Bayesian Analysis (2015) **10**, Number 1, pp. 31–51

Asymptotic Properties of Bayesian Predictive Densities When the Distributions of Data and Target Variables are Different

Fumiyasu Komaki[*†]

Abstract. Bayesian predictive densities when the observed data x and the target variable y to be predicted have different distributions are investigated by using the framework of information geometry. The performance of predictive densities is evaluated by the Kullback–Leibler divergence. The parametric models are formulated as Riemannian manifolds. In the conventional setting in which x and y have the same distribution, the Fisher–Rao metric and the Jeffreys prior play essential roles. In the present setting in which x and y have different distributions, a new metric, which we call the predictive metric, constructed by using the Fisher information matrices of x and y, and the volume element based on the predictive metric play the corresponding roles. It is shown that Bayesian predictive densities based on priors constructed by using non-constant positive superharmonic functions with respect to the predictive metric asymptotically dominate those based on the volume element prior of the predictive metric.

Keywords: differential geometry, Fisher–Rao metric, Jeffreys prior, Kullback–Leibler divergence, predictive metric.

1 Introduction

Suppose that we have independent observations $x(1), x(2), \ldots, x(N)$ from a probability density $p(x \mid \theta)$ that belongs to a parametric model $\{p(x \mid \theta) \mid \theta \in \Theta\}$, where $\theta = (\theta^1, \theta^2, \ldots, \theta^d)$ is an unknown d-dimensional parameter and Θ is the parameter space. The random variable y to be predicted is independently distributed according to a density $\tilde{p}(y \mid \theta)$ in a parametric model $\{\tilde{p}(y \mid \theta) \mid \theta \in \Theta\}$, possibly different from $\{p(x \mid \theta) \mid \theta \in \Theta\}$, with the same parameter θ. The objective is to construct a predictive density $\hat{p}(y; x^N)$ for y by using $x^N := (x(1), \ldots, x(N))$. The performance of $\hat{p}(y; x)$ is evaluated by the Kullback–Leibler divergence

$$D(\tilde{p}(y \mid \theta), \hat{p}(y; x^N)) := \int \tilde{p}(y \mid \theta) \log \frac{\tilde{p}(y \mid \theta)}{\hat{p}(y; x^N)} \mathrm{d}y$$

*Department of Mathematical Informatics, Graduate School of Information Science and Technology, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN, komaki@mist.i.u-tokyo.ac.jp

†RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, JAPAN

 DOI: 10.1214/14-BA886

from the true density $\tilde{p}(y \mid \theta)$ to the predictive density $\hat{p}(y; x^N)$. The risk function is given by

$$\mathrm{E}\Big[D(\tilde{p}(y \mid \theta), \hat{p}(y; x^N)) \Big| \theta\Big] = \iint p(x^N \mid \theta)\tilde{p}(y \mid \theta) \log \frac{\tilde{p}(y \mid \theta)}{\hat{p}(y; x^N)} \mathrm{d}y \mathrm{d}x^N.$$

It is widely recognized that plug-in densities $\tilde{p}(y \mid \hat{\theta})$ constructed by replacing the unknown parameter θ by an estimate $\hat{\theta}(x^N)$ may not perform very well and that Bayesian predictive densities

$$\tilde{p}_\pi(y \mid x^N) := \frac{\int \tilde{p}(y \mid \theta) p(x^N \mid \theta)\pi(\theta)\mathrm{d}\theta}{\int p(x^N \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

constructed by using a prior π perform better than plug-in densities. If the value of θ is given, there is no specific meaning of considering the conditional density of y given x^N since the obvious relation $p(y \mid x, \theta) = p(y \mid \theta)$ holds. However, if θ is unknown, Bayesian predictive densities $p_\pi(y \mid x^N)$ constructed by introducing a prior density $\pi(\theta)$ on the parameter space are useful to approximate the true density $p(y \mid \theta)$ as discussed in Aitchison and Dunsmore (1975) and Geisser (1993). In fact, there exists a predictive density whose asymptotic risk is smaller than that of a plug-in density unless the mean mixture curvature of the model manifold vanishes, see Komaki (1996) and Hartigan (1998) for details. The choice of π becomes important especially when the sample size N is not very large. Although the Jeffreys prior is a widely known default prior, it does not perform satisfactorily especially when the unknown parameter is multidimensional as Jeffreys himself pointed out.

Komaki (2001) constructed a Bayesian predictive density incorporating the advantage of shrinkage methods for the multivariate normal model. See also George et al. (2006) for useful results for the normal model.

In the conventional setting in which the distributions of $x(i)$, $i = 1, \dots, N$, and y are the same, asymptotic theory of prediction based on general parametric models has been studied by using the framework of information geometry, see Komaki (1996). In information geometry, a parametric statistical model is regarded as a differentiable manifold, which we call the model manifold, and the parameter space is regarded as a coordinate system of the manifold, see Amari (1985). The Fisher–Rao metric is a Riemannian metric based on the Fisher information matrix on the model manifold. The Jeffreys prior $\pi_\mathrm{J}(\theta)$ corresponds to the volume element of the model manifold associated with the Fisher–Rao metric. When the distributions of $x(i)$, $i = 1, \dots, N$, and y are the same, the asymptotic difference between the risks of $\tilde{p}_\pi(y \mid x^N)$ and $\tilde{p}_\mathrm{J}(y \mid x^N)$ is given by

$$\begin{aligned} N^2 \Big[& \mathrm{E}\big\{D(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N) \,\big|\, \theta\big\} - \mathrm{E}\big\{D(\tilde{p}(y \mid \theta), \tilde{p}_\mathrm{J}(y \mid x^N) \,\big|\, \theta\big\}\Big] \\ &= \frac{\Delta\left(\frac{\pi}{\pi_\mathrm{J}}\right)}{\left(\frac{\pi}{\pi_\mathrm{J}}\right)} - \frac{1}{2}\sum_{i=1}^d \sum_{j=1}^d g^{ij} \frac{\partial_i \left(\frac{\pi}{\pi_\mathrm{J}}\right) \partial_j \left(\frac{\pi}{\pi_\mathrm{J}}\right)}{\left(\frac{\pi}{\pi_\mathrm{J}}\right)^2} + \mathrm{o}(1) = 2\frac{\Delta\left(\frac{\pi}{\pi_\mathrm{J}}\right)^{\frac{1}{2}}}{\left(\frac{\pi}{\pi_\mathrm{J}}\right)^{\frac{1}{2}}} + \mathrm{o}(1), \end{aligned} \quad (1)$$

where ∂_i denotes $\partial/\partial\theta^i$, $g_{ij} := \mathrm{E}\{\partial_i \log p(x \mid \theta)\partial_j \log p(x \mid \theta) \mid \theta\}$, g^{ij} denotes the (i,j)-element of the inverse of the $d \times d$ matrix (g_{ij}), and Δ is the Laplacian, see Komaki (2006). The Laplacian Δ on a Riemannian manifold endowed with a metric g_{ij} is defined by

$$\Delta f = |g|^{-1/2} \sum_i \sum_j \partial_i(|g|^{1/2} g^{ij} \partial_j f) = \sum_i \sum_j \nabla_i^{(0)}(g^{ij}\partial_j f), \tag{2}$$

where $|g|$ is the determinant of the $d \times d$ matrix (g_{ij}), f is a smooth real function on Θ, and $\nabla_i^{(0)}$ denotes the covariant derivative, defined in the next section. The indices $i, j, k \ldots$ run from 1 to d. Note that both the definition (2) of the Laplacian and the definition $\Delta f = -|g|^{-1/2} \sum_i \sum_j \partial_i(|g|^{1/2} g^{ij} \partial_j f)$ that differs in sign are widely adopted in the mathematics literature, although it is confusing. Because of (1), if there exists a non-constant positive superharmonic function f, i.e. a non-constant positive function satisfying $\Delta f \leq 0$ for every θ, on the model manifold, then the Bayesian predictive density based on the prior density defined by $\pi = f\pi_{\mathrm{J}}$ asymptotically dominates that based on the Jeffreys prior. Here, the Riemannian geometric structure of the model manifold based on the Fisher–Rao metric plays a fundamental role.

In practical applications, it often occurs that observed data $x(i)$, $i = 1, \ldots, N$, and the target variable y to be predicted have different distributions. Regression models are a typical example. Suppose that we observe $x = W\theta + \varepsilon$, where W is a given $n \times d$ matrix ($n \geq d$), and predict $y = Z\theta + \varepsilon$, where Z is a given $m \times d$ matrix and $\theta = (\theta^1, \cdots, \theta^d)$ is an unknown parameter. Then, the Fisher information matrices for the same parameter θ based on $p(x \mid \theta)$ and $\tilde{p}(y \mid \theta)$ are different. Similar situations also occur in nonlinear regression problems. Kobayashi and Komaki (2008) and George and Xu (2008) showed that shrinkage priors are useful for constructing Bayesian predictive densities for linear regression models when the observations are normally distributed with known variance. However, it has been difficult to construct useful priors for general models other than the normal models when x and y have different distributions.

In the present paper, we study asymptotic theory for the setting in which $x(i)$, $i = 1, \ldots, N$, and y have different distributions. Although several asymptotic properties of predictive distributions for such a setting are studied by Fushiki et al. (2004), the result corresponding to (1) has not been explored. The generalization is not straightforward because two different differential geometric structures, one for $p(x \mid \theta)$ and the other for $\tilde{p}(y \mid \theta)$, such as the Fisher–Rao metrics exist in the present setting.

We introduce a new metric $\mathring{g}_{ij}$, which we call the predictive metric, depending on both $p(x \mid \theta)$ and $\tilde{p}(y \mid \theta)$. The predictive metric $\mathring{g}_{ij}$ and the volume element $|\mathring{g}|^{1/2}\mathrm{d}\theta^1 \cdots \mathrm{d}\theta^d$ of it correspond to the Fisher–Rao metric and the Jeffreys prior in the conventional setting.

In Section 2, we obtain an expansion of the difference of the risk functions of Bayesian predictive densities. Each term in the expansion is represented by using geometrical quantities and is invariant with respect to parameter transformations. In Section 3, we introduce the predictive metric $\mathring{g}_{ij}$ and evaluate the asymptotic risk difference between a Bayesian predictive density based on a prior π and that based on the volume element prior $|\mathring{g}|^{1/2}\mathrm{d}\theta^1\cdots\mathrm{d}\theta^d$ of the predictive metric $\mathring{g}_{ij}$. The asymptotic risk difference is represented by using the Laplacian associated with the predictive metric $\mathring{g}_{ij}$. In Section 4, we consider three examples and construct superior priors by using the formula obtained in Section 3.

2 An expansion of the risk of predictive densities

First, we prepare several information geometrical notations to be used. In the following, the quantities associated with the model $\{p(x \mid \theta) \,|\, \theta \in \Theta\}$ are denoted without tilde, and those associated with the model $\{\tilde{p}(y \mid \theta) \,|\, \theta \in \Theta)\}$ are denoted with tilde. We put $l := \log p(x \mid \theta)$ and $\tilde{l} := \log \tilde{p}(y \mid \theta)$. The Fisher–Rao metrics on the model manifolds $\{p(x \mid \theta) \,|\, \theta \in \Theta\}$ and $\{\tilde{p}(y \mid \theta) \,|\, \theta \in \Theta\}$ are given by

$$g_{ij}(\theta) := \mathrm{E}\left(\partial_i l \partial_j l \,\middle|\, \theta\right) \quad \text{and} \quad \tilde{g}_{ij}(\theta) := \mathrm{E}\big(\partial_i \tilde{l} \partial_j \tilde{l} \,\big|\, \theta\big),$$

respectively. The (i, j)-elements of the inverses of the $d \times d$ matrices (g_{ij}) and $(\tilde{g}_{ij})$ are denoted by g^{ij} and $\tilde{g}^{ij}$, respectively. We define

$$\begin{aligned}
T_{ijk}(\theta) &:= \mathrm{E}\Big(\partial_i l \partial_j l \partial_k l \,\Big|\, \theta\Big), \quad \widetilde{T}_{ijk}(\theta) := \mathrm{E}\Big(\partial_i \tilde{l} \partial_j \tilde{l} \partial_k \tilde{l} \,\Big|\, \theta\Big), \\
\Gamma^{(\mathrm{e})}_{ijk}(\theta) &:= \mathrm{E}\Big(\partial_i \partial_j l \partial_k l \,\Big|\, \theta\Big), \quad \tilde{\Gamma}^{(\mathrm{e})}_{ijk}(\theta) := \mathrm{E}\Big(\partial_i \partial_j \tilde{l} \partial_k \tilde{l} \,\Big|\, \theta\Big), \\
\Gamma^{(\mathrm{m})}_{ijk}(\theta) &:= \mathrm{E}\Big(\partial_i \partial_j l \partial_k l \,\Big|\, \theta\Big) + T_{ijk}(\theta), \quad \tilde{\Gamma}^{(\mathrm{m})}_{ijk}(\theta) := \mathrm{E}\Big(\partial_i \partial_j \tilde{l} \partial_k \tilde{l} \,\Big|\, \theta\Big) + \widetilde{T}_{ijk}(\theta), \\
\Gamma^{(0)}_{ijk}(\theta) &:= \frac{1}{2}\Big\{\Gamma^{(\mathrm{e})}_{ijk}(\theta) + \Gamma^{(\mathrm{m})}_{ijk}(\theta)\Big\} = \frac{1}{2}\left\{\partial_i g_{jk}(\theta) + \partial_j g_{ki}(\theta) - \partial_k g_{ij}(\theta)\right\}, \\
\tilde{\Gamma}^{(0)}_{ijk}(\theta) &:= \frac{1}{2}\Big\{\tilde{\Gamma}^{(\mathrm{e})}_{ijk}(\theta) + \tilde{\Gamma}^{(\mathrm{m})}_{ijk}(\theta)\Big\} = \frac{1}{2}\left\{\partial_i \tilde{g}_{jk}(\theta) + \partial_j \tilde{g}_{ki}(\theta) - \partial_k \tilde{g}_{ij}(\theta)\right\},
\end{aligned}$$

and

$$\tilde{Q}_{ijkl}(\theta) := \mathrm{E}\Big(\partial_i \tilde{l} \partial_j \tilde{l} \partial_k \tilde{l} \partial_l \tilde{l} \,\Big|\, \theta\Big).$$

Here, $\Gamma^{(\mathrm{e})}_{ijk}$ are the e-connection coefficients, $\Gamma^{(\mathrm{m})}_{ijk}$ are the m-connection coefficients, and $\Gamma^{(0)}_{ijk}$ are the Riemannian connection coefficients. The relations

$$\partial_i g_{jk} = \Gamma^{(\mathrm{e})}_{ijk} + \Gamma^{(\mathrm{m})}_{ikj}, \quad \text{and} \quad \partial_i \tilde{g}_{jk} = \tilde{\Gamma}^{(\mathrm{e})}_{ijk} + \tilde{\Gamma}^{(\mathrm{m})}_{ikj} \tag{3}$$

represent the duality between $\Gamma^{(\mathrm{e})}_{ijk}$ and $\Gamma^{(\mathrm{m})}_{ijk}$ with respect to the metric g_{ij}, and the duality between $\tilde{\Gamma}^{(\mathrm{e})}_{ijk}$ and $\tilde{\Gamma}^{(\mathrm{m})}_{ijk}$ with respect to the metric $\tilde{g}_{ij}$, respectively.

Covariant derivatives $\nabla^{(\mathrm{e})}_i u^j$, $\nabla^{(0)}_i u^j$, and $\nabla^{(\mathrm{m})}_i u^j$ of a vector field u^j with respect to the connection coefficients $\Gamma^{(\mathrm{e})k}_{ij}$, $\Gamma^{(0)k}_{ij}$, and $\Gamma^{(\mathrm{m})k}_{ij}$ are defined by $\nabla^{(\mathrm{e})}_i u^j := \partial_i u^j + \sum_k \Gamma^{(\mathrm{e})j}_{ik} u^k$, $\nabla^{(0)}_i u^j := \partial_i u^j + \sum_k \Gamma^{(0)j}_{ik} u^k$, and $\nabla^{(\mathrm{e})}_i u^j := \partial_i u^j + \sum_k \Gamma^{(\mathrm{e})j}_{ik} u^k$, respectively, where $\Gamma^{(\mathrm{e})j}_{ik} = \sum_l \Gamma^{(\mathrm{e})}_{ikl} g^{jl}$, $\Gamma^{(0)j}_{ik} = \sum_l \Gamma^{(0)}_{ikl} g^{jl}$, and $\Gamma^{(\mathrm{m})j}_{ik} = \sum_l \Gamma^{(\mathrm{m})}_{ikl} g^{jl}$. In the same way, the covariant derivatives $\widetilde{\nabla}^{(\mathrm{e})}_i u^j$, $\widetilde{\nabla}^{(0)}_i u^j$, and $\widetilde{\nabla}^{(\mathrm{m})}_i u^j$, with respect to the connection coefficients $\tilde{\Gamma}^{(\mathrm{e})j}_{ik} = \sum_l \tilde{\Gamma}^{(\mathrm{e})}_{ikl}\tilde{g}^{jl}$, $\tilde{\Gamma}^{(0)j}_{ik} = \sum_l \tilde{\Gamma}^{(0)}_{ikl}\tilde{g}^{jl}$, and $\tilde{\Gamma}^{(\mathrm{m})j}_{ik} = \sum_l \tilde{\Gamma}^{(\mathrm{m})}_{ikl}\tilde{g}^{jl}$, are defined.

Theorem 1 below is used in the following sections.

Theorem 1. The difference between the risk functions of Bayesian predictive densities $\tilde{p}_\pi(y \mid x^N)$ and $\tilde{p}_{\pi'}(y \mid x^N)$ based on priors $\pi(\theta)\mathrm{d}\theta$ and $\pi'(\theta)\mathrm{d}\theta$, respectively, is given by

$$
\begin{aligned}
N^2 & \Bigl[\mathrm{E}\bigl\{D(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N)\bigr) \bigm| \theta\bigr\} - \mathrm{E}\bigl\{D(\tilde{p}(y \mid \theta), \tilde{p}_{\pi'}(y \mid x^N)\bigr) \bigm| \theta\bigr\}\Bigr] \\
&= \Biggl(\frac{1}{2}\sum_{i,j}\tilde{g}_{ij}u^i_\pi u^j_\pi + \sum_{i,j,k}\tilde{g}_{ij}g^{jk}\widetilde{\nabla}^{(\mathrm{e})}_k u^i_\pi\Biggr) \\
&\quad - \Biggl(\frac{1}{2}\sum_{i,j}\tilde{g}_{ij}u^i_{\pi'} u^j_{\pi'} + \sum_{i,j,k}\tilde{g}_{ij}g^{jk}\widetilde{\nabla}^{(\mathrm{e})}_k u^i_{\pi'}\Biggr) + \mathrm{o}(1), \qquad (4)
\end{aligned}
$$

where

$$
u^i_\pi(\theta) := \sum_k g^{ik}(\theta)\Bigl\{\partial_k \log\pi(\theta) - \sum_j \Gamma^{(\mathrm{e})j}_{kj}(\theta)\Bigr\} + \sum_{k,l} g^{kl}(\theta)\Bigl\{\tilde{\Gamma}^{(\mathrm{m})i}_{kl}(\theta) - \Gamma^{(\mathrm{m})i}_{kl}(\theta)\Bigr\}.
$$

The proof of Theorem 1 is given in the Appendix.

3 Prior construction based on the predictive metric

In this section, we introduce a new metric defined by

$$
\mathring{g}_{ij} := \sum_{k=1}^d \sum_{l=1}^d g_{ik}\tilde{g}^{kl}g_{jl}, \qquad (5)
$$

which we call the predictive metric. Since $\mathring{g}_{ij}$ is positive definite, it can be adopted as a Riemannian metric on Θ. It will be shown that the predictive metric $\mathring{g}_{ij}$, the corresponding volume element

$$
\pi_{\mathrm{P}}(\theta)\mathrm{d}\theta := |\mathring{g}_{ij}|^{\frac{1}{2}}\mathrm{d}\theta = \Bigl|\sum_{k=1}^d\sum_{l=1}^d g_{ik}\tilde{g}^{kl}g_{lj}\Bigr|^{\frac{1}{2}}\mathrm{d}\theta = |g_{ij}||\tilde{g}_{ij}|^{-\frac{1}{2}}\mathrm{d}\theta, \qquad (6)
$$

and the Laplacian $\mathring{\Delta}$ based on $\mathring{g}_{ij}$ play essential roles corresponding to those played by the Fisher–Rao metric g_{ij}, the Jeffreys prior $|g_{ij}|^{1/2}\mathrm{d}\theta$, and the Laplacian Δ based on g_{ij} in the conventional setting where $g_{ij} = \tilde{g}_{ij}$. Here, $|\mathring{g}_{ij}|$, $|g_{ij}|$, and $|\tilde{g}_{ij}|$ denote determinants of $d \times d$ matrices $(\mathring{g}_{ij})$, (g_{ij}), and $(\tilde{g}_{ij})$, respectively. The (i,j)-element of the inverse of the $d \times d$ matrix $(\mathring{g}_{ij})$ is given by $\mathring{g}^{ij} := \sum_{k,l}\tilde{g}_{kl}g^{ik}g^{jl}$.

Here, we give an intuitive meaning of the predictive metric $\mathring{g}_{ij}$ by a nonrigorous argument. In the standard estimation theory, the Fisher-Rao metric g_{ij}, which is the Fisher information matrix, corresponds to the inverse of the asymptotic variance of the maximum likelihood estimator. In the setting we consider, the asymptotic variance of the maximum likelihood estimator based on x^N is $(Ng)^{-1}$, where g is the $d \times d$ matrix (g_{ij}), and the asymptotic variance of the maximum likelihood estimator based on both of x^N and y is $(Ng+\tilde{g})^{-1}$, where $\tilde{g}$ is the $d \times d$ matrix $(\tilde{g}_{ij})$. The inverse of the reduction of the asymptotic variance by observing y in addition to $x(i)$ $(i = 1, \ldots, N)$ are given by $\{(Ng)^{-1} - (Ng + \tilde{g})^{-1}\}^{-1} = N^2 \mathring{g} + \mathrm{O}(N)$, as we see in Example 1 in Section 4, corresponding to the predictive metric $\mathring{g}$.

The Riemannian connection coefficients with respect to the predictive metric $\mathring{g}_{ij}$ are given by

$$\mathring{\Gamma}^{(0)}_{ijk} = \frac{1}{2}\left(\partial_i \mathring{g}_{jk} + \partial_j \mathring{g}_{ki} - \partial_k \mathring{g}_{ij}\right),$$

and we put $\mathring{\Gamma}^{(0)j}_{ik} = \sum_l \mathring{\Gamma}^{(0)}_{ikl} \mathring{g}^{jl}$. Then,

$$\partial_k \log |\mathring{g}_{ij}|^{\frac{1}{2}} = \frac{1}{2}\partial_k \log |\mathring{g}_{ij}| = \sum_{i,j} \frac{1}{2}(\partial_k \mathring{g}_{ij})\mathring{g}^{ij} = \sum_i \mathring{\Gamma}^{(0)i}_{ki}. \tag{7}$$

In the same way, we have

$$\partial_k \log |g_{ij}|^{\frac{1}{2}} = \sum_i \Gamma^{(0)i}_{ki}, \quad \text{and} \quad \partial_k \log |\tilde{g}_{ij}|^{\frac{1}{2}} = \sum_i \tilde{\Gamma}^{(0)i}_{ki}. \tag{8}$$

Thus,

$$\sum_i \mathring{\Gamma}^{(0)i}_{ki} = \partial_k \log |\mathring{g}_{ij}|^{\frac{1}{2}} = \partial_k \log |g_{ij}| - \frac{1}{2}\partial_k \log |\tilde{g}_{ij}| = 2\sum_i \Gamma^{(0)i}_{ki} - \sum_i \tilde{\Gamma}^{(0)i}_{ki}. \tag{9}$$

The Laplacian $\mathring{\Delta}$ with respect to the predictive metric $\mathring{g}_{ij}$ is defined by

$$\begin{aligned}\mathring{\Delta} f = \sum_{i,j} \mathring{\nabla}^{(0)}_i (\mathring{g}^{ij}\partial_j f) &= \sum_{i,j} \partial_i (\mathring{g}^{ij}\partial_j f) + \sum_{i,j,k} \mathring{\Gamma}^{(0)i}_{ik} \mathring{g}^{kj} \partial_j f \\ &= \sum_{i,j} \mathring{g}^{ij}(\partial_i \partial_j f - \sum_k \mathring{\Gamma}^{(0)k}_{ij} \partial_k f),\end{aligned}$$

where $\mathring{\nabla}^{(0)}_i u^j = \partial_i u^j + \sum_k \mathring{\Gamma}^{(0)j}_{ik} u^k$, and f is a real smooth function on Θ.

By using these quantities, we obtain the following theorem corresponding to (1) in the conventional setting.

Theorem 2. The difference between the risk functions of Bayesian predictive densities $\tilde{p}_\pi(y \mid x^N)$ based on a $\pi(\theta)\mathrm{d}\theta$ and $\tilde{p}_{\mathrm{P}}(y \mid x^N)$ based on $\pi_{\mathrm{P}}(\theta)\mathrm{d}\theta$ is given by

$$N^2\Big[\mathrm{E}\big\{D(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N) \,\big|\, \theta\big\} - \mathrm{E}\big\{D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{P}}(y \mid x^N) \,\big|\, \theta\big\}\Big]$$

$$= \frac{\mathring{\Delta}\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)}{\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)} - \frac{1}{2}\sum_{i,j}\mathring{g}^{ij}\frac{\partial_i\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)\partial_j\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)}{\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)^2} + \mathrm{o}(1) = 2\,\frac{\mathring{\Delta}\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)^{\frac{1}{2}}}{\left(\dfrac{\pi}{\pi_{\mathrm{P}}}\right)^{\frac{1}{2}}} + \mathrm{o}(1). \quad (10)$$

The proof of Theorem 2 is given in the Appendix.

If there exists a positive constant c such that $\pi' = c\pi$, we identify the prior π' with π because the posterior densities based on them are identical. In fact, the risk difference (10) between π and π_{P} coincides with that between π' and π_{P}.

Corollary 1. If a positive function $f(\theta)$ is superharmonic with respect to the predictive metric $\mathring{g}$, i.e. $\mathring{\Delta} f(\theta) \leq 0$ for every $\theta \in \Theta$, and the strict inequality holds at a point in Θ, then the Bayesian predictive density based on the prior density $\{f(\theta)\}^2\pi_{\mathrm{P}}(\theta)$ asymptotically dominates the Bayesian predictive density $\tilde{p}_{\mathrm{P}}(y \mid x^N)$ based on the prior density $\pi_{\mathrm{P}}(\theta)$. If there exists a non-constant positive superharmonic function $f(\theta)$ with respect to the predictive metric $\mathring{g}$, then the Bayesian predictive density based on the prior density $\{f(\theta)\}^{2c}\pi_{\mathrm{P}}(\theta)$ $(0 < c < 1)$ asymptotically dominates $\tilde{p}_{\mathrm{P}}(y \mid x^N)$.

Proof. The first statement is a straightforward conclusion from Theorem 2. We show the second statement. The function $\{f(\theta)\}^c\,(0 < c < 1)$ is superharmonic because $\Delta f^c = cf^{-(1-c)}\{\Delta f - (1-c)f^{-1}g^{ij}\partial_i f \partial_j f\} \leq 0$ if $f(\theta)$ is a positive superharmonic function. The strict inequality holds at θ satisfying $\partial_i f(\theta) \neq 0$ for any i. Such θ exists since $f(\theta)$ is a non-constant function. Thus, the second statement follows from the first statement. □

By setting $c = 1/2$, it follows from Corollary 1 that the Bayesian predictive density based on the prior $f(\theta)\pi_{\mathrm{P}}(\theta)$ asymptotically dominates the Bayesian predictive density based on π_{P} if $f(\theta)$ is a non-constant positive superharmonic function.

Note that Corollary 1 also holds if we replace the predictive metric $\mathring{g}$ with another metric $\mathring{g}'$ satisfying $\mathring{g}' = c\mathring{g}$ with a positive constant c. This is because the volume element with respect to $\mathring{g}'$ is proportional to that with respect to $\mathring{g}$ and the relation $\mathring{\Delta}' f = (1/c)\mathring{\Delta} f$ holds, where $\mathring{\Delta}'$ is the Laplacian with respect to $\mathring{g}'$.

4 Examples

In this section, we see three examples. We verify that the results in the previous sections are consistent with several known results in Examples 1 and 2 and obtain some new results in Examples 2 and 3.

Example 1. Normal models

Suppose that x is distributed according to the d-dimensional normal distribution $\mathrm{N}_d(\mu, \Sigma)$ with mean vector μ and covariance matrix $\Sigma = (\Sigma^{ij})$ and that y is distributed according to the d-dimensional normal distribution $\mathrm{N}_d(\mu, \tilde{\Sigma})$ with the same mean vector μ and possibly different covariance matrix $\tilde{\Sigma} = (\tilde{\Sigma}^{ij})$. Here, μ is the unknown parameter and Σ and $\tilde{\Sigma}$ are known.

The Fisher information matrix for $p(x \mid \mu)$ is $(g_{ij}) = (\Sigma_{ij})$ and that for $p(x \mid \mu)$ is $(\tilde{g}_{ij}) = (\tilde{\Sigma}_{ij})$, where (Σ_{ij}) and $(\tilde{\Sigma}_{ij})$ are inverse matrices of (Σ_{ij}) and $(\tilde{\Sigma}_{ij})$, respectively.

Since the coefficients of the predictive metric $\mathring{g}_{ij} = \sum_{k,l} g_{ik}\tilde{g}^{kl}g_{jl}$ do not depend on μ, the volume element with respect to the predictive metric is

$$\pi_{\mathrm{P}}(\mu)\mathrm{d}\mu = \left|\mathring{g}_{ij}\right| \mathrm{d}\mu \propto \mathrm{d}\mu,$$

which is the uniform distribution $\pi_{\mathrm{U}}(\mu) \propto 1$.

Kobayashi and Komaki (2008) and George and Xu (2008) considered shrinkage priors for this model. The Bayesian predictive density $\tilde{p}_\pi(y \mid x^N)$ dominates $\tilde{p}_{\mathrm{U}}(y \mid x^N)$ based on the uniform measure $\pi_{\mathrm{U}}(\mu)$ if $\pi(\mu)$ is a superharmonic function on the Euclidean space $\mathbb{R}^d$ endowed with the metric $\left((N\,g)^{-1} - (Ng + \tilde{g})^{-1}\right)^{-1}$, see Theorem 3.2 in Kobayashi and Komaki (2008). This result holds for every positive integer N.

Since

$$\begin{aligned}\left((Ng)^{-1} - (Ng + \tilde{g})^{-1}\right)^{-1} =&(Ng)^{\frac{1}{2}} \left[I - \left\{I + (Ng)^{-\frac{1}{2}}\tilde{g}(Ng)^{-\frac{1}{2}}\right\}^{-1}\right]^{-1} (Ng)^{\frac{1}{2}} \\ =&(Ng)^{\frac{1}{2}} \left[I - I + (Ng)^{-\frac{1}{2}}\tilde{g}(Ng)^{-\frac{1}{2}} + \mathrm{O}(N^{-2})\right]^{-1} (Ng)^{\frac{1}{2}} \\ =&N^2\, g\, \tilde{g}^{-1}\, g + \mathrm{O}(N)\end{aligned}$$

corresponds to the predictive metric $\mathring{g}$, Theorem 2 is consistent with theoretical and numerical results in Kobayashi and Komaki (2008) and George and Xu (2008).

Example 2. Location-scale models

Suppose that $\phi(x)$ and $\tilde{\phi}(y)$ are probability densities on $\mathbb{R}$ that are symmetric about the origin. Let

$$p(x \mid \mu, \sigma)\mathrm{d}x := \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x \quad \text{and} \quad \tilde{p}(y \mid \mu, \sigma)\mathrm{d}y := \frac{1}{\sigma}\tilde{\phi}\left(\frac{y-\mu}{\sigma}\right)\mathrm{d}y,$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown parameters.

Suppose that we have a set of N independent observations $x(1), \ldots, x(N)$ distributed according to $p(x \mid \mu, \sigma)$. The variable y to be predicted is independently distributed according to $\tilde{p}(y \mid \mu, \sigma)$. The objective is to construct a prior π for a Bayesian predictive density $\tilde{p}_\pi(y \mid x)$.

The Fisher–Rao metrics on the model manifolds $\{p(x \mid \mu, \sigma)\}$ and $\{\tilde{p}(y \mid \mu, \sigma)\}$ are

$$g_{\mu\mu} = \frac{a}{\sigma^2}, \quad g_{\sigma\sigma} = \frac{b}{\sigma^2}, \quad g_{\mu\sigma} = 0,$$

$$\tilde{g}_{\mu\mu} = \frac{\tilde{a}}{\sigma^2}, \quad \tilde{g}_{\sigma\sigma} = \frac{\tilde{b}}{\sigma^2}, \text{ and } \tilde{g}_{\mu\sigma} = 0,$$

respectively, where a and b are positive constants depending on $\phi(x)$, and $\tilde{a}$ and $\tilde{b}$ are positive constants depending on $\tilde{\phi}(y)$.

The predictive metric is given by

$$\mathring{g}_{\mu\mu} = \frac{a^2/\tilde{a}}{\sigma^2}, \quad \mathring{g}_{\sigma\sigma} = \frac{b^2/\tilde{b}}{\sigma^2}, \quad \text{and} \quad \mathring{g}_{\mu\sigma} = 0.$$

Define

$$u := \sqrt{\frac{\tilde{b}}{\tilde{a}}} \frac{a}{b} \mu, \quad v := \sigma \tag{11}$$

by rescaling the location parameter μ. We call this coordinate system (u, v) the upper-half plane coordinates. Then, the predictive metric is represented by

$$\mathring{g}_{uu} = \frac{b^2/\tilde{b}}{v^2}, \quad \mathring{g}_{vv} = \frac{b^2/\tilde{b}}{v^2}, \text{ and } \mathring{g}_{uv} = 0,$$

coinciding with the metric on the Hyperbolic plane $H^2(-\tilde{b}/b^2)$, which is a 2-dimensional complete manifold with constant sectional curvature $-\tilde{b}/b^2$. Thus, the model manifold endowed with the predictive metric $\mathring{g}$ is isometric to $H^2(-\tilde{b}/b^2)$.

The volume element with respect to the predictive metric $\mathring{g}$ is given by

$$\pi_{\mathrm{P}}(\mu, \sigma)\mathrm{d}\mu\mathrm{d}\sigma = |\mathring{g}|^{1/2}\mathrm{d}\mu\mathrm{d}\sigma \propto \frac{1}{\sigma^2}\mathrm{d}\mu\mathrm{d}\sigma$$

and coincides with the Jeffreys priors $|g|^{1/2}\mathrm{d}\mu\mathrm{d}\sigma \propto 1/\sigma^2\mathrm{d}\mu\mathrm{d}\sigma$ for $p(x \mid \mu, \sigma)$ and $|\tilde{g}|^{1/2}\mathrm{d}\mu\mathrm{d}\sigma \propto 1/\sigma^2\mathrm{d}\mu\mathrm{d}\sigma$ for $\tilde{p}(y \mid \mu, \sigma)$.

The Laplacian on the model manifold endowed with the predictive metric $\mathring{g}$ is given by

$$\mathring{\Delta} = \sigma^2 \left(\frac{\tilde{a}}{a^2} \frac{\partial^2}{\partial\mu^2} + \frac{\tilde{b}}{b^2} \frac{\partial^2}{\partial\sigma^2} \right). \tag{12}$$

By Corollary 1, the Bayesian predictive density $\tilde{p}_{\mathrm{R}}(y \mid x)$ based on the prior

$$\pi_{\mathrm{R}}(\mu, \sigma)\mathrm{d}\mu\mathrm{d}\sigma \propto \frac{1}{\sigma}\mathrm{d}\mu\mathrm{d}\sigma$$

asymptotically dominates $\tilde{p}_{\mathrm{P}}(y \mid x)$ based on π_{P} because

$$\mathring{\Delta} \frac{\pi_{\mathrm{R}}(\mu, \sigma)}{\pi_{\mathrm{P}}(\mu, \sigma)} = \mathring{\Delta} \frac{1/\sigma}{1/\sigma^2} = \mathring{\Delta}\sigma = \sigma^2 \left(\frac{\tilde{a}}{a^2} \frac{\partial^2}{\partial\mu^2} + \frac{\tilde{b}}{b^2} \frac{\partial^2}{\partial\sigma^2} \right) \sigma = 0.$$

By Theorem 2, the asymptotic risk difference is

$$
\begin{aligned}
&N^2\left[\mathrm{E}\left\{D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{R}}(y \mid x^N) \,\middle|\, \theta\right\} - \mathrm{E}\left\{D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{P}}(y \mid x^N) \,\middle|\, \theta\right\}\right] \\
&\quad = 2\,\frac{\mathring{\Delta}\left(\dfrac{\pi_{\mathrm{R}}}{\pi_{\mathrm{P}}}\right)^{\frac{1}{2}}}{\left(\dfrac{\pi_{\mathrm{R}}}{\pi_{\mathrm{P}}}\right)^{\frac{1}{2}}} + \mathrm{o}(1) = 2\,\frac{\mathring{\Delta}\sigma^{\frac{1}{2}}}{\sigma^{\frac{1}{2}}} + \mathrm{o}(1) = -\frac{\tilde{b}}{2b^2} + \mathrm{o}(1).
\end{aligned} \tag{13}
$$

In fact, it can be shown that the Bayesian predictive density $\tilde{p}_{\mathrm{R}}(y \mid x)$ exactly dominates $\tilde{p}_{\mathrm{P}}(y \mid x)$ for finite N because π_{P} is the left invariant prior and π_{R} is the right invariant prior with respect to the location-scale group. The Bayesian procedures based on the right invariant prior dominate those based on the left invariant prior in many problems associated with group models as shown in Zidek (1969). The prior π_{R} is also derived as a reference prior, see Berger and Bernardo (1992).

Furthermore, as we see below, the Bayesian predictive density $\tilde{p}_{c,\kappa}(y \mid x)$ based on the prior $\pi_{c,\kappa}$ defined by

$$
\frac{\pi_{c,\kappa}}{\pi_{\mathrm{P}}}(\mu,\sigma) := \frac{2\kappa\sigma}{\dfrac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2 + c(\sigma+\kappa)^2 + (1-c)(\sigma^2+\kappa^2)} \qquad (0 \le c \le 1,\ 0 < \kappa < \infty) \tag{14}
$$

asymptotically dominates $\tilde{p}_{\mathrm{R}}(y \mid x)$ and thus also dominates $\tilde{p}_{\mathrm{P}}(y \mid x)$.

To clarify the meaning of the prior $\pi_{c,\kappa}$, we introduce another coordinate system on the model manifold. Let $(b/\sqrt{\tilde{b}})\rho$ be the Riemannian distance based on the predictive metric $\mathring{g}$ between a point P and an arbitrary fixed point O on $H^2(-\tilde{b}/b^2)$. The direction of P from O is represented by a point τ on the unit circle in the tangent space at O. Then, the point P is represented by ρ and τ, see e.g. Helgason (1984) p. 152. This coordinate system (ρ, τ) is called the geodesic polar coordinates. Then, the predictive metric is given by

$$
\mathring{g}_{\rho\rho} = \frac{b^2}{\tilde{b}}, \quad \mathring{g}_{\tau\tau} = \frac{b^2}{\tilde{b}}(\sinh\rho)^2, \text{ and } \ \mathring{g}_{\rho\tau} = 0.
$$

The Laplacian is represented by

$$
\mathring{\Delta} = \frac{\tilde{b}}{b^2}\left\{\frac{\partial^2}{\partial\rho^2} + \frac{\cosh\rho}{\sinh\rho}\frac{\partial}{\partial\rho} + (\sinh\rho)^{-2}\mathring{\Delta}_{\mathrm{S}}\right\}, \tag{15}
$$

where $\mathring{\Delta}_{\mathrm{S}}$ is the Laplacian on the unit circle in the tangent space at O, see e.g. Helgason (1984) p. 158.

When the upper-half plane coordinate system is adopted, the Riemannian distance $(b/\sqrt{\tilde{b}})\rho$ between (u, v) and $(\bar{u}, \bar{v})$ is represented by

$$
\cosh\rho = \frac{|u-\bar{u}|^2 + v^2 + \bar{v}^2}{2v\bar{v}},
$$

see e.g. Davies (1989) p. 176. Thus, in the original coordinate system (μ,σ), the Riemannian distance $(b/\sqrt{\tilde{b}})\rho$ between and (μ,σ) and $(0,\kappa)$ is

$$\cosh\rho = \frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2+\sigma^2+\kappa^2}{2\sigma\kappa}. \tag{16}$$

Thus, the ratio of prior densities is given by

$$\frac{\pi_{c,\kappa}(\mu,\sigma)}{\pi_{\rm P}(\mu,\sigma)} = \frac{1}{\frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2+\sigma^2+\kappa^2}{2\sigma\kappa}+c} \tag{17}$$

$$= \frac{1}{\cosh\rho + c}. \tag{18}$$

Note that $\pi_{c,\kappa}(\mu,\sigma)/\pi_{\rm P}(\mu,\sigma)$ depends on (μ,σ) only through $\rho(\mu,\sigma)$ defined by (16). Thus, from (15), (18), and Theorem 2, we have

$$\begin{aligned} N^2\Big[&\mathrm{E}\big\{D(\tilde{p}(y\mid\theta),\tilde{p}_{\pi_{c,\kappa}}(y|x)\,\big|\,\theta\big\} - \mathrm{E}\big\{D(\tilde{p}(y\mid\theta),\tilde{p}_{\rm P}(y\mid x)\,\big|\,\theta\big\}\Big] \\ =&2\,\frac{\mathring{\Delta}\left(\frac{\pi_{c,\kappa}}{\pi_{\rm P}}\right)^{\frac{1}{2}}}{\left(\frac{\pi_{c,\kappa}}{\pi_{\rm P}}\right)^{\frac{1}{2}}} + \mathrm{o}(1) = -\frac{\tilde{b}}{b^2}\left\{\frac{1}{2}+c\frac{\pi_{c,\kappa}}{\pi_{\rm P}}+\frac{3}{2}(1-c^2)\left(\frac{\pi_{c,\kappa}}{\pi_{\rm P}}\right)^2\right\}+\mathrm{o}(1) \\ =&-\frac{\tilde{b}}{b^2}\left\{\frac{1}{2}+c\frac{1}{\cosh\rho+c}+\frac{3}{2}(1-c^2)\frac{1}{(\cosh\rho+c)^2}\right\}+\mathrm{o}(1), \end{aligned} \tag{19}$$

and (19) is smaller than (13) when $0\le c<1$ and $0<\kappa<\infty$. The asymptotic risk difference (19) can also be derived from (17) and the Laplacian (12) in the original coordinate system.

By Corollary 1, the Bayesian predictive density $\tilde{p}_{c,\kappa}(y\mid x^N)$ $(0\le c<\infty, 0<\kappa<\infty)$ asymptotically dominates $\tilde{p}_{\rm P}(y\mid x^N)$ since the function (14) is superharmonic for $0\le c<\infty$. However, $\tilde{p}_{c,\kappa}(y\mid x^N)$ asymptotically dominates $\tilde{p}_{\rm R}(y\mid x^N)$ only when $0\le c\le 1$.

Several properties of the function (14) are discussed in Komaki (2007). As $\kappa\to\infty$, the prior $\pi_{c,\kappa}$ converges to the right invariant prior $\pi_{\rm R}$, because

$$\frac{\pi_{c,\kappa}(\mu,\sigma)}{\pi_{\rm P}(\mu,\sigma)} = \frac{1}{\frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2+\sigma^2+\kappa^2}{2\sigma\kappa}+c} \propto \frac{\kappa/2}{\frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2+\sigma^2+\kappa^2}{2\sigma\kappa}+c} \to \sigma$$

when $\kappa\to\infty$. Here, priors are identified up to a positive multiplicative constant. As $\kappa\to 0$, the prior $\pi_{c,\kappa}$ converges to

$$\pi_{\rm C}(\mu,\sigma)\mathrm{d}\mu\mathrm{d}\sigma := \frac{\sigma}{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2+\sigma^2}\frac{1}{\sigma^2}\mathrm{d}\mu\mathrm{d}\sigma = \frac{\sigma^{-1}}{\frac{a^2\tilde{b}}{b^2\tilde{a}}(\frac{\mu}{\sigma})^2+1}\frac{1}{\sigma^2}\mathrm{d}\mu\mathrm{d}\sigma,$$

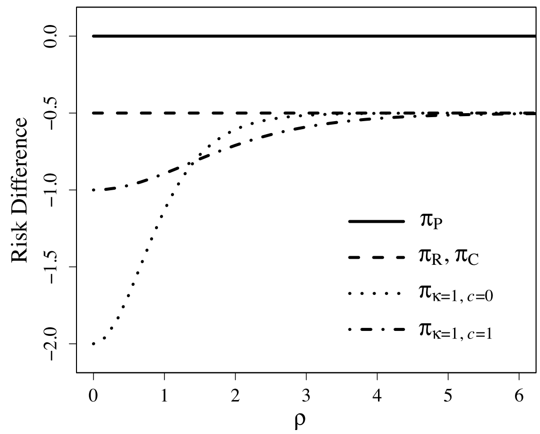

Figure 1: The asymptotic risk difference $N^2[\mathrm{E}\{D(\tilde{p}(y \mid \theta), \tilde{p}_{c,\kappa}(y \mid x^N) \mid \theta\} - \mathrm{E}\{D(\tilde{p}(y \mid \theta), \tilde{p}_{\rm P}(y \mid x^N) \mid \theta\}] + \mathrm{o}(1) = -(\tilde{b}/b^2)\{1/2 + c(\pi/\pi_{\rm P}) + (3/2)(1-c^2)(\pi/\pi_{\rm P})\}^2$ for Bayesian predictive densities based on $\pi_{\rm P}$, $\pi_{\rm R}$, $\pi_{\rm C}$, $\pi_{\kappa=1,c=0}$, and $\pi_{\kappa=1,c=1}$. We put $\tilde{b}/b^2 = 1$ just for simplicity.

because

$$\frac{\pi_{c,\kappa}(\mu,\sigma)}{\pi_{\rm P}(\mu,\sigma)} = \frac{1}{\frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2 + \sigma^2 + \kappa^2}{2\sigma\kappa} + c} \propto \frac{1/(2\kappa)}{\frac{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2 + \sigma^2 + \kappa^2}{2\sigma\kappa} + c} \to \frac{\sigma}{\frac{a^2\tilde{b}}{b^2\tilde{a}}\mu^2 + \sigma^2}$$

when $\kappa \to 0$. The prior density with respect to the rescaled parameter (u, v) defined by (11) is given by

$$\pi_{\rm C}(\mu,\sigma)\mathrm{d}\mu\mathrm{d}\sigma \propto \frac{v^{-1}}{(u/v)^2 + 1}\frac{1}{v^2}\mathrm{d}u\mathrm{d}v. \tag{20}$$

Note that the Cauchy prior for u, discussed by Jeffreys and many researchers, appears in (20). Thus, the class $\pi_{c,\kappa}$ of priors bridges the right invariant prior $\pi_{\rm R}$, coinciding with the reference prior, and the Cauchy prior $\pi_{\rm C}$.

Figure 1 illustrates the difference between the risk functions of Bayesian predictive densities based on $\pi_{\rm R}$, $\pi_{\rm C}$, $\pi_{\kappa=1,c=0}$, and $\pi_{\kappa=1,c=1}$ and the risk function of $\tilde{p}_{\rm P}(y \mid x^N)$. The risk functions of the right invariant prior $\pi_{\rm R}$ and the Cauchy prior $\pi_{\rm C}$ are uniformly smaller than that of $\pi_{\rm P}$. The asymptotic risk of the Cauchy prior $\pi_{\rm C}$ coincides with that of $\pi_{\rm R}$. Furthermore, the asymptotic risks of $\pi_{\kappa=1,c=0}$ and $\pi_{\kappa=1,c=1}$ are smaller than that of $\pi_{\rm R}$ for every (μ,σ). Therefore, the use of $\pi_{\kappa=1,c}$ $(0 \le c \le 1)$ is recommended. The risk of $\pi_{\kappa=1,c=0}$ is smaller than that of $\pi_{\kappa=1,c=1}$ when ρ is small, and vice versa. Thus, there does not exist a unique best value of c. The choice of the value of $0 < \kappa < \infty$ is arbitrary because it corresponds to the center of shrinkage. Finite-sample decision theoretic properties such as admissibility of Bayesian predictive densities $\tilde{p}_{\kappa,c}(y \mid x^N)$ based on proposed priors $\pi_{\kappa,c}$ $(0 < \kappa < \infty,\ 0 \le c \le 1)$ require further research.

Example 3. Poisson models

Suppose that x_i $(i = 1, \dots, d)$ are independently distributed according to the Poisson distribution $\mathrm{Po}(\lambda_i)$ with mean λ_i and that y_i $(i = 1, \dots, d)$ are independently distributed according to the Poisson distribution $\mathrm{Po}(s_i\lambda_i)$ with mean $s_i\lambda_i$. Here, s_i are known positive constants. The unknown parameter is $\theta = (\theta^1, \dots, \theta^d) := (\lambda_1, \dots, \lambda_d)$. The objective is to construct a predictive density for y by using x. This problem in the conventional setting, in which $s_1 = s_2 = \cdots = s_d$, is studied in Komaki (2004).

If $s_i \ll 1$ for each i, then this prediction problem is in the asymptotic setting. The Fisher–Rao metrics corresponding to x and y are given by

$$g_{ij} = \begin{cases} \dfrac{1}{\lambda_i} & (i = j) \\ 0 & (i \neq j) \end{cases} \quad \text{and} \quad \tilde{g}_{ij} = \begin{cases} \dfrac{s_i}{\lambda_i} & (i = j) \\ 0 & (i \neq j) \end{cases},$$

respectively. The predictive metric is

$$\mathring{g}_{ij} = \begin{cases} \dfrac{1}{s_i\lambda_i} & (i = j) \\ 0 & (i \neq j) \end{cases}, \tag{21}$$

and the corresponding volume element is

$$\pi_{\mathrm{P}}(\lambda)\mathrm{d}\lambda := |\mathring{g}|^{1/2}\mathrm{d}\lambda = \left\{\prod_{i=1}^{d} \frac{1}{(s_j\lambda_j)^{1/2}}\right\}\mathrm{d}\lambda \propto \frac{1}{(\lambda_1\cdots\lambda_d)^{1/2}}\mathrm{d}\lambda$$

coinciding with the Jeffreys priors for $p(x \mid \lambda)$ and $\tilde{p}(y \mid \lambda)$.

The Laplacian $\mathring{\Delta}$ based on the predictive metric $\mathring{g}$ is given by

$$\mathring{\Delta} f = \left(\prod_{k=1}^{d} \lambda_k^{1/2}\right) \sum_{i=1}^{d} \frac{\partial}{\partial\lambda_i}\left(\frac{s_i\lambda_i}{\prod_{j=1}^{d}\lambda_j^{1/2}}\frac{\partial f}{\partial\lambda_i}\right) = \sum_i s_i\left(\lambda_i\frac{\partial^2 f}{\partial\lambda_i^2} + \frac{1}{2}\frac{\partial f}{\partial\lambda_i}\right),$$

where f is a smooth real function of λ.

Define

$$\pi_{\mathrm{S}}(\lambda)\mathrm{d}\lambda := \frac{(\lambda_1/s_1 + \cdots + \lambda_d/s_d)^{-(d/2-1)}}{\prod_j {\lambda_j}^{1/2}}\mathrm{d}\lambda \propto (\lambda_1/s_1 + \cdots + \lambda_d/s_d)^{-(d/2-1)}|\mathring{g}|^{1/2}\mathrm{d}\lambda.$$

Then, from

$$\frac{\partial}{\partial\lambda_i}\frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} = \left(-\frac{d}{2}+1\right)\left(\frac{\lambda_1}{s_1} + \cdots + \frac{\lambda_d}{s_d}\right)^{-\frac{d}{2}}\frac{1}{s_i} \tag{22}$$

and

$$\frac{\partial^2}{\partial\lambda_i^2}\frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} = \left(-\frac{d}{2}+1\right)\left(-\frac{d}{2}\right)\left(\frac{\lambda_1}{s_1} + \cdots + \frac{\lambda_d}{s_d}\right)^{-\frac{d}{2}-1}\left(\frac{1}{s_i}\right)^2,$$

we have

$$\mathring{\Delta}\frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} = \sum_i s_i \left(\lambda_i \frac{\partial^2}{\partial \lambda_i^2} \frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} + \frac{1}{2} \frac{\partial}{\partial \lambda_i} \frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} \right) = 0. \tag{23}$$

Since $\pi_{\mathrm{S}}/\pi_{\mathrm{P}}$ is a non-constant positive superharmonic function of λ, the Bayesian predictive density $\tilde{p}_{\mathrm{S}}(y \mid x)$ based on π_{S} asymptotically dominates $\tilde{p}_{\mathrm{P}}(y \mid x)$ by Corollary 1.

The model manifold endowed with the predictive metric $\mathring{g}_{ij}$ is isometric to the first orthant $\mathbb{R}^n_+ = \{(x^1, \cdots, x^n) : x^1 > 0, x^2 > 0, \cdots, x^n > 0\}$ of the Euclidean space $\mathbb{R}^n$, as we see below. Define

$$\xi^{i'} = 2\sqrt{\frac{\theta^{i'}}{s_{i'}}} \quad (i' = 1, \dots, d).$$

Then,

$$\frac{\partial \theta^i}{\partial \xi^{i'}} = \begin{cases} (s_{i'}\lambda_{i'})^{\frac{1}{2}} & (i = i') \\ 0 & (i \neq i'). \end{cases}$$

Thus, from (21), the coefficients of the metric with respect to $(\xi^{i'})$ are given by

$$\mathring{g}_{i'j'} = \sum_{i,j} \frac{\partial \theta^i}{\partial \xi^{i'}} \mathring{g}_{ij} \frac{\partial \theta^j}{\partial \xi^{j'}} = \begin{cases} 1 & (i' = j') \\ 0 & (i' \neq j'). \end{cases}$$

This coincides with the usual metric on $\mathbb{R}^n_+$.

Here, the function

$$\|\xi\|^{-d+2} \propto \frac{\pi_{\mathrm{S}}(\lambda)}{\pi_{\mathrm{P}}(\lambda)} = \left(\frac{\lambda_1}{s_1} + \cdots + \frac{\lambda_d}{s_d} \right)^{-\frac{d}{2}+1}$$

of ξ is the Green function of the heat equation on $\mathbb{R}^n$ and plays an essential role in Bayesian methods for model manifolds isometric to the Euclidean space. For example, the prior density $\|\mu\|^{-d+2}$ for the d-dimensional Normal model $N_d(\mu, I_d)$, where μ is the d-dimensional unknown mean vector and I_d is the $d \times d$ identity matrix, is known as the Stein prior.

The Bayesian predictive density based on π_{P} is

$$\begin{aligned} \tilde{p}_{\mathrm{P}}(y \mid x) &= \frac{\displaystyle\int \prod_{i=1}^d \left\{ \frac{{\lambda_i}^{x_i}}{x_i!} \mathrm{e}^{-\lambda_i} \frac{(s_i\lambda_i)^{y_i}}{y_i!} \mathrm{e}^{-s_i\lambda_i} \right\} \prod_{j=1}^d {\lambda_j}^{-1/2} \mathrm{d}\lambda}{\displaystyle\int \prod_{i=1}^d \left(\frac{{\lambda_i}^{x_i}}{x_i!} \mathrm{e}^{-\lambda_i} \right) \prod_{j=1}^d {\lambda_j}^{-1/2} \mathrm{d}\lambda} \\ &= \frac{\displaystyle\prod_{i=1}^d \left\{ \frac{s_i^{y_i}}{(1+s_i)^{x_i+y_i+1/2}} \frac{\Gamma(x_i + y_i + 1/2)}{x_i! y_i!} \right\}}{\displaystyle\prod_{i=1}^d \frac{\Gamma(x_i + 1/2)}{x_i!}} \end{aligned}$$

$$= \prod_{i=1}^{d} \left\{ \frac{s_i^{y_i}}{(1+s_i)^{x_i+y_i+1/2}} \frac{\Gamma(x_i+y_i+1/2)}{\Gamma(x_i+1/2) y_i!} \right\},$$

where $\mathrm{d}\lambda := \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d$.

The Bayesian predictive density based on π_{S} is

$$\tilde{p}_{\mathrm{S}}(y \mid x) = \frac{\int \prod_{i=1}^{d} \left\{ \frac{\lambda_i{}^{x_i}}{x_i!} \mathrm{e}^{-\lambda_i} \frac{(s_i\lambda_i)^{y_i}}{y_i!} \mathrm{e}^{-s_i\lambda_i} \right\} \left(\sum_{j=1}^{d} \frac{\lambda_j}{s_j} \right)^{-(d/2-1)} \prod_{k=1}^{d} \lambda_k{}^{-1/2} \mathrm{d}\lambda}{\int \prod_{i=1}^{d} \left(\frac{\lambda_i{}^{x_i}}{x_i!} \mathrm{e}^{-\lambda_i} \right) \left(\sum_{j=1}^{d} \frac{\lambda_j}{s_j} \right)^{-(d/2-1)} \prod_{k=1}^{d} \lambda_k{}^{-1/2} \mathrm{d}\lambda}$$

$$= \frac{\int \prod_{i=1}^{d} \left\{ \frac{s_i^{y_i} \lambda_i^{x_i+y_i-1/2}}{y_i!} \mathrm{e}^{-(1+s_i)\lambda_i} \right\} \left\{ \int_0^\infty u^{\frac{d}{2}-2} \exp\Bigl(-u \sum_j \frac{\lambda_j}{s_j}\Bigr) \mathrm{d}u \right\} \mathrm{d}\lambda}{\int \prod_{i=1}^{d} \left(\lambda_i{}^{x_i-1/2} \mathrm{e}^{-\lambda_i} \right) \left\{ \int_0^\infty u^{\frac{d}{2}-2} \exp\Bigl(-u \sum_j \frac{\lambda_j}{s_j}\Bigr) \mathrm{d}u \right\} \mathrm{d}\lambda}$$

$$= \frac{\int_0^\infty u^{\frac{d}{2}-2} \prod_{i=1}^{d} (1+s_i+u/s_i)^{-(x_i+y_i+1/2)} \mathrm{d}u}{\int_0^\infty u^{\frac{d}{2}-2} \prod_{i=1}^{d} (1+u/s_i)^{-(x_i+1/2)} \mathrm{d}u} \prod_{i=1}^{d} \frac{s_i^{y_i} \Gamma(x_i+y_i+1/2)}{y_i! \Gamma(x_i+1/2)}.$$

We have the asymptotic risk difference

$$\begin{aligned} N^2 & \Bigl[\mathrm{E}\bigl\{ D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{S}}(y \mid x) \bigm| \theta \bigr\} - \mathrm{E}\bigl\{ D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{P}}(y \mid x) \bigm| \theta \bigr\} \Bigr] \\ & = \frac{\mathring{\Delta}(\pi_{\mathrm{S}}/\pi_{\mathrm{P}})}{\pi_{\mathrm{S}}/\pi_{\mathrm{P}}} - \frac{1}{2} \sum_{i,j} \mathring{g}^{ij} \frac{\partial_i(\pi_{\mathrm{S}}/\pi_{\mathrm{P}}) \partial_j(\pi_{\mathrm{S}}/\pi_{\mathrm{P}})}{(\pi_{\mathrm{S}}/\pi_{\mathrm{P}})^2} + \mathrm{o}(1) \\ & = -\frac{1}{2} \left(\frac{d}{2} - 1 \right)^2 \left(\frac{\lambda_1}{s_1} + \cdots + \frac{\lambda_d}{s_d} \right)^{-1} + \mathrm{o}(1) \end{aligned} \tag{24}$$

by Theorem 2, (22), (23), and

$$\mathring{g}^{ij} = \begin{cases} s_i\lambda_i & (i = j) \\ 0 & (i \neq j). \end{cases}$$

The asymptotic risk difference (24) depends on λ only through $\lambda_1/s_1 + \cdots + \lambda_d/s_d$. When $\lambda_1/s_1 + \cdots + \lambda_d/s_d$ is small the improvement is large, and it converges to zero as $\lambda_1/s_1 + \cdots + \lambda_d/s_d$ goes to infinity.

It can be shown that π_{S} dominates π_{P} in the sense of infinitesimal prediction, and we can construct a Bayesian predictive density dominating $\tilde{p}_{\mathrm{P}}(y \mid x)$ for arbitrary $s_i > 0$

$(i = 1, \ldots, d)$ by modifying the prior π_{S}. Finite sample properties of this prior will be discussed in a another paper by using an approach different from the asymptotic methods in the present paper.

In Examples 1, 2, and 3, the volume element based on the predictive metric $\mathring{g}_{ij}$ coincides with the Jeffreys priors based on g_{ij} and $\tilde{g}_{ij}$, i.e. $|\mathring{g}_{ij}(\theta)|^{1/2} \propto |g_{ij}(\theta)|^{1/2} \propto |\tilde{g}_{ij}(\theta)|^{1/2}$, although the three metrics are different. In general, if two metrics g_{ij} and $\tilde{g}_{ij}$ satisfy the relation

$$\tilde{g}_{ij}(\theta) = \sum_{k,l} g_{kl}(\theta) A_i^k A_j^l, \tag{25}$$

where (A_j^i) is a $d \times d$ regular matrix not depending on θ, then

$$|\tilde{g}_{ij}|^{\frac{1}{2}} = |A_k^l||g_{ij}|^{\frac{1}{2}}, \quad \text{and} \quad |\mathring{g}_{ij}|^{\frac{1}{2}} = |g_{ij}||\tilde{g}_{ij}|^{-\frac{1}{2}} = |A_k^l|^{-1}|g_{ij}|^{\frac{1}{2}}$$

and the volume elements based on g_{ij}, $\tilde{g}_{ij}$, and $\mathring{g}_{ij}$ are proportional to each other. The relation (25) appears in many examples as in Examples 1, 2, and 3.

Appendix. Proofs of Theorems 1 and 2

First, we prepare a preliminary result, Theorem A1, to prove Theorem 1.

Asymptotic properties of predictive densities in the conventional setting in which $x(i)$, $i = 1, \ldots, N$, and y have the same distribution have been studied, see Komaki (1996), Hartigan (1998), and Sweeting et al. (2006).

Fushiki et al. (2004) generalized these results for the setting in which $x(i)$, $i = 1, \ldots, N$, and y have different distributions. The Bayesian predictive density is expanded as

$$\begin{aligned}
\tilde{p}_\pi(y \mid x^N) &= \tilde{p}(y \mid \hat{\theta}_{\mathrm{mle}}) + \frac{1}{2N} \sum_{i,j} g^{ij}(\hat{\theta}_{\mathrm{mle}}) \left\{ \partial_i \partial_j \tilde{p}(y \mid \hat{\theta}_{\mathrm{mle}}) - \sum_k \tilde{\Gamma}_{ij}^{(\mathrm{m})k} \partial_k \tilde{p}(y \mid \hat{\theta}_{\mathrm{mle}}) \right\} \\
&+ \frac{1}{2N} \sum_k \left[\sum_{i,j} g^{ij}(\hat{\theta}_{\mathrm{mle}}) \left\{ \tilde{\Gamma}_{ij}^{(\mathrm{m})k}(\hat{\theta}_{\mathrm{mle}}) - \Gamma_{ij}^{(\mathrm{m})k}(\hat{\theta}_{\mathrm{mle}}) \right\} \right. \\
&\left. + 2 \sum_i g^{ik}(\hat{\theta}_{\mathrm{mle}}) \left\{ \partial_i \log \pi(\hat{\theta}_{\mathrm{mle}}) - \sum_j \Gamma_{ij}^{(\mathrm{e})j}(\hat{\theta}_{\mathrm{mle}}) \right\} \right] \partial_k \tilde{p}(y \mid \hat{\theta}_{\mathrm{mle}}) + \mathrm{o_p}(N^{-1}),
\end{aligned} \tag{26}$$

where $\hat{\theta}_{\mathrm{mle}}$ is the maximum likelihood estimator, and $\partial_i := \partial/\partial\theta^i$. The estimator minimizing the Bayes risk $\int \mathrm{E}[D\{\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x)\}|\theta]\pi(\theta)\mathrm{d}\theta$ is given by

$$\hat{\theta}_\pi^i = \hat{\theta}_{\mathsf{mle}}^i + \frac{1}{N} w_\pi^i(\hat{\theta}_{\mathsf{mle}}) + \mathrm{o_p}(N^{-1}), \tag{27}$$

where

$$w^i_\pi(\theta) := \sum_k g^{ik}(\theta)\Big\{\partial_k \log \pi(\theta) - \sum_j \Gamma^{(\mathrm{e})j}_{kj}(\theta)\Big\} + \frac{1}{2}\sum_{k,l} g^{kl}(\theta)\Big\{\tilde{\Gamma}^{(\mathrm{m})i}_{kl}(\theta) - \Gamma^{(\mathrm{m})i}_{kl}(\theta)\Big\}, \tag{28}$$

which is a covariant vector.

The expansion of the risk function of a Bayesian predictive density $\tilde{p}_\pi(y \mid x^N)$ up to the order N^{-2} is given in Theorem A1 below. The expansion is invariant in the sense that each term is a scalar not depending on parametrization. In Theorem A1, we put

$$\begin{aligned}
v^{(\mathrm{e})}_{ij}(x;\theta) :=& \partial_i\partial_j \log p(x \mid \theta) + g_{ij}(\theta) - \sum_k \Gamma^{(\mathrm{e})k}_{ij}(\theta)\partial_k \log p(x \mid \theta), \\
\tilde{v}^{(\mathrm{m})}_{ij}(y;\theta) :=& \frac{1}{\tilde{p}(y \mid \theta)}\Big\{\partial_i\partial_j \tilde{p}(y \mid \theta) - \sum_k \tilde{\Gamma}^{(\mathrm{m})k}_{ij}\partial_k \tilde{p}(y \mid \theta)\Big\}, \\
T^{ijk} :=& \sum_{l,m,n} T_{lmn} g^{il} g^{jm} g^{kn}, \quad \text{and} \quad v^{(\mathrm{e})ik} := \sum_{j,l} v^{(\mathrm{e})}_{jl} g^{ij} g^{kl}.
\end{aligned}$$

Here, $v^{(\mathrm{e})}_{ij}$ and $\tilde{v}^{(\mathrm{m})}_{ij}$ are vectors orthogonal to the model manifolds $\{p(x \mid \theta) \,|\, \theta \in \Theta)\}$ and $\{\tilde{p}(y \mid \theta) \,|\, \theta \in \Theta)\}$, respectively. These vectors are closely related to the curvature of the manifolds.

Theorem A1. The expected Kullback–Leibler divergence from the true density $\tilde{p}(y \mid \theta)$ to the Bayesian predictive density $\tilde{p}_\pi(y \mid x^N)$ based on a prior $\pi(\theta)$ is expanded as

$$\begin{aligned}
&\mathrm{E}\Big\{D\big(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N)\big)\Big|\,\theta\Big\} \\
&= \frac{1}{2N}\sum_{i,j} \tilde{g}_{ij} g^{ij} + \frac{1}{2N^2}\sum_{i,j} \tilde{g}_{ij} u^i_\pi u^j_\pi + \frac{1}{N^2}\sum_{i,j,k} \tilde{g}_{ij} g^{jk} \widetilde{\nabla}^{(\mathrm{e})}_k u^i_\pi \\
&\quad + \frac{1}{2N^2}\sum_{i,j,k,l} \mathrm{E}\left(v^{(\mathrm{e})ik} v^{(\mathrm{e})jl} \,\Big|\, \theta\right) g_{kl}\tilde{g}_{ij} \\
&\quad - \frac{1}{2N^2}\sum_{i,j,k,l} \mathrm{E}\left(\tilde{v}^{(\mathrm{m})}_{ij} \tilde{v}^{(\mathrm{m})}_{kl} \,\Big|\, \theta\right) g^{ij} g^{kl} - \frac{1}{3N^2}\sum_{i,j,k} \widetilde{T}_{ijk} T^{ijk} \\
&\quad + \frac{3}{4N^2}\sum_{i,j,k,l} \tilde{Q}_{ijkl} g^{ij} g^{kl} - \frac{1}{N^2}\sum_{i,j,k,l} \mathrm{E}\left(\partial_i \tilde{l} \partial_j \tilde{l}\, \tilde{v}^{(\mathrm{m})}_{kl} \,\Big|\, \theta\right) g^{ik} g^{jl} \\
&\quad + \frac{1}{4N^2}\sum_{i,j,k,l} \mathrm{E}\left(\tilde{v}^{(\mathrm{m})}_{ij} \tilde{v}^{(\mathrm{m})}_{kl} \,\Big|\, \theta\right) g^{ik} g^{jl} \\
&\quad + \frac{1}{4N^2}\sum_{i,j,k,l,m,n} \tilde{g}_{ij} (\tilde{\Gamma}^{(\mathrm{m})i}_{kl} - \Gamma^{(\mathrm{m})i}_{kl})(\tilde{\Gamma}^{(\mathrm{m})j}_{mn} - \Gamma^{(\mathrm{m})j}_{mn}) g^{km} g^{ln} \\
&\quad - \frac{1}{N^2}\sum_{i,j,k,l,m} \widetilde{T}_{ijk} (\tilde{\Gamma}^{(\mathrm{m})k}_{lm} - \Gamma^{(\mathrm{m})k}_{lm}) g^{il} g^{jm} + \mathrm{o}(N^{-2}),
\end{aligned} \tag{29}$$

where

$$u_\pi^i(\theta) := \sum_k g^{ik}(\theta)\Big\{\partial_k \log \pi(\theta) - \sum_j \Gamma^{(\mathrm{e})j}_{kj}(\theta)\Big\} + \sum_{k,l} g^{kl}(\theta)\Big\{\tilde{\Gamma}^{(\mathrm{m})i}_{kl}(\theta) - \Gamma^{(\mathrm{m})i}_{kl}(\theta)\Big\}.$$

Outline of the Proof. Expansions of the risk functions corresponding to (29) when the distributions of $x(i)$, $i = 1, \dots, N$, and y are the same are obtained by Komaki (1996) for curved exponential families by using differential geometrical notions and by Hartigan (1998) for general models under rigorous regularity conditions. Fushiki et al. (2004) obtained several related results when when the distributions of $x(i)$, $i = 1, \dots, N$, and y are different. The expansion (29) is shown by lengthy calculations parallel to those in Komaki (1996) and Hartigan (1998) by using the results such as (26), (27), and (28) obtained by Fushiki et al. (2004). □

The quantity $\sum_{i,j,k,l} \mathrm{E}\left(v^{(\mathrm{e})ik} v^{(\mathrm{e})jl} \,\middle|\, \theta\right) g_{kl}$ is the Efron curvature (Efron, 1975) of the model manifold $\{p(x \mid \theta) \,|\, \theta \in \Theta\}$ at θ, and $\sum_{i,j,k,l} \mathrm{E}\left(\tilde{v}^{(\mathrm{m})}_{ij} \tilde{v}^{(\mathrm{m})}_{kl} \,\middle|\, \theta\right) g^{ij} g^{kl}$ is the mixture mean curvature discussed in Komaki (1996) of the model manifold $\{\tilde{p}(y \mid \theta) \,|\, \theta \in \Theta\}$ at θ.

Proof of Theorem 1. The desired result is obvious from Theorem A1 because (29) has the form

$$\begin{aligned}
&\mathrm{E}\Big\{D\big(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N)\big) \,\Big|\, \theta\Big\} \\
&\quad = \frac{1}{2N} \sum_{i,j} \tilde{g}_{ij} g^{ij} + \frac{1}{2N^2} \sum_{i,j} \tilde{g}_{ij} u_\pi^i u_\pi^j + \frac{1}{N^2} \sum_{i,j,k} \tilde{g}_{ij} g^{jk} \widetilde{\nabla}^{(\mathrm{e})}_k u_\pi^i \\
&\qquad + \text{terms independent of } \pi + \mathrm{o}(N^{-2}).
\end{aligned} \tag{30}$$

□

To derive Theorem 1, it is sufficient to show (30). Much less calculation is needed to verify (30) than to obtain all the explicit terms in (29).

Proof of Theorem 2. Let $f(\theta) := \pi(\theta)/\pi_{\mathrm{P}}(\theta)$. Since

$$\frac{1}{2} \sum_i \sum_j \mathring{g}^{ij} \partial_i \log f \partial_j \log f + \mathring{\Delta} \log f = \frac{\mathring{\Delta} f}{f} - \frac{1}{2} \sum_i \sum_j \mathring{g}^{ij} \frac{\partial_i f \partial_j f}{f^2} = 2 \frac{\mathring{\Delta} f^{\frac{1}{2}}}{f^{\frac{1}{2}}}, \tag{31}$$

it is sufficient to show that the left-hand side of (10) is equal to $(1/2)\mathring{g}^{ij} \partial_i \log f \partial_j \log f + \mathring{\Delta} \log f$.

From (6) and (7), we have

$$\partial_i \log \pi = \partial_i \log f + \partial_i \log \pi_{\mathrm{P}} = \partial_i \log f + \sum_{j,k} \mathring{\Gamma}^{(0)}_{ijk} \mathring{g}^{jk}.$$

Let $r^j := \sum_{k,l} g^{kl} \left(\tilde{\Gamma}^{(\mathrm{m})j}_{kl} - \Gamma^{(\mathrm{m})j}_{kl} \right)$ and $s^i := \sum_{k,j} g^{ik} \big(\mathring{\Gamma}^{(0)j}_{kj} - \Gamma^{(\mathrm{e})j}_{kj} \big)$. Then, from (28),

$$\begin{aligned} u^i_\pi &= \sum_{i,k} g^{ik} \left(\partial_k \log \pi - \sum_j \Gamma^{(\mathrm{e})j}_{kj} \right) + r^i \\ &= \sum_k g^{ik} \left(\partial_k \log f + \sum_j \mathring{\Gamma}^{(0)j}_{kj} - \sum_j \Gamma^{(\mathrm{e})j}_{kj} \right) + r^i \\ &= \sum_k g^{ik} \partial_k \log f + s^i + r^i. \end{aligned} \tag{32}$$

Thus, when $\pi = \pi_{\mathrm{P}}$, $u^i_{\mathrm{P}} = s^i + r^i$. From (4), we have

$$\begin{aligned} & N^2 \bigg(\mathrm{E}\big[D(\tilde{p}(y \mid \theta), \tilde{p}_\pi(y \mid x^N) \big] - \mathrm{E}\big[D(\tilde{p}(y \mid \theta), \tilde{p}_{\mathrm{P}}(y \mid x^N) \big] \bigg) \\ =& \frac{1}{2} \sum_{i,j} \tilde{g}_{ij} u^i_\pi u^j_\pi + \sum_{i,j,k} \tilde{g}_{ij} g^{jk} \left(\partial_k u^i_\pi + \sum_l \tilde{\Gamma}^{(\mathrm{e})i}_{kl} u^l_\pi \right) \\ & - \frac{1}{2} \sum_{i,j} \tilde{g}_{ij} u^i_{\mathrm{P}} u^j_{\mathrm{P}} - \sum_{i,j,k} \tilde{g}_{ij} g^{jk} \left(\partial_k u^i_{\mathrm{P}} + \sum_l \tilde{\Gamma}^{(\mathrm{e})i}_{kl} u^l_{\mathrm{P}} \right) + \mathrm{o}(1) \\ =& \frac{1}{2} \sum_{i,j} \tilde{g}_{ij} (\sum_k g^{ik} \partial_k \log f + s^i + r^i)(\sum_l g^{jl} \partial_l \log f + s^j + r^j) \\ & - \frac{1}{2} \sum_{i,j} \tilde{g}_{ij} (s^i + r^i)(s^j + r^j) \\ & + \sum_{i,j,k} \tilde{g}_{ij} g^{jk} \left\{ \sum_l \partial_k (g^{il} \partial_l \log f) + \sum_{l,m} \tilde{\Gamma}^{(\mathrm{e})i}_{kl} g^{lm} \partial_m \log f \right\} + \mathrm{o}(1) \\ =& \frac{1}{2} \sum_{i,j} \mathring{g}^{ij} \partial_i \log f \partial_j \log f + \sum_{i,j,k} \tilde{g}_{ij} g^{ik} (\partial_k \log f)(s^j + r^j) \\ & + \sum_{i,j,k,l} \tilde{g}_{ij} g^{ik} \partial_k (g^{jl} \partial_l \log f) + \sum_{i,j,k,l,m} \tilde{g}_{ij} g^{jk} \tilde{\Gamma}^{(\mathrm{e})i}_{jk} g^{lm} \partial_m \log f + \mathrm{o}(1). \end{aligned} \tag{33}$$

Let $L_i := \partial_i \log f$. From (31), it is sufficient to show that

$$\sum_{i,j,k} \tilde{g}_{ij} g^{ik} L_k (s^j + r^j) + \sum_{i,j,k,l} \tilde{g}_{ij} g^{ik} \partial_k (g^{jl} L_l) + \sum_{j,k,l,m} g^{jk} \tilde{\Gamma}^{(\mathrm{e})}_{klj} g^{lm} L_m \tag{34}$$

is equal to $\mathring{\Delta}\log f = \sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) + \sum_{i,j,k}\mathring{\Gamma}^{(0)i}_{ij}\mathring{g}^{jk}L_k$. Since

$$0 = \partial_i \delta^l_f = \partial_i(\sum_m g^{lm}g_{mn}) = \sum_m (\partial_i g^{lm})g_{mn} + \sum_m g^{lm}(\partial_i g_{mn}),$$

we have

$$\partial_i g^{lm} = -\sum_{j,k} g^{jl}g^{km}(\partial_i g_{jk}).$$

Thus, from

$$\begin{aligned}\sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) &= \sum_{i,j,k,l}\partial_i(g^{ik}g^{jl}\tilde{g}_{kl}L_j)\\ &= \sum_{i,j,k,l}(\partial_i g^{ik})g^{jl}\tilde{g}_{kl}L_j + \sum_{i,j,k,l}(\partial_i\tilde{g}_{kl})g^{ik}g^{jl}L_j + \sum_{i,j,k,l}g^{ik}\tilde{g}_{kl}\partial_i(g^{jl}L_j),\end{aligned}$$

we have

$$\begin{aligned}\sum_{i,j,k,l}\tilde{g}_{kl}g^{ik}\partial_i(g^{jl}L_j) =& \sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) - \sum_{i,j,k,l}(\partial_i g^{ik})g^{jl}\tilde{g}_{kl}L_j - \sum_{i,j,k,l}(\partial_i\tilde{g}_{kl})g^{ik}g^{jl}L_j\\ =& \sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) + \sum_{i,j,m,n}(\partial_i g_{mn})g^{im}\mathring{g}^{jn}L_j - \sum_{i,j,k,l}(\partial_i\tilde{g}_{kl})g^{ik}g^{jl}L_j.\end{aligned}$$

Hence, because of the duality (3) of the e-connection and the m-connection, (34) is equal to

$$\begin{aligned}&\sum_{i,j,k,l}\tilde{g}_{ij}g^{ik}L_k g^{jl}\Big(\sum_m \mathring{\Gamma}^{(0)m}_{lm} - \sum_m \Gamma^{(\mathrm{e})m}_{lm}\Big) + \sum_{i,j,k,l,m}\tilde{g}_{ij}g^{ik}L_k g^{lm}\big(\tilde{\Gamma}^{(\mathrm{m})j}_{lm} - \Gamma^{(\mathrm{m})j}_{lm}\big)\\ &\quad + \sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) + \sum_{i,j,k,l}(\partial_i g_{jl})g^{il}\mathring{g}^{jk}L_k - \sum_{i,j,k,l}(\partial_i\tilde{g}_{jl})g^{il}g^{jk}L_k + \sum_{j,k,l,m}g^{jk}\tilde{\Gamma}^{(\mathrm{e})}_{klj}g^{lm}L_m\\ &= \sum_{i,j}\partial_i(\mathring{g}^{ij}L_j) + \sum_{k,l,m}\mathring{g}^{kl}\mathring{\Gamma}^{(0)m}_{lm}L_k\\ &\quad - \sum_{i,j,k,l}\left(\Gamma^{(\mathrm{e})}_{ijl} + \Gamma^{(\mathrm{m})}_{ilj} - \partial_i g_{jl}\right)g^{il}\mathring{g}^{jk}L_k + \sum_{i,j,k,l}\left(\tilde{\Gamma}^{(\mathrm{e})}_{ijl} + \tilde{\Gamma}^{(\mathrm{m})}_{ilj} - \partial_i\tilde{g}_{jl}\right)g^{il}g^{jk}L_k\\ &= \mathring{\Delta}\log f.\end{aligned}$$

□

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: Cambridge University Press. 32

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. New York: Springer-Verlag. 32

Berger, J. O. and Bernardo, J. M. (1992). "On the development of reference priors (with discussion)." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 35–60. New York: Oxford University Press. 40

Davies, E. B. (1989). *Heat Kernels and Spectral Theory*. Cambridge: Cambridge University Press. 41

Efron, B. (1975). "Defining curvature of a statistical problem (with applications to second order efficiency)." *Annals of Statistics*, 3: 1189–1242. 48

Fushiki, T., Komaki, F., and Aihara, K. (2004). "On parametric bootstrapping and Bayesian prediction." *Scandinavian Journal of Statistics*, 31: 403–416. 33, 46, 48

Geisser, S. (1993). *Predictive Inference: An Introduction*. New York: Chapman & Hall. 32

George, E. I., Liang, F., and Xu, X. (2006). "Improved minimax prediction under Kullback–Leibler loss." *Annals of Statistics*, 34: 78–91. 32

George, E. I. and Xu, X. (2008). "Predictive density estimation for multiple regression." *Econometric Theory*, 24: 528–544. 33, 38

Hartigan, J. A. (1998). "The maximum likelihood prior." *Annals of Statistics*, 26: 2083–2103. 32, 46, 48

Helgason, S. (1984). *Groups and Geometric Analysis*. Orlando, FL: Academic Press. 40

Kobayashi, K. and Komaki, F. (2008). "Bayesian shrinkage prediction for the regression problem." *Journal of Multivariate Analysis*, 99: 1888–1905. 33, 38

Komaki, F. (1996). "On asymptotic properties of predictive distributions." *Biometrika*, 83: 299–313. 32, 46, 48

— (2001). "A shrinkage predictive distribution for multivariate normal observables." *Biometrika*, 88: 859–864. 32

— (2004). "Simultaneous prediction of independent Poisson observables." *Annals of Statistics*, 32: 1744–1769. 43

— (2006). "Shrinkage priors for Bayesian prediction." *Annals of Statistics*, 34: 808–819. 33

— (2007). "Bayesian prediction based on a class of shrinkage priors for location-scale models." *Annals of the Institute of Statistical Mathematics*, 59: 135–146. 41

Sweeting, T. J., Datta, G. S., and Ghosh, M. (2006). "Nonsubjective priors via predictive relative entropy regret." *Annals of Statistics*, 34: 441–468. 46

Zidek, J. V. (1969). "A representation of Bayesian invariant procedures in terms of Haar measure." *Annals of the Institute of Statistical Mathematics*, 21: 291–308. 40

Acknowledgments

The author appreciates constructive comments of the associate editor. This research was partially supported by Grant-in-Aid for Scientific Research (23650144, 26280005).