Electronic Journal of Statistics
Vol. 18 (2024) 3310–3326
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2277

Predictive densities for multivariate normal models based on extended models and shrinkage Bayes methods

Michiko Okudo and Fumiyasu Komaki

Department of Mathematical Informatics, The University of Tokyo,
e-mail: okudo@mist.i.u-tokyo.ac.jp; komaki@mist.i.u-tokyo.ac.jp

Abstract: We investigate predictive densities for multivariate normal models with unknown mean vectors and known covariance matrices. Bayesian predictive densities based on shrinkage priors often have complex representations, although they are effective in various problems. We consider extended normal models with mean vectors and covariance matrices as parameters, and adopt predictive densities that belong to the extended models including the original normal model. We adopt predictive densities that are optimal with respect to the posterior Bayes risk in the extended models. The proposed predictive density based on a superharmonic shrinkage prior is shown to dominate the Bayesian predictive density based on the uniform prior under a loss function based on the Kullback–Leibler divergence when the variance of future samples is sufficiently large. Our method provides an alternative to the empirical Bayes method, which is widely used to construct tractable predictive densities.

MSC2020 subject classifications: Primary 62C10, 62C12; secondary 62F15.
Keywords and phrases: Bayes extended estimator, empirical Bayes, extended plug-in density, Stein's prior.

Received December 2022.

1. Introduction

Suppose that we have independent observations $x_1, \ldots, x_n$ from a d-dimensional multivariate normal model $\mathrm{N}_d(\mu, I_d)$, $\mu \in \mathbb{R}^d$. By sufficiency reduction, it is sufficient to consider the setting in which we have a single observation x distributed according to $\mathrm{N}_d(\mu, uI_d)$, where $u > 0$ is known and fixed. We address the problem of predicting a future outcome y following a d-dimensional multivariate normal distribution $\mathrm{N}_d(\mu, vI_d)$, $\mu \in \mathbb{R}^d$, $v > 0$ with the same mean vector μ by using a predictive density $\hat{p}(y \mid x)$ that depends on x. The variance v is known and possibly differs from u. The performance of a predictive density $\hat{p}(y \mid x)$ is evaluated by the Kullback–Leibler divergence

$$D\{p(y;\mu,vI_d);\hat{p}(y \mid x)\} = \int p(y;\mu,vI_d)\log\frac{p(y;\mu,vI_d)}{\hat{p}(y \mid x)}dy,$$

where $p(y;\mu,\Sigma)$ $(\mu\in\mathbb{R}^d, \Sigma\in\mathbb{R}^{d\times d})$ is the density of $\mathrm{N}_d(\mu,\Sigma)$.

There are two widely used methods to construct predictive densities: Bayesian predictive densities and plug-in densities. Bayesian predictive densities are expressed as

$$p_\pi(y\mid x)=\int p(y;\mu,vI_d)p_\pi(\mu\mid x)d\mu,$$

where $p_\pi(\mu\mid x)$ is the posterior density

$$p_\pi(\mu\mid x)=\frac{p(x;\mu,vI_d)\pi(\mu)}{\int p(x;\mu,vI_d)\pi(\mu)d\mu}$$

based on a prior density $\pi(\mu)$. Bayesian predictive densities do not belong to a tractable finite-dimensional family unless a conjugate prior is adopted. On the other hand, plug-in predictive densities can easily be obtained by plugging an estimator $\hat{\mu}$ such as maximum likelihood estimators or Bayes estimators, of the unknown parameter μ of the density $p(y;\mu,vI_d)$ of y. However, Bayesian predictive densities are preferable to plug-in densities in many examples.

Shrinkage methods are effective both in estimation and in prediction for normal models with unknown mean vectors μ. Bayes estimators based on Stein's prior $\pi_{\mathrm{S}}(\mu)\propto\|\mu\|^{-(d-2)}$ dominate the maximum likelihood estimator $\hat{\mu}_{\mathrm{mle}}=x$ when $d\geq 3$ under the squared error loss $\|\mu-\hat{\mu}\|^2$ [12]. Kullback–Leibler loss $D\{p(y;\mu,vI_d);p(y;\hat{\mu},vI_d)\}$ for a plugin density $p(y;\hat{\mu},vI_d)$ is proportional to the squared error loss. Priors that "shrink" the posterior density to a certain point such as the origin or to a subspace, are called shrinkage priors. If a function $\pi(\mu)$ satisfies the inequality

$$\Delta\pi(\mu):=\sum_{i=1}^{d}\frac{\partial^2}{\partial\mu_i^2}\pi(\mu)\leq 0,$$

then $\pi(\mu)$ is said to be superharmonic. Bayes estimators based on nonconstant superharmonic priors dominate the maximum likelihood estimator [12]. The density π_{S} shrinks the posterior to the origin and satisfies

$$\Delta\pi_{\mathrm{S}}(\mu)=-\delta(\mu),$$

where δ denotes the Dirac delta function, in the framework of Schwartz's distribution theory, see e.g. [6] p. 74. In this sense, π_{S} is a superharmonic function. The maximum likelihood estimator coincides with the Bayes estimator based on the uniform prior $\pi_{\mathrm{U}}(\mu)\propto 1$.

A parallel result regarding Bayesian prediction is obtained by [8], and the Bayesian predictive density based on Stein's prior dominates the Bayesian predictive density based on the uniform prior. Bayesian predictive densities based on superharmonic priors dominate the Bayesian predictive density based on π_{U} [4]. Other important shrinkage priors for multivariate normal models with unknown mean include shrinkage priors for regression problems [5, 7] and singular value shrinkage priors for matrix-variate normal models [10]. The Bayesian predictive density $\hat{p}_{\mathrm{U}}$ based on π_{U} has the simple form $\mathrm{N}_d(x,(u+v)I_d)$. On the

other hand, Bayesian predictive densities based on shrinkage priors generally do not have such simple forms.

The empirical Bayes method is another method of constructing predictive densities with reasonable risk performance and small computational cost. An empirical Bayes method for approximating a Bayesian predictive density based on Stein's prior is studied by [13]. The use of plug-in densities for prediction is also discussed in [3]. Stein's prior is represented as a mixture of normal distributions:

$$\pi_{\mathrm{S}}(\mu) \propto \|\mu\|^{-(d-2)} = \frac{2}{\Gamma(d/2-1)} \int_0^\infty (2\tau)^{-d/2} \exp\left(-\frac{\|\mu\|^2}{2\tau}\right) d\tau, \tag{1.1}$$

where $\Gamma(d/2-1)$ denotes the Gamma function. The representation (1.1) is used to construct the Bayesian predictive density based on Stein's prior in [8]. In [13], Bayesian predictive densities based on a prior

$$\pi(\mu; \hat{\tau}(x)) = (2\pi\hat{\tau}(x))^{-d/2} \exp\left(-\frac{\|\mu\|^2}{2\hat{\tau}(x)}\right)$$

with an estimator $\hat{\tau}(x)$ are constructed. The predictive density based on the empirical Bayes method is expressed as

$$\int p(y; \mu, vI_d)\pi(\mu; \hat{\tau}(x))d\mu. \tag{1.2}$$

Therefore, the empirical Bayes method is regarded as an approximation of the full Bayes method in which a prior is adopted for the hyperparameter τ. The predictive density (1.2) that is obtained by the empirical Bayes method is also a normal distribution.

The computational difference between full Bayes and empirical Bayes lies in tha fact that empirical Bayes methods requires only one plug-in distribution to compute the predictive density. Approximating $p_\pi(y \mid x)$ by empirical Bayes saves computational cost and it is effective when predicting densities for many future samples and when d is large.

We present an alternative to the empirical Bayes method to construct tractable predictive densities based on shrinkage priors. We consider an "extended" model including the original model $\mathrm{N}_d(\mu, vI_d)$ with the fixed v. Normal models such as $\mathrm{N}_d(\mu, \xi I_d)$ $(\xi > 0)$ and $\mathrm{N}_d(\mu, \Sigma)$ $(\Sigma \in \mathbb{R}^{d\times d})$ are adopted as the extended models for y. We denote the predictive densities in those extended models as extended plug-in densities. The resulting predictive densities are optimal with respect to the posterior Bayes risk in the extended models. Our method is based on a combination of extended plug-in densities for curved exponential families [11] and shrinkage priors. We can construct predictive densities not only in the normal model $\mathrm{N}_d(\mu, \xi I_d)$ $(\xi > 0)$ like empirical Bayes method in [13], but also in the larger normal model $\mathrm{N}_d(\mu, \Sigma)$ $(\Sigma \in \mathbb{R}^{d\times d}, \Sigma \succ 0)$. This approach could apply to various models besides the normal models that can be embedded in larger exponential families.

We show that the Kullback–Leibler risk difference of an extended plug-in predictive density based on a prior and the Bayesian predictive density based on the uniform prior reduces to the Kullback–Leibler risk difference of the corresponding Bayes estimators in the limit $1/v \to 0$.

Thus, our predictive density asymptotically dominates the Bayesian predictive density based on the uniform prior if the performance of the predictive densities is evaluated in the limit $1/v \to 0$. The numerical simulations suggest that the proposed predictive density performs better than the Bayesian predictive density based on the uniform prior even if $1/v$ is not close to 0.

2. Bayes extended estimators

2.1. Extended models and estimators

We investigate extended plug-in densities in extended models as predictive densities. We consider two extended models:

$$\mathcal{E}_1 = \{\mathrm{N}_d(\mu, \xi I_d) \mid \mu \in \mathbb{R}^d,\ \xi > 0\},$$

and

$$\mathcal{E}_2 = \{\mathrm{N}_d(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}, \Sigma \succeq 0\},$$

which include the original model $\mathcal{P} := \mathrm{N}_d(\mu, vI_d)$ with the known v. In the first extended model $\mathcal{E}_1$, the variance ξ is a parameter in contrast to fixed v in $\mathcal{P}$.

The second extended model $\mathcal{E}_2$ allows all positive semidefinite covariance matrices Σ. The inclusion relation is $\mathcal{P} \subseteq \mathcal{E}_1 \subseteq \mathcal{E}_2$. Other extended models such as $\{\mathrm{N}_d(\mu, D) \mid \mu \in \mathbb{R}^d, D : d\text{-dimensional diagonal matrix}\}$ can be considered in the same manner.

Although the original model $\mathcal{P}$ is a full exponential family, it can be formulated as a curved exponential family that is embedded in the extended models $\mathcal{E}_1$ or $\mathcal{E}_2$. Thus, we can choose a predictive density that belongs to $\mathcal{E}_1$ or $\mathcal{E}_2$ instead of the original model $\mathcal{P}$. For a density function

$$p(y;\theta) = b(y)\exp(s(y)^\top\theta - \psi(\theta))$$

of an exponential family $\mathcal{E}$, the expectation parameter is

$$\eta(\theta) = \mathrm{E}[s(y)].$$

When the original model $\mathcal{P}$ is a curved exponential family

$$p(y;\omega) = b(y)\exp(s(y)^\top\theta(\omega) - \psi(\theta(\omega)))$$

that is embedded in $\mathcal{E}$, we denote the posterior mean of η based on a prior π as the Bayes extended estimator and write it as $\hat{\eta}_\pi$. It minimizes the posterior Bayes risk of $p(y;\hat{\eta})$ about $\hat{\eta}$, thus it is reasonable to consider the extended plug-in densities that belong to $\mathcal{E}$ and not to $\mathcal{P}$ [11] for prediction. In other words,

TABLE 1
Extended plug-in distributions and extended Bayes estimators with respect to a prior density π.

Extended model	Expectation parameters	Bayes extended estimators	Extended plug-in distribution
$\mathcal{E}_1$: $\mathrm{N}_d(\mu, \xi I_d)$	$(\mu, d\xi + \mu^\top \mu)$	$\hat{\mu}_\pi = \mathrm{E}_\pi[\mu \mid x]$, $\hat{\xi}_\pi = v + (\mathrm{E}_\pi[\mu^\top \mu \mid x] - \hat{\mu}_\pi^\top \hat{\mu}_\pi)/d$	$\mathrm{N}_d(\hat{\mu}_\pi, \hat{\xi}_\pi I_d)$
$\mathcal{E}_2$: $\mathrm{N}_d(\mu, \Sigma)$	$(\mu, \Sigma + \mu\mu^\top)$	$\hat{\mu}_\pi = \mathrm{E}_\pi[\mu \mid x]$ $\hat{\Sigma}_\pi = vI_d + \mathrm{E}_\pi[\mu\mu^\top \mid x] - \hat{\mu}_\pi \hat{\mu}_\pi^\top$	$\mathrm{N}_d(\hat{\mu}_\pi, \hat{\Sigma}_\pi)$

extended plug-in densities with $\hat{\eta}_\pi$ are the closest to the Bayesian predictive densities with respect to the posterior Bayes risk.

We obtain the expectation parameters of the extended models $\mathcal{E}_1$ and $\mathcal{E}_2$. Bayes extended estimators are their posterior means. The results are shown in Table 1.

The expectation parameters of the extended model $\mathcal{E}_1 = \mathrm{N}_d(\mu, \xi I_d)$ are

$$\eta = (\mathrm{E}[y], \mathrm{E}[y^\top y]) = (\mu, d\xi + \mu^\top \mu).$$

The expectation parameter for a density in $\mathcal{P} \subset \mathcal{E}_1$ is $\eta = (\mu, dv + \mu^\top \mu)$, where v is known and fixed. Thus, the posterior mean of η is

$$\hat{\eta}_\pi = (\mathrm{E}_\pi[\mu \mid x],\ dv + \mathrm{E}_\pi[\mu^\top \mu \mid x]), \tag{2.1}$$

where $\mathrm{E}_\pi[\cdot \mid x]$ denotes the expectation with respect to the posterior density of μ based on a prior π. Although the prior and posterior measures are probability measures on $\mathcal{P}$, an extended plug-in distribution with the posterior mean $\mathrm{E}_\pi[\eta \mid x]$ belongs to $\mathcal{E}_1$, not to $\mathcal{P}$, which consequently has a favourable effect on the predictive performance.

By plugging (2.1) into $\eta = (\mu, d\xi + \mu^\top \mu)$, we obtain the extended plug-in density $\mathrm{N}_d(\hat{\mu}_\pi, \hat{\xi}_\pi I_d)$, with respect to $\mathcal{E}_1$, where $\hat{\mu}_\pi = \mathrm{E}_\pi[\mu \mid x]$ and $\hat{\xi}_\pi = v + (\mathrm{E}_\pi[\mu^\top \mu \mid x] - \hat{\mu}_\pi^\top \hat{\mu}_\pi)/d$.

Similarly, the expectation parameter of the second extended model $\mathcal{E}_2$ is

$$\eta = (\mathrm{E}[y], \mathrm{E}[yy^\top]) = (\mu, \Sigma + \mu\mu^\top).$$

Thus, the extended plug-in density with the posterior mean $\hat{\eta}_\pi$ based on $\mathcal{E}_2$ is $\mathrm{N}_d(\hat{\mu}_\pi, \hat{\Sigma}_\pi)$, where $\hat{\Sigma}_\pi = vI_d + \mathrm{E}_\pi[\mu\mu^\top \mid x] - \hat{\mu}_\pi \hat{\mu}_\pi^\top$.

We obtain the extended plug-in densities with respect to the uniform prior $\pi_{\mathrm{U}}(\mu) = 1$. As the posterior density with respect to π_{U} is

$$p_{\mathrm{U}}(\mu \mid x) = p(x; \mu, uI_d)\pi_{\mathrm{U}}(\mu),$$

we obtain

$$\hat{\mu}_{\mathrm{U}} = \mathrm{E}_{\pi_{\mathrm{U}}}[\mu \mid x] = x$$

and

$$\hat{\xi}_{\mathrm{U}} = v + \mathrm{E}_{\pi_{\mathrm{U}}}[\mu^\top \mu \mid x]/d - x^\top x/d = u + v.$$

Thus, the extended plug-in distribution $\mathrm{N}_d(\hat{\mu}_{\mathrm{U}}, \hat{\xi}_{\mathrm{U}} I_d) = \mathrm{N}_d(\hat{\mu}_{\mathrm{U}}, (u+v) I_d)$ based on $\mathcal{E}_1$ is identical to the Bayesian predictive density $\hat{p}_{\mathrm{U}}$ based on π_{U} mentioned in Section 1. As it is optimal with respect to the posterior Bayes risk among all distributions and included in $\mathcal{E}_1$, the extended plug-in density based on $\mathcal{E}_2$ is also identical to $\mathrm{N}_d(\hat{\mu}_{\mathrm{U}}, \hat{\xi}_{\mathrm{U}} I_d)$.

Bayesian predictive densities based on shrinkage priors do not belong to normal models, thus we investigate extended plug-in densities based on shrinkage priors including Stein's prior $\pi_{\mathrm{S}} \propto \|\mu\|^{-(d-2)}$.

2.2. Posterior mean representations

We evaluate posterior means that were described in the previous subsection. Let

$$m_\pi(x) = \int p(x; \mu, uI_d)\pi(\mu)d\mu,$$

which is the marginal density of x. The derivatives of model density functions are given by

$$\begin{aligned}
\nabla p(x; \mu, uI_d) &= \frac{1}{u}(\mu - x)p(x; \mu, uI_d), \\
\nabla^2 p(x; \mu, uI_d) &= -\frac{1}{u}p(x; \mu, uI_d)I_d + \frac{1}{u^2}(\mu - x)(\mu - x)^\top p(x; \mu, uI_d),
\end{aligned}$$

and the Laplacian is given by

$$\Delta p(x; \mu, uI_d) = -\frac{d}{u}p(x; \mu, uI_d) + \frac{1}{u^2}(\mu - x)^\top(\mu - x)p(x; \mu, uI_d),$$

where, for a function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_d}(x)\right)^\top, \quad \Delta f(x) := \sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}(x)$$

and $\nabla^2 f$ is the Hessian matrix whose (i, j) element is

$$(\nabla^2 f(x))_{ij} := \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

The posterior mean of μ is evaluated in [1] as

$$\begin{aligned}
\hat{\mu}_\pi &= \int \mu \frac{p(x; \mu, uI_d)\pi(\mu)}{\int p(x; \tilde{\mu}, uI_d)\pi(\tilde{\mu})d\tilde{\mu}} d\mu \\
&= x + \int (\mu - x) \frac{p(x; \mu, uI_d)\pi(\mu)}{\int p(x; \tilde{\mu}, uI_d)\pi(\tilde{\mu})d\tilde{\mu}} d\mu = x + u\nabla \log m_\pi. \qquad (2.2)
\end{aligned}$$

The posterior mean of $(\mu - x)(\mu - x)^\top$ is expressed as

$$\int (\mu - x)(\mu - x)^\top \frac{p(x; \mu, uI_d)\pi(\mu)}{\int p(x; \tilde{\mu}, uI_d)\pi(\tilde{\mu})d\tilde{\mu}} d\mu$$

$$= \int \{u^2 \nabla^2 p(x;\mu,uI_d) + up(x;\mu,uI_d)I_d\} \frac{\pi(\mu)}{\int p(x;\tilde{\mu},uI_d)\pi(\tilde{\mu})d\tilde{\mu}} d\mu$$
$$= u^2 \frac{\nabla^2 m_\pi}{m_\pi} + uI_d. \tag{2.3}$$

Thus, the posterior mean of $(\mu - x)^\top(\mu - x)$ is

$$\int (\mu - x)^\top(\mu - x) \frac{p(x;\mu,uI_d)\pi(\mu)}{\int p(x;\tilde{\mu},uI_d)\pi(\tilde{\mu})d\tilde{\mu}} d\mu = u^2 \frac{\Delta m_\pi}{m_\pi} + du. \tag{2.4}$$

From (2.2), (2.3), and (2.4), the estimators $\hat{\xi}_\pi$ and $\hat{\Sigma}_\pi$ are given by

$$\begin{aligned}\hat{\xi}_\pi &= v + (\mathrm{E}_\pi[\mu^\top\mu \mid x] - \hat{\mu}_\pi^\top\hat{\mu}_\pi)/d \\ &= v + \mathrm{E}_\pi[(\mu - x)^\top(\mu - x) \mid x]/d - (\hat{\mu}_\pi - x)^\top(\hat{\mu}_\pi - x)/d \\ &= u + v + h_\pi(x),\end{aligned} \tag{2.5}$$

where

$$h_\pi(x) \coloneqq \frac{u^2}{d}\frac{\Delta m_\pi}{m_\pi} - \frac{u^2}{d}\frac{\|\nabla m_\pi\|^2}{m_\pi^2},$$

and

$$\begin{aligned}\hat{\Sigma}_\pi &= vI_d + \mathrm{E}_\pi[\mu\mu^\top \mid x] - \hat{\mu}_\pi^\top\hat{\mu}_\pi \\ &= vI_d + \mathrm{E}_\pi[(\mu - x)(\mu - x)^\top \mid x] - (\hat{\mu}_\pi - x)(\hat{\mu}_\pi - x)^\top \\ &= (u + v)I_d + H_\pi(x),\end{aligned} \tag{2.6}$$

where

$$H_\pi(x) \coloneqq u^2\left(\frac{\nabla^2 m_\pi}{m_\pi} - \frac{\nabla m_\pi \nabla m_\pi^\top}{m_\pi^2}\right).$$

Note that $\hat{\xi}_\pi$ is greater than the model variance v. If π is superharmonic, $\hat{\xi}_\pi$ is smaller than $u + v$, which is the variance of the Bayesian predictive density based on the uniform prior. It can be shown that $\Delta m_\pi \le 0$ holds if $\Delta\pi \le 0$ as follows. We have

$$\frac{\partial}{\partial x_i} m_\pi(x) = \int \frac{\mu - x_i}{u} p(x;\mu,uI_d)\pi(\mu)d\mu = \int p(x;\mu,uI_d)\frac{\partial}{\partial \mu_i}\pi(\mu)d\mu.$$

The last equation comes from Stein's lemma. Thus,

$$\frac{\partial^2}{\partial x_i^2} m_\pi(x) = \int p(x;\mu,uI_d)\frac{\partial^2}{\partial \mu_i^2}\pi(\mu)d\mu$$

and we obtain

$$\Delta m_\pi(x) = \int p(x;\mu,uI_d)\Delta\pi(\mu)d\mu.$$

Therefore, when π is a superharmonic function, m_π is also superharmonic and

$$\hat{\xi}_\pi(x) = u + v + h_\pi(x) \leq u + v.$$

On the other hand, because

$$\hat{\xi}_\pi = v + (\mathrm{E}[\mu^\top \mu \mid x] - \hat{\mu}_\pi^\top \hat{\mu}_\pi)/d = v + \mathrm{E}[(\mu - \hat{\mu}_\pi)^\top (\mu - \hat{\mu}_\pi) \mid x]/d,$$

we have $\hat{\xi}_\pi(x) \geq v$. Because $\mathrm{tr}\hat{\Sigma}_\pi = d\hat{\xi}_\pi$, the average of the eigenvalues of $\hat{\Sigma}_\pi$ is also smaller than the variance $u+v$ of the Bayesian predictive density based on the uniform prior.

3. Risk for infinitesimal prediction

We compare the Kullback–Leibler risk of the extended plug-in densities with Bayes extended estimators and that of the Bayesian predictive density $p_{\mathrm{U}}(y \mid x)$ based on the uniform prior. The Bayesian predictive density $p_{\mathrm{U}}(y \mid x)$ is included in the extended normal model $\mathrm{N}_d(\mu, \xi I_d)$ $(\xi \in \mathbb{R})$ and it is minimax. It is desirable to obtain predictive densities belonging to the extended models that perform better than $p_{\mathrm{U}}(y \mid x)$.

The risk function of $\hat{p}(y \mid x)$ is

$$R(\mu; \hat{p}) = \mathrm{E}[D\{p(y; \mu, vI_d); \hat{p}(y \mid x)\}] = \int p(x; \mu, uI_d) D\{p(y; \mu, vI_d); \hat{p}(y \mid x)\} dx.$$

For the predictive densities $\hat{p}_1(y \mid x)$ and $\hat{p}_2(y \mid x)$, we have

$$D\{p(y; \mu, vI_d); \hat{p}_1(y \mid x)\} - D\{p(y; \mu, vI_d); \hat{p}_2(y \mid x)\} = \int p(y; \mu, vI_d) \log \frac{\hat{p}_2(y \mid x)}{\hat{p}_1(y \mid x)} dy.$$

We introduce the time variables $s := 1/u$ and $t := 1/v$, which can be regarded as the numbers of observations and the number of future samples, respectively. To appreciate the time interpretation of s and t, consider a Gaussian process Z_τ $(\tau \geq 0)$ defined by the stochastic differential equation

$$\mathrm{d}Z_\tau = \mu \mathrm{d}\tau + \mathrm{d}B_\tau \quad (\tau \geq 0),$$

where $Z_0 = 0$ and B_τ $(\tau \geq 0)$ is a standard d-dimensional Brownian motion. Consequently, the distribution of $(1/\tau)Z_\tau$ is $\mathrm{N}(\mu, (1/\tau)I_d)$. Thus, our problem is equivalent to a problem in which we observe $(1/s)Z_s$ and predict $(1/t)(Z_{s+t} - Z_s)$. Therefore, s and t correspond to the observation time and prediction time, respectively. Let $\hat{\mu}_{t,\pi}$ be the posterior mean of μ based on observation Z_t and prior π.

In this setting, the relationship between prediction risk and estimation risk used in [2] is represented by

$$R(\mu; p_{\mathrm{U}}) - R(\mu; p_\pi) = \int_s^{s+t} \frac{\mathrm{E}_\tau[\|\tau^{-1} Z_\tau - \mu\|^2] - \mathrm{E}_\tau[\|\hat{\mu}_{\tau,\pi} - \mu\|^2]}{2} \mathrm{d}\tau, \quad (3.1)$$

where $\mathrm{E}_\tau[\cdot]$ means taking expectation about $\mathrm{N}_d(\mu, (1/\tau)I_d)$. This shows that the risk difference of the Bayesian predictive densities is represented as the integration of the estimation risk difference from s to $s+t$. The relationship of prediction risk and estimation risk is also considered in [3].

The relation (3.1) shows that

$$R(\mu; p_{\mathrm{U}}) - R(\mu; p_\pi) \geq 0$$

holds if

$$\mathrm{E}_\tau[\|\tau^{-1}Z_\tau - \mu\|^2] - \mathrm{E}_\tau[\|\hat{\mu}_{\tau,\pi} - \mu\|^2] \geq 0$$

for all $\tau > 0$. Thus, if π is a superharmonic prior, p_π dominates p_{U}. Since

$$\begin{aligned}\lim_{t\to 0}\frac{1}{t}\int_s^{s+t}\frac{\mathrm{E}_\tau[\|\tau^{-1}Z_\tau - \mu\|^2] - \mathrm{E}_\tau[\|\hat{\mu}_{\tau,\pi} - \mu\|^2]}{2}\mathrm{d}\tau \\ = \frac{1}{2}\big(\mathrm{E}_s[\|\tau^{-1}Z_s - \mu\|^2] - \mathrm{E}_s[\|\hat{\mu}_{s,\pi} - \mu\|^2]\big),\end{aligned}$$

the estimation risk difference $\mathrm{E}_s[\|s^{-1}Z_s - \mu\|^2] - \mathrm{E}_s[\|\hat{\mu}_{s,\pi} - \mu\|^2]$ corresponds to the infinitesimal-prediction risk difference.

Subsequently, we consider the relationship between the risk of extended plug-in densities and that of Bayes extended estimators. We compare the risk functions of extended plug-in predictive densities with Bayes extended estimators based on superharmonic priors and the uniform prior π_{U}. Recall that the extended plug-in densities $p(y; \hat{\mu}_{\mathrm{U}}, \hat{\xi}_{\mathrm{U}})$ and $p(y; \hat{\mu}_{\mathrm{U}}, \hat{\Sigma}_{\mathrm{U}})$ based on the uniform prior π_{U} coincide with the Bayesian predictive density based on π_{U}. We show that the infinitesimal prediction risk difference of extended plug-in predictive densities at $\tau = s$ is the risk difference between the corresponding Bayes extended estimators. This shows that the extended plug-in distributions with $(\hat{\mu}_\pi, \hat{\xi}_{t,\pi})$ and $(\hat{\mu}_\pi, \hat{\Sigma}_{t,\pi})$, where the subscript t is added to the densities to clarify their dependency on it, have better performance than $\hat{p}_{\mathrm{U}}$ if t is small enough and π is a superharmonic prior for $d \geq 3$. From (2.2), $\hat{\mu}_\pi$ does not depend on t.

Theorem 3.1. *For a prior π, denote the Kullback–Leibler risk of $p(y; \hat{\mu}_\pi, \hat{\xi}_{t,\pi} I_d)$ and $p(y; \hat{\mu}_\pi, \hat{\Sigma}_{t,\pi})$ as*

$$R_t(\mu; \hat{\mu}_\pi, \hat{\xi}_{t,\pi}) := \int p(x; \mu, s^{-1}I_d) D\{p(y; \mu, t^{-1}I_d); p(y; \hat{\mu}_\pi(x), \hat{\xi}_{t,\pi}(x) I_d)\} dx.$$

and

$$R_t(\mu; \hat{\mu}_\pi, \hat{\Sigma}_{t,\pi}) := \int p(x; \mu, s^{-1}I_d) D\{p(y; \mu, t^{-1}I_d); p(y; \hat{\mu}_\pi(x), \hat{\Sigma}_{t,\pi}(x))\mathrm{d}x,$$

respectively. Then,

$$\lim_{t\to 0}\frac{\partial}{\partial t}\left\{R_t(\mu; \hat{p}_{t,\mathrm{U}}) - R_t(\mu; \hat{\mu}_\pi, \hat{\xi}_{t,\pi})\right\} = \frac{\mathrm{E}[\|x - \mu\|^2] - \mathrm{E}[\|\hat{\mu}_\pi - \mu\|^2]}{2} \tag{3.2}$$

and

$$\lim_{t\to 0}\frac{\partial}{\partial t}\left\{R_t(\mu;\hat{p}_{t,\mathrm{U}})-R_t(\mu;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})\right\}=\frac{\mathrm{E}[\|x-\mu\|^2]-\mathrm{E}[\|\hat{\mu}_\pi-\mu\|^2]}{2} \tag{3.3}$$

hold.

Proof. The risk difference between $p(y;\hat{\mu}_\pi,\hat{\xi}_{t,\pi}I_d)$ and $p_{\mathrm{t,U}}(y\mid x)=p(y;x,(s^{-1}+t^{-1})I_d)$ is given by

$$\begin{aligned}
R_t(\mu;\hat{p}_{t,\mathrm{U}})-R_t(\mu;\hat{\mu}_\pi,\hat{\xi}_{t,\pi})&=\mathrm{E}_{x,y|t}\left[\log\frac{p(y;\hat{\mu}_\pi,\hat{\xi}_{t,\pi}I_d)}{p(y;x,(s^{-1}+t^{-1})I_d)}\right]\\
&=\mathrm{E}_{x,y|t}\left[-\frac{d}{2}\log\frac{\hat{\xi}_{t,\pi}}{s^{-1}+t^{-1}}-\frac{1}{2\hat{\xi}_{t,\pi}}(y-\hat{\mu}_\pi)^\top(y-\hat{\mu}_\pi)\right.\\
&\qquad\left.+\frac{1}{2(s^{-1}+t^{-1})}(y-x)^\top(y-x)\right]\\
&=\mathrm{E}_{x,y|t}\left[-\frac{d}{2}\log\frac{\hat{\xi}_{t,\pi}}{s^{-1}+t^{-1}}-\frac{1}{2\hat{\xi}_{t,\pi}}(y-\hat{\mu}_\pi)^\top(y-\hat{\mu}_\pi)\right]+\frac{d}{2},
\end{aligned} \tag{3.4}$$

where the expectation about (x,y) is denoted as $\mathrm{E}_{x,y|t}[\cdot]$. We evaluate the differential of the risk difference with respect to t. From (2.5), we have

$$\frac{\partial\hat{\xi}_{t,\pi}}{\partial t}=\frac{\partial}{\partial t}\{s^{-1}+t^{-1}+h_\pi(x)\}=-t^{-2}.$$

Thus,

$$\begin{aligned}
\frac{\partial}{\partial t}\log\frac{\hat{\xi}_{t,\pi}}{s^{-1}+t^{-1}}&=\frac{1}{\hat{\xi}_{t,\pi}}\frac{\partial\hat{\xi}_{t,\pi}}{\partial t}-\frac{1}{s^{-1}+t^{-1}}\frac{\partial(s^{-1}+t^{-1})}{\partial t}\\
&=-t^{-2}\left(\frac{1}{\hat{\xi}_{t,\pi}}-\frac{1}{s^{-1}+t^{-1}}\right).
\end{aligned} \tag{3.5}$$

We differentiate the rest of (3.4) and obtain

$$\begin{aligned}
&\frac{\partial}{\partial t}\mathrm{E}_{x,y|t}\left[-\frac{1}{2\hat{\xi}_{t,\pi}}(y-\hat{\mu}_\pi)^\top(y-\hat{\mu}_\pi)\right]=-\frac{1}{2}\frac{\partial}{\partial t}\mathrm{E}_x\left[\frac{dt^{-1}+(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)}{\hat{\xi}_{t,\pi}}\right]\\
&\quad=-\frac{1}{2}\mathrm{E}_x\left[-dt^{-2}\frac{1}{\hat{\xi}_{t,\pi}}+dt^{-1}\frac{t^{-2}}{\hat{\xi}_{t,\pi}^2}+\frac{t^{-2}}{\hat{\xi}_{t,\pi}^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\right].
\end{aligned} \tag{3.6}$$

From (3.4), (3.5), and (3.6), we obtain

$$\begin{aligned}
&\frac{\partial}{\partial t}\left\{R_t(\mu;\hat{p}_{t,\mathrm{U}})-R_t(\mu;\hat{\mu}_\pi,\hat{\xi}_{t,\pi})\right\}\\
&=\frac{1}{2}\mathrm{E}_x\left[dt^{-2}\left(\frac{1}{\hat{\xi}_{t,\pi}}-\frac{1}{s^{-1}+t^{-1}}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad + dt^{-2}\frac{1}{\hat{\xi}_{t,\pi}} - dt^{-1}\frac{t^{-2}}{\hat{\xi}_{t,\pi}^2} - \frac{t^{-2}}{\hat{\xi}_{t,\pi}^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\Bigg] \\
&= \frac{1}{2}\mathrm{E}_x\Bigg[dt^{-2}\left(\frac{2}{s^{-1}+t^{-1}+h_\pi} - \frac{1}{s^{-1}+t^{-1}}\right) \\
&\qquad - dt^{-1}\frac{t^{-2}}{(s^{-1}+t^{-1}+h_\pi)^2} - \frac{t^{-2}}{(s^{-1}+t^{-1}+h_\pi)^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\Bigg] \\
&= \frac{1}{2}\mathrm{E}_x\Bigg[dt^{-1}\left\{2 - \frac{2(s^{-1}+h_\pi)}{s^{-1}+t^{-1}+h_\pi} - 1 + \frac{s^{-1}}{s^{-1}+t^{-1}}\right\} \\
&\qquad - dt^{-1}\left\{1 - \frac{t^2(s^{-1}+h_\pi)^2 + 2t(s^{-1}+h_\pi)}{(1+ts^{-1}+th_\pi)^2}\right\} \\
&\qquad - \frac{t^{-2}}{(s^{-1}+t^{-1}+h_\pi)^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\Bigg] \\
&= \frac{1}{2}\mathrm{E}_x\Bigg[d\left\{-\frac{2(s^{-1}+h_\pi)}{1+ts^{-1}+th_\pi} + \frac{s^{-1}}{1+ts^{-1}} + \frac{t(s^{-1}+h_\pi)^2 + 2(s^{-1}+h_\pi)}{(1+ts^{-1}+th_\pi)^2}\right\} \\
&\qquad - \frac{1}{(1+ts^{-1}+th_\pi)^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\Bigg] \\
&= \frac{1}{2}\mathrm{E}_x\Bigg[d\left\{s^{-1} - \frac{ts^{-2}}{1+ts^{-1}} - \frac{t(s^{-1}+h_\pi)^2}{(1+ts^{-1}+th_\pi)^2}\right\} \\
&\qquad - \frac{1}{(1+ts^{-1}+th_\pi)^2}(\mu-\hat{\mu}_\pi)^\top(\mu-\hat{\mu}_\pi)\Bigg].
\end{aligned}
$$

Thus, from

$$
\lim_{t\to 0}\frac{\partial}{\partial t}\left\{R_t(\mu;\hat{p}_{t,\mathrm{U}}) - R_t(\mu;\hat{\mu}_\pi,\hat{\xi}_{t,\pi})\right\} = \frac{ds^{-1} - \mathrm{E}_x[\|\hat{\mu}_\pi-\mu\|^2]}{2},
$$

the desired result (3.2) is obtained.

Next, the risk difference between the extended plug-in density $p(y;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})$ and $p_{\mathrm{U}}(y\mid x)$ is

$$
\begin{aligned}
R_t(\mu;p_{\mathrm{U}}) - R_t(\mu;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi}) &= \mathrm{E}_{x,y|t}\left[\log\frac{p(y;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})}{p(y;x,(s^{-1}+t^{-1})I_d)}\right] \\
&= \mathrm{E}_{x,y|t}\Bigg[-\frac{1}{2}\log\frac{|\hat{\Sigma}_{t,\pi}|}{(s^{-1}+t^{-1})^d} - \frac{1}{2}(y-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-1}(y-\hat{\mu}_\pi) \\
&\qquad + \frac{1}{2(s^{-1}+t^{-1})}(y-x)^\top(y-x)\Bigg]
\end{aligned}
$$

$$= \mathrm{E}_{x,y|t}\left[-\frac{1}{2}\log\frac{|\hat{\Sigma}_{t,\pi}|}{(s^{-1}+t^{-1})^d} - \frac{1}{2}(y-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-1}(y-\hat{\mu}_\pi)\right] + \frac{d}{2}. \tag{3.7}$$

From (2.6),

$$\frac{\partial}{\partial t}\hat{\Sigma}_{t,\pi} = \frac{\partial}{\partial t}\big\{(s^{-1}+t^{-1})I_d + H_\pi(x)\big\} = -t^{-2}I_d. \tag{3.8}$$

Thus,

$$\frac{\partial}{\partial t}\log|\hat{\Sigma}_{t,\pi}| = \mathrm{tr}\Big\{\hat{\Sigma}_{t,\pi}^{-1}(-t^{-2})I_d\Big\} = -t^{-2}\mathrm{tr}\hat{\Sigma}_{t,\pi}^{-1}$$

and

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{E}_{x,y|t}[(y-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-1}(y-\hat{\mu}_\pi)] &= \frac{\partial}{\partial t}\mathrm{E}_x[t^{-1}\mathrm{tr}(\hat{\Sigma}_{t,\pi}^{-1}) + (\mu-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-1}(\mu-\hat{\mu}_\pi)]\\ &= -t^{-2}\mathrm{tr}\,\hat{\Sigma}_{t,\pi}^{-1} + t^{-1}\mathrm{tr}(t^{-2}\hat{\Sigma}_{t,\pi}^{-2}) + t^{-2}(\mu-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-2}(\mu-\hat{\mu}_\pi).\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}&\frac{\partial}{\partial t}\mathrm{E}_{x,y|t}\left[\log\frac{p(y;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})}{p(y;x,(s^{-1}+t^{-1})I_d)}\right]\\ &= \frac{\partial}{\partial t}\mathrm{E}_{x,y|t}\left[-\frac{1}{2}\log\frac{|\hat{\Sigma}_{t,\pi}|}{(s^{-1}+t^{-1})^d} - \frac{1}{2}(y-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-1}(y-\hat{\mu}_\pi)\right]\\ &= \mathrm{E}_x\left[-\frac{1}{2}\Big(-t^{-2}\mathrm{tr}\,\hat{\Sigma}_{t,\pi}^{-1} - d\frac{-t^{-2}}{s^{-1}+t^{-1}}\Big)\right.\\ &\qquad\left.-\frac{1}{2}\Big\{-t^{-2}\mathrm{tr}\,\hat{\Sigma}_{t,\pi}^{-1} + t^{-1}\mathrm{tr}(t^{-2}\hat{\Sigma}_{t,\pi}^{-2}) + t^{-2}(\mu-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-2}(\mu-\hat{\mu}_\pi)\Big\}\right]\\ &= \mathrm{E}_x\left[\frac{dt^{-1}}{2}\Big(\frac{s^{-1}}{s^{-1}+t^{-1}}-1\Big) + t^{-1}\mathrm{tr}(t^{-1}\hat{\Sigma}_{t,\pi}^{-1})\right.\\ &\qquad\left.-\frac{t^{-1}}{2}\mathrm{tr}(t^{-2}\hat{\Sigma}_{t,\pi}^{-2}) - \frac{1}{2}t^{-2}(\mu-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-2}(\mu-\hat{\mu}_\pi)\right].\end{aligned}$$

Let

$$A_\pi := s^{-1}I_d + s^{-2}\left(\frac{\Delta^2 m_\pi}{m_\pi} - \frac{\Delta m_\pi \Delta m_\pi^\top}{m_\pi^2}\right).$$

Then,

$$t\hat{\Sigma}_{t,\pi} = I_d + tA_\pi$$

and

$$t^{-1}\hat{\Sigma}_{t,\pi}^{-1} = (I_d + tA_\pi)^{-1}.$$

When t is small enough, all absolute values of the eigenvalues of tA_π are smaller than 1 and

$$t^{-1}\hat{\Sigma}_{t,\pi}^{-1} = \sum_{i=0}^{\infty}(-1)^i(tA_\pi)^i.$$

In the same manner, let

$$B_\pi := 2A_\pi + tA_\pi^2,$$

and we have

$$(t\hat{\Sigma}_{t,\pi})^2 = I_d + tB_\pi,$$

and when t is small enough that all absolute values of the eigenvalues of tB_π are smaller than 1,

$$t^{-2}\hat{\Sigma}_{t,\pi}^{-2} = (I_d + tB_\pi)^{-1} = \sum_{i=0}^{\infty}(-1)^i(tB_\pi)^i.$$

Therefore, when $t > 0$ is small enough,

$$\begin{aligned}
&\frac{\partial}{\partial t}\mathrm{E}_{x,y|t}\left[\log\frac{p(y;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})}{p(y;x,(s^{-1}+t^{-1})I_d)}\right]\\
&= \mathrm{E}_x\left[\frac{dt^{-1}}{2}\Big(\frac{s^{-1}}{s^{-1}+t^{-1}}-1\Big) + t^{-1}\mathrm{tr}(t^{-1}\hat{\Sigma}_{t,\pi}^{-1}) - \frac{t^{-1}}{2}\mathrm{tr}(t^{-2}\hat{\Sigma}_{t,\pi}^{-2})\right.\\
&\quad\left. - \frac{1}{2}t^{-2}(\mu-\hat{\mu}_\pi)^\top\hat{\Sigma}_{t,\pi}^{-2}(\mu-\hat{\mu}_\pi)\right]\\
&= \frac{dt^{-1}}{2}\frac{s^{-1}}{s^{-1}+t^{-1}} + \mathrm{E}_x\left[-\frac{dt^{-1}}{2} + t^{-1}\mathrm{tr}\Big\{\sum_{i=0}^{\infty}(-1)^i(tA_\pi)^i\Big\}\right.\\
&\quad - \frac{t^{-1}}{2}\mathrm{tr}\Big\{\sum_{j=0}^{\infty}(-1)^j(tB_\pi)^j\Big\}\\
&\quad\left. - \frac{1}{2}(\mu-\hat{\mu}_\pi)^\top\Big\{\sum_{j=0}^{\infty}(-1)^j(tB_\pi)^j\Big\}(\mu-\hat{\mu}_\pi)\right]\\
&= \frac{d}{2}\frac{s^{-1}}{1+ts^{-1}} + \mathrm{E}_x\left[-\frac{dt^{-1}}{2} + \mathrm{tr}\Big\{t^{-1}I_d - A_\pi + \sum_{i=2}^{\infty}(-1)^i t^{i-1}A_\pi^i\Big\}\right.\\
&\quad - \frac{1}{2}\mathrm{tr}\Big\{t^{-1}I_d - 2A_\pi - tA_\pi^2 + \sum_{j=2}^{\infty}(-1)^j t^{j-1}B_\pi^j\Big\}\\
&\quad\left. - \frac{1}{2}(\mu-\hat{\mu}_\pi)^\top\Big\{I_d + \sum_{j=1}^{\infty}(-1)^j(tB_\pi)^j\Big\}(\mu-\hat{\mu}_\pi)\right].
\end{aligned}$$

Thus, from

$$\lim_{t\to 0}\frac{\partial}{\partial t}\mathrm{E}_{x,y|t}\left[\log\frac{p(y;\hat{\mu}_\pi,\hat{\Sigma}_{t,\pi})}{p(y;x,(s^{-1}+t^{-1})I_d)}\right] = \frac{ds^{-1} - \mathrm{E}_x[\|\hat{\mu}_\pi-\mu\|^2]}{2},$$

the desired result (3.3) is obtained. □

4. Numerical experiments

We compare the Kullback–Leibler risks of the extended plug-in densities based on Stein's prior π_{S}, the Bayesian predictive density $p_{\mathrm{U}}(y \mid x)$ based on the uniform prior π_{U}, the Bayesian predictive density $p_{\mathrm{S}}(y \mid x)$ based on π_{S}, and an empirical Bayes method studied in [13]. Observation x is distributed according to $\mathrm{N}_d(\mu, uI_d)$ with $\mu \in \mathbb{R}^d$ and $u > 0$, and a future sample y comes from a normal distribution $\mathrm{N}_d(\mu, vI_d)$ with the same mean μ and with a possibly different variance $v > 0$. In Theorem 3.1, we observe that the proposed methods based on a superharmonic prior dominate $p_{\mathrm{U}}(y \mid x)$ when $1/v$ is close to 0. In these experiments, we numerically evaluate the Kullback–Leibler risks for finite $v > 0$. Although we are interested in the risk comparison among predictive densities that can be obtained by simple computations, we also simulate the Kullback–Leibler risk of the Bayesian predictive density $p_{\mathrm{S}}(y \mid x)$ based on π_{S} to verify the approximate performance of those plug-in densities.

When Stein's prior is employed, the extended estimators $\hat{\mu}_\pi$, $\hat{\xi}_\pi$ and $\hat{\Sigma}_\pi$ are given by

$$\begin{aligned}
\hat{\mu}_\pi &= F_1 x, \\
\hat{\xi}_\pi &= v + F_1 u + \frac{x^\top x}{d}(F_2 - F_1^2), \\
\hat{\Sigma}_\pi &= vI_d + F_1 u I_d + x^\top x (F_2 - F_1^2)
\end{aligned}$$

where

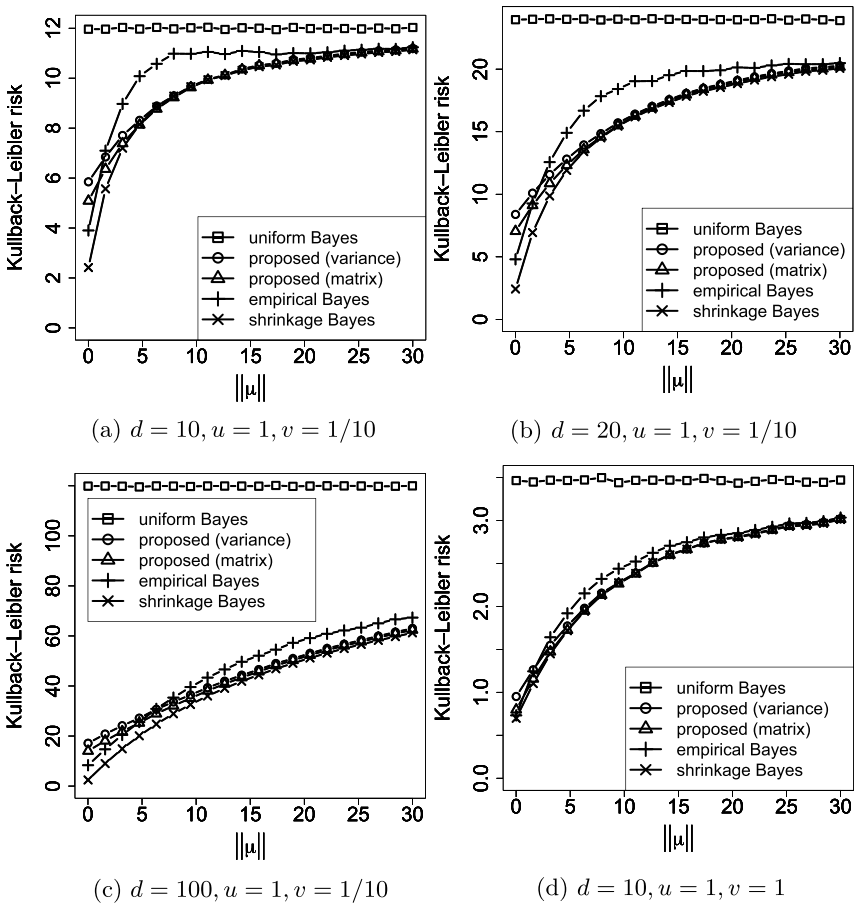
$$\begin{aligned}
F_1 &= 1 - 2\frac{\phi_{d+2}(\|x\|/\sqrt{u})}{\phi_d(\|x\|/\sqrt{u})}, \\
F_2 &= 1 + 4\frac{\phi_{d+4}(\|x\|/\sqrt{u}) - \phi_{d+2}(\|x\|/\sqrt{u})}{\phi_d(\|x\|/\sqrt{u})}
\end{aligned}$$

and

$$\phi_d(a) = a^{-d+2}\int_0^{a^2/2} s^{d/2-2}\exp(-s)ds \quad (a \geq 0).$$

These evaluations of the extended estimators follow the mixture representation (1.1) of Stein's prior. For comparison, we employed the empirical Bayes method $\hat{p}_{p-3}$ from the numerical analysis in [13]. The Kullback–Leibler risks are computed by taking the average of 5000 trials.

The simulation results are shown in Figure 1. As expected, the risk of $p_{\mathrm{S}}(y \mid x)$ is the smallest, whereas the risk of $p_{\mathrm{U}}(y \mid x)$, which is the only method in this experiment that does not employ a shrinkage prior, is the largest. The risk of $p_{\mathrm{U}}(y \mid x)$ is much larger than that of any other methods in Figure 1c. The four competitors that do not need complex representations are the two extended plug-in densities, the empirical Bayes predictive density, and $p_{\mathrm{U}}(y \mid x)$. Among these, the extended plug-in density $p(y; \hat{\mu}_\pi, \hat{\Sigma}_\pi)$ exhibits the best performance unless $\|\mu\|$ is very close to 0. The risk performance of the proposed extended



(a) $d = 10, u = 1, v = 1/10$ (b) $d = 20, u = 1, v = 1/10$

(c) $d = 100, u = 1, v = 1/10$ (d) $d = 10, u = 1, v = 1$

Fig 1: Kullback–Leibler risks of extended plug-in densities with $(\hat{\mu}_\pi, \hat{\xi}_\pi)$ and $(\hat{\mu}_\pi, \hat{\Sigma}_\pi)$, empirical Bayes method in [13], and Bayesian predictive densities p_{U} and p_{S}.

plug-in densities approaches that of $p_{\mathrm{S}}(y \mid x)$ more rapidly than the empirical Bayes as $\|\mu\|$ increases.

Figure 2 presents a magnification of the effects in Figure 1d of the choice of the extended models by showing the risk differences of $p(y; \hat{\mu}_\pi, \hat{\xi}_\pi)$, $p(y; \hat{\mu}_\pi, \hat{\Sigma}_\pi)$, and $\hat{p}_{\mathrm{S}}(y \mid x)$. The extended spaces to which extended plug-in densities $p(y; \hat{\mu}_\pi, \hat{\xi}_\pi)$ and $p(y; \hat{\mu}_\pi, \hat{\Sigma}_{t,\pi})$ belong are $\mathrm{N}_d(\mu, \xi I_d)$ and $\mathrm{N}_d(\mu, \Sigma)$, respectively, and their dimensions are $d + 1$ and $d + d(d+1)/2 = d^2/2 + (3/2)d$, respectively. The Bayesian predictive density $p_{\mathrm{S}}(y \mid x)$ does not belong to any of the finite-dimensional models. The risk comparison demonstrates that $p(y; \hat{\mu}_\pi, \hat{\Sigma}_\pi)$ performs slightly better than $p(y; \hat{\mu}_\pi, \hat{\xi}_\pi I_d)$, which suggests that a larger extended model results in a better performance of the predictive density with the

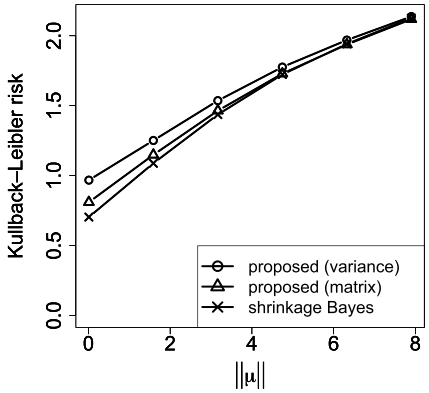


Fig 2: Kullback–Leibler risks of extended plug-in densities with $(\hat{\mu}_\pi, \hat{\xi}_\pi)$, $(\hat{\mu}_\pi, \hat{\Sigma}_\pi)$ and a Bayesian predictive density p_{S} when $d = 10, u = 1, v = 1$.

Bayes extended estimator.

5. Discussions

We have investigated the construction of Bayes extended estimators for the prediction of multivariate normal models with unknown mean μ. Investigating the admissibility of the proposed extended plug-in densities would be a future problem, and the admissibility would depend on the choice of superharmonic priors. In Theorem 3.1, we consider the condition $t \to 0$, and the dominance result for a large t should also be investigated in future research.

Predictive densities for linear regression models with unknown variance under α-divergence loss are studied in [9]. The development of Bayes extended estimators for α-divergence loss in similar contexts merits further investigation.

Acknowledgments

The authors would like to thank the associate editor and two reviewers for their constructive comments. This work was supported in part by JSPS KAKENHI Grant Numbers JP20K23316 and JP22H00510.

References

[1] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics* **42** 855–903. MR0286209

[2] Brown, L. D., George, E. I. and Xu, X. (2008). Admissible predictive density estimation. *The Annals of Statistics* **36** 1156–1170. MR2418653

[3] Fourdrinier, D., Marchand, É., Righi, A. and Strawderman, W. E. (2011). On improved predictive density estimation with parametric constraints. *Electronic Journal of Statistics* **5** 172–191. MR2792550
[4] George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Annals of Statistics* **34** 78–91. MR2275235
[5] George, E. I. and Xu, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory* **24** 528–544. MR2391619
[6] John, F. (1978). *Partial Differential Equations*, 3rd ed. Springer, New York. MR0514404
[7] Kobayashi, K. and Komaki, F. (2008). Bayesian shrinkage prediction for the regression problem. *Journal of Multivariate Analysis* **99** 1888–1905. MR2466542
[8] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415
[9] Maruyama, Y. and Strawderman, W. E. (2012). 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*, **8** 42–57. Institute of Mathematical Statistics. MR3202501
[10] Matsuda, T. and Komaki, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrika* **102** 843–854. MR3431557
[11] Okudo, M. and Komaki, F. (2021). Bayes extended estimators for curved exponential families. *IEEE Transactions on Information Theory* **67** 1088–1098. MR4232003
[12] Stein, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics* 345–381. MR0381062
[13] Xu, X. and Zhou, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *Journal of Multivariate Analysis* **102** 1417–1428. MR2819959