Geometric properties of noninformative priors based on the chi-square divergence

Fuyuhiko Tanaka[1,2]*

[1]Center for Education in Liberal Arts and Sciences, Osaka University, Suita, Japan, [2]Center for Quantum Information and Quantum Biology, Osaka University, Suita, Japan

Recently, a noninformative prior distribution that is different from the Jeffreys prior was derived as an extension of Bernardo's reference prior based on the chi-square divergence. We summarize this result in terms of information geometry and clarify some geometric properties. Specifically, we show that it corresponds to a parallel volume element and can be written as a power of the Jeffreys prior in flat model manifolds.

KEYWORDS

noninformative prior, Jeffreys prior, reference prior, alpha-parallel prior, objective prior, chi-square divergence

1. Introduction

The problem of noninformative prior in Bayesian statistics is to determine what kind of probability distribution (often called a *noninformative prior* or an *objective prior*) is desirable on a statistical model in the absence of information about the parameters. In theory, though not in practice, it is essentially a problem of small-sample statistics, which has been under consideration for a long time [1–4].

Theoretical research on noninformative priors dates back to Jeffreys [3], and currently, a noninformative prior proposed by him, called the *Jeffreys prior*, is the standard noninformative prior. Theoretical justification of the Jeffreys prior comes from the theory of reference priors, which were originally proposed by Bernardo [5] decades ago when considering the maximization of the mutual information between the parameter and the outcome. Many related studies in this direction have since been reported [for review, see, e.g., Berger et al. [6]].

On the contrary, there are several criteria for considering noninformative priors. For example, Komaki [7, 8] has proposed objective priors to improve the performance of Bayesian predictive densities. Some significant results were presented by his co-workers, including the author [e.g., noninformative priors on time series models have been proposed [9, 10]]. From the viewpoint of information geometry, Takeuchi and Amari [11] proposed an α-parallel prior. For a recent review of other noninformative priors, see, e.g., Ghosh [12].

Recently, considering a certain extension of Bernardo's reference prior, Liu et al. [13] showed that a prior distribution different from the Jeffreys prior can be derived. Since it is based on the chi-square divergence, we call it χ^2-*prior* for convenience. Apart from the Jeffreys prior, the geometric properties of χ^2-prior are yet to be discussed.

In the present study, we investigate the derivation by Lie et al. of χ^2-prior from the viewpoint of information geometry. We put emphasis on the invariance of the theory under reparametrization (coordinate transformation in differential geometry). While we follow their derivation, we rewrite the asymptotic expansion in geometric terms, which makes the problem easier to understand. We also derive the tensor equations that χ^2-prior and an

α-parallel prior satisfy. As a consequence, we find that χ^2-prior agrees with an α-parallel prior for $\alpha = \frac{1}{2}$, i.e., the $\frac{1}{2}$-parallel prior.

Basic definitions and notation are given in Section 2. We also review some noninformative priors in terms of information geometry. In Section 3, we rewrite the asymptotic expansion by Liu et al. [13] in geometric terms to simplify their argument. In Section 4, we briefly review α-parallel priors, clarify a relation between χ^2-prior and α-parallel prior, and derive a formula of an α-parallel prior in γ-flat models. Finally, concluding remarks are given in Section 5.

2. Preliminaries

We briefly review some definitions and notation of information geometry [for details, refer to textbooks on information geometry [14, 15]]. We also review some noninformative priors in terms of information geometry.

For a given statistical model, we would like to consider noninformative prior distributions defined in a manner independent of parametrization. For this reason, it is convenient to introduce differential geometrical quantities into our discussion, i.e., to consider them from the viewpoint of information geometry.

2.1. Basic definitions of information geometry

Suppose that a statistical model $\mathcal{M} = \{p(x;\theta): \theta \in \Theta \subset \mathbf{R}^p\}$ is given, which is regarded as a p-dimensional differential manifold and called a *statistical model manifold* (though it will be called simply a *model* where no confusion is possible). As usual, all necessary regularity conditions are assumed.

We also define the Riemannian metric and affine connections on the manifold $\mathcal{M}$. Let $l = \log p(x;\theta)$ denote the log-likelihood function.

Definition 1. The Riemannian metric $g_{ij} = g(\partial_i, \partial_j)$ is defined as

$$g_{ij} = E[\partial_i l \partial_j l],$$

where $\partial_i l = \frac{\partial l}{\partial \theta^i} = \frac{\partial \log p(x;\theta)}{\partial \theta^i}$ and $E[\cdot]$ denotes expectation with respect to observation x. The above quantities are also called the *Fisher information matrix* in statistics. Thus, we often call the above metric the *Fisher metric*.

The statistical cubic tensor and the coefficients of the e-connection are defined as

$$T_{ijk} = E[\partial_i l \partial_j l \partial_k l],$$
$$\overset{(e)}{\Gamma}_{ij,k} = E\left[\partial_i \partial_j l \partial_k l\right].$$

Definition 2. For every real α, p^3 quantities

$$\overset{(\alpha)}{\Gamma}_{ij,k} = \overset{(e)}{\Gamma}_{ij,k} + \frac{1-\alpha}{2} T_{ijk}$$

define an affine connection, which is called the α-*connection*.

We identify an affine connection with its coefficients below. Connection coefficients with upper indices are obtained by

$$\overset{(e)}{\Gamma}{}_{ij}^{\ k} = \overset{(e)}{\Gamma}{}_{ij,l}\, g^{lk},$$
$$\overset{(\alpha)}{\Gamma}{}_{ij}^{\ k} = \overset{(e)}{\Gamma}{}_{ij}^{\ k} + \frac{1-\alpha}{2} T_{ijl} g^{lk},$$

where g^{ij} is the inverse matrix of the Fisher metric g_{ij}, and we have used Einstein's summation convention [see, e.g., Amari and Nagaoka [14] for details].

Conventionally, when $\alpha = 1$, we call it the *e-connection* and when $\alpha = -1$, we call it the *m-connection* and denote it as $\overset{(m)}{\Gamma}{}_{ij,k}$, i.e.,

$$\overset{(m)}{\Gamma}{}_{ij,k} = \overset{(e)}{\Gamma}{}_{ij,k} + T_{ijk}.$$

It is well-known that α-connection and $-\alpha$-connection are mutually dual with respect to the Fisher metric. (In a Riemannian manifold with an affine connection Γ, another affine connection Γ^* is said to be dual with respect to Γ if it satisfies $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}$. For equivalent definitions, see, e.g., Amari and Nagaoka [14], Chap. 3.) When $\alpha = 0$, the self-dual connection is called the *Levi-Civita connection*, which defines a parallel transport that keeps the Riemannian metric invariant. The Levi-Civita connection is defined by the sum of the partial derivative of the metric, and its explicit form is given by

$$\overset{(0)}{\Gamma}{}_{ij,k} = \overset{(e)}{\Gamma}{}_{ij,k} + \frac{1}{2} T_{ijk} = \frac{1}{2}(\partial_j g_{ki} + \partial_i g_{kj} - \partial_k g_{ij}). \quad (1)$$

2.2. Useful identities for alpha-connections

In the present study, the following identities are useful. They are obtained in a straightforward manner; thus, their proofs are omitted.

Lemma 1. Let $m_{ijk} = E[\partial_i \partial_j \partial_k l]$. Then,

$$m_{ijk} = -\partial_i g_{jk} - \overset{(e)}{\Gamma}{}_{jk,i},$$

and

$$\partial_k g_{ij} = \overset{(e)}{\Gamma}{}_{ki,j} + \overset{(e)}{\Gamma}{}_{kj,i} + T_{ijk},$$
$$= \overset{(e)}{\Gamma}{}_{ki,j} + \overset{(m)}{\Gamma}{}_{kj,i}.$$

hold.

The first equation yields relation (Equation 1). The last equation shows the duality of e- and m-connections directly and is generalized to $\pm\alpha$-connections.

Lemma 2. For mutually dual connections, the following identities hold.

$$\partial_k g_{ij} = \overset{(\alpha)}{\Gamma}{}_{ki,j} + \overset{(-\alpha)}{\Gamma}{}_{kj,i}, \quad (2)$$
$$\overset{(\alpha)}{\Gamma}{}_{ki,j} - \overset{(-\alpha)}{\Gamma}{}_{ki,j} = -\alpha T_{kij}. \quad (3)$$

Using Lemma 1 and Equation (1), we obtain the *Bartlett identity*, which is well-known in mathematical statistics.

Lemma 3. For m_{ijk}, T_{ijk}, and, the first derivative of Fisher metric, g_{ij}, the following holds.

$$m_{ijk} = \frac{1}{2}\left(T_{ijk} - \partial_k g_{ij} - \partial_j g_{ik} - \partial_i g_{jk}\right). \quad (4)$$

2.3. Prior distributions and volume elements

In Bayesian statistics, for a given statistical model $\mathcal{M}$, we need a probability distribution over the model parameter space, which is called *prior distribution*, or simply a *prior*. We often denote a prior density as π ($\pi(\theta) \geq 0$ and $\int_\Theta \pi(\theta)\,\theta = 1$).

A volume element on a p-dimensional model manifold corresponds to a prior density function over the parameter space ($\theta \in \Theta \subset \mathbf{R}^p$) in a one-to-one manner. For a prior $\pi(\theta)$, its corresponding volume element ω is a p-form (differential form of degree p) and is written as

$$\omega = \pi(\theta)d\theta^1 \wedge \cdots \wedge d\theta^p$$

in the local coordinate system.

For example, in two-dimensional Euclidian space ($p = 2$), the volume element is given by $\omega = dx \wedge dy$ in Cartesian coordinates (x, y). In polar coordinates (r, θ), it is written as $\omega = rdr \wedge d\theta$.

Then, under the coordinate transformation $\theta \rightarrow \xi$, how do the probability density on the parameter space and its ratio change? From the formula for the p-dimensional volume element, it is written as

$$\pi(\xi) = \pi(\theta)\left|\frac{\partial\theta}{\partial\xi}\right|, \quad (5)$$

where $\left|\frac{\partial\theta}{\partial\xi}\right|$ denotes the Jacobian. In differential geometry, such quantities are called *tensor densities*.

From the above Equation (5), we see that the ratio of two probability densities, say $\frac{\pi_1(\theta)}{\pi_2(\theta)}$, is invariant under reparametrization.

2.4. Noninformative priors defined by equations

We briefly summarize some of the prior studies on noninformative priors in Bayesian statistics. Basically, a noninformative prior is often defined as the solution of a partial differential equation (PDE) derived from fundamental principles. If it is independent of parametrization, then it usually has a geometrical meaning. The defining equation itself is expected to be invariant under every coordinate transformation.

2.4.1. Tensor equations

Before proceeding, we briefly review the definition of tensor on a manifold [for strict modern definitions, see, e.g., Kobayashi and Nomizu [16], Chap. 1].

For simplicity, we assume that the manifold admits global coordinates Θ, and each point is specified by θ. We fix some nonnegative integers r and s. Suppose that a set of p^{r+s} functions of the parameter θ

$$A^{b_1\ldots b_s}_{a_1\ldots a_r}(\theta), \qquad a_1,\ldots,a_r; b_1,\ldots,b_s = 1,\ldots,p$$

is given, and these functions also have a representation in a different coordinate system, say ξ. Suppose they satisfy the following equation:

$$\tilde{A}^{\beta_1\ldots\beta_s}_{\alpha_1\ldots\alpha_r}(\xi) = \tilde{\Lambda}^{a_1}_{\alpha_1}\ldots\tilde{\Lambda}^{a_r}_{\alpha_r}\Lambda^{\beta_1}_{b_1}\ldots\Lambda^{\beta_s}_{b_s}A^{b_1\ldots b_s}_{a_1\ldots a_r}(\theta),$$

where $\Lambda^\beta_b = \frac{\partial\xi^\beta}{\partial\theta^b}$ denotes the Jacobi matrix and $\tilde{\Lambda}^a_\alpha = \frac{\partial\theta^a}{\partial\xi^\alpha}$ denotes the inverse. Then these functions are called a *type (s, r) tensor field*, or simply a *tensor*.

Some specific types have established names. For example, a type $(0, 0)$ tensor is called a *scalar (field)* and a type $(1, 0)$ tensor is called a *vector (field)*. In particular, the ratio of two prior densities is a scalar. For a differential one-form, which is written as $A = A_j d\theta^j$, the set of components A_j is regarded as a contravariant vector [type $(0, 1)$ tensor].

For a type (s, r) tensor A, which often includes a derivative, we refer to an equation like $A = 0$ as a *tensor equation*. Usually, such a tensor A is derived using some differential operators, and the component-wise form yields a PDE. The component-wise form is given as

$$A^{b_1\ldots b_s}_{a_1\ldots a_r}(\theta) = 0, a_1,\ldots,a_r; b_1,\ldots,b_s = 1,\ldots,p.$$

By definition, tensor equations are invariant under coordinate transformation (reparametrization). When we show that $A^{b_1\ldots b_s}_{a_1\ldots a_r}(\theta) = 0$ for one coordinate system, say θ, then, for another coordinate system, say ξ, due to multilinearity,

$$\tilde{A}(\xi)^{\beta_1\ldots\beta_s}_{\alpha_1\ldots\alpha_r} = \tilde{\Lambda}^{a_1}_{\alpha_1}\ldots\tilde{\Lambda}^{a_r}_{\alpha_r}\Lambda^{\beta_1}_{b_1}\ldots\Lambda^{\beta_s}_{b_s}A(\theta)^{b_1\ldots b_s}_{a_1\ldots a_r} = 0$$

holds. Tensor equations are often written in the form $A = B$.

2.4.2. Noninformative priors

Now let us explain about noninformative priors [see, e.g., Robert [4] for more details]. As mentioned before, we need to set a prior distribution over the parameter space for a given statistical model in Bayesian statistics. If we have certain information on the parameter in advance, then the prior should reflect this, and such a prior is often called a *subjective prior*. If not, we adopt a certain criterion and use a prior obtained through the criterion. Such priors are called *noninformative priors*.

The definition of a noninformative prior, which is often written as a PDE, should not depend on a specific parametrization (a coordinate system of the model manifold). If we claim to have no information on the parameter, then we do not determine which parametrization is natural. Based on this viewpoint, we take several examples of noninformative priors defined through a PDE with a certain criterion. Some equations defining a noninformative prior are not tensor equations and their solutions, that is, noninformative priors, do not satisfy the equation in another coordinate system.

2.4.3. Uniform prior

The uniform prior $\pi_U(\theta)$ over the parameter space Θ would be the most naive noninformative prior. This idea dates back to Laplace and has been criticized [3]. The uniform prior is given by a solution of the following PDE:

$$\frac{\partial}{\partial \theta^i} \log \pi(\theta) = 0. \tag{6}$$

Clearly, the above PDE (Equation 6) is not a tensor equation. In other words, it is not invariant under reparametrization. While the solution for the original parameter θ is constant, $\pi_U(\theta) \propto 1$, the solution for another parameter ξ is obtained by

$$\tilde{\pi}_U(\xi) = \pi_U(\theta)\big|_{\theta=f(\xi)} \times \left|\frac{\partial \theta}{\partial \xi}\right| = \left|\frac{\partial \theta}{\partial \xi}\right|.$$

Thus, the final form does not satisfy the PDE (Equation 6) for ξ any more. That is,

$$\frac{\partial}{\partial \xi^\alpha} \log \tilde{\pi}_U(\xi) \neq 0.$$

2.4.4. Jeffreys prior

Let us modify the above PDE (Equation 6) slightly so that it is invariant under coordinate transformation. Thus, we obtain the following PDE:

$$\frac{\partial}{\partial \theta^i} \log\{\pi(\theta)/\sqrt{g}\} = 0, \tag{7}$$

where g denotes the determinant of the Fisher metric. The solution, which is given as a constant times $\sqrt{g}$, is called the *Jeffreys prior* [3]. It is the most famous noninformative prior in Bayesian statistics. Let $\pi_J(\theta)(\propto \sqrt{g})$ denote the Jeffreys prior from here on. It is the straightforward extension of the uniform prior.

As Jeffreys himself pointed out, it is not necessarily reasonable to adopt the Jeffreys prior as an objective prior in a higher dimensional parametric model. This is one of the reasons to propose noninformative priors under a fundamental criterion [see, e.g., Robert [4] and references therein].

Note that the following identity for the Riemannian metric tensor will be useful:

$$\partial_i \log \pi_J = \frac{1}{2} \partial_i \log g,$$
$$= \frac{1}{2} g^{jk} \partial_i g_{jk}. \tag{8}$$

2.4.5. First moment matching prior

The moment matching prior was proposed by Ghosh and Liu [17]. From the original article, we obtain a PDE in terms of information geometry.

Theorem 1. Ghosh and Liu's moment matching prior is given by the solution of the following PDE:

$$\partial_i \log\left(\frac{\pi(\theta)}{\pi_J(\theta)}\right) - \frac{1}{2} g^{jk}(\theta) \overset{(e)}{\Gamma}_{jk,i}(\theta) = 0.$$

From the aforementioned form, it is clearly not a tensor equation, and thus, the PDE is not invariant under reparametrization. Indeed, while the first term of the LHS is a $(0, 1)$ tensor, the second term is not.

Proof. First, from the formula in Ghosh and Liu [17] (Section 3, p. 193), we obtain

$$n(\hat{\theta}^m_\pi - \hat{\theta}^m_{ML}) = \left(U^m + \frac{1}{2}V^m\right) + o_P(1),$$

where $\hat{\theta}^m_{ML}$ and $\hat{\theta}^m_\pi$ are the MLE and the posterior mean of θ, respectively, $U^m = g^{ml}\partial_l \log \pi$ and $V^m = g^{ml}g^{jk}m_{ljk}$. Therefore, the condition of the first moment matching is given by

$$U^m + \frac{1}{2}V^m = 0.$$

Multiplying both sides with Fisher matrix g_{im}, we obtain an equivalent equation as follows:

$$\partial_i \log \pi + \frac{1}{2} g^{jk} m_{ijk} = 0.$$

Therefore, using Lemma 1 and Equation (8), we obtain

$$\partial_i \log \pi = -\frac{1}{2} g^{jk} m_{ijk}$$
$$= -\frac{1}{2} g^{jk} \left(-\partial_i g_{jk} - \overset{(e)}{\Gamma}_{jk,i}\right)$$
$$= \partial_i \log \sqrt{g} + \frac{1}{2} g^{jk} \overset{(e)}{\Gamma}_{jk,i}.$$

Since we may replace $\sqrt{g}$ with π_J in the last expression, we can obtain

$$\partial_i \log(\pi/\pi_J) - \frac{1}{2} g^{jk} \overset{(e)}{\Gamma}_{jk,i} = 0.$$

Remark 1. For the exponential family with the natural parameter θ, it is known that $\overset{(e)}{\Gamma}_{jk,i} \equiv 0$. When all connection coefficients vanish, the coordinate system is called *affine*. In this sense, the natural parameter is called the *e-affine coordinate*. From the above equation, in this parametrization, the moment matching prior agrees with the Jeffreys prior. However, if we begin with a different parametrization, then we obtain a prior which is different from the Jeffreys prior. As a specific example, let us consider the binomial model with the success probability η ($0 < \eta < 1$) in Ghosh [12] (Section 5.2, p. 199). Thereafter, the moment matching prior for η is given by $\pi_M(\eta) \propto \eta^{-1}(1-\eta)^{-1}$. However, taking the natural parameter $\theta = \log(\eta/(1-\eta))$, the moment matching prior for θ is given by the Jeffreys prior, $\pi_J(\theta) \propto e^{\theta/2}(1+e^\theta)^{-1}$. It is rewritten as $\pi_J(\eta) \propto \eta^{-1/2}(1-\eta)^{-1/2}$, which is different from $\pi_M(\eta)$.

2.4.6. Chi-square prior

Liu et al. [13] developed an extension of the reference prior by replacing the KL-divergence in the original definition with the general alpha-divergence. As an exceptional case, we obtain a prior which is different from the Jeffreys prior. The PDE is given by

$$\partial_i \log\left(\frac{\pi(\theta)}{\pi_J(\theta)}\right) = -\frac{1}{4} T_i, \tag{9}$$

where $T_i = T_{ijk}g^{jk}$ is a type $(0, 1)$ tensor. Thus, the above PDE is a tensor equation. Its derivation and details are explained in the next section.

Definition 3. [Liu et al. [13]]. If the PDE (Equation 9) has a solution, then we call the prior distribution χ^2-*prior*. We denote χ^2-prior as π_{χ^2}.

As we will see later, π_{χ^2} does not necessarily exist. However, usual statistical models satisfy a necessary and sufficient condition for the existence of π_{χ^2}. These models are invariant under coordinate transformation.

3. Derivation of chi-square prior in terms of information geometry

Liu et al. [13] derived the PDE (Equation 9) that π_{χ^2} should satisfy by considering the maximization of a functional of a prior π based on χ^2-divergence. In the present section, we review their result and rewrite the functional in terms of information geometry. As a result, we obtain a more explicit form and a better interpretation of the maximization.

3.1. Extension of the reference prior

As an underlying principle, Bernardo [5] adopted construction of the minimax code in information theory to derive noninformative priors. After that, the noninformative prior is defined as the input source distribution that maximizes the mutual information between the parameter and the outcome. This prior is called a *(Bernardo's) reference prior*. Under some conditions, his idea has been strictly formulated by several authors [18, 19] (for a review, see, e.g., Berger et al. [6]).

In one of the many studies and variants of reference priors, recently Liu et al. [13] adopted the α-divergence instead of the KL-divergence in Bernardo's argument and obtained a generalized result.

Definition 4. Let $p(x)$ and $q(x)$ be probability densities. For a fixed real parameter α, the α-*divergence* from p to q is defined as

$$D_\alpha(p; q) = \frac{1}{\alpha(1-\alpha)} \left\{ 1 - \int p(x)^\alpha q(x)^{1-\alpha} dx \right\} (\geq 0). \quad (10)$$

Remark 2. In the textbook on information geometry by Amari [20], the following parametrization is used because of the emphasis on the duality:

$$\widetilde{D}_\beta(p; q) = \frac{4}{1-\beta^2} \left\{ 1 - \int q(x)^{\frac{1+\beta}{2}} p(x)^{\frac{1-\beta}{2}} dx \right\}, \quad (11)$$

where we write β instead of α. We adopt the parametrization of α in Equation (10). For example, χ^2-divergence corresponds to $\alpha = -1$ in Equation (10) and $\beta = 3$ in Equation (11). More explicitly, the relation $\alpha = \frac{1-\beta}{2}$ (and thus, $1 - \alpha = \frac{1+\beta}{2}$) holds.

When $\alpha = 0, 1$, taking the limit, the α-divergence reduces to the KL-divergence.

Now, let us see the definition of the noninformative prior proposed by Liu et al. [13]. Under regularity conditions (e.g., the compactness of the parameter space Θ), they considered the maximization of the following functional of a prior density π as follows:

$$J[\pi] = \int E \left[D_\alpha \left(\pi(\cdot); \pi(\cdot|X) \right) | \theta \right] \pi(\theta) d\theta, \quad (12)$$

where $E[\cdot|\theta]$ denotes expectation with respect to $p(X|\theta)$, and the expression emphasizes that the parameter θ is fixed in the integral. Under their criterion, the maximizer of $J[\pi]$ is adopted as a noninformative prior.

Following Liu et al. [13], we rewrite the above functional Equation (12) in a more simple form as follows:

$$\alpha(1-\alpha)J[\pi] = 1 - \int E \left[\pi(\theta|X)^{-\alpha} | \theta \right] \pi(\theta)^{\alpha+1} d\theta.$$

Depending on the sign of $\alpha(1 - \alpha)$, our problem reduces to maximization or minimization of the expectation $E[\pi(\theta|X)^{-\alpha} | \theta]$. Clearly, it is not solved explicitly for general cases. Thus, as usual, we adopt the approximation of the expectation term under the assumption that $X = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} p(x|\theta)$ with $n \to \infty$.

3.2. Asymptotic expansion of the expectation term

Except for $\alpha = -1$ (χ^2-divergence), the maximization of $J[\pi]$ reduces to that of the first-order term in the following expansion (Theorem 2), which yields the Jeffreys prior for $-1 < \alpha < 1$. However, for χ^2-divergence, we need to evaluate the second-order term since the first-order term is constant.

First, we present a key result in Liu et al. [13]. Some notation in their result follows ours. For example, the Fisher information matrix and its determinant are denoted as g_{ij} and g, respectively. The dimension of the parameter θ is denoted as p. Please refer to the original article for technical details.

Theorem 2. [Liu et al. [13], Theorem 3.1] The expectation term $E[\pi(\theta|X)^{-\alpha} | \theta]$ in the functional $J[\pi]$ can be rewritten as

$$E[\pi(\theta|X)^{-\alpha} | \theta] = (2\pi)^{p\alpha/2} n^{-\frac{p\alpha}{2}} g^{-\frac{1}{2}\alpha} (1-\alpha)^{-\frac{p}{2}}$$
$$\left[1 + \frac{1}{n} \left\{ \cdots + s(\theta) \right\} + o(n^{-1}) \right],$$

where the $1/n$ part in braces $\{\cdots\}$ is given by

$$\frac{\alpha}{1-\alpha} \partial_j g^{ij} \cdot \partial_i \log \pi + \frac{-\alpha^2}{2(1-\alpha)} g^{ij} \frac{1}{g} \partial_j g \, \partial_i \log \pi$$
$$- \frac{\alpha^2}{2(1-\alpha)} m_{ijk} (\partial_l \log \pi) g^{ij} g^{kl}$$
$$- \frac{\alpha}{2} g^{ij} \partial_i \log \pi \cdot \partial_j \log \pi + \frac{2\alpha - \alpha^2}{2(1-\alpha)} g^{ij} \frac{\partial_i \partial_j \pi}{\pi} + s(\theta).$$

The last term $s(\theta)$ does not include the prior density π.

From Theorem 2, for a positive constant C_n and a sufficiently large n, the functional Equation (12) is approximated by

$$\alpha(1-\alpha)J[\pi] \simeq 1 - C_n \int g^{-\frac{1}{2}\alpha}\pi^{\alpha+1}d\theta.$$

When $-1 < \alpha < 1$, the maximization yields $\pi \propto g^{\frac{1}{2}}$, that is, the Jeffreys prior. When $\alpha < -1$, rather, the Jeffreys prior minimizes the functional $J[\pi]$.

However, at the boundary point $\alpha = -1$ (χ^2-divergence), the above first-order term becomes a constant independent of π. In this case, we need to evaluate the second-order term more carefully.

3.3. Rewriting Liu et al.'s Theorem 3.1 in geometrical terms

Now let us rewrite the second-order term of the asymptotic expansion in Theorem 2 in terms of information geometry. We fix $\alpha = -1$, and from here on, consider only the case for χ^2-divergence.

Although our approach differs from that in the original article, the final PDE agrees with their result. The difference and our contribution are discussed in the next subsection.

We summarize how we rewrite each term to obtain the final result (Theorem 3) later. First, we rewrite $\frac{\partial_i\partial_j\pi}{\pi}$ by using the following relation:

$$\frac{\partial_i\partial_j\pi}{\pi} = \partial_i\partial_j\log\pi + \partial_i\log\pi \cdot \partial_j\log\pi.$$

After that, we replace a prior density π with the density ratio $h = \pi/\pi_J$, where $\pi_J = \sqrt{g}$. The terms including the prior density π and its derivatives are expected to be written using the scalar function $\log h$. Indeed, this expectation is correct, and we obtain the final form after tedious, lengthy, and straightforward calculation. Because we use partial integrals in transforming the original form of the asymptotic expansion, the integral symbol remains in the expression below.

Theorem 3. [Liu et al. [13]], Corollary of Theorem 3.1.

$$\int E[\pi(\theta|X)|\theta]d\theta$$
$$= (2\pi)^{-p/2}n^{\frac{p}{2}}(2)^{-\frac{p}{2}}\left[\int \sqrt{g}\theta + \frac{1}{n}\int\{\cdots + s(\theta)\}\sqrt{g}d\theta + o(n^{-1})\right],$$

where the $1/n$ part in square brackets is given by

$$\frac{1}{n}\int\left\{-\frac{1}{4}\left\|d\log h + \frac{T}{4}\right\|^2 + \frac{1}{4}\left\|d\log\pi_J - \frac{T}{4}\right\|^2 + s(\theta)\right\}\sqrt{g}d\theta,$$
$$(13)$$

in which, we set $T := T_i d\theta^i$ and the norm of one-form A is defined as $\|A\|^2 := A_iA_jg^{ij}$.

The above one-form T is called the *Tchebychev form* in affine geometry [see, e.g., p. 58 in Simon et al. [21]].

From Theorem 3, maximizing $J[\pi]$ over the set of all prior densities is equivalent to maximizing the above integral with

respect to a scalar function h when $n \to \infty$. Since the second and third terms inside braces $\{\cdots\}$ in Equation (13) are independent of h, the expression achieves the maximum if the first term vanishes, that is,

$$d\log h + \frac{T}{4} = 0 \qquad (14)$$

holds. Thus, we obtain an equation of a differential one-form which determines χ^2-prior. In a proper coordinate system, the component-wise form of equation (14) is given by

$$\frac{\partial}{\partial\theta^i}\log h = -\frac{1}{4}T_i,$$

which agrees with the original PDE (Equation 9) derived in the previous study.

Finally, we discuss the existence of χ^2-prior. Generally, χ^2-prior does not necessarily exist on a statistical model. The existence of a χ^2-prior on a given model is equivalent to the existence of the solution of PDE (Equation 9).

A solution of PDE (Equation 9) exists if and only if T_i satisfies the following condition:

$$\partial_jT_i - \partial_iT_j = 0, \qquad (15)$$

which is called an *integrability condition* and is well-known. As a simplification, we may write $dT = 0$.

A bit surprisingly, the above condition (Equation 15) agrees with the condition that the $\alpha(\neq 0)$-parallel prior exists [11]. This implies a certain relationship between χ^2-prior and an α-parallel prior. Indeed, its expectation is correct and χ^2-prior is shown to be the $\frac{1}{2}$-parallel prior, which is the theme in the next section.

3.4. Discussion

We here discuss the difference between the original result obtained by Liu et al. [13] and the present study.

First, the PDE they obtained for χ^2-prior is not in the form of tensor equations. They gave a PDE for $\log\pi$ as follows

$$\partial_i\log\pi = -\frac{1}{4}T_{ijk}g^{jk} + \frac{1}{2}g^{-1}\partial_ig, \qquad (16)$$

instead of our PDE (Equation 9) [Liu et al. [13], p. 357, Equation (48)]. Both sides of Equation (16) are not tensors, i.e., not invariant under coordinate transformation.

Second, although Liu et al. obtained the asymptotic expansion as in Theorem 2 (Theorem 3.1 in the original article), their approach to derive the PDE (Equation 16) is not sufficient. Strictly speaking, they only show that π_{χ^2} satisfying the PDE (Equation 16) achieves the extreme value asymptotically. They did not organize messy terms and utilized the variational method in an *ad hoc* manner to derive the PDE (Equation 16). Moreover, their approach does not exclude the possibility of achieving the minimum.

Our approach shows more directly that π_{χ^2} satisfying the PDE (Equation 16)achieves the maximum of the functional asymptotically. Using the square completion for the one-form $d\log h$, we show that π_{χ^2} maximizes the functional $J[\pi]$ when $n \to \infty$.

In addition, our underlying philosophy is the invariance principle under coordinate transformation. Clearly, the expected χ^2-divergence from a prior to its posterior is independent of parametrization. Thus, we naturally expect that the $O(n^{-1})$ term is independent of parametrization, i.e., represented by geometrical quantities. As a result, we obtain a simpler expression (Equation 13) in Theorem 3. This is a good example of how organizing from the viewpoint of information geometry can simplify various terms and make the structure of the problem easier to understand.

As for derivation of fundamental PDEs, we point out a formal analogy between general relativity and ours. Historically speaking, Hilbert showed that the Einstein equation is derived from Einstein–Hilbert action integral $S[g_{ab}]$, where g_{ab} is the pseudo-Riemannian metric on the time-space manifold [see, e.g., Wald [22], Appendix E.1]. In our problem, we take the expected χ^2-divergence from a prior to its posterior instead of $S[g_{ab}]$. The maximization of $J[\pi]$ and minimization of $S[g_{ab}]$ yield the tensor equation (Equation 16) and the Einstein equation, respectively.

4. Relation between chi-square priors and alpha-parallel priors

In this section, we show that χ^2-prior is the $\frac{1}{2}$-parallel prior, a special case of an α-parallel prior. As we shall see later, an α-parallel prior is defined through an α-parallel volume element and was proposed by Takeuchi and Amari [11]. Among several existence conditions for an α-parallel prior, we focus on the PDE of $\log \pi$ and rewrite it in terms of the log ratio $\log h$.

In the exponential family, χ^2-prior and α-parallel priors were derived by the two author groups, and Takeuchi and Amari [11] and Liu et al. [13], respectively. We also generalize this result for any α-flat model.

4.1. Alpha-parallel priors

Takeuchi and Amari [11] introduced a family of geometric priors called α-parallel priors, which include the well-known Jeffreys prior and maximum likelihood (ML) prior [23]. We briefly review basic definitions and related results on α-parallel priors below.

4.1.1. Equiaffine connection

First, we recall the definition of equiaffine connection in affine geometry. Let us consider a p-dimensional orientable smooth manifold $\mathcal{M}$ with an affine connection ∇. We shall say that a torsion-free affine connection ∇ is *equiaffine* when there exists a parallel volume element, that is, a nonvanishing p-form ω such that $\nabla \omega = 0$.

One necessary and sufficient condition for ∇ to be equiaffine is

$$R_{ijk}^{\nabla \ k} = 0, \tag{17}$$

where $R_{ijk}^{\nabla \ l}$ is the Riemann–Christoffel curvature tensor with respect to the connection ∇. The condition (Equation 17) is slightly weaker than the condition that an affine manifold is flat, $R_{ijk}^{\nabla \ l} = 0$.

4.1.2. Definition of alpha-parallel prior

Here, we develop an aforementioned argument in statistical models. Since statistical models have a family of affine connections in a natural manner, we expect that the condition of being an equiaffine connection is obtained as a property of model manifolds rather than one of affine connections.

Let a p-dimensional statistical model manifold $\mathcal{M}$ be given. We assume that it is covered by a single coordinate, say, $\theta \in \Theta \subseteq \mathbf{R}^p$, orientable, and simply connected.

Definition 5. Suppose that there exists a parallel volume element ω for a fixed α, i.e., $\overset{(\alpha)}{\nabla} \omega = 0$. Therefore, in a coordinate system, say, $\theta = (\theta^1, \ldots, \theta^p)$, α-parallel volume element ω is represented as

$$\omega \propto \pi(\theta) d\theta^1 \wedge \cdots \wedge d\theta^p,$$

where π is a nonnegative function over the parameter space Θ. We call π an α-*parallel prior*.

Some examples of α-parallel prior are as follows: when $\alpha = 1$, the 1-parallel prior (also called the e-parallel prior) is the so-called ML prior proposed by Hartigan [23]; and when $\alpha = 0$, the 0-parallel prior is the Jeffreys prior. As we shall see later, the 0-parallel prior is exceptional and always exists on a statistical model. Indeed, a 0-parallel volume element, $\sqrt{g} d\theta^1 \wedge \cdots \wedge d\theta^p$, is known as the invariant volume element on a Riemannian manifold $(\mathcal{M}, g_{ij})$ with the Levi-Civita connection (0-connection in information geometry). Note that an α-parallel prior could be an improper prior. For other properties of α-parallel priors, see Takeuchi and Amari [11].

4.1.3. Existence conditions of alpha-parallel prior

In statistical models, we obtain a deeper result for the existence of an α-parallel prior. First, we note that the relation

$$R_{ijk}^{\overset{(\alpha)}{\nabla} k} = -\frac{\alpha}{2}(\partial_i T_j - \partial_j T_i) \tag{18}$$

holds for every α. From the necessary and sufficient condition for the existence of α-parallel prior (Equation 17), we find that the 0-parallel prior ($\alpha = 0$) necessarily exists. For $\alpha \neq 0$, we introduce the concept of statistically equiaffine.

Definition 6. A statistical model manifold $\mathcal{M}$ is said to be *statistically equiaffine* [11], when the cubic tensor T_{ijk} satisfies the following condition:

$$\partial_j T_i - \partial_i T_j = 0. \tag{19}$$

Observing the existence condition for α-parallel prior (Equation 17) and the relation (Equation 18), we easily obtain the following theorem.

Theorem 4. For a statistical model manifold $\mathcal{M}$, the following conditions are equivalent:

(a) an α-parallel volume element exists for all α,
(b) an α-parallel volume element exists for α ($\neq 0$),
(c) $\mathcal{M}$ is statistically equiaffine.

Note that a weaker condition (b) implies stronger conditions (a) and (c). The usual statistical models have been shown to be statistically equiaffine [11]. An important statistical model that is not statistically equiaffine is the ARMA model in time series analysis [24].

4.2. Chi-square prior is the half-parallel prior

Now let us consider a relation between α-parallel prior and χ^2-prior. To compare them, we focus on the following PDE for an α-parallel prior:

$$\partial_i \log \pi = g^{jk} \overset{(\alpha)}{\Gamma}_{ij,k} \qquad (20)$$

[Takeuchi and Amari [11], Proposition 1, p. 1016, Equation (7)].

Since both sides of the PDE (Equation 20) are not tensors, its invariance under coordinate transformation is not clear. Thus, we introduce a one-form (geometrical quantity) derived from a scalar function $h = \pi/\pi_J$ and modify the equation.

Theorem 5. The above PDE (Equation 20) is equivalent to the following tensor equation:

$$\partial_i \log h = -\frac{\alpha}{2} T_i. \qquad (21)$$

When we set $T = T_i d\theta^i$, then the above equation (21) can be rewritten as

$$d \log h + \frac{\alpha}{2} T = 0,$$

which is the equation of a differential one-form.

Proof. Using Equation (8), we rewrite the PDE (Equation 20) as follows:

$$\partial_i \log h = \left(-\frac{1}{2} \partial_i g_{jk} + \overset{(\alpha)}{\Gamma}_{ij,k} \right) g^{jk}. \qquad (22)$$

Therefore, using Lemma 2, we modify the RHS of Equation (22):

$$
\begin{aligned}
\left(-\frac{1}{2} \partial_i g_{jk} + \overset{(\alpha)}{\Gamma}_{ij,k} \right) g^{jk} &= \left(-\frac{1}{2} \overset{(\alpha)}{\Gamma}_{ij,k} - \frac{1}{2} \overset{(-\alpha)}{\Gamma}_{ik,j} + \overset{(\alpha)}{\Gamma}_{ij,k} \right) g^{jk}, \\
&= \left(-\frac{1}{2} \overset{(-\alpha)}{\Gamma}_{ik,j} + \frac{1}{2} \overset{(\alpha)}{\Gamma}_{ki,j} \right) g^{jk}, \\
&= -\frac{1}{2} \alpha T_{ikj} g^{jk}, \\
&= -\frac{\alpha}{2} T_i.
\end{aligned}
$$

Surprisingly, the PDE defining π_{χ^2} (Equation 9) agrees with Equation (21) with $\alpha = \frac{1}{2}$. Thus, χ^2-prior derived by Liu et al. [13] is the $\frac{1}{2}$-parallel prior. This finding is interesting in two ways.

First, to Bayesian statistics, it is a new example where the formulation in terms of information geometry is useful to research on noninformative priors [for several examples, see Komaki [7] and Tanaka and Komaki [9]]. Liu et al. [13] derived the PDE (Equation 9) by considering one extension of the reference prior with χ^2-divergence. Their starting point is completely independent of the

geometry of statistical models. In spite of this, χ^2-prior has a good geometrical interpretation: it is volume element invariant under the parallel transport with respect to the $\frac{1}{2}$-connection.

Second, to information geometry, it would be the first specific example where only the $\frac{1}{2}$-connection makes sense in statistical applications. In information geometry, the meaning of each connection among α-connections has not been clarified enough except for specific alphas ($\alpha = 0, \pm 1$). In Takeuchi and Amari [11], α-parallel priors were not proposed as noninformative priors. Rather, they regarded the Jeffreys prior as the 0-parallel prior and extended it to every α. Except for $\alpha = 0$, only the $\frac{1}{2}$-parallel prior is interpreted as a noninformative prior.

4.3. General form of alpha-parallel priors in statistically equiaffine models

Let us derive a general form of α-parallel priors in statistically equiaffine models. In the following, we denote an α-parallel prior as π_α. For example, $\pi_{1/2} = \pi_{\chi^2}$ and $\pi_0 = \pi_J$.

First, we briefly review some formulas for α-parallel priors derived by several authors [11, 25]. According to Matsuzoe et al. [25], there exists a scalar function ϕ that satisfies $T_a = \partial_a \phi$ on a statistically equiaffine model manifold. Therefore, using this function ϕ, a general solution of the PDE (Equation 21) is given by $h(\theta) \propto \exp\left\{ -\frac{\alpha}{2}\phi(\theta) \right\}$. Thus, we obtain α-parallel prior π_α as having the following form:

$$\pi_\alpha(\theta) \propto \exp\left\{ -\frac{\alpha}{2}\phi(\theta) \right\} \pi_J(\theta).$$

In the exponential family (e-flat model), Takeuchi and Amari show that α-parallel priors are representable as a power of Jeffreys prior π_J for every α [Takeuchi and Amari [11], Example 2, p. 1017].

Here, we extend their result for the exponential family to a $\gamma (\gamma \neq 1)$-flat model. We use the parameter γ instead of α because the two parameters may be different.

Theorem 6. Let $\gamma \neq 0$. Suppose that a statistical model manifold $(\mathcal{M}, g, \overset{(\gamma)}{\nabla}, \overset{(-\gamma)}{\nabla})$ is γ-flat. Then, there exists an α-parallel prior π_α for every α. In a γ-affine coordinate system, $\{\theta^i\}$, it is written as a power of Jeffreys prior π_J, that is,

$$\pi_\alpha(\theta) \propto \pi_J(\theta)^{1 - \frac{\alpha}{\gamma}}$$

holds.

Proof. Since the model is γ-flat, we take a γ-affine coordinate system, say, $\{\theta^i\}$. Then, $\overset{(\gamma)}{\Gamma}_{ij,k} = 0$ and from Equation (3) in Lemma 2,

$$T_{ijk} = \frac{1}{\gamma} \overset{(-\gamma)}{\Gamma}_{ij,k} \qquad (23)$$

holds.

On the contrary, from a result in Amari and Nagaoka [14], Section 3.3, there exists a scalar function $\psi(\theta)$ such that

$$g_{ij} = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\psi,$$

$$\overset{(-\gamma)}{\Gamma}_{ij,k} = \frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k}\psi$$

for the γ-affine coordinates $\{\theta^i\}$. This implies that

$$\overset{(-\gamma)}{\Gamma}_{ij,k} = \partial_k g_{ij} = \frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k}\psi. \tag{24}$$

Therefore, using Eqs (23) and (24), we can rewrite T_i as follows:

$$\begin{aligned}
T_i &= g^{jk}T_{ijk} \\
&= \frac{1}{\gamma}g^{jk}\overset{(-\gamma)}{\Gamma}_{ij,k}, \\
&= \frac{1}{\gamma}g^{jk}\partial_i g_{jk}, \\
&= \frac{2}{\gamma}\frac{\partial}{\partial\theta^i}\log\sqrt{g}.
\end{aligned}$$

Clearly, T satisfies the condition (Equation 19), and thus, the model is statistically affine. Therefore, Theorem 4 implies that there exists an α-parallel prior π_α for every α.

Now, let us obtain an explicit form by using π_J. Substituting the above T_i into the RHS of Equation (21),

$$\frac{\partial}{\partial\theta^i}\log h = -\frac{\alpha}{\gamma}\frac{\partial}{\partial\theta^i}\log\pi_J,$$

and thus, we obtain

$$h(\theta) = \frac{\pi_\alpha(\theta)}{\pi_J(\theta)} \propto \pi_J(\theta)^{-\frac{\alpha}{\gamma}}.$$

In particular, we get

$$\pi_\alpha(\theta) \propto \pi_J(\theta)^{1-\frac{\alpha}{\gamma}}.$$

It is true only in the γ-affine coordinate system $\{\theta^i\}$ that π_α is equal to a power of the Jeffreys prior. Since the above argument is not invariant under coordinate transformation, we take the Jacobian into consideration in another coordinate.

Theorem 6 includes previous results. Liu et al. [13], Example 1, corresponds to the case when $\alpha = 1/2$ and $\gamma = 1$. Takeuchi and Amari [11], Example 2 (p. 1017), corresponds to the case when $\gamma = 1$.

5. Conclusion

In the present study, we investigated the derivation by Lie et al. of χ^2-prior from the viewpoint of information geometry. We showed that χ^2-prior agrees with the $\frac{1}{2}$-parallel prior (α-parallel prior for $\alpha = \frac{1}{2}$), which gives a geometrical interpretation. In addition, in our formulation, using the log ratio $\log\pi/\pi_J$, which is invariant under reparametrization, simplifies a PDE defining a noninformative prior π in Bayesian analysis.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Bernardo JM. Reference analysis. In: Dey KK, Rao CR, editors. *Handbook of Statistics*, Vol. 25. Amsterdam: Elsevier (2005), p. 17–90. doi: 10.1016/S0169-7161(05)25002-2

2. Berger J. *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer (1985). doi: 10.1007/978-1-4757-4286-2

3. Jeffreys H. *Theory of Probability*. Oxford: Oxford University Press (1961).

4. Robert CP. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer (2001).

5. Bernardo JM. Reference posterior distributions for Bayesian inference. *J R Statist Soc B*. (1979) 41:113. doi: 10.1111/j.2517-6161.1979.tb01066.x

6. Berger JO, Bernardo JM, Sun D. The formal definition of reference priors. *Ann Statist*. (2009) 37:905–38. doi: 10.1214/07-AOS587

7. Komaki F. Shrinkage priors for Bayesian prediction. *Ann Statist.* (2006) 34:808–19. doi: 10.1214/009053606000000010

8. Komaki F. Bayesian predictive densities based on latent information priors. *J Stat Plan Inf.* (2011) 141:3705–15. doi: 10.1016/j.jspi.2011.06.009

9. Tanaka F, Komaki F. Asymptotic expansion of the risk difference of the Bayesian spectral density in the autoregressive moving average model. *Sankhya Series A.* (2011) 73(A):162–84. doi: 10.1007/s13171-011-0005-1

10. Tanaka F. Superharmonic priors for autoregressive models. *Inf Geom.* (2018) 1:215–35. doi: 10.1007/s41884-017-0001-1

11. Takeuchi J, Amari S. Alpha-parallel prior and its properties. *IEEE Trans Info Theory.* (2005) 51:1011–23. doi: 10.1109/TIT.2004.842703

12. Ghosh M. Objective priors: an introduction for frequentists. *Stat Sci.* (2011) 26:187–202. doi: 10.1214/10-STS338

13. Liu R, Chakrabarti A, Samanta T, Ghosh JK, Ghosh M. On divergence measures leading to Jeffreys and other reference priors. *Bayesian Anal.* (2014) 9:331–70. doi: 10.1214/14-BA862

14. Amari S, Nagaoka H. *Methods of Information Geometry*. Oxford: AMS (2000).

15. Amari S. *Information Geometry and Its Applications*. Tokyo: Springer-Verlag (2016). doi: 10.1007/978-4-431-55978-8

16. Kobayashi S, Nomizu K. *Foundations of Differential Geometry I*. New York, NY: Wiley (1969).

17. Ghosh M, Liu R. Moment matching priors. *Sankhya Series A.* (2011) 73(A):185–201. doi: 10.1007/s13171-011-0012-2

18. Clarke BS, Barron AR. Information-theoretic asymptotics of Bayes methods. *IEEE Trans Inform Theory.* (1990) 36:453–71. doi: 10.1109/18.54897

19. Clarke BS, Barron AR. Jeffreys' prior is asymptotically least favorable under entropy risk. *J Stat Plan Inference.* (1994) 41:37–60. doi: 10.1016/0378-3758(94)90153-8

20. Amari S. *Differential Geometrical Methods in Statistics*. Oxford: Springer-Verlag (1985). doi: 10.1007/978-1-4612-5056-2

21. Simon U, Schwenk-Schellschmidt A, Viesel H. *Introduction to the Affine Differential Geometry of Hypersurfaces*. Tokyo: Lecture Notes of the Science University of Tokyo (1991).

22. Wald RM. *General Relativity*. Chicago, IL: The University of Chicago Press (1984). doi: 10.7208/chicago/9780226870373.001.0001

23. Hartigan JA. The maximum likelihood prior. *Ann Statist.* (1998) 26:2083–103. doi: 10.1214/aos/1024691462

24. Tanaka F. Curvature form on statistical model manifolds and its application to Bayesian analysis. *J Stat Appl Probab.* (2012) 1:35–43. doi: 10.12785/jsap/010105

25. Matsuzoe H, Takeuchi J, Amari S. Equiaffine structures on statistical manifolds and Bayesian statistics. *Differ Geom Appl.* (2006) 24:567–78. doi: 10.1016/j.difgeo.2006.02.003