IMS Collections
**Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman**
Vol. 8 (2012) 42–56

DOI: 10.1214/11-IMSCOLL803

# Bayesian predictive densities for linear regression models under $\alpha$-divergence loss: Some results and open problems

**Yuzo Maruyama**[1] **and William, E. Strawderman**[2]

*University of Tokyo and Rutgers University*

**Abstract:** This paper considers estimation of the predictive density for a normal linear model with unknown variance under $\alpha$-divergence loss for $-1 \leq \alpha \leq 1$. We first give a general canonical form for the problem, and then give general expressions for the generalized Bayes solution under the above loss for each $\alpha$. For a particular class of hierarchical generalized priors studied in Maruyama and Strawderman (2005, 2006) for the problems of estimating the mean vector and the variance respectively, we give the generalized Bayes predictive density. Additionally, we show that, for a subclass of these priors, the resulting estimator dominates the generalized Bayes estimator with respect to the right invariant prior, i.e., the best (fully) equivariant minimax estimator when $\alpha = 1$.

## Contents



## 1. Introduction

We begin with the standard normal linear regression model setup

$$y \sim N_n(X\beta, \sigma^2 I_n), \tag{1.1}$$

[1]Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi Chiba, 277-8568, Japan, e-mail: maruyama@csis.u-tokyo.ac.jp
[2]Department of Statistics and Biostatistics, Rutgers University, 501 Hill Center, Busch Campus, 110 Frelinghuysen Road Piscataway, NJ 08854-8019, USA, e-mail: straw@stat.rutgers.edu
*AMS 2000 subject classifications:* Primary 62C20, 62J07; secondary 62F15
*Keywords and phrases:* shrinkage prior, Bayesian predictive density, alpha-divergence, Stein effect

where $y$ is an $n \times 1$ vector of observations, $X$ is an $n \times k$ matrix of $k$ potential predictors where $n > k$ and rank $X = k$, and $\beta$ is a $k \times 1$ vector of unknown regression coefficients, and $\sigma^2$ is unknown variance. Based on observing $y$, we consider the problem of giving the predictive density $p(\tilde{y}|\beta, \sigma^2)$ of a future $m \times 1$ vector $\tilde{y}$ where

$$\tilde{y} \sim N_m(\tilde{X}\beta, \sigma^2 I_m). \tag{1.2}$$

Here $\tilde{X}$ is a fixed $m \times k$ design matrix of the same $k$ predictors in $X$, and the rank of $\tilde{X}$ is assumed to be $\min(m, k)$. We also assume that $y$ and $\tilde{y}$ are conditionally independent given $\beta$ and $\sigma^2$. Note that in most earlier papers on such prediction problems, $\sigma^2$ is assumed known, mainly because this typically makes the problem less difficult. However, the assumption of unknown variance is more realistic, and we treat this more difficult case in this paper. In the following we denote by $\psi$ all the unknown parameters $\{\beta, \sigma^2\}$.

For each value of $y$, a predictive estimate $\hat{p}(\tilde{y}; y)$ of $p(\tilde{y}|\psi)$ is often evaluated by the Kullback-Leibler (KL) divergence

$$D_{KL}\{\hat{p}(\tilde{y}; y), p(\tilde{y}|\psi)\} = \int p(\tilde{y}|\psi) \log \frac{p(\tilde{y}|\psi)}{\hat{p}(\tilde{y}; y)} d\tilde{y}, \tag{1.3}$$

which is called the KL divergence loss from $p(\tilde{y}|\psi)$ to $\hat{p}(\tilde{y}; y)$. The overall quality of the procedure $\hat{p}(\tilde{y}; y)$ for each $\psi$ is then conveniently summarized by the KL risk

$$R_{KL}(\hat{p}(\tilde{y}; y), \psi) = \int D_{KL}\{\hat{p}(\tilde{y}; y), p(\tilde{y}|\psi)\} p(y|\psi) dy, \tag{1.4}$$

where $p(y|\psi)$ is the density of $y$ in (1.1). [1] showed that the Bayesian solution with respect to the prior $\pi(\psi)$ under $D_{KL}$ loss given by (1.3) is what is called the Bayesian predictive density

$$\hat{p}_\pi(\tilde{y}; y) = \frac{\int p(\tilde{y}|\psi) p(y|\psi) \pi(\psi) d\psi}{\int p(y|\psi)\pi(\psi) d\psi} = \int p(\tilde{y}|\psi)\pi(\psi|y) d\psi, \tag{1.5}$$

where

$$\pi(\psi|y) = \frac{p(y|\psi)\pi(\psi)}{\int p(y|\psi)\pi(\psi) d\psi}.$$

For the prediction problems in general, many studies suggest the use of the Bayesian predictive density rather than plug-in densities of the form $p(\tilde{y}|\hat{\psi}(y))$, where $\hat{\psi}$ is an estimated value of $\psi$. In our setup of the problem, [12] showed that the Bayesian predictive density with respect to the right invariant prior is best equivariant and minimax. Although the Bayesian predictive density with respect to the right invariant prior is a good default procedure, it has been shown to be inadmissible in some cases. Specifically, when $\sigma^2$ is assumed to be known and

$$\begin{aligned} &m \geq k \geq 3, \ n = mN, \\ &X = 1_N \otimes \tilde{X} = (\tilde{X}', \ldots, \tilde{X}')', \end{aligned} \tag{AS1}$$

where $N$ is an positive integer, $1_N$ is an $N \times 1$ vector each component of which is one, and $\otimes$ is the Kronecker product, [11] showed that the shrinkage Bayesian predictive density with respect to the harmonic prior

$$\pi_{S,0}(\psi) = \pi(\beta) = \{\beta' \tilde{X}' \tilde{X} \beta\}^{1-k/2} \tag{1.6}$$

dominates the best invariant Bayesian predictive density with respect to

$$\pi_{I,0}(\psi) = \pi(\beta) = 1. \tag{1.7}$$

[9] extended [11] result to general shrinkage priors including [17] prior. As pointed out in the above, we will assume that the variance $\sigma^2$ is unknown in this paper. The first decision-theoretic result in the unknown variance case was derived by [10]. He showed that, under the same assumption of [11] given by (AS1), the Bayesian predictive density with respect to the shrinkage prior

$$\pi_{S,1}(\psi) = \pi(\beta, \sigma^2) = \{\beta' \tilde{X}' \tilde{X} \beta\}^{1-k/2} \{\sigma^2\}^{-2} \tag{1.8}$$

dominates the best invariant predictive density which is the Bayesian predictive density with respect to the right invariant prior

$$\pi_{I,1}(\psi) = \pi(\beta, \sigma^2) = \{\sigma^2\}^{-1}. \tag{1.9}$$

From a more general viewpoint, the KL-loss given by (1.3) is in the class of $\alpha$-divergence introduced by [6] and defined by

$$D_\alpha\{\hat{p}(\tilde{y}; y), p(\tilde{y}|\psi)\} = \int f_\alpha\left(\frac{\hat{p}(\tilde{y}; y)}{p(\tilde{y}|\psi)}\right) p(\tilde{y}|\psi) d\tilde{y}, \tag{1.10}$$

where

$$f_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2}(1 - z^{(1+\alpha)/2}) & |\alpha| < 1 \\ z \log z & \alpha = 1 \\ -\log z & \alpha = -1. \end{cases}$$

Clearly the KL-loss given by (1.3) corresponds to $D_{-1}$. [5] showed that a generalized Bayesian predictive density under $D_\alpha$ loss is

$$\hat{p}_{\pi,\alpha}(\tilde{y}; y) \propto \begin{cases} \left[\int p^{(1-\alpha)/2}(\tilde{y}|\psi)\pi(\psi|y)d\psi\right]^{2/(1-\alpha)} & \alpha \neq 1 \\ \exp\{\int \log p(\tilde{y}|\psi)\pi(\psi|y)d\psi\} & \alpha = 1. \end{cases} \tag{1.11}$$

Hence the Bayesian predictive density of the form (1.5) may not perform well under $\alpha$-divergence with $\alpha \neq -1$. As [3] pointed out in the estimation problem, decision-theoretic properties often seem to depend on the general structure of the problem (the general type of problem (location, scale), and the dimension of the parameter space) and on the prior in a Bayesian-setup, but not on the loss function. In fact, we will show, under (AS1) and $D_1$ loss, the predictive density with respect to the same shrinkage prior given by (1.8) improves on the best invariant predictive density with respect to (1.9) (See Section 4). From this viewpoint, we are generally interested in how robust the Stein effect already established under $D_\alpha$ loss for a specific $\alpha$ is. For example, we can find some concrete problems as follows.

**Problem 1** Under the assumption (AS1) and $D_\alpha$ loss for $-1 < \alpha < 1$, does the predictive density with respect to the same shrinkage prior given by (1.8) improve on the best invariant predictive density with respect to (1.9)?

**Problem 2-1** Under $D_1$ loss, even if $k = 1, 2$, the best invariant predictive density remains inadmissible because an improved non-Bayesian predictive density is easily found. (See Section 4.) Can we determine improved Bayesian predictive densities for this case ($k = 1, 2$)?

**Problem 2-2** Under $k = 1, 2$ and $D_\alpha$ loss with $-1 \leq \alpha < 1$, does the best invariant predictive density keep inadmissibility? If so, which Bayesian predictive density improve the best invariant predictive density?

In this paper, a main focus is on Problem 2-1 and 2-2. For Problem 2-1, we will give an exact solution. We could not completely solve Problem 2-2 in this paper, but by a natural extension of the shrinkage prior considered for Problem 2-1 ($D_1$ loss), we provide a class of predictive densities which will hopefully lead to the solution. In addition, Problem 1 remains open.

The organization of this paper is as follows. We treat not only simple design matrices like (AS1) but also general ones as in the beginning of this section. In order to make structure clearer, Section 2 gives its canonical form. In Section 3, we consider a natural extension of a hierarchical prior which was originally proposed in [17] and [14] for the problem of estimating $\beta$. Using it, we will construct a Bayesian predictive density under $D_\alpha$ loss for $-1 \leq \alpha < 1$ and $\alpha = 1$. In Section 4, we show that a subclass of the Bayesian predictive densities proposed in Section 3 is minimax under $D_1$ loss even if $k$ is small. Section 5 gives concluding remarks.

## 2. A canonical form

In the section, we reduce the problem to a canonical form. To simplify expressions it is helpful to rotate the problem via the following transformation. First we note that for the observation $y$, sufficient statistics are

$$\hat{\beta}_U = (X'X)^{-1}X'y \sim N_k(\beta, \sigma^2(X'X)^{-1}),$$
$$S = \|(I - X(X'X)^{-1}X')y\|^2 \sim \sigma^2\chi^2_{n-k},$$

where $\hat{\beta}_U$ and $S$ are independent.

**Case I:** When $m \geq k$, let $M$ be a nonsingular $k \times k$ matrix which simultaneously diagonalizes matrices $X'X$ and $\tilde{X}'\tilde{X}$,

$$M'(X'X)^{-1}M = \mathrm{diag}(d_1, \ldots, d_k), \quad MM' = \tilde{X}'\tilde{X},$$

where $d_1 \geq \cdots \geq d_k$. Let $V = M'\hat{\beta}_U$ and $\theta = M'\beta$.

**Case II:** When $m < k$, there exists an $(k-m) \times k$ matrix $\tilde{X}_*$ such that $(\tilde{X}', \tilde{X}_*')'$ is a $k \times k$ non-singular matrix and also $\tilde{X}(X'X)^{-1}\tilde{X}_*'$ is an $m \times (k-m)$ zero matrix. Further there exists an $m \times m$ orthogonal matrix $P$ which diagonalizes $\sigma^2\tilde{X}(X'X)^{-1}\tilde{X}'$, the covariance matrix of $\tilde{X}\hat{\beta}_U$, i.e.,

$$P'\tilde{X}(X'X)^{-1}\tilde{X}'P = \mathrm{diag}(d_1, \ldots, d_m),$$

where $d_1 \geq \cdots \geq d_m$. There also exists a $(k-m) \times (k-m)$ matrix $P_*$ such that

$$P_*'\tilde{X}_*(X'X)^{-1}\tilde{X}_*'P_* = I_{k-l}.$$

Put

$$\begin{pmatrix} V \\ V_* \end{pmatrix} = \begin{pmatrix} P' & 0 \\ 0 & P_*' \end{pmatrix} \begin{pmatrix} \tilde{X} \\ \tilde{X}_* \end{pmatrix} \hat{\beta}_U,$$

so that $V$ and $V_*$ are independent and have multivariate normal distributions $N_m(P'\tilde{X}\beta, \sigma^2 D)$ and $N_{k-m}(P_*'\tilde{X}_*\beta, \sigma^2 I_{k-m})$ respectively. Let $\theta = P'\tilde{X}\beta$ and $\mu = P_*'\tilde{X}_*\beta$.

In summary, a canonical form of the prediction problem is as follows. We observe

$$V \sim N_l(\theta, \eta^{-1}D), \quad V_* \sim N_{k-l}(\mu, \eta^{-1}I), \quad \eta S \sim \chi^2_{n-k} \tag{2.1}$$

where $\eta = \sigma^{-2}$, $l = \min(k, m)$, $D = \mathrm{diag}(d_1, \ldots, d_l)$ and $d_1 \geq \cdots \geq d_l$. When $m \geq k$, $V_*$ is null. Then the problem is to give a predictive density of an $m$-dimensional future observation

$$\tilde{Y} \sim N_m(Q\theta, \eta^{-1}I_m), \tag{2.2}$$

where $Q$ is an $m \times l$ matrix, which is given by

$$Q = \begin{cases} P & \text{if } m < k \\ \tilde{X}(M')^{-1} & \text{if } m \geq k, \end{cases}$$

$Q'Q = I_l$. Notice that, under the assumption (AS1), $D$ becomes $N^{-1}I_k$, $V_*$ is 0, and $Q$ becomes $\tilde{X}(\tilde{X}'\tilde{X})^{-1/2}$.

The distribution of $\tilde{y}$ in (2.2) is the same as in (1.2), so it is just the $\tilde{y}$'s that have been transformed. In the remainder of the paper, we will consider the problem in its canonical form, (2.1) and (2.2). For brevity we will use the notation $\hat{p}(\tilde{y}|y)$.

## 3. A class of generalized Bayes predictive densities

In this section, we consider the following class of hierarchical prior densities, $\pi(\theta, \mu, \eta)$, for the canonical model given by (2.1) and (2.2).

$$\begin{aligned} &\theta|\eta, \lambda \sim N_l\left(0, \eta^{-1}\left(D^{-1} + \{(1-\alpha)/2\}I_l\right)^{-1}(C/\lambda - I_l)\right), \\ &\mu|\eta, \lambda \sim N_{k-l}\left(0, \eta^{-1}(\gamma/\lambda - 1)I_{k-l}\right), \\ &\qquad \eta \propto \eta^a, \quad \lambda \propto \lambda^a(1-\lambda)^b I_{(0,1)}(\lambda), \end{aligned} \tag{3.1}$$

where $C = \mathrm{diag}(c_1, \ldots, c_l)$ with $c_i \geq 1$ for $1 \leq i \leq l$, $b = b(\alpha) = (1-\alpha)m/4 + (n-k)/2 - 1$ and $\gamma \geq 1$. The integral which appears in the Bayesian predictive density below is well-defined when $a > -k/2 - 1$. An essentially equivalent class was considered for the problem of estimating $\theta$ and $\sigma^2$ in [14, 15] respectively. When $m \geq k$, the prior on $\mu$ is null and we have only to eliminate $\|V_*\|^2/\gamma$ from the representation of the Bayesian solution in the following theorems 3.1 and 3.2, in order to have the corresponding result.

### 3.1. Case i: $\alpha \in [-1, 1)$

**Theorem 3.1.** *The generalized Bayes predictive density under $D_\alpha$ loss with respect to the prior (3.1) is given by*

$$\hat{p}_\alpha(\tilde{y}|y) \propto \hat{p}_{\{U,\alpha\}}(\tilde{y}|y) \times \check{p}_\alpha(\tilde{y}|y), \tag{3.2}$$

*where*

$$\begin{aligned} \hat{p}_{\{U,\alpha\}}(\tilde{y}|y) &= \left\{(\tilde{y} - Qv)'\Sigma_U^{-1}(\tilde{y} - Qv) + s\right\}^{-m/2-(n-k)/(1-\alpha)}, \\ \check{p}_\alpha(\tilde{y}|y) &= \left\{(\tilde{y} - Q\hat{\theta}_B)'\Sigma_B^{-1}(\tilde{y} - Q\hat{\theta}_B) + R + \frac{\|v_*\|^2}{\gamma} + s\right\}^{-\frac{k+2a+2}{1-\alpha}}, \end{aligned} \tag{3.3}$$

*and where*

$$
\begin{aligned}
\Sigma_U &= \{2/(1-\alpha)\}I + QDQ', \\
\hat{\theta}_B &= (C-I)(C+(1-\alpha)D/2)^{-1}v, \\
\Sigma_B &= \{2/(1-\alpha)\}I + Q(C-I)D(C+\{(1-\alpha)/2\}D)^{-1}Q', \\
R(v) &= v'(\{(1-\alpha)/2\}D+I)D^{-1}(C+\{(1-\alpha)/2\}D)^{-1}v.
\end{aligned} \tag{3.4}
$$

*Proof.* See Appendix. □

The first term $\hat{p}_{\{U,\alpha\}}(\tilde{y}|y)$ is the best invariant predictive density, and is Bayes with respect to the right invariant prior $\pi(\theta,\mu,\eta) = \eta^{-1}$. Up to normalization, $\hat{p}_{\{U,\alpha\}}(\tilde{y}|y)$ is multivariate-$t$ with the mean $Qv = \tilde{X}\hat{\beta}_U$. We omit the straightforward calculation. [12] show that $\hat{p}_{\{U,\alpha\}}(\tilde{y}|y)$ has a constant minimax risk.

The second term, $\check{p}_\alpha$, is a pseudo multivariate-$t$ density with mean vector $Q\hat{\theta}_B$. Since $\|\hat{\theta}_B\| \le \|v\|$, $\check{p}_\alpha$ induces a shrinkage effect toward the origin. The complexity of this term is considerably reduced by the choice $C = I$, in which case $\hat{\theta}_B = 0$, $\Sigma_B = \{2/(1-\alpha)\}I$ and $R(v) = v'D^{-1}v$. However, the covariance matrix of $v$, $\sigma^2 D$, is diagonal, not necessarily a multiple of $I$, so that introduction of $C \neq I$ seems reasonable. Indeed in the context of ridge regression, [4] and [14] have argued that shrinking unstable components more than stable components is reasonable. An ascending sequence of $c_i$'s leads to this end. This additional complexity, while perhaps not pleasing, is nevertheless potentially useful.

### *3.2. Case ii: α = 1*

**Theorem 3.2.** *The generalized Bayes predictive distribution under $D_1$ divergence with respect to the prior (3.1) is a normal distribution $N_m(\hat{\theta}_{\nu,C}, \hat{\sigma}^2_{\nu,C} I_m)$ where*

$$
\begin{aligned}
\hat{\theta}_{\nu,C} &= \left(I - \frac{\nu}{\nu+1+W}C^{-1}\right)V, \\
\hat{\sigma}^2_{\nu,C} &= \left(1 - \frac{\nu}{\nu+1+W}\right)\frac{S}{n-k},
\end{aligned}
$$

*and where $W = \{V'C^{-1}D^{-1}V + \|V_*\|^2/\gamma\}/S$ and $\nu = (k+2a+2)/(n-k)$.*

*Proof.* See Appendix. □

It is quite interesting to note that the Bayesian predictive density $\hat{p}_\alpha(\tilde{y}|y)$ for $\alpha \in [-1,1)$ given in Section 3.1 converges to $\phi_m(\tilde{y}, Q\hat{\theta}_{\nu,C}, \hat{\sigma}^2_{\nu,C})$ as $\alpha \to 1$ where $\phi_m(\cdot,\xi,\tau^2)$ denotes the $m$-variate normal density with the mean vector $\xi$ and the covariance matrix $\tau^2 I_m$.

Since the Bayes solution is the plug-in predictive density as shown in Theorem 3.2, we pay attention only to the properties of plug-in predictive densities under $D_1$ loss. The $\alpha$-divergence with $\alpha = 1$, from $\phi_m(\tilde{y}, Q\hat{\theta}, \hat{\sigma}^2)$, the predictive normal density with plug-in estimators $\hat{\theta}$ and $\hat{\sigma}^2$, to $\phi_m(\tilde{y}, Q\theta, \sigma^2)$, the true normal density,

is given by

$$
\begin{aligned}
&\int \log \frac{\phi_m(\tilde{y}, Q\hat{\theta}, \hat{\sigma}^2)}{\phi_m(\tilde{y}, Q\theta, \sigma^2)} \phi_m(\tilde{y}, Q\hat{\theta}, \hat{\sigma}^2) d\tilde{y} \\
&= \int \left\{ -\frac{m}{2} \log \frac{\hat{\sigma}^2}{\sigma^2} + \frac{\|\tilde{y} - Q\theta\|^2}{2\sigma^2} - \frac{\|\tilde{y} - Q\hat{\theta}\|^2}{2\hat{\sigma}^2} \right\} \phi_m(\tilde{y}, Q\hat{\theta}, \hat{\sigma}^2) d\tilde{y} \\
&= -\frac{m}{2} \log \frac{\hat{\sigma}^2}{\sigma^2} - \frac{m}{2} + \int \left\{ \frac{\|\tilde{y} - Q\hat{\theta} + Q\hat{\theta} - Q\theta\|^2}{2\sigma^2} \right\} \phi_m(\tilde{y}, Q\hat{\theta}, \hat{\sigma}^2) d\tilde{y} \qquad (3.5) \\
&= \frac{\|\hat{\theta} - \theta\|^2}{2\sigma^2} + \frac{m}{2} \left\{ \frac{\hat{\sigma}^2}{\sigma^2} - \log \frac{\hat{\sigma}^2}{\sigma^2} - 1 \right\} \\
&= \frac{1}{2} \left\{ L_1(\hat{\theta}, \theta, \sigma^2) + m L_2(\hat{\sigma}^2, \sigma^2) \right\}.
\end{aligned}
$$

In (3.5), $L_1$ denotes the scale invariant quadratic loss,

$$L_1(\hat{\theta}, \theta, \sigma^2) = \frac{(\hat{\theta} - \theta)'(\hat{\theta} - \theta)}{\sigma^2},$$

for $\theta$, and $L_2$ denotes Stein's or entropy loss,

$$L_2(\hat{\sigma}^2, \sigma^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log \frac{\hat{\sigma}^2}{\sigma^2} - 1,$$

for $\sigma^2$. Hence when the prediction problem under $\alpha$-divergence with $\alpha = 1$ is considered from the Bayesian point of view, the Bayesian solution is the normal distribution with plug-in Bayes estimators and the prediction problem reduces to the simultaneous estimation problem of $\theta$ and $\sigma^2$ under the sum of losses as in (3.5).

## 4. Improved minimax predictive densities under $D_1$

In this section, we give analytical results on minimaxity under $D_1$ loss. As pointed out in the previous section, the prediction problem under $D_1$ loss, reduces to the simultaneous estimation problem of $\theta$ and $\sigma^2$ under the sum of losses as in (3.5). Clearly the UMVU estimators of $\theta$ and $\sigma^2$ are $\hat{\theta}_U = V$ and $\hat{\sigma}^2_U = S/(n-k)$. These are also generalized Bayes estimators with respect to the the right invariant prior $\pi(\theta, \mu, \eta) = \eta^{-1}$ and are hence minimax. The constant minimax risk is given by $\mathrm{MR}_{\theta,\sigma^2}$, where

$$\mathrm{MR}_{\theta,\sigma^2} = \frac{1}{2} \left\{ \mathrm{tr} D + m \left( \log \gamma - \frac{\Gamma'(\gamma)}{\Gamma(\gamma)} \right) \right\} \qquad (4.1)$$

and $\gamma = (n-k)/2$.

Recall that from observation $y$, there exist independent sufficient statistics given by (2.1):

$$V \sim N_l(\theta, \eta^{-1} D), \quad V_* \sim N_{k-l}(\mu, \eta^{-1} I), \quad \eta S \sim \chi^2_{n-k},$$

where $\eta = \sigma^{-2}$, $l = \min(k, m)$, $D = \mathrm{diag}(d_1, \ldots, d_l)$ and $d_1 \geq \cdots \geq d_l$. When $m \geq k$, $V_*$ is empty.

In the variance estimation problem of $\sigma^2$ under $L_2$, [16] showed that $S/(n-k)$ is dominated by

$$\hat{\sigma}^2_{ST} = \min \left( \frac{S}{n-k}, \frac{V' D^{-1} V + S}{l + n - k} \right), \qquad (4.2)$$

for any combination of $\{n, k, m\}$ including $l = \min(k, m) = 1$. Hence, in the simultaneous estimation problem of $\theta$ and $\sigma^2$, we easily see that $\{\hat{\theta}_U, \hat{\sigma}_U^2\}$ is dominated by $\{\hat{\theta}_U, \hat{\sigma}_{ST}^2\}$ and hence have the following result.

**Proposition.** *The estimator* $\{\hat{\theta}_U, \hat{\sigma}_U^2\}$ *is inadmissible for any combination of* $\{n, k, m\}$.

The improved solution, $\{\hat{\theta}_U, \hat{\sigma}_{ST}^2\}$, is not Bayes. When $l \geq 3$ and

$$l - 2 \leq 2\left(d_1^{-1} \textstyle\sum_{i=1}^{l} d_i - 2\right), \tag{4.3}$$

we can construct a Bayesian solution using our earlier studies as follows. In the estimation problem of $\theta$ under $L_1$, [14] showed that the generalized Bayes estimator of $\theta$ with respect to the harmonic-type prior

$$\pi_{S,1}(\theta, \eta) = \{\theta' D^{-1} \theta\}^{1-l/2} \tag{4.4}$$

improves on the UMVU estimator $\hat{\theta}_U$ when $l \geq 3$ and (4.3) is satisfied. In the variance estimation problem of $\sigma^2$ under $L_2$, although [15] did not state so explicitly, they showed that the generalized Bayes estimator of $\sigma^2$ with respect to the same prior (4.4) dominates the UMVU estimator $\hat{\sigma}_U^2$ when $l \geq 3$. Hence the prior (4.4) gives an improved Bayesian solution in the simultaneous estimation problem of $\theta$ and $\sigma^2$ when $l \geq 3$ and (4.3) is satisfied. (Note that under the special assumption (AS1) introduced in Section 1, $D$ becomes the multiple of identity matrix and hence (4.3) is automatically satisfied.)

However, in the above construction of the Bayesian solution, two assumptions, $l \geq 3$ and (4.3) are needed. Further even if $m < k$ and $V_*$ exists, the Bayes procedure does not depend on $V_*$. This is not desirable because the statistic $V_*$ has some information about $\eta$ or $\sigma^2$. In fact, the Stein-type estimator of variance

$$\hat{\sigma}_{ST*}^2 = \min\left(\frac{S}{n-k}, \frac{\|V_*\|^2 + S}{n-l}\right), \tag{4.5}$$

as well as $\{\hat{\sigma}_{ST}^2\}$ dominates $\hat{\sigma}_U^2$ and hence $\{\hat{\theta}_U, \hat{\sigma}_{ST*}^2\}$ also dominates $\{\hat{\theta}_U, \hat{\sigma}_U^2\}$ in the simultaneous estimation problem.

Now we show that a subclass of the generalized Bayes procedure under $D_1$ given in Section 3.2 improves on the generalized Bayes procedure with respect to the right invariant prior. We assume neither $l \geq 3$ nor (4.3). Additionally the proposed procedure does depend on $V_*$ if it exists.

**Theorem 4.1.** *The generalized Bayes estimators of Theorem 3.2,*

$$\hat{\theta}_{\nu,C} = \left(I - \frac{\nu}{\nu + 1 + W} C^{-1}\right) V$$
$$\hat{\sigma}_{\nu,C}^2 = \left(1 - \frac{\nu}{\nu + 1 + W}\right) \frac{S}{n-k},$$

*where* $W = \{V' C^{-1} D^{-1} V + \|V_*\|^2/\gamma\}/S$, *dominate the UMVU estimators* ($V$ *and* $S/(n-k)$) *under the loss* (3.5) *if* $\gamma \geq 1$ *and* $0 < \nu \leq \min(\nu_1, \nu_2, \nu_3)$ *where*

$$\nu_1 = 4\frac{\sum(d_i/c_i) - 2\max(d_i/c_i) + m/(n-k)}{2\max(d_i/c_i)(n-k+2) + m}$$
$$\nu_2 = \frac{4\{\sum(d_i/c_i) - \max(d_i/c_i)\} + 2m/(n-k)}{(n-k-2)\max(d_i/c_i) + m}$$
$$\nu_3 = \frac{4}{m}\sum\frac{d_i}{c_i}.$$

*Proof.* See Appendix. □

Clearly $\nu_2$ and $\nu_3$ are always positive. Now consider $\nu_1$. Assume $\nu_1$ is negative for fixed $C_0$. But there exits $g_0 > 1$ such that $C = g_0 C_0$ makes $\nu_1$ positive. Hence we can choose an increasing sequence of $c_i$'s which guarantees the minimaxity of $(\hat{\theta}_{\nu,C}, \hat{\sigma}^2_{\nu,C})$ and increased shrinkage of unstable components.

**Remark 1.** *We make some comments about domination results under $D_1$ loss for the case of a known variance, say $\sigma^2 = 1$. By (2.1) and (3.5), the prediction problem under $D_1$ loss reduces to the problem of estimating an $l$-dimensional mean vector $\theta$ under the quadratic loss $L_1(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$ in the case where there exists a sufficient statistic $V \sim N_l(\theta, D)$. It is well known that the UMVU estimator $V$ is admissible when $l = 1, 2$, and inadmissible when $l \geq 3$. Minimax admissible estimators for $l \geq 3$ have been proposed by many researchers including [17], [2], [8], and [13]. On the other hand, for KL (i.e. $D_{-1}$) loss, [9] used some techniques including the heat equation and Stein's identity, and eventually found a new identity which links KL risk reduction to Stein's unbiased estimate of risk reduction. Based on this link, they obtained sufficient minimaxity conditions on the Bayesian predictive density. Hence we expect that there should exist an analogous relationship between the prediction problem under $D_\alpha$ loss for $|\alpha| < 1$, and the problem of estimating the mean vector. As far as we know, this is still an open problem.*

## 5. Concluding remarks

In this paper we have studied the construction and behavior of generalized Bayes predictive densities for normal linear models with unknown variance under $\alpha$-divergence loss. In particular we have shown that the best equivariant, (Bayes under the right invariant prior) and minimax predictive density under $D_1$ is inadmissible in all dimensions and for all residual degrees of freedom. We have found a class of improved hierarchical generalized Bayes procedures, which gives a solution to Problem 2-1 of Section 1.

The domination results in this paper are closely related to those in [14, 15] for the respective problems of estimating the mean vector under the quadratic loss and the variance under Stein's loss. In fact a key observation that aids the current development is that the Bayes estimator under $D_1$ loss is a plug-in estimator, specifically a normal density with mean vector and variance closely related to those of the above papers, and that $D_1$ loss is the sum of a quadratic loss in the mean and Stein's loss for the variance.

We expect that an extension of a hierarchical prior given in Section 3.1, for the prediction problem under $D_\alpha$ loss for $-1 \leq \alpha < 1$, can form a basis to solve Problem 2-2 of Section 1. We have been less successful in extending the domination results to the full class of $\alpha$-divergence losses.

## Appendix A: Appendix section

### *A.1. Proof of Theorem 3.1*

The Bayesian predictive density $\hat{p}_\alpha(\tilde{y}|y)$ under $D_\alpha$ divergence for general $\alpha \in [-1, 1)$ is proportional to

$$\text{(A.1)} \qquad \left[ \iiint \{p(\tilde{y}|\theta, \eta)\}^{\frac{1-\alpha}{2}} p(v|\theta, \eta) p(v_*|\mu, \eta) p(s|\eta) \pi(\theta, \mu, \eta) d\theta d\mu d\eta \right]^{\frac{2}{1-\alpha}},$$

and hence the the integral in brackets can be written as

$$\begin{aligned}
&\iiiint \eta^{\frac{m}{2}\frac{1-\alpha}{2}} \exp\left(-\frac{\eta}{2}\frac{1-\alpha}{2}\|\tilde{y}-Q\theta\|^2\right)\eta^{\frac{n-k}{2}}\exp\left(-\frac{\eta s}{2}\right)\\
&\quad\times\eta^{\frac{l}{2}}\exp\left(-\frac{\eta}{2}(v-\theta)'D^{-1}(v-\theta)\right)\eta^{\frac{k-l}{2}}\exp\left(-\frac{\eta}{2}\|v_*-\mu\|^2\right)\\
&\quad\times\frac{\lambda^{l/2}\eta^{l/2}}{\prod(c_i-\lambda)^{1/2}}\exp\left(-\frac{\eta}{2}\theta'\left(D^{-1}+\frac{1-\alpha}{2}I_l\right)(C/\lambda-I_l)^{-1}\theta\right)\\
&\quad\times\left(\frac{\lambda\eta}{\gamma-\lambda}\right)^{(k-l)/2}\exp\left(-\frac{\eta}{2}\frac{\lambda\|\mu\|^2}{\gamma-\lambda}\right)\eta^a\lambda^a(1-\lambda)^b d\theta d\mu d\eta d\lambda.
\end{aligned}\tag{A.2}$$

To simplify integration with respect to $\theta$, we first re-express those terms involving $\theta$ by completing the square, and neglecting, for now, the factor $\eta(1-\alpha)/4$. Let $D_* = \{(1-\alpha)/2\}D$. Then

$$\begin{aligned}
&\|\tilde{y}-Q\theta\|^2+(v-\theta)'D_*^{-1}(v-\theta)+\theta'(I+D_*^{-1})(C/\lambda-I)^{-1}\theta\\
&=\theta'(I+D_*^{-1})(I-C^{-1}\lambda)^{-1}\theta-2\theta'(Q'\tilde{y}+D_*^{-1}v)+\|\tilde{y}\|^2+v'D_*^{-1}v\\
&=\{\theta-(I+D_*^{-1})^{-1}(I-C^{-1}\lambda)(Q'\tilde{y}+D_*^{-1}v)\}'\{(I+D_*^{-1})(I-C^{-1}\lambda)^{-1}\}\\
&\qquad\times\{\theta-(I+D_*^{-1})^{-1}(I-C^{-1}\lambda)(Q'\tilde{y}+D_*^{-1}v)\}\\
&\quad-(Q'\tilde{y}+D_*^{-1}v)'\{(I+D_*^{-1})^{-1}(I-C^{-1}\lambda)\}(Q'\tilde{y}+D_*^{-1}v)+\|\tilde{y}\|^2+v'D_*^{-1}v.
\end{aligned}$$

The "residual term",

$$-(Q'\tilde{y}+D_*^{-1}v)'\{(I+D_*^{-1})^{-1}(I-C^{-1}\lambda)\}(Q'\tilde{y}+D_*^{-1}v)+\|\tilde{y}\|^2+v'D_*^{-1}v,$$

may be expressed as $A+\lambda\{B-A\}$, where

$$\begin{aligned}
A&=A(\tilde{y},v,D_*,Q)\\
&=\|\tilde{y}\|^2+v'D_*^{-1}v-(Q'\tilde{y}+D_*^{-1}v)'(I+D_*^{-1})^{-1}(Q'\tilde{y}+D_*^{-1}v)\\
&=\{2/(1-\alpha)\}(\tilde{y}-Qv)'\Sigma_U^{-1}(\tilde{y}-Qv),
\end{aligned}\tag{A.3}$$

with $\Sigma_U$ given by (3.4) and

$$\begin{aligned}
B&=B(\tilde{y},v,C,D_*,Q)\\
&=\|\tilde{y}\|^2+v'D_*^{-1}v-(Q'\tilde{y}+D_*^{-1}v)'(I+D_*^{-1})^{-1}(I-C^{-1})(Q'\tilde{y}+D_*^{-1}v)\\
&=\{2/(1-\alpha)\}\left\{(\tilde{y}-Q\hat{\theta}_B)'\Sigma_B^{-1}(\tilde{y}-Q\hat{\theta}_B)\right\}\\
&\qquad+\{2/(1-\alpha)\}\left\{v'(\{(1-\alpha)/2\}D+I)D^{-1}(C+\{(1-\alpha)/2\}D)^{-1}v\right\},
\end{aligned}\tag{A.4}$$

where $\hat{\theta}_B$ and $\Sigma_B$ are given by (3.4). The third equality in (A.3) and (A.4) will be proved in Lemma 1 below. Similarly we may re-express the terms involving $\mu$ as

$$\|v_*-\mu\|^2+\frac{\lambda\|\mu\|^2}{\gamma-\lambda}=\frac{\gamma}{\gamma-\lambda}\|\mu\left(1-\{1-\lambda/\gamma\}v_*\right)\|^2+\lambda\frac{\|v_*\|^2}{\gamma}.$$

After integration with respect to $\theta$ and $\mu$, the integral given by (A.2) is proportional to

$$\begin{aligned}
&\iint \eta^{(1-\alpha)m/4+n/2+a}\lambda^{k/2+a}(1-\lambda)^b \\
&\qquad \exp\left(-\frac{\eta}{2}\left\{\frac{1-\alpha}{2}A+s+\lambda\left(\frac{1-\alpha}{2}(B-A)+\frac{\|v_*\|^2}{\gamma}\right)\right\}\right)d\eta d\lambda \\
&\propto \int_0^1 \lambda^{k/2+a}(1-\lambda)^{(1-\alpha)m/4+(n-k)/2-1} \\
&\quad \left\{\frac{1-\alpha}{2}A+s+\lambda\left(\frac{1-\alpha}{2}(B-A)+\frac{\|v_*\|^2}{\gamma}\right)\right\}^{-(1-\alpha)m/4-n/2-a-1} d\lambda.
\end{aligned} \tag{A.5}$$

Note that in an identity given by Maruyama and Strawderman [14, p. 1758],

$$\begin{aligned}
&\int_0^1 \lambda^\alpha(1-\lambda)^\beta(1+w\lambda)^{-\gamma}d\lambda \\
&\qquad = \frac{1}{(w+1)^{\alpha+1}}\int_0^1 t^\alpha(1-t)^\beta\left\{1-\frac{tw}{w+1}\right\}^{-\alpha-\beta+\gamma-2}dt,
\end{aligned} \tag{A.6}$$

the integral of the right-hand side reduces to the beta function $Be(\alpha+1,\beta+1)$ when $-\alpha-\beta+\gamma-2=0$. Hence the integral (A.5) is proportional to

$$\left(\frac{1-\alpha}{2}A+s\right)^{-(1-\alpha)m/4-(n-k)/2}\left(\frac{1-\alpha}{2}B+s+\frac{\|v_*\|^2}{\gamma}\right)^{-k/2-a-1}. \tag{A.7}$$

Since the Bayesian predictive density $\hat{p}_\alpha(\tilde{y}|y)$ with respect to the prior $\pi(\theta,\mu,\eta)$ is a multiple of the integral (A.7) to the $2/(1-\alpha)$ power, the theorem follows.

**Lemma 1.** *Let $F$ and $D_*$ be diagonal matrices and $Q'Q=I$. Then*

$$\begin{aligned}
&G(\tilde{y},v,F,D_*,Q) \\
&= \|\tilde{y}\|^2+v'D_*^{-1}v-(Q'\tilde{y}+D_*^{-1}v)'(I+D_*^{-1})^{-1}F(Q'\tilde{y}+D_*^{-1}v),
\end{aligned}$$

*has the form*

$$\begin{aligned}
&\{\tilde{y}-QF(I+D_*(I-F))^{-1}v\}'\{I+QFD_*(I+D_*(I-F))^{-1}Q'\}^{-1} \\
&\qquad\times\{\tilde{y}-QF(I+D_*(I-F))^{-1}v\} \\
&\quad +v'(D_*+1)(I-F)D_*^{-1}(I+D_*(I-F))^{-1}v.
\end{aligned}$$

*Proof.* The function $G(\tilde{y},v,F,D_*,Q)$ can be re-expressed as

$$\begin{aligned}
G=\tilde{y}'(I-Q(I+D_*^{-1})^{-1}FQ')\tilde{y}-2\tilde{y}'Q(I+D_*)^{-1}Fv \\
+v'D_*^{-1}\{I-(I+D_*)^{-1}F\}v.
\end{aligned}$$

Since

$$(I-Q(I+D_*^{-1})^{-1}FQ')^{-1}=I+QFD_*(I+D_*(I-F))^{-1}Q', \tag{A.8}$$

we obtain

$$\{I+QFD_*(I+D_*(I-F))^{-1}Q'\}Q(I+D_*)^{-1}F=QF(I+D_*(I-F))^{-1}$$

and

$$F(I+D_*)^{-1}Q'\{I+QFD_*(I+D_*(I-F))^{-1}Q'\}Q(I+D_*)^{-1}F$$
$$=F^2(I+D_*)^{-1}(I+D_*(I-F))^{-1}.$$

Hence

$$G=\{\tilde{y}-QF(I+D_*(I-F))^{-1}v\}'(I+QFD_*(I+D_*(I-F))^{-1}Q')^{-1}$$
$$\times\{\tilde{y}-QF(I+D_*(I-F))^{-1}v\}$$
$$-v'F^2(I+D_*)^{-1}(I+D_*(I-F))^{-1}v+v'D_*^{-1}\{I-(I+D_*)^{-1}F\}v.$$

Since the matrix for the quadratic form of $v$ in the "residual term" can be written as

$$D_*^{-1}\{I-(I+D_*)^{-1}F\}-F^2(I+D_*)^{-1}(I+D_*(I-F))^{-1}$$
$$=(D_*+I)(I-F)D_*^{-1}(I+D_*(I-F))^{-1},$$

the lemma follows. □

### *A.2. Proof of Theorem 3.2*

The Bayes predictive density $\hat{p}_\alpha(\tilde{y}|y)$ under the divergence $D_\alpha$ for $\alpha=1$ is proportional to

$$\exp\left\{\iiint \log p(\tilde{y}|\theta,\eta)p(v|\theta,\eta)p(v_*|\mu,\eta)p(s|\eta)\pi(\theta,\mu,\eta)d\theta d\mu d\eta\right\}$$
$$\propto\exp\left\{\int\left(-\eta\frac{\|\tilde{y}-Q\theta\|^2}{2}\right)\pi(\theta,\mu,\eta|v,v_*,s)d\theta d\mu d\eta\right\} \tag{A.9}$$
$$\propto\exp\left(-\frac{E(\eta|v,v_*,s)}{2}\left\|\tilde{y}-Q\frac{E[\eta\theta|v,v_*,s]}{E[\eta|v,v_*,s]}\right\|^2\right).$$

Hence the Bayes solution with respect to the prior density $\pi(\theta,\mu,\sigma^2)$ under $D_1$ is the plug-in normal density

$$\hat{p}_\alpha(\tilde{y}|y)=\phi_m(\tilde{y},Q\hat{\theta}_\pi,\hat{\sigma}_\pi^2),$$

where $\phi_m(\cdot,Q\hat{\theta}_\pi,\hat{\sigma}_\pi^2)$ denotes the $m$-variate normal density with the mean vector $Q\hat{\theta}_\pi$ and the covariance matrix $\hat{\sigma}_\pi^2 I_m$ and where $\hat{\theta}_\pi$ and $\hat{\sigma}_\pi^2$ are given by

$$\hat{\theta}_\pi=\frac{E[\eta\theta|y]}{E[\eta|y]}=v-\frac{D\nabla_v m(v,v_*,s)}{2\{\partial/\partial s\}m(v,v_*,s)},$$
$$\hat{\sigma}_\pi^2=\frac{1}{E[\eta|y]}=-\frac{m(v,v_*,s)}{2\{\partial/\partial s\}m(v,v_*,s)}, \tag{A.10}$$

and where $m(v,v_*,s)$ is the marginal density given by

$$m(v,v_*,s)=\iiint p(v|\theta,\eta)p(v_*|\mu,\eta)p(s|\eta)\pi(\theta,\mu,\eta)d\theta d\mu d\eta. \tag{A.11}$$

Now we consider the marginal density of $(v,v_*,s)$ with respect to the prior $\pi(\theta,\mu,\eta)$, (3.1) with $\alpha=1$. Using essentially the same calculations as in Section 3.1, we obtain the marginal density in the relatively simple form

$$m(v,v_*,s)\propto s^{-(n-k)/2}(v'C^{-1}D^{-1}v+\|v_*\|^2/\gamma+s)^{-(k/2+a+1)}. \tag{A.12}$$

From the expression in (A.10), a straightforward calculation gives the the estimators of $\theta$ and $\sigma^2$ in the simple form

$$\begin{aligned}\hat{\theta}_{\nu,C} &= \left(I - \frac{\nu}{\nu+1+W}C^{-1}\right)V,\\ \hat{\sigma}^2_{\nu,C} &= \left(1 - \frac{\nu}{\nu+1+W}\right)\frac{S}{n-k},\end{aligned} \tag{A.13}$$

where $W = \{V'C^{-1}D^{-1}V + \|V_*\|^2/\gamma\}/S$, respectively. This completes the proof.

### *A.3. Proof of Theorem 4.1*

[14] showed that, under the $L_1$ loss, the risk function of a general shrinkage estimator

$$\hat{\theta}_\phi = \left(I - \frac{\phi(W)}{W}C^{-1}\right)V$$

with suitable $\phi$ is given by

$$\begin{aligned}&E\left[L_1(\hat{\theta}_\phi,\theta,\sigma^2)\right] = E\left[\frac{\|\hat{\theta}_\phi-\theta\|^2}{\sigma^2}\right]\\ &= E\left[\frac{\|V-\theta\|^2}{\sigma^2}\right] + E\left[\frac{\phi(W)}{W}\left\{\psi(V,V_*,C,D,\nu)\Big((n-k+2)\phi(W)\right.\right.\\ &\quad\left.\left.+4\left\{1-\frac{W\phi'(W)}{\phi(W)}(1+\phi(W)\right\}\Big) - 2\sum_{i=1}^{l}\frac{d_i}{c_i}\right\}\right],\end{aligned}$$

where

$$\psi(v,v_*,C,D,\nu) = \frac{v'C^{-2}v}{v'C^{-1}D^{-1}v + \|v_*\|^2/\gamma}.$$

For $\phi_\nu(w) = \nu w/(\nu+1+w)$, we have

$$\begin{aligned}&(n-k+2)\phi(w) + 4\left\{1-\frac{w\phi'(w)}{\phi(w)}(1+\phi(w)\right\}\\ &= \frac{\{(n-k+2)\nu+4\}w^2 + (\nu+1)\{\nu(n-k-2)+4\}w}{(1+\nu+w)^2}\end{aligned}$$

which is always positive when $n-k-2\geq 0$. Since $\psi$ is bounded from above by $\max_{1\leq i\leq l} d_i/c_i$, the risk function of $\hat{\theta}_\nu$ satisfies

$$\begin{aligned}&E\left[L_1(\hat{\theta}_{\nu,C},\theta,\sigma^2)\right]\\ &\leq \mathrm{MR}_\theta + E\left[\frac{\nu}{1+\nu+W}\left\{-2\sum\frac{d_i}{c_i} + \max\frac{d_i}{c_i}\frac{\{(n-k+2)\nu+4\}W^2}{(1+\nu+W)^2}\right.\right.\\ &\quad\left.\left.+\max\frac{d_i}{c_i}\frac{(\nu+1)\{\nu(n-k-2)+4\}W}{(1+\nu+W)^2}\right\}\right],\end{aligned}$$

where $\mathrm{MR}_\theta = \mathrm{tr}D$.

Next we consider the risk function of $\hat{\sigma}^2_\phi = (1-\phi(W)/W)S/(n-k)$ where $0 < \phi(w)/w < 1$, which is given by

$$\begin{aligned} E[L_2(\hat{\sigma}^2_\nu, \sigma^2)] &= E\left[\left(1-\frac{\phi(W)}{W}\right)\frac{S}{(n-k)\sigma^2} - \log\frac{S}{(n-k)\sigma^2}\right. \\ &\quad \left. - \log\left(1-\frac{\phi(W)}{W}\right) - 1\right] \\ &= \mathrm{MR}_{\sigma^2} + E\left[-\frac{\phi(W)}{W\sigma^2}\frac{S}{n-k} - \log\left(1-\frac{\phi(W)}{W}\right)\right]. \end{aligned}$$

Here $\mathrm{MR}_{\sigma^2} = \log\gamma - \Gamma'(\gamma)/\Gamma(\gamma)$ and $\gamma = (n-k)/2$. By the chi-square identity (See e.g. [7]),

$$E\left[\frac{\phi(W)S}{W\sigma^2}\right] = E\left[(n-k+2)\frac{\phi(W)}{W} - 2\phi'(W)\right].$$

Also using the relation

$$-\log(1-x) = \sum_{i=1}^{\infty}\frac{x^i}{i} \le x + \frac{1}{2}\frac{x^2}{1-x},$$

for $0 < x < 1$, we have

$$\begin{aligned} &E[L_2(\hat{\sigma}^2_\nu, \sigma^2)] \\ &\le \mathrm{MR}_{\sigma^2} + E\left[\frac{\phi(W)}{W}\left\{\frac{2}{n-k}\left(\frac{W\phi'(W)}{\phi(W)} - 1\right) + \frac{1}{2}\max\frac{\phi(w)/w}{1-\phi(w)/w}\right\}\right]. \end{aligned}$$

For $\phi(w) = \nu w/(\nu+1+w)$, one gets

$$\begin{aligned} &E[L_2(\hat{\sigma}^2_{\nu,C}, \sigma^2)] \\ &\le \mathrm{MR}_{\sigma^2} + E\left[\frac{\nu}{1+\nu+W}\left\{-\frac{2}{n-k}\frac{W}{1+\nu+W} + \frac{\nu}{2}\right\}\right]. \end{aligned}$$

Hence

$$\frac{1}{2}E\left[L_1(\hat{\theta}_{\nu,C}, \theta, \sigma^2)\right] + \frac{m}{2}E[L_2(\hat{\sigma}^2_{\nu,C}, \sigma^2)] \le \mathrm{MR}_{\theta,\sigma^2} - \frac{\nu}{2}E\left[\frac{\psi(W)}{(1+\nu+W)^3}\right],$$

where $\mathrm{MR}_{\theta,\sigma^2}$ is the minimax risk given by (4.1) and

$$\begin{aligned} \psi(w) = \frac{w^2}{2}&\left(4\left\{\sum\frac{d_i}{c_i} - 2\max\frac{d_i}{c_i} + \frac{m}{n-k}\right\} - \nu\left\{2\max\frac{d_i}{c_i}(n-k+2) + m\right\}\right) \\ &+ (\nu+1)w\left(4\left\{\sum\frac{d_i}{c_i} - \max\frac{d_i}{c_i}\right\} + \frac{2m}{n-k} - \nu\left\{(n-k-2)\max\frac{d_i}{c_i} + m\right\}\right) \\ &+ \frac{(1+\nu)^2}{2}\left(4\sum\frac{d_i}{c_i} - \nu m\right) \ge 0. \end{aligned}$$

Hence the theorem follows.

## References


[1] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353 (52 ##12174)

[2] Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226. MR0397940 (53 ##1795)

[3] Brown, L. D. (1979). A heuristic method for determining admissibility of estimators—with applications. *Ann. Statist.* **7** 960–994. MR536501 (80j:62008)

[4] Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* **8** 1036–1056. MR585702 (82a:62021)

[5] Corcuera, J. M. and Giummolè, F. (1999). A generalized Bayes rule for prediction. *Scand. J. Statist.* **26** 265–279. MR1707658 (2000f:62061)

[6] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345 (36 ##2428)

[7] Efron, B. and Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21. MR0403001 (53 ##6814)

[8] Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (1998). On the construction of Bayes minimax estimators. *Ann. Statist.* **26** 660–671. MR1626063 (99e:62102)

[9] George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.* **34** 78–91. MR2275235 (2008h:62034)

[10] Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Inst. Statist. Math.* **61** 531–542. MR2529965

[11] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415

[12] Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inform. Theory* **50** 2708–2726. MR2096988 (2005f:94040)

[13] Maruyama, Y. (2004). Stein's idea and minimax admissible estimation of a multivariate normal mean. *J. Multivariate Anal.* **88** 320–334. MR2025616 (2004m:62026)

[14] Maruyama, Y. and Strawderman, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Ann. Statist.* **33** 1753–1770. MR2166561 (2006f:62012)

[15] Maruyama, Y. and Strawderman, W. E. (2006). A new class of minimax generalized Bayes estimators of a normal variance. *J. Statist. Plann. Inference* **136** 3822–3836. MR2299167 (2008e:62029)

[16] Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* **16** 155–160. MR0171344 (30 ##1575)

[17] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939 (53 ##1794)