata presented in section 2.

1. Bayesian predictive density is not parameterization invariant unless the sample size is sufficiently large to wipe out the effect of the prior distribution. This lack of invariance can be problematic in practice (Lele, 2020). For example, one can (deviously) choose a parameterization such that Bayesian predictive distribution comes close to what one wants. This is the same as someone choosing a prior distribution to support pre-determined conclusions.
2. Bayesian predictive density automatically reflects how informative the observed data are. This is one of the attractive features of the Bayesian approach. It does not average over good and bad samples as the unconditional variance does in the Classical approach. Bayesian approach awards the researcher if the sample is informative and punishes when it is bad.
3. Bayesian predictive density does not lead to correct predictive coverage in general. This is obvious because different prior distributions lead to different post-predictive distributions. All of them cannot have correct predictive coverage. In general, the validity of the confidence intervals or prediction intervals can be rigorously proved only for large samples. What is a large sample and if one has it in practice is never known.
4. Ideas, such as cross validation can be used to test the validity of the predictive density. Thus, these inferential statements are fully testable.
5. If the post-predictive density does not appear to have good coverage properties, we cannot say whether it is due to the incorrect data generating mechanism or due to the prior distribution. Thus, it does not guide us to modify the data generating model. This is another important practical limitation of the Bayesian approach.

To summarize, in order to quantify uncertainty in the Bayesian paradigm one has to answer the question: What is the prior distribution? The Bayesian uncertainty statements reflect personal beliefs and hence are not transferable to anyone else, unless you happen to have the same prior beliefs. Uncertainty reflected in the posterior distribution has no aleatory meaning and hence is not probeable. Furthermore, predictive statements based on the Bayesian predictive densities are not guaranteed to have correct coverage. Another important limitation of the Bayesian approach is the lack of model diagnostics and suggestions for possible model modification. It can diagnose whether the model fits the observed data or not but, when the model does not fit the observed data, it cannot localize the errors in the model specification.

5. EVIDENTIAL PARADIGM AND QUANTIFICATION OF UNCERTAINTY

We will now study the Evidential paradigm and uncertainty quantification. For detailed introduction to the Evidential paradigm (see Royall, 1997). For an easily accessible and ecologically oriented version (see Taper and Ponciano, 2016 or Dennis et al., 2019). The Evidential paradigm is still in its infancy in terms of real life applications. However, we can observe certain general properties and study it in the light of the desiderata in section 2.

Royall (1997) claims that statistical inference addresses three different questions:

1. Given these data, what is the strength of evidence for one hypothesis vis-a-vis an alternative?
2. Given these data, how do we change our beliefs?
3. Given these data, what decision should we make?

Royall (1997) uses the likelihood function to quantify the strength of evidence in the data.

5.1. Likelihood Function

Suppose $Y_1, Y_2, ..., Y_n$ are independent, identically distributed random variables with $Y \sim f(.;\theta)$ where $\theta \subset \Theta$. The likelihood function is given by: $L(\theta; y_{(n)}) = \prod f(y_i; \theta)$. Recall that likelihood is a function of θ and the data $y_{(n)} = (y_1, y_2, ..., y_n)$ are considered fixed.

5.2. The Law of the Likelihood

Let θ_1, θ_2 denote two specific values of the parameters. Then the strength of evidence for θ_2 vs. θ_1 is given by the likelihood ratio

$$LR(\theta_2, \theta_1) = \frac{L(\theta_2; y_{(n)})}{L(\theta_1; y_{(n)})}$$

with values larger than 1 implying θ_2 is better supported than θ_1 and vice versa.

Strength of evidence can be seen to be a comparison of the divergence between the true model and the two competing hypotheses (Lele, 2004; Taper and Lele, 2004; Dennis et al., 2019; Ponciano and Taper, 2019). The law of the likelihood corresponds to using the Kullback-Leibler divergence but other measures, such as the Hellinger divergence, Jeffrey's divergence, etc. also lead to appropriate quantification of the strength of evidence with some important robustness properties (Lele, 2004; Markatou and Sofikitou, 2019).

The Evidential paradigm is fundamentally different from the Classical paradigm in that it concentrates not on the control of error probabilities but on the measure of distance of the proposed models (hypotheses) from the true model. See Taper et al. (2019) for further discussion. For example, one fixes a cut-off point K that indicates the strength of evidence (difference in the divergences) that will be considered "strong evidence" *a priori*. The choice of the value of K is of the experimenter. Given such a cut-off point, if the LR is larger than this cut-off point, we say that we have strong evidence for θ_2. If it is $< 1/K$, then we say that θ_1 has strong evidence. Anything in between, we say that we have weak evidence and neither hypothesis is strongly supported. For any fixed cut-off value, one computes probabilities of misleading evidence, weak evidence and study their behavior as sample size changes. In contrast to the Classical approach where probability of type I error remains fixed at the a priori level α, probabilities of both weak and misleading evidence converge to zero as the sample size increases. See the papers by Royall (2000) and Dennis et al. (2019) for more detailed discussion on this point. Evidential approach can be extended to the case of evidence for parameter of interest in the presence of nuisance parameters using the concept of profile likelihood (Royall, 1997; Royall and Tsou, 2003).

Much of the discussion above is in the context of comparing two specified parameter values. But it is easy to construct evidential intervals (e.g., Royall, 1997; Bandyopadhyay et al., 2016; Jerde et al., 2019) that provide a range of values that are "well-supported" by the data. This is, in spirit, similar to confidence intervals and credible intervals. Notice that these intervals reflect the information in the data appropriately: Highly informative data lead to shorter evidential intervals and vice versa in the regression example of section 3.

5.3. Evidential Intervals

Let us consider a single parameter case. An evidence interval for θ at level $1/K$ is given by:

$$\left\{ \theta : \frac{L(\theta; y_{(n)})}{L(\hat{\theta}; y_{(n)})} > (1/K) \right\}$$

for a fixed value of $K > 0$. This can be generalized to evidential sets for multi-parameter situation in a straight forward fashion.

How do we quantify uncertainty in the Evidential paradigm when the inferential statements are made about the parameters of the underlying process? The probabilities of misleading evidence, weak evidence and strong evidence as defined by Royall (1997) are pre-data quantities. He does not provide any explicit suggestions as to how to report the uncertainty of the strength of evidence once the data are obtained. Should one discuss coverage probabilities of evidential intervals? As we have argued throughout this paper, without such quantification of uncertainty, the inferential statements are incomplete. Taper and Lele (2011) attempt to answer this question using bootstrapping to compute the post-data error probabilities. Taper et al. (2019) use bootstrapping to compute the distribution over the range of values strength of evidence could have taken had the experiment been replicated. Such calculations seem to be enormously informative and useful in practice. However, any such calculation requires answering the same question that Classical paradigm faced: which experiment do we replicate? Hence, although Royall's formulation quantifies the strength of evidence that satisfies the likelihood principle (but see Lele, 2004), any computation of the uncertainty in the strength of evidence seems to face the same philosophical problems Classical paradigm faces. Do we gain anything by using the Evidential approach? An affirmative answer is provided in Dennis et al. (2019) in the context of model selection. Can we use prediction to resolve this problem in general?

The evidential paradigm can also be used for prediction using various versions of predictive likelihood (e.g., Bjornstad, 1990). Let us look closely at one such predictive likelihood (Mathiasen, 1979) and our suggestion for its modification. Following Shen et al. (2018), an intuitively appealing version of evidential predictive density may be defined as follows:

5.3.1. Evidential Predictive Density

$$f_E^p(y_{n+1}|y_{(n)}) = \frac{\int f(y_{n+1}; \theta) L(\theta; y_{(n)}) d\theta}{\int L(\theta; y_{(n)}) d\theta}$$

Where y_{n+1} is the potential value of the new observation. It is necessary to assume that the integral in the denominator is finite. This may not be the case if the parameter space is infinite.

Evidential predictive density, in this formulation, is a weighted average of the data generating mechanism with weights proportional to the evidence for various parameter values in the observed data. This predictive density, in form, is identical to the predictive density one obtains with a uniform distribution as a prior distribution. However, if the parameter space is not finite, such a prior distribution is not mathematically valid as it does not integrate to 1. Let us look at the evidential predictive density a little more closely. First notice that the numerator is nothing but the likelihood function where data are now augmented by y_{n+1}. Thus, the evidential predictive density can be written as:

$$f_E^p(y_{n+1}|y_{(n)}) = \frac{\int L(\theta; y_{(n+1)})d\theta}{\int L(\theta; y_{(n)})d\theta}$$

Let $\tilde{\theta}$ denote the value of θ that maximizes $L(\theta; y_{(n+1)})$ and $I(\tilde{\theta})$ denote its Hessian, matrix of second derivatives, evaluated at $\tilde{\theta}$. Similarly, let $\hat{\theta}$ denote the value of θ that maximizes $L(\theta; y_{(n)})$ and $I(\hat{\theta})$ denote its Hessian evaluated at $\hat{\theta}$. The difference in $\tilde{\theta}$ and $\hat{\theta}$ is the effect of having a future observation equal to y_{n+1}. Now we will use the Laplace approximation described in Tierney and Kadane (1986) to evaluate the evidential predictive density approximately as:

$$f_E^p(y_{n+1}|y_{(n)}) \doteq \left|I(\tilde{\theta})\right|^{0.5} \left|I(\hat{\theta})\right|^{-0.5} \frac{L(\tilde{\theta}; y_{(n+1)})}{L(\hat{\theta}; y_{(n)})}$$

The evidential predictive density as defined above is *not* parameterization invariant (Bjornstad, 1990). Because Evidential predictive density and Bayesian predictive density are the same when one can impose a uniform prior distribution, this result also implies that Bayesian predictive density is not parameterization invariant in general. From a scientific perspective, it is clear that parameterization invariance is of fundamental importance (e.g., Bjornstad, 1990). See also Lele (2020) for practical consequences of lack of invariance in wildlife management.

Suppose we consider the part of the above approximation that is parameterization invariant as an estimate of the predictive density, namely,

$$f_E^p(y_{n+1}|y_{(n)}) = \frac{L(\tilde{\theta}; y_{(n+1)})}{L(\hat{\theta}; y_{(n)})}$$

In the following, we will call this as the evidential predictive density. Notice that the evidential predictive density is proportional to the predictive likelihood defined by Mathiasen (1979), namely $L(\tilde{\theta}; y_{(n+1)})$. Bjornstad (1990) suggests using normalized version of the predictive likelihood, namely $f_E^p(y_{n+1}|y_{(n)}) = \frac{L(\tilde{\theta}; y_{(n+1)})}{\int L(\tilde{\theta}^*; y^*_{(n+1)})dy^*_{n+1}}$ for predictive density and shows that it has good coverage properties. In our case, instead of the integral in the denominator, we use $L(\hat{\theta}; y_{(n)})$ as an approximate normalizing constant.

Let us now look at our linear regression problem to see how the evidential predictive density compares with the true density of the new observation. **Figure 2** illustrates the comparison between evidential predictive density and the true density for a new observation and for different sample sizes.

In the **Supplementary Material**, we have provided an R code that can be used to reproduce such a figure for different values of X_{new} and other variations. It is clear from this figure that evidential predictive density is a reasonable, but not very accurate, approximation of the true density of the new observation. The area under the approximate Evidential predictive density is generally not equal to 1 and that may be the reason for the discrepancy. But such standardization breaks down the invariance property. The approximation, as expected, improves with sample size. An extensive simulation study of the performance of the Evidential predictive density involving various distributions, dependent data etc. will be needed to see if this approach is better than other approaches in terms of prediction coverage or density approximation. One can, however, study the properties theoretically. The likelihood for the parameter is only interpretable in a comparative fashion as a likelihood ratio. It will be interesting to see if the Evidential predictive density ratios, that correspond to profile predictive likelihood ratios, will approximate the true predictive density ratios.

5.4. Important Properties of the Evidential Predictive Density

1. In the following, we show that this estimator is a consistent estimator of the true density $f(y_{n+1}|y_{(n)}; \theta_T)$. This is an essential property that has to be satisfied by all predictive densities. The result follows as long as the MLEs $\tilde{\theta}$ and $\hat{\theta}$ are consistent estimators of θ_T, the true parameter value.

$$\begin{aligned} f_E^p(y_{n+1}|y_{(n)}) &= \frac{L(\tilde{\theta}; y_{(n+1)})}{L(\hat{\theta}; y_{(n)})} \\ &= \frac{f(y_{n+1}|y_{(n)}; \tilde{\theta})L(\tilde{\theta}; y_{(n)})}{L(\hat{\theta}; y_{(n)})} \longrightarrow f(y_{n+1}|y_{(n)}; \theta_T) \end{aligned}$$

   as $n \longrightarrow \infty$. This is a "pointwise convergence in probability" result. It would be useful to obtain a uniform convergence result.
2. The evidential predictive density is parameterization invariant. This follows by the parameterization invariance of the likelihood function.
3. The evidential predictive density, as defined above, does not require integration, numerical or otherwise.
4. The evidential predictive density is easy to use for dependent data, such as the time series or spatial data commonly occurring in ecology and other applied sciences.
5. The evidential predictive density uses neither the sampling distribution nor the posterior distribution of the estimator, thus avoids both the specification of the experiment that is to be repeated under the Classical paradigm or choice of the prior distribution that should be chosen under the Bayesian paradigm. Evidential predictive density depends only on specification of the data generating mechanism.
6. Dealing with random effects, missing data etc. is simply a prediction problem and hence evidential predictive density can be used for analyzing hierarchical models. Thus, this approach is applicable to many ecologically interesting problems.

7. The asymptotic validity of the evidential predictive density does not depend on the asymptotic sampling distribution or asymptotic posterior distribution. It only depends on the consistency of the MLE which is a much more relaxed assumption than existence of the asymptotic distribution.
8. The evidential predictive density is conditionally appropriate. It conditions on the appropriate ancillary statistics automatically by using the likelihood function in its entirety. Highly informative data lead to tighter prediction intervals and vice versa automatically.
9. The main disadvantage of the evidential predictive density, as defined above, is that it is not guaranteed to be a probability density function. That is, it may not integrate to 1 exactly when integrated over the range of Y. Given the consistency result, this is only a small sample problem. Initial simulations suggest that even for small samples, this may not be a major problem. Similar problem arises for some non-parametric density estimators based on orthogonal polynomials (e.g., Prakasa Rao, 1983) without causing many problems in practice. A simple solution is to normalize the predictive likelihood using $\int L(y|y_{(n)}; \theta\tilde{)}dy$. This integral exists if the range of Y is finite. Simulation results in Bjornstad (1990) suggest that predictive likelihood has good coverage properties for reasonable sample size.
10. It is not completely clear how to use general evidence functions in lieu of the likelihood function in the above formulation. If such an extension is possible, one may be able to make such inferences robust against outliers.

6. CONCLUSIONS

We studied three different ways to quantify uncertainty in inferential statements. We can summarize our findings as follows.

- Classical paradigm uncertainty quantification depends on deciding which experiment to replicate. Unfortunately this leads to problems related to the pre- vs. post-data uncertainty. The Classical uncertainty quantification does not always reflect what the data at hand says about the parameter or future observations. It averages the uncertainty over all possible realizations of the process and hence punishes those who happen to have good data and awards those with bad data. This is scientifically inappropriate.
- Bayesian paradigm eschews aleatory probability and uses epistemic probability to quantify uncertainty. Bayesian approach does not need to answer the question of which experiment to replicate and reflects the information in the data at hand without averaging over what other data might have been, but were not, observed. But it requires specifying a prior distribution. Specifying a prior distribution leads to the problems of subjectivity, aside from the specification of the data generating mechanism, and possibility of untestable mis-specification. The optimality claims about the Bayesian inference are somewhat vacuous because there is no general recipe to find the prior distribution that leads to such optimal decisions.
- The Evidential paradigm addresses the issue of conditioning on the observed data appropriately. It does not require hypothetical replications of the experiment to obtain uncertainty quantification about the observables. Evidential quantification of uncertainty is aleatory, and hence falsifiable in practice, that depends only on the data generating mechanism and the choice of the evidence function. One of the reasonable objections to the classical paradigm is that the idea of replication makes no sense when analyzing time series or spatial-time series data. However, evidential support intervals, error probabilities and evidential predictive density are applicable in a straight forward fashion to dependent data, hierarchical models and other more complex situations.

  The Evidential paradigm, unlike the Classical and Bayesian paradigm, has not been extensively field tested in wide range of practical situations. Its operational feasibility is largely unknown and needs to be explored. For some examples of its applications, see Jerde et al. (2019) for an important ecological application in the study of allometry and Taper et al. (2019) in model selection for linear regression analysis. Ironically, these applications point out that reporting the strength of evidence for different models needs to be bolstered by quantification of the reliability of the estimate of the strength of evidence. If this were to be the case in other situations, it will inevitably lead to the problem of addressing the question: which experiment do we replicate? and the associated conditionality conundrum. May be we have not escaped the shackles of the hypothetical replication of experiments when it comes to making inferential statements about parameters, a statistical construct. On the other hand, evidential predictive approach seems to satisfy most of the desiderata. Although promising, jury is still out for the evidential paradigm.

In conclusion, we show that to quantify uncertainty in statistical inference, one has to choose either a specification of the sampling distribution (conditional or unconditional) or a prior distribution on the parameters or an evidence function. As scientists and statisticians, we must understand and reflect upon the implications of each of these choices.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SL conceived of the project and conducted the analysis and writing of the paper.

ACKNOWLEDGMENTS

The author thanks the insightful comments and various discussions he had with Mark L. Taper, Brian Dennis, and Jose M. Ponciano on the topic of foundations of statistical inference and ecology.

SUPPLEMENTARY MATERIAL

The R code for the simulations and figures are available at https://github.com/jmponciano/LELE_SophiesChoice.

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* 62, 547–554.

Bandyopadhyay, P., Brittan, G., and Taper, M. (2016). *Belief, Evidence, and Uncertainty*. New York, NY: Springer.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343–365.

Berger, J., and Wolpert, R. (1988). *The Likelihood Principle*. Hayward, CA: The Institute of Mathematical Statistics.

Billheimer, D. (2019). Predictive inference and scientific reproducibility. *Am. Stat.* 73, 291–295. doi: 10.1080/00031305.2018.1518270

Bjornstad, J. F. (1990). Predictive likelihood: a review. *Stat. Sci.* 5, 242–254. doi: 10.1214/ss/1177012175

Brittan, G. Jr., and Bandyopadhyay, P. S. (2019). Ecology, evidence, and objectivity: in search of a bias-free methodology. *Front. Ecol. Evol.* 7:399. doi: 10.3389/fevo.2019.00399

Buehler, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Stat.* 30, 845–863.

Casella, G., and Berger, R. L. (2002). *Statistical Inference. 2nd Edn.* Pacific Grove, CA: Duxbury Press. 337–472.

Casella, G., and Goustis, C. (1995). Frequentist post data inference. *Int. Stat. Rev.* 63, 325–344.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecol. Lett.* 8, 2–14. doi: 10.1111/j.1461-0248.2004.00702.x

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Stat.* 29, 357–372.

Cox, D. R. (1975). "Prediction intervals and empirical Bayes confidence intervals," in *Perspectives in Probability and Statistics*, ed J. Gani (London: Academic Press), 47–55.

Datta, G. S., and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* 82, 37–45.

Dennis, B. (1996). Discussion: should ecologists become Bayesians? *Ecol. Appl.* 6, 1095–1103.

Dennis, B. (2004). "Statistics and the scientific method in ecology (with commentary)," in *The Nature of Scientific Evidence*, eds M. L. Taper and S. R. Lele (Chicago, IL: University of Chicago Press), 327–378.

Dennis, B., Ponciano, J. M., Taper, M. L., Lele, S. R. (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Evol.* 7:372. doi: 10.3389/fevo.2019.00372

Efron, B., and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed vs. expected Fisher information. *Biometrika* 65, 457–482.

Efron, B., and Tibshirani, R. (1993). *An Introduction to Bootstrap*. London: Chapman and Hall.

Fisher, R. A. (1955). Statistical methods and scientific induction. *J. R. Stat. Soc. B* 17, 69–78.

Fox, C. R., and Ulkumen, G. (2011). "Distinguishing two dimensions of uncertainty," in *Perspectives on Thinking, Judging, and Decision Making*, eds W. Brun, G. Keren, G. Kirkeboen, and H. Montgomery (Oslo: Universitetsforlaget), 21–35.

Hamilton, J. D. (1986). A standard error for the estimated state vector of a state-space model. *J. Econometr.* 33, 387–397.

Harris, I. (1989). Predictive fit for natural exponential families. *Biometrika* 76, 675–684.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. 2nd Edn.* New York, NY: Springer.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspect. Psychol. Sci.* 7, 645–654. doi: 10.1177/1745691612464056

Jerde, C. L., Kraskura, K., Eliason, E. J., Csik, S. R., Stier, A. C., and Taper, M. L. (2019). Strong evidence for an intraspecific metabolic scaling coefficient near 0.89 in fish. *Front. Physiol.* 10:1166. doi: 10.3389/fphys.2019.01166

Kery, M., and Royle, A. (2016). *Applied Hierarchical Modelling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*. New York, NY: Elsevier.

Lawless, F., and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* 92, 529–542. doi: 10.1093/biomet/92.3.529

Lehmann, E. L. (1995). Neyman's statistical philosophy. *Probabil. Math. Stat.* 15, 29–36.

Lejeune, M., and Faulkenberry, G. D. (1982). A simple predictive density function. *J. Am. Stat. Assoc.* 77, 654–657.

Lele, S. R. (2004). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence*, eds M. L. Taper and S. R. Lele (Chicago, IL: University of Chicago Press), 191–216.

Lele, S. R. (2020). Consequences of lack of parameterization invariance of non-informative Bayesian analysis for wildlife management: Survival of San Joaquin kit fox and declines in amphibian populations. *Front. Ecol. Evol.* 7:501. doi: 10.3389/fevo.2019.00501

Lele, S. R., and Dennis, B. (2009). Bayesian methods for hierarchical models: are ecologists making a Faustian bargain. *Ecol. Appl.* 19, 581–584. doi: 10.1890/08-0549.1

Leonard, T. (1982). Comment on "A simple predictive density function" by M. Lejeune and G. D. Faulkenberry. *J. Am. Stat. Assoc.* 77, 657–658.

Lindsay, B. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conf. Ser. Probabil. Stat.* 5, I-163.

Markatou, M., and Sofikitou, E. M. (2019). Statistical distances and the construction of evidence functions for model adequacy. *Front. Ecol. Evol.* 7:447. doi: 10.3389/fevo.2019.00447

Mathiasen, P. E. (1979). Prediction functions. *Scand. J. Stat.* 6, 1–21.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. A* 236, 333–380.

Ponciano, J. M., and Taper, M. L. (2019). Model projections in model space: a geometric interpretation of the AIC allows estimating the distance between truth and approximating models. *Front. Ecol. Evol.* 7:413. doi: 10.3389/fevo.2019.00413

Prakasa Rao, B. L. S. (1983). *Non-parametric Functional Estimation*. New York, NY: Academic Press.

Ramsey, F., and Schafer, D. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxburry Press.

Robert, C. P. (1994). *The Bayesian Choice*. New York, NY: Springer-Verlag. 436.

Royall, R., and Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *J. R. Stat. Soc. B Stat. Methodol.* 65, 391–404. doi: 10.1111/1467-9868.00392

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Primer*. London: Chapman and Hall.

Royall, R. M. (2000). On the probability of observing misleading evidence. *J. Am. Stat. Assoc.* 95, 760–768. doi: 10.1080/01621459.2000.10474264

Royall, R. M., and Cumberland, W. G. (1985). Conditional coverage properties of finite population confidence interval. *J. Am. Stat. Assoc.* 80, 355–359.

Royle, J. A., and Dorazio, R. M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data From Populations, Metapopulations and Communities*. London, UK: Elsevier.

Schweder, T., and Hjort, N. (2016). *Confidence, Likelihood and Probability*. Cambridge: Cambridge University Press.

Shen, J., Liu, R. L., and Xie, M. (2018). Prediction with confidence? A general framework for predictive inference. *J. Stat. Plan. Infer.* 195, 126–140. doi: 10.1016/j.jspi.2017.09.012

Smith, R. L. (1998). "Bayesian and frequentist approaches to parametric predictive inference," in *Bayesian Statistics*, Vol. 6, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford, UK: Oxford University Press), 589–612.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. B Methodol.* 39, 44–47.

Taper, M. L., Brittan, G. Jr., and Bandyopadhyay, P. S. (2019). *Statistical Inference and the Plethora of Probability Paradigms: A Principled Pluralism*. PhilArchive copy v2. Available online at: https://philarchive.org/archive/TAPSIAv2

Taper, M. L., and Lele, S. R. (2004). "The nature of scientific evidence: a forward looking synthesis," in *The Nature of Scientific Evidence*, eds M. L. Taper and S. R. Lele (Chicago, IL: University of Chicago Press), 525–551.

Taper, M. L., and Lele, S. R. (2011). "Evidence, evidence functions and error probabilities," in *Handbook for Philosophy of Statistics*, Vol. 7, eds M. R. Forster and P. S. Bandopadhyay (Amsterdam: Elsevier Press), 513–532.

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29. doi: 10.1007/s10144-015-0533-y

Teicher, H. (1961). Identifiability of Mixtures. *Ann. Math. Stat.* 32, 244–248.

Tierney, L., and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* 81, 82–86.

Vidoni, P. (1995). A simple predictive density based on the p^* formula. *Biometrika* 82, 855–863.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Stat.* 14, 1261–1295.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fevo.2021.614374

The Evidential Statistics of Genetic Assembly: Bootstrapping a Reference Sequence

Yukihiko Toquenaga[1]* and *Takuya Gagné*[2]†

[1] Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan, [2] Graduate School of Biological Sciences, University Tsukuba, Tsukuba, Japan

The reference sequences play an essential role in genome assembly, like type specimens in taxonomy. Those references are also samples obtained at some time and location with a specific method. How can we evaluate or discriminate uncertainties of the reference itself and assembly methods? Here we bootstrapped 50 random read data sets from a small circular genome of a *Escherichia coli* bacteriophage, phiX174, and tried to reconstruct the reference with 14 free assembly programs. Nine out of 14 assembly programs were capable of circular genome reconstruction. Unicycler correctly reconstructed the reference for 44 out of 50 data sets, but each reconstructed contig of the failed six data sets had minor defects. The other assembly software could reconstruct the reference with minor defects. The defect regions differed among the assembly programs, and the defect locations were far from randomly distributed in the reference genome. All contigs of Trinity included one, but Minia had two perfect copies other than an imperfect reference copy. The centroid of contigs for assembly programs except Unicycler differed from the reference with 75bases at most. Nonmetric multidimensional scaling (NMDS) plots of the centroids indicated that even the reference sequence was located slightly off from the estimated location of the true reference. We propose that the combination of bootstrapping a reference, making consensus contigs as centroids in an edit distance, and NMDS plotting will provide an evidential statistic way of genetic assembly for non-fragmented base sequences.

Keywords: NGS, PhiX174, consensus sequences, bootstrapping, nonmetric multidimensional scaling

Edited by:
Dapeng Wang,
University of Oxford, United Kingdom

Reviewed by:
Xiaofan Zhou,
South China Agricultural University, China
Rayan Chikhi,
Centre National de la Recherche Scientifique (CNRS), France

****Correspondence:***
Yukihiko Toquenaga
toque@biol.tsukuba.ac.jp

†Present address:
Takuya Gagne,
Alphadrive Co., Ltd., Tokyo, Japan

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *06 October 2020*
Accepted: *01 July 2021*
Published: *26 July 2021*

Citation:
Toquenaga Y and Gagné T (2021) The Evidential Statistics of Genetic Assembly: Bootstrapping a Reference Sequence.
Front. Ecol. Evol. 9:614374.
doi: 10.3389/fevo.2021.614374

1. INTRODUCTION

Assume that you got multiple non-fragmented base sequences assembled from data generated with next-generation sequencing (NGS) or more advanced methods. We further assume that we do not have available reference sequences for the material. QUAST (http://bioinf.spbau.ru/quast) and similar tools would recommend choosing longer sequences as plausible ones. But the length of the sequence itself does not guarantee how the sequences resemble or correspond to the correct sequence. Here we propose an evidential statistical method for inferring true sequence by bootstrapping and Nonmetric Multidimensional Scaling (NMDS) plotting with the assembled non-fragmented sequences.

1.1. Background

In evidential statistics, we never seek the true model for a specific data set. Instead, we choose models supported by the given data set (Edwards, 1992; Royall, 1997). The information-theoretic approach also neglects to chase the true model for increasing the prediction ability of selected models (Burnham and Anderson, 1998; Konishi and Kitagawa, 2004; Akaike et al., 2007). Even in Bayesian statistics, the true model is believed to be included in their models with parameter distributions. But the true model for a specific data set still plays an essential role in biology using base sequence data.

DNA or RNA sequencing is rather conservative. It relies on reference sequences often obtained with decades-old sequence techniques (e.g., Sung, 2017). Assume that you get short segments of sequences (called "reads") with NGS methods. Then you have to align them correctly for constructing the whole sequence. Reads are inherently erroneous, and you will have to use several approaches to reconstruct the entire sequence. The reconstructed sequences could be divergent. But those that resemble the reference are promising candidates. Here the reference sequences play the role of type specimens in taxonomical identification (Ballouz et al., 2019).

But it is well known that different assembly programs return different assembly results for a given read data (e.g., Salzberg et al., 2012). Researchers use reference sequences or check annotation of known genes for correcting resultant sequences (Sung, 2017; Sohn and Nam, 2018). High variation in sequence results among assembly programs is mainly caused by applying only heuristic approaches, such as the base-by-base approach, de Bruijn graph, and String graph (Sung, 2017; Sohn and Nam, 2018). Random searching is the core for those approaches, but somehow some assembly programs return the same contig data (both contents and order of sequences) for a given read data set. Others return different contigs for a given data in multiple trials.

For example, the insect mitochondrial genome is a compact circular molecule typically 15–18 kb in size (Cameron, 2014). Recently two genome sequences of mitochondria of *Acanthoselides obtectus* were proposed: one with 16,130 bp (Yao et al., 2017) but the other with 26,613 bp (Sayadi et al., 2017). The latter sequence includes repetitive spacer sequences (**Figure 1**). Usually, repetitive or duplicated sequences are targets to collapse by assemblers. But the researchers published the longer mitochondria claim that they used a new and reliable long-read technique, and hence, the repetitive sequences are real (Sayadi et al., 2017). From now on, we will have to deal with the two references for mtDNA of *A. obtectus*.

What will happen if the reference sequence is not reliable or not exist at all? How can we choose promising ones among divergent reconstructed sequences? Conventionally, researchers believe that the reference sequences are correct. But the references are also samples obtained at some time and location with a specific method. How can we measure the reliability of a reference sequence? Conversely, can we measure the reliance of an assembly method by assembling random reads generated with the reference sequence itself? The random read generation is what we call the bootstrapping of the reference sequence. We propose that bootstrapping can evaluate and characterize assembly programs by using distance measurements that play essential roles in the evidential statistics (Lindsay, 2004).

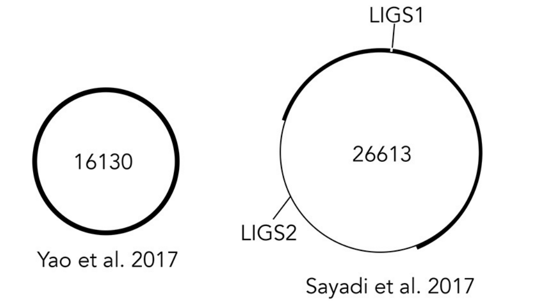


FIGURE 1 | Two schematic mtDNA of *Acanthoscelides obtectus* with different lengths. There are short (1) and long (2) long intergenic spacers (LIGS1 and LIGS2) in the longer one. Other than the spacers, two mtDNAs are identical.

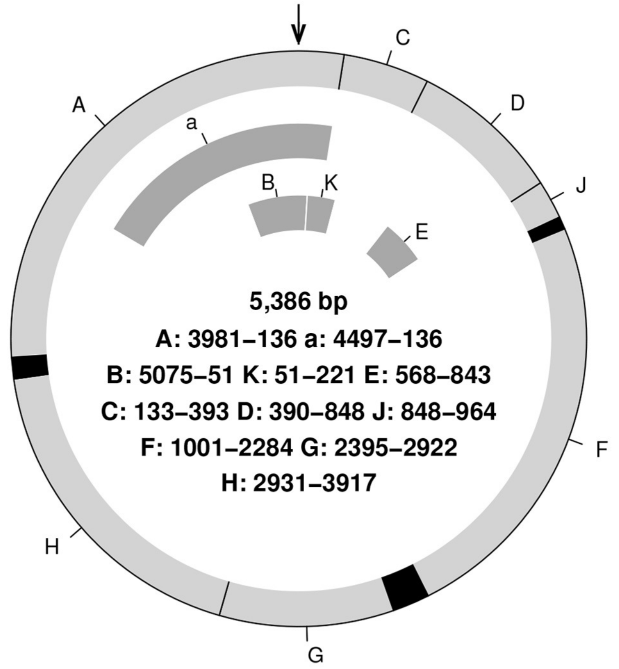

C
A
D
a
B
K
J
E
5,386 bp
A: 3981–136 a: 4497–136
B: 5075–51 K: 51–221 E: 568–843
C: 133–393 D: 390–848 J: 848–964
F: 1001–2284 G: 2395–2922
H: 2931–3917
F
H
G

FIGURE 2 | A schematic genome structure and gene sizes of phiX174. There are 11 genes, and black regions are intergenic spacers. The arrow indicates the origin. Note that there is a 4-base overlap between genes A and C.

1.2. Preliminary Analysis With the phiX174 Reference

We tried to reconstruct the genome sequence of a very simple life, phiX174, an *Escherichia coli* bacteriophage. (**Figure 2**). Several read data sets for phiX174 are available on the internet (e.g., https://github.com/gigascience/galaxy-bgisoap/tree/master/test-data/phiX174). In our preliminary examination with one of the sample read data, 13 free and frequently used assembly programs (all programs in **Table 1** except Unicycler) fairly well reconstructed the reference sequence; defect proportion was

TABLE 1 | List of assembly programs and their characteristics.

Name	Version	v-status	rseed	Method	Circular	No. of copy
A5miseq	20160825	v	Doc	S,HB	Yes	1
ABySS	2.1.0	v	Time	HB	No	–
Fermi	1.1	v	None	S	Yes	1
IDBA	1.1.3	v	Self	HB	Yes	1
Megahit	1.1.3	v	Fixed	HB	Yes	1
Minia	2.0.7	c	Time	HB	Yes	3
Mira	4.9.6	v*	Time	BB	Yes	1
Platanus	2.2.2	c	None	HB	Yes	1
Ray	2.3.1	v	Time	HB	No	–
SOAP*denovo*	2.04.r240	c	cmout	HB	No	–
SPAdes	3.14.0	c	Fixed	HB	Yes	1
Trinity	2.11.0	v*	Fixed	BB,HB	Yes	2
Velvet	1.1	c*	Time	EB	Yes	1
Unicycler	0.4.9b	–	–	HB	Yes	1

*Columns are respectively names of programs, variable statuses of resultant contigs whether they return constant(c)/variable(v) contigs for the same read data set, random seed settings, methods of contig assembly, circular capability, no. of copies of reference, and the proportion of defects region of contigs assembled from 50 randomly generated read data. Random seed statuses are: doc, documented in the manual; time, time as seed; none, nothing specified; self, self-made random generator; cmout, random seed setting commented out; fixed, fixed random seed used. Methods are EB, Eulerian de Bruijn graph; HB, Hamiltonian de Bruijn graph; BB, base by base; S, string graph. *Indicates that SRR7700817 was used for checking contig output variability.*

<5%. But none of them returned the perfect sequence of the reference. Most defects were concentrated at the origin (locus = 0) of the circular phiX174 genome.

We obtained similar results when we reconstructed random reads generated from the reference with a random read simulator, ART (art-illumina Q version) ver. 2.5.8 (Huang et al., 2011). None of the assembly software perfectly reconstructed the reference. Increasing the coverage did not change the results. Again defects were concentrated at the locus zero of the reference. But those results might be caused by an artifact; the random reads were generated, assuming that the reference genome was linear. Both edges inevitably had minimal coverage that caused concentrated defects at the head or tail of the sequence. We should prepare reads generated from the reference assumed to be circular. We also want to exclude all errors specific to NGS methods while generating random read data.

1.3. Flow of This Article

We first show how to obtain hypothetical read data sets for bootstrapping the phiX174 reference sequence in the following sections. Next, we introduce and characterize 14 free assembly programs. Then we explain how to analyze resultant contig sequences. We also introduce the way of constructing consensus sequences from the resulting contigs. The consensus sequences were then plotted in an edit distance space with an NMDS method. For the best-performing assembly program, we tried to reconstruct mtDNAs of *A. obtectus* in Yao et al. (2017) and Sayadi et al. (2017). Results are reported according to the same order. We discuss the possibility of estimating the true reference sequence from the NMDS plots of the consensus and the reference sequences based on the evidential statistics.

2. MATERIALS AND METHODS

2.1. Read Simulators

We surveyed 24 sequence simulators (Alosaimi et al., 2020) and found that only two of them capable of generating reads for circular references. One of them adopts GUI, so we have to use the other one, GemSIM ver. 1.6 (McElroy et al., 2012). But GemSIM cannot specify random seeds. Moreover, GemSIM cannot stop errors specific to sequencers. We had to generate hypothetical random reads without any INDEL and sequence-specific errors. So we made an Illumina read simulator free from any kinds of errors by ourselves with Ruby's programming language (ringreads.rb, ref.rb, and doRingreads.rb).

The simulator accepts a FASTA file of a circular reference DNA sequence. A random read generation started from a location randomly selected within the reference. Then the read was extended to the prespecified read length. Next, a new starting point was located apart from the endpoint of the read with the extent of an insertion length. Then the "paired" read with the same length was generated, but this new read was transformed to its reverse complement. In this way, paired-end reads separated with the given insertion length were generated. This procedure was repeated for the circular genome until the minimal coverage for each locus exceeded the pre-specified coverage value. The Ruby scripts are available at: https://tivoli.ska.life.tsukuba.ac.jp/~toque/to9ue/ringreads.

2.2. Assembly Programs

We collected 14 free assembly programs (see **Table 1**). All programs were installed from source codes or binary distributions. We used Mac OS X 10. 14.6 on a Mac mini (2018) and iMac (Retina 5K, 27-inch, 2020) as our computational environment. We obtained source codes of all software irrespective of the ways of installation. We focused on program performance only at the contig construction level because our target genome of phiX174 was short enough, and each reconstructed sequence was almost always a single contig. We did not have to apply any polishing processes, either. Other than specifying lengths of reads and insertion, we used default parameter settings for each assembly program. For Unicycler, we applied normal model, https://github.com/rrwick/Unicycler.

For all programs except Unicycler, we checked whether each program returned the same contig for the same read data set for multiple trials. For this purpose, we used reads of *Homo sapience* chromosome 3 (30CJCAAXX_4_[12].fq.gz) available at: http://sjackman.ca/abyss-activity. Please consult the link for parameter settings for k-mers and insertion lengths. Mira, Velvet, and Trinity could not handle the chromosome 3 reads, so we used a smaller data set of Human Mitochondrial DNA from Postmortem Brain and Blood (SRR7700817 in SRX4559088) for those three programs.

An assembly program was judged as constant only if both the order and the content of contigs were the same in multiple assembly trials. The program was judged as variable otherwise,

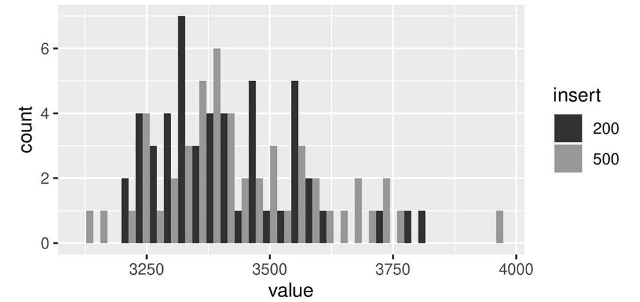


FIGURE 3 | Frequency distributions of the number of random reads for 200 and 500 insertion lengths.

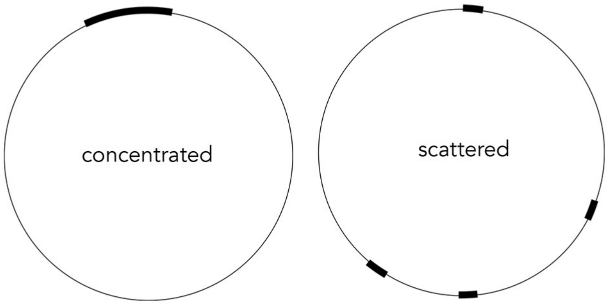


FIGURE 4 | Two hypothetical patterns of defect locations. One is concentrated at a location, but defect locations are scattered in the other.

or it was judged as variable even if only the order of the same contig sets was different. We also checked random seed settings specified in the source codes of the assembly programs. Methods for contig reconstruction were examined for each assembly program based on the description in Sohn and Nam (2018), Sung (2017), software manuals, and publications on the software (Chevreux et al., 1999; Boisvert et al., 2010; Grabherr et al., 2011; Kajitani et al., 2014; Coli et al., 2015; Li et al., 2015). Only 11 out of 14 assembly programs could handle reads generated from the circular reference (**Table 1**). So we applied randomly generated read data only for those 11 assembly programs. Please consult doASSEMBLER_NAME.rb at: https://tivoli.ska.life.tsukuba.ac.jp/~toque/to9ue/ringreads that provide parameters for assembly programs. For Unicycler, we performed assembly with and without polishing with the pilon algorithm using "–no_pilon" option.

2.3. Bootstrapping Reads From the Reference

We generate 50 random sets of reads from the phiX174 reference (accession no. NC_001422), in each of which the data structure was the same as paired-end data of an available read data set (https://github.com/gigascience/galaxy-bgisoap/tree/master/test-data/phiX174); read length = 90, insert length = [200, 500], and the minimum coverage = 20. For each random read data, we tried to assemble contigs with the 11 assembly programs. We aligned the resultant contigs to the reference sequence and examined unique sequences among the 50 contigs for each assembly program. The number of reads for insertion length of 200 ranged from 3,216 to 3,812. Those for insertion length of 500 ranged from 3,132 to 3,964 (**Figure 3**).

2.4. Analyzing Contig Sequences

If a resultant contig is different from the reference and a little longer, we would have two possibilities (**Figure 4**). One is that defective parts are scattered within the contig. The other is that defects are concentrated at a specific region, as in the preliminary experimental results in which we treated the reference as a linear genome. If we apply BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for the known 11 genes of the phiX174, we would not be able to reconstruct several genes in the former case. On the contrary, we could not reconstruct only a couple of genes in the latter case. We performed BLAST searches for the 11 genes for the resultant contigs generated with the 11 assembly software to distinguish the two scenarios. We used the rBLAST library (https://github.com/mhahsler/rBLAST) for it.

The BLAST search of genes, which adopts the Smith-Waterman algorithm, can not correctly determine whether a given contig succeeded in reconstructing the reference because it arbitrarily inserts gaps or deletion for sequence comparison. What we want to do is precisely compare a contig and the reference without any insertion and deletions. To do so, we used the **diffobj** library of R (https://cran.r-project.org/web/packages/diffobj/index.html). Functions of the **diffobj** work just like the diff command of UNIX. We can pinpoint the defective parts among the contig with this functionality. But we should apply the diff operation to two sequences that start at the same origin.

Resultant contigs started from 5" to 3" arbitrarily. So we first have to align all contigs to the direction of the phiX174 reference. Then we have to find the true origin corresponding to the locus zero of the reference because assembly programs returned linear contigs starting from arbitrary origins. To do so, we have to use some distance measurements for comparing two sequences. We applied the Levenshtein edit distance [$LED_{x,y(i,j)}$] defined by three operations: deletion, insertion, and substitution (Equation 1).

$$LED_{x,y(i,j)} = \begin{cases} max(i,j) & if\ min(i,j)=0, \\ min \begin{cases} LED_{x,y}(i-1,j)+1 & (deletion) \\ LED_{x,y}(i,j-1)+1 & (insertion) \\ LED_{x,y}(i-1,j-1)+1_{(x_i \neq y_j)} & (substitution) \end{cases} & otherwise. \end{cases} \quad (1)$$

where $1_{(x_i \neq y_j)}$ is the indicator function equals to zero when $x_i = y_j$ and unity otherwise. $LED_{x,y}(i,j)$ is the distance between the first i characters of x and the first j characters of y. Normalized

Levenshtein edit distance can be obtained by dividing the raw edit distance with $max[length(x), length(y)]$. We used the **stringsim** function provided by the **stringdist** library (https://cran.r-project.org/web/packages/stringdist/index.html) for R ver. 3.6.3. (R Core Team, 2018).

To find the true locus zero in a contig, we chose a tentative origin randomly within a linear contig. We decided on another random starting point if the similarity between the contig and the phiX174 reference was higher than a pre-specified threshold. Otherwise, we chose the next origin by bit-wise walking to the right or left direction. We chose the direction so that the similarity between the contig and the reference increased. We stopped the process and defined the location as the locus zero if we attained the maximum similarity to the reference. We applied **diffobj** functions against the reference and the contig starting from the locus zero.

2.4.1. Consensus Sequence

Consensus sequences that we want to reconstruct differ from those obtained with conventional bioinformatics software, such as DECIPHER for R (Wright, 2016). Conventional applications insert gaps for reconstructing consensus sequences after the alignment of sequences of the same length. But our consensus sequences should have no gaps. As a result, the size of our consensus sequences was indefinite before the reconstruction.

To construct such a consensus sequence of assembly software from 50 contigs of randomly generated read data, we made a program equipped with GPU genetic algorithm (GPU-GA) to search the nearest neighborhood to all 50 contigs within a Levenshtein edit distance space. For GPU-GA calculation, we made an R library named **gpuga** in which We applied OpenCL (https://www.khronos.org/opencl/) for applying to GPU hardware including non-NVIDIA products. Note that you should use R ver. 3.x for running **gpuga**. The **gpuga** package is available at https://tivoli.ska.life.tsukuba.ac.jp/~toque/to9ue/ringreads.

We first listed up the longest common substrings among the 50 contigs. We used those common substrings for masking from INDEL operations during GA calculations. Next, we copied each of the 50 contigs 20 times for constructing the initial population of 1,000 bit-strings, each of which represents its specific contig sequence. We let evolve the bit string population for 100 generations with setting 0.0001 and 0.002 respectively for mutation and crossing over rates per bit. The fitness value is the sum of edit distance from the initial 50 contigs. After 100 generations, we applied bit-shift to the evolved bit-string population and then tried another ten generations of evolution to avoid being trapped in local optima.

After obtaining the consensus sequences of assembly programs, we reconstructed NMDS plots of sequences for examining relative locations against the reference. According to Ponciano and Taper (2019), we can obtain reliable estimates of the generating (true) model by plotting candidate models in a distance space with NMDS methods. A critical difference from Ponciano and Taper (2019) is that we do not have parametric generating functions for reconstructing contigs from the reference, and we cannot apply estimating methods for neg-cross and neg-selfentropies. But if we can assume that $h^2 = 0$ in Equation 9 (Ponciano and Taper, 2019), we can estimate the true reference location as the origin (0,0) in the reconstructed NMDS spaces.

Multiple NMDS plots may be derived from the same data. In our preliminary examination, a general NMDS method applicable for sequence data (**nmds**, Taguchi and Oono, 2005) could not converge to a common spatial configuration. Dr. Mark L. Taper kindly recommended using **mds** function in **smacof** library (https://cran.r-project.org/web/packages/smacof/vignettes/smacof.pdf) for NMDS plottings, based on his experience in Ponciano and Taper (2019), compared to other NMDS functions available in R. Different NMDS functions adopt different stress functions being minimized. We checked that 2D NMDS plots created with **metaMDS** function in **vegan** library (https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/metaMDS), **isoMDS** (https://www.rdocumentation.org/packages/MASS/versions/7.3-51.6/topics/isoMDS), and **sammon** (https://www.rdocumentation.org/packages/MASS/versions/7.3-51.6/topics/sammon) in **MASS** library were rotationally symmetric with that created with the **mds** function of **smacof** library.

2.5. Reconstructing mtDNA of *A. obtectus*

We created random reads from the sequence data of mtDNA of *A. obtectus* for both references respectively proposed in Yao et al. (2017) (accession no. KX825864) and Sayadi et al. (2017) (accession no. MF925724). We adopted the same lengths of reads (90) and insertion (200 or 500) and the coverage (20) as for the phiX174. Then we applied for the best assembly program in the phiX174 trial for reconstructing the mtDNA of *A. obtectus*. We analyzed the resultant contigs in a similar way as those for the phiX174 data sets. But the mtDNA sequences are more than three times longer than that of phiX174, which hindered analyzing methods, such as using R **diffobj** libraries.

3. RESULTS

3.1. Assembly Performance

Most contigs generated with assembly software were longer than the reference sequence. After the BLAST searches, most contigs had a pattern with defects concentrated on a small region. In other words, each contig contained an almost perfect copy of the reference. Analyses with **diffobj** searches confirmed this result, and the proportion of discrepancy equals to the following equation.

$$1 - \frac{L_r}{L_c}$$

where L_r and L_c are lengths of the reference and a contig, respectively.

Contigs generated by assembly methods consist of two groups based on the size: monomer and polymer ones (**Figure 5**). Monomer contigs are those with <6,200 bases. Polymer contigs include those with two- or threefold lengths of the reference.

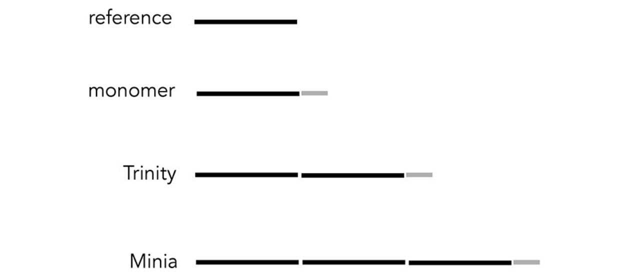


FIGURE 5 | Schematic linear representation of resultant contigs assuming that defect concentrated at the end of gene A. A monomer contig consisted of nearly a whole part of the reference (black) and a region mixing the rest of the reference and defects (gray). Trinity contigs included another complete reference. Minia contigs contained two perfect copies of the reference.

In the following, we explain the state of contigs for each assembly method.

3.1.1. Polymer Contigs

Trinity returned all different contigs against 50 random read data sets, but the size of all of them was 10,795 bases. Each contig contained a perfect and an imperfect reference sequence concatenated in a line. Each imperfect reference had an extra paste margin, and its location was scattered all over among the 11 genes. The distribution was weakly biased to gene H (Chisq = 11.724, df = 6, $P = 0.06841$, **Figure 6**).

Minia returned 24 unique contigs with all of which had 16,189 bases. Each contig consisted of two complete and one incomplete copy of the reference sequence concatenated linearly. The defect parts were scattered among the incomplete reference but heavily biased to gene G compared to the reference (Chisq = 33.595, df = 6, $P = 8.055E\text{-}06$, **Table 2**, **Figure 6**).

3.1.2. Monomer Contigs

A5miseq returned 11 unique contigs ranging from 5,461 to 5,465 bases. Fourteen and 36 contigs, respectively, had defects at genes A and H (Chisq = 56.996, df = 6, $P = 1.831e\text{-}10$, **Figure 6**). On the contrary, fermi returned four contigs for each random read data set. There were 191 unique contigs among them, ranging from 5,454 to 5,473 bases. Only 33 contigs among them passed BLAST searches for the reconstruction of 11 genes. But those 33 contigs had defects at small portions within gene D. Other contigs failed to reconstruct one of the genes of a (A), K, C, and D (not E) in the BLAST searches. One contig failed gene a (A). Fifty-two contigs failed gene K. Ninety-one contigs failed gene C. 15 contigs failed gene D (not E). Because of the high failure rate, we did not conduct **diffobj** analyses for Fermi contigs.

IDBA returned only three unique contigs ranging from 5,417 to 5,427 bases. Defects of each contig concentrated at genes H, A (not a), and E (and D), respectively (Chisq = 166.49, df = 6, $P = 2.2e\text{-}16$, **Figure 6**). Megahit also returned only three unique contigs with the common base length of 5,417. Defects were heavily concentrated at genes G and H (Chisq = 388.67, df = 6, $P = 2.2e\text{-}16$, **Figure 6**). Platanus returned seven unique contigs ranging from 5,430 to 5,436 bases. Defects were heavily concentrated at gene G (Chisq = 405.87, df = 6, $P = 2.2e\text{-}16$, **Figure 6**). SPAdes returned a single unique contig of 5,441 bases with a defect at gene H (Chisq = 203.88, df = 6, $P = 2.2e\text{-}16$, **Figure 6**).

Unicycler returned eight unique contigs ranging from 5,380 to 5,386. Those with 5,386 bases, which is the size of the reference genome, completely reconstructed the reference. Others failed to reconstruct genes F and G (Chisq = 14.691, df = 6, $P = 0.0228$). In summary, 44 out of 50 (88%) contigs were the perfect copy of the reference sequence. Velvet returned 39 unique contigs. All of them were with 5,416 bases and failed to reconstruct narrow regions of gene K overlapped with the end regions of genes A, a, and the beginning of gene C (Chisq = 160.23, df = 6, $P = 2.2e\text{-}16$). The performance of Unicycler did not change a lot when we even stopped the polishing process with the pilon algorithm; 39 out of 50 (78%) contigs were still the perfect copy of the reference.

Mira returned 49 unique contigs with 5,803–6,164 bases. Thirty-four of them completely reconstructed all 11 genes. Discrepancy regions for the 34 contigs were scattered all around the reference genome. Among the rest of 15 contigs, two failed to reconstruct gene A other than the region of gene a. Five could not reconstruct gene a. Six and two could not reconstruct genes F and H, respectively. Because of the high rate of failure, we did not conduct **diffobj** analyses for Mira contigs.

3.2. Consensus Sequence

We reconstructed consensus sequences for those assembly methods that returned almost perfect copy or copies of the reference; we excluded those of Mira and Fermi because those software returned variable contigs with variable defects. **Table 3** shows defect starting locations and similarities against the reference for those consensus contigs. As expected, the consensus sequences resemble each methods' majorities, and their similarity to the reference is more than 0.986. That means the number of defects or extra bases was about 75 at most. There was no specific region for defects; incomplete reconstructions occurred at genes A, C, D, G, and H.

Figure 9 shows the NMDS plot of monomer consensus contigs with the reference. The **mds** function of **smacof** library returned the same plot for multiple trials. As expected, the reference (and Unicycler) location was close to the origin (0,0) in the NMDS plot. Megahit, velvet, and IDBA were moderately apart from the origin. On the contrary, A5miseq, SPAdes, and platanus were fairly apart from the origin. Especially, A5miseq was isolated from the cluster of the other six programs.

3.3. mtDNA of *A. obtectus*

Unicycler, the best performer in the assembly programs, returned a unique contig for each random data set. It reconstructed the same sequence with accession no. KX825864 in 38 out of 50 (76%) data sets. Unicycler returned the sequence with the same length (16,130) with the reference. For the six data sets of the rest, the sequence length was 16,129. For the two of the rest, the sequence length was 16,127. For the last data set, the length was 16,126.

But Unicycler returned a longer and multiple much shorter (a 10th of the longer one at most) contigs for random data

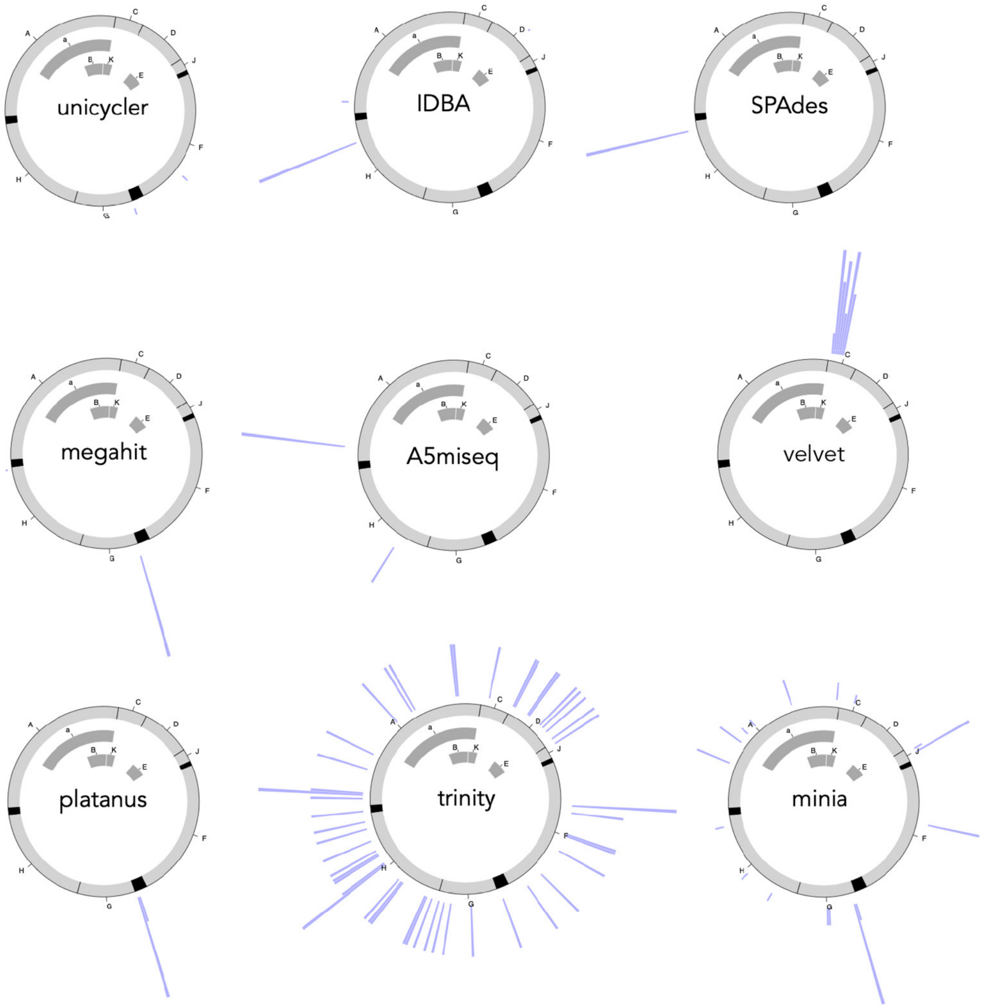


FIGURE 6 | Circular histograms of defect locations on the reference for the nine assembly software.

sets created from the reference of MF925724. The length of the long contig ranged from 16,576 to 17,284 (**Figure 7**), which was enough longer than that of the reference of KX825864. The number of short contigs ranged from one to eight (mode = 5, **Figure 8**). Results were similar even we applied mode = bold to obtain the minimum number of contigs.

4. DISCUSSION

Reconstructed contigs contained only a slightly incomplete reference genome of phiX174. Assembly programs were good at assembling the length of 5.4 ks bases but merely failed to glue the final contigs' edges. For making a ring from the resultant linear contig, software inserted extra bases (the left panel of **Figure 4**). Interestingly, there is not much freedom for the gluing positions for the software that reconstructed monomer contigs. Contrarily, defect locations were scattered among the genome for those generating polymer contigs (**Figure 4**). But the occurrences of the defective regions were not proportional to the size of genes (**Table 2**).

Unicycler almost correctly reconstructed the reference. It is not so surprising because this assembly software is specially developed for a circular genome. Interestingly, its backend software, SPAdes, could not achieve similar performance. It is also interesting that SPAdes returned a unique answer for all data set, but Unicycler returned variable solutions. I am not sure the monomorphic behavior of SPAdes is caused by the fact that its random seed is fixed to 42, an enigmatic number (e.g., Adams, 1980), for random seeds at several places in its source codes. SPAdes, IDBA, megahit A5miseq, velvet, and Platanus returned quite a similar sequence to the reference. Defective locations were highly concentrated: genes H, A, and G for the programs. Velvet specifically had defects at gene C. These defect specificity does not seem to be related to assembly methods nor random seed specification (**Table 1**).

Minia and Trinity returned polymer contigs. Interestingly, the two programs' contigs included two and one complete copy

TABLE 2 | Distribution of defects among the seven gene regions. For reference, gene sizes are indicated.

Method	A	C	D	J	F	G	H	χ^2	P
phiX174	1,539	261	459	117	1,284	528	987	–	–
Minia	6	1	0	3	2	9	3	33.595	8.055E-06
Trinity	12	1	8	0	8	4	16	11.724	0.06841
A5miseq	36	0	0	0	0	0	14	56.996	1.831e-10
IDBA	3	0	1	0	0	0	46	166.49	2.2e-16
Megahit	0	0	0	0	0	49	1	388.67	2.2e-16
Platanus	0	0	0	0	0	50	0	405.87	2.2e-16
Spades	0	0	0	0	0	0	50	203.88	2.2e-16
Unicycler	0	0	0	0	3	3	0	14.691	0.0228
Velvet	30	20	0	0	0	0	0	160.23	2.2e-16

A Chisquare and a P-value are results of Chisquare analysis of the distribution of genes that include defects compared with gene sizes of the reference.

TABLE 3 | The starting location of defects and similarity to the reference for each consensus sequence.

Method	Loci	Similarity	Gene
A5miseq	4,142	0.986447	A
Fermi	–	–	–
IDBA	3,715	0.994277	H
Megahit	2,461	0.994277	G
Minia	13,217 (2,445)	0.998085	G
Mira	–	–	–
Platanus	2,437	0.991349	G
Spades	3,851	0.989892	H
Trinity	527	0.997869	D
Unicycler	–	1.0	–
Velvet	137	0.994461	C

Note that the length of each consensus sequence equals 5,386/similarity.

of the reference with an incomplete one. Those copies were concatenated linearly. The enlargement of the contigs occurred by a quite different mechanism from what happened in the mitochondrial genome of *A. obtectus* (Sayadi et al., 2017; Yao et al., 2017); no regions like the repetitive spacer sequences were detected in contigs for Minia and Trinity (**Figure 5**). The mtDNA structure rather resembles that of monomer contigs, but the defects of the latter were much shorter than LIGSs in the mtDNAs.

Unicycler fairly succeeded in reconstructing the mtDNA of *A. obtectus* for Yao et al. (2017). The success rates dropped only 12% for the reference with the threefold reference. But the longer mtDNA of *A. obtectus* proposed by Sayadi et al. (2017) could not be reconstructed correctly with Unicycler. Resultant contigs could not converge to a single contig. Moreover, the longest contigs were much longer than the reference of Yao et al. (2017). Including LIGS added different features to the reference genome even though the LIGS are entirely separated from the known gene regions. The skewness of **Figure 7** to the shorter contigs (the left direction) may indicate a bias for estimating shorter contigs, also criticized in Sayadi et al. (2017).

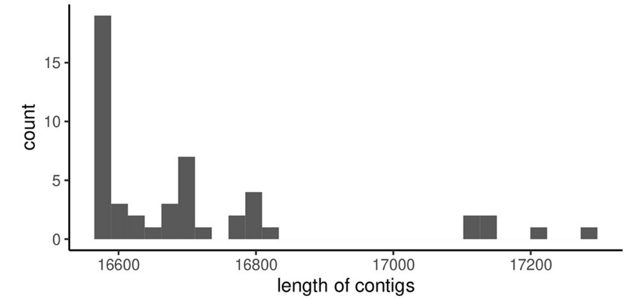


FIGURE 7 | The length of longer contigs generated by Unicycler for random read data generated from MF925724 reference.

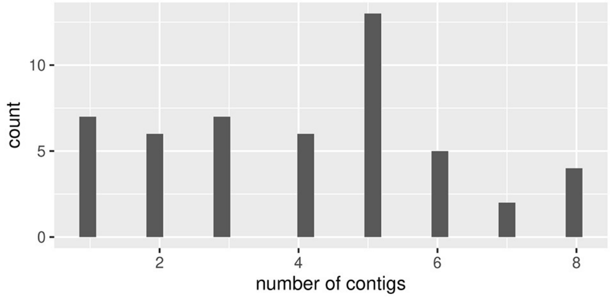


FIGURE 8 | The number of short contigs generated by Unicycler for random read data generated from MF925724 reference.

Evidential statistics uses the evidence functions (e.g., likelihood ratios and consistent information criteria) to quantify the strength of evidence in the data (Royall, 1997; Lele, 2004). If each reconstructed contigs plays the role of a model in a statistical sense, distances among contigs or those from the reference can be a statistical loss function (Lindsay, 2004). If we take the linear or circular DNA/RNA sequences as models, the Levenshtein edit distance can be an appropriate distance measurement among the models and references. The consensus sequence or the centroid in the edit distance for contigs generated by assembly programs is a statistical representation of the specific assembly methods. The consensus sequence of Unicycler coincided with the reference for the short mtDNA of *A. obtectus* (KX825864) as well as phiX174.

Other than Unicycler, reconstructed contigs created by the assembly programs showed a variation in distance from the reference. How can this variation be evaluated? The behavior of consensus sequences may propose two contrasting understandings. One is simply the bias of the assembly programs or incorrectness of the reference itself. This possibility is convincing from the biased defective locations (**Figure 6**). The other interpretation might arise if we take the consensus sequence as a tentative true model. **Figure 9** represents plots for the reference and consensus contigs in the reconstructed NMDS

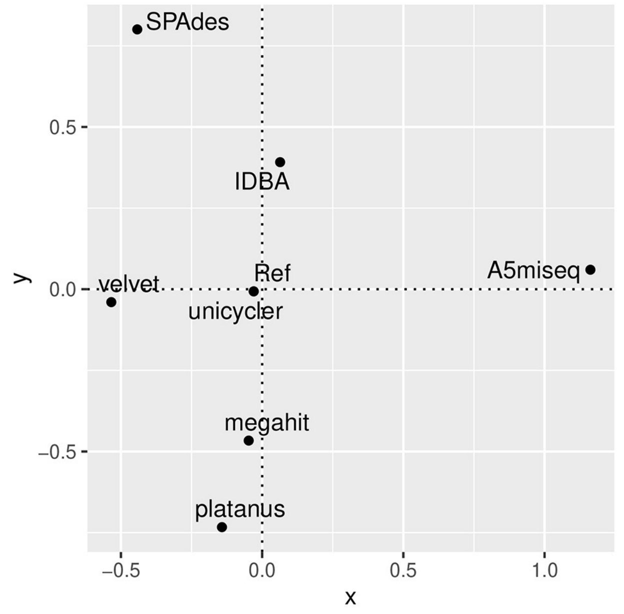


FIGURE 9 | Non-metric multidimensional scaling (NMDS) plot in the edit distance space of the reference and monomer consensus contigs reconstructed with the seven assembly programs using the mds function of smacof library. Note that Unicycler was located at the same place with the reference (Ref).

space. As we expected, the reference and the consensus contig of Unicycler have their position close to the origin (0,0).

These results may suggest a way of finding the true reference sequence when you assembly a novel but none-fragmented genome. Analyzing with multiple assembly programs, you should reconstruct consensus contigs of each assembly program. Then you plot the consensus contigs in a reliable NMDS plot. If there is no consensus contig near the origin, create candidate sequences by taking centroids of those consensus contigs until you obtain a sequence sufficiently close to the origin. But the reference (and that of Unicycler) was not located precisely at the origin in the NMDS plot. Because the reference is another sample from the true model, the origin indicated by the NMDS methods might show the true reference locations. Once we accept the consensus contigs' variation in the bootstrapping of a reference, we will need an alternative representation of the references.

Conventionally, a reference sequence has been represented as a single base string. We should express the uncertainty of the reference for incorporating the variable parts in the reference, as we see in our bootstrapping results. Although some researchers claim a necessity for being aware of the stochastic aspects at each base locus (e.g., O'Rawe et al., 2015), assembly algorithms incorporating stochastic locus have never been proposed. A regular expression using the IUPAC nucleotide code (e.g., Paris and Després, 2012) might be an alternative way to express the stochasticity of sequence loci. The flexibility of each locus is expressed with one of the 15 patterns in the IUPAC code. On the contrary, the flexibility of length is characterized by rules of regular expression. The major drawback of using such regular expressions is the lack of standard measurements of distances against given sequences or regular expressions. It might be a little bit rude to apply Levenshtein edit distance against those regular expressions. Approximate regular expression matching is proposed at most (Belazzougui and Raffinot, 2013).

We proposed an evidential statistics approach consists of three steps; bootstrapping a non-fragmented base sequence, reconstructing consensus sequence from the assembled ones, and plotting the consensus sequences with NMDS. In this new approach, we still obey the one-base-for-one-locus rule. Those consensus sequences are centroids of bootstrapped references and can be taken as approximations of the regular expression with the IUPAC nucleotide code. They might be more evidential for a given read data set than a legacy reference obtained with hopefully reliable but an old sequence method. The proposed method relies on a strong assumption in which we have already got non-fragmented sequences for inferring the true reference. But once you could obtain non-fragmented sequences, and if we can improve the analyzing method using R **diffobj** libraries more efficiently, our method can be applicable for much longer genomes than that of phiX174.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/_001422; https://www.ncbi.nlm.nih.gov/genbank/KX825864; https://www.ncbi.nlm.nih.gov/genbank/MF925724.

AUTHOR CONTRIBUTIONS

YT designed this project and made R and ruby scripts for assembling and analyzing contigs and wrote the manuscript. TG created **gpuga** library and R and python scripts with which he reconstructed consensus sequences for simulated contigs. All authors contributed to the article and approved the submitted version.

FUNDING

This study was partly supported by JSPS KAKENHI Grant Numbers 17H04612 and 18K06410.

ACKNOWLEDGMENTS

We thank Drs. M.L. Taper and J.M. Ponciano for giving us to contribute a paper for this feature on evidential statistics. We thank Dr. T. Tanaka for the valuable discussion on the current status of the assembly of read data sets generated with NGS. We thank Hikaru Kuwabara, who did preliminary experiments assuming the phiX174 reference was linear. We are grateful to members of the TOQUE lab. at Univ. Tsukuba for their debate and helpful ideas at the early stage of our manuscript. Finally, we thank Dr. H. Yasue for providing us the information on assembly failures of some read data sets.

REFERENCES

Adams, D. (1980). *The Ultimate Hitchhiker's Guide to the Galaxy*. New York, NY: Harmony Books.

Akaike, H., Amari, S., Kabashima, Y., Kitagawa, G., and Shimodaira, H. (2007). *Akaike Information Criterion AIC: Modeling, Prediction and Knowledge Discovery (in Japanese)*. Tokyo: Kyoritsu.

Alosaimi, S., Bandiang, A., van Biljon, N., Awany, D., Thami, P. K., Tchamga, M. S., et al. (2020). A broad survey of DNA sequence data simulation tools. *Brief. Func. Genom*. 19, 49–59. doi: 10.1093/bfgp/elz033

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol*. 20, 1–9. doi: 10.1186/s13059-019-1774-4

Belazzougui, D., and Raffinot, M. (2013). Approximate regular expression matching with multi-strings. *J. Discr. Algorith*. 18, 14–21. doi: 10.1016/j.jda.2012.07.008

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comp. Biol*. 17, 1519–1533. doi: 10.1089/cmb.2009.0238

Burnham, K. P., and Anderson, D. R. (1998). *Model Selection and Multimodel Inference: A Practical Infromation-Theoretic Approach*. New York, NY: Springer. doi: 10.1007/978-1-4757-2917-7_3

Cameron, S. L. (2014). Insect mitochondrial genomics: implications for evolution and phylogeny. *Annu. Rev. Entomol*. 59, 95–117. doi: 10.1146/annurev-ento-011613-162007

Chevreux, B., Wetter, T., and Suhai, S. (1999). "Genome sequence assembly using trace signals and additional sequence information," in *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB), Vol. 99* (Leipzig), 45–56.

Coli, D., Jospin, G., and Darling, A. E. (2015). A5-miseq: an updated pipeline to assemble microbial genomes from illumina miseq data. *Bioinfomatics* 31, 587–589. doi: 10.1093/bioinformatics/btu661

Edwards, A. W. F. (1992). *Likelihood: Expanded Edition*. Baltimore, MD: Johns Hopkins Paperbacks.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat Biotechnol*. 29, 644–652. doi: 10.1038/nbt.1883

Huang, W., Li, L., Myers, J., and Marth, G. (2011). Art: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi: 10.1093/bioinformatics/btr708

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*. 24, 1384–1395. doi: 10.1101/gr.170720.113

Konishi, S., and Kitagawa, G. (2004). *Information Criteria (in Japanese)*. Tokyo: Asakura Publisher.

Lele, S. R. (2004). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence*, eds M. L. Taper and S. R. Lele (Chicago, IL: Chicago Press), 191–216. doi: 10.7208/chicago/9780226789583.003.0007

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and com- plex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Lindsay, B. G. (2004). "Statistical distances as loss functions in assessing model adequacy," in *The Nature of Scientific Evidence*, eds M. L. Taper and S. R. Lele (Chicago, IL: Chicago Press), 439–487. doi: 10.7208/chicago/9780226789583.003.0014

McElroy, K. E., Luciani, F., and Thomas, T. (2012). Gemsim: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* 13:74. doi: 10.1186/1471-2164-13-74

O'Rawe, J. A., Ferson, S., and Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends in Genet*. 31, 61–66. doi: 10.1016/j.tig.2014.12.002

Paris, M., and Després, L. (2012). "Data production and analysis in population genomics. Methods in molecular biology," in *In Silico Fingerprinting (ISIF): A User-Friendly In Silico AFLP Program*, eds F. Pompanon and A. Bonin (New York, NY: Springer), 55–64. doi: 10.1007/978-1-61779-870-2_4

Ponciano, J. M., and Taper, M. L. (2019). Model projections in model space: a geometric interpretation of the AIC allows estimating the distance between truth and approximating models. *Front. Ecol. Evol*. 7:413. doi: 10.3389/fevo.2019.00413

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). Gage: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 22, 557–567. doi: 10.1101/gr.131383.111

Sayadi, A., Immonen, E., Tellgren-Roth, C., and Arqvist, G. (2017). The evolution of dark matter in the mitogenome of seed beetles. *Genome Biol. Evol*. 9, 2697–2706. doi: 10.1093/gbe/evx205

Sohn, J.-I., and Nam, J.-W. (2018). The present and future of *de novo* whole-genome assembly. *Brief. Bioinformatics* 19, 23–40. doi: 10.1093/bib/bbw096

Sung, W.-K. (2017). *Algorithms for Next-Generation Sequencing*. Boca Raton, FL: CRC Press. doi: 10.1201/9781315374352

Taguchi, Y., and Oono, Y. (2005). Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics* 21, 730–740. doi: 10.1093/bioinformatics/bti067

Wright, E. (2016). Using decipher v2.0 to analyze big biological sequence data in r. *R J*. 8, 352–359. doi: 10.32614/RJ-2016-025

Yao, J., Yang, H., and Dai, R. (2017). Characterization of the complete mitochondrial genome of Acanthoscelides obtectus (coleoptera: Chrysomelidae: Bruchinae) with phylogenetic analysis. *Genetica* 145, 397–408. doi: 10.1007/s10709-017-9975-9

Conflict of Interest: TG is employed by Alphadrive Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 06 December 2021
doi: 10.3389/fevo.2021.679155

Assessing the Global and Local Uncertainty of Scientific Evidence in the Presence of Model Misspecification

Mark L. Taper[1,2], Subhash R. Lele[3], José M. Ponciano[2], Brian Dennis[4,5] and Christopher L. Jerde[6]*

[1] Department of Ecology, Montana State University, Bozeman, MT, United States, [2] Department of Biology, University of Florida, Gainesville, FL, United States, [3] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada, [4] Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID, United States, [5] Department of Mathematics and Statistical Science, University of Idaho, Moscow, ID, United States, [6] Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA, United States

OPEN ACCESS

Edited by:
Ichiro Ken Shimatani,
Institute of Statistical Mathematics (ISM), Japan

Reviewed by:
Hirohisa Kishino,
The Research Institute of Evolutionary Biology, Japan
Yusaku Ohkubo,
Institute of Statistical Mathematics (ISM), Japan
Keisuke Yano,
Institute of Statistical Mathematics (ISM), Japan

****Correspondence:***
Mark L. Taper
MarkLTaper@gmail.com

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

Received: *11 March 2021*
Accepted: *13 September 2021*
Published: *06 December 2021*

Citation:
Taper ML, Lele SR, Ponciano JM, Dennis B and Jerde CL (2021) Assessing the Global and Local Uncertainty of Scientific Evidence in the Presence of Model Misspecification. Front. Ecol. Evol. 9:679155. doi: 10.3389/fevo.2021.679155

Scientists need to compare the support for models based on observed phenomena. The main goal of the evidential paradigm is to quantify the strength of evidence in the data for a reference model relative to an alternative model. This is done via an evidence function, such as ΔSIC, an estimator of the sample size scaled difference of divergences between the generating mechanism and the competing models. To use evidence, either for decision making or as a guide to the accumulation of knowledge, an understanding of the uncertainty in the evidence is needed. This uncertainty is well characterized by the standard statistical theory of estimation. Unfortunately, the standard theory breaks down if the models are misspecified, as is commonly the case in scientific studies. We develop non-parametric bootstrap methodologies for estimating the sampling distribution of the evidence estimator under model misspecification. This sampling distribution allows us to determine how secure we are in our evidential statement. We characterize this uncertainty in the strength of evidence with two different types of confidence intervals, which we term "global" and "local." We discuss how evidence uncertainty can be used to improve scientific inference and illustrate this with a reanalysis of the model identification problem in a prominent landscape ecology study using structural equations.

Keywords: evidential confidence intervals, unconditional and conditional inference, information criteria, model selection, non-parametric bootstrap, pre- and post-data inference, profile likelihood, reliability

1. INTRODUCTION

> When a person supposes that he knows, and does not know; this appears to be the great source of all the errors of the intellect.
>
> Plato. The Sophist. 360 B.C.E Translated by Benjamin Jowett

One of the main goals of scientific inference is to delineate and understand the underlying mechanism of a phenomenon of interest. In practice, scientists have several different hypotheses or proposed mechanisms, and want to use the observed data to quantify the strength of evidence for one mechanism over the alternatives. The evidential approach to statistical and scientific inference uses estimates of the difference of the divergences from the true mechanism to the competing mechanisms to quantify the strength of evidence in the observed data for one mechanism over the other. The evidence function is an estimator of the sample size scaled divergence difference between two candidate statistical mechanisms (Lele, 2004a). Importantly, evidence functions can be applied pairwise to multiple models to determine the support for multiple alternative mechanisms.

Dennis et al. (2019) demonstrates that evidential inference makes fewer errors than does the Neyman-Pearson hypothesis testing (NPHT) approach at all but the very smallest sample sizes. This is true if the models being compared are "correctly specified." Evidential inference is even more strongly favored if the model set is "misspecified" (definitions of correctly specified and misspecified model sets follow below). Unfortunately, Dennis et al. also shows that the probability of error depends on the nature of model misspecification and can be large. Ponciano and Taper (2019) demonstrates that the entire geometry of the model set and unknown generating process influences inference. Lele (2020a) discusses that uncertainty in science can be usefully expressed in multiple fashions. The current paper is written as a unifying response to these three papers and will be most clear if read in conjunction with them. The goal of the current paper is to introduce empirical measures of evidential uncertainty that are both valid and estimable in the presence of model misspecification.

Various papers (Lele, 2004a; Taper and Lele, 2011; Taper and Ponciano, 2016; Jerde et al., 2019) discuss the desiderata that an evidence function should satisfy. In comparing a reference model to an alternative, the log-likelihood ratio (LLR) is the most commonly used evidence function. An evidence function is usually constructed so that if the realized value of the evidence function, the observed evidence, is larger than a pre-specified positive threshold value (k_{R}), we say that data strongly support the reference model. If it is below a negative threshold value (k_{A}) (i.e., closer to negative infinity), data strongly support the alternative model. If the evidence function is in between these two thresholds, data are said to be unable to distinguish between the two models.

A commonly used alternative to the evidential framework, Neyman-Pearson tests, accords a special statistical status to the null model in that the type I error probability is fixed (does not depend on sample size) and the *p*-value is calculated with only the null model. Consequently, a variety of inferential distortions can famously occur when Neyman-Pearson testing is used for purposes beyond its working specifications (Dennis et al., 2019). By contrast, in the evidential framework no special status is accorded either the reference or alternative model. The designations of reference or alternative serve only to help an analyst understand which model is supported (relative to the other) by positive or negative evidence and do not confer any differences in statistical properties. Royall (1997, 2000) considers the situation where the reference and alternative models are fully specified, that is, there are no parameters with unknown values that need to be estimated from the data. Under the assumption that the reference model is the true generating mechanism, he uses the asymptotic distribution of the LLR to compute the probability of misleading evidence, that is the probability that observed evidence would strongly support the alternative (i.e., wrong) mechanism. He also considers the probability of weak evidence, that is the probability of being unable to distinguish between the two mechanisms. Following the results of Godambe (1960), Lele (2004a) shows that, under regularity conditions, among all evidence functions the LLR is optimal in the sense that the rate at which the probability of strong evidence converges to 1 is the fastest. These error probabilities, especially the probability of weak evidence, are useful for pre-experiment decisions on sample size (Strug et al., 2007) or optimal designing of experiments.

Dennis et al. (2019) recognizes the reality that most models are only approximations and hence the true generating mechanism is likely to be neither the reference nor the alternative model. Following Dennis et al. (2019), we consider a model misspecified if the data distribution it predicts cannot be made to match the distribution of the true generating process by appropriate parameterization. A model set is misspecified if all of its members are misspecified. In practice, the model sets used in science are almost always misspecified to some degree and may be badly misspecified particularly during early exploration of scientific phenomena.

The asymptotic distribution of the LLR under model misspecification (Vuong, 1989; Sayyareh et al., 2011; Dennis et al., 2019) depends on the geometry of the misspecification, that is, how the true generating mechanism and the two competing model spaces relate to each other. In scientific studies, instead of fully specified reference and alternative models, one generally has reference and alternative model spaces, a set of parametric models whose parameters need to be estimated using the observed data. Such a set forms a space because its elements have geometrical relationships such as divergences between them. Dennis et al. (2019) uses the asymptotic distribution of the LLR to compute the error probabilities in comparing model spaces when the true generating model might be outside the specified model spaces. The current paper lists all possible topologies, i.e., configurations, for the generating mechanism and competing model spaces and corresponding asymptotic distributions of the LLR. One important feature of these asymptotic distributions is that the means of these distributions increase toward infinity at rates proportional to sample size, *n,* whereas the standard deviations increase toward infinity at rates proportional to $n^{1/2}$, producing tail probabilities (probabilities of misleading evidence) that converge to zero (because the coefficient of variation goes to zero). Thus, in all evidential comparisons using the LLR, as the sample size increases, probability of strong evidence for the best approximating mechanism converges to 1 and all other error probabilities converge to 0 (Dennis et al., 2019).

As discussed by Royall (1997, 2000, 2004), this behavior of the error probabilities is in stark contrast to the classical Neyman-Pearson approach where the probability of type I error remains constant for all sample sizes. The consequence to the applied scientist is that the true generating mechanism is rejected in favor of a misspecified null some fraction of the time regardless of the amount of data collected. Of course, classical statistical inference does not stop at hypothesis testing. It also computes the sampling distribution of the estimator of the effect size. Unlike the probability of type I error in hypothesis testing, as the sample size increases, the sampling distribution does concentrate around the true effect size, thus leading to the correct inference. Royall (2000) and Dennis et al. (2019) obtain this sampling distribution asymptotically. Several excellent papers (Linhart, 1988; Shimodaira, 1998; Ng and Joe, 2016) construct confidence intervals for evidence under model misspecification using the asymptotic theory of White (1982) and Vuong (1989). Our experience in simulations is that the distribution of evidence does not approach its asymptotic form until sample size is quite large (Jerde et al., 2019; Taper et al., 2019).

Again, the goal of this paper is to obtain a fuller understanding of uncertainty in observed evidence under realistic sample sizes by estimating the finite sample sampling distribution of the strength of evidence under model misspecification via non-parametric bootstrap. In an earlier paper, Taper and Lele (2011) had suggested the use of non-parametric bootstrap to understand finite sample uncertainty in observed evidence when the true generating mechanism may be different than the reference and alternative models. This current paper is a detailed exploration of this suggestion.

The non-parametric bootstrap is a computational approach (Hall, 1986, 1987; Efron and Tibshirani, 1993) used to get a finite sample approximation to the sampling distribution of a statistic that is valid under model misspecification. Generally, the sampling distribution of the estimator is far more useful for supporting scientific arguments than is a hypothesis test by itself (Xie and Singh, 2013; Schweder, 2018).

An inferential statement is any statement about the model parameters, form of the underlying mechanism, or a future outcome. An inferential statement becomes a statistical inferential statement only when a measure of uncertainty is attached to it (Cox, 1958). An accessible review of various approaches to quantifying uncertainty in an inferential statement is available in Lele (2020a). The classical frequentist inference uses aleatory probability (frequency of an event under hypothetical infinite replication of experiment) to quantify uncertainty of an inferential statement. To obtain the aleatory uncertainty of an inferential statement, a critical question that needs to be answered is: which experiment/sampling design do we (hypothetically) repeat? Lele (2020a) uses the simple linear regression model to illustrate the distinction between the global (also known as unconditional, pre-data or, pre-experiment) and local (also known as conditional, post-data or, post-experiment) uncertainty. In this paper, we augment that illustration by comparing the differences between global and local uncertainty in mark-recapture analysis and in structural equations.

Although the unconditional/conditional distinction has been in the theoretical statistics literature since Fisher (1936), the difference has not been well understood by ecologists and scientists in general. To the extent that the difference has been recognized at all it has been common to ascribe unconditional inference to frequentists and conditional inference to Bayesians. However, we agree with Goutis and Casella (1995) that: “In any experiment both pre-data inferences and post-data inferences are important, and each can be made within either frequentist or Bayesian paradigms, which perhaps shows that the frequentist/Bayesian distinction is not as fundamental as the pre-data/post-data distinction.”

In the ecological literature, both kinds of intervals have been used, often without an awareness of the distinction. This is a mistake, because the two kinds of intervals answer different scientific questions. In the discussion, we expand on the interpretation of the two intervals.

Here we consider the evidential approach to model selection under model misspecification. As was described in Dennis et al. (2019), the reference and the alternative models are not fully specified. There are parameters with values that need to be estimated and hence the set-up discussed in Royall (1997) must be altered. Because these two competing models may involve different number of parameters, an unmodified LLR is not an appropriate evidence function, and the LLR needs to be penalized for the number of parameters to be estimated (Akaike, 1973). Furthermore, to make the error probabilities of misleading and weak evidence to converge to 0 as sample size increases, we also need to moderate the penalty by a function of the sample size that grows to infinity at a rate between $\log(\log(n))$ and n (Nishii, 1988). The appropriate evidence functions for the model selection problem are based on the consistent information criteria (IC) such as the Schwarz's Information Criterion (SIC)[1] (Schwarz, 1978) that incorporates both the sample size and the number of parameters in its penalty term. Inconsistent criteria, such as the Akaike Information Criterion (AIC), tend to overfit at all sample sizes and do not lead to valid evidence functions due to the absence of an augmentation of the penalty by the sample size. Note that despite having a sample size correction, the AICc (Hurvich and Tsai, 1989) is not consistent. Its sample size correction is aimed at correcting small sample bias, not large sample inconsistency. We will return to this point in the discussion.

All of the above measures are based on the Kullback-Leibler divergence. However, one can potentially use any divergence measure and with appropriate (i.e., consistent) sample size and parameter number penalty function, one can create a valid evidence function. The evidence function is, as will be made clear later, a *scaled and penalized difference between the estimates of divergences of two models each to the generating process*.

In this paper, we show that model selection based on a bootstrap bias corrected information criterion known as the extended information criterion (EIC) (e.g., Kitagawa and Konishi, 2010) is strongly connected to various bias corrections of the profile likelihood (e.g., Pace and Salvan, 2006). We combine these two ideas with the use of a consistent penalty and show that a non-parametric bootstrap approach can be used to obtain finite sample and consistent estimates of global and local uncertainty in the observed strength of evidence for the reference model vis-à-vis the alternative model. The mathematical details are given in Section 4. As a consequence of this development, we will use as

[1] The SIC is frequently referred to as the BIC or Bayesian Information Criterion. Since we use the criterion as one of a series of criteria, all with frequentist derivations (Nishii, 1988) we use the notation SIC to avoid Bayesian implications.

our evidence function the mean of a bootstrapped distribution of ΔSICs.

Pace and Salvan (2006) and Kitagawa and Konishi (2010) use the bootstrap only for computing the bias correction factor. In contrast, we also use the entire sampling distribution to obtain valid, finite sample, global and local confidence intervals for the strength of evidence. That is, our confidence intervals will also be based on the quantiles of a bootstrapped distribution of ΔSICs.

These confidence intervals are extremely helpful in drawing scientific conclusions (Tukey, 1960). For example, if most of the sampling distribution is above the threshold, we have not only strong evidence, but it is also very unlikely to be strong by chance. We define such evidence as secure. If the sampling distribution is such that a substantial portion is below the threshold, the observed evidence may be strong, but it cannot be considered secure, and more data may be needed to clarify the situation.

Hoping to stimulate practicing scientists with the utility of our approach before they encounter the mathematics of our methods, this paper proceeds as follows: In Section 2, we discuss the implications of uncertainty in evidence and the use of sampling distributions of the strength of evidence in drawing scientific conclusions in detail. In Section 3 we apply these ideas in a reanalysis of a prominent ecological experiment analyzed using structural equations models (SEM) and discuss the scientific implications of the uncertainty in the strength of evidence. Section 4 describes the underlying mathematical concepts and the methodology for computing finite sample, global and local sampling distributions of the strength of evidence for model selection. In Section 5, we validate the methodology using simulations for model selection in linear regression. In Section 6, we discuss implications of the uncertainty quantification of the strength of evidence for the pursuance of science and suggest avenues for further research. Section 7 concludes.

2. SCIENTIFIC INFERENCE UNDER EVIDENTIAL UNCERTAINTY

First, we note that simulations as well as the analytical results in Dennis et al. (2019) show that the sampling variability in evidence can be substantial. Hence using empirical evidence without a measure of uncertainty can be dangerous in practice leading to overconfidence, wrong decisions, misleading inferences, and misguided scientific enquiry. Furthermore, under model misspecification, evidence functions, such as the LLR and others become detached from model-based estimates of error probabilities and are just measures of relative plausibility (Barnard, 1949; Fisher, 1922, 1960; Sprott, 2000). Non-parametric confidence intervals on the strength of inference then allow us to reattach our inferences to probability measures, although there is a considerable difference in what those probabilities mean between global and local inference. Before discussing the methodology to quantify global and local uncertainties in evidence and their real-world applications, let us first discuss how the sampling distribution of the strength of evidence could be used to draw scientific conclusions.

Royall (1997) considers three categories of strength of evidence: Strong evidence for a reference model, strong evidence for the alternative model, and weak evidence when the strength of evidence cannot distinguish between the two models. Often in ecological analysis, one finds the strength of evidence that is neither so weak that one feels comfortable saying one cannot distinguish between the models nor so strong that one is willing to stake a reputation on it. Hence, we suggest using five categories for strength of evidence, inserting categories of prognostic evidence for the reference model and prognostic evidence for the alternative. See **Box 1** for a more complete discussion.

One final difference between Royall's characterization of the strength of evidence and our characterization is that Royall considered the strength of evidence a ratio of likelihoods. We, on the other hand always consider strength of evidence as differences on a logarithmic scale (see discussion in Barnard, 1949). This ties our conceptualization more closely with information theory and the comparison of divergences.

This seemingly small difference marks large differences between our current understanding and that expressed in Royall (1997). We differ from Royall primarily in two intertwined but distinct issues. The first is the utility and scope of the "likelihood principle" (LP). And the second is the usefulness of measures of "pre-data" and "post-data" uncertainty.

Royall's (1997) evidence is developed axiomatically from the "likelihood principle" (Birnbaum, 1962). We do not deny the likelihood principle within the context it was originally stated: "We deliberately delimit and idealize the present discussion by considering only models whose adequacy is postulated and is not in question" (Birnbaum, 1962). Unfortunately, this means that the likelihood principle and everything that follows from it is silent on what happens if models are at all misspecified. We agree with Sprott (2000, p. 105) that "Since few scientists would claim that the model and surrounding assumptions are exactly correct, particularly in the latter situation, the domain of scientific application of LP seems extremely narrow."

We develop evidence as the difference of estimates of the distance of a modeled distribution to the generating process's distribution. This definition is compatible with model misspecification. Further, as we have previously demonstrated (Lele, 2004a; Taper, 2004; Dennis et al., 2019 in this research topic), under correct model specification, along with both models being simple hypotheses (i.e., no parameters with unknown values), this definition is compatible with the Royall's likelihood ratio definition of evidence, if one uses the Kullback-Leibler divergence as a distance measure. We also suggest and use distances that are different from KL distance. That negates the likelihood principle in its purest form. For example, design seems to play a role (Lele, 2004a; and our discussion in Section 6).

Royall's commitment to the likelihood principle entails a stance supporting the irrelevance of uncertainty estimates of evidence based on sample space probabilities, such as pre- and post-data error probabilities. Nevertheless, Royall sets great stock by his argument that you don't need to worry about the probability of misleading evidence post data, because it will always be small if the LR evidence is large. Royall's argument falls short when there are parameters with values to be estimated and/or when there is model misspecification. We have previously argued (Taper and Lele, 2011; Dennis et al., 2019) that pre- and post-data measures of uncertainty are useful for scientists

BOX 1 | Categories of strength of evidence.

Often in ecological analysis, one finds evidence that is neither so weak that one feels comfortable saying one cannot distinguish between the models at all nor so strong that one is willing to stake a reputation on it. Thus, to the thresholds k_A and k_R we add the thresholds k_a and k_r. Evidence between the thresholds k_A and k_a and between k_r and k_R could reasonably be called moderate, but to avoid a clash in abbreviations with the error category of misleading evidence, we will call such evidence prognostic. Now evidence is divided into five categories: strong evidence for the alternative model, prognostic evidence for the alternative model, evidence so weak that it is best to say that neither model is favored, prognostic evidence for the reference model, and strong evidence for the reference model.

(1) Strong evidence for the reference model if the strength of evidence is larger than k_R.
(2) Prognostic evidence for the reference model if the strength of evidence is between k_r and k_R.
(3) Weak evidence favoring neither model if the strength of evidence is between k_a and k_r.
(4) Prognostic evidence for the alternative model if the strength of evidence is between k_A and k_a.
(5) Strong evidence for the alternative model if the strength of evidence is less than k_A.

Royall (1997) pointed out that on occasion, one can have strong evidence that one model, say the reference, in your comparison is closer to the generating process than the other, say the alternative, when in fact it is the alternative that is truly closer to the generating process. Royall called such counterfactual evidence "misleading." With the weaker category of prognostic evidence, it is even more likely that evidence that is counterfactual will be estimated. We designate counterfactual prognostic evidence as "confusing evidence." With real data, one does not know if strong evidence is in fact misleading, or if prognostic evidence is confusing. However, in design and validation studies, whether analytic or computational, the researcher does know when evidence is misleading or confusing, and these categories are very helpful (see Section 5).

It is important to realize that the sign of evidence only indicates which model is estimated to be closer to the generating process, positive for the reference model and negative for the alternative. Previously in the literature, k_A has been set symmetrically to $-k_R$. In specific cases, there could be reason for asymmetry in thresholds, either because of asymmetry in probability models or because of decision cost. For simplicity, we adopt symmetric thresholds with $-k_p$ and k_p indicating the thresholds between weak evidence and prognostic evidence for the alternative and reference models respectively. Similarly, $-k_S$ and k_S are the thresholds between prognostic evidence and strong evidence for the alternative and reference models. The boundaries for our categories then become: strong evidence for the alternative = $-k_S$, prognostic evidence for the alternative = $-k_p$, prognostic evidence for the reference = k_p, and strong evidence for the reference = k_S. Jerde et al. (2019) discuss interpretations for levels of evidence. Following their recommendations, we define $k_p \equiv 4$ and $k_S \equiv 7$.

While we have introduced thresholds, it is important to realize that these are not the absolute accept/reject thresholds of NPHT. They create descriptive categories to help us think, like the names of colors. Light with a wavelength of 521 nm is called a green while that with a wavelength of 519 is called a cyan, but the difference is slight. These thresholds should be thought of "as more what you call guidelines, than actual rules"[2] (Bruckheimer and Verbinski, 2003).

We note finally that Dennis et al. (2019) used a reversed direction for the evidence scale, in order to compare more clearly evidence analysis with Neyman-Pearson hypothesis testing. Dennis et al. posed a correspondence between the reference model in evidence analysis and a NPHT null hypothesis, along with a correspondence between the alternative models, to study error properties of the two analysis approaches. It was convenient to define evidence strength for the alternative to increase as the evidence function moved in the positive direction (by simply reversing the difference of SICs) instead of the negative direction. This defined evidence for the alternative model to be in concordance with the direction favoring the alternative hypothesis in NPHT according to the generalized likelihood ratio statistic (G^2), allowing easy study of errors with the well-known asymptotic distributions of G^2. Either direction for evidence favoring the alternative model can be used provided one stays consistent within an application. In the present paper, it is convenient to adopt the convention described earlier in this box, because errors will be estimated by bootstrapping rather than by asymptotic distributions of G^2.

to think about. Even in the correct specification case where the (post-data) probability of misleading evidence is bounded by 1/LR, other uncertainty measures are useful for study planning and probing the extent of the results. In the more usual case of model misspecification, estimation of the probability of misleading evidence is not simply a matter of transforming the evidence. We have shown (Dennis et al., 2019) that it also depends on the geometry of the model set and the generating process. Importantly, the probability of misleading evidence is not guaranteed to be small—it can be as large as 0.5. Thus, measures of the uncertainty of evidence are a critical complement to an estimate of evidence. Further, to be useful, such measures *must* be estimable in the presence of model misspecification. In this work, we show that non-parametric bootstrap greatly expands the options, capabilities and the nature of the inferential problem under which estimating these measures is possible.

We are not alone in our insistence on a measure of uncertainty in evidence. Alan Birnbaum, after being an early advocate of Hacking's (Hacking, 1965) LR formulation of statistical evidence, strongly repudiated it in Birnbaum (1970, 1972) on the grounds of its lack of confidence measures.

> *If there has been 'one rock in a shifting scene' or general statistical thinking and practice in recent decades, it has not been the likelihood concept, as Edwards suggests, but rather the concept by which confidence limits and hypothesis tests are usually interpreted, which we may call the confidence concept of statistical evidence. This concept is not part of the Neyman-Pearson theory of tests and confidence region estimation, which denies any role to concepts of statistical evidence, as Neyman consistently insists. The confidence concept takes from the Neyman-Pearson approach techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data. (The absence of a comparable property in the likelihood and Bayesian approaches is widely regarded as a decisive inadequacy.)*
>
> *Birnbaum (1970)*

We believe that the current paper rehabilitates statistical evidence by coupling it with an estimate of confidence.

2.1. Understanding Global and Local Uncertainty in Evidence

Confidence intervals are a mainstay in ecological inference, increasingly and justifiably so (Johnson, 1999; Ponciano et al., 2009; Halsey, 2019; Holland, 2019; Fieberg et al., 2020). They transmit a more complete and interpretable representation of the information in data than do hypothesis tests. A confidence

[2] As voiced by the character Hector Barbossa https://www.youtube.com/watch?v=6GMkuPiIZ2k&ab_channel=JesseDB

interval is a range of values for a statistic, a function of the data, that is expected to cover (capture, include) an estimation target a given per cent of the time (e.g., 95%) under repetition of a specified hypothetical experiment (Neyman, 1937). The target of an interval is something in nature about which we would like to make an inference such as a population parameter or a function of a parameter.

For evidence, there are both local and global intervals that can be calculated (see Section 4 for details). In order to understand confidence intervals for evidence, it is important to realize that not only are the interval widths different, but that the targets are also different.

The global target is the difference between the divergences of the best possible representations of the two models to the natural generating process. The uncertainty in the global interval includes the sampling uncertainty for the data, model estimation uncertainty given the data, and uncertainty due to model set misspecification.

The local target is the evidence *in the observed data* for the best possible representation of one model over the best possible representation of the other model. The uncertainty in the local interval represents just the model estimation uncertainties given the observed data, and uncertainty due to model set misspecification.

Global intervals reflect the variation in the estimates if independent experiments are conducted in a manner like the original experiment. The local intervals reflect the informativeness of the specific experimental outcome in hand.

The local interval can capitalize on lucky samples to make precise inferences about the strength of evidence for the reference model relative to the alternative model. On the other hand, with unlucky samples where the parameter estimate may be far from the truth, the local intervals also end up making precise but misleading inferential statements. Global intervals, because they average over all possible datasets, tend to be wider than the local intervals. They are conservative in their uncertainty quantification, making strong inferential statements only cautiously. That does not mean that the global intervals are without use. Scientific results need to be validated by independent replication. A global interval indicates how discrepant the results of a repetition of the experiments could be from the original before contradicting your results and hence protects against the possibility of being contradicted. A worked example of global and local intervals in a mark recapture analysis can be found in **Box 2**.

2.2. Interpreting Evidential Uncertainty

Generally, desirable properties in confidence intervals are proper coverage and given proper coverage, shortness of length (Casella and Berger, 2002). A confidence interval can either cover the target or it can miss it. If the interval fails to cover the target, it can either be entirely above the target (miss high) or entirely below it (miss low) (see **Figure 1**). It is often, but not always, considered desirable if intervals that miss the target value are distributed equally above and below it. Evidence is one of the cases where an equal distribution of non-coverage is undesirable. In this context missing high is superior to missing low. Both types of intervals misrepresent the confidence one should have in the evidence, but the high miss is at least always indicating a correct assessment while a low miss could be supporting an incorrect assessment. Of course, this is assuming the expected evidence is positive, as in **Figure 1**, if the expected evidence were negative, the desirability of missing high and low would be reversed. Really, we mean that it is better for the interval to miss its target distally from 0 than to miss proximally to 0. However, in this simulation study the evidential comparisons are arranged so the reference model is always the better model as to keep the language of missing high and low less confusing.

The categories of evidence introduced in **Box 1** suggest useful ways to apply confidence intervals for strength of evidence to scientific inference. Scientifically, the paramount question is: is the evidence veridical (i.e., in agreement with fact) or is it misleading? The intervals we propose estimating can give us confidence in our answer. We propose that if the proximal bound of this confidence interval is distal to k_S that it be considered “very secure.” If the proximal bound falls between k_S and k_p then the evidence should be considered “secure.” Finally, if the proximal bound is proximal to k_p or the interval overlaps 0 the evidence is “insecure.”

These three levels of strength of evidence and two levels of security of evidence create six heuristic categories:

1. Strong and very secure (SV): The point estimate of evidence (e.g., ΔSIC) is strong and the lower bound of uncertainty indicates that we have confidence that the target (true evidence) is also strong.
2. Strong and secure (SS): The point estimate of evidence is strong, and we are confident that the true target is at least prognostic. There is very little chance that this evidence is misleading.
3. Strong but insecure (SI): The point estimate of evidence is strong, but we cannot be confident that the target is not weak.
4. Prognostic and secure (PS): The point estimate of evidence is prognostic, and we can be confident that the target is at least prognostic.
5. Prognostic but insecure (PI): The point estimate of evidence is prognostic, but we are not confident that the target is not weak.
6. Weak and insecure (WI): The point estimate of evidence is weak and thus by definition, we are not confident that the target is not weak.

As sample size increases, a majority of the sampling distribution lies above the strong evidence threshold and the probability of obtaining evidence that is not SS diminishes to 0 (Dennis et al., 2019). There is, of course, the pathological case where two models are equally divergent from the true generating process. Were this curiosity ever to occur, then each model would be strongly and securely selected with probability 0.5. It is arguable that, even in such a situation, no error has occurred, as in each case a model closest to the generating process has been selected. Substantial discussion on interpreting statistical evidence when augmented with confidence intervals is given in **Box 3**.

BOX 2 | Global and local intervals in mark/recapture analysis.
In ecology, where uncertainty in the study systems is ubiquitous, it is common practice to formulate a scientific hypothesis in the form of a simplified probabilistic model of how the data arose. This simplification allows the analysts to focus the inferential process on a typically small set of quantities bearing strong ecological or management importance. Such simplifications are in fact conceptual restrictions on how the data arose and are used to formulate the likelihood function. Multiple uncertainty simplifications/restrictions are incorporated in the form of multiple conditioning layers. Take for instance a simple closed population mark-recapture experiment where in a first visit to a study area, a number of animals of the species of interest are marked and released. In a second visit, a sample of animals from the same population are captured and the number of previously marked animals in that sample recorded. Under that setting, different levels of conditioning restrict more and more the sampling uncertainty while keeping the focus on the same inferential quantity of interest—the total population size. We prefer the terms "global' and "local" because they evoke the scope of inference that can be addressed by each type of uncertainty. The sampling distribution for global uncertainty is computed using the entire sample space whereas "local" uncertainty is computed using a relevant subset of the sample space (Buehler, 1959).

The key question in global and local inference is what components of your data do you want to be considered fixed (or given) and what components do you want to be considered random (or representative). A completely unconstrained interval is considered global. Intervals with constraints are considered local. An alternative way of approaching this question, which may be clearer for some, is to recognize that a confidence interval represents the variability in hypothetically repeated experiments. When you treat a component as fixed or random, you are specifying different hypothetical experiments. One of the goals of confidence intervals is to define what estimates a skeptic who tries to replicate the experiment might obtain. Different types of experimental conditions that the skeptic might use dictate the choice of the interval.

We illustrate the concepts of global and local inference using the familiar problem of population size estimation using the Lincoln-Peterson estimator. We use the data from a published experiment on iguana population density to create a realistic framework along with some R commands to demark the global and local differences clearly in the calculations. The data and a more complete treatment can be found in Powell and Gale (2015).

Below is a mark-recapture data set, describing one re-sampling occasion. On day 3 of their experiment 131 individuals, *n*, are captured and 116, *x*, of these have previously been marked. Initially (days 0, 1 and 2) *m* = 221 individuals have been captured, marked and released:

```
m <- 221
n <- 131
x <- 116
```

From these data we estimate a total population size using the Lincoln-Petersen estimator. Thus, the target for point and interval estimation is the true population size. As it happens, the same estimator is obtained whether you assume that: (1) Both *m* and *n* are fixed. (2) *m* is considered fixed, but *n* is not. And (3) Both *m* and *n* are considered random.

While the estimate of the total population for these three cases is identical, the uncertainty around it is not. Each set of assumptions fully determines the confidence intervals. We demonstrate this via parametric bootstrap (PB) because of how the levels of randomness enter at each stage is much more perspicuous in the PB code than in the corresponding analytic formulae.

Parametric Bootstrap
Compute the Lincoln-Petersen estimator for the sample at hand as well as the nuisance parameter phi.hat (the capture probability)

```
t.hat <- floor((n*m)/x)
print(t.hat)
```

```
## [1] 249
```

```
phi.hat <- n/t.hat # estimated capture probability
print(phi.hat)
```

```
## [1] 0.5261044
```

Now let's set our PB simulation parameters to these two estimates:

```
t.true <- t.hat
phi.true <- phi.hat
```

Next, set the total number of simulations

```
B <- 10000
```

and then create empty arrays to store the three types of estimates

```
# Lincoln Petersen constrained on m and n (ultimate local: fixed m and n)
LP.mn.bt <-  rep(NA,B)

#Lincoln Petersen constrained on m (local-fixed- m, but global-random- n)
LP.m.bt <-  rep(NA,B)

#Lincoln Petersen unconstrained (Global m and global n i.e. both are random)
LP.bt <-  rep(NA,B)
```

Finally, just turn the crank on the PB iterations and store them:

```
for (i in 1:B){

 #### Simulating data and computing t.hat under the first assumption:
 X.mn <- rhyper(nn=1, m=m,n=(t.true-m),k=n) #constrained on m and n
 LP.mn.bt[i] <- m*n/X.mn

 #### Simulating data and computing t.hat under the second assumption
 N <- rbinom(n=1,size=t.true,prob=phi.true) # unconstrained
 X.m <- rbinom(n=1, size=min(m,N), prob=m/t.true)  #constrained on m but not n
 LP.m.bt[i] <-  m*N/X.m

 #### Simulating data and computing t.hat under the third assumption
```

(Continued)

BOX 2 | (Continued)

```
  M <- rbinom(n=1,size=t.true,prob=phi.true) # unconstrained
  X <- rbinom(n=1, size=min(M,N), prob=M/t.true)    # not constrained on either m or n
  LP.bt[i]   <-  M*N/X
}


# Throw out the outcomes for which x=0. A result of x=0 is possible, but gives
# an infinite estimate of population size.
LP.mn.bt <- LP.mn.bt[is.finite(LP.mn.bt)]
LP.m.bt <- LP.m.bt[is.finite(LP.m.bt)]
LP.bt <- LP.bt[is.finite(LP.bt)]
```

It is instructive to look at the sample spaces for these three estimators:

$$\begin{aligned}
&\text{Sample Spaces :}\\
&\text{LP.bt} : \Omega_G = \{M \in \{0, \cdots, T\}, N \in \{0, \cdots, T\}, X \in \{\max(0, N - (T - M)), \cdots, \min(M, N)\}\}\\
&\text{LP.m.bt} : \Omega_{L_1} = \{m, N \in \{0, \cdots, T\}, X \in \{\max(0, N - (T - m)), \cdots, \min(m, N)\}\}\\
&\text{LP.mn.bt} : \Omega_{L_2} = \{m, n, X \in \{\max(0, n - (T - m)), \cdots, \min(m, n)\}\},\\
&\text{where } T \text{ is the true population size.}
\end{aligned}$$

The sample spaces are all possible data sets that the simulations could generate under each of the model assumptions. The sample space for LP.m.bt is nested within that of LP.mn.bt, which is itself nested within the sample space of LP.bt. Clearly, global and local are relative terms. LP.m.bt is local with respect to LP.bt, but global with respect to LP.mn.bt.

The sampling distributions for the three estimators are plotted in the figure below. We now have three different confidence intervals. Which is right? Statistics by itself cannot answer that question. These three intervals represent the uncertainty in the hypothetical repetition of three different experiments. In the type 1 experiment, with *m* and *n* constrained, the only thing that can vary experiment to experiment is the number of marked animals in the final day sample.

In type 2, the number of previously marked individuals is constrained but not the final day sample size. The hypothetical experiment is repeated only for the final day; varying numbers of individuals as well as varying numbers of marked animals may be captured on the final day. In type 3, the entire hypothetical experiment is repeated. The number of marked individuals, the number of captured individuals, and the number of marked individuals in the second sample may all vary.

The appropriate interval depends on the kind of uncertainty you are trying to represent. The first interval answers the question: How different the estimators of the total population could be if someone else replicated the experiment such that the total number of marked individuals and total number of captures are identical to your experiment? This can happen in a field survey where the total number of marked animals and total number of captures is fixed by design, *a priori*. These numbers may depend on the budget the researcher might have for capturing animals for marking and for recapturing.

In some situations, such as camera trap surveys, the total number of marked animals may be fixed by design but the total number of captures, by the nature of the survey technique, is random. The second interval considers this possibility and allows for the randomness in the number of captures to compute the uncertainty in the total population size estimator. In the case of fish surveys, the number of fish caught in the traps or by electrofishing for marking is necessarily random and so is the number of fish in the sample afterwards. In this case, the third interval will be appropriate.

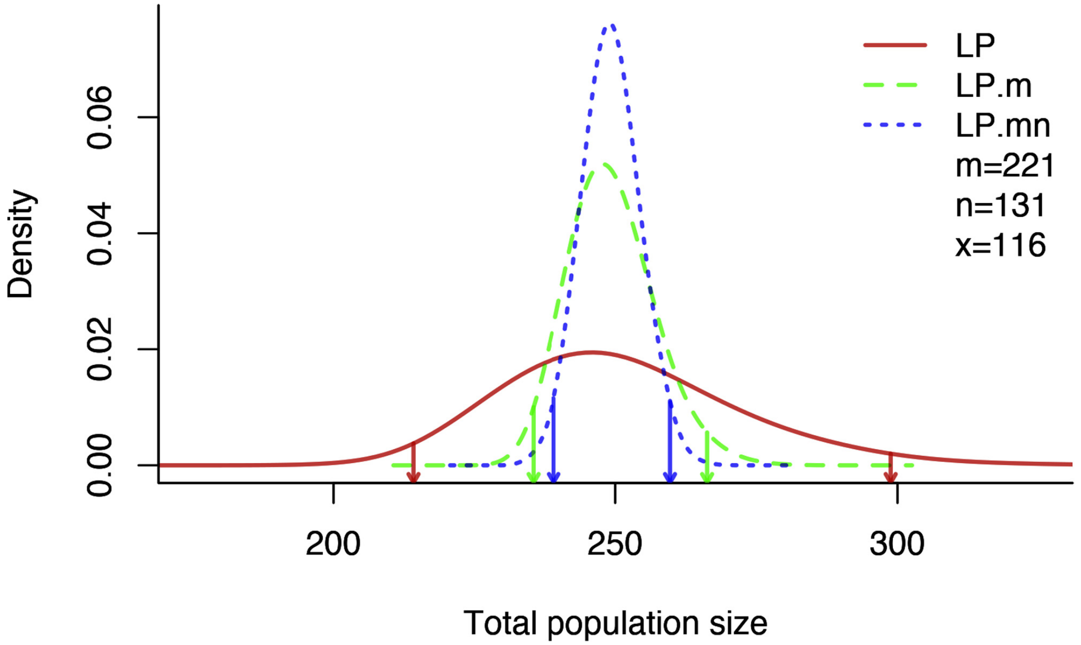


Figure Box 2.1 | Sampling distributions and 95% confidence intervals of total population size estimates for three levels of conditioning in Lincoln Peterson estimates. The ML estimate for all three models is 249. The confidence 2.5 and 97.5% limits are indicated by the vertical lines dropped from each curve to the x-axis. The intervals become increasingly shorter as the models (hypothetical experiments) become more constrained. Here, as is generally but not universally the case, the intervals are completely nested.

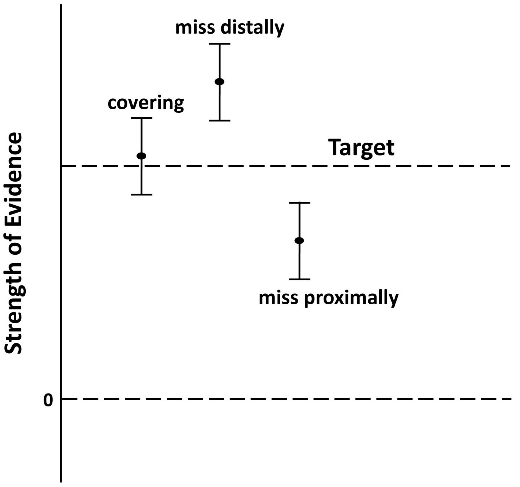


FIGURE 1 | Hypothetical coverage of confidence intervals for evidence. The strength of evidence is the value of an evidence function relating two models and a data set. Typical evidence functions are LLR or the difference of information criterion values, ΔICs. In our worked example (Section 3) we use the Schwarz information criterion. ΔSIC_{RA} values greater than 0 indicate support in the data for the reference model relative to the alternative. These values are indicated by dots in the figure. The vertical bars indicate confidence intervals for the strength of evidence. The target for a confidence interval on the strength of evidence is a penalized scaled divergence difference (see Section 4.1), loosely this is the expected evidence. By design, a perfect confidence interval, at say the 95% confidence level, will fail to cover its target 5% of the time. If a confidence interval that misses its target is entirely more distant from 0 than is its target, we say that it misses distally, otherwise we say that it misses proximally. We will also speak of the bound of a confidence interval for evidence that is closest to 0 as the proximal bound.

3. EXAMPLE: UNCERTAINTY IN A STRUCTURAL EQUATIONS MODELS ANALYSIS OF POST-FIRE RECOVERY OF PLANT DIVERSITY

To probe the effectiveness of bootstrapping evidence in realistically complex problems, we revisit the classic analysis of Grace and Keeley (2006). These authors used structural equation modeling to study the impact of landscape, environment, and community factors on the recovery after fire of shrubland plant diversity.

A recent article on developing causal models (Grace and Irvine, 2020) revisits the 2006 study and takes a more moderate stance than the original paper: "Subsequent SEM studies (Keeley et al., 2008) have enhanced our confidence in the general inferences drawn from the original study. That said, we would not claim that all our parameter values are unbiased causal estimates without further evidence to support such inferences." We believe that had Grace and Keeley had the tools for estimating the two kinds of evidential uncertainties we have developed here a much more nuanced understanding could have been gained—even from the original data—as to which paths were likely to be supported by future work and which were potentially non-replicable.

3.1 Example Choice

There are reasons why SEM is growing in influence in environmental informatics, ecology and evolution. First, SEM allows for legitimate causal inference in situations both in observational studies (Grace, 2008; Bollen and Pearl, 2013; Grace and Irvine, 2020) and where experimental manipulation has been performed (Grace et al., 2009; Breitsohl, 2019). In fact, path analysis, the precursor to SEM, was first developed by Sewall Wright (1934) to expose causal effects to statistical inference. Second, because it is designed for estimation of a network of causal effects, SEM is well suited for analyses of the complex patterns of influence often found in environmental science, ecology and evolution (e.g., Grace and Pugesek, 1997). Third, SEM recognizes that many observables may be recorded with measurement error (Bollen, 1989). The ability to incorporate measurement error in an analysis eliminates an important source of bias that has plagued environmental science, ecology and evolution (Taper and Marquet, 1996; Cheng and Van Ness, 1999). Implicit in the incorporation of measurement error is the ability to consider latent variables (i.e., unobserved, and potentially unobservable variables) (Grace and Bollen, 2008; Grace et al., 2010). Fourth, causal paths and latent variables allow linking scientific theory and statistical analysis in a particularly perspicuous fashion (Grace and Bollen, 2008; Grace et al., 2010; Laughlin and Grace, 2019). Because of these beneficial features, SEM is being utilized in growing number of applications in environmental informatics, ecology and evolution. The explosive growth of SEM in ecology is documented in Laughlin and Grace (2019).

Despite its many advantages for scientific thinking, SEM does present some inferential difficulties (Tomarken and Waller, 2003). Information can flow between variables by multiple pathways. As a consequence, the fit of alternative models and therefore the evidence between them can vary considerably with small changes in the configurations of the data. This uncertainty in evidence needs to be quantified.

A final reason for the choice of the Grace and Keeley example is the excellence of the original study. The observations were collected under the direction of Jon Keeley, while the analysis was conducted by James Grace. Jon Keeley is a very experienced empirical ecologist, while Grace has been a leading proponent the application of SEM to ecological systems. Both are scientists of great distinction. We do not seek to cavil at pedestrian research but look to see what bootstrapping of evidence can add to a well done scientific analysis.

3.2. Example Description

Keeley et al. (2005) and Grace and Keeley (2006) describe the data collection in detail. In brief, 90 sites in southern California were surveyed for 5 years following wildfire. Seven variables were observed indicating 7 latent variables (see **Table 1**). Variables were transformed to generate approximate linear homoscedastic relationships.

BOX 3 | Interpreting evidence using confidence intervals.

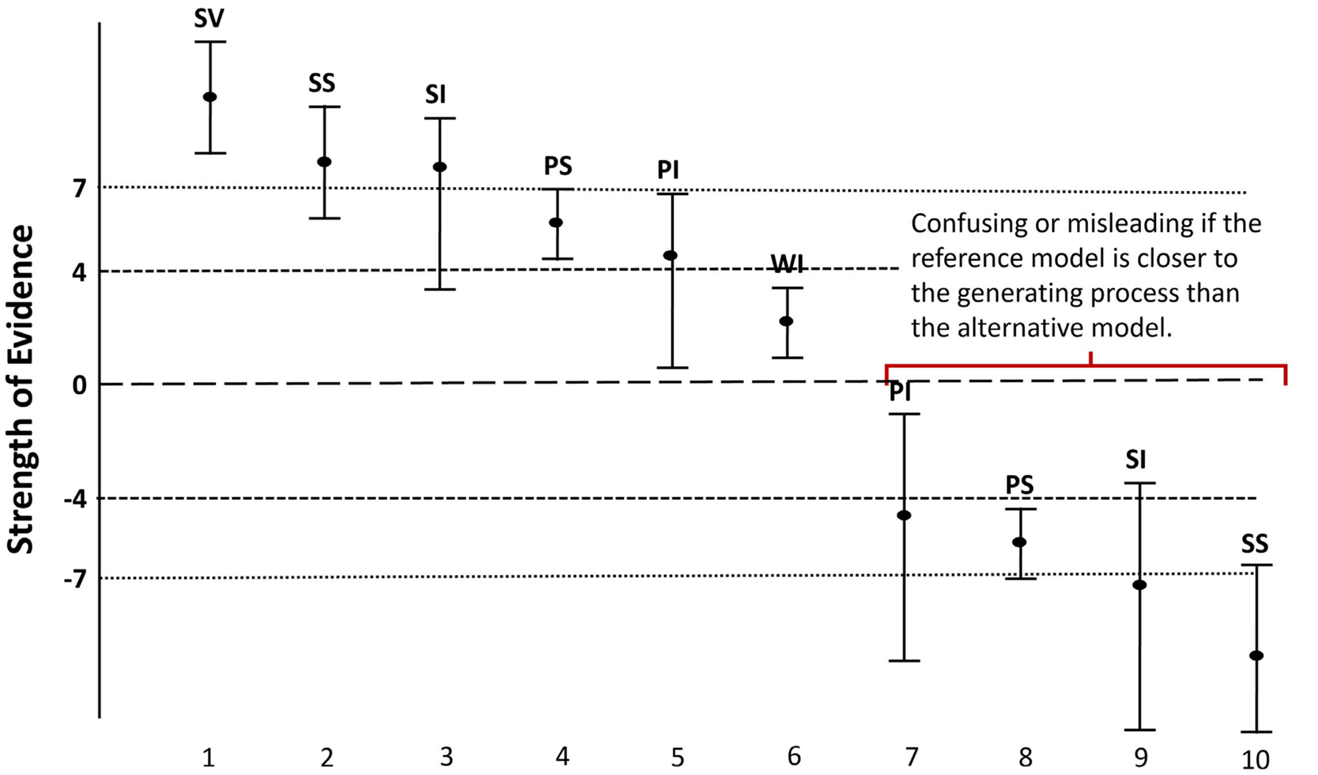


Figure Box 3.1 | depicts some hypothetical confidence intervals for the strength of evidence. The boundaries for the evidential categories are set as: strong evidence for the alternative = $-k_S = -7$, prognostic evidence for the alternative = $-k_P = -4$, prognostic evidence for the reference = $k_P = 4$, and strong evidence for the reference = $k_{\ p}$ = 7.

In interval 1, the observed evidence (e.g., ΔSIC), indicated by the filled oval, is strong and the lower bound for the confidence interval is above the strong evidence threshold. This evidence is designated strong and very secure (SV)—the reference model is strongly supported as being closer to the generating process than the alternative and there is almost no chance that sampling variation would upset this identification. In this case, the researcher may reasonably conclude that no further work is needed regarding model identification in this particular model contrast. Possibly, further work may be indicated to improve parameter estimate precision in the identified better model.

In interval 2, the observed evidence is above the strong evidence threshold, and the proximal bound is greater than the prognostic evidence threshold. We call this situation "strong but secure" (SS). This implies that the reference model is strongly supported, and it is unlikely (but plausible) that this is due to sampling variation. Cautious but optimistic interpretation is indicated, and if possible, more data should be collected to confirm the conclusions.

In interval 3, the observed evidence is above the strong evidence threshold, but the proximal bound is less than the prognostic evidence threshold. We call this situation "strong but insecure" (SI). This implies that while the reference model is strongly supported, it is uncertain due to sampling variation. Very cautious interpretation is indicated, and if possible, more data should be collected to confirm the conclusions.

In interval 4, the observed evidence is less than the strong evidence threshold, and the proximal bound is greater than the prognostic evidence threshold. We call this situation "prognostic but secure" (PS). This implies that while the reference model has only moderate support, it is unlikely that this is due to sampling variation. In this case, the distal bound is less than the strong evidence threshold. It is likely that both models explain the data nearly equally well, but with a slight edge to the favored model.

In interval 5, the observed evidence is less than the strong evidence threshold, and the proximal bound is less than the prognostic evidence threshold. We call this situation "prognostic but insecure" (PI). This implies that the reference model has only moderate support and even this may be due to sampling variation. The primary implication is that more data is needed either within the context of the current experiment or by combining these results with the results of other experiments.

In interval 6 the evidence is weak and insecure (WI). The models are not differentiated by the data. The researcher should collect more data in order to identify the models. The researcher should of course recognize that not all data is equally informative and seek data that will distinguish the two models (e.g., Cooper et al., 2008). Another choice that could be made, particularly if large amounts of data have already been collected, is to decide that both models are adequate for the intended purposes (Lindsay, 2004; Markatou and Sofikitou, 2019).

Intervals 7, 8, 9, and 10 are reflections of intervals 5, 4, 3, and 2, only in this case they are misleading. The designation C stands for confusing evidence, which is prognostic evidence for the wrong model. The designation M stands for misleading evidence, which is strong evidence for the wrong model.

Interval 10 is a researcher's worst case. The evidence is strong, secure and misleading. The researcher should try to avoid this situation both by experimental design (large sample size, treatments or observations that strongly differentiate between the models) and by analytic design (higher strong and marginal evidence thresholds).

(Continued)

BOX 3 | (Continued)

In practice, we do not know if the evidence is misleading or not. For this discussion, we consider "negative" evidence as misleading but in fact, it only indicates that it supports the alternative model—unless one knows the location of the generating process (see Ponciano and Taper, 2019). Simulations (Section 5.2 and Taper et al., 2019) show that for global evidence strong but secure misleading evidence occurs very rarely—regardless of whether the model set is correctly specified or misspecified. For local evidence if the model set is correctly specified, secure misleading evidence is exceedingly rare. However, under model misspecification secure misleading evidence occurs more frequently, although it is still not common. We present an explicit comparison of global and local inference under correctly specified and misspecified models in Section 5.2.

TABLE 1 | Descriptions of variables from Grace and Keeley (2006).

Observed variable G&K name	G&K Data file name	Latent variable G&K name	Single character abbrev. TLPD&J	Measurement error assumed
Distance from coast	Distance	Landscape Position	L	No
Age	Age	Stand Age	A	No
Community heterogeneity	Hetero	Heterogeneity	H	Yes
Abiotic optimum	Abiotic	Local abiotic conditions	C	No
Fire index 1	Firesev	Fire severity	F	Yes
Species/plot	Rich	Richness	R	No
Total cover	Cover	Plant cover	P	No

3.3. Model Naming Conventions

We will use a model naming convention that indicates latent variable regression structure. The single character abbreviation for a variable will be followed by "." and then by the abbreviations for the variables it is regressed on. Regressions with different response variables will be separated by "_."

If a latent is isolated, that is it is neither a response nor a predictor in any regression in the model, its character would be entered in the model name but not followed by a "." We don't consider any such models, because we are picking up the Grace and Keeley reanalysis mid-stream, after they eliminated a variable called "Community Type" from their analysis. Alphabetical order will be imposed so that a path model uniquely determines a name. Thus the Grace and Keeley best model can be named: "A.L_C.L_F.A_H.L_P.F_R.CHLP" (see **Figure 2** and **Table 1**).

3.4. Example Reanalysis

Dr. Grace kindly provided the original data set and his original code (written using R package lavaan). In our reanalysis we use the R package lava (version 1.6.7). The estimates of the standardized coefficients from the two packages agree to at least the 5 decimal places reported by lava. Grace and Keeley determine their best model based on several factors including theoretical background, chi-square model adequacy tests, generalized likelihood ratio tests between nested models, and inspection of deviations between observed and model implied covariances. Grace and Keeley note the consistency of their model identification with identification based on information criterion.

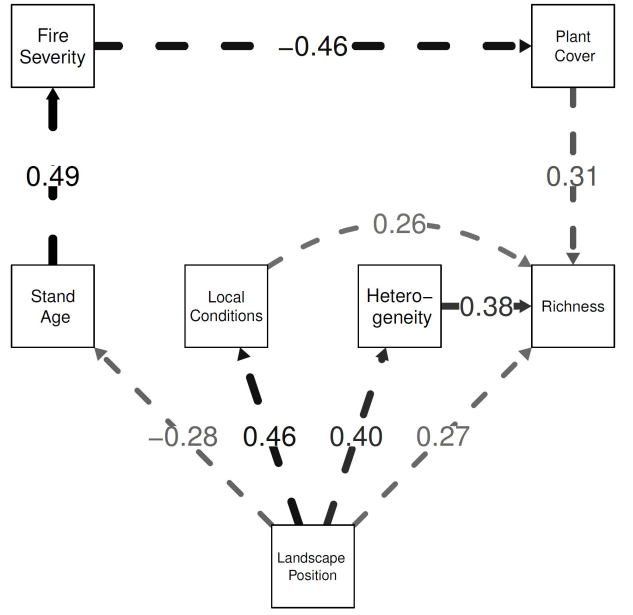


FIGURE 2 | The estimated final, simplified model explaining plant diversity. Arrows indicate causal influences. The standardized coefficients are indicated by path labels and widths. Weak paths with coefficients of magnitude less than 0.30 are shown in gray.

The strong theoretical relationship between ΔICs, the difference of information criterion values, and the likelihood ratio test statistic has been noted before (e.g., Burnham and Anderson, 2002; Lele and Taper, 2012; Taper and Ponciano, 2016). What differs between the approaches are the assumptions and warrants that tie the statistics to scientific inference. These differences can lead to substantive differences in inference from the same data and essentially the same statistic. With a NP test you inference is a categorical accept or reject if your p-value is 0.051, just the wrong side of alpha of 0.05 your reject. If you have a ΔIC of 6.9, you don't reject it instead you give a more elaborate discussion: "Well the evidence doesn't quite reach our arbitrary strong evidence threshold, but it is very strong prognostic evidence." We will return to this in the discussion (see also **Box 1**). Here we focus on the impact of uncertainty in evidence for one model over another given the data on reasonable scientific inference.

TABLE 2 | Models compared in our reanalysis of the Grace and Keeley (2006) structural equation analysis of diversity recovery after fire.

Full name	Description
A.L_C.L_F.A_H.L_P.F_R.CHLP	GKBM (G&K best model)
A.L_C.L_H.L_P.F_R.CHLP	GKBM - F∼A
A.L_F.A_H.L_P.F_R.CHLP	GKBM - C∼L
A.L_C.L_F.A_H.L_R.CHLP	GKBM - P∼F
A.L_C.L_F.A_P.F_R.CHLP	GKBM - H∼L
A.L_C.L_F.A_H.L_P.F_R.CLP	GKBM - R∼H
A.L_C.L_F.A_H.L_P.F_R.CHL	GKBM - R∼P
C.L_F.A_H.L_P.F_R.CHLP	GKBM - A∼L
A.L_C.L_F.A_H.L_P.F_R.HLP	GKBM - R∼C
A.L_C.L_F.A_H.L_P.F_R.CHP	GKBM - R∼L
A.L_C.L_F.A_H.L_P.F_R.CFHLP	GKBM + R∼F. Clarifies G&K question 4
A.L_C.L_F.A_H.L_P.AF_R.CHLP	GKBM + P∼A. Clarifies G&K question 7
A.L_C.L_F.A_H.L_P.F_R.ACHLP	GKBM + R∼A. G&K Model D
A.L_C.L_F.A_H.L_P.FL_R.CHLP	GKBM + P∼L. Added because of covariance residuals
A.L_C.L_F.A_H.L_P.AFL_R.CHLP	GKBM + P∼AL. Added because of covariance residuals

The left-hand column gives the model's full name, which indicates the complete path structure. The right-hand column describes how the model relates to the Grace and Keeley best model.

3.4.1. Models Considered

Statistical evidence, at least defining the term as in the Royall (1997), Lele (2004a), Taper and Ponciano (2016), and Brittan and Bandyopadhyay (2019) tradition, is not unary, but binary: It measures the support (Edwards, 1992) for one model over another model that is given by data. The models we compare are listed in **Table 2**.

The first model is the Grace and Keeley best model (GKBM). The next 9 models are deletion models that each differ from the best model by the absence of a single path. These models are listed in order (strongest to weakest) of the strength of the effect in the best model (as measured by the coefficient z-statistic). Comparison of each of these models with the GKBM will probe the question of whether the deleted path belongs in "best model." The last 5 models are addition models that each differ from the GKBM by the presence of 1 or 2 paths. Comparison of each of the addition models with the GKBM probes the question of whether that/those paths should be included in a "best model."

3.4.2. Example Reanalysis Results

The results of our reanalysis are presented in **Figure 3**, which plots the evidence (ΔSIC) and its uncertainty for the GKBM relative to each of the deletion models, and **Figure 4**, which shows GKBM evidence and uncertainty relative to the addition models.

The first three model comparisons are rock solid. They all have strong and secure global evidence and strong and very secure local evidence. Not only does this data set strongly favor including these three paths, but replication of the experiment—in the same environment—will almost always reach the same conclusion.

The next two comparisons (GKBM - H∼L and GKBM - R∼H) both have strong and secure local evidence for including their paths, but globally, they are insecure. We have good reason to believe that these paths represent real causal effects, but need to advise researchers seeking to replicate this experiment to increase sample size to avoid equivocal results.

Then a comparison (GKBM - R∼P) with evidence, both global and local, that is strong but insecure. Here the global interval crosses the 0 line. Researchers should consider the possibility that the path may be weaker than estimated or may be non-existent.

The next two comparisons have barely prognostic evidence for their paths, but are insecure both globally and locally, with intervals that substantially overlap the line separating evidence for one model versus evidence for the other. The final comparison has positive but weak evidence for inclusion of the path. It is by definition insecure. The local evidence interval falls entirely between the two prognostic evidence thresholds. There is evidence for the path, but it is just a bit more than a toss-up.

Whether or not the last 3 paths should be included in a model is a judgment call for the reporting researchers based on the costs both practical and intellectual of including false paths or omitting true paths. For these deletion paths, a nudge might be given toward including them because the evidence favors the more complex model despite the *SIC* evidence function being used having a slight bias at small sample size toward compact models.

All five addition models have global evidence that is weak and insecure but that leans toward the more compact GKBM. However, all the global intervals overlap the separatrix at 0, and three of the intervals even overlap the marginal evidence thresholds for including the paths. The local evidence shifts slightly further toward the GKBM.

At this sample size, there is no compelling statistical reason to include any of the addition paths in the "best model," but there is also no compelling statistical reason not to. The slight tilt toward the GKBM may represent nothing more that the SIC bias toward compact models. It is very hard statistically to distinguish between the true absence of a path and the presence of a weak path. It would take a sample size of more than 1,000 for there to be an expectation of global strong and secure SIC evidence for the absence of a path even if it was truly absent. On the other hand, because the coefficient of variation of local evidence declines at a much faster rate than that of global evidence (n^{-1} versus $n^{-1/2}$) even a modest increase in sample size may allow local identification of weak effects. In the case of the Grace and Keeley example the breadth of the conditional intervals indicates that the sample size is marginal in a statistical sense—despite the Herculean effort represented.

Models are single entities, but they are entities built from components. In our experience, a great deal of insight into how components function in models can be found by estimating the evidence for a model including the component relative to the same model without that component. In all 14 model comparisons, the weight of evidence tilts toward the GKBM. We agree with Grace and Keeley that A.L_C.L_F.A_H.L_P.F_R.CHLP is the "best model" (at least out of those considered) to describe the structural relationships

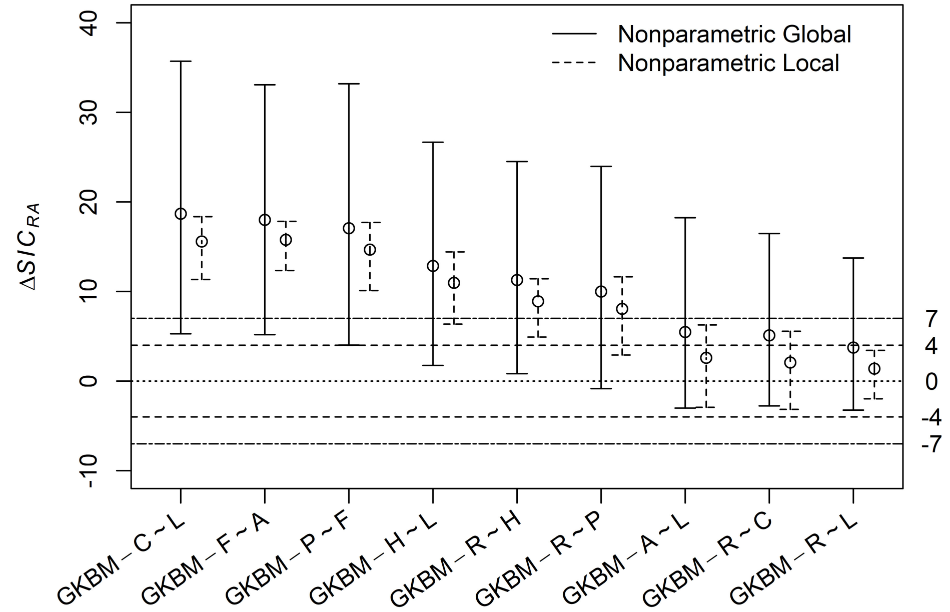


FIGURE 3 | Evidential uncertainty intervals comparing the Grace and Keeley best model with 9 models, each that deletes one of the paths in the GKBM. For each model comparison, the open circle indicates the observed evidence, the solid error bar indicates the global uncertainty, the dashed error bars show the local uncertainty. These are approximate 90% confidence intervals based on 4000 non-parametric bootstraps. The strong evidence thresholds are indicated by dot-dash horizontal limit lines at 7 and -7, while the prognostic evidence thresholds are indicated at dashed limit lines at 4 and -4. Positive values of the ΔSIC_{RA} indicate evidence for the GKBM, as the reference model, relative to the alternative model, while negative values indicate evidence for the alternative model relative to the GKBM. The separatrix between these two regions is the dotted horizontal limit at 0.

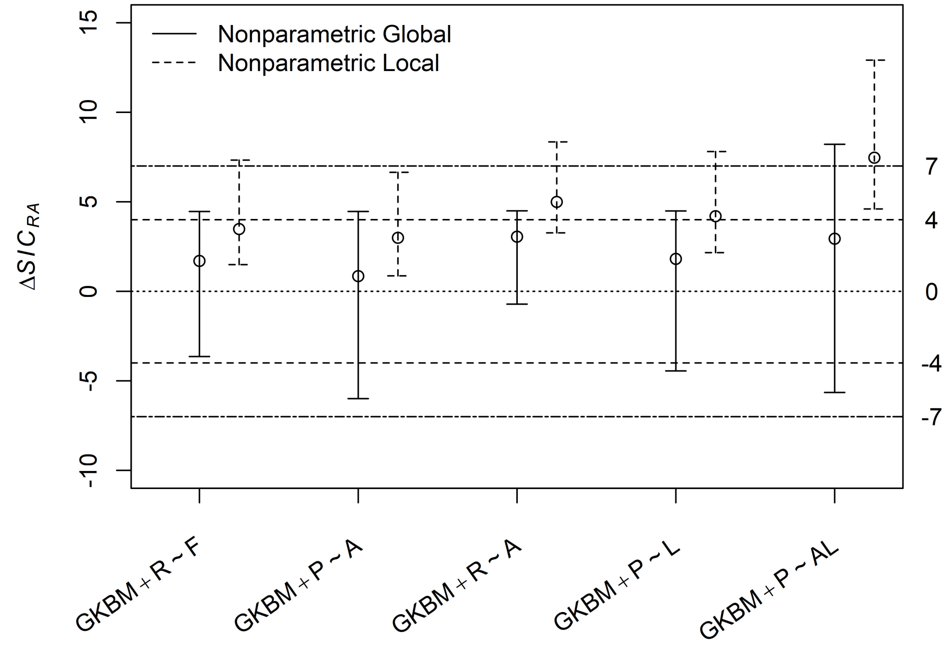


FIGURE 4 | Evidential uncertainty intervals comparing the Grace and Keeley best model with 5 models, each that adds one or two paths to the GKBM.

in this data set. Grace and Keeley chose in 2006 to interpret the empirical results of their study narrowly. “Ultimately, results and interpretations presented in this paper are based on the model judged to be the best representation of the data” (Grace and Keeley, 2006). Here we do disagree with Grace and Keeley. Our analysis has shown that even within a small list of *a priori* models, drawn from their own back-ground theory, there are multiple plausible models whose interpretation should be considered. To interpret only a single best model is like choosing to use only a parameter point estimate without considering its uncertainty. It is simple, but over-confidence can be generated.

4. MATHEMATICAL DEVELOPMENT

In this section, we develop the statistical justification and estimation algorithms for the confidence intervals for evidence that we use in this paper. A reader satisfied with a simulation-based justification could skip to Section 5, at least on first reading.

Different statistical divergences could be used to construct model adequacy measures and thus evidence functions (see Lele, 2004a; Markatou and Sofikitou, 2019). Each will have its own properties, and each could be useful in different circumstances. In this paper we focus on the Kullback-Leibler divergence (KLD) as it leads to the information criteria, evidence functions already in common use. The treatment of uncertainty for other divergences and evidence functions should parallel that for the KLD. The mathematical notation, definitions, and assumptions used in our treatment are given in **Box 4**.

Commonly, either confidence or credible intervals are used to quantify uncertainty in parameter estimates. A very general method of constructing confidence intervals is hypothesis test inversion (Casella and Berger, 2002). If your test is a generalized likelihood ratio test then the set $\left\{\theta, 2\left(l_{m_{\hat{\theta}}}\left(\underline{x}\right) - l_{m_\theta}\left(\underline{x}\right)\right) < \chi^2_{p,(1-\alpha)}\right\}$ is an approximate $100\,(1-\alpha)\,\%$ confidence interval if θ is of dimension 1 or confidence region if θ is of dimension > 1 (Pawitan, 2001).

If one is interested in inference on a subset of the parameters in a multidimensional parameter vector θ, one can partition the parameter vector as $\theta = [\gamma, \lambda]$, where γ is a vector of the parameters of interest, often of dimension 1, and λ is a vector of all the other parameters. A profile log-likelihood (for a given γ) can be calculated as $l_p(\gamma\,;\underline{x}) = \max_{\lambda} l_m\left(\underline{x};\gamma\,,\lambda\right)$, that is by maximizing over λ. It is argued (Cox and Reid, 1987) that maximization of the profile likelihood leads to inconsistent estimators of the parameters of interest because it does not appropriately penalize for the cost of the estimation of the incidental parameters. Various bias corrections or penalty terms for the profile likelihood have been suggested (Pace and Salvan, 2006).

The connection between profile likelihood and model selection becomes obvious if one considers that the parameter of interest could be nothing more than an index for the models considered. In **Box 5** we use this connection to develop and justify global and local uncertainty in the evidence for one model over another. We point out that these penalties for parameter estimation are similar to the penalties employed in information criteria. A general parametric bootstrap approach to calculating an approximate penalty for the profile likelihood is described in Pace and Salvan (2006).

4.1. Divergence Difference, Penalized Divergence Difference, and Evidence Functions

We start with describing precisely the quantities that we want to estimate (targets) and their estimators. An estimator is a function of a random variable and thus describes a probability distribution. An estimator applied to a particular data set produces an estimate, which is a realization from the distribution of estimator.

To understand the bias and uncertainty in an estimator, one needs to compare estimates to estimation targets. For much inference, the targets are obvious. For evidence (which is an estimate), the target was not obvious to us and so to understand the quality of our evidence estimate we begin by first carefully defining what its target is. Then we describe how one can obtain the sampling distribution of these estimators, either asymptotically as was done by Royall (1997, 2000) and Dennis et al. (2019) or by non-parametric bootstrap as was suggested by Taper and Lele (2011).

4.1.1. Fully Specified Competing Models

Consider the case where the competing models are fully specified. In the following, we explicitly define the target quantity, its estimator (the evidence function) and the estimate (observed value of the evidence function). As has been discussed in various papers (Lele, 2004a; Taper and Lele, 2004, 2011; Dennis et al., 2019), the sample size scaled difference between the divergences from the true generating mechanism and the two competing hypothesized mechanisms, namely, $\Delta D_{Pn}(g, M_R, M_A, n) = 2n\{K(g, M_A) - K(g, M_R)\} + c_n(p_A - p_R)$ is of great interest. We call this the penalized scaled divergence difference (see **Box 4**, definition 19). This is an unknown quantity because in practice, we do not know the true generating mechanism $g(.)$.

In this formulation, because of the sample size multiplier $2n$, $\Delta D_{Pn}(g, m_R, m_A, n)$ converges to $\pm\infty$ or 0 as the sample size increases. We use the above formulation to be consistent with the discussion in Dennis et al. (2019) and information-based model selection criteria.

One could, alternatively, standardize the evidence so that it converges to a constant: 0 if the two models are equidistant from the true generating model, a positive number if m_R is closer to $g(.)$ or a negative number if m_A is closer to $g(.)$ as was done in Lele (2004a). One can also use other forms of divergences such as the Hellinger divergence to quantify evidence (Lele, 2004a) to make it model robust or outlier robust.

Given the data $\underline{X}$, a natural estimator of $\Delta D_{Pn}(g, m_R, m_A, n)$, termed the evidence function (Lele, 2004a), is a sample sized scaled difference of the KLD estimators (**Box 4**, definition 21) $2n\{K(g, m_A; \underline{X}) - K(g, m_R; \underline{X})\}$. Notice that, with the KL divergence, the unknown density $g(.)$ gets canceled while taking the difference and does not need to be estimated explicitly. Hence the estimate of the sample size scaled divergence difference, under the KLD, is: $Ev^{raw}(m_{R,} m_A; \hat{g}_{n,\underline{x}}, \underline{x}) = -2\left(l_{m_A}\left(\underline{x}\right) - l_{m_R}\left(\underline{x}\right)\right)$.

In the following, we will describe the use of non-parametric bootstrap to calculate a more accurate estimate of the evidence for the reference model relative to the alternative than the raw evidence and also to quantify uncertainty in the estimated

BOX 4 | Mathematical notations, definitions, and assumptions.

The notation in this box is more verbose than commonly used to allow the reader to track fine distinctions among generating process, distribution estimators, estimated distributions for a particular sample, true parameters, parameter estimators and parameter estimates given a particular sample.

(1) Data are assumed to be suitable for non-parametric bootstrapping. For this paper we further assume that the data are independently and identically distributed (i.i.d.).

(2) Probability density function (pdf) or probability mass function (pmf) representing the true generating mechanism is denoted $g(.)$. Its cumulative distribution function (cdf) is denoted as $F_g(\cdot)$.

(3) Observed data: $\underline{x} = (x_1, x_2, ..., x_n)$, where n denotes the sample size.

(4) Random variables: $\underline{X} = (X_1, X_2, ..., X_n)$.

(5) The pdfs/pmfs for reference (R) and alternative (A) models are denoted by $m_R(.)$ and $m_A(.)$, respectively. For example, m_R is $N(\mu = 5, \sigma = 1)$. Note, these are fully specified models.

(6) If the reference and alternative model are not fully specified, then they represent model spaces denoted M_R and M_A respectively. In that case each of M_R and M_A is a collection of models. For example, $M_R = N(\mu, \sigma))$ with μ in $(-\infty, \infty)$ and σ in $(0, \infty)$.

(7) $F_g^{(n)}(t; \underline{X}) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$ is the empirical estimator of the cdf of $g(.)$ for a random vector of length n. Here $I(A)$ is the indicator function for event A. Denote a corresponding numerically smoothed density as $g_{n,\underline{X}}(.)$.

(8) $\hat{F}_g^{(n)}(t; \underline{x}) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t)$, the empirical estimate of the cdf of $g(.)$ for an observed vector of length n. Denote a corresponding numerically smoothed density as $\hat{g}_{n,\underline{x}}(.)$.

(9) The KLD between two specified continuous models, where the reference model is m_R is $K(m_R, m_A) = \int(\log(m_R(x)) - \log(m_A(x)))\, m_R(x)\, dx$. In general, for any two models (discrete, continuous or piecewise continuous) we write $K(m_1, m_2) = \int(\log(m_1(x)) - \log(m_2(x)))\, dF_{m_1}(x)$.

(10) The KLD orthogonal projection of a probability distribution, such as a fully specified model, $s(.)$ onto a model space M is $m_s^* = \underset{m \in M}{\arg\min} K(s(.), m)$ (see **Figure 3** in Ponciano and Taper, 2019). This model is the closest approximation to $s(.)$ in the model space M.

(11) If $s(.) \in M \Rightarrow m_s^*(.) \equiv s(.)$. If the generating process is in either M_R or M_A that is if either $g(.) \in M_R$ or $g(.) \in M_A$ then the model set $\{M_R, M_A\}$ is considered correctly specified, as in the foundations of much classical statistics (e.g., Neyman and Pearson, 1933; Wilks, 1938; Wald, 1943).

(12) The log-likelihood function for the observed data, $\underline{x}$, under $g(.)$ is $l_g(\underline{x}) = \sum_{i=1}^{n} \log(g(x_i))$
The log-likelihood function for the observed data under a model $m(.)$ is $l_m(\underline{x}) = \sum_{i=1}^{n} \log(m(x_i))$. $\hat{m}_{\underline{x}}(.)$ is the model with parameter values that maximizes $l_m(\underline{x})$.

(13) Conceptually, $\hat{m}_{\underline{x}}(.)$ is the same model as $m^*_{\hat{g}_{n,\underline{x}}}$. The first notation is more familiar, the second emphasizes that the maximum likelihood model is a projection of the model to the empirical density. Asymptotically these estimates will be identical, but there will be slight numerical differences at finite sample size due to the smoothing in $\hat{g}_{n,\underline{x}}$.

(14) The KLD estimator of the divergence of a model, m, from the generating process, g is given as

$$K(\hat{g}_{n,\underline{X}}, m; \underline{X}) = \int \log\left(\hat{g}_{n,\underline{X}}(t)\right) dF_g^{(n)}(t; \underline{X}) - \int \log(m(t))\, dF_g^{(n)}(t; \underline{X}) = S_{\hat{g}_{n,\underline{X}}, \hat{g}_{n,\underline{X}}} - S_{\hat{g}_{n,\underline{X}}, m}$$

where $S_{\hat{g}_{n,\underline{X}}, \hat{g}_{n,\underline{X}}}$ is the neg-self-entropy of the generating process and $S_{g_{n,\underline{X}}, m}$ is the neg-cross-entropy from the generating process to the model m. Note, an estimator is the function of a random variable (i.e., $\underline{X}$) that returns an estimate for a particular realization of the random variable.

(15) The KLD estimate of the divergence of a model, m, from the generating process, g:

$$K(\hat{g}_{n,\underline{x}}, m; \underline{x}) = \int \log\left(\hat{g}_{n,\underline{x}}(t)\right) d\hat{F}_g^{(n)}(t; \underline{x}) - \int \log(m(t))\, d\hat{F}_g^{(n)}(t; \underline{x}) = S_{\hat{g}_{n,\underline{x}}, \hat{g}_{n,\underline{x}}} - S_{\hat{g}_{n,\underline{x}}, m} .$$

where $S_{\hat{g}_{n,\underline{x}}\hat{g}_{n,\underline{x}}}$ is the neg-self-entropy of the empirical distribution.

(16) The KLD projection estimator of the divergence of a model space, M, from the generating process, g: $K(\hat{g}_{n,\underline{X}}, M; \underline{X}) = S_{\hat{g}_{n,\underline{X}}, \hat{g}_{n,\underline{X}}} - S_{\hat{g}_{n,\underline{X}}, m^*_{\hat{g}_{n,\underline{X}}}}$

(17) The KLD projection estimate of the divergence of a model space, M, from the generating process, g: $K(\hat{g}_{n,\underline{x}}, M; \underline{x}) = S_{\hat{g}_{n,\underline{x}}, \hat{g}_{n,\underline{x}}} - S_{\hat{g}_{n,\underline{x}}, m^*_{\hat{g}_{n,\underline{x}}}}$

(18) One estimate for $K(\hat{g}_{n,\underline{x}}, M; \underline{x})$ is $S_{\hat{g}_{n,\underline{x}}\hat{g}_{n,\underline{x}}} - l_{\hat{m}(\underline{x})}$, see discussion in definition (13). Bias correction for this estimate is the goal of information criteria. We employ the consistent family of bias correction terms $c_n p$, where c_n is a function of n growing strictly between loglog(n) and n. And, p is the parametric dimension of M (Nishii, 1988).

(19) The global penalized scaled divergence difference target: $\Delta D_{Pn}(g, M_R, M_A, n) = 2n\{K(g, M_A) - K(g, M_R)\} + c_n(p_A - p_R)$ (see definition 16). The target is the quantity for which we attempt to find both a central estimate and an uncertainty measure (see discussion in Section 4.1). Note that for fully specified model comparisons, the penalty term is 0, and $\Delta D_{Pn}(g, m_R, m_A, n) = 2n\{K(g, m_A) - K(g, m_R)\}$

(20) The local penalized scaled divergence difference target, $\Delta d_{Pn}(g, M_R, M_A, \underline{x}) = 2n\{K(g, M_A, \underline{x}) - K(g, M_R, \underline{x})\} + c_n(p_A - p_R)$ (see definition 17).

(21) The global penalized divergence difference estimator, $\Delta D_{Pn}(\hat{g}_{n,\underline{X}}, M_R, M_A, \underline{X}) = E_{\hat{g}_{n,\underline{X}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{Y}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{Y})\} + c_n(p_A - p_R)\right)$. Note that inside the expectation $\underline{Y}$ is a random vector drawn from $\hat{g}_{n,\underline{X}}$.

(22) The local penalized divergence difference estimator, $\Delta d_{Pn}(\hat{g}_{n,\underline{X}}, M_R, M_A, \underline{x}) = E_{\hat{g}_{n,\underline{X}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{x}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{x})\} + c_n(p_A - p_R)\right)$. Note that inside the expectation $\underline{Y}$ is a random vector drawn from $\hat{g}_{n,\underline{X}}$.

(23) The global evidence estimate,

$$\begin{aligned} Ev_G(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}) &= E_{\hat{g}_{n,\underline{x}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{Y}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{Y})\} + c_n(p_A - p_R)\right) \\ &= E_{\hat{g}_{n,\underline{x}}}\left(-2\{l_{\hat{m}_{A_{\underline{Y}}}}(\underline{Y}) - l_{\hat{m}_{R_{\underline{Y}}}}(\underline{Y})\} + c_n(p_A - p_R)\right) \end{aligned}$$

. Note that inside the expectation $\underline{Y}$ is a random vector drawn from $\hat{g}_{n,\underline{x}}$ and that the maximum likelihood estimate, $\hat{m}_{\underline{x}}$, has been substituted for $m^*_{\hat{g}_{n,\underline{x}}}$ (see definitions 13 and 18). Both the estimated models and the data from which the likelihoods are calculated are random. Thus, variation in Ev_G is due to both variation in $\underline{Y}$ and to variation in the estimates of $\hat{m}_{A_{\underline{Y}}}$ and $\hat{m}_{R_{\underline{Y}}}$. Non-parametric bootstrap will be used to estimate the expectation and its uncertainty estimation and for further bias reduction. Positive values for evidence indicate that the reference model is supported over the alternative model (see discussion Box 1).

(Continued)

BOX 4 | (Continued)

(24) The local evidence estimate, $$Ev_L\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right) = E_{\hat{g}_{n,\underline{x}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{x}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{x})\} + c_n(p_A - p_R)\right)$$ $$= E_{\hat{g}_{n,\underline{x}}}\left(-2\{l_{\hat{m}_{A_{\underline{Y}}}}(\underline{x}) - l_{\hat{m}_{R_{\underline{Y}}}}(\underline{x})\} + c_n(p_A - p_R)\right).$$ Note that inside the expectation $\underline{Y}$ is a random vector drawn from $\hat{g}_{n,\underline{x}}$ and that the maximum likelihood estimate, $\hat{m}_{\underline{x}}$, has been substituted for $m^*_{\hat{g}_{n,\underline{x}}}$ (see definition 18). Here the estimated models are random, but the data from which the likelihoods are calculated are fixed. Thus, variation in Ev_L is due only to variation in the estimates of $\hat{m}_{A_{\underline{Y}}}$ and $\hat{m}_{R_{\underline{Y}}}$. Non-parametric bootstrap will be used to estimate the expectation and its uncertainty estimation and for further bias reduction. Positive values for evidence indicate that the reference model is supported over the alternative model (see discussion Box 1).

(25) The raw evidence, $$Ev^{raw}(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}) = 2n\{K(\hat{g}_{n,\underline{x}}, M_A, \underline{x}) - K(\hat{g}_{n,\underline{x}}, M_R, \underline{x})\} + c_n(p_A - p_R)$$ $$\approx -2\{l_{\hat{m}_{A_{\underline{x}}}}(\underline{x}) - l_{\hat{m}_{R_{\underline{x}}},}(\underline{x})\} + c_n(p_A - p_R)$$. Note that no bootstrapping is done nor expectation taken. This is an information criterion as generally used.

evidence as was suggested in Taper and Lele (2011). **Box 6** lists an explicit algorithm for this bootstrap.

Instead of the LLR as the estimated evidence, we use the expectation (mean) of the density function of the bootstrap evidence as the estimated evidence. This could be estimated as the bootstrap average evidence. For a slight increase in accuracy, we calculate the expectation by numerically integrating over an estimated density function for the bootstrapped evidence. We use the R package kde1d (version 1.0.2, Nagler and Vatter, 2019), which uses univariate local polynomial(log-quadratic) kernel density estimators. Our validation tests support the literature (Geenens and Wang, 2018) on the strength of this method. We find that confidence bounds are located more accurately with kde1d quantiles than with raw bootstrap quantiles, BCa quantiles, or with calibrated (double bootstrap) quantiles (see Efron and Tibshirani, 1993 for description of these methods) and that estimated distributions are more accurate (in integrated squared error) than standard kernel density estimation.

We note a few important features of the bootstrapping procedure described in **Box 5**. When the models are fully specified the log-likelihood ratio is a U-statistic (Serfling, 1984) and hence it is an unbiased estimator of the target quantity. However, divergences other than KLD may lead to biased estimators of the target quantity. In which case, the mean of the bootstrap distribution is a bias corrected estimate of the target quantity. Also, if the models are not fully specified, it is well known that the log-likelihood ratio is a biased estimator of the target quantity (Akaike, 1973). The mean of the bootstrap distribution of the log-likelihood ratio corrects for bias (Ishiguro et al., 1997).

We do not discuss the case of fully specified models any further but move on to the interesting case where parameters need to be estimated.

4.1.2. Competing Models With Unknown Parameter Values

Next, we consider the problem of model selection where there are unknown parameter values that need to be estimated. When we are dealing with model selection, the quantity of interest is scaled divergence difference penalized for the complexity of the models. We consider global penalized scaled divergence differences of the form: $\Delta D_{Pn}(\hat{g}_{n,\underline{X}}, M_{R,} M_A, n) = E_{g_{n,\underline{X}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{Y}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{Y})\} + c_n(p_A - p_R)\right)$, where c_n is a function of the sample size that converges to infinity at the rate strictly between $\log(\log(n))$ and n (Nishii, 1988), p_R and p_A are the number of unknown quantities (parameters) in the models that are estimated using the data. For example, for the Schwarz Information Criterion (SIC), $c_n = \log(n)$. This constraint guarantees that the information criterion will be a consistent criterion; that is, asymptotically it will lead to identifying the model in the model space that is closest to the true generating mechanism. We include the multiplier 2 to keep it consistent with common information criteria. We emphasize again that, the target, $\Delta D_{Pn}(g, M_{R,} M_A, n)$, is unknown in practice.

Assuming that the observations in the data are independent, identically distributed random variables, using the SIC (a.k.a. Bayesian Information Criterion or BIC) sample size correction, and using the maximum log-likelihood as an estimator of the KLD of a model to the generating process, leads to the evidence function $Ev_G\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right) \approx E_{\hat{g}_{n,\underline{x}}}\left(-2\{l_{\hat{m}_{A_{\underline{Y}}}}(\underline{Y}) - l_{\hat{m}_{R_{\underline{Y}}}}(\underline{Y})\} + c_n(p_A - p_R)\right)$, where $Y_i \sim \hat{g}_{n,\underline{x}}$ (see definition 23 **Box 4**), and $\hat{m}_{R_{\underline{Y}}}$ and $\hat{m}_{A_{\underline{Y}}}$ are those models in M_R and M_A that are closest to $\hat{F}_g^{(n)}(.)$, the empirical CDF based on the data $\underline{Y} = (Y_1, X_2, ..., Y_n)$, a random vector of length n from $\hat{g}_{n,\underline{x}}$. Note that inside the expectation $\underline{Y}$ is a random vector drawn from $\hat{g}_{n,\underline{x}}$ and that the maximum likelihood estimate, $\hat{m}$, has been substituted for m^* (see definition 18). Variation in Ev_G is due to variation in $\hat{m}_{A_{\underline{Y}}}$, $\hat{m}_{R_{\underline{Y}}}$, and $\underline{Y}$. We calculate the expectation by numerically integrating over an estimated density function for the bootstrapped ΔSIC_{RA}s. We use the R package kde1d for the density estimation. **Figure 5** presents a schematic of this development.

We point out that, except for the nuance of kernel density smoothing, the algorithm we describe above for Ev_G is the EIC algorithm of Ishiguro et al. (1997) applied to ΔICs rather than directly to log-likelihoods. Kitagawa and Konishi (2010) point out that the bootstrap bias correction can be applied to any functional, not just the log-likelihood. The use of the expectation of the sampling distributions of ΔICs, which already contain an analytic bias correction, adds another layer of bias correction. Accordingly, the evidence should be 3rd order accurate (Kitagawa and Konishi, 2010).

Similarly, the local evidence function $Ev_L\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right)$ is an estimate of the local penalized scaled divergence difference, $\Delta d_{Pn}(g_{n,\underline{X}}, M_{R,} M_A, \underline{x}) = E_{g_{n,\underline{X}}}\left(2n\{K(g_{n,\underline{Y}}, M_A, \underline{x}) - K(g_{n,\underline{Y}}, M_R, \underline{x})\} + c_n(p_A - p_R)\right)$

BOX 5 | Adjusted profile likelihood for model selection inference.

Readers can see Meeker and Escobar (1995) for a brief introduction to profile likelihood in the context of confidence interval construction and Pierce and Bellio (2017) for a substantial review of practical likelihood adjustments. A gentle introduction to model selection through information criteria can be found in Anderson (2008), with more technically robust discussions in Burnham and Anderson (2002) or Konishi and Kitagawa (2008).

A general parametric bootstrap approach to calculating an approximate penalty for the profile likelihood is described in Pace and Salvan (2006) and outlined below.

Let $M_\varphi, \varphi = 1, 2, ..., S$ denote S distinct model spaces. The goal of model selection is to use the data to select the best model space. The form of the best model space is used to draw various statistical and scientific inferences about the generating mechanism.

First, we show that model selection procedure can be looked upon as a profile likelihood estimation procedure. Let $\{\underline{\theta}_1, \underline{\theta}_2, ..., \underline{\theta}_S\}$ denote the parameters for the respective model spaces $(M_1, M_2, ..., M_S)$. Denote the dimension of $\underline{\theta}_\varphi$ by p_φ.

A universal model space, that is simply a union of the model spaces, may be written as $M = \{f(x; \varphi, \underline{\theta}_\varphi), \varphi = 1, 2, ..., S\}$. In this notation, $f(x; 1, \underline{\theta}_1)$ indicates the parametric form of the probability model in the first model space, say $LogNormal(\mu, \sigma^2)$, $f(x; 2, \underline{\theta}_2)$ denotes the parametric form of the probability model in the second model space, say $Gamma(\mu, \phi)$, and so on. The parameter φ, which is a discrete parameter, is simply an index for the model space. Thus, model selection can be viewed as selecting a particular value of φ. In model selection problem, the index parameter φ is of interest and model parameters $\underline{\theta}_\varphi$ are the incidental parameters. The profile likelihood of the index parameter φ can be written as: $l_p(\varphi, \hat{\underline{\theta}}_\varphi; \underline{x}) = \underset{\underline{\theta}_\varphi}{max} \sum_{i=1}^{n} \log f(x_i; \varphi, \underline{\theta}_\varphi)$.

In the familiar example of the maximum likelihood estimator of the variance σ^2 in the multiple linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ independent, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} \right)^2$. This is a biased estimator and bias is pronounced when the number of covariates is large. A bias corrected profile likelihood yields the usual unbiased estimator with the divisor $(n - p - 1)$, instead of n. We lose $(p + 1)$ degrees of freedom because we spend some of the information in the data to estimate the nuisance parameters $(\beta_0, \beta_1, ..., \beta_p)$.

We describe the Pace-Salvan approach for the general profile likelihood case where the parameter of interest may or may not be discrete. To reflect this generality, for the description of the Pace-Salvan approach, we make a slight change in the notation. We use γ for the parameter of interest, λ for the incidental parameters and $h(.)$ denotes the parametric probability function presumed to be the data generating mechanism.

Let $X \sim h(., \gamma, \lambda)$. Let the parameter of interest, γ, be of dimension 1 and the nuisance parameter λ be a vector of any dimension that does not depend on the sample size. Let $\underline{x} = (x_1, x_2, ..., x_n)$ be a random sample of size n from $h(., \gamma, \lambda)$. The log-profile likelihood for γ is defined as $l_p(\gamma; \underline{x}) = \underset{\lambda}{max} \sum_{i=1}^{n} \log(h(\gamma, \lambda; x_i))$.

Model selection based on the maximum of this profile likelihood would correspond to selecting the model space that maximizes the log-likelihood but without any penalty for the number of parameters in the model. This procedure is known to lead to what is termed an inconsistent model selection procedure. The reason for the inconsistency is that this profile likelihood is a biased estimator of the expected Kullback-Leibler divergence (Akaike, 1973; see discussion in Ponciano and Taper, 2019). The inconsistency of and the bias correction used in information-based model selection bears strong similarity to the inconsistency and bias correction in the profile likelihood estimators (e.g., Severini, 2000; Pace and Salvan, 2006) suggested in a very different context.

Following Pace and Salvan (2006), the adjusted profile likelihood, adjusted for the effects of estimation of the nuisance parameter λ, can be computed, assuming the presumed model is the true generating mechanism, using parametric bootstrap as follows:

(1) Estimate the full parameter vector $(\hat{\gamma}, \hat{\lambda})$.

(2) For each bootstrap iteration $b \in \{1, \cdots, B\}$

  (a) Generate a random sample of size n from $h(.; \widehat{\gamma}, \hat{\lambda})$ denoted by $\underline{x}_b = (x_{b,1}, ..., x_{b,n})$.

  (b) For these new data and for a fixed value of γ, obtain $\hat{\lambda}_b(\gamma)$ by $\underset{\lambda}{max} \sum_{i=1}^{n} \log(h(\gamma, \lambda; x_{b,i}))$.

(3) Compute the simulation adjusted profile likelihood as: $l_{SA}(\gamma; \underline{x}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \log(h(\gamma, \hat{\lambda}_b(\gamma); x_i))$. We point out specifically that the likelihood is evaluated for the original data $\underline{x}$ but with the parameters $(\gamma, \hat{\lambda}_b(\gamma))$ that are estimated using the bootstrap data.

Pace and Salvan (2006) suggest using $l_{SA}(\gamma; \underline{x})$, instead of $l_p(\gamma; \underline{x})$ to conduct statistical inference for γ, the parameter of interest. Most importantly, they use sophisticated mathematics to show that the adjustment achieved by $l_{SA}(\gamma; \underline{x})$ is locally (conditionally, post-data, post-experiment) appropriate. Note that following Efron and Tibshirani (1993) description of bootstrap bias correction, one may use $l_A(\gamma; \underline{x}) = 2l_p(\gamma; \underline{x}) - l_{SA}(\gamma; \underline{x})$. It follows from the results in Section 3.4 of Pace and Salvan (2006) that these two versions are equivalent up to $O(n^{-1})$ and that the difference between these central estimates is small compared to the uncertainty. We use the mean of the bootstrap distribution as our central estimate to be consistent with both Pace and Salvan (2006) and Kitagawa and Konishi (2010). There is reason to believe that the median of the bootstrap distribution might have superior theoretical properties (De Blasi and Schweder, 2018), but we will pursue this in another paper.

We point out that these penalties to the profile likelihood for parameter estimation are similar to the penalties employed in information criteria. In the information theoretic literature, non-parametric bootstrap bias corrections have been developed as the extended information criterion (EIC) (Ishiguro et al., 1997; Konishi and Kitagawa, 2008; Kitagawa and Konishi, 2010). There are two important, differences between the basic (EIC) and the Pace-Salvan adjusted profile likelihood. First, EIC uses non-parametric bootstrap whereas Pace and Salvan use parametric bootstrap. The use of non-parametric bootstrap relaxes the assumption that the parametric model is the true generating mechanism. Model misspecification is built into the EIC correction. And second, bias correction in EIC is a global (unconditional, pre-data, pre-experiment) adjustment, averaging over the variation from one experiment to other, whereas the Pace-Salvan adjustment is a local (conditional, post-data, post-experiment) adjustment that evaluates the likelihood at the observed data $\underline{x}$ but is averaged over variation of the incidental parameter estimates from one bootstrap sample to the other.

The bias correction for the EIC can be decomposed into three components: D_1, D_2, D_3 (Kitagawa and Konishi, 2010). One component, D_2, has expectation 0 and is discarded in the EIC_2, the variance reduced form of the EIC. The EIC bootstrap bias correction can be applied not just to the likelihood of the data, but to any functional of the data. Some algebra on equations 44 and 51 of Kitagawa and Konishi (2010) shows that $Ev_L\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right) = EIC_2\left(\Delta SIC_{RA}\left(\underline{x}\right)\right) + D_1\left(\Delta SIC_{RA}\left(\underline{x}\right)\right)$. We have found numerically that $D_1\left(\Delta SIC_{RA}\left(\underline{x}\right)\right)$ is a small term that appears to have mean at or near 0, at least under the conditions that we have investigated. The SIC includes an analytic bias correction to the likelihood accounting for the number of parameters estimated. Thus, that $D_1\left(\Delta SIC_{RA}\left(\underline{x}\right)\right)$ is small in these cases does not mean that D_1 is always unimportant, just that we are in a region of model space where the analytic bias correction works well. Central estimates for evidence and uncertainty intervals could be based on the entire EIC_2. We will explore these connections elsewhere.

(Continued)

BOX 5 | (Continued)

The Pace-Salvan adjusted profile likelihood, the EIC and the EIC$_2$ use the bootstrap distribution only to compute the bias correction factor. We stress the use of the entire bootstrap distribution to quantify uncertainty in the evidence. A non-parametric bootstrap procedure similar to the Pace-Salvan approach yields local uncertainty while a bootstrap similar to the EIC can give us global uncertainty.

(see **Box 4** definition 22), and $Ev_L\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right) \approx E_{\hat{g}_{n,\underline{x}}}\left(-2\{l_{\hat{m}_{A_{\underline{Y}}}}\left(\underline{x}\right) - l_{\hat{m}_{R_{\underline{Y}}}}\left(\underline{x}\right)\} + c_n(p_A - p_R)\right)$ (see **Box 4** definition 24). The difference between the global and local is that in the calculation of the global evidence the observed data, $\underline{x}$, are considered as a realization of a random vector, $\underline{X}$, both in the estimation of the models to be compared and in the data on which they are compared. While in the local evidence, the data vector is considered random in the estimation of the models but fixed in the data on which they are compared.

It is well established in statistics that providing an estimate of an unknown quantity is not sufficient; one must provide uncertainty associated with such an estimate. We use aleatory probability to quantify this uncertainty (Lele, 2020a). In quantifying the pre-experiment uncertainty in evidence, we ask the question: How variable would the evidence be if we were to repeat the experiment? This is represented by the global (pre-experiment) sampling distribution of the evidence function. This distribution does not depend on the particular data set in hand.

When the competing models are fully specified and the reference model is the true model, Royall (1997, 2000) used the asymptotic Normal distribution of the LLR to approximate the sampling distribution of the evidence function and calculate the error probabilities. In Dennis et al. (2019), we derived the asymptotic distributions of the evidence function when the competing models are not fully specified and the true model is not part of the competing model spaces to approximate the sampling distribution and compute the error probabilities.

4.2. Uncertainty in Evidence

An important element common to all of our bootstrap procedures is that the complete evidence functions are the objects bootstrapped, not the component divergences. Thus, if the difference of information criterion values is the evidence function used, such a bootstrap will produce a single distribution of ΔICs rather than two distributions of IC values. This is necessary because the geometry of model misspecification (Dennis et al., 2019; Ponciano and Taper, 2019, see also **Table 3**) can create covariances (positive and negative) between the component divergences. These need to be captured by a bootstrap for it to accurately reflect the uncertainty in evidence. The non-parametric bootstrap method for the two cases described above is as follows.

4.2.1. Global Uncertainty in Evidence for the Fully Specified Models

Notice that in the bootstrap procedure in Section 4.1.1, we are bootstrapping the difference in the log-likelihood jointly and not each component separately. Evidence, innately, is a comparison between two quantities. Clearly uncertainty in evidence involves not just the variances of each component but also covariance between them. The uncertainty reflected in the bootstrap distribution accounts for the covariance also. Thus, if the two models are positively correlated with each other, the uncertainty is reduced whereas if they are negatively correlated, the uncertainty is higher than the sum of variances. This, thus, takes into account the geometry of the model spaces appropriately, even when the models are fully specified. The quantiles of the smoothed bootstrap density of $Ev^{raw}\left(m_R, m_A; \hat{g}_{n,\underline{x}}, \underline{x}_b\right)$ give us confidence intervals for evidence (see **Box 6** for an explicit algorithm).

4.2.2. Global Uncertainty in Evidence for Model Spaces With Unknown Parameter Values

Bootstrapping can also be used to obtain global confidence intervals for evidence with estimated parameters. The only difference is that the quantity bootstrapped is $Ev_G^{raw}(M_{\mathrm{R}}, M_{\mathrm{A}}; \hat{g}_{n,\underline{x}}, \underline{x}_b)$, which is, in this paper, a difference of information criterion values (see **Box 6** for an explicit algorithm).

4.2.3. Local Uncertainty in Evidence

Lele (2020a) reviewed the philosophical problems associated with global (pre-experiment) uncertainty and discussed the use of local (post-experiment) uncertainty in the context of linear regression. To recap, suppose we have only one covariate and we are fitting a linear regression through origin model. That is, the data are (x_i, y_i), $i = 1, 2, \ldots, n$ and we fit the model $Y_i = \beta X_i + \varepsilon_{\mathrm{i}}$ where $\varepsilon_{\mathrm{i}} \sim N(0, \sigma^2)$ are independent, identically distributed random variables. The maximum likelihood estimator of β is, $\hat{\beta} = \sum Y_i X_i \big/ \sum X_i^2$.

The question is: what is the variance of $\hat{\beta}$? If we consider the covariates to be random (this is the case when the experiment is not a designed experiment but an observational study), then $var(\hat{\beta}) = \sigma^2 E\left(1 \big/ \sum X_i^2\right)$. If $X_i \sim N(0, 1)$, then $var(\hat{\beta}) = \sigma^2/(n-2)$. This variance, which we term the global variance, is sometimes called an unconditional or pre-data variance. On the other hand, if we consider the covariates to be fixed, as is the case in designed experiments, $var(\hat{\beta}|x_1, x_2, \ldots, x_n) = \sigma^2/\left\{\sum x_i^2\right\}$. This variance, which we call the local variance, is sometimes called the conditional or post-data variance.

The conditional variance is the variance most ecologists use when conducting regression analysis. Notice that conditional variance depends on the configuration of covariates the researcher observes in their particular data set. If the covariate values are highly dispersed, the slope is extremely well estimated; on the other hand, if the observed covariates values are not very different from each other, the slope is estimated with large uncertainty.

The local (conditional) variance makes intuitive sense: good data, strong inference; bad data, weak inference. It is argued (e.g., Goutis and Casella, 1995) that the global (unconditional) inference does not reflect this differentiated inferential value of

BOX 6 | Bootstrap algorithms for global and local evidence uncertainty.

All of the bootstraps described in this box can be performed using the R function KKICv, which we supply in Supplemental Material.

Evidence uncertainty for specified models:

(1) Obtain a random sample of size *n* with replacement from the original sample. This bootstrap sample is denoted by $\underline{x}_b = (x_{b1}, x_{b2}, ..., x_{bn})$.

(2) Evaluate the evidence at the bootstrap sample, namely, $Ev^{raw}\left(m_R, m_A; \hat{g}_{n,\underline{x}}, \underline{x}_b\right) = -2\left(l_{m_A}\left(\underline{x}_b\right) - l_{m_R}\left(\underline{x}_b\right)\right)$.

(3) Repeat steps 1 and 2 B times and accumulate to get the set of results $\{Ev^{raw}(m_R, m_A; \hat{g}_{n,\underline{x}}, \underline{x}_b), b = 1, 2, ..., B\}$.

(4) Estimate the density function of the $\{Ev^{raw}(m_R, m_A; \hat{g}_{n,\underline{x}}, \underline{x}_b)\}$ in 3). Quantiles of this density yield confidence intervals for the evidence.

(5) Calculate $Ev(m_{R,} m_A; \hat{g}_{n,\underline{x}}, \underline{x})$ as the expectation (mean) of the estimated density from step 4.

Global evidence uncertainty estimation:

(1) Obtain a simple random sample of size n with replacement from the observed data $\underline{x}$. Let us denote this by $\underline{x}_b = (x_{b,1}, x_{b,2}, ..., x_{b,n})$.

(2) Based on this bootstrap data, estimate the model parameters for each model space. Let us denote these models by $\hat{m}_{R,b}$ and $\hat{m}_{A,b}$. These are projections of the empirical CDF of the bootstrap data onto the corresponding model spaces.

(3) Compute and store $Ev^{raw}(M_{\mathrm{R},} M_{\mathrm{A}}; \hat{g}_{n,\underline{x}_b}, \underline{x}_b) = -2\{l_{\hat{m}_{A,\underline{x}_b}}\left(\underline{x}_b\right) - l_{\hat{m}_R\underline{x}_b,}\left(\underline{x}_b\right)\} + c_n(p_A - p_R)$. The smoothed density of $Ev^{raw}(M_{\mathrm{R},} M_{\mathrm{A}}; \hat{g}_{n,\underline{x}}, \underline{x}_b),\ b = 1, 2, ..., B$ is the bootstrap estimate of the sampling distribution of $Ev^{raw}(M_{\mathrm{R},} M_{\mathrm{A}}; \hat{g}_{n,\underline{x}_b}, \underline{x}_b)$.

(4) Quantiles of the smoothed density of $Ev^{raw}(M_{\mathrm{R},} M_{\mathrm{A}}; \hat{g}_{n,\underline{x}_b}, \underline{x}_b)$ give us confidence intervals for evidence.

(5) Calculate $Ev_G(M_{R,} M_A; \hat{g}_{n,\underline{x}}, \underline{x})$ as the expectation (mean) of the estimated density from step 4.

Local evidence uncertainty estimation:

(1) Generate a random sample with replacement and of size *n* from the observed data. Let us denote this by $\underline{x}_b = (x_{b,1}, x_{b,2}, ..., x_{b,n})$.

(2) Re-estimate the parameters using the bootstrap sample. Let us denote them by $\hat{m}_{R,b}$ and $\hat{m}_{A,b}$.

(3) Compute $Ev^{raw}(M_{R,} M_A; \hat{g}_{\underline{x}_b}, \underline{x}) = -2\{l_{\hat{m}_{A_{\underline{x}_b}}}\left(\underline{x}\right) - l_{\hat{m}_{R_{\underline{x}_b}}}\left(\underline{x}\right)\} + c_n(p_A - p_R)$.

(4) Use the quantiles of the smoothed bootstrap distribution of $Ev^{raw}(M_{R,} M_A; \hat{g}_{\underline{x}_b}, \underline{x})$ to quantify uncertainty of the strength of local evidence.

(5) Calculate $\mathrm{E}_L(M_{R,} M_A; \hat{g}_{n,\underline{x}}, \underline{x})$ as the expectation (mean) of the estimated density from step 4

We find it remarkable that a non-parametric bootstrap can be used to quantify local/conditional/post-data uncertainty. We explained how this occurs in definitions 23 and 24 in Box 4, but the point is important enough that we reiterate here in the comparison of bootstrap algorithms. The key is to realize that for estimated models the data are used in two fashions: first to estimate the parameters for each of the models, and second to calculate the strength of evidence for one model over another. Compare step 3 of the global and local bootstraps. The global bootstrap generates a large number of alternative data sets and for each iteration uses the same bootstrapped data to both estimate the models and calculate the evidence. On the other hand, the local bootstrap while also bootstrapping the data and reestimating models based on the bootstrapped data, only uses the *original data* for calculating the evidence. There is a relevant subset involved. It is the original data. Thus, as we say in the paper, the local bootstrap represents uncertainty in evidence due to uncertainty in model estimation and does not include sampling variation.

the observed data appropriately. Even if the researcher happens to have good, dispersed covariates, the global variance does not recognize that happy event and increases the variance because the researcher, in another replication of the experiment could have observed less dispersed covariates and vice versa. We note that the pairwise resampling used in bootstrap inference for regression gives the unconditional variance and is robust against mean as well as error structure misspecification. On the other hand regression bootstrap based on residuals provides conditional inference but is only robust against error model misspecification (Efron and Tibshirani, 1993).

For local uncertainty, the sample space over which the variation is considered is a subset of the total sample space. This is called a "relevant subset" (Buehler, 1959). Such a relevant subset is often determined using an ancillary statistic. An ancillary statistic is a function of the data whose distribution does not depend on the parameters. There are, often, multiple ancillary statistics (Basu, 1964; Pena et al., 1992) and hence relevant subsets are not necessarily unique. In our opinion, the appropriateness of the relevant subset is determined based on the type of future experimental replication one envisions. Different future experiments determine different relevant subsets as was the case in the Mark-Capture-Recapture example in **Box 2**.

It has been argued that local (post-experiment, post-data, conditional) confidence intervals are preferable as the measure of uncertainty because they reflect the informativeness of the data at hand appropriately. If the data are highly informative, the local confidence intervals are shorter than the global confidence intervals and if the data are not informative, the local confidence intervals appropriately are wider than the global confidence intervals. Again, this argument hinges on the model being correctly specified.

Some august statisticians (e.g., Royall, 2004) argue the local interval is the only one that should be used irrespective of the design because design is an ancillary statistic and has no impact on the inference once the data are obtained. If the data are highly informative either by design or by chance, we should be quite confident about our estimate of the total population size, irrespective of what other experimenters might observe. It can be shown (see review in Lele, 2020b) that prediction of a new observation based on local uncertainty is more accurate than prediction based on global uncertainty. However, this result also depends on correct model specification.

On the other hand, other equally august statisticians (e.g., Cox, 2004 in his discussion of Royall, 2004) claim design should play a role in uncertainty quantification. We agree with this

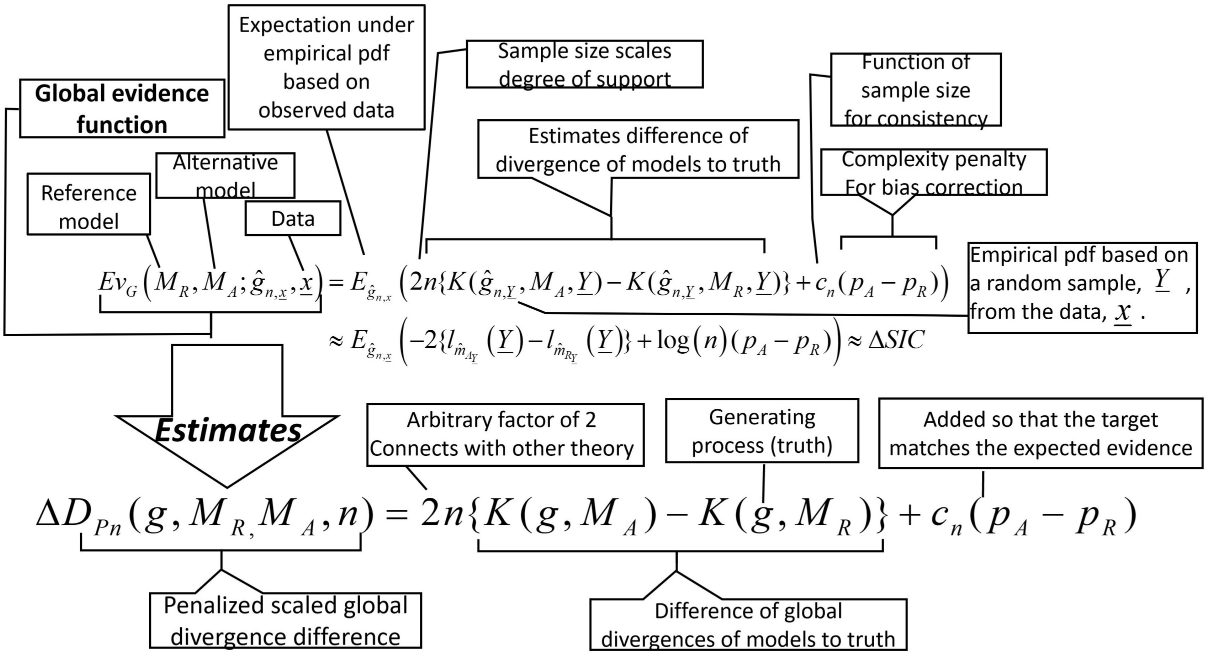

Global evidence function
Expectation under empirical pdf based on observed data
Sample size scales degree of support
Function of sample size for consistency
Estimates difference of divergence of models to truth
Complexity penalty For bias correction
Reference model
Alternative model
Data
Empirical pdf based on a random sample, $\underline{Y}$, from the data, $\underline{x}$.
$Ev_G\left(M_R, M_A; \hat{g}_{n,\underline{x}}, \underline{x}\right) = E_{\hat{g}_{n,\underline{x}}}\left(2n\{K(\hat{g}_{n,\underline{Y}}, M_A, \underline{Y}) - K(\hat{g}_{n,\underline{Y}}, M_R, \underline{Y})\} + c_n(p_A - p_R)\right)$
$\approx E_{\hat{g}_{n,\underline{x}}}\left(-2\{l_{\hat{m}_{A_{\underline{Y}}}}(\underline{Y}) - l_{\hat{m}_{R_{\underline{Y}}}}(\underline{Y})\} + \log(n)(p_A - p_R)\right) \approx \Delta SIC$
Estimates
Arbitrary factor of 2 Connects with other theory
Generating process (truth)
Added so that the target matches the expected evidence
$\Delta D_{Pn}(g, M_{R,} M_A, n) = 2n\{K(g, M_A) - K(g, M_R)\} + c_n(p_A - p_R)$
Penalized scaled global divergence difference
Difference of global divergences of models to truth

FIGURE 5 | A schematic indicating how an evidence function relates to its global target. See supplement for a similar figure for local evidence. The principal differences between the figures are that for global evidence the target does not (and must not) depend on the data while for local evidence the target does (and must). Reflecting this difference, the global evidence function resamples the observed data to calculate likelihoods while the local does not.

latter opinion on the importance of design. Both because the interpretation of uncertainty intervals should depend on the potential type of the future experimental replication, and thus so should the choice of the ancillary statistics or relevant subsets. And because, as we show in Section 5.2, the accuracy of the local interval depends on correct model specification to a greater degree than does the global.

4.2.4. Local Uncertainty When Comparing Two Model Spaces

Local evidence uncertainty in the comparison of model spaces is calculated similarly to global evidence uncertainty. Data sets are repeatedly reconstructed by bootstrapping the original data. With each bootstrapped, data set model parameters for both reference and alternative models are reestimated, and an evidence value comparing the models is calculated. The critical distinction between global and local uncertainty is that in the local calculations the likelihood for each bootstrapped model is evaluated using the original data not the bootstrapped data (see **Box 6** for an explicit algorithm).

In Section 5, we use simulations to study the coverage properties of the global and local sampling distributions. Both the cases of linear regression and structural equation models are investigated.

5. SIMULATION VALIDATION

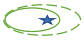
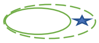
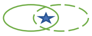
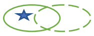
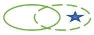
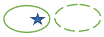
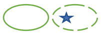
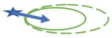
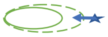
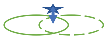
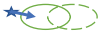
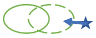
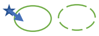
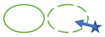
If new statistical approaches are proposed, the scientific community has a legitimate expectation that they will be validated both mathematically, and computationally (Devezer et al., 2021). For a procedure that generates confidence intervals, whether global or local, to be a legitimate frequentist procedure, they need to cover/capture their targets at least at the specified level (Casella, 1992). The fundamental difference between global and local inference is that a global target cannot depend on the data at hand, while a local target must depend on the data at hand.

Globally we want our intervals to cover the global penalized scaled divergence difference: $\Delta D_{Pn}(g_{n,\underline{X}}, M_{R,} M_A, \underline{X}) = E_{g_{n,\underline{X}}}\left(2n\{K(g_{n,\underline{Y}}, M_A, \underline{Y}) - K(g_{n,\underline{Y}}, M_R, \underline{Y})\} + c_n(p_A - p_R)\right)$. Locally we want our intervals to cover the local penalized scaled divergence difference: $\Delta d_{Pn}(g_{n,\underline{X}}, M_{R,} M_A, \underline{x}) = E_{g_{n,\underline{X}}}\left(2n\{K(g_{n,\underline{Y}}, M_A, \underline{x}) - K(g_{n,\underline{Y}}, M_R, \underline{x})\} + c_n(p_A - p_R)\right)$. For the Kullback-Leibler divergence, this is approximately $-2\{l_{\hat{m}_{A_{\underline{x}_b}}}(\underline{x}) - l_{\hat{m}_{R_{\underline{x}_b}}}(\underline{x})\} + c_n(p_A - p_R)$, the penalized scaled LLR for the observed data under the best approximating models in the two competing spaces to the true generating mechanism. We note this is identical to what is considered the target likelihood in the general profile likelihood literature, e.g., Section 3.1 of Pace and Salvan (2006).

5.1. Global and Local Coverages in Alternate Model Space Topologies

There are 14 possible topologies for a reference model space, an alternative model space and a generating process. The model spaces compared can be nested, overlapping, or disjoint. If the model comparison is correctly specified, the generating process will be in at least one of the model spaces. If the comparison is misspecified then the generating process will be in neither

TABLE 3 | The behavior of our global and local uncertainty procedures in all 14 possible model specification topologies.

Case	g location	Asymptotic distribution	Exemplar	G par	Global coverage	Global length mean (SD)	Local coverage	Local length mean (SD)
					95%/90%	95%/90%	95%/90%	95%/90%
1		Chi-square	$Ev(g = m_{001}, M_R = M_{001}, M_A = M_{011}; x)$	0.00 0.00 0.15	0.00 0.00	8.18 (3.67) 6.48 (3.16)	0.99 0.97	6.66 (1.17) 4.90 (0.86)
2		Non-central chi-square	$Ev(g = m_{011}, M_R = M_{001}, M_A = M_{011}; x)$	0.00 0.30 0.15	0.95 0.88	22.79 (7.42) 19.06 (6.29)	0.98 0.95	8.12 (1.39) 6.03 (1.00)
3		Weighted sum of chi-square	$Ev(g = m_{010}, M_R = M_{110}, M_A = M_{011}; x)$	0.00 0.30 0.00	1.00 1.00	13.75 (4.03) 10.58 (3.41)	0.99 0.97	10.94 (1.37) 7.84 (0.95)
4		Normal	$Ev(g = m_{110}, M_R = M_{110}, M_A = M_{011}; x)$	0.60 0.30 0.00	0.95 0.90	44.89 (5.91) 37.62 (4.97)	0.98 0.94	13.29 (1.4) 9.90 (0.97)
5		Normal	$Ev(g = m_{011}, M_R = M_{110}, M_A = M_{011}; x)$	0.00 0.30 0.15	0.98 0.93	18.23 (5.71) 14.51 (4.96)	0.98 0.96	11.12 (1.35) 8.02 (0.95)
6		Normal	$Ev(g = m_{110}, M_R = M_{110}, M_A = M_{001}; x)$	0.60 0.30 0.00	0.96 0.92	48.00 (5.53) 40.25 (4.66)	0.97 0.93	14.23 (1.42) 10.69 (1.01)
7		Normal	$Ev(g = m_{001}, M_R = M_{110}, M_A = M_{001}; x)$	0.00 0.00 0.15	0.95 0.85	20.67 (5.55) 16.56 (4.78)	0.99 0.96	12.9 (1.36) 9.53 (0.98)
8		Weighted sum of chi-square	$Ev(g = m_{111}, M_R = M_{001}, M_A = M_{011}; x)$	0.05 0.05 0.15	0.00 0.00	8.11 (3.58) 6.42 (3.09)	0.97 0.93	6.73 (1.18) 4.95 (0.85)
9		Normal	$Ev(g = m_{111}, M_R = M_{001}, M_A = M_{011}; x)$	0.05 0.30 0.15	0.94 0.88	22.73 (7.12) 19.01 (6.02)	0.99 0.97	8.08 (1.36) 6.01 (0.99)
10		Weighted sum of chi-square	$Ev(g = m_{111}, M_R = M_{110}, M_A = M_{011}; x)$	0.05 0.30 0.05	0.99 0.98	15.1 (4.69) 11.75 (4.02)	0.97 0.93	10.92 (1.41) 7.84 (0.94)
11		Normal	$Ev(g = m_{111}, M_R = M_{110}, M_A = M_{011}; x)$	0.60 0.30 0.05	0.96 0.90	45.47 (6.09) 38.09 (5.12)	0.99 0.97	13.38 (1.42) 9.98 (1.01)
12		Normal	$Ev(g = m_{111}, M_R = M_{110}, M_A = M_{011}; x)$	0.05 0.30 0.15	0.99 0.96	18.98 (5.88) 15.14 (5.09)	0.98 0.94	11.08 (1.37) 8.01 (0.98)
13		Normal	$Ev(g = m_{111}, M_R = M_{110}, M_A = M_{001}; x)$	0.60 0.30 0.05	0.95 0.92	49.05 (5.9) 41.1 (4.97)	0.98 0.96	14.33 (1.44) 10.77 (1.01)
14		Normal	$Ev(g = m_{111}, M_R = M_{110}, M_A = M_{001}; x)$	0.05 0.05 0.15	0.95 0.88	22.55 (5.6) 18.18 (4.8)	0.97 0.94	12.93 (1.36) 9.56 (0.95)

In the g location figures the solid ellipse indicates the reference model space while dashed ellipse indicate the alternative model space. For the correctly specified comparisons, cases 1–7, the star indicates the location of the generating process. For the misspecified comparisons, the arrow indicates the location of the projection from the generating process to the model spaces. The asymptotic distribution refers to the unpenalized likelihood ratio statistic (often denoted G2); the penalty term for converting G2 to an evidence function produces location-shifted versions of the asymptotic distributions (Dennis et al., 2019). The covariates are three N(0,1) random vectors and are held constant over all simulations. For each line, the coefficients $(\beta_1, \beta_2, \beta_3)$in the generating model of the three covariates (there are no interactions) are given in the column g par. In all simulations the intercept is 2.0 and the error standard deviation is 1. The sample size for all simulations in this table is 100, a realistic size for ecological studies, and one that meets most common rules of thumb for multiple regression. Coverage proportions were estimated using 1,000 trials for each case. Coverage is reported for nominal 95 and 90% kde1d intervals. Mean interval length and its standard deviation is also reported.

model space. **Table 3** describes coverage and interval length for the global and local confidence intervals of the strength of evidence for model comparisons in each of these topologies in a simple multiple regression example (see the table legend for simulation details).

A number of interesting patterns can be observed in **Table 3**. In 12 of the 14 possible model space topologies, the global intervals cover reasonably, with actual coverages close to nominal coverages. Cases 1 and 8, however, have no coverage! Case 1 is the topology of nested models with the generating process in the reduced model. The asymptotic distribution for this case is chi-square. Case 8 represents the misspecified analog of Case 1, the approximating models are nested with the generating process closest to the reduced model. The asymptotic distribution for case 8 is a weighted sum of chi-square. This is a very flexible distribution, and in this case generates a distribution indistinguishable from a chi-square distribution. Alarm at this complete lack of coverage in these two cases is somewhat reduced by recognizing that the target $(\Delta D_{Pn}(g, M_{R,} M_A, \underline{X}))$ is the boundary of these chi-square distributions and hence impossible to capture with finite sampling.

On the other hand, the local confidence intervals for evidence behave well in all 14 possible model space topologies. In all cases

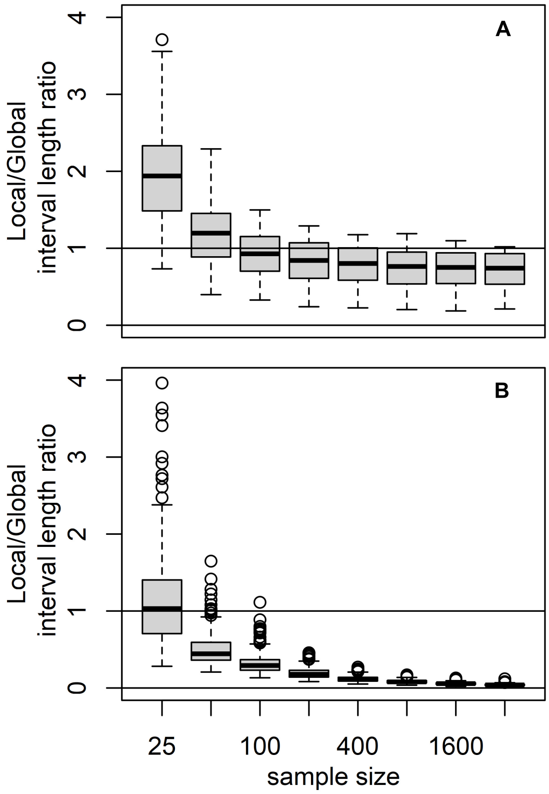


FIGURE 6 | Box plots of the ratios of the local and global interval lengths as a function sample size. Each box summarizes the results for 1,000 simulations. In **(A)**, all parameters (except for sample size) are set to those case 1 in **Table 3**. In **(B)**, all parameters (except for sample size) are set to those of case 4 in **Table 3**.

local interval coverage exceeds the nominal levels. Overcovering is acceptable in approximate confidence intervals, particularly if interval length is narrow. In all cases of **Table 1**, the average lengths of the local intervals are less than that of global intervals. This is not always the case. For very small sample size, the average local interval length may exceed the average global interval length (see **Figure 6**).

5.2. Sample Size and Interval Lengths

In the linear models example of **Table 3**, global and local intervals respond quite differently to changes in sample size. These differences are explored in **Figures 6**, **7**. **Figure 6A** shows box plots of the ratio of local interval length to global interval length over a range of increasing sample size for the case of case 1 from **Table 1**. The models compared are nested and the generating process is in the reduced model. The asymptotic distribution of evidence is chi-squared. At lower sample sizes the local interval length generally exceeds the global length. At higher sample sizes the local interval is generally shorter than the global interval, with the ratio appearing to approach a limit of at about 0.6. Model topologies shown in case 1 and 8 of **Table 3** behave in this fashion.

Figure 6B represents case 4 from **Table 1**. The models compared are overlapping with the generating process located in the non-overlapping portion of the reference model. The interval length behavior here is very different from that in panel A. Local intervals exceed global intervals only at the smallest sample sizes. Further, the local/global interval length ratio rapidly decreases toward 0 (rate $1\big/\sqrt{n}$). All model topologies except those of cases 1 and 8 behave in this fashion.

For both global and local evidence, the expectation grows linearly with sample size. The standard deviation in global evidence grows as the square root of sample size. On the other hand, the standard deviation in local evidence approaches a constant as sample size increases (Kitagawa and Konishi, 2010). These differences have considerable impact on inference and experimental design.

The ability of the global interval to distinguish the observed evidence from 0 grows very slowly with sample size. On the other hand, the local interval will be able to detect real difference from 0 or either of our two thresholds with relatively small sample sizes. Nevertheless, in both global and local cases, the coefficient of variation in evidence goes to 0 as sample size grows to infinity.

5.3. Model Set Misspecification and Evidential Uncertainty

Here we demonstrate the effect of model set misspecification on the uncertainty of evidence with simulations based on the Grace and Keeley example. We look at four different conditions of model set adequacy: (A) correctly specified comparison with very strong evidence, (B) correctly specified comparison with strong evidence, (C) a mildly misspecified model comparison, and (D) a badly misspecified model comparison.

In case A), we compare the model that is the GKBM without the weakest path (GKBM – R∼L) with a model that is the GKBM without the second weakest path (GKBM – R∼C). The data in these simulations are generated from the estimated (GKBM – R∼L). The generating process is in the compared model set; therefore, the comparison is correctly specified.

In case B) we estimate and compare the same models as in case A. The generating model has same form as in case A (all the same paths are present) but one of the coefficients (R∼P) has been weakened from 0.299 to 0.205. The model set is still correctly specified, but the penalize divergence differences (whether global or local, see definitions 19 and 20) between the compared models is less than in case A. Consequently, the distribution of realized evidences (definitions 23 and 24) will be shifted to lower values.

Case C) compares the same models as in case A) {GKBM – R∼L, GKBM – R∼C}. The data are generated by the GKBM. Since the generating process (GKBM) is quite close to one of the models in the model set (GKBM – R∼L), the comparison is only mildly misspecified.

Finally, in case D) we compare a model that is the GKBM without the second strongest path (GKBM – C ∼ L) with a model that is the GKBM without the strongest path (GKBM – F ∼ A). As in case B), the data are generated by the GKBM. Since the generating process (GKBM) is quite different from both of the models in the model set, the comparison is badly misspecified.

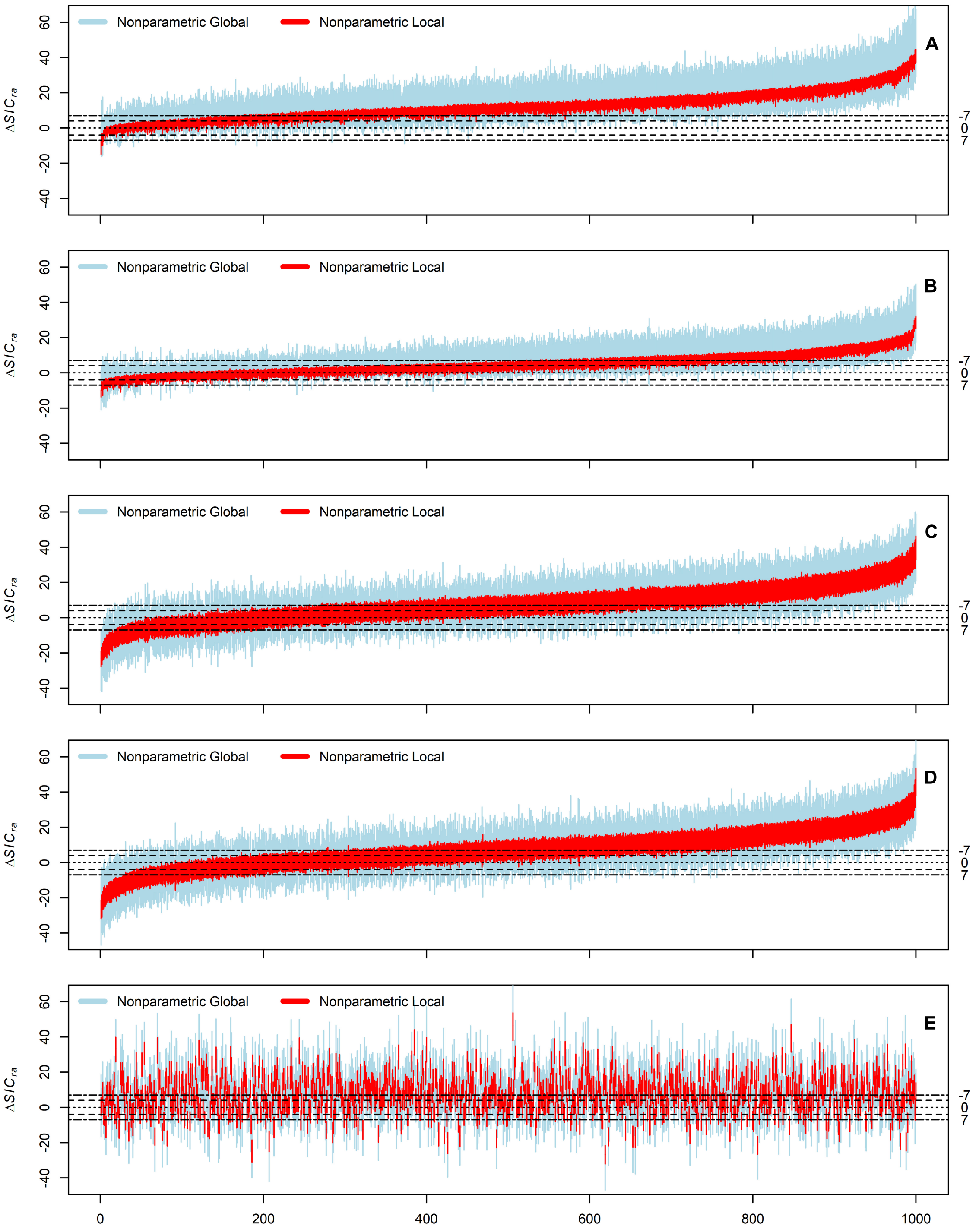


FIGURE 7 | Global (blue) and local (red) 90% confidence intervals for the 1,000 simulations for cases **(A–D)** Described in the text. In **(A–D)** simulations are ordered by mean smoothed local evidence. Panel **(E)** presents the same data as **(D)** but ordered by the raw evidence. **(A–C)** Not shown reordered because with a Spearman correlation of ≥0.997 between raw and mean smoothed evidence in these cases, there is no perceptible change in the figures.

Table 4 indicates that, at least in this example, under correct model specification, a researcher is very unlikely to obtain secure misleading evidence using either interval. On the other hand, the researcher is more likely to correctly obtain strong and secure evidence using the conditional interval than with the unconditional interval. If the model set is misspecified, secure misleading evidence becomes a possibility, and much more so using the conditional interval than the unconditional interval.

Interestingly, the average reliability (proportion of the time correct model is identified) is always slightly greater using the local evidence distribution rather than when using the global evidence distribution. This agrees with the previous results (Aitchison, 1975; Royall and Cumberland, 1985; Vidoni, 1995) that indicate predictive accuracy is greater using conditional inference.

The table gives the impression that there is little difference between mildly and badly misspecified model sets regarding evidence. But this is only because the choice of the mean of the smoothed bootstrapped ΔSIC as the measure of the strength of evidence rather than the raw ΔSIC has profound impact. **Figure 7** presents the same data used to calculate **Table 4** in another fashion. Here both the global and local intervals are explicitly plotted for each 1000 trials in the simulations of cases A, B, C, and D. The trials are sorted along the x axis by the mean smoothed bootstrapped strength of evidence. Panel E plots the same simulations and intervals as panel D, however, in this case the trials are sorted by the raw ΔSIC—not by the mean smoothed bootstrapped ΔSIC. We do not show plots with similar reordering for panels A, B, and C because in these cases the differences between the raw ΔSIC and the mean of the smoothed bootstrap are not visually perceptible.

In cases A, B, and C the difference between raw ΔSIC and smoothed mean bootstrapped ΔSIC are quite small and the correlation of raw ΔSIC and mean smoothed bootstrapped ΔSIC are greater than 0.99. Thus, there is almost no impact of choice of evidence measure in these cases with correct and mild misspecification. In the badly misspecified case D, there is a large average difference between raw ΔSIC and the mean smoothed bootstrapped ΔSIC and almost no correlation between them. Further, when using the raw ΔSIC, the location of the security intervals becomes almost unrelated to the strength of evidence. Consequently, the raw ΔSIC has almost no ability to securely identify the best model.

6. DISCUSSION

Historically, the appeal of classical Neyman-Pearson testing has been the appearance of a strong control of error probabilities. Dennis et al. (2019) show this apparent control to be an illusion for the great majority of cases of interest in ecological science where models are misspecified. Under model misspecification, the realized error rate for a NP test can be less than or greater than its nominal rate. In some realistic cases the probability of error in a NP test can even increase with increasing sample size. Evidential analysis is superior to NP testing in that the total error rate always decreases with increasing sample size, both under correct model specification and under model misspecification.

However, Dennis et al. (2019) further points out that evidence is not entirely immune to problems due to model misspecification. Under misspecification, the probability of strong misleading evidence is not directly calculable because the generating process is not one of the models compared and is not even known. This current paper demonstrates that evidential error rates can be estimated even under model misspecification using non-parametric bootstrapping techniques (at least for independent data). Our approach to the bootstrapping of evidence differs from that used in the EIC (Konishi and Kitagawa, 1996; Ishiguro et al., 1997) in that we bootstrap the evidential comparison as a unit (see definitions 23 and 24 **Box 4**) whereas the EIC compares bootstrapped components. The joint bootstrapping allows us to estimate the impact of model set misspecification on evidential uncertainty more effectively. In this paper, we have only addressed the case of independently distributed data. We expect, however, that this approach can be extended to other data structures with the use of subtler bootstrapping methods (Lele, 1991, 2003; Lahiri, 2003).

It is important for scientists seeking to use and interpret these measures of uncertainty to understand the two intervals, global and local, are quantifying two different kinds of uncertainty. Statistical evidence is an estimate of the relationship between two models and the generating process. It is a penalized sample size scaled estimate of the difference of the divergences of two models

TABLE 4 | Models compared and generating process for each model set are described in the text.

Case	Model set adequacy	Interval type	Evidential security categories									Average reliability
			MS	CS	MI	CI	W	PI	SI	PS	SS	
A	Correctly specified	Global	0	0	0.001	0	0.069	0.108	0.447	0	0.375	0.944
		Local	0	0	0	0.001	0.069	0.105	0.101	0	0.724	0.975
B	Correctly specified	Global	0	0	0.001	0.003	0.345	0.195	0.371	0	0.085	0.834
		Local	0.001	0	0	0.003	0.336	0.202	0.152	0	0.306	0.877
C	Mildly mis-specified	Global	0.003	0	0.042	0.066	0.260	0.140	0.390	0	0.099	0.720
		Local	0.034	0	0.012	0.063	0.256	0.148	0.126	0	0.361	0.775
D	Badly mis-specified	Global	0.003	0	0.068	0.050	0.261	0.114	0.400	0	0.104	0.711
		Local	0.046	0	0.025	0.050	0.260	0.115	0.137	0	0.367	0.761

The bootstrap mean evidence is used as the strength of evidence. Each row lists the proportions each security category occurs in 1,000 simulations and the overall reliability. Security in each row is determined either by the unconditional evidential confidence intervals or the conditional evidential confidence intervals. The categories of security are: MS, misleading and secure; CS, confusing and secure; MI, misleading and insecure; W, weak; PI, prognostic and insecure); SI, strong and insecure; PS, prognostic and secure; SS, strong and secure. Reliability is the proportion of times the best model is correctly identified—by any strength of evidence—averaged over all trials.

from the generating process (truth). Valid confidence intervals of an estimate tell us how confident we are that the estimation target lies within the interval. In the global case, our estimate of evidence is the mean of the global bootstrap distribution of evidence, but the estimation target is the true penalized scaled divergence difference (**Box 4**, definition 19). In the local case our estimate of evidence is the mean of the local bootstrap distribution of evidence, but the target is the true evidence in the data without model estimation error (**Box 4**, definition 20).

For badly misspecified model comparisons local inference has strong and secure but *misleading* evidence more often than global inference. Nevertheless, we are in a position to make scientific inferences about the true relationships of our compared models to the generating process, backed by an uncertainty measure warrant.

Both the global and local evidence confidence intervals are important to science because they answer different questions. The global interval is a confidence interval on the true penalized scaled divergence difference. This speaks directly to the relative ability of our models to represent nature. The resampling is non-parametric to accommodate model misspecification. Further, the intervals incorporate both sample and model estimation uncertainty.

The global uncertainty we offer answers the question of how dissimilar to the current evidence we would expect new evidence to be if our experiment were to be repeated. This is the interval that other researchers should consider when trying to decide if their new results call the current results into question.

On the other hand, the local uncertainty tells you how confident you are in your evidence given the data you have collected. This might be the interval to use if you intend to take an action based on the results.

Replication is often seen as a pillar of science as a social activity (e.g., Johnson, 2002). But, what to replicate and how to measure is not always clearly understood. Which interval should a scientist use? Unfortunately, a univocal recommendation is not possible. The local interval is tremendously appealing because it is so short and because its overall reliability is greater (see **Table 4**). However, to justify inference based on it alone, the scientist needs to be able to defend the assumption of approximately correct model set specification. In the rough and tumble world of ecology this will rarely be possible, except for tightly controlled experiments with well understood error structures. The global interval presents an appraisal of the replicability of the scientist's results. If the global interval has been presented, the local interval can be a useful indication of how good the results could possibly be. For the accumulation of understanding through science, the global interval may be preferable. This preference is grounded in our opening quote from Plato. Using the global interval, you will accept wrong statements less frequently than when using the local interval. However, in a decision context, where costs and benefits are explicit, the local inference's property of making correct predictions more often than global inference might be important.

Hopefully, our recommendation to focus on the global interval will be only temporary. We expect that often model sets could be misspecified, but close enough to correctly specified that the local interval would be a justifiable improvement over the global interval. Research into diagnostics to identify these cases is called for (Cook and Weisberg, 1982). Useful diagnostics will involve more than measures of the adequacy of single models (e.g., Markatou and Sofikitou, 2019) they must somehow include measures of the geometry of the generating process and the competing models (Dennis et al., 2019; Ponciano and Taper, 2019).

In the meantime, little is practically lost. We agree with Goutis and Casella (1995) that "In any experiment both pre-data inferences and post-data inferences are important." Our inferential strategy is a hybrid of local and global (conditional and unconditional). Our primary tool is the strength of evidence, which is local (i.e., conditional). The evidence expresses clearly what the data we have says about the relationships among nature and our models. Our secondary tools are our pair of measures of the security of the evidence. If we choose a global (that is unconditional measure) we gain an honest, if perhaps overly conservative, insight into the degree that chance, experimental/sample design, and model misspecification may have influenced our evidence. If we choose a local (that is a conditional measure) we gain a more precise understanding of the information in the data, at the risk of overconfidence due to model misspecification. Much of statistics both classical and Bayesian relies on conditional inference and thus might be over-confident in its conclusions in the face of potential model misspecification (see also Yang and Zhu, 2018).

While the global uncertainty, either calculated from asymptotic theory or from the non-parametric bootstrap is a useful statistic, it should not be interpreted too literally. As Fisher (1945a,b, 1955, 1956, 1960) long argued (see Rubin, 2020; Devezer et al., 2021 for detailed discussions) an exact repetition of an experiment is not possible in many branches of science. Certainly, this is true in ecology and environmental science, where heterogeneity and temporal data abound. To paraphrase Heraclitus, you can't electrofish the same river twice. A more realistic understanding of global uncertainty would come from a metanalysis of the actual repetition of modestly sized experiments distributed in space and time than from a single large experiment. As an example, Jerde et al. (2019) conducts an evidential comparison of models for the intra-specific allometry of metabolic rate in fish using a database of 25 high quality studies, with 55 independent trials, across 16 fish species.

Jerde et al. (2019) use evidential support intervals in their analysis of the allometry of metabolic rate in fish. These intervals are post-data/conditional/local intervals. We wish to point out that, while both are useful, evidential support intervals and confidence intervals for evidence are different. Evidential support intervals indicate the range of parameter values in a model space that are not differentiated from the best estimate at a specified strength of evidence. Confidence intervals make a statement that at the specified probability a random interval, whose randomness stems from sample space probabilities, contains the true parameter value (Dennis, 2004). Under correct model specification, the support interval indicates over what range of parameter values the relative plausibility of the best estimate relative to the parameter value is less than the designated strength of evidence.

Under a correct model assumption, a ΔAIC interval is directly transformable into a confidence interval on the strength of evidence using Wilks-Wald hypothesis test inversion (see Dennis et al., 2019). The confidence level of this transformed interval will depend only on the chosen strong evidence threshold, k_{R}. On the other hand, the level of an evidence confidence interval corresponding to a ΔSIC interval will be a function of both k_{R} and $\log(n)$. As n increases the confidence level will increase. This parametric confidence interval is preferred if a true model assumption is justified. Using a nonparametric confidence interval rather than a evidence interval acknowledges that your model set may be misspecified.

Global and local confidence intervals for the strength of evidence, at least as we have developed them in this paper, are used in the comparisons of model spaces. The intervals discussed are on the space of strength of evidence values; they are not on the space of parameter values nor are they on the space of predictions. We have seen that the interpretation of local and global intervals on evidence requires deep consideration of the scientific questions being asked. Complexities also arise for conditional and unconditional intervals for parameters and for predictions. We defer to another paper a unified discussion of the effects of post-data and pre-data intervals on science.

As laid out in Royall (1997) and Dennis et al. (2019), one of the great strengths of the evidential approach relative to NPHT, is that M, the probability of misleading evidence goes to 0 as sample size increases whereas α, the corresponding uncertainty measure in NPHT, remains constant. It is a commonplace in introductory mathematical statistics courses that hypothesis tests and confidence intervals are inter-convertible. Given this, a reasonable question to ask is: By incorporating confidence intervals have we somehow given up the superior error structure of evidence? The answer to this question is no. The NPHT freights both its measure of the strength of evidence and its measure of uncertainty onto α. We use 2 measures; the primary is evidence and as sample size increases this will go to either $+\infty$ or $-\infty$. Our second measure is the standard deviation of evidence and, as discussed in Section 5.2, this does not grow as rapidly as the evidence itself. As a consequence, the probability of making an error of assignment—at any specified level of confidence—also goes to 0 as sample size increases.

A literature has developed that constructs confidence sets (i.e., confidence intervals on discrete parameters) in model identification (Hansen et al., 2011; Ferrari and Yang, 2015; Sayyareh, 2017; Li et al., 2019; Zheng et al., 2019; Liu et al., 2021). These papers differ from the current work in several important fashions. First, the confidence intervals being considered are not even related. Our work constructs a confidence interval on a continuous parameter, the strength of evidence between models. The parameter in the model confidence sets literature is a discrete parameter of model inclusion. Second, one feature of the confidence set approach is that specification of the entire model set is essential to interpretation of a confidence set. This is a drawback that is shared by Bayesian model selection and model averaging. Our worked example in Section 3 makes 14 evidential comparisons. Should some sort of adjustment be made? If the analyst is willing to specify the model set, multiple comparison adjustments are appropriate in evidential comparisons, particularly when massive numbers of comparisons or badly misspecified model sets are involved. Fortunately, there are several features of the evidential paradigm that allow it to respond to multiple comparisons with more grace and less cost than classical hypothesis testing approaches. Evidential multiple comparisons have been extensively discussed in Strug and Hodge (2006a,b) and Taper and Lele (2011). These reviews were written from the standpoint of correctly specified model sets with the probability of misleading evidence being estimated by Royall's universal bound (Royall, 1997). We hope to soon write a paper on evidential multiple comparisons that utilizes the ability of our non-parametric bootstrap to estimate the probability of misleading evidence in the face of model misspecification (Taper et al., 2019; Liu et al., 2021).

Another attribute of the model confidence set papers is that they all make their selections based on some form of NPHT. We suspect that these confidence sets inherit the stringent properties of multiple comparisons in NPHTs rather than the more permissive properties of evidential multiple comparison. We look forward to investigating this in more detail in the future.

Due to limitations of space, the topic of this paper is treated strictly as a development of evidentialist statistics using a frequentist notion of probability. When epistemic comparisons are made, they are to NPHT. Readers interested in better understanding the relative epistemic character of evidential statistics, error statistics (classical hypothesis testing), and Bayesian statistics might explore some of Dennis (2004), Lele (2004a,b, 2010, 2020a), Taper and Lele (2004), Efron (2005), Lele and Allen (2006), Lele et al. (2007, 2010), Lele and Dennis (2009), Ponciano et al. (2009, 2012), Bandyopadhyay and Forster (2011), Bandyopadhyay et al. (2016), Taper and Ponciano (2016), Mayo (2018), and Brittan and Bandyopadhyay (2019) as examples of a vast battleground of literature on the topic.

7. CONCLUSION

Neither the Bayesian nor classical frequentist statistical toolkits appear adequate for the increasingly complex challenges of the future. In the long run, neither our models nor our data, nor our conclusions are static. We need to look at multiple models realizing that we do not know truth and evolve these models toward better approximations of truth with the accumulation of data and use of evidence as a selection function.

We have produced both global and local uncertainty measures that are easily calculated for many analyses using the R-code that we supply in **Supplementary Material**. Further, by creating three categories for the strength of evidence coupled with three categories for the security of evidence we have constructed a conceptual language that allows scientists a statistically valid way to talk, and publish, about interesting results that are not yet conclusive.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MT wrote the R Code. MT and SL jointly wrote the first draft. All authors contributed to the many draft revisions and conceived of this study jointly.

FUNDING

MT, SL, BD, and JP were all partially funded by a grant from the Japan Society for the Promotion of Science to Professor Ichiro Ken Shimatani. JP was partially supported by two grants from the US National Institute of Health grant number to University of Florida NIH R01 GM103604 to JP and NIH 1R01GM117617 to Profs. Jason K. Blackburn and J. M. Ponciano. The participation of CJ in this research was partially funded by USAID (AID-OAA-A-00057), NASA (NNX14AR62A), BOEM (MC15AC00006), and NOAA's support of the Santa Barbara Channel Marine Biodiversity Observation Network.

ACKNOWLEDGMENTS

Dr. James B. Grace kindly contributed the original data set from Grace and Keeley (2006). Drs. Prasanta Bandyopadhyay, James B. Grace, Jake M. Ferguson, and David Staples all read the manuscript carefully and provided many helpful discussions and comments. The editor, IS, and three reviewers provided comments that greatly improved the manuscript.

SUPPLEMENTARY MATERIAL

The files KKICv.R, KKICv.demo.R, GK.d.rds, and README.md can be found at https://github.com/jmponciano/mltaper-bootstrap.

Supplementary Figure 1 | A schematic indicating how a local evidence function relates to its target. See the manuscript body for a similar figure for global evidence. The principal differences between the figures are that for global evidence the target does not (and must not) depend on the data while for local evidence the target does (and must). Reflecting this difference, the global evidence function resamples the observed data to calculate likelihoods while the local does not.

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* 62, 547–554.

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. N. Petrov, and F. Csaki (Budapest: Akademiai Kiado).

Anderson, D. R. (2008). *Model Based Inference in the Life Sciences: a Primer on Evidence*. Berlin: Springer Science & Business Media.

Bandyopadhyay, P. S., Brittan, G., and Taper, M. L. (2016). *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference*. Berlin: Springer.

Bandyopadhyay, P. S., and Forster, M. R. (eds) (2011). *Philosophy of Statistics*. Amsterdam: Elsevier.

Barnard, G. A. (1949). Statistical inference. *J. R. Statist. Soc. Series B-Statistical Methodol.* 11, 115–149.

Basu, D. (1964). Recovery of ancillary information. *Sankhya* 26, 3–16.

Birnbaum, A. (1962). On foundations of statistical-inference. *J. Am. Statist. Assoc.* 57, 269–306.

Birnbaum, A. (1970). Statistical methods in scientific inference. *Nature* 225:1033.

Birnbaum, A. (1972). More on concepts of statistical evidence. *J. Am. Statist. Assoc.* 67, 858–861.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.

Bollen, K. A., and Pearl, J. (2013). "Eight myths about causality and structural equation models," in *Handbook of causal analysis for social research*, ed. L. Morgan Stephen (Dordrecht: Springer).

Breitsohl, H. (2019). Beyond ANOVA: an introduction to structural equation models for experimental designs. *Organ. Res. Methods* 22, 649–677. doi: 10.1016/j.addbeh.2018.08.030

Brittan, G., and Bandyopadhyay, P. S. (2019). Ecology, evidence, and objectivity: in search of a bias-free methodology. *Front. Ecol. Evol.* 7:399. doi: 10.3389/fevo.2019.00399

Bruckheimer, J., and Verbinski, G. (2003). *Pirates of the Caribbean: The Curse of the Black Pearl*. Burbank, CA: Walt Disney Pictures.

Buehler, R. 'J. (1959). Some validity criteria for statistical inferences. *Ann. Mathematical Statist.* 30, 845–863.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*, 2nd Edn. New York, NY: Springer-Verlag.

Casella, G. (1992). Conditional inference from confidence sets. *Lecture Notes-Monograph Series* 17, 1–12.

Casella, G., and Berger, R. L. (2002). *Statistical Inference*, 2nd Edn. Boston, MA: Cenage Learning.

Cheng, C. L., and Van Ness, J. W. (1999). *Statistical Regresion with Measurement Error*, 1st Edn. London: Arnold.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cooper, L. N., Lee, A. H., Taper, M. L., and Horner, J. R. (2008). Relative growth rates of predator and prey dinosaurs reflect effects of predation. *Proc. R. Soc. B-Biol. Sci.* 275, 2609–2615. doi: 10.1098/rspb.2008.0912

Cox, D. R. (1958). *Planning of Experiments*. Oxford: Wiley.

Cox, D. R. (2004). "Commentary on the likelihood paradigm for statistical evidence by R. Royall," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: University of Chicago Press).

Cox, D. R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. Series B (Methodological)* 49, 1–39.

De Blasi, P., and Schweder, T. (2018). Confidence distributions from likelihoods by median bias correction. *J. Statist. Plann. Inference* 195, 35–46.

Dennis, B. (2004). "Statistics and the scientific method in ecology," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: The University of Chicago Press).

Dennis, B., Ponciano, J. M., Taper, M. L., and Lele, S. R. (2019). Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Front. Ecol. Evol.* 7:372. doi: 10.3389/fevo.2019.00372

Devezer, B., Navarro, D. J., Vandekerckhove, J., and Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *R. Soc. Open Sci.* 8:200805. doi: 10.1098/rsos.200805

Edwards, A. W. F. (1992). *Likelihood. Expanded Edition*. Cambridge: Cambridge University Press.

Efron, B. (2005). Bayesians, frequentists, and scientists [Editorial Material]. *J. Am. Statist. Assoc.* 100, 1–5. doi: 10.1198/016214505000000033

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Ferrari, D., and Yang, Y. H. (2015). Confidence sets for model selection by F-testing. *Statistica Sinica* 25, 1637–1658. doi: 10.5705/ss.2014.110

Fieberg, J. R., Vitense, K., and Johnson, D. H. (2020). Resampling-based methods for biologists. *Peerj* 8:e908.

Fisher, R. (1955). Statistical methods and scientific induction. *J. R. Statist. Soc. Series B-Statist. Methodol.* 17, 69–78.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London Series A* 222, 309–368.

Fisher, R. A. (1936). Uncertain inference. *Sci. Monthly* 43, 402–410.

Fisher, R. A. (1945a). A new test for 2X2 tables. *Nature* 156, 388–388.

Fisher, R. A. (1945b). The logical inversion of the notion of the random variable. *Sankhya* 7, 129–132.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.

Fisher, R. A. (1960). Scientific thought and the refinement of human reasoning. *J. Operat. Res. Soc. Japan* 3, 1–10.

Geenens, G., and Wang, C. (2018). Local-Likelihood transformation kernel density estimation for positive random variables. *J. Computat. Graph. Statist.* 27, 822–835.

Godambe, V. P. (1960). An optimum property of regular maximum-likelihood estimation. *Ann. Mathematical Stat.* 31, 1208–1211.

Goutis, C., and Casella, G. (1995). Frequentist post-data inference. *Int. Statist. Rev.* 63, 325–344. doi: 10.1890/13-1291.1

Grace, J. B. (2008). Structural equation modeling for observational studies. *J. Wildlife Manag.* 72, 14–22.

Grace, J. B., Anderson, T. M., Olff, H., and Scheiner, S. M. (2010). On the specification of structural equation models for ecological systems. *Ecol. Monographs* 80, 67–87.

Grace, J. B., and Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: the role of composite variables. *Environ. Ecol. Statist.* 15, 191–213.

Grace, J. B., and Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology* 101:e02962. doi: 10.1002/ecy.2962

Grace, J. B., and Keeley, J. E. (2006). A structural equation model analysis of postfire plant diversity in California shrublands. *Ecol. Appl.* 16, 503–514. doi: 10.1890/1051-0761(2006)016[0503:asemao]2.0.co;2

Grace, J. B., and Pugesek, B. H. (1997). A structural equation model of plant species richness and its application to a coastal wetland. *Am. Nat.* 149, 436–460.

Grace, J. B., Youngblood, A., and Scheiner, S. M. (2009). "Structural equation modeling and ecological experiments," in *Real World Ecology: Large-Scale and Long-Term Case Studies and Methods*, eds S. Miao, S. Carstenn, and M. Nungesser (Berlin: Springer Science).

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hall, P. (1986). On the bootstrap and confidence-intervals. *Ann. Statist.* 14, 1431–1452.

Hall, P. (1987). On the bootstrap and likelihood-based confidence-regions. *Biometrika* 74, 481–493.

Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* 15:20190174. doi: 10.1098/rsbl.2019.0174

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica* 79, 453–497. doi: 10.3982/ecta5771

Holland, S. M. (2019). Estimation, not significance. *Paleobiology* 45, 1–6.

Hurvich, C. M., and Tsai, C. L. (1989). Regression and time-series model selection in small samples. *Biometrika* 76, 297–307.

Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Institute Statist. Mathematics* 49, 411–434. doi: 10.1111/1541-0420.00020

Jerde, C. L., Kraskura, K., Eliason, E. J., Csik, S., Stier, A. C., and Taper, M. L. (2019). Strong evidence for an intraspecific metabolic scaling coefficient near 0.89 in fish. *Front. Physiol.* 10:1166. doi: 10.3389/fphys.2019.01166

Johnson, D. H. (1999). The insignificance of statistical significance testing. *J. Wildlife Manag.* 63, 763–772.

Johnson, D. H. (2002). The importance of replication in wildlife research. *J. Wildlife Manag.* 66, 919–932.

Keeley, J. E., Baer-Keeley, M., and Fotheringham, C. J. (2005). Alien plant dynamics following fire in mediterranean-climate California shrublands. *Ecol. Appl.* 15, 2109–2125.

Keeley, J. E., Brennan, T., and Pfaff, A. H. (2008). Fire severity and ecosytem responses following crown fires in California shrublands. *Ecol. Appl.* 18, 1530–1546. doi: 10.1890/07-0836.1

Kitagawa, G., and Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Ann. Institute Statistical Mathemat.* 62:209.

Konishi, S., and Kitagawa, G. (1996). GeneralisedGeneralized information criteria in model selection. *Biometrika* 83, 875–890.

Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York, NY: Springer.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. New York, NY: Springer.

Laughlin, D. C., and Grace, J. B. (2019). Discoveries and novel insights in ecology using structural equation modeling. *Ideas Ecol. Evol.* 12, 28–34.

Lele, S. (1991). Jackknifing linear estimating equations - asymptotic theory and applications in stochastic-processes. *J. R. Statist. Soc. Series B-Methodol.* 53, 253–267.

Lele, S. R. (2003). Impact of bootstrap on the estimating functions. *Statist. Sci.* 18, 185–190.

Lele, S. R. (2004a). "Elicit data, not prior: on using expert opinion in ecological studies," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: University of Chicago Press).

Lele, S. R. (2004b). "Evidence functions and the optimality of the law of likelihood," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: The University of Chicago Press).

Lele, S. R. (2010). Model complexity and information in the data: could it be a house built on sand? *Ecology* 91, 3493–3496. doi: 10.1890/10-0099.1

Lele, S. R. (2020a). Consequences of lack of parameterization invariance of non-informative Bayesian analysis for wildlife management: survival of San Joaquin kit fox and declines in amphibian populations. *Front. Ecol. Evol.* 7:501. doi: 10.3389/fevo.2019.00501

Lele, S. R. (2020b). How should we quantify uncertainty in statistical inference? *Front. Ecol. Evol.* 8:35. doi: 10.3389/fevo.2020.00035

Lele, S. R., and Allen, K. L. (2006). On using expert opinion in ecological analyses: a frequentist approach. *Environmetrics* 17, 683–704. doi: 10.1002/env.786

Lele, S. R., and Dennis, B. (2009). Bayesian methods for hierarchical models: are ecologists making a Faustian bargain? *Ecol. Appl.* 19, 581–584. doi: 10.1890/08-0549.1

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* 10, 551–563. doi: 10.1111/j.1461-0248.2007.01047.x

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *J. Am. Statist. Assoc.* 105, 1617–1625.

Lele, S. R., and Taper, M. L. (2012). "Information criteria in ecology," in *Encyclopedia of Theoretical Ecology*, eds A. Hastings and L. Gross (Berkeley: University of California Press).

Li, Y., Luo, Y. T., Ferrari, D., Hu, X. N., and Qin, Y. C. (2019). Model confidence bounds for variable selection. *Biometrics* 75, 392–403. doi: 10.1111/biom.13024

Lindsay, B. G. (2004). "Statistical distances as loss functions in assessing model adequacy," in *The Nature of Scientific Evidence: Statistical, philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: University of Chicago Press). doi: 10.3390/e20060464

Linhart, H. (1988). A test whether 2 AICs differ significantly. *South African Statist. J.* 22, 153–161.

Liu, X. H., Li, Y. Y., and Jiang, J. M. (2021). Simple measures of uncertainty for model selection. *Test* 30, 673–692.

Markatou, M., and Sofikitou, E. M. (2019). Statistical distances and the construction of evidence functions for model adequacy. *Front. Ecol. Evol.* 7:447447. doi: 10.3389/fevo.2019.00447447

Mayo, D. G. (2018). *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

Meeker, W. Q., and Escobar, L. A. (1995). Teaching about approximate confidence-regions based on maximum-likelihood-estimation. *Am. Statist.* 49, 48–53.

Nagler, T., and Vatter, T. (2019). *kde1d: Univariate Kernel Density Estimation. R Package Version 1.0.2.* Available online at: https://CRAN.R-project.org/package=kde1d

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. London Series A Mathemat. Phys. Sci.* 236, 333–380.

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London Series A* 231, 289–337.

Ng, C. T., and Joe, H. (2016). Comparison of non-nested models under a general measure of distance. *J. Statist. Plann. Inference* 170, 166–185. doi: 10.1016/j.jspi.2015.10.004

Nishii, R. (1988). Maximum-Likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.* 27, 392–403.

Pace, L., and Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *J. Statist. Plann. Inference* 136, 3554–3564.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford: Oxford University Press.

Pena, E. A., Rohatgi, V. K., and Szekely, G. J. (1992). On the non-existence of ancillary statistics. *Statist. Probab. Lett.* 15, 357–360.

Pierce, D. A., and Bellio, R. (2017). Modern likelihood-frequentist inference. *Int. Statist. Rev.* 85, 519–541.

Ponciano, J. M., Burleigh, G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biol.* 61, 955–972. doi: 10.1093/sysbio/sys055

Ponciano, J. M., and Taper, M. L. (2019). Model projections in model space: a geometric interpretation of the AIC allows estimating the distance between truth and approximating models. *Front. Ecol. Evol.* 7:413. doi: 10.3389/fevo.2019.00413

Ponciano, J. M., Taper, M. L., Dennis, B., and Lele, S. R. (2009). Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology* 90, 356–362. doi: 10.1890/08-0967.1

Powell, L. A., and Gale, G. A. (2015). *Estimation of Parameters for Animal Populations: a Primer for the Rest of US*. Lincoln, NE: Caught Napping Publications.

Royall, R. M. (1997). *Statistical Evidence: a Likelihood Paradigm*. London: Chapman & Hall.

Royall, R. M. (2000). On the probability of observing misleading statistical evidence. *J. Am. Statist. Assoc.* 95, 760–780.

Royall, R. M. (2004). "The likelihood paradigm for statistical evidence," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: The University of Chicago Press).

Royall, R. M., and Cumberland, W. G. (1985). Conditional coverage properties of finite population confidence-intervals. *J. Am. Statist. Assoc.* 80, 355–359. doi: 10.1093/jssam/smv031

Rubin, M. (2020). Repeated sampling from the same population? a critique of Neyman and Pearson's responses to Fisher. *Eur. J. Philos. Sci.* 10:42.

Sayyareh, A. (2017). Non parametric multiple comparisons of non nested rival models. *Commun. Statistics-Theory Methods* 46, 8369–8386. doi: 10.1080/03610926.2016.1179759

Sayyareh, A., Obeidi, R., and Bar-Hen, A. (2011). Empiricial comparison between some model selection criteria. *Commun. Statistics-Simulat. Comput.* 40, 84–98. doi: 10.1080/03610918.2010.530367

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464. doi: 10.1007/978-3-319-10470-6_18

Schweder, T. (2018). Confidence is epistemic probability for empirical science. *J. Statist. Plann. Inference* 195, 116–125.

Serfling, R. J. (1984). Generalized L-statistics, M-statistics, and R-statistics. *Ann. Statist.* 12, 76–86.

Severini, T. A. (2000). The likelihood ratio approximation to the conditional distribution of the maximum likelihood estimator in the discrete case. *Biometrika* 87, 939–945.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Institute Statistical Mathemat.* 50, 1–13.

Sprott, D. A. (2000). *Statistical Inference in Science*. New York, NY: Springer-Verlag.

Strug, L. J., and Hodge, S. E. (2006a). An alternative foundation for the planning and evaluation of linkage analysis I. decoupling 'error probabilities' from 'measures of evidence'. *Hum. Heredity* 61, 166–188. doi: 10.1159/000094709

Strug, L. J., and Hodge, S. E. (2006b). An alternative foundation for the planning and evaluation of linkage analysis II. implications for multiple test adjustments. *Hum. Heredity* 61, 200–209. doi: 10.1159/000094775

Strug, L. J., Rohde, C. A., and Corey, P. N. (2007). An introduction to evidential sample size calculations. *Am. Statist.* 61, 207–212.

Taper, M. L. (2004). "Model identification from many candidates," in *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, eds M. L. Taper and S. R. Lele (Chicago, ILL: The University of Chicago Press).

Taper, M. L., and Lele, S. R. (eds) (2004). *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. Chicago, ILL: The University of Chicago Press.

Taper, M. L., and Lele, S. R. (2011). "Evidence, evidence functions, and error probabilities," in *Philosophy of Statistics*, eds P. S. Bandyopadhyay and M. R. Forster (Oxford: Elsevier).

Taper, M. L., Lele, S. R., Ponciano, J.-M., and Dennis, B. (2019). Assessing the uncertainty in statistical evidence with the possibility of model misspecification using a non-parametric bootstrap. *arXiv [Preprints]*. Available online at: https://arxiv.org/ftp/arxiv/papers/1911/1911.06421.pdf

Taper, M. L., and Marquet, P. A. (1996). How do species really divide resources? *Am. Nat.* 147, 1072–1086.

Taper, M. L., and Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Popul. Ecol.* 58, 9–29.

Tomarken, A. J., and Waller, N. G. (2003). Potential problems with "well fitting" models. *J. Abnorm. Psychol.* 112, 578–598.

Tukey, J. W. (1960). Conclusions vs decisions. *Technometrics* 2, 423–433.

Vidoni, P. (1995). A simple predictive density based on the p*-formula. *Biometrika* 82, 855–863.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi: 10.1002/jbmr.3576

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Trans. Am. Math Soc.* 54, 426–482.

White, H. (1982). Maximum-likelihood estimation of mis-specified models. *Econometrica* 50, 1–25. doi: 10.2307/1912526

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Mathemat. Statist.* 9, 60–62. doi: 10.1186/1471-2156-10-72

Wright, S. S. (1934). The method of path coefficients. *Ann. Mathemat. Statist.* 5, 161–215.

Xie, M. G., and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int. Statist. Rev.* 81, 3–39. doi: 10.1002/jrsm.1471

Yang, Z., and Zhu, T. (2018). Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U S A.* 115, 1854–1859. doi: 10.1073/pnas.1712673115

Zheng, C., Ferrari, D., and Yang, Y. H. (2019). Model selection confidence sets by likelihood ratio testing. *Statist. Sinica* 29, 827–851. doi: 10.5705/ss.202017.0006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership