Ann Inst Stat Math (2009) 61:531–542
DOI 10.1007/s10463-007-0163-z

Improved prediction for a multivariate normal distribution with unknown mean and variance

Kengo Kato

Received: 13 February 2007 / Revised: 11 October 2007 / Published online: 10 January 2008

Abstract The prediction problem for a multivariate normal distribution is considered where both mean and variance are unknown. When the Kullback–Leibler loss is used, the Bayesian predictive density based on the right invariant prior, which turns out to be a density of a multivariate t-distribution, is the best invariant and minimax predictive density. In this paper, we introduce an improper shrinkage prior and show that the Bayesian predictive density against the shrinkage prior improves upon the best invariant predictive density when the dimension is greater than or equal to three.

Keywords Bayesian prediction · Kullback–Leibler divergence · Multivariate normal distribution · Multivariate t-distribution · Right invariant prior · Shrinkage prior · Star ordering

1 Introduction

Let $X_{(n)} = (X_1, \ldots, X_n)$ be independent random vectors from a d-dimensional multivariate normal distribution $N_d(\mu, \sigma^2 I_d)$ where $\mu \in \mathbb{R}^d$ and $\sigma > 0$ are unknown parameters, and Y be another independent random vector from the same distribution. We denote $p(x_{(n)}|\mu, \sigma)$ and $p(y|\mu, \sigma)$ for densities of $X_{(n)}$ and Y, respectively. We assume $n \geq 2$.

Based on the observation $X_{(n)} = x_{(n)}$, we consider the problem of constructing a predictive density $\hat{p}(y|x_{(n)})$ for Y. The Kullback–Leibler divergence

$$L\left\{(\mu, \sigma), \hat{p}(\cdot|x_{(n)})\right\} = \int p(y|\mu, \sigma) \log \frac{p(y|\mu, \sigma)}{\hat{p}(y|x_{(n)})} \mathrm{d}y$$

K. Kato (✉)
Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
e-mail: kato_ken@hkg.odn.ne.jp

is adopted as a loss function, and a predictive density $\hat{p}(y|x_{(n)})$ is evaluated by its expected loss or risk function

$$R\left\{(\mu,\sigma),\hat{p}\right\} = \int p(x_{(n)}|\mu,\sigma) L\left\{(\mu,\sigma),\hat{p}(\cdot|x_{(n)})\right\} \mathrm{d}x_{(n)}.$$

There are two major methods to obtain predictive densities. One is to construct the plug-in density $p(y|\hat{\mu},\hat{\sigma})$, where $\hat{\mu}$ and $\hat{\sigma}$ are estimates based on $x_{(n)}$. Another is to construct the Bayesian predictive density defined as

$$\hat{p}_{\pi}(y|x_{(n)}) = \frac{\iint p(y|\mu,\sigma)p(x_{(n)}|\mu,\sigma)\pi(\mu,\sigma)\mathrm{d}\mu\mathrm{d}\sigma}{\iint p(x_{(n)}|\mu,\sigma)\pi(\mu,\sigma)\mathrm{d}\mu\mathrm{d}\sigma},$$

with a prior $\pi(\mu,\sigma)$. It follows from Aitchison (1975) that for a proper π, $\hat{p}_{\pi}$ minimizes the Bayes risk.

For prediction problems in general, many studies have recommended the use of Bayesian predictive densities rather than plug-in densities (Geisser 1993; Komaki 1996). In the present problem, it can be shown from the arguments in Aitchison (1975) that the plug-in densities $p(y|\hat{\mu},\hat{\sigma})$, where $\hat{\mu} = n^{-1}\sum_{j=1}^{n} x_j$ and $\hat{\sigma}$ is the square root of the maximum likelihood estimate or the unbiased estimate of σ^2 based on $x_{(n)}$, are dominated by the Bayesian predictive density $\hat{p}_R$ defined below.

When a Bayesian procedure is used, the choice of a prior is an important problem. Non-informative priors such as Jeffreys priors are often used to construct Bayesian predictive densities. The Jeffreys prior coincides with the left invariant prior $\pi_L(\mu,\sigma) = 1/\sigma^{d+1}$ in the present setting (Robert 2001). However, as shown in Liang and Barron (2004), the best invariant and minimax predictive density is given by the Bayesian predictive density $\hat{p}_R$ based on the right invariant prior $\pi_R(\mu,\sigma) = 1/\sigma$. It will be explicitly verified in the next section that $\hat{p}_R$ dominates $\hat{p}_L$, which is the Bayesian predictive density based on π_L.

Although $\hat{p}_R$ would be considered as a good default procedure, it has not been addressed whether $\hat{p}_R$ is admissible. From analogous arguments in parameter estimation, it can be conjectured that $\hat{p}_R$ is inadmissible when $d \geq 3$.

For a d-dimensional multivariate normal distribution $N_d(\mu,\sigma^2 I_d)$ with unknown μ and known σ, Komaki (2001) showed that when $d \geq 3$, the Bayesian predictive density based on the improper shrinkage prior $\pi_S(\mu) = \|\mu\|^{-(d-2)}$ dominates the Bayesian predictive density $\hat{p}_U$ based on the uniform prior $\pi_U(\mu) = 1$, which is the best invariant predictive density with respect to the translation group. George et al. (2006) and Brown et al. (2007) have obtained several conditions for priors which yield admissible predictive densities dominating $\hat{p}_U$. Their results suggest fundamental similarities between the prediction problem under the Kullback–Leibler loss and the problem of estimating a multivariate normal mean under the quadratic loss.

It should be pointed out that when σ is unknown, the best invariant predictive density turns out to be a density of a multivariate t-distribution and hence does not belong to the normal model, which is a difference from the case where σ is known. It is thus a substantially new task to show the dominance over the best invariant predictive density when σ is unknown. Of course, from a practical point of view, it is a worthwhile challenge to derive an improved predictive density for a multivariate normal model where both mean and variance are unknown.

In the present paper, we introduce an improper shrinkage prior of the form

$$\pi_{LT}(\mu, \eta)\mathrm{d}\mu\mathrm{d}\eta \propto \|\mu\|^{-(d-2)}\sigma^{-1}\mathrm{d}\mu\mathrm{d}\sigma,$$

where $\eta = \sigma^{-2}$, which shrinks the mean vector toward the origin compared with π_R. We show that the Bayesian predictive density based on the introduced prior dominates $\hat{p}_R$ when $d \geq 3$. Hence $\hat{p}_R$ is shown to be inadmissible. This prior was originally considered in Lin and Tsai (1973) for estimation of a multivariate normal mean. It seems interesting that the shrinkage method still leads to an exactly superior predictive distribution when σ is unknown. The method considered here is applicable to the normal linear model.

The organization of this paper is as follows. In Sect. 2, we first summarize properties of the predictive densities based on the left and right invariant priors. The main theorem, Theorem 3, is stated in Sect. 2.2. The proof of this theorem is provided in Sect. 3. The proof uses a somewhat new technique, namely the star ordering of distribution functions. In Sect. 3.1, we briefly explain the star ordering and its related notion, the dispersive ordering, prior to the proof of Theorem 3.

Although only the one-step prediction is discussed in the present paper, our result holds when we consider to predict m random vectors $Y_{(m)} = (Y_1, \ldots, Y_m)$, where $Y_1, \ldots, Y_m$ are independently distributed as $N_d(\mu, \sigma^2 I_d)$.

2 Main results

2.1 Prediction with the left and right invariant priors

We first consider the left and right invariant priors, and briefly summarize their properties. The predictive density based on the right invariant prior $\pi_R(\mu, \sigma) = 1/\sigma$ is given by

$$\hat{p}_R(y|x_{(n)}) = \frac{\Gamma(nd/2)}{\pi^{\frac{d}{2}}(s_1^2)^{\frac{d}{2}}\Gamma\{(n-1)d/2\}}\left(\frac{n}{n+1}\right)^{\frac{d}{2}}\left\{1 + \frac{\|y-\bar{x}\|^2}{\left(1+\frac{1}{n}\right)s_1^2}\right\}^{-\frac{nd}{2}},$$

where $\bar{x} = n^{-1}\sum_{j=1}^n x_j$ and $s_1^2 = \sum_{j=1}^n \|x_j - \bar{x}\|^2$. Note that $\hat{p}_R$ is a density of a multivariate t-distribution with $(n-1)d$ degrees of freedom.

In this setting, a predictive density $\hat{p}(y|x_{(n)})$ is said to be invariant if $b^d\hat{p}\{b(y-a)|b(x_{(n)}-a)\} = \hat{p}(y|x_{(n)})$ for any $a \in \mathbb{R}^d$ and $b > 0$, where the notation $x_{(n)} - a$ denotes $x_1 - a, \ldots, x_n - a$. The next theorem is given in Liang and Barron (2004).

Theorem 1 (Liang and Barron 2004) *For $n \geq 2$, the Bayesian predictive density $\hat{p}_R$ is the best invariant and minimax predictive density under the Kullback–Leibler loss.*

The left invariant prior $\pi_L(\mu, \sigma) = 1/\sigma^{d+1}$ coincides with the Jeffreys prior. The predictive density based on π_L is given by

$$\hat{p}_L(y|x) = \frac{\Gamma\{(n+1)d/2\}}{\pi^{\frac{d}{2}}(s_1^2)^{\frac{d}{2}}\Gamma(nd/2)}\left(\frac{n}{n+1}\right)^{\frac{d}{2}}\left\{1+\frac{\|y-\bar{x}\|^2}{\left(1+\frac{1}{n}\right)s_1^2}\right\}^{-\frac{(n+1)d}{2}},$$

which in turn is the density of a multivariate t-distribution with nd degrees of freedom.

Although Jeffreys priors are widely used in Bayesian prediction, Theorem 1 implies that $\hat{p}_L$ is not as good as $\hat{p}_R$ since $\hat{p}_L$ is invariant. In fact, the dominance of $\hat{p}_R$ over $\hat{p}_L$ is explicitly shown by a direct calculation as follows:

Let $s_2^2 = \{n/(n+1)\}\|y-\bar{x}\|^2$. Then,

$$\log\frac{\hat{p}_R(y|x_{(n)})}{\hat{p}_L(y|x_{(n)})} = \log\frac{B(nd/2, d/2)}{B\{(n-1)d/2, d/2\}} - \log\left(\frac{s_1^2}{s_1^2+s_2^2}\right)^{\frac{d}{2}}.$$

Since $s_1^2/(s_1^2+s_2^2)$ is distributed as $Beta\{(n-1)d/2, d/2\}$, Jensen's inequality yields that the risk difference $R\left\{(\mu,\sigma), \hat{p}_L\right\} - R\left\{(\mu,\sigma), \hat{p}_R\right\} = E_{\mu,\sigma}\{\log(\hat{p}_R/\hat{p}_L)\}$ is positive.

We summarize this fact as a corollary.

Corollary 1 *$\hat{p}_L$ is dominated by $\hat{p}_R$ under the Kullback–Leibler loss.*

2.2 Improved prediction

We introduce an improper shrinkage prior $\pi_{LT}(\mu,\eta)$ defined as

$$\mu|(\eta,\lambda) \sim N_d\left(0, \frac{1-\lambda}{\lambda}\eta^{-1}I_d\right),$$
$$(\eta,\lambda) \sim \eta^{-2}\lambda^{-2}, \quad \eta > 0,\ 0 < \lambda < 1,$$

where $\eta = \sigma^{-2}$. Note that $\pi_{LT}(\mu,\eta)\mathrm{d}\mu\mathrm{d}\eta \propto \|\mu\|^{-(d-2)}\sigma^{-1}\mathrm{d}\mu\mathrm{d}\sigma$. This prior was originally considered in Lin and Tsai (1973) for estimation of a multivariate normal mean.

Theorem 2 *The Bayesian predictive density based on π_{LT} $(d \geq 3)$ is given by*

$$\hat{p}_{LT}(y|x_{(n)}) = \frac{\Gamma\{(n+1)d/2-1\}}{\pi^{\frac{d}{2}}(s_1^2)^{\frac{d}{2}}\Gamma(nd/2-1)}\left(\frac{n}{n+1}\right)$$
$$\times\frac{\int_0^1 t^{\frac{d}{2}-2}\left\{1+\frac{n}{n+1}\frac{\|y-\bar{x}\|^2}{s_1^2}+\frac{(n+1)\|\frac{n\bar{x}+y}{n+1}\|^2}{s_1^2}t\right\}^{-\frac{(n+1)d}{2}+1}\mathrm{d}t}{\int_0^1 t^{\frac{d}{2}-2}\left\{1+\frac{n\|\bar{x}\|^2}{s_1^2}t\right\}^{-\frac{nd}{2}+1}\mathrm{d}t}. \tag{1}$$

Proof We write $p(x_{(n)}|\mu,\eta)$ and $p(y|\mu,\eta)$ in place of $p(x_{(n)}|\mu,\sigma)$ and $p(y|\mu,\sigma)$, respectively. The Bayesian predictive density based on π_{LT} is given by

$$\hat{p}_{LT}(y|x_{(n)}) = \frac{\iint p(y|\mu,\eta)p(x_{(n)}|\mu,\eta)\pi_{LT}(\mu,\eta)\mathrm{d}\mu\mathrm{d}\eta}{\iint p(x_{(n)}|\mu,\eta)\pi_{LT}(\mu,\eta)\mathrm{d}\mu\mathrm{d}\eta}, \tag{2}$$

and we calculate the denominator and the numerator of (2).

First, the denominator of (2) is

$$\begin{aligned}&\iint p(x_{(n)}|\mu,\eta)\pi_{LT}(\mu,\eta)\mathrm{d}\mu\mathrm{d}\eta\\ &\quad= \frac{1}{(2\pi)^{\frac{(n+1)d}{2}}}\iiint \eta^{\frac{(n+1)d}{2}-2}\lambda^{\frac{d}{2}-2}(1-\lambda)^{-\frac{d}{2}}\mathrm{e}^{-\frac{\eta}{2}\left(s_1^2+n\|\bar{x}-\mu\|^2+\frac{\lambda}{1-\lambda}\|\mu\|^2\right)}\mathrm{d}\lambda\mathrm{d}\mu\mathrm{d}\eta.\end{aligned} \tag{3}$$

Making the transformation $\lambda/(1-\lambda) = nt/(1-t)$ with $\mathrm{d}\lambda = n\mathrm{d}t/\{1+(n-1)t\}^2$ and using the relation

$$\|\bar{x}-\mu\|^2 + \frac{t}{1-t}\|\mu\|^2 = \frac{1}{1-t}\|\mu-(1-t)\bar{x}\|^2 + t\|\bar{x}\|^2,$$

we can rewrite the right-hand side of (3) as

$$\begin{aligned}&\frac{n^{\frac{d}{2}-1}}{(2\pi)^{\frac{(n+1)d}{2}}}\int_0^1\int_0^\infty\int_{\mathbb{R}^d}\eta^{\frac{(n+1)d}{2}-2}t^{\frac{d}{2}-2}(1-t)^{-\frac{d}{2}}\mathrm{e}^{-\frac{\eta}{2}(s_1^2+n\|\bar{x}\|^2t)}\\ &\qquad\times\mathrm{e}^{-\frac{n\eta}{2(1-t)}\|\mu-(1-t)\bar{x}\|^2}\mathrm{d}\mu\mathrm{d}\eta\mathrm{d}t\\ &\quad=\frac{1}{n(2\pi)^{\frac{nd}{2}}}\int_0^1\int_0^\infty\eta^{\frac{nd}{2}-2}t^{\frac{d}{2}-2}\mathrm{e}^{-\frac{\eta}{2}(s_1^2+n\|\bar{x}\|^2t)}\mathrm{d}\eta\mathrm{d}t\\ &\quad=\frac{\Gamma(nd/2-1)}{2n\pi^{\frac{nd}{2}}}\int_0^1 t^{\frac{d}{2}-2}(s_1^2+n\|\bar{x}\|^2t)^{-\frac{nd}{2}+1}\mathrm{d}t.\end{aligned} \tag{4}$$

Next, note that

$$p(y|\mu,\eta)p(x_{(n)}|\mu,\eta) = \left(\frac{\eta}{2\pi}\right)^{\frac{(n+1)d}{2}}\mathrm{e}^{-\frac{\eta}{2}\left\{s_1^2+\frac{n}{n+1}\|y-\bar{x}\|^2+(n+1)\left\|\mu-\frac{n\bar{x}+y}{n+1}\right\|^2\right\}}.$$

Then, the numerator of (2) is similarly calculated as follows:

$$\begin{aligned}&\iint p(y|\mu,\eta)p(x_{(n)}|\mu,\eta)\pi_{LT}(\mu,\eta)\mathrm{d}\mu\mathrm{d}\eta = \frac{\Gamma\{(n+1)d/2-1)\}}{2(n+1)\pi^{\frac{(n+1)d}{2}}}\\ &\quad\times\int_0^1 t^{\frac{d}{2}-2}\left\{s_1^2+\frac{n}{n+1}\|y-\bar{x}\|^2+(n+1)\left\|\frac{n\bar{x}+y}{n+1}\right\|^2t\right\}^{-\frac{(n+1)d}{2}+1}\mathrm{d}t.\end{aligned} \tag{5}$$

Combining (4) and (5) gives the expression (1). □

Now, we state our main theorem of this paper. The proof of the theorem is given in the next section.

Theorem 3 *For $n \geq 2$ and $d \geq 3$, the inequality*

$$R\left\{(\mu,\sigma),\hat{p}_R\right\} - R\left\{(\mu,\sigma),\hat{p}_{LT}\right\} > 0$$

holds for all $\mu \in \mathbb{R}^d$ and $\sigma > 0$, i.e., $\hat{p}_R$ is dominated by $\hat{p}_{LT}$ and hence inadmissible.

2.3 Simulation studies

It is of interest to investigate the behaviors of the risk differences between $\hat{p}_{LT}$ and $\hat{p}_R$ for several values of d and n. The risk differences $R\left\{(\mu,\sigma),\hat{p}_R\right\} - R\left\{(\mu,\sigma),\hat{p}_{LT}\right\}$ for $d = 3, 5, 7$ and $n = 5, 10$ are given in Fig. 1a and b.

It can be verified from these figures that the risk gain of $\hat{p}_{LT}$ is larger when d is big or n is small. The proposed predictive density $\hat{p}_{LT}$ is thus especially recommended in these situations.

3 Proof of Theorem 3

3.1 Star and dispersive orderings

In this subsection, we introduce some notions of stochastic orderings, known as star ordering and dispersive ordering, which will be used to prove our main result. For a distribution function F on $\mathbb{R}$, F^{-1} denotes its left continuous inverse function.

Definition 1 Let F and G be distribution functions on $\mathbb{R}$. Then,

- F is star-ordered with respect to G (written as $F \leq_{\star} G$) if $G^{-1}(p)/F^{-1}(p)$ is nondecreasing in $p \in (0,1)$,
- F is less dispersed than G (written as $F \leq_{\mathrm{disp}} G$) if $F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha)$ for all $0 < \alpha \leq \beta < 1$.

When U and V are random variables with distribution functions F and G respectively, we also write $U \leq_{\star} V$ if $F \leq_{\star} G$, and $U \leq_{\mathrm{disp}} V$ if $F \leq_{\mathrm{disp}} G$. The next lemma states a correspondence between the star ordering and dispersive ordering for positive random variables.

Lemma 1 *Suppose U and V are random variables positive in probability* 1. *If their distribution functions are continuous with their supports being intervals, then,*

$$U \leq_{\star} V \Leftrightarrow -\log U \leq_{\mathrm{disp}} -\log V. \tag{6}$$

Proof Define $W = -\log U$. Let F_U and F_W be the distribution functions of U and W, respectively. Then since $F_W(w) = P(-\log U \leq w) = P(U \geq \mathrm{e}^{-w}) =$

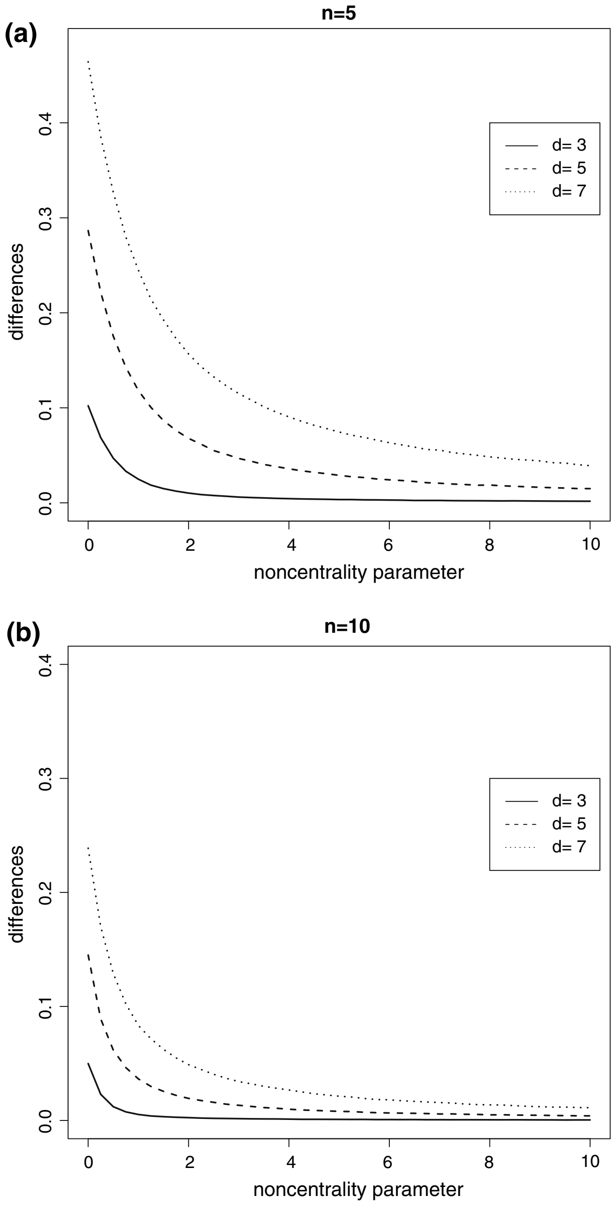

Fig. 1 Risk differences $R\left\{(\mu,\sigma),\hat{p}_R\right\}-R\left\{(\mu,\sigma),\hat{p}_{LT}\right\}$ for $d=3,5,7$ and $n=5,10$. 'noncentrality parameter' denotes $\|\mu\|^2/\sigma^2$

$1-F_U(\mathrm{e}^{-w})$, we have $F_W^{-1}(p)=-\log F_U^{-1}(1-p)$ for $p\in(0,1)$. Also, define $Z=-\log V$ and let F_V nd F_Z be the distribution functions of V and Z, respectively. Again, it follows $F_Z^{-1}(p)=-\log F_V^{-1}(p)$ for $p\in(0,1)$.

Now, by the definition of the dispersive ordering, $W \leq_{\text{disp}} Z$ is equivalent to

$$\begin{aligned}&-\log F_U^{-1}(1-\beta)+\log F_U^{-1}(1-\alpha)\\ &\quad\leq -\log F_V^{-1}(1-\beta)+\log F_V^{-1}(1-\alpha)\ \text{ for } 0<\alpha\leq\beta<1,\end{aligned}$$

which is equivalent to

$$\log\frac{F_V^{-1}(1-\beta)}{F_U^{-1}(1-\beta)}\leq\log\frac{F_V^{-1}(1-\alpha)}{F_U^{-1}(1-\alpha)}\quad\text{for } 0<\alpha\leq\beta<1. \tag{7}$$

Since the condition (7) means that $F_V^{-1}(p)/F_U^{-1}(p)$ is nondecreasing in $p\in(0,1)$, we obtain the equivalence (6). □

For every function f with domain $I\subset\mathbb{R}$ and for every $c\in\mathbb{R}$, we define the function f_c by $f_c(u)=f(u-c)$, $u\in\{v+c;\, v\in I\}$. The number of sign changes of f in I is defined by

$$S^-(f)=\sup S^-\{f(u_1),\ldots,f(u_m)\} \tag{8}$$

where $S^-(a_1,\ldots,a_m)$ is the number of sign changes of the indicated sequence, zero terms being discarded, and the supremum in (8) is extended over all sets $u_1<\cdots<u_m$ such that $u_j\in I$ and $m<\infty$.

The next theorem given in Shaked (1982) provides a useful tool for proving the dispersive ordering between two distribution functions.

Theorem 4 (Shaked 1982) *Let F and G be two absolutely continuous distribution functions with support $[0,\infty)$ and let f and g be the corresponding densities. If*

$$S^-(f_c-g)\leq 2 \tag{9}$$

for every $c>0$, with the sign sequence being $-,+,-$ in case of equality, and if $F(u)\geq G(u)$ for all $u>0$, then $F\leq_{\text{disp}} G$.

The next lemma, which will be used in the proof of Theorem 3, is a slight extension of Lemma 1 of Jeon et al. (2006).

Lemma 2 *Let $U\sim Beta(\alpha,\gamma_1)$ and $V\sim Beta(\alpha,\gamma_2)$ with $\alpha>0$ and $1<\gamma_1<\gamma_2$. Then, $U\leq_\star V$.*

Proof From Lemma 1, we need to show that

$$-\log U\leq_{\text{disp}}-\log V. \tag{10}$$

The densities of $-\log U$ and $-\log V$ are

$$f(u)=\frac{1}{B(\alpha,\gamma_1)}\mathrm{e}^{-\alpha u}(1-\mathrm{e}^{-u})^{\gamma_1-1},\quad g(v)=\frac{1}{B(\alpha,\gamma_2)}\mathrm{e}^{-\alpha v}(1-\mathrm{e}^{-v})^{\gamma_2-1},$$

for $u>0$ and $v>0$, respectively.

First, since

$$\frac{g(u)}{f(u)} \propto (1-\mathrm{e}^{-u})^{\gamma_2-\gamma_1}$$

is nondecreasing in $u>0$, $F(u) \geq G(u)$ holds for all $u>0$.

Let $c>0$. For $u>c$, the sign of $f_c(u)-g(u)$ is the same as the sign of

$$\log f_c(u) - \log g(u) = A + \alpha c + (\gamma_1 - 1)\log(1-\mathrm{e}^c\mathrm{e}^{-u}) - (\gamma_2-1)\log(1-\mathrm{e}^{-u}),$$

where $A = \log\{B(\alpha,\gamma_2)/B(\alpha,\gamma_1)\}$. Define

$$h(w) = A + \alpha c + (\gamma_1-1)\log(1-\mathrm{e}^c w) - (\gamma_2-1)\log(1-w)$$

for $0<w<\mathrm{e}^{-c}$ and differentiate h to obtain

$$h'(w) = -(\gamma_1-1)\frac{\mathrm{e}^c}{1-\mathrm{e}^c w} + (\gamma_2-1)\frac{1}{1-w}.$$

It is seen that the equation $h'(w)=0$ has at most one root in $0<w<\mathrm{e}^{-c}$ and $h(w)\to-\infty$ as $w\to\mathrm{e}^{-c}$ since $\gamma_1>1$. Then it is seen that the conditions of Theorem 4 are satisfied. Therefore the ordering (10) is established. □

3.2 Proof of Theorem 3

We here provide the proof of Theorem 3. For notational convenience, we write $\bar{x}_n$ as $\bar{x}$ and $\bar{x}_{n+1}$ as $(n\bar{x}+y)/(n+1)$. Then,

$$\begin{aligned}
&\log\frac{\hat{p}_{LT}(y|x_{(n)})}{\hat{p}_R(y|x_{(n)})} \\
&= \left(\frac{d}{2}-1\right)\log\left(\frac{n+1}{n}\right) - \left(\frac{d}{2}-1\right)\log\left(1+\frac{s_2^2}{s_1^2}\right) \\
&\quad + \log\frac{1}{B(d/2-1,nd/2)}\int_0^1 t^{\frac{d}{2}-2}\left\{1+\frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}t\right\}^{-\frac{(n+1)d}{2}+1}\mathrm{d}t \\
&\quad - \log\frac{1}{B\{d/2-1,(n-1)d/2\}}\int_0^1 t^{\frac{d}{2}-2}\left\{1+\frac{n\|\bar{x}_n\|^2}{s_1^2}t\right\}^{-\frac{nd}{2}+1}\mathrm{d}t. \qquad (11)
\end{aligned}$$

Applying the change of variables $s = \frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}t$ to the second integral in the right-hand side of (11), we obtain

$$\begin{aligned}\int_0^1 & t^{\frac{d}{2}-2}\left\{1+\frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}t\right\}^{-\frac{(n+1)d}{2}+1}\mathrm{d}t \\ &= \left\{\frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}\right\}^{-\left(\frac{d}{2}-1\right)}\int_0^{\frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}} s^{\frac{d}{2}-2}(1+s)^{-\frac{(n+1)d}{2}+1}\mathrm{d}s. \qquad (12)\end{aligned}$$

Making the transformation $s = u/(1-u)$ with $\mathrm{d}s = (1-u)^{-2}\mathrm{d}u$ to the integral in the right-hand side of (12), we have

$$\begin{aligned}\int_0^{\frac{(n+1)\|\bar{x}_{n+1}\|^2}{s_1^2+s_2^2}} & s^{\frac{d}{2}-2}(1+s)^{-\frac{(n+1)d}{2}+1}\mathrm{d}s \\ &= \int_0^{\frac{(n+1)\|\bar{x}_{n+1}\|^2}{(n+1)\|\bar{x}_{n+1}\|^2+s_1^2+s_2^2}} u^{\frac{d}{2}-2}(1-u)^{\frac{nd}{2}-1}\mathrm{d}u.\end{aligned}$$

Again, applying the changes of variables to the third integral in the right-hand side of (11) in the similar way, we finally obtain the expression

$$\begin{aligned}\log\frac{\hat{p}_{LT}(y|x_{(n)})}{\hat{p}_R(y|x_{(n)})} &= \left(\frac{d}{2}-1\right)\left\{\log(\|\bar{x}_n\|^2)-\log(\|\bar{x}_{n+1}\|^2)\right\} \\ &\quad + \log\frac{1}{B(d/2-1,nd/2)}\int_0^{\frac{(n+1)\|\bar{x}_{n+1}\|^2}{(n+1)\|\bar{x}_{n+1}\|^2+s_1^2+s_2^2}} t^{\frac{d}{2}-2}(1-t)^{\frac{nd}{2}-1}\mathrm{d}t \\ &\quad - \log\frac{1}{B\{d/2-1,(n-1)d/2\}}\int_0^{\frac{n\|\bar{x}_n\|^2}{n\|\bar{x}_n\|^2+s_1^2}} t^{\frac{d}{2}-2}(1-t)^{\frac{(n-1)d}{2}-1}\mathrm{d}t.\end{aligned}$$

Now, define

$$F_n(u) = \frac{1}{B(d/2-1,nd/2)}\int_0^u t^{\frac{d}{2}-2}(1-t)^{\frac{nd}{2}-1}\mathrm{d}t,$$

and F_{n-1} in the same manner. Then, the risk difference is expressed as

$$\begin{aligned}& R\left\{(\mu,\sigma),\hat{p}_R\right\} - R\left\{(\mu,\sigma),\hat{p}_{LT}\right\} \\ &\quad = E_{\mu,\sigma}\left\{\log(\hat{p}_{LT}/\hat{p}_R)\right\} \\ &\quad = \left(\frac{d}{2}-1\right)\left[E_{\mu,\sigma}\left\{\log(\|\bar{X}_n\|^2)\right\} - E_{\mu,\sigma}\left\{\log(\|\bar{X}_{n+1}\|^2)\right\}\right] \\ &\qquad + E\left\{\log F_n\left(\frac{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2}}{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2}+\chi^2_{nd}}\right)\right\} \\ &\qquad - E\left\{\log F_{n-1}\left(\frac{\chi^2_{d,n\|\mu\|^2/\sigma^2}}{\chi^2_{d,n\|\mu\|^2/\sigma^2}+\chi^2_{(n-1)d}}\right)\right\},\end{aligned}$$

where $\chi^2_{l,\xi}$ is a random variable having the noncentral χ^2-distribution with l degrees of freedom and noncentrality parameter ξ, χ^2_m is a random variable having the χ^2-distribution with m degrees of freedom independent of $\chi^2_{l,\xi}$.

From Lemma 1 of Komaki (2001), it follows that

$$E_{\mu,\sigma}\left\{\log(\|\bar{X}_n\|^2)\right\} - E_{\mu,\sigma}\left\{\log(\|\bar{X}_{n+1}\|^2)\right\} > 0$$

for all $\mu \in \mathbb{R}^d$ and $\sigma > 0$. Hence it is enough to show

$$\begin{aligned} &E\left\{\log F_n\left(\frac{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2}}{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2} + \chi^2_{nd}}\right)\right\} \\ &\qquad -E\left\{\log F_{n-1}\left(\frac{\chi^2_{d,n\|\mu\|^2/\sigma^2}}{\chi^2_{d,n\|\mu\|^2/\sigma^2} + \chi^2_{(n-1)d}}\right)\right\} \geq 0. \end{aligned} \tag{13}$$

Since $F_n(u)$ is a nondecreasing function, it is seen that

$$E\left\{\log F_n\left(\frac{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2}}{\chi^2_{d,(n+1)\|\mu\|^2/\sigma^2} + \chi^2_{nd}}\right)\right\} \geq E\left\{\log F_n\left(\frac{\chi^2_{d,n\|\mu\|^2/\sigma^2}}{\chi^2_{d,n\|\mu\|^2/\sigma^2} + \chi^2_{nd}}\right)\right\},$$

which implies the inequality (13) holds if

$$\begin{aligned} &E\left\{\log F_n\left(\frac{\chi^2_{d,n\|\mu\|^2/\sigma^2}}{\chi^2_{d,n\|\mu\|^2/\sigma^2} + \chi^2_{nd}}\right)\right\} \\ &\qquad -E\left\{\log F_{n-1}\left(\frac{\chi^2_{d,n\|\mu\|^2/\sigma^2}}{\chi^2_{d,n\|\mu\|^2/\sigma^2} + \chi^2_{(n-1)d}}\right)\right\} \geq 0. \end{aligned}$$

Since this difference can be written as

$$\begin{aligned} &\sum_{j=0}^{\infty} \mathrm{e}^{-\tau}\frac{\tau^j}{j!}\Bigg[\frac{1}{B(d/2+j, nd/2)}\int_0^1 \{\log F_n(u)\}\, u^{\frac{d}{2}+j-1}(1-u)^{\frac{nd}{2}-1}\mathrm{d}u \\ &\qquad -\frac{1}{B\{d/2+j, (n-1)d/2\}}\int_0^1 \{\log F_{n-1}(u)\}\, u^{\frac{d}{2}+j-1}(1-u)^{\frac{(n-1)d}{2}-1}\mathrm{d}u\Bigg], \end{aligned}$$

where $\tau = n\|\mu\|^2/2\sigma^2$, it suffices to show that

$$\begin{aligned} &\frac{1}{B(d/2+j, nd/2)}\int_0^1 \{\log F_n(u)\}\, u^{\frac{d}{2}+j-1}(1-u)^{\frac{nd}{2}-1}\mathrm{d}u \\ &\quad -\frac{1}{B\{d/2+j, (n-1)d/2\}}\int_0^1 \{\log F_{n-1}(u)\}\, u^{\frac{d}{2}+j-1}(1-u)^{\frac{(n-1)d}{2}-1}\mathrm{d}u \geq 0 \end{aligned} \tag{14}$$

for each j.

Making the transformation $p = F_n(u)$ with

$$\mathrm{d}u = B(d/2-1, nd/2)\left\{F_n^{-1}(p)\right\}^{-\frac{d}{2}+2}\left\{1-F_n^{-1}(p)\right\}^{-\frac{nd}{2}+1}\mathrm{d}p,$$

we rewrite the first term of the left-hand side of (14) as

$$\frac{B(d/2-1, nd/2)}{B(d/2+j, nd/2)}\int_0^1 (\log p)\left\{F_n^{-1}(p)\right\}^{j+1}\mathrm{d}p.$$

Similarly, we can see that the second term of the left side of (14) is expressed as

$$\frac{B\{d/2-1, (n-1)d/2\}}{B\{d/2+j, (n-1)d/2\}}\int_0^1 (\log p)\left\{F_{n-1}^{-1}(p)\right\}^{j+1}\mathrm{d}p.$$

Note that both $\frac{B(d/2-1,nd/2)}{B(d/2+j,nd/2)}\left\{F_n^{-1}(p)\right\}^{j+1}$ and $\frac{B\{d/2-1,(n-1)d/2\}}{B\{d/2+j,(n-1)d/2\}}\left\{F_{n-1}^{-1}(p)\right\}^{j+1}$ are probability density functions on (0, 1). From Lemma 2, $F_n^{-1}(p)/F_{n-1}^{-1}(p)$ is nondecreasing in $p \in (0, 1)$. Since $p \mapsto \log p$ is nondecreasing, we obtain the desired inequality. Therefore, the proof of Theorem 3 is completed. □

Acknowledgments The author would like to thank Professor Tatsuya Kubokawa for his encouragement and helpful suggestions. He also would like to thank the anonymous referee for significant suggestions for improvement on presentation.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, *62*, 545–554.

Brown, L. D., George, E. I., Xu, X. (2007). Admissible predictive density estimation. *Annals of Statistics*, to appear.

Geisser, S. (1993). *Predictive inference: an introduction*. New York: Chapman and Hall.

George, E. I., Liang, F., Xu, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Annals of Statistics*, *34*, 78–91.

Jeon, J., Kochar, S., Park, C. G. (2006). Dispersive ordering-some applications and examples. *Statistical Papers*, *47*, 227–247.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, *83*, 299–313.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, *88*, 859–864.

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.

Lin, P. E., Tsai, H. L. (1973). Generalized Bayes minimax estimations of the multivariate normal mean with unknown covariance matrix. *Annals of Statistics*, *1*, 142–145.

Robert, C. P. (2001). *The Bayesian choice* (2nd ed.). New York: Springer.

Shaked, M. (1982). Dispersive ordering of distribution. *Journal of Applied Probability*, *19*, 310–320.