Harmonic Bayesian Prediction Under α -Divergence

Maruyama, Yuzo
Matsuda, Takeru
Ohnishi, Toshio

(Citation)
IEEE Transactions on Information Theory,65(9):5352-5366

(Issue Date)
2019-05-07

(Resource Type)
journal article

(Version)
Accepted Manuscript

Harmonic Bayesian Prediction under α-divergence

Yuzo Maruyama, Takeru Matsuda and Toshio Ohnishi

***Abstract*—We investigate Bayesian shrinkage methods for constructing predictive distributions. We consider the multivariate normal model with a known covariance matrix and show that the Bayesian predictive density with respect to Stein's harmonic prior dominates the best invariant Bayesian predictive density, when the dimension is greater than or equal to three. Alpha-divergence from the true distribution to a predictive distribution is adopted as a loss function.**

***Index Terms*—Bayesian predictive density, harmonic prior, minimaxity**

I. Introduction

Let $X \sim N_d(\mu, v_x I)$ and $Y \sim N_d(\mu, v_y I)$ be independent d-dimensional multivariate normal vectors with common unknown mean μ. We assume that $d \geq 3$ and that v_x and v_y are known. Let $\phi(\cdot, \sigma^2)$ be the probability density of $N_d(0, \sigma^2 I)$. Then the probability density of X and that of Y are $\phi(x-\mu, v_x)$ and $\phi(y-\mu, v_y)$, respectively.

Based on only observing $X = x$, we consider the problem of obtaining a predictive density $\hat{p}(y\,|\,x)$ for Y that is close to the true density $\phi(y-\mu, v_y)$. In most earlier papers on such prediction problems, a predictive density $\hat{p}(y\,|\,x)$ is evaluated by

$$\begin{aligned} & D_{\mathrm{KL}} \left\{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,x) \right\} \\ &= \int_{\mathbb{R}^d} \phi(y-\mu, v_y) \log \frac{\phi(y-\mu, v_y)}{\hat{p}(y\,|\,x)} \mathrm{d}y, \end{aligned} \tag{1}$$

which is called the Kullback-Leibler divergence loss (KL-div loss) from $\phi(y-\mu, v_y)$ to $\hat{p}(y \mid x)$. The overall quality of the procedure $\hat{p}(y\,|\,x)$ for each μ is then summarized by the Kullback-Leibler divergence risk

$$\begin{aligned} & R_{\mathrm{KL}} \{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,\cdot) \} \\ &= \int_{\mathbb{R}^d} D_{\mathrm{KL}} \left\{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,x) \right\} \phi(x-\mu, v_x) \mathrm{d}x. \end{aligned} \tag{2}$$

Aitchison [1] showed that the Bayesian solution with respect to a prior $\pi(\mu)$ under KL-div loss given by (1) is the Bayesian predictive density

$$\hat{p}_\pi(y\,|\,x) = \int_{\mathbb{R}^d} \phi(y-\mu, v_y) \pi(\mu\,|\,x) \mathrm{d}\mu,$$

where $\pi(\mu\,|\,x) = \phi(x-\mu, v_x)\pi(\mu)/m_\pi(x, v_x)$ is the posterior density corresponding to $\pi(\mu)$ and

$$m_\pi(x, v) = \int_{\mathbb{R}^d} \phi(x-\mu, v)\pi(\mu) \mathrm{d}\mu$$

is the marginal density of $X \sim N_d(\mu, vI)$ under the prior $\pi(\mu)$.

For prediction problems in general, many studies suggest the use of the Bayesian predictive density rather than plug-in densities of the form

$$\phi(y - \hat{\mu}(x), v_y),$$

where $\hat{\mu}(x)$ is an estimated value of μ. Liang and Barron [2] showed that the Bayesian predictive density with respect to the uniform prior

$$\pi_{\mathrm{U}}(\mu) = 1, \tag{3}$$

which is given by

$$\begin{aligned} \hat{p}_{\mathrm{U}}(y\,|\,x) &= \int_{\mathbb{R}^d} \phi(y-\mu, v_y) \pi_{\mathrm{U}}(\mu\,|\,x) \mathrm{d}\mu \\ &= \phi(y-x, v_x+v_y) \end{aligned} \tag{4}$$

is best invariant and minimax. Although using the best invariant Bayesian predictive density is generally a good default procedure, it has been shown to be inadmissible in some cases. Specifically, Komaki [3] showed that the Bayesian predictive density with respect to Stein's [4] harmonic prior

$$\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$$

dominates the best invariant Bayesian predictive density $\hat{p}_{\mathrm{U}}(y \,|\, x)$, that is,

$$\begin{aligned} & R_{\mathrm{KL}} \{ \phi(y-\mu, v_y) \mid\mid \hat{p}_{\mathrm{U}}(y\,|\,\cdot) \} \\ & \quad - R_{\mathrm{KL}} \{ \phi(y-\mu, v_y) \mid\mid \hat{p}_{\mathrm{H}}(y\,|\,\cdot) \} \geq 0. \end{aligned}$$

George *et al.* [5] extended Komaki's [3] result to general shrinkage priors including Strawderman's [6] prior.

From a more general viewpoint, the KL-div loss given by (1) is in the class of α-divergence loss (α-div loss) and defined by

$$\begin{aligned} & D_\alpha \left\{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,x) \right\} \\ &= \int_{\mathbb{R}^d} f_\alpha \left(\frac{\hat{p}(y\,|\,x)}{\phi(y-\mu, v_y)} \right) \phi(y-\mu, v_y) \mathrm{d}y, \end{aligned} \tag{5}$$

where

$$f_\alpha(z) = \begin{cases} \left\{ 4/(1-\alpha^2) \right\} \left\{ 1 - z^{(1+\alpha)/2} \right\}, & |\alpha| < 1, \\ z \log z, & \alpha = 1, \\ -\log z, & \alpha = -1. \end{cases} \tag{6}$$

When $\alpha = -1$, we have

$$D_{-1} \left\{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,x) \right\} = D_{\mathrm{KL}} \left\{ \phi(y-\mu, v_y) \mid\mid \hat{p}(y\,|\,x) \right\},$$

Manuscript received January xx, 201x; revised January xx, 201x; accepted January xx, 201x. Date of publication January xx, 201x; date of current version January xx, 201x. This work was supported in part by JSPS KAKENHI under Grant 25330035, Grant 16K00040, Grant 26330044, and Grant 17K00053.

Yuzo Maruyama is with University of Tokyo, Tokyo 153-8902, Japan (e-mail: maruyama@mi.u-tokyo.ac.jp).

Takeru Matsuda is with University of Tokyo, Tokyo 113-8656, Japan, and also with RIKEN Center for Brain Science, Saitama 351-0198, Japan. (e-mail: matsuda@mist.i.u-tokyo.ac.jp).

Toshio Ohnishi is with Kyushu University, Fukuoka 819-0395, Japan (e-mail: ohnishi@econ.kyushu-u.ac.jp).

Digital Object Identifier

where D_{KL} is given by (1). When $\alpha = 0$, we have $f_0(z) = 4(1 - z^{1/2})$ and

$$\begin{aligned} & D_0 \{\phi(y-\mu, v_y) \,||\, \hat{p}(y\,|\,x)\} \\ & = 2 \int_{\mathbb{R}^d} \left\{ \hat{p}^{1/2}(y\,|\,x) - \phi^{1/2}(y-\mu, v_y) \right\}^2 \mathrm{d}y, \end{aligned}$$

where $\sqrt{D_0 \{\phi(y-\mu, v_y) \,||\, \hat{p}(y\,|\,x)\}}/2$ is the Hellinger distance between $\hat{p}(y\,|\,x)$ and $\phi(y-\mu, v_y)$. As in the Kullback-Leibler divergence risk given by (2), the overall quality of the procedure $\hat{p}(y \mid x)$ for each μ is summarized by the α-divergence risk

$$\begin{aligned} & R_\alpha\{\phi(y-\mu, v_y) \,||\, \hat{p}(y\,|\,\cdot)\} \\ & = \int_{\mathbb{R}^d} D_\alpha \{\phi(y-\mu, v_y) \,||\, \hat{p}(y\,|\,x)\} \, \phi(x-\mu, v_x) \mathrm{d}x. \end{aligned}$$

Here, following Cichocki and Amari [7], we provide a brief review of KL-div and α-div from an information-theoretic viewpoint. The α-div was originally proposed by Chernoff [8] and has been extensively investigated and extended by Amari [9], [10], [11] and other researchers. Recall that the most well-known divergences belong to the class of Csiszár's f-divergences [12] and/or the class of Brègman divergences [13]. The KL-div given by (1) is known as the only divergence belonging to the intersection of the two classes. The α-div can be derived from the f-divergence and as shown by Amari [10] using some tricks also from Brègman divergence [13]. Hence, following KL-div as the first choice, α-div seems the second choice among a class of most well-known divergences. In this paper, we will consider statistical decision theory of Bayesian predictive density under α-div loss for general $\alpha \in (-1, 1)$ and robustness of minimaxity over $\alpha \in [-1, 1]$.

Corcuera and Giummolè [14] showed that a Bayesian predictive density under α-div loss is

$$\hat{p}_\pi(y\,|\,x;\alpha) \propto \begin{cases} \left\{ \int_{\mathbb{R}^d} \phi^{\frac{1-\alpha}{2}}(y-\mu, v_y) \phi(x-\mu, v_x) \pi(\mu) \mathrm{d}\mu \right\}^{\frac{2}{1-\alpha}} \\ \text{for } -1 \le \alpha < 1, \\ \exp\left(\int_{\mathbb{R}^d} \{\log \phi(y-\mu, v_y)\} \, \phi(x-\mu, v_x) \pi(\mu) \mathrm{d}\mu \right) \\ \text{for } \alpha = 1. \end{cases} \tag{7}$$

By (7), in the prediction problem under α-div loss with $\alpha = 1$ from the Bayesian point of view, the Bayesian solution is the normal density

$$\hat{p}_\pi(y\,|\,x;1) = \phi(y - \hat{\mu}_\pi(x), v_y),$$

where $\hat{\mu}_\pi(x)$ is the posterior mean given by

$$\hat{\mu}_\pi(x) = \int_{\mathbb{R}^d} \mu \pi(\mu\,|\,x) \mathrm{d}\mu = x + v_x \nabla_x \log m(x, v_x)$$

with $\nabla_x = (\partial/\partial x_1, \ldots, \partial/\partial x_d)$. In general, the Bayesian prediction problem under $\alpha = 1$ reduces to the estimation problem under the KL-div loss in the case of the exponential family density. This is because the exponential family density is closed under the calculation in (7) with $\alpha = 1$, as pointed out in Yamagimoto and Ohnishi [15].

As demonstrated in Maruyama and Strawderman [16], the α-div loss in the case of $\alpha = 1$ is written as

$$D_1 \{\phi(y-\mu, v_y) \,||\, \phi(y - \hat{\mu}_\pi(x), v_y)\} = \frac{\|\hat{\mu}_\pi(x) - \mu\|^2}{2v_y},$$

and hence the prediction problem under $\alpha = 1$ reduces to the estimation problem of μ under the quadratic loss. Stein [17] showed that

$$\begin{aligned} & E_X \left[\|\hat{\mu}_\pi(X) - \mu\|^2 \right] \\ & = d v_x + 4 v_x^2 E_X \left[\frac{\Delta_x m_\pi^{1/2}(X, v_x)}{m_\pi^{1/2}(X, v_x)} \right], \end{aligned}$$

where $\Delta_x = \sum_{i=1}^d \partial^2/\partial x_i^2$. Hence the risk difference under $\alpha = 1$ is expressed as

$$\begin{aligned} & R_1\{\phi(y-\mu, v_y) \,||\, \hat{p}_{\text{U}}(y\,|\,\cdot;1)\} \\ & \quad - R_1\{\phi(y-\mu, v_y) \,||\, \hat{p}_\pi(y\,|\,\cdot;1)\} \\ & = \frac{2v_x^2}{v_y} E_X \left[- \frac{\Delta_x m_\pi^{1/2}(X, v_x)}{m_\pi^{1/2}(X, v_x)} \right]. \end{aligned} \tag{8}$$

Under the KL-div loss or α-div loss with $\alpha = -1$, [5] showed that the risk difference is given by

$$\begin{aligned} & R_{-1}\{\phi(y-\mu, v_y) \,||\, \hat{p}_{\text{U}}(y\,|\,\cdot;-1)\} \\ & \quad - R_{-1}\{\phi(y-\mu, v_y) \,||\, \hat{p}_\pi(y\,|\,\cdot;-1)\} \\ & = 2 \int_{v_*}^{v_x} E_Z \left[- \frac{\Delta_z m_\pi^{1/2}(Z, v)}{m_\pi^{1/2}(Z, v)} \right] \mathrm{d}v, \end{aligned} \tag{9}$$

where $\hat{p}_{\text{U}}(y\,|\,x;-1)$ is given by (4), $Z \sim N_d(\mu, vI)$ and $v_* = v_x v_y/(v_x + v_y)$. From this viewpoint, [5] and Brown *et al.* [18] considered the prediction problem under α-div loss in two extreme cases $\alpha = \pm 1$ and found a beautiful relationship of risk differences for two cases via $\Delta_z \{m_\pi(z, v)\}^{1/2}$ for some v. Under both risks R_1 and R_{-1}, any shrinkage prior of the satisfier of the superharmonicity

$$\Delta_z m_\pi^{1/2}(z, v) \le 0 \quad \text{for } \begin{cases} \forall v \in (v_*, v_x) \text{ for } \alpha = -1, \\ v = v_x \text{ for } \alpha = 1, \end{cases} \tag{10}$$

implies an improvement over the best invariant Bayesian procedure. As in [17], the superharmonicity of $\pi(\mu)$, $\Delta_\mu \pi(\mu) \le 0$, implies the superharmonicity of $m_\pi(z, v)$, $\Delta_z m_\pi(z, v) \le 0$. Further the superharmonicity of $m_\pi(z, v)$ implies the superharmonicity of $\{m_\pi(z, v)\}^{1/2}$. Hence the harmonic prior $\pi_{\text{H}}(\mu) = \|\mu\|^{-(d-2)}$ gives the superharmonicity of $\{m_\pi(z, v)\}^{1/2}$.

Because of the relationship given by (8), (9) and (10), it is of great interest to find the corresponding link via $\Delta_z \{m_\pi(z, v)\}^{1/2}$ for α-div loss with general $\alpha \in (-1, 1)$ so that the superharmonicity of $\{m_\pi(z, v)\}^{1/2}$ implies minimaxity or equivalently the improvement over the best invariant Bayesian procedure, which is the motivation of this paper. In other words, we are interested in a kind of robustness of the minimaxity result via superharmonicity uniformly for $\alpha \in [-1, 1]$, where the theory under two extreme cases $\alpha = \pm 1$ has been already established. To our knowledge, decision-theoretic properties 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, as Brown [19] pointed out in the estimation problem.

In this paper, we investigate the risk difference, $\mathrm{diff} R_{\alpha,\mathrm{U},\pi}$, in the case of α-div loss, defined by

$$\begin{aligned}\mathrm{diff} R_{\alpha,\mathrm{U},\pi} &= R_\alpha \left\{\phi(y-\mu, v_y) \mid\mid \hat{p}_{\mathrm{U}}(y \mid \cdot; \alpha)\right\} \\ &\quad - R_\alpha \left\{\phi(y-\mu, v_y) \mid\mid \hat{p}_{\pi}(y \mid \cdot; \alpha)\right\}.\end{aligned} \tag{11}$$

In (11), $\hat{p}_\pi(y \mid x; \alpha)$ is given by (7) and $\hat{p}_{\mathrm{U}}(y \mid x; \alpha)$ is the Bayesian predictive density under the uniform prior (3), the form of which will be derived in (16) of Section II. As a generalization of [2]'s result, $\hat{p}_{\mathrm{U}}(y \mid x; \alpha)$ for general $\alpha \in (-1, 1)$ is best invariant and minimax, as shown in Appendix A. Further, analyzing $\mathrm{diff} R_{\alpha,\mathrm{U},\pi}$, we provide some asymptotic results (Theorem 2.4) and a non-asymptotic decision-theoretic result (Theorem 3.2).

Asymptotic results We show not only somewhat expected relationship

$$\begin{aligned}\lim_{\alpha\to 1-0} \mathrm{diff} R_{\alpha,\mathrm{U},\pi} &= \mathrm{diff} R_{1,\mathrm{U},\pi}, \\ \lim_{\alpha\to -1+0} \mathrm{diff} R_{\alpha,\mathrm{U},\pi} &= \mathrm{diff} R_{-1,\mathrm{U},\pi},\end{aligned} \tag{12}$$

where $\mathrm{diff} R_{1,\mathrm{U},\pi}$ and $\mathrm{diff} R_{-1,\mathrm{U},\pi}$ are given in (8) and (9) respectively, but also the asymptotic relationship for general $\alpha \in (-1, 1)$,

$$\lim_{v_x/v_y\to +0} \mathrm{diff} R_{\alpha,\mathrm{U},\pi} = \mathrm{diff} R_{1,\mathrm{U},\pi}. \tag{13}$$

Hence, the asymptotic situation $v_x/v_y \to 0$ corresponds to the case $\alpha \to 1$ and $\Delta_z\{m_\pi(z, v)\}^{1/2}$ plays an important role for general $\alpha \in (-1, 1)$.

Non-asymptotic result We particularly investigate a decision-theoretic property of the Bayesian predictive density with respect to $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ under α-div loss with general $\alpha \in (-1, 1)$. We show that, the Bayesian predictive density with respect to $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ dominates the best invariant Bayesian predictive density with respect to $\pi_{\mathrm{U}}(\mu) = 1$ if

$$\frac{v_x}{v_y} \le \begin{cases} \dfrac{d+2}{d(1+\alpha)} \text{ if } \dfrac{2}{1-\alpha} \text{ is a positive integer}, \\ \left(\dfrac{2}{1-\alpha}\right)^2 \dfrac{d+2}{d} \dfrac{1-\{\kappa - 2/(1-\alpha)\}}{2\kappa(\kappa-1)} \text{ otherwise}, \end{cases}$$

where κ is the smallest integer larger than $2/(1-\alpha)$.

The organization of this paper is as follows. In Section II, we derive the exact form of $\hat{p}_\pi(y \mid x; \alpha)$, propose a general sufficient condition for $\mathrm{diff} R_{\alpha,\mathrm{U},\pi} \ge 0$, where $\mathrm{diff} R_{\alpha,\mathrm{U},\pi}$ is given by (11), and demonstrate the asymptotic relationship described in (12) and (13). In Section III, we propose the non-asymptotic result under the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ described above. Some technical proofs are given in Sections A–D of Appendix.

II. A General Sufficient Condition for Minimaxity

A. *Bayesian predictive density under α-divergence loss*

As in (7), the Bayes predictive density under α-div loss is

$$\begin{aligned}&\hat{p}_\pi(y \mid x; \alpha) \\ &\propto \left\{\int_{\mathbb{R}^d} \phi(x-\mu, v_x)\, \phi^\beta(y-\mu, v_y)\pi(\mu)\mathrm{d}\mu\right\}^{1/\beta},\end{aligned} \tag{14}$$

where

$$\beta = \frac{1-\alpha}{2}.$$

Clearly, it follows from $\alpha \in (-1, 1)$ that $0 < \beta < 1$. Let

$$\gamma = \frac{1}{1+\beta v_x/v_y}.$$

Since the relation of completing squares with respect to μ, for $\phi(x-\mu, v_x)\, \phi^\beta(y-\mu, v_y)$, is given by

$$\begin{aligned}&\frac{1}{v_x}\|x-\mu\|^2 + \frac{\beta}{v_y}\|y-\mu\|^2 \\ &= \frac{1}{v_x}\left(\|x-\mu\|^2 + \frac{1-\gamma}{\gamma}\|y-\mu\|^2\right) \\ &= \frac{1}{v_x}\left(\frac{1}{\gamma}\|\mu - \{\gamma x + (1-\gamma)y\}\|^2 \right.\\ &\qquad \left. - \frac{\|\gamma x + (1-\gamma)y\|^2}{\gamma} + \|x\|^2 + \frac{1-\gamma}{\gamma}\|y\|^2\right) \\ &= \frac{1}{v_x}\left\{\frac{1}{\gamma}\|\mu - \{\gamma x + (1-\gamma)y\}\|^2 + (1-\gamma)\|y-x\|^2\right\} \\ &= \frac{1}{v_x\gamma}\|\mu - \{\gamma x + (1-\gamma)y\}\|^2 + \beta\frac{\gamma}{v_y}\|y-x\|^2,\end{aligned}$$

we have the identity,

$$\begin{aligned}&\phi(x-\mu, v_x)\, \phi^\beta(y-\mu, v_y) \\ &= \gamma^{(1-\beta)d/2}\phi(\gamma x + (1-\gamma)y - \mu, v_x\gamma) \\ &\quad \times \phi^\beta(y-x, v_y/\gamma).\end{aligned} \tag{15}$$

Under the uniform prior $\pi_{\mathrm{U}}(\mu) = 1$, we have, from (15),

$$\begin{aligned}&\int_{\mathbb{R}^d} \phi(x-\mu, v_x)\, \phi^\beta(y-\mu, v_y)\pi_{\mathrm{U}}(\mu)\mathrm{d}\mu \\ &= \gamma^{(1-\beta)d/2}\phi^\beta(y-x, v_y/\gamma)\end{aligned}$$

in (14). Therefore the Bayesian predictive density under the uniform prior is

$$\hat{p}_{\mathrm{U}}(y \mid x; \alpha) = \phi(y-x, v_y/\gamma) = \phi(y-x, v_y + \beta v_x), \tag{16}$$

which is the target predictive density so that the risk difference

$$\begin{aligned}\mathrm{diff} R_{\alpha,\mathrm{U},\pi} &= R_\alpha \left\{\phi(y-\mu, v_y) \mid\mid \hat{p}_{\mathrm{U}}(y \mid \cdot; \alpha)\right\} \\ &\quad - R_\alpha \left\{\phi(y-\mu, v_y) \mid\mid \hat{p}_{\pi}(y \mid \cdot; \alpha)\right\}\end{aligned}$$

is going to be investigated in this paper. As shown in Appendix A, $\hat{p}_{\mathrm{U}}(y \mid x; \alpha)$ for general $\alpha \in (-1, 1)$ is best invariant and minimax, which is regarded as a generalization of [2]'s minimaxity result. Hence $\hat{p}_\pi(y \mid x; \alpha)$ with $\mathrm{diff} R_{\alpha,\mathrm{U},\pi} \ge 0$ for all $\mu \in \mathbb{R}^d$ is minimax.

The exact form of Bayes predictive density $\hat{p}_\pi(y \mid x; \alpha)$ for (14) with normalizing constant, which is regarded as a generalization of Theorem 1 of [3] as well as Lemma 2 of [5], is provided as follows.

Theorem 2.1: The Bayes predictive density under $\pi(\mu)$ is

$$\hat{p}_\pi(y \mid x; \alpha) = \frac{m_\pi^{1/\beta}(\gamma x + (1-\gamma)y, v_x\gamma)}{E_{Z_1}\left[m_\pi^{1/\beta}(x + \xi Z_1, v_x\gamma)\right]}\hat{p}_{\mathrm{U}}(y \mid x; \alpha), \tag{17}$$

where $Z_1 \sim N_d(0, I)$ and

$$\xi = (1-\gamma)(v_y/\gamma)^{1/2}. \tag{18}$$

Proof: By (14), (15) and (16), we have

$$\hat{p}_\pi(y\,|\,x;\alpha) \propto \phi(y-x, v_y/\gamma) m_\pi^{1/\beta}(\gamma x + (1-\gamma)y, v_x\gamma). \quad (19)$$

The normalizing constant of (19) is

$$\begin{aligned}
&\int_{\mathbb{R}^d} \phi(y-x, v_y/\gamma) m_\pi^{1/\beta}(\gamma x + (1-\gamma)y, v_x\gamma)\mathrm{d}y \\
&= \int_{\mathbb{R}^d} \phi(z_1, 1) m_\pi^{1/\beta}\left(x + (1-\gamma)(v_y/\gamma)^{1/2} z_1, v_x\gamma\right)\mathrm{d}z_1 \\
&= E_{Z_1}\left[m_\pi^{1/\beta}(x + \xi Z_1, v_x\gamma)\right],
\end{aligned}$$

where the first equality is from the transformation, $z_1 = (\gamma/v_y)^{1/2}(y-x)$. ■

B. A general sufficient condition for minimaxity

In the following, as a generalization of the Bayes predictive density, we consider

$$\hat{p}_f(y\,|\,x;\alpha) = \frac{f(\gamma x + (1-\gamma)y)}{E_{Z_1}[f(x+\xi Z_1)]}\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha) \quad (20)$$

where $f : \mathbb{R}^d \to \mathbb{R}_+$ is general. As in the proof of Theorem 2.1, $\int \hat{p}_f(y \mid x;\alpha)\mathrm{d}y = 1$ follows. Also $\hat{p}_f(y \mid x;\alpha)$ is nonnegative for any $y \in \mathbb{R}^d$ and hence $\hat{p}_f(y\,|\,x;\alpha)$ is regarded as a predictive density.

By the definition of the α-div loss given by (5), the risk difference between $\hat{p}_{\mathrm{U}}$ and $\hat{p}_f$ is written as

$$\begin{aligned}
&\mathrm{diff} R_{\alpha,\mathrm{U},f} \\
&= R_\alpha\{\phi(y-\mu, v_y) \,||\, \hat{p}_{\mathrm{U}}(y\,|\,\cdot;\alpha)\} \\
&\quad - R_\alpha\{\phi(y-\mu, v_y) \,||\, \hat{p}_f(y\,|\,\cdot;\alpha)\} \\
&= \frac{1}{\beta(1-\beta)}\int_{\mathbb{R}^{2d}} \phi(x-\mu, v_x)\phi(y-\mu, v_y) \\
&\quad \times \left\{\left(\frac{\hat{p}_f(y\,|\,x;\alpha)}{\phi(y-\mu,v_y)}\right)^{1-\beta} - \left(\frac{\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha)}{\phi(y-\mu,v_y)}\right)^{1-\beta}\right\}\mathrm{d}x\mathrm{d}y.
\end{aligned} \quad (21)$$

Then we have the following result.

Theorem 2.2:

1) The risk difference $\mathrm{diff} R_{\alpha,\mathrm{U},f}$ given by (21) is written by $E[\rho(W,Z)]$ where $W \sim N_d(\mu, v_x\gamma)$, $Z \sim N_d(0, I)$, W and Z are independent, and

$$\rho(w,z) = \frac{4\gamma^{(1-\beta)d/2}}{\beta^2 f^{\beta-1}(w)}\int_0^\xi t\frac{-\Delta_w \varrho(w+tz;t;f)}{\varrho^{2/\beta-1}(w+tz;t;f)}\mathrm{d}t \quad (22)$$

where

$$\varrho(u;t;f) = \{E_{Z_1}[f(tZ_1+u)]\}^{\beta/2}, \quad (23)$$

for $Z_1 \sim N_d(0, I)$.

2) A sufficient condition for $\mathrm{diff} R_{\alpha,\mathrm{U},f} \geq 0$ for $\forall\mu \in \mathbb{R}^d$ is

$$\Delta_u \varrho(u;t;f) \leq 0 \quad \forall u \in \mathbb{R}^d,\ 0 \leq \forall t \leq \xi.$$

Proof: Part 2 easily follows from Part 1 and, in the following, we show Part 1.

By (15), (16), and (20), the integrand of (21) is rewritten as

$$\begin{aligned}
&\left\{\left(\frac{\phi(y-\mu,v_y)}{\hat{p}_f(y\,|\,x;\alpha)}\right)^{\beta-1} - \left(\frac{\phi(y-\mu,v_y)}{\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha)}\right)^{\beta-1}\right\} \\
&\quad \times \phi(y-\mu,v_y)\phi(x-\mu,v_x) \\
&= \gamma^{(1-\beta)d/2}\left\{\left(\frac{E_{Z_1}[f(x+\xi Z_1)]}{f(\gamma x+(1-\gamma)y)}\right)^{\beta-1} - 1\right\} \\
&\quad \times \phi(\gamma x + (1-\gamma)y - \mu, v_x\gamma)\phi(y-x, v_y/\gamma).
\end{aligned}$$

By the change of variables, $w = \gamma x + (1-\gamma)y$ and $z = -(\gamma/v_y)^{1/2}(y-x)$, where Jacobian of the matrix below is $(\gamma/v_y)^{d/2}$,

$$\begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} \gamma I_d & (1-\gamma)I_d \\ (\gamma/v_y)^{1/2}I_d & -(\gamma/v_y)^{1/2}I_d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},$$

the risk difference is expressed as

$$\begin{aligned}
&\frac{\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}E_{W,Z}\left[\left(E_{Z_1}\left[\frac{f(W+\xi(Z_1+Z))}{f(W)}\right]\right)^{\beta-1} - 1\right] \\
&= \frac{\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}E_W\left[f(W)^{1-\beta}\{g(\xi;W) - g(0;W)\}\right] \\
&= \frac{\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}E_W\left[f(W)^{1-\beta}\int_0^\xi \frac{\partial}{\partial t}g(t;W)\mathrm{d}t\right],
\end{aligned} \quad (24)$$

where $\xi = (1-\gamma)(v_y/\gamma)^{1/2}$ as in (18), $W \sim N_d(\mu, v_x\gamma I)$, $Z_1 \sim N_d(0, I)$, $Z \sim N_d(0, I)$ and

$$g(t;w) = E_Z\left[E_{Z_1}[f(w+t\{Z_1+Z\})]^{\beta-1}\right].$$

In the following, $E_{Z_1}[f] = E_{Z_1}[f(w+t\{Z_1+z\})]$ for notational simplicity. Then we have

$$\begin{aligned}
&\frac{\partial}{\partial t}g(t;w) \\
&= E_Z\left[\frac{\partial}{\partial t}\{E_{Z_1}[f]\}^{\beta-1}\right] \\
&= (\beta-1)E_Z\left[\{E_{Z_1}[f]\}^{\beta-2}E_{Z_1}[(Z_1+Z)^{\mathrm{T}}\nabla_w f]\right] \\
&= (\beta-1)E_Z\left[\{E_{Z_1}[f]\}^{\beta-2}(E_{Z_1}[Z_1^{\mathrm{T}}\nabla_w f] \right. \\
&\qquad \left. + Z^{\mathrm{T}}E_{Z_1}[\nabla_w f])\right].
\end{aligned} \quad (25)$$

In (25), we have

$$\begin{aligned}
E_{Z_1}[Z_1^{\mathrm{T}}\nabla_w f] &= E_{Z_1}\left[Z_1^{\mathrm{T}}\frac{1}{t}\nabla_{z_1} f\right] = \frac{1}{t}E_{Z_1}[\Delta_{z_1} f] \\
&= tE_{Z_1}[\Delta_w f] = t\Delta_w E_{Z_1}[f]
\end{aligned} \quad (26)$$

where the second equality follows from the Gauss divergence theorem. Similarly we have

$$\begin{aligned}
&(\beta-1)E_Z\left[\{E_{Z_1}[f]\}^{\beta-2}Z^{\mathrm{T}}E_{Z_1}[\nabla_w f]\right] \\
&= (\beta-1)E_Z\left[\{E_{Z_1}[f]\}^{\beta-2}Z^{\mathrm{T}}\frac{1}{t}E_{Z_1}[\nabla_z f]\right] \\
&= \frac{1}{t}(\beta-1)E_Z\left[\{E_{Z_1}[f]\}^{\beta-2}Z^{\mathrm{T}}\nabla_z E_{Z_1}[f]\right] \\
&= \frac{1}{t}E_Z\left[Z^{\mathrm{T}}\nabla_z\{E_{Z_1}[f]\}^{\beta-1}\right] \\
&= \frac{1}{t}E_Z\left[\Delta_z\{E_{Z_1}[f]\}^{\beta-1}\right] \\
&= tE_Z\left[\Delta_w\{E_{Z_1}[f]\}^{\beta-1}\right],
\end{aligned} \quad (27)$$

where the fourth equality follows from the Gauss divergence theorem. By (25), (26) and (27), we have

$$\frac{\partial}{\partial t}g(t;w) = tE_Z\Big[\Delta_w\left\{E_{Z_1}[f]\right\}^{\beta-1} + (\beta-1)\left\{E_{Z_1}[f]\right\}^{\beta-2}\Delta_w E_{Z_1}[f]\Big]. \tag{28}$$

Recall the formula of Laplacian for a function $h(u)$,

$$\frac{\Delta_u h^a(u)}{ah^a(u)} = \frac{\Delta_u h(u)}{h(u)} + (a-1)\|\nabla_u \log h(u)\|^2, \tag{29}$$

for $a \neq 0$. Then, in (28), we have

$$\begin{aligned}
&\Delta_w\left\{E_{Z_1}[f]\right\}^{\beta-1} + (\beta-1)\left\{E_{Z_1}[f]\right\}^{\beta-2}\Delta_w E_{Z_1}[f] \\
&= \frac{(\beta-1)}{\left\{E_{Z_1}[f]\right\}^{1-\beta}} \times\left(2\frac{\Delta_w E_{Z_1}[f]}{E_{Z_1}[f]} + (\beta-2)\|\nabla_w \log E_{Z_1}[f]\|^2\right) \\
&= \frac{2(\beta-1)}{\left\{E_{Z_1}[f]\right\}^{1-\beta}} \times\left(\frac{\Delta_w E_{Z_1}[f]}{E_{Z_1}[f]} + (\beta/2-1)\|\nabla_w \log E_{Z_1}[f]\|^2\right) \\
&= \frac{2(\beta-1)}{\left\{E_{Z_1}[f]\right\}^{1-\beta}}\frac{\Delta_w\left\{E_{Z_1}[f]\right\}^{\beta/2}}{(\beta/2)\left\{E_{Z_1}[f]\right\}^{\beta/2}} \\
&= \frac{4(\beta-1)}{\beta}\frac{\Delta_w\left\{E_{Z_1}[f]\right\}^{\beta/2}}{\left\{E_{Z_1}[f]\right\}^{1-\beta/2}}.
\end{aligned} \tag{30}$$

By (24), (28) and (30), we completes the proof. ■

Remark 2.1: In the previous version of this article as well as [5], not only the Stein identity but also the heat equation

$$\frac{\partial}{\partial v}\phi(u,v) = \frac{1}{2}\Delta_u\phi(u,v),$$

was efficiently applicable for deriving a nice expression of the risk difference, like Part 1 of Theorem 2.2. It seemed to us that the heat equation was an additional necessary tool for investigating the Stein phenomenon of predictive density. But it is not true, the heat equation is no longer necessary. As seen in the proof of Theorem 2.2, only the Stein identity or the Gauss divergence theorem is the key, as in the Stein "estimation" problem.

The superharmonicity of f implies the superharmonicity of $E_{Z_1}[f(tZ_1+u)]$. Furthermore, using the relationship (29), we see that the superharmonicity of $E_{Z_1}[f(tZ_1+u)]$ implies the superharmonicity of

$$\varrho(u;t;f) = \left\{E_{Z_1}[f(tZ_1+u)]\right\}^{\beta/2}$$

for $\beta \in (0,1)$. Hence, for Part 2 of Theorem 2.2, we have the following corollary.

Corollary 2.1: Suppose $f : \mathbb{R}^d \to \mathbb{R}_+$ is superharmonic. Then the predictive density $\hat{p}_f(y\,|\,x;\alpha)$ given by (20) as

$$\hat{p}_f(y\,|\,x;\alpha) = \frac{f(\gamma x + (1-\gamma)y)}{E_{Z_1}[f(x+\xi Z_1)]}\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha),$$

dominates $\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha)$.

In Section III, we will investigate the properties of the Bayesian predictive density $\hat{p}_\pi(y\,|\,x;\alpha)$ where

$$f(u) = \{m_\pi(u, v_x\gamma)\}^{1/\beta}$$

is assumed in Theorem 2.2 and Corollary 2.1. Actually in this case, Corollary 2.1 is not useful since the superharmonicity of $\{m_\pi(u,v_x\gamma)\}^{1/\beta}$ for $\beta \in (0,1)$ is very restrictive. Recall the relationship given by (29). For example, the superharmonicity of $m_\pi(u,v_x\gamma)$ does not imply the superharmonicity of $\{m_\pi(u,v_x\gamma)\}^{1/\beta}$. Hence, in Section III, we will seriously consider the superharmonicity of

$$\varrho(u;t;m_\pi^{1/\beta}) = \Big\{E_{Z_1}\Big[\{m_\pi(tZ_1+u,v_x\gamma)\}^{1/\beta}\Big]\Big\}^{\beta/2}.$$

Further, when $1/\beta = 2/(1-\alpha)$ is not an integer, $E_{Z_1}\left[\{m_\pi(tZ_1+u,v_x\gamma)\}^{1/\beta}\right]$ in Part 2 of Theorem 2.2 is not tractable for our current methodology in Section III. Thus we propose a variant of Theorem 2.2 with $f(u) = \{m_\pi(u,v_x\gamma)\}^{1/\beta}$, for a non-integer $1/\beta$ as follows. Let κ be the smallest integer among integers which is strictly greater than $1/\beta$,

$$\kappa = \min\{n \in \mathbb{Z} \mid n > 1/\beta\}.$$

Then $\kappa - 1 < 1/\beta < \kappa$. From Jensen's inequality, we have

$$\begin{aligned}
&E_{Z_1}\Big[m_\pi^{1/\beta}(w+\xi(Z_1+Z), v_x\gamma)\Big] \\
&= E_{Z_1}\Big[\{m_\pi^\kappa(w+\xi(Z_1+Z), v_x\gamma)\}^{1/(\beta\kappa)}\Big] \\
&\leq \{E_{Z_1}[m_\pi^\kappa(w+\xi(Z_1+Z), v_x\gamma)]\}^{1/(\beta\kappa)},
\end{aligned} \tag{31}$$

since $0 < 1/(\beta\kappa) < 1$ and hence

$$\begin{aligned}
&R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_{\mathrm{U}}(y\,|\,\cdot;\alpha)\} \\
&\quad - R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_\pi(y\,|\,\cdot;\alpha)\} \\
&\geq \frac{\gamma^{(1-\beta)d/2}}{\beta(1-\beta)} \\
&\quad\times E_{W,Z}\left[E_{Z_1}\left[\frac{m_\pi^\kappa(W+\xi(Z_1+Z),v_x\gamma)}{m_\pi^\kappa(W,v_x\gamma)}\right]^{\frac{\beta-1}{\beta\kappa}} - 1\right].
\end{aligned}$$

Applying the same technique starting (24) through (30) to the lower bound above, we have a variant of Part 2 of Theorem 2.2.

Theorem 2.3: Assume $1/\beta$ is not a positive integer. Let κ be the smallest integer greater than $1/\beta$. A sufficient condition for $\mathrm{diff}R_{\alpha,\mathrm{U},\pi} \geq 0$ is

$$\begin{aligned}
&\Delta_u\left\{E_{Z_1}[m_\pi^\kappa(tZ_1+u,v_x\gamma)]\right\}^{c(\beta)/\kappa} \leq 0, \\
&\forall u \in \mathbb{R}^d,\ 0 \leq \forall t \leq \xi
\end{aligned}$$

where $Z_1 \sim N_d(0,I)$ and

$$c(\beta) = \frac{\kappa - 1/\beta + 1}{2} \in (1/2, 1). \tag{32}$$

C. Asymptotics of the risk difference

In this subsection, using Theorem 2.2 with $f = m_\pi^{1/\beta}$, we investigate asymptotics of the risk difference

$$\begin{aligned}
\mathrm{diff}R_{\alpha,\mathrm{U},\pi} &= R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_{\mathrm{U}}(y\,|\,\cdot;\alpha)\} \\
&\quad - R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_\pi(y\,|\,\cdot;\alpha)\}
\end{aligned}$$

where $\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha)$ and $\hat{p}_\pi(y\,|\,x;\alpha)$ are given by (16) and (17), respectively. In the following theorem, we relate $\mathrm{diff}R_{\alpha,\mathrm{U},\pi}$ to $\mathrm{diff}R_{-1,\mathrm{U},\pi}$ given in (9) and $\mathrm{diff}R_{1,\mathrm{U},\pi}$ given in (8).

Theorem 2.4:

1) $\lim_{\alpha\to -1+0} \mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \mathrm{diff}R_{-1,\mathrm{U},\pi}$.
2) $\lim_{\alpha\to 1-0} \mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \mathrm{diff}R_{1,\mathrm{U},\pi}$.
3) $\lim_{v_x/v_y\to +0} \dfrac{v_y}{v_x}\mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \dfrac{v_y}{v_x}\mathrm{diff}R_{1,\mathrm{U},\pi}$.

Proof: [Part 1] Let $v_* = v_x v_y/(v_x+v_y)$. When $\alpha\to -1$ or equivalently $\beta\to 1$, we have

$$\gamma \to \frac{1}{1+v_x/v_y} = \frac{v_*}{v_x} \text{ and } \xi^2 \to \frac{v_x^2}{v_x+v_y} = v_x - v_*$$

and hence

$$\frac{2\gamma^{(1-\beta)d/2}}{\beta^2}\{m_\pi(w, v_x\gamma)\}^{1/\beta-1} \to 2, \tag{33}$$

which are parts of $\rho(w,z)$ given by (22). Further, in $\varrho(t;u)$ given by (23), we have

$$\begin{aligned} E_{Z_1}\left[m_\pi(tZ_1+u, v_x\gamma)\right] &= m_\pi(u, v_x\gamma + t^2) \\ &\to m_\pi(u, v_* + t^2). \end{aligned} \tag{34}$$

By (33) and (34), we have $\varrho(t;u)\to m_\pi^{1/2}(u, v_*+t^2)$ and

$$\begin{aligned} &E_Z[\rho(w,Z)] \\ &\to 4\int_0^{\sqrt{v_x-v_*}}\int_{\mathbb{R}^d} t\frac{-\Delta_u m_\pi^{1/2}(u,v_*+t^2)}{m_\pi^{1/2}(u,v_*+t^2)}\phi(u-w,t^2)\mathrm{d}u\mathrm{d}t \\ &= 2\int_0^{v_x-v_*}\int_{\mathbb{R}^d}\frac{-\Delta_u m_\pi^{1/2}(u,v_*+s)}{m_\pi^{1/2}(u,v_*+s)}\phi(u-w,s)\mathrm{d}u\mathrm{d}s. \end{aligned} \tag{35}$$

By (35), we have

$$\begin{aligned} &E_{W,Z}[\rho(W,Z)] \\ &\to 2\int_{\mathbb{R}^d}\mathrm{d}w\,\phi(w-\mu,v_*) \\ &\quad\times\left\{\int_0^{v_x-v_*}\int_{\mathbb{R}^d}\frac{-\Delta_u m_\pi^{1/2}(u,v_*+s)}{m_\pi^{1/2}(u,v_*+s)}\phi(u-w,s)\mathrm{d}u\mathrm{d}s\right\} \\ &= 2\int_0^{v_x-v_*}\mathrm{d}s \\ &\quad\times\left\{\int_{\mathbb{R}^d}\frac{-\Delta_u m_\pi^{1/2}(u,v_*+s)}{m_\pi^{1/2}(u,v_*+s)}\phi(u-\mu,v_*+s)\mathrm{d}u\right\} \\ &= 2\int_{v_*}^{v_x} E_Z\left[-\frac{\Delta_z m_\pi^{1/2}(Z,v)}{m_\pi^{1/2}(Z,v)}\right]\mathrm{d}v, \end{aligned}$$

where $Z\sim N_d(\mu, vI)$. The last equality follows from [5]'s result which was already explained in (9) of Section I. Hence we have

$$\lim_{\alpha\to -1+0}\mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \mathrm{diff}R_{-1,\mathrm{U},\pi}.$$

[Parts 2 and 3] Consider the asymptotic situation where

$$(1-\alpha)v_x/v_y \to 0 \Leftrightarrow \beta(v_x/v_y)\to 0 \Leftrightarrow \gamma\to 1.$$

Note that $E_Z[\rho(w,Z)]$ is rewritten as the product $\rho_1(w)\rho_2(w)$ where

$$\rho_1(w) = \frac{2\gamma^{(1-\beta)d/2}}{\beta^2}\{m_\pi(w,v_x\gamma)\}^{1/\beta-1}\xi^2,$$

$$\begin{aligned} \rho_2(w) &= \frac{2}{\xi^2}\int_0^{\xi} t\left\{\int_{\mathbb{R}^d}\frac{-\Delta_u\varrho(t;u)}{\varrho^{2/\beta-1}(t;u)}\phi(u-w,t^2)\mathrm{d}u\right\}\mathrm{d}t \\ &= \frac{1}{\xi^2}\int_0^{\xi^2}\left\{\int_{\mathbb{R}^d}\frac{-\Delta_u\varrho(\sqrt{s};u)}{\varrho^{2/\beta-1}(\sqrt{s};u)}\phi(u-w,s)\mathrm{d}u\right\}\mathrm{d}s. \end{aligned}$$

Since ξ^2 is rewritten as

$$\xi^2 = \frac{(1-\gamma)^2 v_y}{\gamma} = \left(\frac{1-\gamma}{\gamma}\right)^2 v_y\gamma = \frac{v_x^2}{v_y}\beta^2\gamma, \tag{36}$$

we have

$$\rho_1(w) = 2\frac{v_x^2}{v_y}\gamma^{(1-\beta)d/2+1}\{m_\pi(w,v_x\gamma)\}^{1/\beta-1}$$

and

$$\lim_{\gamma\to 1}\rho_1(w) = 2\frac{v_x^2}{v_y}\{m_\pi(w,v_x)\}^{1/\beta-1}. \tag{37}$$

When $\gamma\to 1$, we have $\xi^2\to 0$ by (36) and hence

$$\begin{aligned} \lim_{\gamma\to 1}\rho_2(w) &= \lim_{s\to 0}\left\{\int_{\mathbb{R}^d}\frac{-\Delta_u\varrho(\sqrt{s};u)}{\varrho^{2/\beta-1}(\sqrt{s};u)}\phi(u-w,s)\mathrm{d}u\right\} \\ &= \int_{\mathbb{R}^d}\lim_{s\to 0}\left(\frac{-\Delta_u\varrho(\sqrt{s};u)}{\varrho^{2/\beta-1}(\sqrt{s};u)}\right)\delta(u-w)\mathrm{d}u, \end{aligned} \tag{38}$$

where $\delta(\cdot)$ is the Dirac delta function. By (38) and

$$\begin{aligned} \lim_{\substack{s\to 0\\ \gamma\to 1}}\varrho(\sqrt{s};u) &= \left\{\int_{\mathbb{R}^d} m_\pi^{1/\beta}(u_1+u, v_x\gamma)\delta(u_1)\mathrm{d}u_1\right\}^{\beta/2} \\ &= m_\pi^{1/2}(u,v_x), \end{aligned}$$

we have

$$\lim_{\gamma\to 1}\rho_2(w) = \left(-\Delta_w m_\pi^{1/2}(w,v_x)\right)m_\pi^{1/2-1/\beta}(w,v_x). \tag{39}$$

By (37) and (39), we have

$$\begin{aligned} \lim_{\gamma\to 1}E_Z[\rho(w,Z)] &= \lim_{\gamma\to 1}\rho_1(w)\rho_2(w) \\ &= 2\frac{v_x^2}{v_y}\frac{-\Delta_w m_\pi^{1/2}(w,v_x)}{m_\pi^{1/2}(w,v_x)}, \end{aligned}$$

which implies that

$$\begin{aligned} &\lim_{\alpha\to 1}\mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \mathrm{diff}R_{1,\mathrm{U},\pi} \\ &= 2\frac{v_x^2}{v_y}E\left[\frac{-\Delta_w m_\pi^{1/2}(W,v_x)}{m_\pi^{1/2}(W,v_x)}\right], \end{aligned}$$

and

$$\begin{aligned} &\lim_{v_x/v_y\to 0}\frac{v_y}{v_x}\mathrm{diff}R_{\alpha,\mathrm{U},\pi} = \frac{v_y}{v_x}\mathrm{diff}R_{1,\mathrm{U},\pi} \\ &= 2v_x E\left[\frac{-\Delta_w m_\pi^{1/2}(W,v_x)}{m_\pi^{1/2}(W,v_x)}\right]. \end{aligned}$$

∎

Therefore the asymptotic situation $v_x/v_y\to 0$ corresponds to the case $\alpha\to 1$ and $\Delta_z\{m_\pi(z,v)\}^{1/2}$ plays an important role for general $\alpha\in(-1,1)$.

III. IMPROVEMENT UNDER THE HARMONIC PRIOR

Under the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$, let

$$m_{\mathrm{H}}(w, v) = \int_{\mathbb{R}^d} \phi(w-\mu, v)\pi_{\mathrm{H}}(\mu)\mathrm{d}\mu.$$

Let ν be an integer larger than or equal to 2. The superharmonicity related to $E_{Z_1}[m_{\mathrm{H}}^{\nu}(tZ_1+u, v)]$ with $Z_1 \sim N_d(0, I)$ is as follows.

Theorem 3.1: Let $c \in (0,1)$ and $Z_1 \sim N_d(0, I)$. Let ν be an integer larger than or equal to 2. Then, we have

$$\Delta_u \left\{E_{Z_1}\left[m_{\mathrm{H}}^{\nu}(tZ_1+u, v)\right]\right\}^{c/\nu} \le 0, \quad \forall u \in \mathbb{R}^d,$$

when

$$0 \le t \le \left(\frac{(d+2)(1-c)v}{d\nu(\nu-1)}\right)^{1/2}. \tag{40}$$

Proof: See Section B in Appendix. ■

When $1/\beta$ is an integer larger than or equal to 2, namely,

$$\alpha = 0, 1/3, 1/2, 3/5, 2/3, \dots,$$
$$\beta = 1/2, 1/3, 1/4, 1/5, 1/6, \dots,$$

let $\nu = 1/\beta$, $v = v_x\gamma$ and $c = 1/2$ in Theorem 3.1 and compare (40) in Theorem 3.1 with $0 \le t^2 \le \xi^2 = \beta^2 v_x^2\gamma/v_y$ in Theorem 2.2. If

$$\frac{\beta^2 v_x}{v_y} v_x\gamma \le \frac{(d+2)(1-c)}{d\nu(\nu-1)} v_x\gamma$$

or equivalently

$$\frac{v_x}{v_y} \le \frac{d+2}{d(1+\alpha)} = \frac{d+2}{2d(1-\beta)},$$

$m_{\mathrm{H}}(w, v_x\gamma)$ satisfies the sufficient condition of Theorem 2.2 and we have the following result of the Bayesian predictive density with respect to Stein's harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$, which is given by

$$\hat{p}_{\mathrm{H}}(y\,|\,x;\alpha) = \frac{m_{\mathrm{H}}^{1/\beta}(\gamma x + (1-\gamma)y, v_x\gamma)}{E_{Z_1}\left[m_{\mathrm{H}}^{1/\beta}(x+\xi Z_1, v_x\gamma)\right]}\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha).$$

Theorem 3.2: Suppose $2/(1-\alpha)$ is an positive integer for $\alpha \in (-1, 1)$. Suppose

$$\frac{v_x}{v_y} \le \frac{d+2}{d(1+\alpha)}. \tag{41}$$

Then, under α-div loss, the Bayesian predictive density $\hat{p}_{\mathrm{H}}(y\,|\,x;\alpha)$ with respect to the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ dominates the best invariant Bayesian predictive density $\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha) = \phi(y-x, v_y/\gamma)$.

Remark 3.1: For any $d \ge 3$ and $\alpha \in (-1, 1)$, we have

$$\frac{d+2}{d(1+\alpha)} > \frac{1}{2}.$$

Note that, in most typical situations,

$$\frac{v_x}{v_y} \le \frac{1}{2},$$

is easily assumed as follows. Suppose that we have a set of observations $x_1, \dots, x_n$ from $N_d(\mu, \sigma^2 I)$. An unobserved set $x_{n+1}, \dots, x_{n+m}$ from the same distribution is predicted by using a predictive density as a function of $x_1, \dots, x_n$. From sufficiency,

$$x = n^{-1}\sum\nolimits_{i=1}^{n} x_i \sim N_d(\mu, \sigma^2 I/n),$$
$$y = m^{-1}\sum\nolimits_{i=1}^{m} x_{n+i} \sim N_d(\mu, \sigma^2 I/m)$$

and clearly $v_x/v_y = m/n$ in this case. Since, m is typically 1 or 2 whereas n is relatively large, the condition (41) is satisfied.

When $1/\beta = 2/(1-\alpha)$ is not an integer, Theorem 2.3 can be applied. Let κ be the smallest integer greater than $1/\beta$. Suppose

$$\beta^2 \frac{v_x}{v_y} v_x\gamma \le \frac{(d+2)\{1-c(\beta)\}v_x\gamma}{d\kappa(\kappa-1)}, \tag{42}$$

where $c(\beta)$ is given by (32) as $c(\beta) = c(\{1-\alpha\}/2) = \{\kappa - 2/(1-\alpha)+1\}/2$, the left-hand side is the upper bound of t of Theorem 2.3 and the right-hand side is the upper bound of t of Theorem 3.1. When

$$\frac{v_x}{v_y} \le \left(\frac{2}{1-\alpha}\right)^2 \frac{d+2}{d}\frac{1-\{\kappa-2/(1-\alpha)\}}{2\kappa(\kappa-1)},$$

which is equivalent to (42), $m_{\mathrm{H}}(w, v_x\gamma)$ satisfies the sufficient condition of Theorem 2.3 and we have the following result.

Theorem 3.3: Suppose $2/(1-\alpha)$ is not an positive integer for $\alpha \in (-1, 1)$. Let κ be the smallest integer greater than $2/(1-\alpha)$. Suppose

$$\frac{v_x}{v_y} \le \left(\frac{2}{1-\alpha}\right)^2 \frac{d+2}{d}\frac{1-\{\kappa-2/(1-\alpha)\}}{2\kappa(\kappa-1)}. \tag{43}$$

Then the Bayesian predictive density $\hat{p}_{\mathrm{H}}(y\,|\,x;\alpha)$ with respect to the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ dominates the best invariant Bayesian predictive density $\hat{p}_{\mathrm{U}}(y \mid x;\alpha) = \phi(y - x, v_y/\gamma)$.

A. Discussion

By the definition of κ, we have

$$\kappa - 1 < \frac{2}{1-\alpha} < \kappa.$$

As $2/(1-\alpha) \uparrow \kappa$, the upper bound given by (43) approaches $(d+2)/\{d(1+\alpha)\}$ which is exactly the upper bound given by (41) of Theorem 3.2. On the other hand, as $2/(1-\alpha) \downarrow \kappa - 1$, the upper bound given by (43) approaches 0. Figure 1 gives a graph of behavior of the upper bound of v_x/v_y under $d = 4$ for improvement in Theorems 3.2 and 3.3. When $\alpha = -1$, for any ratio v_x/v_y, minimaxity has been established by [3], which is conformable to the limit, $\lim_{\alpha\to-1}(d+2)/\{d(1+\alpha)\} = \infty$. The undesirable discontinuity, found in Figure 1, is due to Jensen's inequality (31) which was not used in the proof of Theorem 2.2.

Figure 2 gives a graph of the risk difference

$$\begin{aligned} \mathrm{diff}R_{\alpha,\mathrm{U},\mathrm{H}} = {} & R_\alpha\left\{\phi(y-\mu, v_y) \,||\, \hat{p}_{\mathrm{U}}(y\,|\cdot;\alpha)\right\} \\ & - R_\alpha\left\{\phi(y-\mu, v_y) \,||\, \hat{p}_{\mathrm{H}}(y\,|\cdot;\alpha)\right\}, \end{aligned} \tag{44}$$

for $d = 4$, $v_y = 1$ and the following $16 = 4 \times 4$ combinations of α and v_x

$$\alpha = -\frac{1}{2}, 0, \frac{2}{5}, \frac{1}{2}, \quad v_x = b\frac{d+2}{d(1+\alpha)} \text{ with } b = \frac{1}{4}, 1, 4, 16.$$

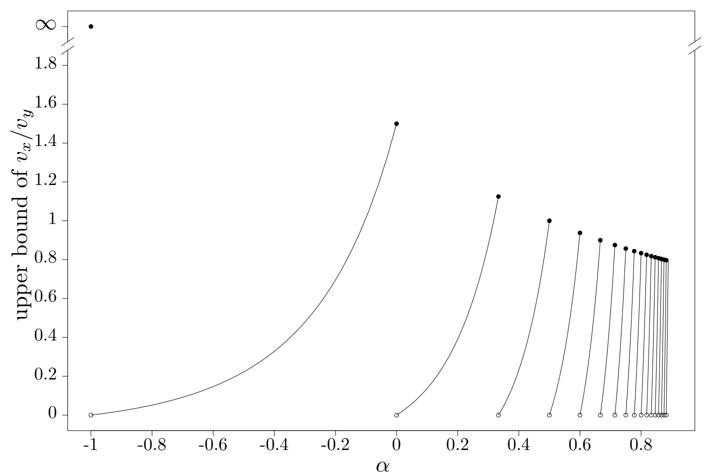

Fig. 1. The upper bound of v_x/v_y in Theorems 3.2 and 3.3 in the case $d = 4$

For each case, $\mathrm{diff} R_{\alpha,\mathrm{U},\mathrm{H}}$ is numerically calculated for $\|\mu\| = 0, 1, 2, 5, 10$. Details of the Monte Carlo calculation and the meaning of error bars in Figure 2 are explained in Appendix D and Remark 3.2 below, respectively. Note Theorem 3.2 guarantees minimaxity of $\hat{p}_{\mathrm{H}}(y \mid x; \alpha)$ only when $b \leq 1$ and $2/(1-\alpha) \in \mathbb{Z}$. As in Figure 2, $\mathrm{diff} R_{\alpha,\mathrm{U},\mathrm{H}}$ for large $\|\mu\|$ may be negative when $\alpha = 1/2$ is large. Hence, there is a possibility that the Bayesian predictive density $\hat{p}_{\mathrm{H}}(y \mid x; \alpha)$ with respect to the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ is not minimax when α and v_x/v_y are large. This phenomenon seems related to a recent work by Mukherjee and Johnstone [20] and hence we make a brief comment in Remark 3.3 below.

Note, $2/(1-\alpha) \notin \mathbb{Z}$ for $\alpha = -1/2, 2/5$, and $2/(1-\alpha) \in \mathbb{Z}$ for $\alpha = 0, 1/2$. Since $\mathrm{diff} R_{\alpha,\mathrm{U},\mathrm{H}}$ for $b = 1$ seems non-negative even for $\alpha = -1/2, 2/5$, and there is little difference among four cases with $b = 1$, we can naturally make a conjecture that the upper bound of v_x/v_y for improvement, $(d+2)/\{d(1+\alpha)\}$, of Theorem 3.2 is still valid even for $2/(1-\alpha) \notin \mathbb{Z}$. In order to prove it theoretically, the methodology for appropriately treating $E_{Z_1}\left[\{m_{\mathrm{H}}(tZ_1 + u, v_x\gamma)\}^{2/(1-\alpha)}\right]$ for $2/(1-\alpha) \notin \mathbb{Z}$ is needed and it remains an open problem.

Remark 3.2: In Figure 2, the numerical values of risk differences are presented with Monte Carlo error bars. Here, the error bar is defined as

$$[\bar{L}_T - s_T/\sqrt{T}, \bar{L}_T + s_T/\sqrt{T}]$$

where T is the Monte Carlo sample size,

$$\bar{L}_T = \frac{L_1 + \cdots + L_T}{T}, \text{ and } s_T^2 = \frac{L_1^2 + \cdots + L_T^2}{T} - \bar{L}_T^2.$$

Namely, the probability that the true value of risk is contained in the error bar is approximately 68% from the central limit theorem. Therefore, even if the upper bound of the error bar is smaller than zero, it does not immediately mean that the risk difference is actually negative.

Remark 3.3: In Remark 3.1, we discussed the ratio v_x/v_y. Here is also a remark related to the ratio v_x/v_y. Mukherjee and Johnstone [21], [20] considered estimating the predictive density under KullbackLeibler loss in an l_0 sparse Gaussian sequence model. [21] explicitly expressed the first order minimax risk along with its exact constant and derived, asymptotically least favorable priors and optimal predictive density estimates. Also [21] pointed out that the future-to-past variance ratio $r = v_y/v_x$ (Note that Theorem 3.2 is stated in terms of v_x/v_y) is an important parameter of the predictive estimation problem. The minimax risk increases as r decreases: we need to estimate the future observation density based on increasingly noisy past observations (in relative terms, $r = v_y/v_x$), and so the difficulty of the density estimation problem increases. In the same setting, [20] found proper Bayes predictive density with asymptotic minimaxity in sparse models. A big surprise is the existence of a phase transition in the future-to-past variance ratio. For smaller r, the natural discrete prior loses asymptotical optimality. Instead, for smaller r, a "bi-grid" prior recovering asymptotic minimaxity was proposed as an alternative.

In our case, Theorems 3.2 and 3.3 guarantee minimaxity under smaller v_x/v_y or equivalently larger r. When b is large in most graphs of Figure 2, the risk difference, $\mathrm{diff} R_{\alpha,\mathrm{U},\mathrm{H}}$ given by (44), is typically negative for larger $\|\mu\|$. Hence there is a possibility that the Bayesian predictive density $\hat{p}_{\mathrm{H}}(y \mid x; \alpha)$ with respect to the harmonic prior $\pi_{\mathrm{H}}(\mu) = \|\mu\|^{-(d-2)}$ is not minimax when v_x/v_y is large. However, taking the error bars of the risk difference into account (see also Remark 3.2), we cannot take sides whether there is a phase transition or not.

A possible direction for future research is to consider asymptotics of the risk difference as $v_x/v_y \to \infty$, which could not be successfully derived this time. If the asymptotic expression suggests an existence of a phase transition, natural directions for future research include the derivation of theoretical boundary of v_x/v_y of the phase transition and the proposal of an alternative with minimaxity when $\hat{p}_{\mathrm{H}}(y \mid x; \alpha)$ is not minimax. The phase transition is also related to Remark 3.4 below.

Remark 3.4: In the same problem setting, Ghosh *et al.* [22] considered minimaxity of the empirical Bayes predictive density given by

$$\hat{p}_\tau(y \mid x; \alpha) = \phi(y - \{\delta_\tau(x)\}, v_y + \beta v_x),$$

where

$$\delta_\tau(x) = \left(1 - \frac{\tau(\|x\|^2/v_x)}{\|x\|^2/v_x}\right) x.$$

They showed that the predictive density $\hat{p}(y \mid x; \tau)$ dominates the best equivariant predictive density

$$\hat{p}_{\mathrm{U}}(y \mid x; \alpha) = \phi(y - x, v_y + \beta v_x)$$

if the following two conditions on τ are satisfied;

$$\begin{cases} \tau(t) \in (0, 2(d-2)) \\ \tau(t) \text{ is differentiable nondecreasing in } t. \end{cases} \tag{45}$$

In the estimation problem, this type of the sufficient condition, (45), for improvement on the best equivariant procedure is known as Baranchik condition [23]. Interestingly there is no restriction on v_x/v_y in (45), or equivalently there is no phase transition in [20]'s sense.

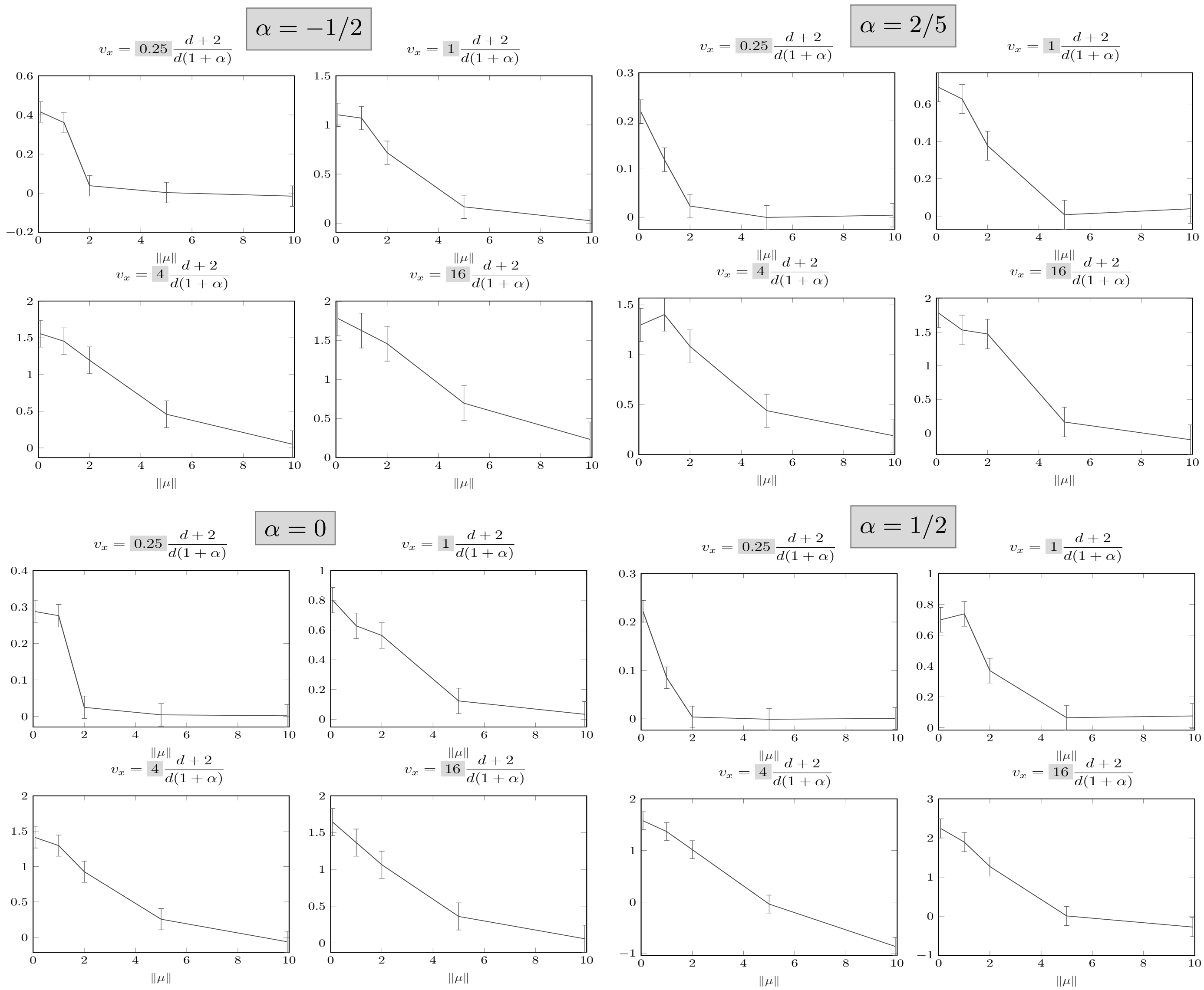

Fig. 2. Risk differences for $\alpha = -1/2, 0, 2/5, 1/2$

APPENDIX A
MINIMAXITY OF $\hat{p}_{\mathrm{U}}(y \mid x; \alpha)$

In this section, we show that

$$\hat{p}_{\mathrm{U}}(y \mid x; \alpha) = \phi(y - x, v_y/\gamma) = \phi(y - x, v_y + \beta v_x)$$

is minimax, by following Sections II and III of [2]. We start with the definition of invariance under location shift.

Definition A.1: A predictive density $\hat{p}(y \mid x)$ is invariant under location shift, if for all $a \in \mathbb{R}^d$ and all x, y, $\hat{p}(y + a \mid x + a) = \hat{p}(y \mid x)$.

Hence any invariant predictive density should be of the form

$$\hat{p}(y \mid x) = q(y - x)$$

which satisfies

$$\int_{\mathbb{R}^d} q(y)\mathrm{d}y = 1.$$

Clearly $\hat{p}_{\mathrm{U}}(y \mid x; \alpha)$ is invariant under location shift. Note that invariant procedures have constant risk since the risk of the invariant predictive density $q(y - x)$ is

$$\begin{aligned} & R_\alpha\{\phi(y-\mu, v_y) \mid\mid q(y - \cdot)\} \\ & = \int_{\mathbb{R}^d} \phi(x - \mu, v_x) \\ & \quad \times \left(\int_{\mathbb{R}^d} f_\alpha \left(\frac{q(y-x)}{\phi(y-\mu, v_y)} \right) \phi(y - \mu, v_y) \mathrm{d}y \right) \mathrm{d}x \\ & = \int_{\mathbb{R}^d} \phi(z_x, v_x) \\ & \quad \times \left(\int_{\mathbb{R}^d} f_\alpha \left(\frac{q(z_y - z_x)}{\phi(z_y, v_y)} \right) \phi(z_y, v_y) \mathrm{d}z_y \right) \mathrm{d}z_x \end{aligned} \tag{46}$$

where $z_x = x - \mu$ and $z_y = y - \mu$, which does not depend on μ. More specifically, the risk of the invariant predictive density $q(y - x)$ is as follows.

Lemma A.1: The risk of an invariant predictive density $q(y-$

x) is

$$\begin{aligned}&R_\alpha\{\phi(y-\mu,v_y)\,||\,q(y-\cdot)\}\\&=\frac{1-\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}+\gamma^{(1-\beta)d/2}D_\alpha\left\{\phi(z,v_y/\gamma)\,||\,q(z)\right\}.\end{aligned}\tag{47}$$

Proof: By (46) and the definition of α-div loss,

$$\begin{aligned}&\beta(1-\beta)R_\alpha\{\phi(y-\mu,v_y)\,||\,q(y-\cdot)\}\\&=1-\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}q^{1-\beta}(y-x)\phi^\beta(y,v_y)\phi(x,v_x)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

By the identity (15) with $\mu=0$, we have

$$\begin{aligned}&\phi\left(x,v_x\right)\phi^\beta(y,v_y)\\&=\gamma^{(1-\beta)d/2}\phi(\gamma x+(1-\gamma)y,v_x\gamma)\phi^\beta(y-x,v_y/\gamma),\end{aligned}$$

and hence

$$\begin{aligned}&\beta(1-\beta)R_\alpha\{\phi(y-\mu,v_y)\,||\,q(y-\cdot)\}\\&=1-\gamma^{(1-\beta)d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}q^{1-\beta}(y-x)\\&\quad\times\phi^\beta(y-x,v_y/\gamma)\phi(\gamma x+(1-\gamma)y,v_x\gamma)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

By the change of variables,

$$\begin{pmatrix}w\\z\end{pmatrix}=\begin{pmatrix}\gamma I_d&(1-\gamma)I_d\\-I_d&I_d\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

where Jacobian of the matrix is 1, we have

$$\begin{aligned}&R_\alpha\{\phi(y-\mu,v_y)\,||\,q(y-\cdot)\}\\&=\frac{1}{\beta(1-\beta)}\Big\{1-\gamma^{(1-\beta)d/2}\\&\quad\times\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}q^{1-\beta}(z)\phi^\beta(z,v_y/\gamma)\phi(w,v_x\gamma)\mathrm{d}z\mathrm{d}w\Big\}\\&=\frac{1}{\beta(1-\beta)}\left\{1-\gamma^{(1-\beta)d/2}\int_{\mathbb{R}^d}q^{1-\beta}(z)\phi^\beta(z,v_y/\gamma)\mathrm{d}z\right\}\\&=\frac{1-\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}+\gamma^{(1-\beta)d/2}D_\alpha\left\{\phi(z,v_y/\gamma)\,||\,q(z)\right\}.\end{aligned}$$

■

In (47) of Lemma A.1, $D_\alpha\left\{\phi(z,v_y/\gamma)\,||\,q(z)\right\}$ is non-negative and equals zero if and only if $q(z)=\phi(z,v_y/\gamma)$. Hence the best invariant procedure is $\hat{p}_{\mathrm{U}}(y\mid x;\alpha)=\phi(y-x,v_y/\gamma)$, where the constant risk is

$$\frac{1-\gamma^{(1-\beta)d/2}}{\beta(1-\beta)}.$$

Since the risk is constant for invariant predictive density, the best invariant $\hat{p}_{\mathrm{U}}(y\mid x;\alpha)$ is the minimax procedure among all invariant procedures. If a constant risk procedure is shown to have an extended Bayes property defined below, then it is, in fact, minimax over all procedures. See Theorem 5.18 of Berger [24] and Theorem 5.1.12 of Lehmann and Casella [25] for the detail.

Definition A.2: A predictive procedure $\hat{p}_*(y\mid x)$ is called extended Bayes, if there exists a sequence of Bayes procedures $\hat{p}_{\pi_c}(y\,|\,x;\alpha)$ with proper prior densities $\pi_c(\mu)$ for $c=1,\dots,$ such that their Bayes risk differences go to zero, that is,

$$\begin{aligned}&\lim_{c\to\infty}\left(\int_{\mathbb{R}^d}R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_*(y\,|\,\cdot)\}\pi_c(\mu)\mathrm{d}\mu\right.\\&\qquad\left.-\int_{\mathbb{R}^d}R_\alpha\{\phi(y-\mu,v_y)\,||\,\hat{p}_{\pi_c}(y\,|\,\cdot;\alpha)\}\pi_c(\mu)\mathrm{d}\mu\right)=0.\end{aligned}$$

Recall that

$$\hat{p}_\pi(y\,|\,x;\alpha)\propto\left\{\int_{\mathbb{R}^d}\phi^\beta(y-\mu,v_y)\phi(x-\mu,v_x)\pi(\mu)\mathrm{d}\mu\right\}^{1/\beta}\tag{48}$$

for $\beta=(1-\alpha)/2$ and $\alpha\in(-1,1)$. Under the prior $\mu\sim N_d(0,\{cv_x\gamma\}I)$ with the density $\pi_c(\mu)=\phi(\mu,cv_x\gamma)$, we have the identity

$$\begin{aligned}&\phi^\beta(y-\mu,v_y)\phi\left(x-\mu,v_x\right)\phi(\mu,cv_x\gamma)\\&=\left(\frac{1+c\gamma}{1+c}\right)^{d(1-\beta)/2}\phi\left(\mu-c\frac{\gamma x+(1-\gamma)y}{1+c},\frac{cv_x\gamma}{1+c}\right)\\&\quad\times\phi^\beta\left(y-\frac{c\gamma x}{1+c\gamma},v_y\frac{1+c}{1+c\gamma}\right)\phi\left(x,v_x(1+c\gamma)\right)\end{aligned}\tag{49}$$

and hence

$$\begin{aligned}&\left\{\int_{\mathbb{R}^d}\phi^\beta(y-\mu,v_y)\phi(x-\mu,v_x)\pi(\mu)\mathrm{d}\mu\right\}^{1/\beta}\\&=\left\{\left(\frac{1+c\gamma}{1+c}\right)^{d(1-\beta)/2}\phi\left(x,v_x(1+c\gamma)\right)\right\}^{1/\beta}\\&\quad\times\phi\left(y-\frac{c\gamma x}{1+c\gamma},v_y\frac{1+c}{1+c\gamma}\right).\end{aligned}\tag{50}$$

By (48) and (50), the Bayesian solution is

$$\hat{p}_{\pi_c}(y\,|\,x;\alpha)=\phi\left(y-\frac{c\gamma}{1+c\gamma}x,v_y\frac{1+c}{1+c\gamma}\right).$$

Furthermore, by the identity (49), the product of $\beta(1-\beta)$ and the Bayes risk of $\hat{p}_{\pi_c}(y\,|\,x;\alpha)$, is given by

$$\begin{aligned}&1-\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left\{\frac{\hat{p}_{\pi_c}(y\,|\,x;\alpha)}{\phi(y-\mu,v_y)}\right\}^{1-\beta}\\&\quad\times\phi\left(x-\mu,v_x\right)\phi(y-\mu,v_y)\phi(\mu,cv_x\gamma)\mathrm{d}x\mathrm{d}y\mathrm{d}\mu\\&=1-\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{d}x\mathrm{d}y\mathrm{d}\mu\,\phi\left(x-\mu,v_x\right)\phi(\mu,cv_x\gamma)\\&\quad\times\phi^{1-\beta}\left(y-\frac{c\gamma}{1+c\gamma}x,v_y\frac{1+c}{1+c\gamma}\right)\phi^\beta(y-\mu,v_y)\\&=1-\left(\frac{1+c\gamma}{1+c}\right)^{d(1-\beta)/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{d}\mu\mathrm{d}y\mathrm{d}x\\&\quad\times\phi\left(\mu-c\frac{\gamma x+(1-\gamma)y}{1+c},\frac{cv_x\gamma}{1+c}\right)\\&\quad\times\phi\left(y-\frac{c\gamma x}{1+c\gamma},v_y\frac{1+c}{1+c\gamma}\right)\phi\left(x,v_x(1+c\gamma)\right)\\&=1-\left(\frac{1+c\gamma}{1+c}\right)^{d(1-\beta)/2},\end{aligned}$$

which approaches $1-\gamma^{(1-\beta)d/2}$ as c goes to infinity. Hence $\hat{p}_{\mathrm{U}}(y\,|\,x;\alpha)$ is extended Bayes and hence minimax.

Appendix B
Proof of Theorem 3.1

Recall the identity

$$\|\mu\|^{-(d-2)}=b\int_0^\infty g^{d/2-2}\exp\left(-g\frac{\|\mu\|^2}{2v}\right)\mathrm{d}g$$

for any $v>0$, where $b=1/\{\Gamma(d/2-1)2^{d/2-1}v^{d/2-1}\}$. Then we have

$$\begin{aligned}
&m_{\mathrm{H}}(w,v)\\
&=\int_{\mathbb{R}^d}\phi(w-\mu,v)\|\mu\|^{-(d-2)}\mathrm{d}\mu\\
&=b\int_0^\infty g^{d/2-2}\mathrm{d}g\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}v^{d/2}}\\
&\quad\times\exp\left(-\frac{\|w-\mu\|^2}{2v}-g\frac{\|\mu\|^2}{2v}\right)\mathrm{d}\mu \qquad (51)\\
&=b\int_0^\infty\frac{g^{d/2-2}}{(1+g)^{d/2}}\exp\left(-\frac{g\|w\|^2}{2(g+1)v}\right)\mathrm{d}g\\
&=b\int_0^1\lambda^{d/2-2}\exp\left(-\frac{\lambda\|w\|^2}{2v}\right)\mathrm{d}\lambda,
\end{aligned}$$

where the third equality is from the relation of completing squares with respect to μ

$$\begin{aligned}
&\|w-\mu\|^2+g\|\mu\|^2\\
&=(g+1)\|\mu-w/(g+1)\|^2+\{g/(g+1)\}\|w\|^2
\end{aligned}$$

and the fourth equality is from the transformation $\lambda=g/(g+1)$.

Note that $m^\nu_{\mathrm{H}}(w,v)$ for a positive integer ν is expressed as

$$\begin{aligned}
&m^\nu_{\mathrm{H}}(w,v)\\
&=b^\nu\int_{\mathcal{D}_\nu}\prod_{i=1}^{\nu}\lambda_i^{d/2-2}\exp\left(-\frac{\sum_{i=1}^\nu\lambda_i\|w\|^2}{2v}\right)\prod\mathrm{d}\lambda_i,
\end{aligned}$$

where $\mathcal{D}_\nu$ is ν-dimensional unit hyper-cube. In the following, $\mathrm{d}\lambda$ denotes $\prod_{i=1}^\nu\mathrm{d}\lambda_i$ for notational simplicity. Furthermore the subscript and superscript of $\prod$ and $\sum$ is omitted for simplicity if they are $i=1$ and $i=\nu$ respectively. Hence $m^\nu_{\mathrm{H}}(w,v)$ in the above is written as

$$m^\nu_{\mathrm{H}}(w,v)=b^\nu\int_{\mathcal{D}_\nu}\prod\lambda_i^{d/2-2}\exp\left(-\frac{\sum\lambda_i\|w\|^2}{2v}\right)\mathrm{d}\lambda.$$

For the calculation of

$$E_{Z_1}\left[m^\nu_{\mathrm{H}}(tZ_1+u,v)\right]=\int_{\mathbb{R}^d}m^\nu_{\mathrm{H}}(x+u,v)\phi(x,t^2)\mathrm{d}x$$

under $Z_1\sim N_d(0,I)$, note the relation of completing squares with respect to x,

$$\begin{aligned}
&\frac{(\sum\lambda_i)\|x+u\|^2}{v}+\frac{\|x\|^2}{t^2}=\frac{1}{v}\left\{\sum\lambda_i\|x+u\|^2+s\|x\|^2\right\}\\
&=\frac{1}{v}\left\{\left(\sum\lambda_i+s\right)\left\|x+\frac{\sum\lambda_i}{\sum\lambda_i+s}u\right\|^2+\frac{s\sum\lambda_i}{\sum\lambda_i+s}\|u\|^2\right\},
\end{aligned} \qquad (52)$$

where $s=v/t^2$. Then, by (52), we have

$$\begin{aligned}
&\frac{t^d}{b^\nu v^{d/2}}E_{Z_1}[m^\nu_{\mathrm{H}}(tZ_1+u,v)]\\
&=\int_{\mathcal{D}_\nu}\frac{\prod\lambda_i^{d/2-2}}{(\sum\lambda_i+s)^{d/2}}\exp\left(-\frac{s\sum\lambda_i}{v(\sum\lambda_i+s)}\frac{\|u\|^2}{2}\right)\mathrm{d}\lambda.
\end{aligned}$$

Re-define $u:=\{s/v\}^{1/2}u$ and let

$$\begin{aligned}
&\psi(u;\nu,s)\\
&=\int_{\mathcal{D}_\nu}\frac{\prod\lambda_i^{d/2-2}}{(\sum\lambda_i+s)^{d/2}}\exp\left(-\frac{\sum\lambda_i}{\sum\lambda_i+s}\frac{\|u\|^2}{2}\right)\mathrm{d}\lambda.
\end{aligned} \qquad (53)$$

By (29), the super-harmonicity of $\{E_{Z_1}\left[m^\nu_{\mathrm{H}}(tZ_1+u,v)\right]\}^{c/\nu}$ with respect to $u\in\mathbb{R}^d$ is equivalent to

$$\left(\frac{c}{\nu}-1\right)\|\nabla_u\psi\|^2+\psi\Delta_u\psi\le 0,\quad\forall u\in\mathbb{R}^d.$$

The integrand of ψ given by (53) is denoted by

$$\begin{aligned}
\zeta(\lambda)&=\zeta(\lambda_1,\dots,\lambda_\nu)\\
&=\frac{\prod\lambda_i^{d/2-2}}{(\sum\lambda_i+s)^{d/2}}\exp\left(-\frac{\sum\lambda_i}{\sum\lambda_i+s}z\right)
\end{aligned}$$

where $z=\|u\|^2/2$. Then we have

$$\frac{\partial}{\partial u_j}\psi=-u_j\int\zeta(\lambda)\frac{\sum\lambda_i}{\sum\lambda_i+s}\mathrm{d}\lambda,$$

for $j=1,\dots,d$ and

$$\frac{\partial^2}{\partial u_j^2}\psi=\int\zeta(\lambda)\left\{-\frac{\sum\lambda_i}{\sum\lambda_i+s}+u_j^2\left(\frac{\sum\lambda_i}{\sum\lambda_i+s}\right)^2\right\}\mathrm{d}\lambda.$$

Noting $z=\|u\|^2/2$, we have

$$\begin{aligned}
\|\nabla_u\psi\|^2&=2z\left(\int\zeta(\lambda)\frac{\sum\lambda_i}{\sum\lambda_i+s}\mathrm{d}\lambda\right)^2\\
&=2\nu^2z\left(\int\zeta(\lambda)\frac{\lambda_1}{\sum\lambda_i+s}\mathrm{d}\lambda\right)^2
\end{aligned} \qquad (54)$$

and

$$\begin{aligned}
&\Delta_u\psi\\
&=-d\int\zeta(\lambda)\frac{\sum\lambda_i}{\sum\lambda_i+s}\mathrm{d}\lambda+2z\int\zeta(\lambda)\left(\frac{\sum\lambda_i}{\sum\lambda_i+s}\right)^2\mathrm{d}\lambda\\
&=-d\nu\int\zeta(\lambda)\frac{\lambda_1}{\sum\lambda_i+s}\mathrm{d}\lambda+2\nu z\int\zeta(\lambda)\frac{\lambda_1^2}{(\sum\lambda_i+s)^2}\mathrm{d}\lambda\\
&\quad+2\nu(\nu-1)z\int\zeta(\lambda)\frac{\lambda_1\lambda_2}{(\sum\lambda_i+s)^2}\mathrm{d}\lambda.
\end{aligned} \qquad (55)$$

In (54) and (55), the second equalities are from symmetry with respect to λ_i's.

Let

$$\begin{aligned}
\rho(j_1,j_2,l)&=\int_{\mathcal{D}_\nu}\lambda_1^{j_1}\lambda_2^{j_2}\left(\sum\lambda_i+s\right)^l\zeta(\lambda)\mathrm{d}\lambda,\\
\eta(j_2,l)&=\int_{\mathcal{D}_{\nu-1}}\lambda_2^{j_2}\left(1+\sum\nolimits_{i=2}\lambda_i+s\right)^l\\
&\quad\times\zeta(1,\lambda_2,\dots,\lambda_\nu)\prod_{i=2}\mathrm{d}\lambda_i,
\end{aligned}$$

where j_1 and j_2 are nonnegative integers. Then $\|\nabla_u\psi\|^2$ and $\Delta_u\psi$ given by (54) and (55) is rewritten as

$$\begin{aligned}
\|\nabla_u\psi\|^2&=2\nu^2z\rho(1,0,-1)^2,\\
\Delta_u\psi&=-d\nu\rho(1,0,-1)+2\nu z\rho(2,0,-2)\\
&\quad+2\nu(\nu-1)z\rho(1,1,-2).
\end{aligned} \qquad (56)$$

Here are some useful relationships and inequalities.

Lemma B.1:

$$\begin{aligned}
&sz\rho(j_1,j_2,l)\\
&=-\eta(j_2,l+2)+(j_1+d/2-2)\rho(j_1-1,j_2,l+2)
\end{aligned}$$

$$+ (l - d/2 + 2)\rho(j_1, j_2, l+1) \qquad \text{for } j_1 \geq 1, \tag{57}$$
$$\rho(0,0,l) = \nu\rho(1,0,l-1) + s\rho(0,0,l-1), \tag{58}$$
$$\begin{aligned}\rho(1,0,l) &= \rho(2,0,l-1) + (\nu-1)\rho(1,1,l-1)\\ &\quad + s\rho(1,0,l-1),\end{aligned} \tag{59}$$
$$\eta(0,1) = \eta(0,0) + (\nu-1)\eta(1,0) + s\eta(0,0), \tag{60}$$
$$\eta(0,1)\rho(0,0,-1) \geq \eta(0,0)\rho(0,0,0), \tag{61}$$
$$\frac{\rho(1,0,-1)}{\rho(1,0,0)} \geq \frac{1}{\nu d/(d+2) + s}. \tag{62}$$

Proof: See Section C in Appendix. ∎

Applying the identity (57) to $\|\nabla_u \psi\|^2$ and $\Delta_u \psi$ given in (56), we have

$$\begin{aligned} s\|\nabla_u\psi\|^2 &= 2\nu^2\{sz\rho(1,0,-1)\}\rho(1,0,-1)\\ &= \nu^2\rho(1,0,-1)\\ &\quad \times\{-2\eta(0,1) + (d-2)\rho(0,0,1) - (d-2)\rho(1,0,0)\}\end{aligned}$$

and

$$\begin{aligned} &s\Delta_u\psi\\ &= \nu(\nu-1)\{-2\eta(1,0) + (d-2)\rho(1,0,0) - d\rho(1,1,-1)\}\\ &\quad + \nu\{-2\eta(0,0) + d\rho(1,0,0) - d\rho(2,0,-1)\}\\ &\quad - d\nu s\rho(1,0,-1)\\ &= \nu(\nu-1)(d-2)\rho(1,0,0) - 2\nu\{\eta(0,0) + (\nu-1)\eta(1,0)\},\end{aligned}$$

where the second equality of $s\Delta_u\psi$ follows from (59). Then we have

$$\begin{aligned} &\frac{s}{\nu}\left(\frac{c-\nu}{\nu}\|\nabla_u\psi\|^2 + \psi\Delta_u\psi\right)\\ &\quad = (\nu-c)\rho(1,0,-1)\\ &\qquad \times[2\eta(0,1) - (d-2)\{\rho(0,0,1) - \rho(1,0,0)\}]\\ &\qquad - 2\{\eta(0,0) + (\nu-1)\eta(1,0)\}\rho(0,0,0)\\ &\qquad + (\nu-1)(d-2)\rho(1,0,0)\rho(0,0,0).\end{aligned} \tag{63}$$

By applying (58), (60) and (61), the terms of (63) including $\eta(\cdot,\cdot)$, divided by 2, is

$$\begin{aligned} &(\nu-c)\eta(0,1)\rho(1,0,-1)\\ &\quad - \{\eta(0,0) + (\nu-1)\eta(1,0)\}\rho(0,0,0)\\ &= (\nu-c)\eta(0,1)\rho(1,0,-1) - \{\eta(0,1) - s\eta(0,0)\}\rho(0,0,0)\\ &= (\nu-c)\eta(0,1)\rho(1,0,-1) + s\eta(0,0)\rho(0,0,0)\\ &\quad - \eta(0,1)\{\nu\rho(1,0,-1) + s\rho(0,0,-1)\}\\ &= -c\eta(0,1)\rho(1,0,-1)\\ &\quad - s\{\eta(0,1)\rho(0,0,-1) - \eta(0,0)\rho(0,0,0)\}\\ &\leq 0,\end{aligned} \tag{64}$$

where the first equality follows from (60), the second equality follows from (58) and the inequality follows from (61).

The terms of (63) not including $\eta(\cdot,\cdot)$, divided by $(d-2)$, are rewritten as

$$\begin{aligned} &(\nu-c)\{-\rho(0,0,1) + \rho(1,0,0)\}\rho(1,0,-1)\\ &\quad + (\nu-1)\rho(1,0,0)\rho(0,0,0)\\ &= -(\nu-c)(\nu-1)\rho(1,0,0)\rho(1,0,-1)\\ &\quad - (\nu-c)s\rho(0,0,0)\rho(1,0,-1)\\ &\qquad + (\nu-1)\rho(1,0,0)\rho(0,0,0)\\ &\leq -\left\{\frac{(\nu-c)s}{\nu d/(d+2) + s} - (\nu-1)\right\}\rho(1,0,0)\rho(0,0,0)\\ &= -\frac{(1-c)s - \nu(\nu-1)d/(d+2)}{\nu d/(d+2) + s}\rho(1,0,0)\rho(0,0,0),\end{aligned} \tag{65}$$

which is nonpositive for $s \geq \nu(\nu-1)d/\{(1-c)(d+2)\}$, where the first equality follows from (58) and the inequality follows from (62).

By (64) and (65), we have

$$\left(\frac{c}{\nu} - 1\right)\|\nabla_u\psi\|^2 + \psi\Delta_u\psi \leq 0, \quad \forall u \in \mathbb{R}^d$$

or equivalently

$$\Delta_u\left\{E_{Z_1}\left[m_{\mathrm{H}}^{\nu}(tZ_1 + u, v)\right]\right\}^{c/\nu} \leq 0, \quad \forall u \in \mathbb{R}^d,$$

when $t \leq \{(d+2)(1-c)v/\{d\nu(\nu-1)\}\}^{1/2}$.

Appendix C
Proof of Lemma B.1

[Part of (57)] Note

$$\begin{aligned} &\frac{\partial}{\partial\lambda_1}\exp\left(-\frac{z\sum\lambda_i}{\sum\lambda_i + s}\right)\\ &\quad = -\frac{sz}{(\sum\lambda_i + s)^2}\exp\left(-\frac{z\sum\lambda_i}{\sum\lambda_i + s}\right).\end{aligned}$$

Then, by an integration by parts, we have

$$\begin{aligned} &sz\int_0^1 \lambda_1^{j_1}\lambda_2^{j_2}(\textstyle\sum\lambda_i + s)^l\zeta(\lambda)\mathrm{d}\lambda_1\\ &= -\lambda_2^{d/2-2+j_2}\prod_{i=3}\lambda_i^{d/2-2}\int_0^1 \mathrm{d}\lambda_1\,\lambda_1^{d/2-2+j_1}\\ &\quad \times(\textstyle\sum\lambda_i + s)^{l-d/2+2}\left\{\dfrac{\partial}{\partial\lambda_1}\exp\left(-\dfrac{z\sum\lambda_i}{\sum\lambda_i + s}\right)\right\}\\ &= -\lambda_2^{d/2-2+j_2}\prod_{i=3}\lambda_i^{d/2-2}\\ &\quad \times\left\{\left[\frac{\lambda_1^{d/2-2+j_1}}{(\sum\lambda_i + s)^{-l+d/2-2}}\exp\left(-\frac{z\sum\lambda_i}{\sum\lambda_i + s}\right)\right]_0^1\right.\\ &\quad -(d/2-2+j_1)\int_0^1\lambda_1^{d/2-3+j_1}(\textstyle\sum\lambda_i + s)^{l-d/2+2}\\ &\quad \times\exp\left(-\dfrac{z\sum\lambda_i}{\sum\lambda_i + s}\right)\mathrm{d}\lambda_1\\ &\quad -(l-d/2+2)\int_0^1\lambda_1^{d/2-2+j_1}(\textstyle\sum\lambda_i + s)^{l-d/2+1}\\ &\quad \left.\times\exp\left(-\dfrac{z\sum\lambda_i}{\sum\lambda_i + s}\right)\mathrm{d}\lambda_1\right\}.\end{aligned}$$

(57) follows from integration with respect to $\lambda_2, \ldots, \lambda_\nu$ in the both hand side of the above equality.

[Parts of (58), (59) and (60)] The equalities (58), (59) and (60) easily follows from symmetry with respect to λ_i's.

[Part of (61)] Note that (61) is equivalent to

$$\begin{aligned}&\eta(0,0)\rho(0,0,0)-\eta(0,1)\rho(0,0,-1)\\&=\{\rho(0,0,0)-\rho(0,0,-1)\}\eta(0,1)\\&\quad-\{\eta(0,1)-\eta(0,0)\}\rho(0,0,0)\\&=\int_{\mathcal{D}_{\nu-1}}f_1(\lambda_2,\dots,\lambda_\nu)\prod_{i=2}\mathrm{d}\lambda_i\int_{\mathcal{D}_{\nu-1}}f_2(\xi_2,\dots,\xi_\nu)\prod_{i=2}\mathrm{d}\xi_i\\&\quad-\int_{\mathcal{D}_{\nu-1}}f_3(\lambda_2,\dots,\lambda_\nu)\prod_{i=2}\mathrm{d}\lambda_i\int_{\mathcal{D}_{\nu-1}}f_4(\xi_2,\dots,\xi_\nu)\prod_{i=2}\mathrm{d}\xi_i\\&\le 0,\end{aligned}$$

where

$$\begin{aligned}f_1(\lambda_2,\dots,\lambda_\nu)&=\int_0^1\left(1-\frac{1}{\sum\lambda_i+s}\right)\zeta(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda_1\\f_2(\xi_2,\dots,\xi_\nu)&=(1+\sum\nolimits_{i=2}\xi_i+s)\zeta(1,\xi_2,\dots,\xi_\nu)\\f_3(\lambda_2,\dots,\lambda_\nu)&=(\sum\nolimits_{i=2}\lambda_i+s)\zeta(1,\lambda_2,\dots,\lambda_\nu)\\f_4(\xi_2,\dots,\xi_\nu)&=\int_0^1\zeta(\xi_1,\dots,\xi_\nu)\mathrm{d}\xi_1.\end{aligned}$$

Since both $1-1/(\sum\lambda_i+s)$ and $\sum\lambda_i+s$ are increasing in each of its arguments, we have

$$\begin{aligned}&\left\{1-1/\left(\sum\lambda_i+s\right)\right\}(1+\sum\nolimits_{i=2}\xi_i+s)\\&\le\left\{1-\frac{1}{(\lambda_1\vee1)+\sum_{i=2}(\lambda_i\vee\xi_i)+s}\right\}\\&\quad\times\left\{(\lambda_1\vee1)+\sum\nolimits_{i=2}(\lambda_i\vee\xi_i)+s\right\}\\&=\sum\nolimits_{i=2}(\lambda_i\vee\xi_i)+s,\end{aligned}\tag{66}$$

where $\vee$ is the maximum operator, i.e. $\lambda_i\vee\xi_i=\max(\lambda_i,\xi_i)$. In the following, $\wedge$ denotes the minimum operator, i.e. $\lambda_i\wedge\xi_i=\min(\lambda_i,\xi_i)$. Note that a function h: $\mathbb{R}^\nu\to\mathbb{R}$ is said to be multivariate totally positive of order two (MTP2) if it satisfies

$$\begin{aligned}&h(x_1,\dots,x_\nu)h(y_1,\dots,y_\nu)\\&\le h(x_1\vee y_1,\dots,x_\nu\vee y_\nu)h(x_1\wedge y_1,\dots,x_\nu\wedge y_\nu)\end{aligned}$$

for any $x,y\in\mathbb{R}^\nu$. By Lemma C.1 below, $\zeta(\lambda_1,\dots,\lambda_\nu)$ is MTP2 as a function of ν-variate function and hence the inequality

$$\begin{aligned}&\zeta(\lambda_1,\dots,\lambda_\nu)\zeta(1,\xi_2,\dots,\xi_\nu)\\&\le\zeta(\lambda_1\vee1,\lambda_2\vee\xi_2,\dots,\lambda_\nu\vee\xi_\nu)\\&\quad\times\zeta(\lambda_1\wedge1,\lambda_2\wedge\xi_2,\dots,\lambda_\nu\wedge\xi_\nu)\\&=\zeta(1,\lambda_2\vee\xi_2,\dots,\lambda_\nu\vee\xi_\nu)\\&\quad\times\zeta(\lambda_1,\lambda_2\wedge\xi_2,\dots,\lambda_\nu\wedge\xi_\nu)\end{aligned}\tag{67}$$

follows. By (66) and (67), we have

$$\begin{aligned}&f_1(\lambda_2,\dots\lambda_\nu)f_2(\xi_2,\dots\xi_\nu)\\&\le\int_0^1\Big[\left\{\sum\nolimits_{i=2}(\lambda_i\vee\xi_i)+s\right\}\zeta(1,\lambda_2\vee\xi_2,\dots,\lambda_\nu\vee\xi_\nu)\\&\qquad\times\zeta(\lambda_1,\lambda_2\wedge\xi_2,\dots,\lambda_\nu\wedge\xi_\nu)\Big]\mathrm{d}\lambda_1\\&=f_3(\lambda_2\vee\xi_2,\dots,\lambda_\nu\vee\xi_\nu)f_4(\lambda_2\wedge\xi_2,\dots,\lambda_\nu\wedge\xi_\nu).\end{aligned}$$

From Theorem C.1 below, shown by Karlin and Rinott [26], the theorem follows.

[Part of (62)] By Jensen's inequality, we have

$$\begin{aligned}\frac{\rho(1,0,-1)}{\rho(1,0,0)}&=\int\frac{1}{\lambda_1+\sum_{i=2}^\nu\lambda_i+s}\frac{\lambda_1\zeta(\lambda)}{\rho(1,0,0)}\mathrm{d}\lambda\\&\ge\frac{1}{\dfrac{\rho(2,0,0)}{\rho(1,0,0)}+(\nu-1)\dfrac{\rho(1,1,0)}{\rho(1,0,0)}+s}.\end{aligned}\tag{68}$$

Let f be a probability density given by

$$f(\lambda_1,\dots,\lambda_\nu)=\frac{d}{2}\left(\frac{d}{2}-1\right)^{\nu-1}\lambda_1^{d/2-1}\prod_{i=2}^\nu\lambda_i^{d/2-2},$$

which is clearly MTP2. Also let

$$\begin{aligned}g_1(\lambda_1,\dots,\lambda_\nu)&=\lambda_1,\\g_2(\lambda_1,\dots,\lambda_\nu)&=-\frac{\exp\left(sz/\{\sum\lambda_i+s\}\right)}{(\sum\lambda_i+s)^{d/2}},\end{aligned}$$

which are both increasing increasing in each of its arguments. Hence, by so-called FKG inequality given in Theorem C.2 below,

$$\begin{aligned}&\int_{\mathcal{D}_\nu}g_1(\lambda_1,\dots,\lambda_\nu)g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda\\&\ge\int_{\mathcal{D}_\nu}g_1(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda\\&\quad\times\int_{\mathcal{D}_\nu}g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda\end{aligned}$$

or equivalently

$$\begin{aligned}&\frac{\int_{\mathcal{D}_\nu}g_1(\lambda_1,\dots,\lambda_\nu)g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda}{\int_{\mathcal{D}_\nu}g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda}\\&\le\int_{\mathcal{D}_\nu}g_1(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda,\end{aligned}$$

since $g_2<0$. Since $\rho(2,0,0)/\rho(1,0,0)$ is expressed as

$$\begin{aligned}&\frac{\rho(2,0,0)}{\rho(1,0,0)}\\&=\frac{\int_{\mathcal{D}_\nu}g_1(\lambda_1,\dots,\lambda_\nu)g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda}{\int_{\mathcal{D}_\nu}g_2(\lambda_1,\dots,\lambda_\nu)f(\lambda_1,\dots,\lambda_\nu)\mathrm{d}\lambda},\end{aligned}$$

we have

$$\frac{\rho(2,0,0)}{\rho(1,0,0)}\le\frac{d}{d+2}.\tag{69}$$

Similarly we have

$$\frac{\rho(1,1,0)}{\rho(1,0,0)}\le\frac{d-2}{d}\le\frac{d}{d+2}.\tag{70}$$

Hence, by (68), (69) and (70), we have

$$\frac{\rho(1,0,-1)}{\rho(1,0,0)}\ge\frac{1}{\nu d/(d+2)+s}.$$

Lemma C.1: Let

$$\zeta(\lambda_1,\dots,\lambda_\nu)=\frac{\prod\lambda_i^{d/2-2}}{(\sum\lambda_i+s)^{d/2}}\exp\left(-\frac{\sum\lambda_i}{\sum\lambda_i+s}z\right).$$

Then $\zeta(\lambda_1,\dots,\lambda_\nu)$ is MTP2.

Proof: Note

$$\exp\left(-\frac{\sum\lambda_i}{\sum\lambda_i+s}z\right)=\exp(-z)\exp\left(\frac{sz}{\sum\lambda_i+s}\right).$$

From the form of ζ, we have only to check

$$\begin{aligned}&\left(\sum\lambda_i+s\right)\left(\sum\xi_i+s\right)\\&\geq\left(\sum\lambda_i\vee\xi_i+s\right)\left(\sum\lambda_i\wedge\xi_i+s\right)\end{aligned}$$

or equivalently

$$\left(\sum\lambda_i\right)\left(\sum\xi_i\right)\geq\left(\sum\lambda_i\vee\xi_i\right)\left(\sum\lambda_i\wedge\xi_i\right).$$

We have

$$\begin{aligned}&\left(\sum\lambda_i\right)\left(\sum\xi_i\right)-\left(\sum\lambda_i\vee\xi_i\right)\left(\sum\lambda_i\wedge\xi_i\right)\\&=\sum_{i\neq j}\{\lambda_i\xi_j+\lambda_j\xi_i-(\lambda_i\vee\xi_i)(\lambda_j\wedge\xi_j)\\&\quad-(\lambda_j\vee\xi_j)(\lambda_i\wedge\xi_i)\}.\end{aligned}$$

Without the loss of generality, assume $\lambda_i\geq\xi_i$. Then we have

$$\begin{aligned}&\lambda_i\xi_j+\lambda_j\xi_i-(\lambda_i\vee\xi_i)(\lambda_j\wedge\xi_j)-(\lambda_j\vee\xi_j)(\lambda_i\wedge\xi_i)\\&=\lambda_i\xi_j+\lambda_j\xi_i-\lambda_i(\lambda_j\wedge\xi_j)-(\lambda_j\vee\xi_j)\xi_i\\&=\lambda_i\{\xi_j-(\lambda_j\wedge\xi_j)\}-\xi_i\{(\lambda_j\vee\xi_j)-\lambda_j\}\\&=(\lambda_i-\xi_i)\{\xi_j-(\lambda_j\wedge\xi_j)\}\\&\geq 0,\end{aligned}$$

which completes the proof. ■

Theorem C.1 (Theorem 2.1 of [26]): Let f_1, f_2, f_3 and f_4 be nonnegative functions satisfying for all $x, y \in \mathbb{R}^\nu$

$$f_1(x)f_2(y)\leq f_3(x\vee y)f_4(x\wedge y).$$

Then

$$\int f_1(x)\mathrm{d}x\int f_2(x)\mathrm{d}x\leq\int f_3(x)\mathrm{d}x\int f_4(x)\mathrm{d}x.$$

Theorem C.2 (FKG Inequality, e.g. Theorem 2.3 of [26]): Let $f(x)$ for $x\in\mathbb{R}^\nu$ be a probability density satisfying MTP2. Then for any pair of increasing functions $g_1(x)$ and $g_2(x)$, we have

$$\int g_1(x)g_2(x)f(x)\mathrm{d}x\geq\int g_1(x)f(x)\mathrm{d}x\int g_2(x)f(x)\mathrm{d}x.$$

Appendix D
Numerical Experiments

We numerically computed the risk functions of $\hat{p}_{\mathrm{H}}(y\,|\,x;\alpha)$, Bayesian predictive densities with respect to the harmonic prior, by Monte Carlo method:

$$\begin{aligned}&R_\alpha\left\{\phi(y-\mu,v_y)\,||\,\hat{p}_{\mathrm{H}}(y\,|\,\cdot;\alpha)\right\}\approx\frac{1}{T}\sum_{t=1}^{T}L_t,\\&L_t=f_\alpha\left(\frac{\hat{p}_{\mathrm{H}}(y_t\,|\,x_t;\alpha)}{\phi(y_t-\mu,v_y)}\right),\end{aligned}$$

where $x_1,\cdots,x_T$ and $y_1,\cdots,y_T$ are i.i.d. samples from $N_d(\mu,v_xI)$ and $N_d(\mu,v_yI)$, respectively, and f_α is defined by (6). Let

$$\bar{L}_T=\frac{L_1+\cdots+L_T}{T},\ \text{and}\ s_T^2=\frac{L_1^2+\cdots+L_T^2}{T}-\bar{L}_T^2$$

be the sample mean and sample variance, respectively. We found that s_T^2 may become large, especially when v_x/v_y is large. Thus, we determined the Monte Carlo sample size T adaptively as follows. First, we did Monte Carlo with sample size 100. Next, if the estimated coefficient of variation $s_T/(\sqrt{T}\bar{L}_T)$ was larger than 0.05, we continued sampling of x_t and y_t until the estimated coefficient of variation became smaller than 0.05. In computing the value of Bayesian predictive density with respect to the harmonic prior at some point, we used the formula (17) of Theorem 2.1 with $m_{\mathrm{H}}(w,v)$ given by (51). Here, we computed the denominator of (17) by using Monte Carlo method with adaptively chosen sample size N: first N was set to 10^4 and then increased until the estimated coefficient of variation became smaller than 0.001.

Acknowledgment

The authors thank the referees and associate editor for their constructive and insightful comments that helped improve this manuscript.

References

[1] J. Aitchison, "Goodness of prediction fit," *Biometrika*, vol. 62, no. 3, pp. 547–554, 1975.
[2] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2708–2726, 2004.
[3] F. Komaki, "A shrinkage predictive distribution for multivariate normal observables," *Biometrika*, vol. 88, no. 3, pp. 859–864, 2001.
[4] C. Stein, "Estimation of the mean of a multivariate normal distribution," in *Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II.* Prague: Charles Univ., 1974, pp. 345–381.
[5] E. I. George, F. Liang, and X. Xu, "Improved minimax predictive densities under Kullback-Leibler loss," *Ann. Statist.*, vol. 34, no. 1, pp. 78–91, 2006.
[6] W. E. Strawderman, "Proper Bayes minimax estimators of the multivariate normal mean," *Ann. Math. Statist.*, vol. 42, no. 1, pp. 385–388, 1971.
[7] A. Cichocki and S.-i. Amari, "Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.
[8] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statistics*, vol. 23, pp. 493–507, 1952.
[9] S. Amari, *Differential-geometrical methods in statistics*, ser. Lecture Notes in Statistics. Springer-Verlag, New York, 1985, vol. 28.
[10] ——, "α-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.
[11] ——, *Information geometry and its applications*, ser. Applied Mathematical Sciences. Springer, [Tokyo], 2016, vol. 194.
[12] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[13] L. M. Brègman, "A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming," *Ž. Vyčisl. Mat. i Mat. Fiz.*, vol. 7, pp. 620–631, 1967.
[14] J. M. Corcuera and F. Giummolè, "A generalized Bayes rule for prediction," *Scand. J. Statist.*, vol. 26, no. 2, pp. 265–279, 1999.
[15] T. Yanagimoto and T. Ohnishi, "Bayesian prediction of a density function in terms of e-mixture," *J. Statist. Plann. Inference*, vol. 139, no. 9, pp. 3064–3075, 2009.
[16] Y. Maruyama and W. E. Strawderman, "Bayesian predictive densities for linear regression models under α-divergence loss: some results and open problems," in *Contemporary developments in Bayesian analysis and statistical decision theory: a Festschrift for William E. Strawderman*, ser. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2012, vol. 8, pp. 42–56.

[17] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, 1981.
[18] L. D. Brown, E. I. George, and X. Xu, "Admissible predictive density estimation," *Ann. Statist.*, vol. 36, no. 3, pp. 1156–1170, 2008.
[19] L. D. Brown, "A heuristic method for determining admissibility of estimators—with applications," *Ann. Statist.*, vol. 7, no. 5, pp. 960–994, 1979.
[20] G. Mukherjee and I. M. Johnstone, "On minimax optimality of sparse Bayes predictive density estimates," 2017, eprint arXiv:1707.04380.
[21] ——, "Exact minimax estimation of the predictive density in sparse Gaussian models," *Ann. Statist.*, vol. 43, no. 3, pp. 937–961, 2015.
[22] M. Ghosh, V. Mergel, and G. S. Datta, "Estimation, prediction and the Stein phenomenon under divergence loss," *J. Multivariate Anal.*, vol. 99, no. 9, pp. 1941–1961, 2008.
[23] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Department of Statistics, Stanford University, Tech. Rep. 51, 1964.
[24] J. O. Berger, *Statistical decision theory and Bayesian analysis*, 2nd ed., ser. Springer Series in Statistics. New York: Springer-Verlag, 1985.
[25] E. L. Lehmann and G. Casella, *Theory of point estimation*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 1998.
[26] S. Karlin and Y. Rinott, "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions," *J. Multivariate Anal.*, vol. 10, no. 4, pp. 467–498, 1980.

Yuzo Maruyama received the B.A., M.S. and Ph.D. degrees from the University of Tokyo in 1994, 1996 and 2000, respectively. He was an Assistant Professor with the School of Mathematics, Kyushu University. He is currently a Professor with Mathematics and Informatics Center, The University of Tokyo.

Takeru Matsuda received the B.E., M.S. and Ph.D. degrees from the University of Tokyo in 2012, 2014 and 2017, respectively. He is currently an Assistant Professor with Department of Mathematical Informatics, the University of Tokyo and a Visiting Researcher at RIKEN Center for Brain Science.

Toshio Ohnishi received the B.S. degree from the University of Tokyo in 1993, the M.S. degree from the University of Tsukuba in 2000 and the Ph.D. from SOKENDAI, the Graduate University for Advanced Studies in 2003. He was an Assistant Professor with the Institute of Statistical Mathematics (Japan). He is currently a Professor with the Faculty of Economics, Kyushu University.