AISM (2007) 59:135–146
DOI 10.1007/s10463-006-0102-4

Bayesian prediction based on a class of shrinkage priors for location-scale models

Fumiyasu Komaki

Received: 31 May 2006 / Revised: 8 November 2006 /
Published online: 14 February 2007

Abstract A class of shrinkage priors for multivariate location-scale models is introduced. We consider Bayesian predictive densities for location-scale models and evaluate performance of them using the Kullback–Leibler divergence. We show that Bayesian predictive densities based on priors in the introduced class asymptotically dominate the best invariant predictive density.

Keywords Asymptotic theory · Jeffreys prior · Neyman–Scott model · Right invariant prior · Kullback–Leibler divergence

1 Introduction

Let $p(x)$ be a probability density on $\mathbb{R}$ that is symmetric about 0, and let $x(l)$ be an observation $x(l) = (x_1(l), x_2(l), \ldots, x_{d-1}(l))$ from a probability density

$$p(x_1, x_2, \ldots, x_{d-1}|\mu, \sigma) := \prod_{i=1}^{d-1} p(x_i|\mu_i, \sigma) := \prod_{i=1}^{d-1} \frac{1}{\sigma} p\left(\frac{x_i - \mu_i}{\sigma}\right), \qquad (1)$$

where $\mu = (\mu_1, \ldots, \mu_{d-1}) \in \mathbb{R}^{d-1}$ and $\sigma > 0$ are unknown parameters. In the following, we call the model (1) a multidimensional location-scale model. When $p(x)$ is the standard normal density, (1) is the Neyman–Scott model (Neyman and Scott 1948). The location-scale model introduced by Fisher (1934) is one of the most important examples of group models both in theoretical and practical

F. Komaki (✉)
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
e-mail: komaki@mist.i.u-tokyo.ac.jp

aspects. The Neyman–Scott model has been extensively studied since it is a typical model including many nuisance parameters.

Suppose that we have a set of independent observations $x^{(N)} = (x(1), x(2), \ldots, x(N))$. An unobserved variable $y := x(N+1) = (x_1(N+1), x_2(N+1), \ldots, x_{d-1}(N+1))$ from the same density (1) is predicted by using a predictive density $\hat{p}(y; x^{(N)})$.

We adopt the Kullback–Leibler divergence

$$D\{p(y|\mu,\sigma), \hat{p}(y; x^{(N)})\} := \int p(y|\mu,\sigma) \log \frac{p(y|\mu,\sigma)}{\hat{p}(y; x^{(N)})} \mathrm{d}y, \tag{2}$$

which has a natural information theoretic meaning, as a loss function. The risk function is

$$\mathrm{E}[D(p,\hat{p})|\mu,\sigma] = \int p(x^{(N)}|\mu,\sigma) \int p(y|\mu,\sigma) \log \frac{p(y|\mu,\sigma)}{\hat{p}(y; x^{(N)})} \mathrm{d}y \mathrm{d}x^{(N)}. \tag{3}$$

A widely used method to construct a predictive density is to use a plug-in density $p(y|\hat{\mu}(x^{(N)}), \hat{\sigma}(x^{(N)}))$, where $\hat{\mu}(x^{(N)})$ and $\hat{\sigma}(x^{(N)})$ are appropriate estimators of μ and σ. However, Bayesian predictive densities have better performance than plug-in distributions in many examples, see Aitchison (1975), Aitchison and Dunsmore (1975), Geisser (1993), and Komaki (1996).

When we use the Bayesian procedure, the choice of a prior distribution is an important problem. Non-informative prior distributions or vague prior distributions such as the Jeffreys prior are widely used to construct Bayesian predictive distributions. Improper priors are often used for location-scale models. The discussion based on the Bayes risk

$$\mathrm{E}_\pi[D(p,\hat{p})] = \int \pi(\mu,\sigma) \int p(x^{(N)}|\mu,\sigma) \int p(y|\mu,\sigma) \log \frac{p(y|\mu,\sigma)}{\hat{p}(y; x^{(N)})} \mathrm{d}y \mathrm{d}x^{(N)} \mathrm{d}\mu \mathrm{d}\sigma, \tag{4}$$

where $\pi(\mu,\sigma)$ is a prior density, is not valid when we adopt an improper prior. We evaluate the performance of predictive distributions by using the risk function (3).

The univariate location-scale model is investigated in Example 2 in Komaki (2006), and it is shown that the right invariant prior, based on which the best invariant predictive density is constructed, is a positive superharmonic function on the model manifold.

In the present paper, the multivariate location-scale model is considered. A class of shrinkage priors for location-scale models is introduced and properties of them are investigated by using the results of previous studies on asymptotic properties of predictive densities. Shrinkage priors give more weight to parameter values close to a point or a subspace in the parameter space than the Jeffreys prior does. Information geometrical approach (Amari 1985; Amari and Nagaoka 2000) is useful to investigate Bayesian methods because prior

distributions are naturally regarded as volume elements on model manifolds. Bayesian predictive densities based on priors in the introduced class asymptotically dominate the best invariant one, even when the dimension of the model is 2, corresponding to the univariate location-scale model, although shrinkage methods effectively work when the dimension of the model is large in many examples. Liang and Barron (2004) showed that the best invariant predictive density for the location-scale model is minimax. However, it has been an open problem whether the best invariant predictive density is admissible or not. The result in this paper gives a negative answer to this problem in the asymptotic sense.

In Sect. 2, we show that the model manifold endowed with the Fisher metric is isometric to the Hyperbolic space and introduce useful coordinates on the model manifold. In Sect. 3, we formulate the model as a group model and obtain the right invariant prior. The Bayesian predictive density based on the right invariant prior is the best invariant predictive density. In Sect. 4, we introduce a class of prior densities for multivariate location scale models. The performance of Bayesian predictive densities based on the Jeffreys prior, the right invariant prior, and the proposed priors are evaluated by using asymptotic theory. It is shown that Bayesian predictive densities based on priors in the introduced class asymptotically dominate the best invariant predictive density.

2 The geometry of multivariate location-scale models

We obtain the Fisher information matrix for the multivariate location-scale model and show that the model manifold is isometric to the Hyperbolic space. The parameter space is regarded as a coordinate system of the model manifold.

We put

$$\xi^i = \mu_i \quad (i = 1, \ldots, d-1), \quad \text{and} \quad \xi^d = \sigma.$$

The Fisher metric on the multivariate location-scale model (1) is given by

$$(g_{ij}) = \left(\mathrm{E}\left[\left\{ \frac{\partial}{\partial \xi^i} \log p(x|\xi) \right\} \left\{ \frac{\partial}{\partial \xi^j} \log p(x|\xi) \right\} \middle| \xi \right] \right) = \begin{pmatrix} \frac{\alpha^2}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \frac{\alpha^2}{\sigma^2} & 0 \\ 0 & \cdots & 0 & \frac{(d-1)\beta^2}{\sigma^2} \end{pmatrix},$$

where

$$\alpha^2 := \int \left(\frac{p'(x)}{p(x)} \right)^2 p(x) \mathrm{d}x$$

and

$$\beta^2 := \int \left(x \frac{p'(x)}{p(x)} + 1 \right)^2 p(x) \mathrm{d}x.$$

By rescaling μ as $\xi^i := \alpha/(\sqrt{d-1}\beta)\mu_i$ $(i = 1, \ldots, d-1)$, the metric tensor is given by

$$(g_{ij}) = \begin{pmatrix} \frac{(d-1)\beta^2}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \frac{(d-1)\beta^2}{\sigma^2} & 0 \\ 0 & \cdots & 0 & \frac{(d-1)\beta^2}{\sigma^2} \end{pmatrix}. \tag{5}$$

Then, the model manifold is isometric to the Hyperbolic space $H^d(-1/\{(d-1)\beta^2\})$ defined below (6).

Since the determinant of the $n \times n$ matrix (5) is

$$|g| = \frac{\beta^{2d}(d-1)^d}{\sigma^{2d}},$$

the Jeffreys prior is given by

$$|g|^{1/2} \mathrm{d}\xi^1 \cdots \mathrm{d}\xi^d \propto \sigma^{-d} \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma.$$

We put $\pi_J(\mu, \sigma) = \sigma^{-d}$. The prior density σ^{-1} is better than the Jeffreys prior in many problems, see Sect. 3.

Example The metric on the model manifold of $\mathrm{N}(\mu, \sigma^2)$ is given by

$$\frac{1}{\sigma^2}\mathrm{d}\mu^2 + \frac{2}{\sigma^2}\mathrm{d}\sigma^2 = \frac{2}{\sigma^2}(\mathrm{d}\xi^1)^2 + \frac{2}{\sigma^2}(\mathrm{d}\xi^2)^2,$$

where

$$\xi^1 := \frac{\alpha}{\sqrt{d-1}\beta}\mu = \frac{1}{\sqrt{2}}\mu, \quad \xi^2 = \sigma.$$

Since $d = 2$, $\alpha^2 = 1$, and $\beta^2 = 2$, as is widely known, the model manifold endowed with the Fisher metric is isometric to the Hyperbolic plane $\mathrm{H}^2(-1/2)$, see Amari (1985).

We summarize properties of the hyperbolic space to be used.

The metric on the Hyperbolic space $\mathrm{H}^d(-1/a^2)$ with constant sectional curvature $-1/a^2$ in the upper-half space coordinates

$$(u, v), \quad u = (u_1, u_2, \ldots, u_{d-1}) \in \mathbb{R}^{d-1}, \quad v > 0$$

is given by

$$a^2 \frac{\mathrm{d}{u_1}^2 + \cdots + \mathrm{d}{u_{d-1}}^2 + \mathrm{d}v^2}{v^2}, \tag{6}$$

see Davis (1989) p. 176.

The volume element induced by the metric (6) of $\mathrm{H}^d(-1/a^2)$ is

$$\frac{a^d}{v^d} \mathrm{d}u_1 \wedge \mathrm{d}u_2 \wedge \cdots \wedge \mathrm{d}u_{d-1} \wedge \mathrm{d}v$$

corresponding to the Jeffreys prior $\pi_{\mathrm{J}}(\mu, \sigma)$.

The Laplacian on $\mathrm{H}^d(-1/a^2)$ in the upper-half space coordinates is

$$\Delta = \frac{1}{a^2} \left\{ v^2 \left(\Delta_u + \frac{\partial^2}{\partial v^2} \right) - (d-2) v \frac{\partial}{\partial v} \right\},$$

where Δ_u is the Laplacian on $\mathbb{R}^{d-1}$, since the Laplacian Δ on a manifold (M, g) endowed with a Riemannian metric g_{ij} is defined by

$$\Delta f = \frac{1}{\sqrt{|g|}} \partial_i \left(\sqrt{|g|} g^{ij} \partial_j f \right) = \nabla_i (g^{ij} \partial_j f),$$

where f is a real function on M, see Davis (1989) p. 176.

Another important coordinate system for $\mathrm{H}^d(-1/a^2)$ used in the following is the geodesic polar coordinates. A point on $\mathrm{H}^d(-1/a^2)$ is represented by the distance $a\rho$ and direction from a fixed point O on $\mathrm{H}^d(-1/a^2)$. The direction is represented by a point on the unit sphere in the tangent space of $\mathrm{H}^d(-1/a^2)$ at O.

The Riemannian metric on $\mathrm{H}^2(-1/a^2)$ in the geodesic polar coordinates is given by

$$g = a^2 (\mathrm{d}\rho^2 + \sinh^2 \rho \, \mathrm{d}\tau^2),$$

where $\mathrm{d}\tau^2$ is the Riemannian metric on the unit sphere in the tangent space at O.

The Laplacian in the geodesic polar coordinates is given by

$$\Delta = \frac{1}{a^2} \left\{ \frac{\partial^2}{\partial \rho^2} + (d-1) \frac{\cosh \rho}{\sinh \rho} \frac{\partial}{\partial \rho} + (\sinh \rho)^{-2} \Delta_{\mathrm{S}} \right\}, \tag{7}$$

where Δ_S is the Laplacian on the unit sphere in the tangent space at O, see Helgason (1984) p. 158.

Let $a\rho$ be the Riemannian distance between two points (u, v) and $(\bar{u}, \bar{v})$ in the upper-half space coordinates. It is known that the relations

$$\frac{1}{2} + \frac{1}{2}\cosh\rho = \cosh^2(\rho/2) = \frac{|u-\bar{u}|^2 + (v+\bar{v})^2}{4v\bar{v}} \tag{8}$$

and

$$\cosh\rho = \frac{|u-\bar{u}|^2 + v^2 + \bar{v}^2}{2v\bar{v}}. \tag{9}$$

hold.

3 The best invariant predictive density for multivariate location-scale models

Multivariate location-scale models can be formulated as a group model. We obtain the left and right invariant priors for the model. The right invariant prior is important when we consider Bayesian procedures.

Suppose that ε_i $(i = 1, \ldots, d-1)$ independently follows a probability density $p(x)$ on $\mathbb{R}$ that is symmetric about 0. Let

$$\begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_{d-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & \sigma & 0 & \cdots & 0 \\ \mu_2 & 0 & \sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \mu_{d-1} & 0 & \cdots & 0 & \sigma \end{pmatrix} \begin{pmatrix} 1 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{d-1} \end{pmatrix}.$$

Then $(x_1, x_2, \ldots, x_{d-1})$ follows (1).

We consider a group of matrices

$$\left\{ \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & \sigma & 0 & \cdots & 0 \\ \mu_2 & 0 & \sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \mu_{d-1} & 0 & \cdots & 0 & \sigma \end{pmatrix} \middle| \; \mu_i \in \mathbb{R}\ (i = 1, \ldots, d-1), \sigma > 0 \right\}$$

under the multiplication

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \bar{\mu}_1 & \bar{\sigma} & 0 & \cdots & 0 \\ \bar{\mu}_2 & 0 & \bar{\sigma} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \bar{\mu}_{d-1} & 0 & \cdots & 0 & \bar{\sigma} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & b & 0 & \cdots & 0 \\ a_2 & 0 & b & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{d-1} & 0 & \cdots & 0 & b \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & \sigma & 0 & \cdots & 0 \\ \mu_2 & 0 & \sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \mu_{d-1} & 0 & \cdots & 0 & \sigma \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 + b\mu_1 & b\sigma & 0 & \cdots & 0 \\ a_2 + b\mu_2 & 0 & b\sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{d-1} + b\mu_{d-1} & 0 & \cdots & 0 & b\sigma \end{pmatrix}. \tag{10}$$

This group is identified with the parameter space $\{(\mu,\sigma)|\mu \in \mathbb{R}^{d-1}, \sigma > 0\}$ for the location-scale model. Thus the location-scale model is a group model.

From (10), we have

$$\begin{aligned} \mathrm{d}\bar{\mu}_1 \wedge \cdots \wedge \mathrm{d}\bar{\mu}_{d-1} \wedge \mathrm{d}\bar{\sigma} &= b^d \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma \\ &= \left(\frac{\bar{\sigma}}{\sigma}\right)^d \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma, \end{aligned}$$

and

$$\bar{\sigma}^{-d} \mathrm{d}\bar{\mu}_1 \wedge \cdots \wedge \mathrm{d}\bar{\mu}_{d-1} \wedge \mathrm{d}\bar{\sigma} = \sigma^{-d} \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma.$$

Thus, the left invariant prior is given by

$$\pi_{\mathsf{L}}(\mu,\sigma) \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma = \sigma^{-d} \mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma,$$

which coincides with the Jeffreys prior.

Since

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \bar{\mu}_1 & \bar{\sigma} & 0 & \cdots & 0 \\ \bar{\mu}_2 & 0 & \bar{\sigma} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \bar{\mu}_{d-1} & 0 & \cdots & 0 & \bar{\sigma} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & \sigma & 0 & \cdots & 0 \\ \mu_2 & 0 & \sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \mu_{d-1} & 0 & \cdots & 0 & \sigma \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & b & 0 & \cdots & 0 \\ a_2 & 0 & b & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{d-1} & 0 & \cdots & 0 & b \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \mu_1 + a_1\sigma & b\sigma & 0 & \cdots & 0 \\ \mu_2 + a_2\sigma & 0 & b\sigma & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \mu_{d-1} + a_{d-1}\sigma & 0 & \cdots & 0 & b\sigma \end{pmatrix},$$

we have

$$\begin{aligned} \mathrm{d}\bar{\mu}_1 \wedge \cdots \wedge \mathrm{d}\bar{\mu}_{d-1} \wedge \mathrm{d}\bar{\sigma} &= (\mathrm{d}\mu_1 + a_1\mathrm{d}\sigma) \wedge \cdots \wedge (\mathrm{d}\mu_{d-1} + a_{d-1}\mathrm{d}\sigma) \wedge b\mathrm{d}\sigma \\ &= b\mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma \\ &= \frac{\bar{\sigma}}{\sigma}\mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma. \end{aligned}$$

Since $\bar{\sigma}^{-1}\mathrm{d}\bar{\mu}_1 \wedge \cdots \wedge \mathrm{d}\bar{\mu}_{d-1} \wedge \mathrm{d}\bar{\sigma} = \sigma^{-1}\mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma$, the right invariant prior is given by

$$\pi_{\mathsf{R}}(\mu,\sigma)\mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma = \sigma^{-1}\mathrm{d}\mu_1 \wedge \cdots \wedge \mathrm{d}\mu_{d-1} \wedge \mathrm{d}\sigma \tag{11}$$

The Bayesian predictive density based on the right invariant prior is the best invariant predictive density (Zidek 1969; Komaki 2002), and the best invariant predictive density for the location-scale model is minimax (Liang and Barron 2004), conditioning on at least two observations ($N = 2$). This prior (11) has been recommended when $p(x)$ is the standard normal density also in the context of the reference prior theory, see for example Robert (2001) p. 133.

4 A class of shrinkage priors

We introduce a class of priors defined by

$$\begin{aligned} \frac{f_{c,\bar{u},\bar{v}}}{\pi_{\mathsf{J}}}(\mu,\sigma) &\propto \frac{1}{(\cosh\rho + c)^{d-1}} \\ &= \left(\frac{2v\bar{v}}{|u-\bar{u}|^2 + c(v+\bar{v})^2 + (1-c)(v^2+\bar{v}^2)}\right)^{d-1} \quad (0 \le c \le 1), \end{aligned} \tag{12}$$

where $(u, v) = (\alpha\beta^{-1}(d-1)^{-1/2}\mu, \sigma)$, the point $(\bar{u}, \bar{v})$ is an arbitrarily fixed point that is the center of shrinkage, and $\sqrt{d-1}\beta\rho$ is the Riemannian distance between (u, v) and $(\bar{u}, \bar{v})$. The equality in (12) holds because (8) and (9). When we set $\bar{u} = 0$, we have

$$\frac{f_{c,\bar{v}}}{\pi_{\mathrm{J}}}(\mu, \sigma) \propto \frac{1}{(\cosh\rho + c)^{d-1}} = \left(\frac{2\bar{v}\sigma}{\alpha^2\beta^{-2}(d-1)^{-1}|\mu|^2 + c(\sigma + \bar{v})^2 + (1-c)(\sigma^2 + \bar{v}^2)}\right)^{d-1}.$$

We evaluate the risk of Bayesian predictive densities based on the right invariant prior and priors in the proposed class by using the following theorem on asymptotic properties of predictive distributions. Information geometry provides a useful viewpoint over the results concerning predictive distributions for general models.

Theorem (*Komaki 2006*) *Let $\{p(x|\theta)|\theta \in \Theta\}$ be a parametric statistical model and $f(\theta)$ be a smooth prior density on a model manifold (M, g) endowed with the Fisher metric.*

The risk difference between $p_{\pi_{\mathrm{J}}}(y|x^{(N)})$, the Bayesian predictive density based on the Jeffreys prior, and $p_f(y|x^{(N)})$, the Bayesian predictive density based on f, is given by

$$(\textit{Risk function of } p_{\pi_{\mathrm{J}}}(y|x^{(N)})) - (\textit{Risk function of } p_f(y|x^{(N)})) = -\frac{2}{N^2}\left(\frac{\pi_{\mathrm{J}}}{f}\right)^{1/2} \Delta\left(\frac{f}{\pi_{\mathrm{J}}}\right)^{1/2} + \mathrm{o}(N^{-2}). \tag{13}$$

The Bayesian predictive density based on $f(\theta)$ asymptotically dominates the Bayesian predictive density based on the Jeffreys prior $\pi_{\mathrm{J}}(\theta)$ if and only if

$$\left(\frac{f(\theta)}{\pi_{\mathrm{J}}(\theta)}\right)^{1/2}$$

is a non-constant positive superharmonic function on (M, g). □

See Komaki (2006) for the proof of the theorem above and Komaki (1996) and Hartigan (1998) for related results.

First we evaluate the risk of the Bayesian predictive density based on the right invariant prior. Since

$$\Delta\frac{\pi_{\mathrm{R}}}{\pi_{\mathrm{J}}} = \Delta\sigma^{d-1} = \frac{1}{(d-1)\beta^2}\left\{\sigma^2\left(\frac{(d-1)\beta^2}{\alpha^2}\Delta_\mu + \frac{\partial^2}{\partial\sigma^2}\right) - (d-2)\sigma\frac{\partial}{\partial\sigma}\right\}\sigma^{d-1} = 0,$$

the ratio

$$\frac{\pi_{\mathrm{R}}}{\pi_{\mathrm{J}}} \propto \sigma^{d-1}$$

is a positive harmonic function on the model manifold. If $f(\theta)/\pi_{\mathrm{J}}(\theta)$ is a non-constant positive harmonic function, then $(f(\theta)/\pi_{\mathrm{J}}(\theta))^{1/2}$ is a non-constant positive superharmonic function. The risk difference is given by

$$\begin{aligned}&(\text{Risk function of } p_{\pi_{\mathrm{J}}}(y|x^{(N)})) - (\text{Risk function of } p_{\pi_{\mathrm{R}}}(y|x^{(N)}))\\ &= -\frac{2}{N^2}\left(\frac{\pi_{\mathrm{J}}}{\pi_{\mathrm{R}}}\right)^{1/2} \Delta\left(\frac{\pi_{\mathrm{R}}}{\pi_{\mathrm{J}}}\right)^{1/2} + \mathrm{o}(N^{-2}) = \frac{1}{N^2}\frac{d-1}{2\beta^2} + \mathrm{o}(N^{-2}). \qquad (14)\end{aligned}$$

Here (14) is a positive constant not depending on the parameter (μ, σ). Therefore, the Bayesian predictive density based on the right invariant prior asymptotically dominates that based on the Jeffreys prior. This result is reasonable because the Bayesian predictive density based on the right invariant prior is the best invariant predictive density.

In Bayesian coding theory, it is known that the Bayesian predictive density based on the Jeffreys prior is asymptotically minimax. In this setting, random variables $y_1, y_2, \ldots, y_m$ from a density $p(y|\theta)$ in the model $\{p(y|\theta)|\theta \in \Theta\}$ are predicted by using a Bayesian predictive density. No data are observed before constructing a predictive density. The performance of prediction is evaluated by the Kullback–Leibler divergence

$$\int p(y^{(m)}|\theta) \log \frac{p(y^{(m)}|\theta)}{\int p(y^{(m)}|\theta')\pi(\theta')\mathrm{d}\theta'} \mathrm{d}y^{(m)},$$

where $y^{(m)} = (y_1, y_2, \ldots, y_m)$. It is often assumed that the parameter space is a compact subset of the original natural parameter space. Then the Jeffreys prior is the minimax prior in the asymptotics as m goes to infinity (Clarke and Barron 1994).

In our setting, we observe data $x_1, x_2, \ldots, x_N$ from a probability density $p(x|\theta)$ in the model $\{p(x|\theta)|\theta \in \Theta\}$, and then construct a predictive density to predict $y = x_{N+1}$ from the same density $p(x|\theta)$. The Kullback–Leibler divergence from $p(y|\theta)$ to a predictive density is used to evaluate the predictive density. Then the Bayesian predictive density based on the right invariant prior is best invariant prior and the Jeffreys prior is not minimax in the asymptotics as N goes to ∞.

Next we evaluate the risk of the Bayesian predictive density based on the prior (12). Then, from (7) and (13) we have

$$\begin{aligned}&(\text{Risk function of } p_{\pi_J}(y|x^{(N)})) - (\text{Risk function of } p_{f_{c,\bar{u},\bar{v}}}(y|x^{(N)}))\\ &= -\frac{2}{N^2}\left(\frac{\pi_J}{f_{c,\bar{u},\bar{v}}}\right)^{1/2} \Delta\left(\frac{f_{c,\bar{u},\bar{v}}}{\pi_J}\right)^{1/2} + o(N^{-2})\\ &= \frac{1}{N^2\beta^2}\left\{\frac{\alpha-1}{2} + \frac{(d+1)(1-c^2)}{2(\cosh\rho + c)^2} + \frac{c}{\cosh\rho + c}\right\} \end{aligned} \tag{15}$$

Since (15) is greater than (14), the introduced priors asymptotically dominates the best invariant predictive density even when $d = 2$, corresponding to the univariate location-scale model. Liang and Barron (2004) showed that the best invariant predictive density for the location-scale model is minimax. However, it has been an open problem whether the best invariant (and minimax) predictive density is admissible or not. The preset result gives a negative answer to this problem in the asymptotic sense.

In the limit $\bar{v} \to \infty$, the prior density $f_{c,\bar{v}}$ converges to the right invariant density π_R except for the multiplicative constant not depending on (μ, σ). In the limit $\bar{v} \to 0$, $f_{c,\bar{v}}$ converges to

$$\frac{\sigma^{-d+1}}{\{\alpha^2\beta^{-2}(d-1)^{-1}|\mu/\sigma|^2 + 1\}^{d-1}} \tag{16}$$

except for the multiplicative constant not depending on (μ, σ).

The two different prior densities π_R and (16) are essentially the same function on the hyperbolic space because they are in symmetry with respect to a point on the Hyperbolic space. Therefore the asymptotic risk difference of the Bayesian predictive density based on (16) coincides with (14).

When $d = 2$, (16) is the Cauchy density as a function of μ. It may be interesting to note that Jeffreys (1963) recommends the Cauchy prior for normal mean in some problems.

In the present paper, we studied the multivariate location-scale model, which is a multivariate model with a specific variance–covariance structure. In prediction theory for multivariate models with a general variance–covariance structure, the geometry of the Wishart model manifold plays an essential role. Information geometrical approach to prediction theory for the 2×2 Wishart model is touched on in Komaki (2006), and detailed discussion will be given in another place.

Acknowledgements The author thanks the referees for their helpful comments.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, *62*, 547–554.
Aitchison, J., Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.
Amari, S. (1985). *Differential-geometrical methods in statistics*. Berlin Heidelberg New York: Springer.
Amari, S., Nagaoka, H. (2000). *Methods of information geometry*. Providence: American Mathematical Society.
Clarke, B. S., Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, *41*, 36–60.
Davis, E. B. (1989). *Heat kernels and spectral theory*. Cambridge: Cambridge University Press.
Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society*, A *144*, 285–307.
Geisser, S. (1993). *Predictive inference: an introduction.* New York: Chapman and Hall.
Hartigan, J. A. (1998). The maximum likelihood prior. *Annals of Statistics*, *26*, 2083–2103.
Helgason, S. (1984). *Groups and geometric analysis*. New York: Academic Press.
Jeffreys, H. (1961). *Theory of probability,* 3rd Edn. Oxford: Oxford University Press.
Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, *83*, 299–313.
Komaki, F. (2002). Bayesian predictive distribution with right invariant priors. *Calcutta Statistical Association Bulletin*, *52*, 171–179.
Komaki, F. (2006). Shrinkage priors for Bayesian prediction. *The Annals of Statistics*, *34*, 808–819.
Liang, F., Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection, *IEEE Transactions on Information Theory*, *50*, 2708–2726.
Neyman, J., Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.
Robert, C. P. (2001). *The Bayesian Choice*, 2nd Edn. Berlin Heidelberg New York: Springer.
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.