frontiers in Neuroscience

ORIGINAL RESEARCH
published: 23 January 2017
doi: 10.3389/fnins.2016.00619

Interpretability of Multivariate Brain Maps in Linear Brain Decoding: Definition, and Heuristic Quantification in Multivariate Analysis of MEG Time-Locked Effects

Seyed Mostafa Kia[1], Sandro Vega Pons[2,3], Nathan Weisz[4] and Andrea Passerini[1]*

[1] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, [2] Fondazione Bruno Kessler, Trento, Italy, [3] Pattern Analysis and Computer Vision , Istituto Italiano di Tecnologia, Genova, Italy, [4] Division of Physiological Psychology, Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria

OPEN ACCESS

Edited by:
Alexandre Gramfort,
Université Paris-Saclay (CNRS),
France

Reviewed by:
Denis A. Engemann,
Unicog (CEA/INSERM) and ICM,
France
Mads Jensen,
Aarhus University, Denmark

***Correspondence:**
Seyed Mostafa Kia
seyedmostafa.kia@unitn.it

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *20 September 2016*
Accepted: *27 December 2016*
Published: *23 January 2017*

Citation:
Kia SM, Vega Pons S, Weisz N and Passerini A (2017) Interpretability of Multivariate Brain Maps in Linear Brain Decoding: Definition, and Heuristic Quantification in Multivariate Analysis of MEG Time-Locked Effects. Front. Neurosci. 10:619. doi: 10.3389/fnins.2016.00619

Brain decoding is a popular multivariate approach for hypothesis testing in neuroimaging. Linear classifiers are widely employed in the brain decoding paradigm to discriminate among experimental conditions. Then, the derived linear weights are visualized in the form of multivariate brain maps to further study spatio-temporal patterns of underlying neural activities. It is well known that the brain maps derived from weights of linear classifiers are hard to interpret because of high correlations between predictors, low signal to noise ratios, and the high dimensionality of neuroimaging data. Therefore, improving the interpretability of brain decoding approaches is of primary interest in many neuroimaging studies. Despite extensive studies of this type, at present, there is no formal definition for interpretability of multivariate brain maps. As a consequence, there is no quantitative measure for evaluating the interpretability of different brain decoding methods. In this paper, first, we present a theoretical definition of interpretability in brain decoding; we show that the interpretability of multivariate brain maps can be decomposed into their reproducibility and representativeness. Second, as an application of the proposed definition, we exemplify a heuristic for approximating the interpretability in multivariate analysis of evoked magnetoencephalography (MEG) responses. Third, we propose to combine the approximated interpretability and the generalization performance of the brain decoding into a new multi-objective criterion for model selection. Our results, for the simulated and real MEG data, show that optimizing the hyper-parameters of the regularized linear classifier based on the proposed criterion results in more informative multivariate brain maps. More importantly, the presented definition provides the theoretical background for quantitative evaluation of interpretability, and hence, facilitates the development of more effective brain decoding algorithms in the future.

Keywords: brain decoding, brain mapping, interpretation, model selection, MEG

1. INTRODUCTION

Understanding the mechanisms of the brain has been a crucial topic throughout the history of science. Ancient Greek philosophers envisaged different functionalities for the brain ranging from cooling the body to acting as the seat of the rational soul and the center of sensation (Crivellato and Ribatti, 2007). Modern cognitive science, emerging in the twentieth century, provides better insight into the brain's functionality. In cognitive science, researchers usually analyze recorded brain activity and behavioral parameters to discover the answers of *where*, *when*, and *how* a brain region participates in a particular cognitive process.

To answer the key questions in cognitive science, scientists often employ mass-univariate hypothesis testing methods to test scientific hypotheses on a large set of independent variables (Groppe et al., 2011a; Maris, 2012). Mass-univariate hypothesis testing is based on performing multiple tests, e.g., *t*-tests, one for each unit of the neuroimaging data, i.e., independent variables. The high spatial and temporal granularity of the univariate tests provides fair level of interpretability. On the down side, the high dimensionality of neuroimaging data requires a large number of tests that reduces the sensitivity of these methods after multiple comparison correction (Bzdok et al., 2016). Although techniques such as the non-parametric cluster-based permutation test (Bullmore et al., 1996; Maris and Oostenveld, 2007), by weak rather strong control of family-wise error rate, offer more sensitivity, they still experience low sensitivity to brain activities that are narrowly distributed in time and space (Groppe et al., 2011a,b). The multivariate counterpart of mass-univariate analysis, known generally as multivariate pattern analysis, have the potential to overcome these deficits. Multivariate approaches are capable of identifying complex spatio-temporal interactions between different brain areas with higher sensitivity and specificity than univariate analysis (van Gerven et al., 2009), especially in group analysis of neuroimaging data (Davis et al., 2014).

Brain decoding (Haynes and Rees, 2006) is a multivariate technique that delivers a model to predict the mental state of a human subject based on the recorded brain signal. There are two potential applications for brain decoding: (1) brain-computer interfaces (BCIs) (Wolpaw et al., 2002), and (2) multivariate hypothesis testing (Bzdok et al., 2016). In the first case, a brain decoder with maximum prediction power is desired. In the second case, in addition to the prediction power, extra information on the spatio-temporal nature of a cognitive process is desired. In this study, we are interested in the second application of brain decoding that can be considered a multivariate alternative for mass-univariate hypothesis testing. Further, we mainly focus on the linear brain decoding because of its wider usage in analyzing inherently small sample size and high dimensional neuroimaging data, compared to the complex (Cox and Savoy, 2003; LaConte et al., 2005) and non-transparent (Lipton et al., 2016) non-linear models.

In linear brain decoding, linear classifiers are used to assess the relation between independent variables, i.e., features, and dependent variables, i.e., cognitive tasks (Besserve et al., 2007; Pereira et al., 2009; Lemm et al., 2011). This assessment is performed by solving an optimization problem that assigns weights to each independent variable. Currently, brain decoding is the gold standard in multivariate analysis for functional magnetic resonance imaging (fMRI) (Haxby et al., 2001; Cox and Savoy, 2003; Mitchell et al., 2004; Norman et al., 2006) and magnetoencephalogram/electroencephalogram (MEEG) studies (Parra et al., 2003; Rieger et al., 2008; Carroll et al., 2009; Chan et al., 2011; Huttunen et al., 2013; Vidaurre et al., 2013; Abadi et al., 2015). It has been shown that brain decoding can be used in combination with brain encoding (Naselaris et al., 2011) to infer the causal relationship between stimuli and responses (Weichwald et al., 2015).

In *Brain mapping* (Kriegeskorte et al., 2006) the pre-computed quantities, e.g., univariate statistics or weights of a linear classifier, are assigned to the spatio-temporal representation of neuroimaging data in order to reveal functionally specialized brain regions which are activated by a certain cognitive task. In its multivariate form, brain mapping uses the learned parameters from brain decoding to produce brain maps, in which the engagement of different brain areas in a cognitive task is visualized. Intuitively, the interpretability of a brain decoder refers to the level of information that can be reliably derived by an expert from the resulting maps. From the cognitive neuroscience perspective, a brain map is considered *interpretable* if it enables a scientist to find out the answers of three key questions: "*where*, *when*, and *how* does a brain region contribute to a cognitive function?"

In fact, a classifier only answers the question of *what* is the most likely label of a given unseen sample (Baehrens et al., 2010). This fact is generally known as knowledge extraction gap (Vellido et al., 2012) in the machine learning context. Thus far, many efforts have been devoted to filling the knowledge extraction gap of linear and non-linear data modeling methods in different areas such as computer vision (Bach et al., 2015), signal processing (Montavon et al., 2013), chemometrics (Yu et al., 2015), bioinformatics (Hansen et al., 2011), and neuroinformatics (Haufe et al., 2013). In the context of neuroimaging, the knowledge extraction gap in classification is generally known as the interpretation problem (Sabuncu, 2014; Haynes, 2015; Naselaris and Kay, 2015). Therefore, improving the interpretability of linear brain decoding and associated brain maps is a primary goal in the brain imaging literature (Strother et al., 2014). The lack of interpretability of multivariate brain maps is a direct consequence of low signal-to-noise ratios (SNRs), high dimensionality of whole-scalp recordings, high correlations among different dimensions of data, and cross-subject variability (Besserve et al., 2007; Anderson et al., 2011; Blankertz et al., 2011; Brodersen et al., 2011; Langs et al., 2011; Lemm et al., 2011; Varoquaux et al., 2012; Kauppi et al., 2013; Haufe et al., 2014a; Olivetti et al., 2014; Taulu et al., 2014; Varoquaux and Thirion, 2014; Haynes, 2015; Wang et al., 2015). At present, two main approaches are proposed to enhance the interpretability of multivariate brain maps: (1) introducing new metrics into the model selection procedure, and (2) introducing new hybrid penalty terms for regularization.

The first approach to improving the interpretability of brain decoding concentrates on the model selection procedure. Model selection is a procedure in which the best values for the hyper-parameters of a model are determined (Lemm et al., 2011). The selection process is generally performed by considering the generalization performance, i.e., the accuracy, of a model as the decisive criterion. Rasmussen et al. (2012) showed that there is a trade-off between the spatial reproducibility and the prediction accuracy of a classifier; therefore, the reliability of maps cannot be assessed merely by focusing on their prediction accuracy. To utilize this finding, they incorporated the spatial reproducibility of brain maps in the model selection procedure. An analogous approach, using a different definition of spatial reproducibility, is proposed by Conroy et al. (2013). Beside spatial reproducibility, the stability of the classifiers (Bousquet and Elisseeff, 2002) is another criterion that is used in combination with generalization performance to enhance the interpretability. For example Yu (2013) and Lim and Yu (2016) showed that incorporating the stability of models into cross-validation improves the interpretability of the estimated parameters (by linear models).

The second approach to improving the interpretability of brain decoding focuses on the underlying mechanism of regularization. The main idea behind this approach is two-fold: 1) customizing the regularization terms to address the ill-posed nature of brain decoding problems (where the number of samples is much less than the number of features; Mørch et al., 1997; Varoquaux and Thirion, 2014), and (2) combining the structural and functional prior knowledge with the decoding process so as to enhance the neurophysiological plausibility of the models. Group Lasso (Yuan and Lin, 2006) and total-variation penalty (Tibshirani et al., 2005) are two effective methods using this technique (Rish et al., 2014; Xing et al., 2014). Sparse penalized discriminant analysis (Grosenick et al., 2008), group-wise regularization (van Gerven et al., 2009), smoothed-sparse logistic regression (de Brecht and Yamagishi, 2012), total-variation ℓ_1 penalization (Michel et al., 2011; Gramfort et al., 2013), the graph-constrained elastic-net (Grosenick et al., 2009, 2013), and social-sparsity (Varoquaux et al., 2016) are examples of brain decoding methods in which regularization techniques are employed to improve the interpretability of linear brain decoding models.

Recently, taking a new approach to the problem, Haufe et al. questioned the interpretability of weights of linear classifiers because of the contribution of noise in the decoding process (Bießmann et al., 2012; Haufe et al., 2013, 2014b). To address this problem, they proposed a procedure to convert the linear brain decoding models into their equivalent generative models. Their experiments on the simulated and fMRI/EEG data illustrate that, whereas the direct interpretation of classifier weights may cause severe misunderstanding regarding the actual underlying effect, their proposed transformation effectively provides interpretable maps. Despite the theoretical soundness, the intricate challenge of estimating the empirical covariance matrix of the small sample size neuroimaging data (Blankertz et al., 2011) limits the practical application of this method.

In spite of the aforementioned efforts to improve the interpretability of brain decoding, there is still no formal definition for the interpretability of brain decoding in the literature. Therefore, the interpretability of different brain decoding methods are evaluated either qualitatively or indirectly (i.e., by means of an intermediate property). In qualitative evaluation, to show the superiority of one decoding method over the other (or a univariate map), the corresponding brain maps are compared visually in terms of smoothness, sparseness, and coherency using already known facts (see for example, Varoquaux et al., 2012). In the second approach, important factors in interpretability such as spatio-temporal reproducibility are evaluated to indirectly assess the interpretability of results (see for example, Langs et al., 2011; Rasmussen et al., 2012; Conroy et al., 2013; Kia et al., 2016). Despite partial effectiveness, there is no general consensus regarding the quantification of these intermediate criteria. For example, in the case of spatial reproducibility, different methods such as correlation (Rasmussen et al., 2012; Kia et al., 2016), dice score (Langs et al., 2011), or parameter variability (Conroy et al., 2013; Haufe et al., 2013) are used for quantifying the stability of brain maps, each of which considers different aspects of local or global reproducibility.

With the aim of filling this gap, our contribution is three-fold: (1) Assuming that the true solution of brain decoding is available, we present a theoretical definition of the interpretability. The presented definition is simply based on cosine proximity in the parameter space. Furthermore, we show that the interpretability can be decomposed into the reproducibility and the representativeness of brain maps. (2) As a proof of the concept, we exemplify a practical heuristic based on event-related fields for quantifying the interpretability of brain maps in time-locked analysis of MEG data. (3) Finally, we propose the combination of the interpretability and the performance of the brain decoding as a new Pareto optimal multi-objective criterion for model selection. We experimentally, on both simulated and real data, show that incorporating the interpretability into the model selection procedure provides more reproducible, more neurophysiologically plausible, and (as a result) more interpretable maps. Furthermore, in comparison with a standard univariate analysis, we show the proposed paradigm offers more sensitivity while preserving the interpretability of results.

2. MATERIALS AND METHODS

2.1. Notation and Background

Let $\mathcal{X} \in \mathbb{R}^p$ be a manifold in Euclidean space that represents the input space and $\mathcal{Y} \in \mathbb{R}$ be the output space, where $\mathcal{Y} = \Phi^*(\mathcal{X})$. Then, let $S = \{\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \mid z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ be a training set of n independently and identically distributed (i.i.d) samples drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ based on an unknown Borel probability measure ρ. In the neuroimaging context, $\mathbf{X}$ indicates the trials of brain recording, e.g., fMRI, MEG, or EEG signals, $\mathbf{Y}$ represents the experimental conditions or dependent variables, and we have $\Phi_S : \mathbf{X} \rightarrow \mathbf{Y}$ (note the difference between Φ_S and Φ^*). The goal of brain decoding is to find the function $\hat{\Phi} : \mathbf{X} \rightarrow \mathbf{Y}$ as an estimation of Φ_S. Here on we refer to $\hat{\Phi}$ as a brain decoding model.

As is a common assumption in the neuroimaging context, we assume the true solution of a brain decoding problem is among the family of linear functions $\mathcal{H}$ ($\Phi^* \in \mathcal{H}$). Therefore, the aim of brain decoding reduces to finding an empirical approximation of Φ_S, indicated by $\hat{\Phi}$, among all $\Phi \in \mathcal{H}$. This approximation can be obtained by estimating the predictive conditional density $\rho(\mathbf{Y} \mid \mathbf{X})$ by training a parametric model $\rho(\mathbf{Y} \mid \mathbf{X}, \Theta)$ (i.e., a likelihood function), where Θ denotes the parameters of the model. Alternatively, Θ can be estimated by solving a risk minimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{X}\Theta, \mathbf{Y}) + \lambda\Omega(\Theta) \tag{1}$$

where $\hat{\Theta}$ is the parameter of $\hat{\Phi}$, $\mathcal{L} : \mathbf{Z} \times \mathbf{Z} \rightarrow \mathbb{R}^+$ is the loss function, $\Omega : \mathbb{R}^p \rightarrow \mathbb{R}^+$ is the regularization term, and λ is a hyper-parameter that controls the amount of regularization. There are various choices for Ω, each of which reduces the hypothesis space $\mathcal{H}$ to $\mathcal{H}' \subset \mathcal{H}$ by enforcing different prior functional or structural constraints on the parameters of the linear decoding model (see for example, Tibshirani, 1996b; Tibshirani et al., 2005; Zou and Hastie, 2005; Jenatton et al., 2011). The amount of regularization λ is generally decided using cross-validation or other data perturbation methods in the model selection procedure.

In the neuroimaging context, the estimated parameters of a linear decoding model $\hat{\Theta}$ can be used in the form of a brain map so as to visualize the discriminative neurophysiological effect. Although the magnitude of $\hat{\Theta}$ (i.e., the 2-norm of $\hat{\Theta}$) is affected by the dynamic range of data and the level of regularization, it has no effect on the predictive power and the interpretability of maps. On the other hand, the direction of $\hat{\Theta}$ affects the predictive power and contains information regarding the importance of and relations among predictors. This type of relational information is very useful when interpreting brain maps in which the relation between different spatio-temporal independent variables can be used to describe how different brain regions interact over time for a certain cognitive process. Therefore, we refer to the normalized parameter vector of a linear brain decoder in the unit hypersphere as a multivariate brain map (MBM); we denote it by $\vec{\Theta}$ where $\vec{\Theta} = \frac{\Theta}{\|\Theta\|_2}$ ($\|.\|_2$ represents the 2-norm of a vector).

As shown in Equation (1), learning occurs using the sampled data. In other words, in the learning paradigm, we attempt to minimize the loss function with respect to Φ_S (and not Φ^*) (Cucker and Smale, 2002). Therefore, all of the implicit assumptions (such as linearity) regarding Φ^* might not hold on Φ_S, and vice versa. The *irreducible error* ε is the direct consequence of sampling; it sets a lower bound on the error, where we have:

$$\Phi_S(\mathbf{X}) = \Phi^*(\mathbf{X}) + \varepsilon \tag{2}$$

The distribution of ε dictates the type of loss function $\mathcal{L}$ in Equation (1). For example, assuming a Gaussian distribution with mean 0 and variance σ^2 for ε implies the least squares loss function (Wu et al., 2006).

2.2. Interpretability of Multivariate Brain Maps: Theoretical Definition

In this section, we present a theoretical definition for the interpretability of linear brain decoding models and their associated MBMs. Consider a linearly separable brain decoding problem in an ideal scenario where $\varepsilon = 0$ and $rank(\mathbf{X}) = p$. In this case, the ideal solution of brain decoding, Φ^*, is linear and its parameters Θ^* are *unique* and neurophysiologically *plausible* (van Ede and Maris, 2016). The unique parameter vector Θ^* can be computed as follows:

$$\Theta^* = \Sigma_{\mathbf{X}}^{-1}\mathbf{X}^T\mathbf{Y} \tag{3}$$

where $\Sigma_{\mathbf{X}}$ represents the covariance of $\mathbf{X}$. Using Θ^* as the reference, we define the *strong-interpretability* of an MBM as follows:

Definition 1. An MBM $\vec{\hat{\Theta}}$ associated with a linear brain decoding model $\hat{\Phi}$ is "strongly-interpretable" if and only if $\vec{\hat{\Theta}} \propto \Theta^*$.

It can be shown that, in practice, the estimated solution of a linear brain decoding problem is not strongly-interpretable because of the inherent limitations of neuroimaging data, such as uncertainty (Aggarwal and Yu, 2009) in the input and output space ($\varepsilon \neq 0$), the high dimensionality of data ($n \ll p$), and the high correlation between predictors ($rank(\mathbf{X}) < p$). With these limitations in mind, even though in practice the solution of linear brain decoding is not strongly-interpretable, one can argue that some are more interpretable than others. For example, in the case in which $\Theta^* \propto [0, 1]^T$, a linear classifier where $\vec{\hat{\Theta}} \propto [0.1, 1.2]^T$ can be considered more interpretable than a linear classifier where $\vec{\hat{\Theta}} \propto [2, 1]^T$. This issue raises the following question:

Problem 1. Let S be a training set of n i.i.d samples drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $P(S)$ be the probability of drawing a certain S from $\mathcal{Z}$. Assume $\vec{\hat{\Theta}}$ is the MBM of a linear brain decoding model $\hat{\Phi}$ on S (estimated using Equation 1 for a certain loss function $\mathcal{L}$, regularization term Ω, and hyper-parameter λ). How can we quantify the proximity of $\hat{\Phi}$ to the strongly-intrepretable solution of the brain decoding problem Φ^*?

To answer this question, considering the uniqueness and the plausibility of Φ^* as the two main characteristics that convey its strong-interpretability, we define the interpretability as follows:

Definition 2. Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1. Then, assume α be the angle between $\vec{\hat{\Theta}}$ and $\vec{\Theta}^*$. The "interpretability" ($0 \leq \eta_\Phi \leq 1$) of a linear brain decoding model $\hat{\Phi}$ is defined as follows:

$$\eta_\Phi = \mathbb{E}_{P(S)}[\cos(\alpha)] \tag{4}$$

In practice, only a limited number of samples are available. Therefore, perturbation techniques are used to imitate the sampling procedure. Let $S^1, \ldots, S^m$ be m perturbed training sets drawn from S via a certain perturbation scheme such as jackknife, bootstrapping (Efron, 1992), or cross-validation (Kohavi, 1995). Assume $\hat{\vec{\Theta}}^1, \ldots, \hat{\vec{\Theta}}^m$ are m MBMs estimated on the corresponding perturbed training sets, and α^j ($j = 1, \ldots, m$) be the angle between $\hat{\vec{\Theta}}^j$ and $\vec{\Theta}^*$. Then, the empirical version of Equation (4) can be rewritten as follows:

$$\eta_\Phi = \frac{1}{m} \sum_{j=1}^{m} \cos(\alpha^j) \tag{5}$$

Empirically, the interpretability is the mean of cosine similarities between Θ^* and MBMs derived fro different samplings of the training set (see **Figure 1A** for a schematic illustration). In addition to the fact that employing cosine similarity is a common method for measuring the similarity between vectors, we have another strong motivation for this choice. It can be shown that, for large values of p, the distribution of the dot product in the unit hyper-sphere, i.e., the cosine similarity, converges to a normal distribution with 0 mean and variance of $\frac{1}{p}$, i.e., $\mathcal{N}(0, \sqrt{1/p})$. Due to the small variance for a large enough p values, any similarity value that is significantly larger than zero represents a meaningful similarity between two high dimensional vectors (see Appendix 6.3 for the mathematical demonstration).

In what follows, we demonstrate how the definition of interpretability is geometrically related to the uniqueness and plausibility characteristics of the true solution of the brain decoding problem.

2.3. Interpretability Decomposition into Reproducibility and Representativeness

The trustworthy and informativeness of decoding models are providing two important motivations for improving the interpretability of models (Lipton et al., 2016). The trust of a learning algorithm refers to its ability to converge to a unique solution. On the other hand, the informativeness refers to the level of plausible information that can be derived from a model to assist or advise to a human expert. Therefore, it is expected that the interpretability can be quantified alternatively by assessing its uniqueness and neurophysiological plausibility. In this section, we firstly define the reproducibility and representativeness as measures for quantifying the uniqueness and neurophysiological plausibility of a brain decoding model, respectively. Then we show how these definitions are related to the definition of interpretability.

The high dimensionality and the high correlations between variables are two inherent characteristics of neuroimaging data that negatively affect the uniqueness of the solution of a brain decoding problem. Therefore, a certain configuration of hyper-parameters may result different estimated parameters on different portions of data. Here, we are interested in assessing this variability as a measure for uniqueness. We first define the *main multivariate brain map* as follows:

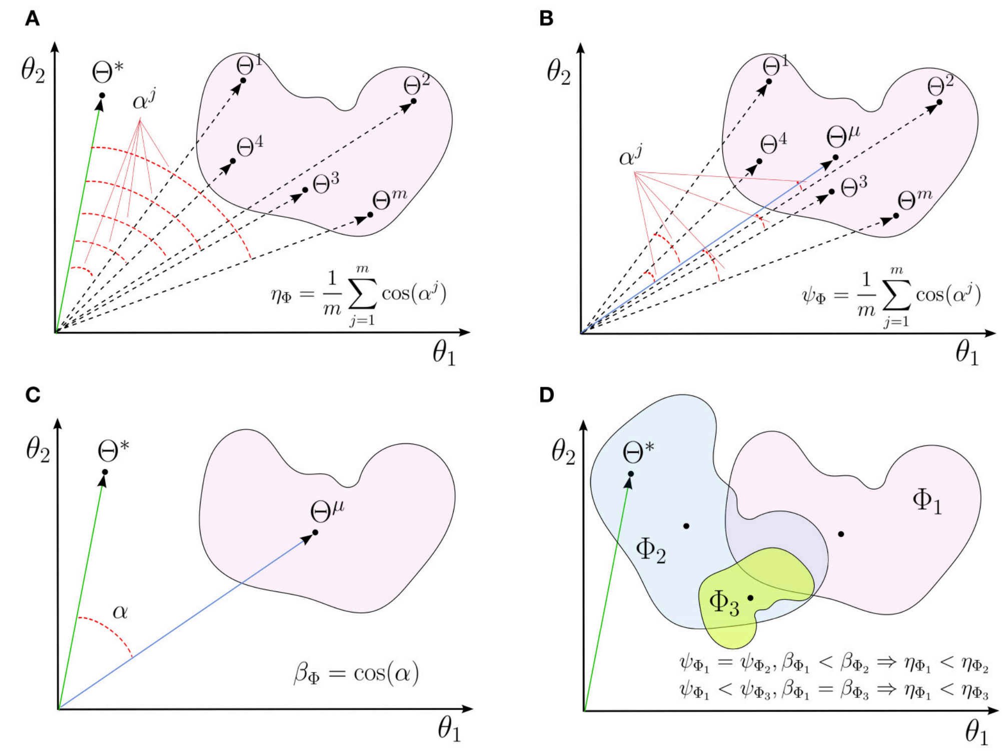

A
θ_2
Θ^*
α^j
Θ^1
Θ^2
Θ^4
Θ^3
Θ^m
$\eta_\Phi = \frac{1}{m}\sum_{j=1}^{m}\cos(\alpha^j)$
θ_1
B
θ_2
α^j
Θ^1
Θ^2
Θ^4
Θ^μ
Θ^3
Θ^m
$\psi_\Phi = \frac{1}{m}\sum_{j=1}^{m}\cos(\alpha^j)$
θ_1
C
θ_2
Θ^*
Θ^μ
α
$\beta_\Phi = \cos(\alpha)$
θ_1
D
θ_2
Θ^*
Φ_1
Φ_2
Φ_3
$\psi_{\Phi_1} = \psi_{\Phi_2}, \beta_{\Phi_1} < \beta_{\Phi_2} \Rightarrow \eta_{\Phi_1} < \eta_{\Phi_2}$
$\psi_{\Phi_1} < \psi_{\Phi_3}, \beta_{\Phi_1} = \beta_{\Phi_3} \Rightarrow \eta_{\Phi_1} < \eta_{\Phi_3}$
θ_1

FIGURE 1 | A schematic illustrations for (A) interpretability (η_Φ), **(B)** reproducibility (ψ_Φ), and **(C)** representativeness (β_Φ) of a linear decoding model in two dimensions. **(D)** The independent effects of the reproducibility and the representativeness of a model on its interpretability.

Definition 3. Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1. The "main multivariate brain map" $\vec{\Theta}^{\mu} \in \mathbb{R}^{p}$ of a linear brain decoding model $\hat{\Phi}$ is defined as:

$$\vec{\Theta}^{\mu} = \frac{\mathbb{E}_{P(S)}[\vec{\hat{\Theta}}]}{\left\| \mathbb{E}_{P(S)}[\vec{\hat{\Theta}}] \right\|_{2}} \tag{6}$$

Assuming θ_{i}^{j} be the ith ($i = 1, \ldots, p$) element of an MBM estimated on the jth ($j = 1, \ldots, m$) perturbed training set, $\vec{\Theta}^{\mu}$ empirically can be estimated by summing up $\vec{\hat{\Theta}}^{j}$s (computed on the perturbed training set S^{j}) in the unit hyper-sphere:

$$\vec{\Theta}^{\mu} = \frac{\left[\sum_{j=1}^{m} \theta_{1}^{j} \; \sum_{j=1}^{m} \theta_{2}^{j} \; \ldots \; \sum_{j=1}^{m} \theta_{p}^{j}\right]^{T}}{\left\| \left[\sum_{j=1}^{m} \theta_{1}^{j} \; \sum_{j=1}^{m} \theta_{2}^{j} \; \ldots \; \sum_{j=1}^{m} \theta_{p}^{j}\right]^{T} \right\|_{2}} \tag{7}$$

$\vec{\Theta}^{\mu}$ provides a reference for quantifying the reproducibility of an MBM:

Definition 4. Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1, and $\vec{\Theta}^{\mu}$ be the main multivariate brain map of $\hat{\Phi}$. Then, assume α be the angle between $\vec{\hat{\Theta}}^{j}$ and $\vec{\Theta}^{\mu}$. The "reproducibility" ψ_{Φ} ($0 \leq \psi_{\Phi} \leq 1$) of a linear brain decoding model $\hat{\Phi}$ is defined as follows:

$$\psi_{\Phi} = \mathbb{E}_{P(S)}[\cos(\alpha)] \tag{8}$$

Let $\vec{\hat{\Theta}}^{1}, \ldots, \vec{\hat{\Theta}}^{m}$ are m MBMs estimated on the corresponding perturbed training sets, and α^{j} ($j = 1, \ldots, m$) be the angle between $\vec{\hat{\Theta}}^{j}$ and $\vec{\Theta}^{\mu}$. Then, the empirical version of Eq. 8 can be rewritten as follows:

$$\psi_{\Phi} = \frac{1}{m} \sum_{j=1}^{m} \cos(\alpha^{j}) \tag{9}$$

In fact, reproducibility provides a measure for quantifying the dispersion of MBMs, computed over different perturbed training sets, from the main multivariate brain map. **Figure 1B** shows a schematic illustration for the reproducibility of a linear brain decoding model.

On the other hand, the similarity between the main multivariate brain map of a decoder and the true solution can be employed as a measure for the neurophysiological plausibility of a model. We refer to this similarity as the *representativeness* of a linear brain decoding model:

Definition 5. Let $\vec{\Theta}^{\mu}$ be the main multivariate brain map of $\hat{\Phi}$. The "representativeness" β_{Φ} ($0 \leq \beta_{\Phi} \leq 1$) of a linear brain decoding model $\hat{\Phi}$ is defined as the cosine similarity between its main multivariate brain map ($\vec{\Theta}^{\mu}$) and the parameters of the true solution ($\vec{\Theta}^{*}$) (see **Figure 1C**):

$$\beta_{\Phi} = \frac{|\vec{\Theta}^{\mu}.\vec{\Theta}^{*}|}{\left\| \vec{\Theta}^{\mu} \right\|_{2} \left\| \vec{\Theta}^{*} \right\|_{2}} \tag{10}$$

As discussed before, the notion of interpretabilty is tightly related to the uniqueness and plausibility, thus to the reproducibility and representativeness, of a decoding model. The following proposition analytically shows this relationship:

Proposition 1. $\eta_{\Phi} = \beta_{\Phi} \times \psi_{\Phi}$.

See Appendix 6.1 for a proof. Proposition 1 indicates the interpretability of a linear brain decoding model can be decomposed into its representativeness and reproducibility. **Figure 1D** illustrates how the reproducibility and the representativeness of a decoding model independently affect its interpretability. Each colored region schematically represents a span of different solutions of the a certain linear model (for example with a certain configuration for its hyper-parameters) on different perturbed training sets. The area of each region schematically visualizes the reproducibility of each model, i.e., the less is the area, the higher is the reproducibility of a model. Further, the angular distance between the centroid of each region (Θ^{μ}) and the true solution (Θ^{*}) visualizes the representativeness of each corresponding model. While Φ_{1} and Φ_{2} have similar reproducibility, Φ_{2} has higher interpretability than Φ_{1} because it is more representative of the true solution. On the other hand, Φ_{1} and Φ_{3} have similar representativeness, however Φ_{3} is more interpretable due to the higher level of reproducibility.

2.4. A Heuristic for Practical Quantification of Interpretability in Time-Locked Analysis of MEG Data

In practice, it is impossible to evaluate the interpretability, as the true solution of the brain decoding problem Φ^{*} is unknown. In this study, to provide a practical proof of theoretical concepts, we exemplify contrast event-related field (cERF) as a neurophysiological plausible heuristic for the true parameters of the linear brain decoding problem (Θ^{*}) in a binary MEG decoding scenario in time domain. Due to the nature of proposed heuristic, its application is limited to the brain responses that are time-locked to the stimulus onset, i.e., the evoked responses.

The MEEG data are a mixture of several simultaneous stimulus-related and stimulus-unrelated brain activities. Assessing the electro/magneto-physiological changes that are time-locked to events of interest is a common approach to the study of MEEG data. In general, unrelated-stimulus brain activities are considered as Gaussian noise with zero mean and variance σ^{2}. One popular approach to canceling the noise component is to compute the average of multiple trials. The assumption is that, when the effect of interest is time-locked to the stimulus onset, the independent noise components can be vanished by means of averaging. It is expected that the average will converge to the true value of the signal with a variance of $\frac{\sigma^{2}}{n}$ (where n is the number of trials). The result of the averaging process consist of a series of positive and negative peaks occurring at a fixed time relative to the event onset, generally known as ERF in the MEG context. These component peaks are reflecting phasic activity that are indexed with different aspects of cognitive processing (Rugg and Coles, 1995)[1].

Assume $\mathbf{X}^+ = \{x_i \in \mathbf{X} \mid y_i = 1\} \in \mathbb{R}^{n^+ \times p}$ and $\mathbf{X}^- = \{x_i \in \mathbf{X} \mid y_i = -1\} \in \mathbb{R}^{n^- \times p}$ be sets of positive and negative samples in a binary MEG decoding scenario. Then, the cERF brain map $\vec{\Theta}^{cERF}$ is computed as follows:

$$\vec{\Theta}^{cERF} = \frac{\frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i}{\left\| \frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i \right\|_2} \tag{11}$$

Generally speaking $\vec{\Theta}^{cERF}$ is a contrast ERF map between two experimental conditions. Using the core theory presented in Haufe et al. (2013), the equivalent generative model for the solution of linear brain decoding, i.e., the activation pattern (A), is unique and we have:

$$A \propto \Sigma_{\mathbf{X}} \hat{\Theta} \tag{12}$$

Assuming $\hat{\Theta}$ to be the solution of least squares in a binary decoding scenario, then the following proposition describes the relation between $\vec{\Theta}^{cERF}$ and the activation pattern A:

Proposition 2. $\vec{\Theta}^{cERF} \propto A$.

See Appendix 6.2 for the proof. Proposition 2.4 shows that, in a binary time-domain MEG decoding scenario, cERF is proportional to the equivalent generative model for the solution of least squares classifier. Furthermore, $\vec{\Theta}^{cERF}$ is proportional to the t-statistic that is widely used in the univariate analysis of neuroimaging data. Using $\vec{\Theta}^{cERF}$ as a heuristic for $\vec{\Theta}^*$, the representativeness can be approximated as follows:

$$\tilde{\beta}_\Phi = \frac{|\vec{\Theta}^{\mu}.\vec{\Theta}^{cERF}|}{\left\|\vec{\Theta}^{\mu}\right\|_2 \left\|\vec{\Theta}^{cERF}\right\|_2} \tag{13}$$

Where $\tilde{\beta}_\Phi$ is an approximation of the actual representativeness β_Φ. In a similar manner, $\vec{\Theta}^{cERF}$ can be used to heuristically approximate the interpretability as follows:

$$\tilde{\eta}_\Phi = \frac{1}{m}\sum_{j=1}^{m} \cos(\gamma^j) \tag{14}$$

where $\gamma_1, \ldots, \gamma_m$ are the angles between $\vec{\hat{\Theta}}^1, \ldots, \vec{\hat{\Theta}}^m$ and $\vec{\Theta}^{cERF}$. It can be shown that $\tilde{\eta}_\Phi = \tilde{\beta}_\Phi \times \psi_\Phi$.

The proposed heuristic is only applicable to the evoked responses in sensor and source space MEEG data. Despite this limitation, cERF provides an empirical example that shows how the presented theoretical definitions can be applied in a real decoding scenario. The choice of the heuristic has a direct effect on the approximation of interpretability and that an inappropriate selection of the heuristic yields a very poor estimation of interpretability. Therefore, the choice of heuristic should be carefully justified based on accepted and well-defined facts regarding the nature of the collected data.

Since the labels are used in the computation of cERF, a proper validation strategy should be employed to avoid the double dipping issue (Kriegeskorte et al., 2009). One possible approach is to exclude the entire test set from the model selection procedure using a nested nested cross-validation strategy. An alternative approach is employing model averaging techniques to neatly get advantage of the whole dataset (Varoquaux et al., 2017). Since our focus is on the model selection, in the remaining text, we implicitly assume the test data is excluded from the experiments, thus, all the experimental results are reported on the training and validation sets.

2.5. Incorporating the Interpretability into Model Selection

The procedure for evaluating the performance of a model so as to choose the best values for hyper-parameters is known as *model selection* (Hastie et al., 2009). This procedure generally involves numerical optimization of the model selection criterion on the training and validation sets (and not the test set). Let U be a set of hyper-parameters, then the goal of model selection procedure reduces to finding the best model configuration $u^* \in U$ that maximizes the model selection criterion (e.g., generalization performance) on the training set S. The most common model selection criterion is based on an estimator of generalization performance, i.e., the predictive power. In the context of brain decoding, especially when the interpretability of brain maps matters, employing the predictive power as the only decisive criterion in model selection is problematic in terms of interpretability of MBMs (Gramfort et al., 2012; Rasmussen et al., 2012; Conroy et al., 2013; Varoquaux et al., 2017). Valverde-Albacete and Peláez-Moreno (2014) experimentally showed that in a classification task optimizing only classification error rate is insufficient to capture the transfer of crucial information from the input to the output of a classifier. This fact highlights the importance of having some control over the estimated model weights in the model selection. Here, we propose a multi-objective criterion for model selection that takes into account both prediction accuracy and MBM interpretability.

Let $\tilde{\eta}_\Phi$ and δ_Φ be the approximated interpretability and the generalization performance of a linear brain decoding model $\hat{\Phi}$, respectively. We propose the use of the *scalarization* technique (Caramia and Dell´ Olmo, 2008) for combining $\tilde{\eta}_\Phi$ and δ_Φ into one scalar $0 \leq \zeta(\Phi) \leq 1$ as follows:

$$\zeta_\Phi = \begin{cases} \frac{\omega_1 \tilde{\eta}_\Phi + \omega_2 \delta_\Phi}{\omega_1 + \omega_2} & \delta_\Phi \geq \kappa \\ 0 & \delta_\Phi < \kappa \end{cases} \tag{15}$$

where ω_1 and ω_2 are weights that specify the level of importance of the interpretability and the performance, respectively. κ is a threshold on the performance that filters out solutions with poor performance. In classification scenarios, κ can be set by adding a small safe interval to the chance level of classification. The hyper-parameters that are optimized based on ζ_Φ are

[1] The application of the presented heuristic to MEG data can be extended to EEG because of the inherent similarity of the measured neural correlates in these two devices. In the EEG context, the ERF can be replaced by the event-related potential (ERP).

Pareto optimal (Marler and Arora, 2004). We hypothesize that optimizing the hyper-parameters based on ζ_Φ, rather only δ_Φ, yields more informative MBMs.

Algorithm 1 summarizes the proposed model selection scheme. The model selection procedure receives the training set S and a set of possible configurations for hyper-parameters U, and returns the best hyper-parameter configuration u^*.

2.6. Experimental Materials

2.6.1. Toy Dataset

We regenerate the simple 2-dimensional toy data presented in Haufe et al. (2013). Assume that the true underlying generative function Φ^* is defined by:

$$\mathcal{Y} = \Phi^*(\mathcal{X}) = \begin{cases} 1 & if \quad x_1 = 1.5 \\ -1 & if \quad x_1 = -1.5 \end{cases}$$

where $\mathcal{X} \in \{[1.5, 0]^T, [-1.5, 0]^T\}$; and x_1 and x_2 represent the first and the second dimension of the data, respectively. Furthermore, assume the data is contaminated by Gaussian noise with co-variance $\Sigma = \begin{bmatrix} 1.02 & -0.3 \\ -0.3 & 0.15 \end{bmatrix}$. In fact, the Gaussian noise adds uncertainty to the input space.

2.6.2. Simulated MEG Data

We simulated two classes of MEG data, each of which composed of 250 epochs with length of $330ms$ at $300Hz$ sampling rate (so that we have 100 time-points). For simplicity, the whole scalp topography are simulated with a single dipole located at -4.7, -3.7, and $5.3cm$ in the RAS (right, anterior, superior) coordinate system. The dipole is oriented toward [1,1,0] direction in the RA plane (see **Figure 2A**). One hundred two magnetometer sensors of Elekta Neuromag system are simulated using a standard forward model algorithm implemented in the Fieldtrip toolbox (Oostenveld et al., 2010). The epochs of the positive class are constructed by adding three components to the dipole time-course: (1) a time-locked ERF effect with a positive $3Hz$ followed by a negative $5\,Hz$ half-cycle sinusoid peaks after $150 \pm 10ms$ and $250 \pm 10ms$ of the epoch onset, respectively; (2) uncorrelated background brain activity that was simulated by summing 50 sinusoids with random frequency from 1 to $125Hz$, and random phase varied between 0 and 2π. Following the data simulation procedure in Yeung et al. (2004), the amplitude of any single frequency component of the signal (the ERF effect and the background noise) is set based on the empirical spectral power of human brain activity to mimic the actual magnetic features of scalp surface; and (3) white Gaussian noise scaled with the root mean squared of the signal in each epoch. The epochs of the negative class are constructed without the ERF effect by adding up only the noise components (i.e., the background activity and the white noise). Therefore, the ERF component is considered as the discriminative ground-truth in our experiments (see **Figure 2B**).

Algorithm 1 The model selection procedure.

```
procedure ModelSelection($S$,$U$)
    Compute $\vec{\Theta}^{cERF}$ on $S$.                         ▷ using Equation (11)
    for all $u_i \in U$ do                            ▷ For all hyper-parameter
                                                        configurations.
        for $j \leftarrow 1, m$ do                      ▷ Data perturbation iterations.
            Partition $S$ into training $S_{tr}$ and validation $S_{vl}$
            subsets via a perturbation method.
            Compute $\hat{\Theta}_j$ on $S_{tr}$ using $u_i$ as the
            hyper-parameter.
        end
        Compute $\delta^i_\Phi$ of $\hat{\Theta}_j$s on $S_{vl}$.
        Compute $\tilde{\eta}^i_\Phi$ of $\hat{\Theta}_j$s using $\vec{\Theta}^{cERF}$.          ▷ using
                                                          Equation (14)
        Compute $\zeta^i_\Phi$.                          ▷ using Equation (15)
    end
    $u^* = \text{argmax}_{u_i \in U}(\zeta_\Phi)$.
    return $u^*$.
```

2.6.3. MEG Data

We use the MEG dataset presented in Henson et al. (2011)[2]. The dataset was also used for the DecMeg2014 competition[3]. In this dataset, visual stimuli consisting of famous faces, unfamiliar faces, and scrambled faces are presented to 16 subjects and fMRI, EEG, and MEG signals are recorded. Here, we are only interested in MEG recordings. The MEG data were recorded using a VectorView system (Elekta Neuromag, Helsinki, Finland) with a magnetometer and two orthogonal planar gradiometers located at 102 positions in a hemispherical array in a light Elekta-Neuromag magnetically shielded room.

Three major reasons motivated the choice of this dataset: (1) It is publicly available. (2) The spatio-temporal dynamic of the MEG signal for face vs. scramble stimuli has been well studied. The event-related potential analysis of EEG/MEG shows that $N170$ occurs $130 - 200ms$ after stimulus presentation and reflects the neural processing of faces (Bentin et al., 1996; Henson et al., 2011). Therefore, the $N170$ component can be considered the ground truth for our analysis. (3) In the literature, non-parametric mass-univariate analysis such as cluster-based permutation tests is unable to identify narrowly distributed effects in space and time (e.g., an $N170$ component; Groppe et al., 2011a,b). These facts motivate us to employ multivariate approaches that are more sensitive to these effects.

As in Olivetti et al. (2014), we created a balanced face vs. scrambled MEG dataset by randomly drawing from the trials of unscrambled (famous or unfamiliar) faces and scrambled faces in equal number. The samples in the face and scrambled face categories are labeled as 1 and -1, respectively. The raw data is high-pass filtered at $1Hz$, down-sampled to $250Hz$, and trimmed from $200ms$ before the stimulus onset to $800ms$ after the stimulus. Thus, each trial has 250 time-points for each of the 306 MEG

[2] The full dataset is publicly available at ftp://ftp.mrc-cbu.cam.ac.uk/personal/rik.henson/wakemandg_hensonrn/.

[3] The competition data are available at http://www.kaggle.com/c/decoding-the-human-brain.

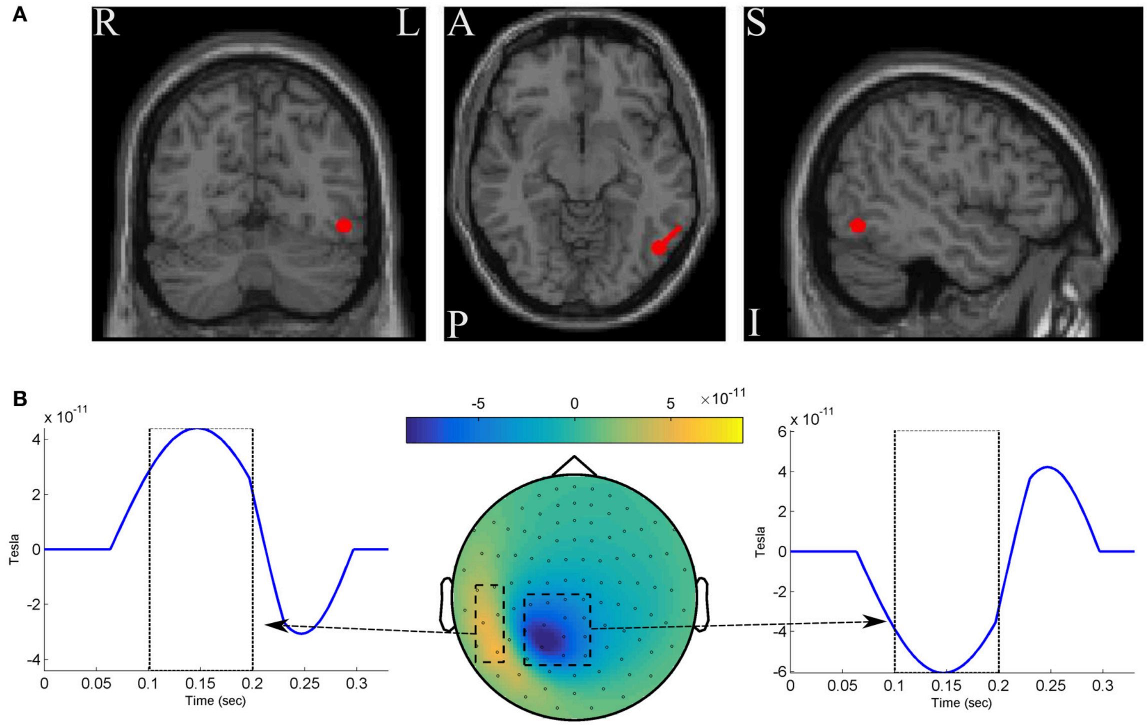


FIGURE 2 | (A) The red circles show the dipole position, and the red stick shows the dipole direction. **(B)** The spatio-temporal pattern of the discriminative ground-truth effect.

sensors (102 magnetometers and 204 planar gradiometers)[4]. To create the feature vector of each sample, we pooled all of the temporal data of 306 MEG sensors into one vector (i.e., we have $p = 250 \times 306 = 76500$ features for each sample). Before training the classifier, all of the features are standardized to have a mean of 0 and standard-deviation of 1.

2.7. Classification and Evaluation

In all experiments, Lasso (Tibshirani, 1996b) classifier with ℓ_1 penalization is used for decoding. Lasso is a very popular classification method in the context of brain decoding, mainly because of its sparsity assumption. The choice of Lasso, as a simple model with only one hyper-parameter, helps us to better illustrate the importance of including the interpretability in the model selection (see the supplementary materials for the results of the elastic-net; Zou and Hastie, 2005 classifier). The solution of decoding is computed by solving the following optimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{X}\Theta, \mathbf{Y}) + \lambda \, \|\Theta\|_1 \tag{16}$$

where $\|.\|_1$ represents the ℓ_1-norm, and λ is the hyper-parameter that specifies the level of regularization. Therefore, the aim of the model selection is to find the best value for λ on the training set S. Here, we try to find the best regularization parameter value among $\lambda = \{0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000\}$.

We use the out-of-bag (OOB) (Wolpert and Macready, 1999; Breiman, 2001) method for computing δ_Φ, ψ_Φ, $\tilde{\beta}_\Phi$, $\tilde{\eta}_\Phi$, and ζ_Φ for different values of λ. In OOB, given a training set $(\mathbf{X}, \mathbf{Y})$, m replications of bootstrap (Efron, 1992) are used to create perturbed training and validation sets (we set $m = 50$)[5]. In all of our experiments, we set $\omega_1 = \omega_2 = 1$ and $\kappa = 0.6$ in the computation of ζ_Φ. Furthermore, we set $\delta_\Phi = 1 - EPE$ where EPE indicates the expected prediction error; it is computed using the procedure explained in Appendix 6.4. Employing OOB provides the possibility of computing the bias and variance of the model as contributing factors in EPE.

3. RESULTS

3.1. Performance-Interpretability Dilemma: A Toy Example

In the definition of Φ^* on the toy dataset discussed in Section 2.6.1, x_1 is the decisive variable and x_2 has no effect on the classification of samples into target classes. Therefore, excluding the effect of noise and based on the theory of the maximal margin classifier (Vapnik and Kotz, 1982), $\vec{\Theta}^* \propto [1, 0]^T$ is the true solution to the decoding problem. By accounting for the effect of noise, solving the decoding problem in $(\mathbf{X}, \mathbf{Y})$ space yields $\vec{\hat{\Theta}} \propto [1/\sqrt{5}, 2/\sqrt{5}]^T$ as the parameters of the linear classifier. Although the estimated parameters on the noisy data provide

[4] The preprocessing scripts in python and MATLAB are available at: https://github.com/FBK-NILab/DecMeg2014/.

[5] The MATLAB code used for experiments is available at https://github.com/smkia/interpretability/.

the best generalization performance for the noisy samples, any attempt to interpret this solution fails, as it yields the wrong conclusion with respect to the ground truth (it says x_2 has twice the influence of x_1 on the results, whereas it has no effect). This simple experiment shows that the most accurate model is not always the most interpretable one, primarily because the contribution of the noise in the decoding process (Haufe et al., 2013). On the other hand, the true solution of the problem $\vec{\Theta}^*$ does not provide the best generalization performance for the noisy data.

To illustrate the effect of incorporating the interpretability in the model selection, a Lasso model with different λ values is used for classifying the toy data. In this example, because $\vec{\Theta}^*$ is known, the exact value of interpretability can be computed using Equation (5). **Table 1** compares the resultant performance and interpretability from Lasso. Lasso achieves its highest performance ($\delta_\Phi = 0.9884$) at $\lambda = 10$ with $\hat{\vec{\Theta}} \propto [0.4636, 0.8660]^T$ (indicated by the black dashed line in **Figure 3**). Despite having the highest performance, this solution suffers from a lack of interpretability ($\eta_\Phi = 0.4484$). By increasing λ, the interpretability improves so that for $\lambda = 500, 1000$ the classifier reaches its highest interpretability by compensating for 0.06 of its performance. Our observation highlights two main points:

1. In the case of noisy data, the interpretability of a decoding model can be possibly incoherent with its performance. Thus, optimizing the parameter of the model based on its performance does not necessarily improve its interpretability. This observation confirms the previous finding by Rasmussen et al. (2012) regarding the trade-off between the spatial reproducibility (as a measure for the interpretability) and the prediction accuracy in brain decoding.
2. If the right criterion is used in the model selection, employing proper regularization technique (sparsity prior, in the case of toy data) leads to more interpretable decoding models.

3.2. Decoding on Simulated MEG Data

With the main aim of comparing the quality of the heuristically approximated interpretability with respect to its actual value, we solve the decoding problem on the simulated MEG data where the ground-truth discriminative effect is known. The ground truth effect $\vec{\Theta}^*$ is used to compute the actual interpretability of the decoding model. On the other hand, interpretability is approximated by means of $\vec{\Theta}^{cERF}$. The whole data simulation and decoding processes are repeated 25 times and the results are summarized in **Figure 4**. **Figures 4A,B** show the actual

TABLE 1 | Comparison between δ_Φ, η_Φ, and ζ_Φ for different λ values on the toy example shows the performance-interpretability dilemma, in which the most accurate classifier is not the most interpretable one.

λ	**0**	**0.001**	**0.01**	**0.1**	**1**	**10**	**50**	**100**	**250**	**500**	**1000**
$\delta(\Phi)$	0.9883	0.9883	0.9883	0.9883	0.9883	**0.9884**	0.9880	0.9840	0.9310	0.9292	0.9292
$\eta(\Phi)$	0.4391	0.4391	0.4391	0.4392	0.4400	0.4484	0.4921	0.5845	0.9968	**1**	**1**
$\zeta(\Phi)$	0.7137	0.7137	0.7137	0.7137	0.7142	0.7184	0.7400	0.7842	0.9639	**0.9646**	**0.9646**
$\hat{\vec{\Theta}} \propto$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4521\\0.8919\end{bmatrix}$	$\begin{bmatrix}0.4532\\0.8914\end{bmatrix}$	$\begin{bmatrix}0.4636\\0.8660\end{bmatrix}$	$\begin{bmatrix}0.4883\\0.8727\end{bmatrix}$	$\begin{bmatrix}0.5800\\0.8146\end{bmatrix}$	$\begin{bmatrix}0.99\\0.02\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$

The bold indicates the best values of different criteria.

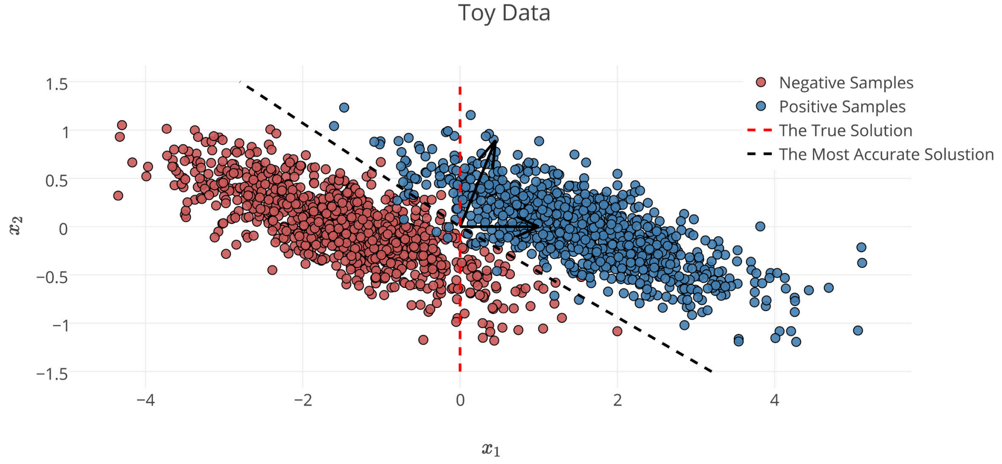


FIGURE 3 | Noisy samples of toy data. The red dashed line shows the true separator based on the generative model (Φ^*). The black dashed line shows the most accurate classification solution. Because of the contribution of noise, any interpretation of the parameters of the most accurate classifier yields a misleading conclusion with respect to the true underlying phenomenon (Haufe et al., 2013).

(η_Φ) and the approximated ($\tilde{\eta}_\Phi$) interpretability for different λ values. Even though $\tilde{\eta}_\Phi$ consistently overestimates η_Φ, there is a significant co-variation (Pearson's correlation p-value $= 9 \times 10^{-4}$) between two measures as λ increases. Thus, despite overestimation problem of the heuristically approximated interpretability values, they are still reliable measures for quantitative comparison between interpretability level of brain decoding models with different hyper-parameters. For example, both η_Φ and $\tilde{\eta}_\Phi$ suggest the decoding model with $\lambda = 50$ as the most interpretable model.

Figure 4C shows brain decoding models at $\lambda = 10$ and $\lambda = 50$ yield equivalent generalization performances (Wilcoxon rank sum test p-value $= 0.08$), while the MBM resulted from $\lambda = 50$ has significantly higher interpretability (Wilcoxon rank sum test p-value $= 4 \times 10^{-9}$). The advantage of this difference in interpretability levels is visualized in **Figure 5** where topographic maps are plotted for the weights of brain decoding models with different λ values by averaging the classifier weights in the time interval of 100–200 ms. The visual comparison shows MBM at $\lambda = 50$ is more similar to the ground-truth map (see **Figure 2B**) than the MBMs computed at other λ values. This superiority is well-reflected in the corresponding approximated interpretability values, that confirms the effectiveness of the interpretability criterion in measuring the level of information in the MBMs.

The results of this experiment confirm again the fact that the generalization performance is not a reliable criterion to measure the level of information learned by a linear classifier. For example consider the decoding model with $\lambda = 1$ in which the performance of the model is significantly above the chance level (see **Figure 4C**) while the corresponding MBM (**Figure 5A**) is completely misrepresents the ground-truth effect (**Figure 2B**).

3.3. Single-Subject Decoding on MEG Data

To investigate the behavior of the proposed model selection criterion ζ_Φ, we benchmark it against the commonly used performance criterion δ_Φ in a single-subject decoding scenario. Assuming $(\mathbf{X}_i, \mathbf{Y}_i)$ for $i = 1, \ldots, 16$ are MEG trial/label pairs for subject i, we separately train a Lasso model for each subject to estimate the parameter of the linear function $\hat{\Phi}_i$, where $\mathbf{Y}_i = \mathbf{X}_i\hat{\Theta}_i$. We represent the optimized solution based on δ_Φ and ζ_Φ by $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$, respectively. We also denote the MBM associated with $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ by $\vec{\hat{\Theta}}_i^\delta$ and $\vec{\hat{\Theta}}_i^\zeta$, respectively. Therefore, for each subject, we compare the resulting decoders and MBMs computed based on these two model selection criteria.

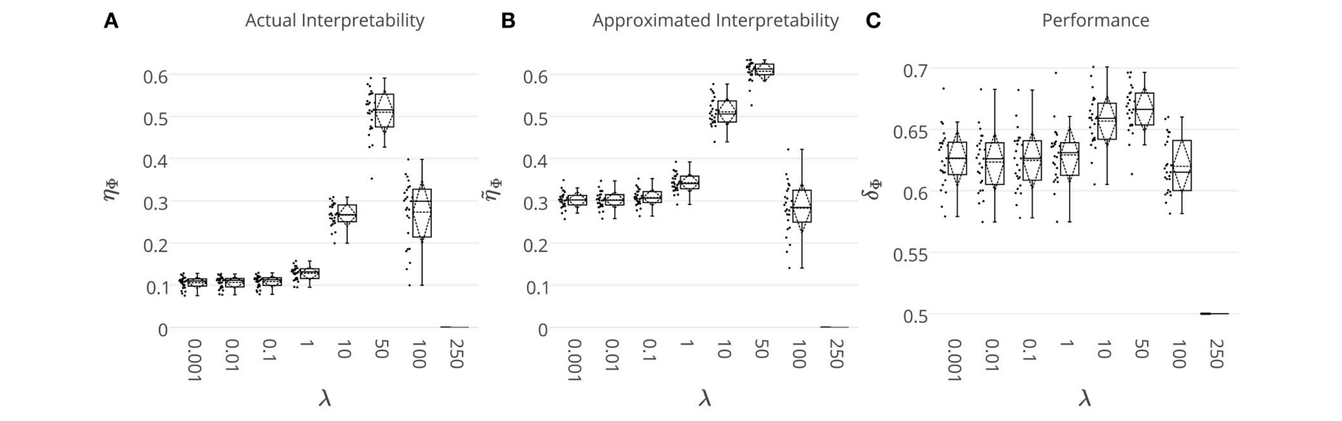


FIGURE 4 | (A) The actual η_Φ, and **(B)** the heuristically approximated interpretability $\tilde{\eta}_\Phi$ of decoding models across different λ values. There is a significant co-variation (Pearson's correlation p-value $= 9 \times 10^{-4}$) between η_Φ and $\tilde{\eta}_\Phi$. **(C)** The generalization performance of decoding models. The box gives the quartiles, while the whiskers give the 5 and 95 percentiles.

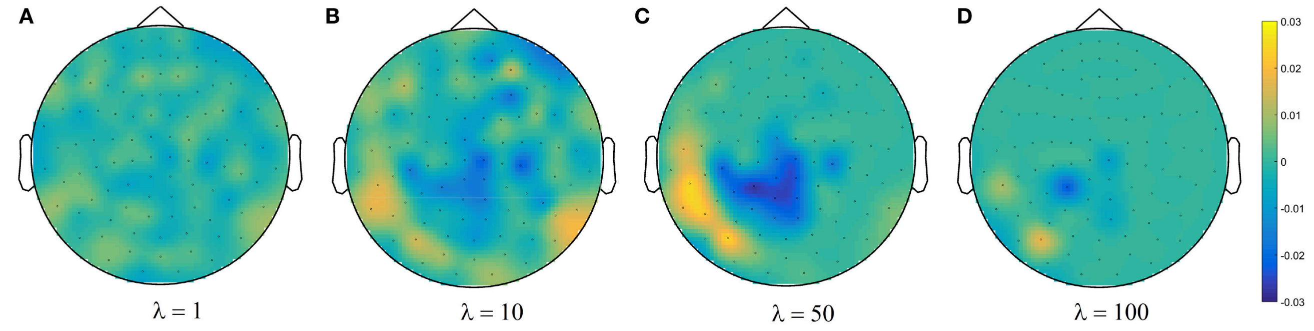


FIGURE 5 | Topographic maps of weights of brain decoding models for (A) $\lambda = 1$, **(B)** $\lambda = 10$, **(C)** $\lambda = 50$, and **(D)** $\lambda = 100$.

Figure 6A represents the mean and standard-deviation of the performance and interpretability of Lasso across 16 subjects for different λ values. The performance and interpretability curves further illustrate the performance-interpretability dilemma of Lasso classifier in the single-subject decoding scenario, in which increasing the performance delivers less interpretability. The average performance across subjects is improved when λ approaches 1, but on the other side, the reproducibility and the representativeness of models declines significantly (see **Figure 6B**; Wilcoxon rank sum test p-value $= 9 \times 10^{-4}$ and 8×10^{-7}, respectively). In fact, in this dataset a higher amount of sparsity increases the gap between the generalization performance and interpretability.

One possible reason behind the performance-interpretability dilemma in this experiment is illustrated in **Figure 6C**. The figure shows the mean and standard deviation of bias, variance, and EPE of Lasso across 16 subjects. The plot shows while the change in bias is correlated with that of EPE (Pearson's correlation coefficient $= 0.9993$), there is anti-correlation between the trends of variance and EPE (Pearson's correlation coefficient $= -0.8884$). Furthermore, it proposes that the effect of variance is overwhelmed by bias in the computation of EPE, where the best performance (minimum EPE) at $\lambda = 1$ has the lowest bias, its variance is higher than for $\lambda = 0.001, 0.01, 0.1$. While this tiny increase in the variance has negligible effect on the EPE of the model, **Figure 6B** shows its significant (Wilcoxon rank sum test p-value $= 8 \times 10^{-7}$) negative effect on the reproducibility of maps from $\lambda = 0.1$ to $\lambda = 1$.

Table 2 summarizes the performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ for 16 subjects. The average result over 16 subjects shows that employing ζ_{Φ} instead of δ_{Φ} in model selection provides higher reproducibility, representativeness, and (as a result) interpretability compensating for 0.04 of performance. The last column of table (δ_{cERF}) summarizes the performance of decoding models over 16 subjects when $\vec{\Theta}^{cERF}$ is used as classifier weights. The comparison illustrates a significant difference (Wilcoxon rank sum test p-value $= 1.5 \times 10^{-6}$) between δ_{cERF} and $\delta(\Phi)$s. These facts demonstrate that $\hat{\vec{\Theta}}^{\zeta}$ is a good compromise between $\hat{\vec{\Theta}}_{\delta}$ and $\vec{\Theta}^{cERF}$ in terms of classification performance and model interpretability.

These results are further analyzed in **Figure 7** where $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ are compared subject-wise in terms of their performance and interpretability. The comparison shows that adopting ζ_{Φ} instead of δ_{Φ} as the criterion for model selection yields higher interpretable models by compensating a negligible degree of performance in 14 out of 16 subjects. **Figure 7A** shows that employing δ_{Φ} provides on average slightly higher accurate models (Wilcoxon rank sum test p-value $= 0.012$) across subjects (0.83 ± 0.05) than using ζ_{Φ} (0.79 ± 0.04). On the other side, **Figure 7B** shows that employing ζ_{Φ} and compensating by 0.04 in the performance provides (on average) substantially higher (Wilcoxon rank sum test p-value$=5.6 \times 10^{-6}$) interpretability across subjects (0.62 ± 0.05) compared to δ_{Φ} (0.31 ± 0.12). For example, in the case of subject 1 (see **Table 2**), using δ_{Φ} in model selection to select the best λ value for the Lasso yields a model with $\delta_{\Phi} = 0.81$ and $\tilde{\eta}_{\Phi} = 0.26$. In contrast, using ζ_{Φ} delivers a model with $\delta_{\Phi} = 0.78$ and $\tilde{\eta}_{\Phi} = 0.63$. This inverse relationship between performance and interpretability is direct consequence of over-fitting of model to the noise structure in a small-sample size brain decoding problem (see Section 3.1).

The advantage of the exchange between the performance and the interpretability can be seen in the quality of MBMs. **Figures 8A,B** show $\hat{\vec{\Theta}}_1^{\delta}$ and $\hat{\vec{\Theta}}_1^{\zeta}$ of subject 1, i.e., the spatio-temporal multivariate maps of the Lasso models with maximum values of δ_{Φ} and ζ_{Φ}, respectively. The maps are plotted for 102 magnetometer sensors. In each case, the time course of weights of classifiers associated with the MEG2041 and MEG1931 sensors are plotted. Furthermore, the topographic maps represent the spatial patterns of weights averaged between 184ms and 236ms after stimulus onset. While $\hat{\vec{\Theta}}_1^{\delta}$ is sparse in time and space, it fails to accurately represent the spatio-temporal dynamic of the N170 component. Furthermore, the multicollinearity problem arising from the correlation between the time course of the MEG2041 and MEG1931 sensors causes extra attenuation of the N170 effect

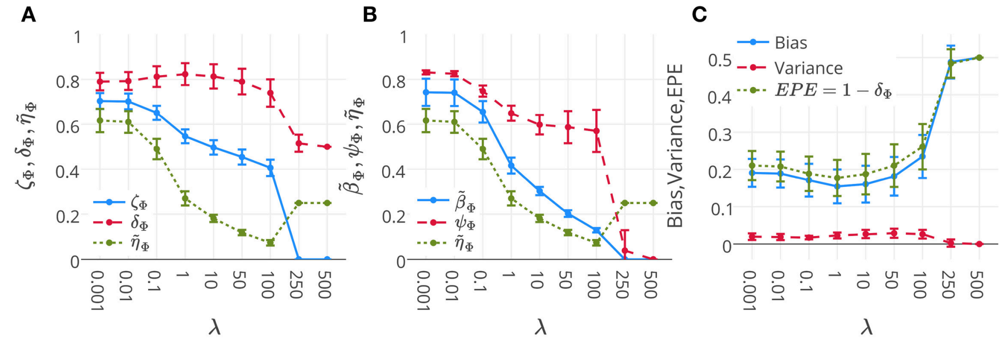


FIGURE 6 | (A) Mean and standard-deviation of the performance (δ_{Φ}), interpretability (η_{Φ}), and ζ_{Φ} of Lasso over 16 subjects. **(B)** Mean and standard-deviation of the reproducibility (ψ_{Φ}), representativeness (β_{Φ}), and interpretability (η_{Φ}) of Lasso over 16 subjects. The interpretability declines because of the decrease in both reproducibility and representativeness (see Proposition 1). **(C)** Mean and standard-deviation of the bias, variance, and EPE of Lasso over 16 subjects. While the change in bias is correlated with that of EPE (Pearson's correlation coefficient = 0.9993), there is anti-correlation between the trend of variance and EPE (Pearson's correlation coefficient = −0.8884).

TABLE 2 | The performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ over 16 subjects.

Subs	Criterion: $\delta(\Phi)$					Criterion: $\zeta(\Phi)$					δ_{cERF}
	$\delta(\Phi)$	$\zeta(\Phi)$	$\tilde{\eta}(\Phi)$	$\tilde{\beta}(\Phi)$	$\psi(\Phi)$	$\delta(\Phi)$	$\zeta(\Phi)$	$\tilde{\eta}(\Phi)$	$\tilde{\beta}(\Phi)$	$\psi(\Phi)$	
1	0.81	0.53	0.26	0.42	0.62	0.78	0.70	0.63	0.76	0.83	0.56
2	0.80	0.70	0.60	0.72	0.83	0.80	0.70	0.60	0.72	0.83	0.54
3	0.81	0.63	0.45	0.64	0.71	0.78	0.71	0.64	0.78	0.83	0.57
4	0.84	0.52	0.20	0.31	0.66	0.76	0.70	0.64	0.77	0.83	0.55
5	0.80	0.54	0.29	0.44	0.65	0.78	0.69	0.61	0.73	0.83	0.54
6	0.79	0.52	0.24	0.39	0.63	0.74	0.67	0.61	0.74	0.82	0.57
7	0.84	0.55	0.27	0.40	0.66	0.81	0.70	0.59	0.71	0.84	0.56
8	0.87	0.55	0.24	0.35	0.68	0.85	0.68	0.52	0.61	0.84	0.56
9	0.80	0.55	0.31	0.46	0.67	0.77	0.67	0.57	0.69	0.82	0.57
10	0.79	0.53	0.26	0.41	0.64	0.77	0.68	0.58	0.70	0.83	0.59
11	0.74	0.65	0.56	0.68	0.82	0.74	0.65	0.56	0.68	0.82	0.53
12	0.80	0.55	0.29	0.46	0.64	0.79	0.70	0.61	0.74	0.83	0.58
13	0.83	0.50	0.18	0.29	0.61	0.77	0.70	0.63	0.76	0.82	0.59
14	0.90	0.58	0.27	0.39	0.68	0.81	0.78	0.74	0.89	0.84	0.62
15	0.92	0.63	0.34	0.48	0.71	0.89	0.78	0.66	0.77	0.86	0.63
16	0.87	0.55	0.23	0.37	0.62	0.81	0.74	0.67	0.81	0.83	0.65
Mean	**0.83±0.05**	0.57 ± 0.05	0.31 ± 0.12	0.45 ± 0.13	0.68 ± 0.07	0.79 ± 0.04	**0.70 ± 0.04**	**0.62 ± 0.05**	**0.74 ± 0.06**	**0.83 ± 0.01**	0.58 ± 0.03

The bold indicates the best mean values over different criteria.

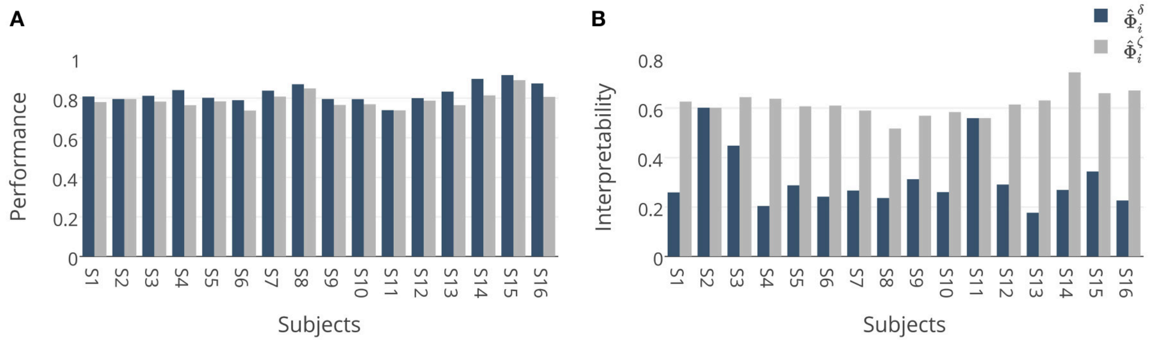


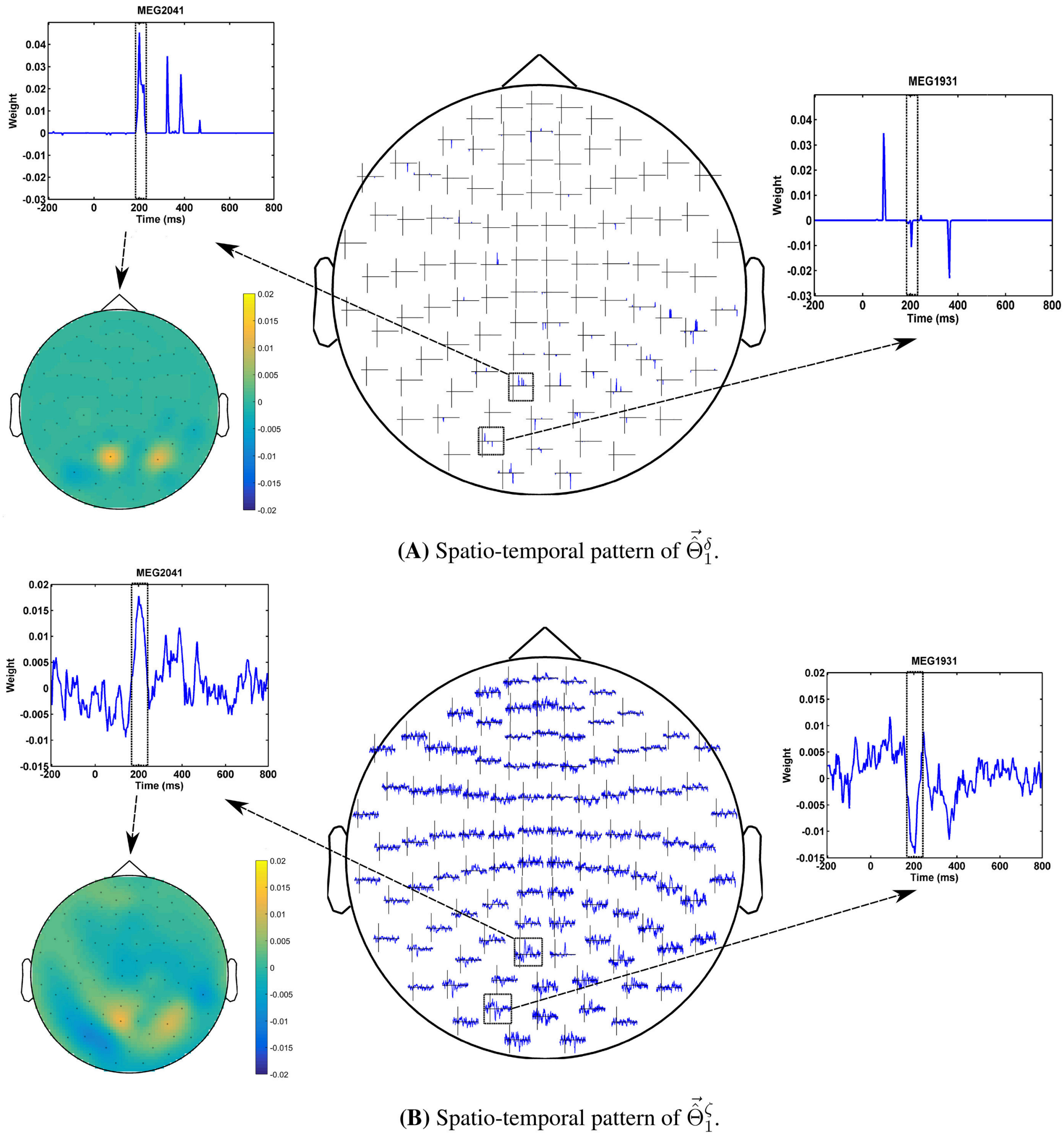
FIGURE 7 | (A) Comparison between generalization performances of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$. Adopting ζ_{Φ} instead of δ_{Φ} in model selection yields (on average) 0.04 less accurate classifiers over 16 subjects. **(B)** Comparison between interpretabilities of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$. Adopting ζ_{Φ} instead of δ_{Φ} in model selection yields on average 0.31 more interpretable classifiers over 16 subjects.

in the MEG1931 sensor. Therefore, the model is unable to capture the spatial pattern of the dipole in the posterior area. In contrast, $\vec{\hat{\Theta}}_1^{\zeta}$ represents the dynamic of the N170 component in time. In addition, it also shows the spatial pattern of two dipoles in the posterior and temporal areas. In summary, $\vec{\hat{\Theta}}_1^{\zeta}$ suggests a more representative pattern of the underlying neurophysiological effect than $\vec{\hat{\Theta}}_1^{\delta}$.

In addition, optimizing the hyper-parameters of brain decoding based on ζ_{Φ} offers more reproducible brain decoders. According to **Table 2**, using ζ_{Φ} instead of δ_{Φ} provides (on average) 0.15 more reproducibility over 16 subjects. To illustrate the advantage of higher reproducibility on the interpretability of maps, **Figure 9** visualizes $\vec{\hat{\Theta}}_1^{\delta}$ and $\vec{\hat{\Theta}}_1^{\zeta}$ over 4 perturbed training sets. The spatial maps (**Figures 9A,C**) are plotted for the magnetometer sensors averaged in the time interval between 184*ms* and 236*ms* after stimulus onset. The temporal maps (**Figures 9B,D**) are showing the multivariate temporal maps of MEG1931 and MEG2041 sensors. While $\vec{\hat{\Theta}}_1^{\delta}$ is unstable in time and space across the 4 perturbed training sets, $\vec{\hat{\Theta}}_1^{\zeta}$ provides more reproducible maps.

3.4. Mass-Univariate Hypothesis Testing on MEG Data

It is shown by Groppe et al. (2011a,b) that non-parametric mass-univariate analysis is unable to detect narrowly distributed effects in space and time (e.g., an *N*170 component). To



(A) Spatio-temporal pattern of $\vec{\hat{\Theta}}_1^{\delta}$.

(B) Spatio-temporal pattern of $\vec{\hat{\Theta}}_1^{\zeta}$.

FIGURE 8 | Comparison between spatio-temporal multivariate maps of (A) the most accurate, and **(B)** the most interpretable classifiers for Subject 1. $\vec{\hat{\Theta}}_1^{\zeta}$ provides a better spatio-temporal representation of the N170 effect than $\vec{\hat{\Theta}}_1^{\delta}$.

illustrate the advantage of the proposed decoding framework for spotting these effects, we performed a non-parametric cluster-based permutation test (Maris and Oostenveld, 2007) on our MEG dataset using Fieldtrip toolbox (Oostenveld et al., 2010). In a single subject analysis scenario, we considered the trials of MEG recordings as the unit of observation in a between-trials experiment. Independent-samples t-statistics are used as the statistics for evaluating the effect at the sample level and to construct spatio-temporal clusters. The maximum of the cluster-level summed t-value is used for the cluster level statistics; the significance probability is computed using a Monte Carlo method. The minimum number of neighboring channels for computing the clusters is set to 2. Considering 0.025 as the two-sided threshold for testing the significance level and repeating the procedure separately for magnetometers and combined-gradiometers, no significant result is found for any of the 16 subjects. This result motivates the search for more sensitive (and, at the same

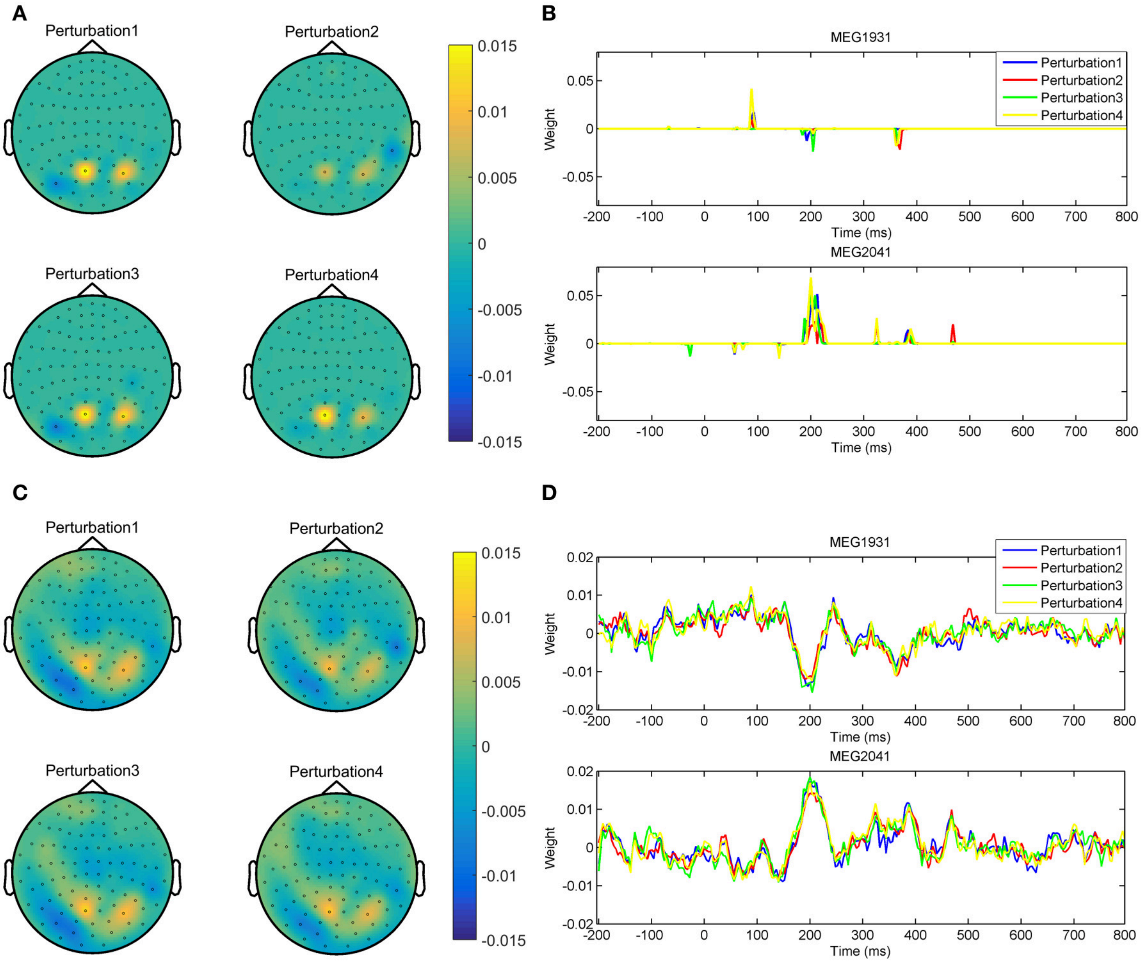


FIGURE 9 | Comparison of the reproducibility of Lasso when δ_Φ and ζ_Φ are used in the model selection procedure. (A,B) show the spatio-temporal patterns represented by $\vec{\tilde{\Theta}}_1^{\delta}$ across the 4 perturbed training sets. **(C,D)** show the spatio-temporal patterns represented by $\vec{\tilde{\Theta}}_1^{\zeta}$ across the 4 perturbed training sets. Employing ζ_Φ instead of δ_Φ in the model selection yields on average 0.15 more reproduciblilty of MBMs.

time, more interpretable) alternatives for univariate hypothesis testing.

4. DISCUSSIONS

4.1. Defining Interpretability: Theoretical Advantages

An overview of the brain decoding literature shows frequent co-occurrence of the terms interpretation, interpretable, and interpretability with the terms model, classification, parameter, decoding, method, feature, and pattern (see the quick meta-analysis on the literature in the supplementary material); however, a formal formulation of the interpretability is never presented. In this study, our primary interest is to present a simple and theoretical definition of the interpretability of linear brain decoding models and their corresponding MBMs. Furthermore, we show the way in which interpretability is related to the reproducibility and neurophysiological representativeness of MBMs. Our definition and quantification of interpretability remains theoretical, as we assume that the true solution of the brain decoding problem is available. Despite this limitation, we argue that the presented definition provides a concrete framework of a previously abstract concept and that it establishes a theoretical background to explain an ambiguous phenomenon in the brain decoding context. We support our argument using an example in the time-domain MEG decoding in which we show how the presented definition can be exploited to heuristically approximate the interpretability. Our experimental results on MEG data shows accounting for the approximated measure of interpretability has a positive effect on the human interpretation of brain decoding models. This example shows how partial prior knowledge regarding the timing and location of neural activity can be used to find more plausible multivariate patterns in data. Furthermore, the proposed decomposition of the interpretability of MBMs into their reproducibility and representativeness explains the relationship between the influential cooperative factors in the interpretability of brain decoding models and highlights the possibility of indirect and partial evaluation of interpretability by measuring these effective factors.

4.2. Application in Model Evaluation

Discriminative models in the framework of brain decoding provide higher sensitivity and specificity than univariate analysis in hypothesis testing of neuroimaging data. Although multivariate hypothesis testing is performed based solely on the generalization performance of classifiers, the emergent need for extracting reliable complementary information regarding the underlying neuronal activity motivated a considerable amount of research on improving and assessing the interpretability of classifiers and their associated MBMs. Despite ubiquitous use, the generalization performance of classifiers is not a reliable criterion for assessing the interpretability of brain decoding models (Rasmussen et al., 2012; Varoquaux et al., 2017). Therefore, considering extra criteria might be required. However, because of the lack of a formal definition for interpretability, different characteristics of linear classifiers are considered as the decisive criterion in assessing their interpretability. Reproducibility (Rasmussen et al., 2012; Conroy et al., 2013), stability selection (Varoquaux et al., 2012; Wang et al., 2015), sparsity (Dash et al., 2015; Shervashidze and Bach, 2015), and neurophysiological plausibility (Afshin-Pour et al., 2011) are examples of related criteria.

Our definition of interpretability helped us to fill this gap by introducing a new multi-objective model selection criterion as a weighted compromise between interpretability and generalization performance of linear models. Our experimental results on single-subject decoding showed that adopting the new criterion for optimizing the hyper-parameters of brain decoding models is an important step toward reliable visualization of learned models from neuroimaging data. It is not the first time in the neuroimaging context that a new metric is proposed in combination with generalization performance for the model selection. Several recent studies proposed the combination of the reproducibility of the maps (Rasmussen et al., 2012; Conroy et al., 2013; Strother et al., 2014) or the stability of the classifiers (Yu, 2013; Lim and Yu, 2016; Varoquaux et al., 2017) with the performance of discriminative models to enhance the interpretability of decoding models. Our definition of interpretability supports the claim that the reproducibility is not the only effective factor in interpretability. Therefore, our contribution can be considered a complementary effort with respect to the state of the art of improving the interpretability of brain decoding at the model selection level.

Furthermore, this work presents an effective approach for evaluating the quality of different regularization strategies for improving the interpretability of MBMs. As briefly reviewed in Section 1, there is a trend of research within the brain decoding context in which the prior knowledge is injected into the decoding process via the penalization term in order to improve the interpretability of decoding models. Thus far, in the literature, there is no *ad-hoc* method to directly compare the interpretability of MBMs resulting from different penalization techniques. Our findings provide a further step toward direct evaluation of interpretability of the currently proposed penalization strategies. Such an evaluation can highlight the advantages and disadvantages of applying different strategies on different data types and facilitates the choice of appropriate methods for a certain application.

4.3. Regularization and Interpretability

Haufe et al. (2013) demonstrated that the weight in linear discriminative models are unable to accurately assess the relationship between independent variables, primarily because of the contribution of noise in the decoding process. The authors concluded that the interpretability of brain decoding cannot be improved using regularization. The problem is primarily caused by the decoding process *per se*, where it minimizes the classification error only considering the uncertainty in the output space (Zhang, 2005; Aggarwal and Yu, 2009; Tzelepis et al., 2015) and not the uncertainty in the input space (or noise). Our experimental results on the toy data (see Section 3.1) shows that if the right criterion is used for selecting the best values for hyper-parameters, appropriate choice of the regularization strategy can still play a significant role in improving the interpretability of results. For example, in the case of toy data, the true generative function behind the sampled data is sparse (see Section 2.6.1), but because of the noise in the data, the sparse model is not the most accurate one. On the other hand, a more comprehensive criterion (in this case, ζ_Φ) that considers also the interpretability of model parameters facilitates the selection of correct prior assumptions about the distribution of the data via regularization. This observation encourages the modification of the conclusion in Haufe et al. (2013) as follows: if the performance of the model is the only criterion in the model selection, then the interpretability cannot necessarily be improved by means of regularization. This modification offers a practical shift in methodology, where we propose to replace the post-processing of weights proposed in Haufe et al. (2013) with refinement of hyper-parameter selection based on the newly developed model selection criterion.

4.4. The Performance-Interpretability Dilemma

The performance-interpretability dilemma refers to the trade-off between the generalization performance and the interpretability of a decoding model. In some applications of brain decoding, such as BCI, a more accurate model (even with no interpretability) is desired. On the other hand, when the brain decoding is employed for hypothesis testing purpose, an astute balance between two factors is more favorable. The presented metric for model selection (ζ_Φ) provides the possibility to maintain this balance. An important question at this point is on the nature of the performance-interpretability dilemma, whether it is model-driven or data-driven? In other words, whether some decoding models (e.g., sparse models) suffer from this deficit, or it is independent from the decoding model and depends on the distribution of data rather assumptions of the decoding model.

Our experiments shed light on the fact that the performance-interpretability dilemma is driven by the *uncertainty* (Aggarwal and Yu, 2009) in data. The uncertainty in data refers to the difference between the true solution of decoding Φ^* and the solution of decoding in sampled data space Φ_S, and is generally consequence of noise in the input or/and output spaces. This gap between Φ^* and Φ_S is also known as irreducible error (see

Equation 2) in the learning theory, and it cannot fundamentally be reduced by minimizing the error. Therefore, any attempt toward improving the classification performance in the sampled data space might increase the irreducible error. As an example, our experiment on the toy data (see Section 3.1) shows the effect of noise in input space on the performance-interpretability dilemma. Improving the performance of the model (i.e., fitting to Φ_S) diverges the estimated solution of decoding $\hat{\hat{\Phi}}$ from its true solution Φ^*, thus reduces the interpretability of the decoding model. Furthermore, our experiments demonstrate that incorporating the interpretability of decoding models in model selection facilitates finding the best match between the decoding model and the distribution of data. For example in classification of toy data, the new model selection metric ζ_Φ selects the more sparse model with a better match to the true distribution of data, despite worse generalization performance.

4.5. Advantage over Mass-Univariate Analysis

Mass-univariate hypothesis testing methods are among the most popular tools for forward inference on neuroimaging data in cognitive neuroscience field. Mass-univariate analyses consist of univariate statistical tests on single independent variables followed by multiple comparison correction. Generally, multiple comparison correction reduces the sensitivity of mass-univariate approaches because of the large number of univariate tests involved. Cluster-based permutation testing (Maris and Oostenveld, 2007) provides a more sensitive univariate analysis framework by making the cluster assumption in the multiple comparison correction. Unfortunately, this method is not able to detect narrow spatio-temporal effects in the data (Groppe et al., 2011a). As a remedy, brain decoding provides a very sensitive tool for hypothesis testing; it has the ability to detect multivariate patterns, but suffers from a low level of interpretability. Our study proposes a possible solution for the interpretability problem of classifiers, and therefore, it facilitates the application of brain decoding in the analysis of neuroimaging data. Our experimental results for the MEG data demonstrate that, although the non-parametric cluster-based permutation test is unable to detect the N170 effect in MEG data, employing ζ_Φ instead of δ_Φ in model selection not only detects the stimuli-relevant information in the data, but also assures both reproducible and representative spatio-temporal mapping of the timing and the location of underlying neurophysiological effect.

4.6. Limitations and Future Directions

Despite theoretical and practical advantages, the proposed definition and quantification of interpretability suffer from some limitations. All of the presented concepts are defined for linear models, with the main assumption that $\Phi^* \in \mathcal{H}$ (where $\mathcal{H}$ is a class of linear functions). This fact highlights the importance of linearizing the experimental protocol in the data collection phase (Naselaris et al., 2011). Extending the definition of interpretability to non-linear models demands future research into the visualization of non-linear models in the form of brain maps. Currently, our findings cannot be directly applied to non-linear models. Furthermore, the proposed heuristic for the time-domain MEG data applies only to binary classification. One possible solution in multiclass classification is to separate the decoding problem into several binary sub-problems. In addition the quality of the proposed heuristic is limited for the small sample size datasets. Of course the proposed heuristic is just an example of possible options for assessing the neurophysiological plausibility of MBMs in time-locked analysis of MEG data, thus, improving the quality of heuristic would be of interest in future researches. Finding physiologically relevant heuristics for other acquisition modalities such as fMRI, or frequency domain MEEG data, can be also considered as possible directions in future work.

5. CONCLUSIONS

We presented a novel theoretical definition for the interpretability of linear brain decoding and associated multivariate brain maps. We demonstrated how the interpretability relates to the representativeness and reproducibility of brain decoding. Although it is theoretical, the presented definition provides a first step toward practical solution for filling the knowledge extraction gap in linear brain decoding. As an example of this major breakthrough, and to provide a proof of concept, a heuristic approach based on the contrast event-related field is proposed for practical evaluation of the interpretability in multivariate recovery of evoked MEG responses. We experimentally showed that adding the interpretability of brain decoding models as a criterion in the model selection procedure yields significantly higher interpretable models by sacrificing a negligible amount of performance. Our methodological and experimental achievements can be considered a complementary theoretical and practical effort that contributes to researches on enhancing the interpretability of multivariate pattern analysis.

AUTHOR CONTRIBUTIONS

SK contributed in developing the theoretical and experimental contents of this study. SV and AP were involved in developing the theoretical machine learning aspects. NW was advising on the MEG experimental aspects.

ACKNOWLEDGMENTS

The authors wish to thank Raffaele Tucciarelli, Emanuele Olivetti, and Paolo Avesani for valuable discussion and comments. We would like to also thank reviewers for their insightful comments that improved our work significantly. This work has been partially funded by the IIT-FBK grant n.0030882/14.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2016.00619/full#supplementary-material

REFERENCES

Abadi, M., Subramanian, R., Kia, S., Avesani, P., Patras, I., and Sebe, N. (2015). DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Trans. Affect. Comput.* 6, 209–222. doi: 10.1109/TAFFC.2015.2392932

Afshin-Pour, B., Soltanian-Zadeh, H., Hossein-Zadeh, G.-A., Grady, C. L., and Strother, S. C. (2011). A mutual information-based metric for evaluation of fMRI data-processing approaches. *Hum. Brain Mapping* 32, 699–715. doi: 10.1002/hbm.21057

Aggarwal, C. C., and Yu, P. S. (2009). A survey of uncertain data algorithms and applications. *IEEE Transac. Knowl. Data Eng.* 21, 609–623. doi: 10.1109/TKDE.2008.190

Anderson, A., Labus, J. S., Vianna, E. P., Mayer, E. A., and Cohen, M. S. (2011). Common component classification: what can we learn from machine learning? *Neuroimage* 56, 517–524. doi: 10.1016/j.neuroimage.2010.05.065

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e130140. doi: 10.1371/journal.pone.0130140

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *J. Mach. Learn. Res.* 11, 1803–1831.

Bentin, S., Allison, T., Puce, A., Perez, E., and McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *J. Cogn. Neurosci.* 8, 551–565.

Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., and Garnero, L. (2007). Classification methods for ongoing EEG and MEG signals. *Biol. Res.* 40, 415–437. doi: 10.4067/S0716-97602007000500005

Bießmann, F., Dähne, S., Meinecke, F. C., Blankertz, B., Görgen, K., Müller, K.-R., et al. (2012). "On the interpretability of linear multivariate neuroimaging analyses: filters, patterns and their relationship," in *Proceedings of the 2nd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging* (Lake Tahoe: Harrahs and Harveys).

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of erp components a tutorial. *Neuroimage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Bousquet, O., and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* 2, 499–526.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Tittgemeyer, M., Buhmann, J. M., et al. (2011). Model-based feature construction for multivariate decoding. *Neuroimage* 56, 601–615. doi: 10.1016/j.neuroimage.2010.04.036

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., et al. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Mag. Reson. Med.* 35, 261–277.

Bzdok, D., Varoquaux, G., and Thirion, B. (2016). Neuroimaging research: from null-hypothesis falsification to out-of-sample generalization. *Educ. Psychol. Meas.* doi: 10.1177/0013164416667982. Available online at: http://journals.sagepub.com/doi/full/10.1177/0013164416667982

Caramia, M., and Dell'Olmo, P. (2008). *Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms*. London: Springer. 11–36. doi: 10.1007/978-1-84800-382-8

Carroll, M. K., Cecchi, G. A., Rish, I., Garg, R., and Rao, A. R. (2009). Prediction and interpretation of distributed neural activity with sparse models. *Neuroimage* 44, 112–122. doi: 10.1016/j.neuroimage.2008.08.020

Chan, A. M., Halgren, E., Marinkovic, K., and Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage* 54, 3028–3039. doi: 10.1016/j.neuroimage.2010.10.073

Conroy, B. R., Walz, J. M., and Sajda, P. (2013). Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fMRI decoding models. *PLoS ONE* 8:e79271. doi: 10.1371/journal.pone.0079271

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

Crivellato, E., and Ribatti, D. (2007). Soul, mind, brain: Greek philosophy and the birth of neuroscience. *Brain Res. Bull.* 71, 327–336. doi: 10.1016/j.brainresbull.2006.09.020

Cucker, F., and Smale, S. (2002). On the mathematical foundations of learning. *Am. Math. Soc.* 39, 1–49. doi: 10.1090/S0273-0979-01-00923-5

Dash, S., Malioutov, D. M., and Varshney, K. R. (2015). "Learning interpretable classification rules using sequential rowsampling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference* (South Brisbane, QLD: IEEE).

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., and Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? how subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage* 97, 271–283. doi: 10.1016/j.neuroimage.2014.04.037

de Brecht, M., and Yamagishi, N. (2012). Combining sparseness and smoothness improves classification accuracy and interpretability. *Neuroimage* 60, 1550–1561. doi: 10.1016/j.neuroimage.2011.12.085

Domingos, P. (2000). A unified bias-variance decomposition for zero-one and squared loss. *AAAI/IAAI* 2000, 564–569. Available online at: http://www.aaai.org/Library/AAAI/2000/aaai00-086.php

Efron, B. (1992). "Bootstrap methods: another look at the jackknife," in *Breakthroughs in Statistics: Methodology and Distribution*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 569–593. doi: 10.1007/978-1-4612-4380-9_41

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). "Identifying predictive regions from fMRI with TV-L1 prior," in *International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (Philadelphia, PA), 17–20.

Gramfort, A., Varoquaux, G., and Thirion, B. (2012). "Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify," in *Machine Learning and Interpretation in Neuroimaging*, eds G. Langs, I. Rish, M. Grosse-Wentrup, and B. Murphy (Berlin: Springer). 9–16.

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology* 48, 1711–1725. doi: 10.1111/j.1469-8986.2011.01273.x

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: simulation studies. *Psychophysiology* 48, 1726–1737. doi: 10.1111/j.1469-8986.2011.01272.x

Grosenick, L., Greer, S., and Knutson, B. (2008). Interpretable classifiers for fMRI improve prediction of purchases. *IEEE Trans. Neural Sys. Rehabilit. Eng.* 16, 539–548. doi: 10.1109/TNSRE.2008.926701

Grosenick, L., Klingenberg, B., Greer, S., Taylor, J., and Knutson, B. (2009). Whole-brain sparse penalized discriminant analysis for predicting choice. *Neuroimage* 47, S58. doi: 10.1016/S1053-8119(09)70232-0

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., and Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with graphnet. *Neuroimage* 72, 304–321. doi: 10.1016/j.neuroimage.2012.12.062

Hansen, K., Baehrens, D., Schroeter, T., Rupp, M., and Müller, K.-R. (2011). Visual interpretation of kernel-based prediction models. *Mol. Inform.* 30, 817–826. doi: 10.1002/minf.201100059

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, Vol. 2. New York, NY: Springer.

Haufe, S., Dähne, S., and Nikulin, V. V. (2014a). Dimensionality reduction for the analysis of brain oscillations. *Neuroimage* 101, 583–597. doi: 10.1016/j.neuroimage.2014.06.073

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2013). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.-D., Blankertz, B., et al. (2014b). "Parameter interpretation, regularization and source localization in multivariate linear models," in *International Workshop on Pattern Recognition in Neuroimaging, (PRNI)* (Tubingen: IEEE), 1–4.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron* 87, 257–270. doi: 10.1016/j.neuron.2015.05.025

Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi: 10.1038/nrn1931

Henson, R. N., Wakeman, D. G., Litvak, V., and Friston, K. J. (2011). A Parametric Empirical Bayesian framework for the EEG/MEG inverse problem: generative models for multisubject and multimodal integration. *Front. Hum. Neurosci.* 5:76. doi: 10.3389/fnhum.2011.00076

Huttunen, H., Manninen, T., Kauppi, J.-P., and Tohka, J. (2013). Mind reading with regularized multinomial logistic regression. *Mach. Visi. Appl.* 24, 1311–1325. doi: 10.1007/s00138-012-0464-y

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* 12, 2777–2824.

Kauppi, J.-P., Parkkonen, L., Hari, R., and Hyvärinen, A. (2013). Decoding magnetoencephalographic rhythmic activity using spectrospatial information. *Neuroimage* 83, 921–936. doi: 10.1016/j.neuroimage.2013.07.026

Kia, S. M., Vega-Pons, S., Olivetti, E., and Avesani, P. (2016). *Multi-Task Learning for Interpretation of Brain Decoding Models.* Cham: Springer International Publishing.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 1137–1143.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., and Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317–329. doi: 10.1016/j.neuroimage.2005.01.048

Langs, G., Menze, B. H., Lashkari, D., and Golland, P. (2011). Detecting stable distributed patterns of brain activation using gini contrast. *Neuroimage* 56, 497–507. doi: 10.1016/j.neuroimage.2010.07.074

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Lim, C., and Yu, B. (2016). Estimation stability with cross validation (escv). *J. Comput. Graphical Statist.* 25, 464–492. doi: 10.1080/10618600.2015.1020159

Lipton, Z. C., Kale, D. C., Elkan, C., Wetzell, R., Vikram, S., McAuley, J., et al. (2016). The mythos of model interpretability. *IEEE Spectrum.*

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology* 49, 549–565. doi: 10.1111/j.1469-8986.2011.01320.x

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Marler, R. T., and Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. *Struc. Multidiscipl. Optimiz.* 26, 369–395. doi: 10.1007/s00158-003-0368-6

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *Imaging IEEE Transac. Med.* 30, 1328–1340. doi: 10.1109/TMI.2011.2113378

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., et al. (2004). Learning to decode cognitive states from brain images. *Mach. Learn.* 57, 145–175. doi: 10.1023/B:MACH.0000035475.85309.1b

Montavon, G., Braun, M., Krueger, T., and Muller, K.-R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *Signal Process. Magaz. IEEE* 30, 62–74. doi: 10.1109/MSP.2013.2249294

Mørch, N., Hansen, L. K., Strother, S. C., Svarer, C., Rottenberg, D. A., Lautrup, B., et al. (1997). "Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover," in *Information Processing in Medical Imaging*, eds J. Duncan and G. Gindi (Berlin; Heidelberg: Springer), 259–270.

Naselaris, T., and Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn. Sci.* 19, 551–554. doi: 10.1016/j.tics.2015.07.005

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cognitive Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Olivetti, E., Kia, S. M., and Avesani, P. (2014). "MEG decoding across subjects," in *International Workshop on Pattern Recognition in Neuroimaging* (Tubingen: IEEE).

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2010). Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869. Available online at: https://www.hindawi.com/journals/cin/2011/156869/cta/

Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., et al. (2003). Single-trial detection in EEG and MEG: keeping it linear. *Neurocomputing* 52–54, 177–183. doi: 10.1016/s0925-2312(02)00821-4

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45, 199–209. doi: 10.1016/j.neuroimage.2008.11.007

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., and Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recogn.* 45, 2085–2100. doi: 10.1016/j.patcog.2011.09.011

Rieger, J. W., Reichert, C., Gegenfurtner, K. R., Noesselt, T., Braun, C., Heinze, H.-J., et al. (2008). Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. *Neuroimage* 42, 1056–1068. doi: 10.1016/j.neuroimage.2008.06.014

Rish, I., Cecchi, G. A., Lozano, A., and Niculescu-Mizil, A. (2014). *Practical Applications of Sparse Modeling.* Cambridge, MA: MIT Press

Rugg, M. D., and Coles, M. G. (1995). *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition.* Oxford: Oxford University Press.

Sabuncu, M. R. (2014). A universal and efficient method to compute maps from image-based prediction models. *Med. Image Comput. Comput. Assist. Intervent.* 17(Pt 3), 353–360. doi: 10.1007/978-3-319-10443-0_45

Shervashidze, N., and Bach, F. (2015). Learning the structure for structured sparsity. *IEEE Trans. Signal Process.* 63, 4894–4902. doi: 10.1109/TSP.2015.2446432

Spruill, M. C. (2007). Asymptotic distribution of coordinates on high dimensional spheres. *Electron. Commune. Probab.* 12, 234–247. doi: 10.1214/ECP.v12-1294

Strother, S. C., Rasmussen, P. M., Churchill, N. W., and Hansen, K. (2014). *Stability and Reproducibility in fMRI Analysis.* New York, NY: Springer-Verlag.

Taulu, S., Simola, J., Nenonen, J., and Parkkonen, L. (2014). "Novel noise reduction methods," *Magnetoencephalography: From Signals to Dynamic Cortical Networks*, eds S. Supek and C. J. Aine (Berlin; Heidelberg: Springer), 35–71. doi: 10.1007/978-3-642-33045-2_2

Tibshirani, R. (1996a). *Bias, Variance and Prediction Error for Classification Rules.* Toronto, ON: University of Toronto; Department of Statistics.

Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Ser. B (Methodol)* 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. Ser. B (Statist. Methodol.)* 67, 91–108. doi: 10.1111/j.1467-9868.2005.00490.x

Tzelepis, C., Mezaris, V., and Patras, I. (2015). Linear maximum margin classifier for learning from uncertain data. *arXiv preprint arXiv:1504.03892.*

Valentini, G., and Dietterich, T. G. (2004). Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *J. Mach. Learn. Res.* 5, 725–775.

Valverde-Albacete, F. J., and Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE* 9:e84217. doi: 10.1371/journal.pone.0084217

van Ede, F., and Maris, E. (2016). Physiological plausibility can increase reproducibility in cognitive neuroscience. *Trends Cogn. Sci.* 20, 567–569. doi: 10.1016/j.tics.2016.05.006

van Gerven, M., Hesse, C., Jensen, O., and Heskes, T. (2009). Interpreting single trial data using groupwise regularisation. *NeuroImage* 46, 665–676. doi: 10.1016/j.neuroimage.2009.02.041

Vapnik, V. N., and Kotz, S. (1982). *Estimation of Dependences Based on Empirical Data*, Vol. 40. New York, NY: Springer-verlag.

Varoquaux, G., Gramfort, A., and Thirion, B. (2012). "Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (Edinburgh), 1375–1382.

Varoquaux, G., Kowalski, M., and Thirion, B. (2016). "Social-sparsity brain decoders: faster spatial sparsity," in *Pattern Recognition in Neuroimaging, 2016 International Workshop on* (IEEE).

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Varoquaux, G., and Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience* 3:28. doi: 10.1186/2047-217X-3-28

Vellido, A., Martin-Guerroro, J., and Lisboa, P. (2012). "Making machine learning models interpretable," in *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* (Bruges), 163–172.

Vidaurre, D., Bielza, C., and Larrañaga, P. (2013). A survey of L1 regression. *Int. Statist. Rev.* 81, 361–387. doi: 10.1111/insr.12023

Wang, Y., Zheng, J., Zhang, S., Duan, X., and Chen, H. (2015). Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification. *Neuroimage* 117, 170–183. doi: 10.1016/j.neuroimage.2015.05.057

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage* 110, 48–59. doi: 10.1016/j.neuroimage.2015.01.036

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Wolpert, D. H., and Macready, W. G. (1999). An efficient method to estimate bagging's generalization error. *Machine Learning* 35, 41–55.

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

Xing, E. P., Kolar, M., Kim, S., and Chen, X. (2014). "High-dimensional sparse structured input-output models, with applications to gwas," in *Practical Applications of Sparse Modeling* (Cambridgem, MA: MIT Press), 37–64.

Yeung, N., Bogacz, R., Holroyd, C. B., and Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: an evaluation of methods. *Psychophysiology* 41, 822–832. doi: 10.1111/j.1469-8986.2004.00239.x

Yu, B. (2013). Stability. *Bernoulli* 19, 1484–1500. doi: 10.3150/13-BEJSP14

Yu, D., Lee, S. J., Lee, W. J., Kim, S. C., Lim, J., and Kwon, S. W. (2015). Classification of spectral data using fused lasso logistic regression. *Chemometrics Intell. Lab. Sys.* 142, 70–77. doi: 10.1016/j.chemolab.2015.01.006

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zhang, J. B. T. (2005). "Support vector classification with input data uncertainty," in *Advances in Neural Information Processing Systems*, eds L. K. Saul and Y. Weiss, and L. Bottou (Cambridge, MA: The MIT Press), 17, 161–168. Available online at: http://papers.nips.cc/paper/2743-support-vector-classification-with-input-data-uncertainty

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 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.

A. APPENDICES

A.1. Proof of Proposition 1

Throughout this proof, we assume that all of the parameter vectors are normalized in the unit hypersphere (see **Figure A1** as an illustrative example in two dimensions). Let $T = \{\vec{\hat{\Theta}}^1, \dots, \vec{\hat{\Theta}}^m\}$ be a set m MBMs, for m perturbed training sets where $\vec{\hat{\Theta}}^i \in \mathbb{R}^p$. Now, consider any arbitrary $p-1$ dimensional hyperplane $\mathcal{A}$ that contains $\vec{\Theta}^\mu$. Clearly, $\mathcal{A}$ divides the p-dimensional parameter space into two subspaces. Let $\triangledown$ and $\blacktriangledown$ be binary operators where $\vec{\Theta}^i \triangledown \vec{\Theta}^k$ indicates that $\vec{\Theta}^i$ and $\vec{\Theta}^k$ are in the same subspace, and $\vec{\Theta}^i \blacktriangledown \vec{\Theta}^k$ indicates that they are in different subspaces. Now, we define $T_U = \{\vec{\Theta}^i \mid \vec{\Theta}^i \triangledown \vec{\Theta}^*\}$ and $T_L = \{\vec{\Theta}^i \mid \vec{\Theta}^i \blacktriangledown \vec{\Theta}^*\}$. Let the cardinality of T_L denoted by $n(T_L)$ be j ($n(T_L) = j$). Thus, $n(T_U) = m - j$. Now, assume that $\measuredangle(\vec{\hat{\Theta}}^i, \mathcal{A}) = \alpha_1, \dots, \alpha_j$ are the angles between $\vec{\hat{\Theta}}^i \in T_L$ and $\mathcal{A}$, and (similarly) $\alpha_{j+1}, \dots, \alpha_m$ for $\vec{\hat{\Theta}}^i \in T_U$ and $\mathcal{A}$. Based on Equation (6), let $\vec{\Theta}^\mu_L$ and $\vec{\Theta}^\mu_U$ be the main maps of T_L and T_U, respectively. Therefore, we obtain $\vec{\Theta}^\mu = \frac{\vec{\Theta}^\mu_L + \vec{\Theta}^\mu_U}{\left\|\vec{\Theta}^\mu_L + \vec{\Theta}^\mu_U\right\|}$ and $\measuredangle(\vec{\Theta}^\mu_L, \mathcal{A}) = \measuredangle(\vec{\Theta}^\mu_U, \mathcal{A}) = \alpha$. Furthermore, assume $\measuredangle(\vec{\Theta}^*, \mathcal{A}) = \gamma$. As a result, $\psi_\Phi = \cos(\alpha)$ and $\beta_\Phi = \cos(\gamma)$. According to Equation (4) and using a cosine similarity definition, we have:

$$
\begin{aligned}
\eta_\Phi &= \frac{1}{m} \sum_{j=1}^{m} \left| \vec{\Theta}^* . \vec{\hat{\Theta}}^j \right| \\
&= \frac{\cos(\gamma + \alpha_1) + \dots + \cos(\gamma + \alpha_j) + \cos(\gamma - \alpha_{j+1}) + \dots + \cos(\gamma - \alpha_m)}{m} \\
&= \frac{\cos(\gamma + \alpha) + \cos(\gamma - \alpha)}{2} \\
&= \frac{\cos(\gamma)\cos(\alpha) - \sin(\gamma)\sin(\alpha) + \cos(\gamma)\cos(\alpha) + \sin(\gamma)\sin(\alpha)}{2} \\
&= \cos(\gamma)\cos(\alpha) = \beta_\Phi \times \psi_\Phi .
\end{aligned}
\tag{A1}
$$

A similar procedure can be used to prove $\tilde{\eta}_\Phi = \tilde{\beta}_\Phi \times \psi_\Phi$ by replacing $\vec{\Theta}^*$ with $\vec{\Theta}^{cERF}$.

A.2. Proof of Proposition 2

According to Haufe et al. (2013), for a linear discriminative model with parameters $\hat{\Theta}$, the unique equivalent generative model can be computed as follows:

$$A \propto \Sigma_{\mathbf{X}} \hat{\Theta} \tag{A2}$$

In a binary ($\mathbf{Y} = \{1, -1\}$) least squares classification scenario, we have:

$$A \propto \Sigma_{\mathbf{X}} \Sigma_{\mathbf{X}}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y} = \mu^+ - \mu^- \tag{A3}$$

where $\Sigma_{\mathbf{X}}$ represents the covariance of the input matrix $\mathbf{X}$, and μ^+ and μ^- are the means of positive and negative samples, respectively. Therefore, the equivalent generative model for the above classification problem can be derived by computing the difference between the mean of samples in two classes that is equivalent to the definition of cERF in time-domain MEG data.

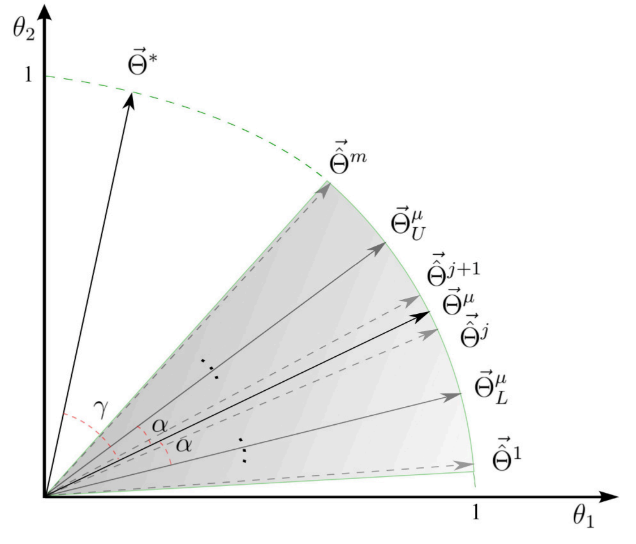


FIGURE A1 | Relation between representativeness, reproducibility, and interpretability in two dimensions.

A.3. The Distribution of Cosine Similarity

The aim of this section is to illustrate that the probability density function (PDF) of the cosine similarity between two randomly drawn vectors in the high dimensional space (large p) is very close to normal distribution with zero mean and small variance. To do this, we first need to find the distribution of dot product in the uniform unit hyper-sphere. Let a and b be two uniformly drawn random vectors from a unit hyper-sphere in $\mathbb{R}^p$. Assuming that γ is the angle between a and b, the distribution of cosine similarity is equivalent to the dot product $< a.b >$. Without loss of generality, let b be along the positive x-axis in the coordinate system. Thus, the dot product $< a.b >$ is the projection of a on the x-axis, i.e., x coordinate of a. Therefore, for a certain value of γ, the dot product is a $p-1$ dimensional hyper-sphere that is orthogonal to the x-axis (the red circle in **Figure A2**) and the PDF of the dot product is the surface area of p dimensional hyper-sphere constructed by the dot products for different γ values (the dashed blue sphere in **Figure A2**). To compute the area of this hyper-sphere we take the sum of the surface area of the p dimensional conical frustums over small intervals dx (the gray area in **Figure A2**):

$$
\begin{aligned}
Pr(-1 \leq x \leq 1) = &\quad 2^{p-2}\pi \int_{-1}^{1} (1 - x^2)^{p-2} \frac{dx}{1 - x^2} \\
= &\quad 2^{p-2}\pi \int_{-1}^{1} (1 - x^2)^{p-3} dx
\end{aligned}
\tag{A4}
$$

where $(1 - x^2)^{p-2}$ is the surface area of the base of the cone (e.g., the perimeter of the red circle in **Figure A2**) and $\frac{dx}{1-x^2}$ is the slope

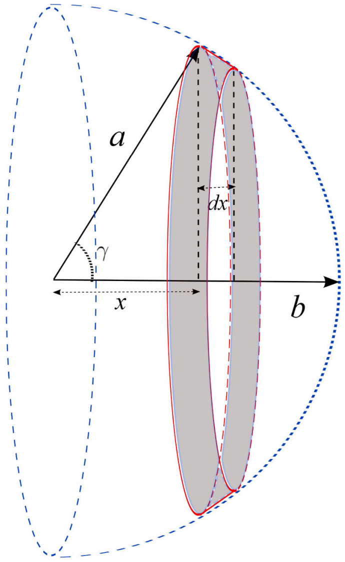


FIGURE A2 | Two-dimensional geometrical illustration for computing the PDF of cosine similarity.

size. Setting $t = \frac{x+1}{2}$ we have:

$$Pr(0 \leq t \leq 1) = 4^{p-2}\pi \int_0^1 t^{\frac{p-3}{2}}(1-t)^{\frac{p-3}{2}}\, dt \tag{A5}$$

which is a Beta distribution, where $\alpha = \beta = \frac{p-1}{2}$, that is a symmetric and unimodal distribution with mean 0.5. Because the PDF of $x = 2t - 1$ can be computed using a linear transformation of the above density function, it can be shown that the distribution of the dot product in unit hyper-sphere, i.e., the cosine similarity, is also a symmetric and unimodal distribution with 0 mean. Based on asymptotic assumption of Spruill (2007), for a large values of p this distribution converges to a normal distribution with $\sigma^2 = \frac{1}{p}$. Therefore assuming large p, the distribution of cosine similarity for uniformly random vectors drawn from p-dimensional unit hyper-sphere is approximately $\mathcal{N}(0, \sqrt{\frac{1}{p}})$.

A.4. Computing the Bias and Variance in Binary Classification

Here, using the out-of-bag (OOB) technique, and based on procedures proposed by Domingos (2000) and Valentini and Dietterich (2004), we compute the expected prediction error (EPE) for a linear binary classifier Φ under bootstrap perturbation of the training set. Let m be the number of perturbed training sets resulting from partitioning $S = (X, Y)$ into $S_{tr} = (X_{tr}, Y_{tr})$ and $S_{vl} = (X_{vl}, Y_{vl})$, i.e., training and validation sets. If $\hat{\Phi}^j$ is the linear classifier estimated from the jth perturbed training set, then the main prediction $\Phi^{\mu}(\mathbf{x}_i)$ for each sample in the dataset can be computed as follows:

$$\Phi^{\mu}(\mathbf{x}_i) = \begin{cases} 1 \ \ if & \frac{1}{k_i}\sum_{j=1}^{k_i} \hat{\Phi}^j(\mathbf{x}_i) \geq \frac{1}{2} \\ 0 & otherwise \end{cases} \tag{A6}$$

where k_i is the number of times that x_i is present in the test set[6].

The computation of bias is challenging because the optimal model Φ^* is unknown. According to Tibshirani (1996a), misclassification error is one of the loss measures that satisfies a Pythagorean-type equality, and:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^{\mu}(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^{\mu}(\mathbf{x}_i)) - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^*(\mathbf{x}_i)) \tag{A7}$$

Because all terms of the above equation are positive, the mean loss between the main prediction and the actual labels can be considered as an upper-bound for the bias:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^{\mu}(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^{\mu}(\mathbf{x}_i)) \tag{A8}$$

Therefore, a pessimistic approximation of bias $B(\mathbf{x}_i)$ can be calculated as follows:

$$B(\mathbf{x}_i) = \begin{cases} 0 \ \ if & \Phi^{\mu}(\mathbf{x}_i) = y_i \\ 1 & otherwise \end{cases} \tag{A9}$$

Then, the unbiased and biased variances (see Domingos, 2000 for definitions) in each training set can be calculated by:

$$V_u^j(\mathbf{x}_i) = \begin{cases} 1 \ \ if & B(\mathbf{x}_i) = 0 \ \ and \ \ \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 & otherwise \end{cases} \tag{A10}$$

$$V_b^j(\mathbf{x}_i) = \begin{cases} 1 \ \ if & B(\mathbf{x}_i) = 1 \ \ and \ \ \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 & otherwise \end{cases} \tag{A11}$$

Then, the expected prediction error of Φ can be computed as follows (ignoring the irreducible error):

$$EPE_{\Phi}(X) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} B(\mathbf{x}_i)}_{Bias} + \underbrace{\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}[V_u^j(\mathbf{x}_i) - V_b^j(\mathbf{x}_i)]}_{Variance} \tag{A12}$$

[6] It is expected that each sample $\mathbf{x}_i \in X$ appears (on average) $k_i \approx \frac{m}{3}$ times in the test sets.