**MATHEMATICAL ENGINEERING
TECHNICAL REPORTS**

Improving on Singular Value Shrinkage Priors and Block-wise Stein Priors

Takeru MATSUDA and Fumiyasu KOMAKI

METR 2017–10 April 2017

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

The METR technical reports are published as a means to ensure timely dissemination of scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the authors or by other copyright holders, notwithstanding that they have offered their works here electronically. It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may not be reposted without the explicit permission of the copyright holder.

Improving on Singular Value Shrinkage Priors and Block-wise Stein Priors

Takeru MATSUDA[1] and Fumiyasu KOMAKI[1, 2]

[1]Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
[2]RIKEN Brain Science Institute
{matsuda,komaki}@mist.i.u-tokyo.ac.jp

April 10, 2017

Abstract

Matsuda and Komaki (2015) developed a singular value shrinkage prior for the mean matrix parameter in the matrix-variate normal distribution. This prior is superharmonic and therefore the generalized Bayes estimator and Bayesian predictive density based on this prior are minimax. In this study, we develop two types of priors that asymptotically dominate the singular value shrinkage prior in both estimation and prediction. The first type, which is motivated from the estimator of Efron and Morris (1976), adds scalar shrinkage whereas the second type adds column-wise shrinkage. When applied to multivariate linear regression, the second type accomplishes response selection or predictor selection. In addition to the singular value shrinkage prior, we show that the block-wise Stein prior is improved asymptotically in a similar way. Numerical results imply that these improvements hold even in finite samples.

1 Introduction

Suppose that we have a matrix observation $Y \in \mathbb{R}^{n \times m}$ whose entries are independent normal random variables $Y_{ij} \sim \mathrm{N}(M_{ij}, 1)$, where $M \in \mathbb{R}^{n \times m}$ is an unknown mean matrix. In the notation of matrix-variate normal distributions by Dawid (1981), it is expressed as $Y \sim \mathrm{N}_{n,m}(M, I_n, I_m)$, where I_k denotes the k-dimensional identity matrix. We assume $n - m - 1 > 0$. We consider the estimation of M under the

Frobenius loss

$$l(M, \hat{M}) = \|\hat{M} - M\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{M}_{ij} - M_{ij})^2.$$

Efron and Morris (1972) proposed

$$\hat{M}_{\mathrm{EM}} = Y \left(I_m - (n - m - 1)(Y^\top Y)^{-1} \right) \tag{1}$$

as an empirical Bayes estimator. They proved that $\hat{M}_{\mathrm{EM}}$ is minimax and dominates the maximum likelihood estimator $\hat{M} = Y$. Let $Y = U \Lambda V^\top$, $U \in \mathbb{R}^{n \times m}$, $V \in \mathbb{R}^{m \times m}$, $\Lambda = \mathrm{diag}(\sigma_1(Y), \ldots, \sigma_m(Y))$ be the singular value decomposition of Y, where $U^\top U = V^\top V = I_m$ and $\sigma_1(Y) \geq \cdots \geq \sigma_m(Y) \geq 0$ are the singular values of Y. Stein (1974) pointed out that $\hat{M}_{\mathrm{EM}}$ shrinks the singular values of Y to zero:

$$\hat{M}_{\mathrm{EM}} = U \hat{\Lambda} V^\top, \quad \hat{\Lambda} = \mathrm{diag}(\sigma_1(\hat{M}_{\mathrm{EM}}), \ldots, \sigma_m(\hat{M}_{\mathrm{EM}})),$$

where

$$\sigma_i(\hat{M}_{\mathrm{EM}}) = \left(1 - \frac{n - m - 1}{\sigma_i(Y)^2} \right) \sigma_i(Y) \quad (i = 1, \ldots, m).$$

Namely, $\hat{M}_{\mathrm{EM}}$ shrinks the singular values for each. Later, Efron and Morris (1976) proved that the modified estimator

$$\hat{M}_{\mathrm{MEM}} = Y \left(I_m - (n - m - 1)(Y^\top Y)^{-1} - \frac{m^2 + m - 2}{\mathrm{tr}(Y^\top Y)} I_m \right) \tag{2}$$

dominates $\hat{M}_{\mathrm{EM}}$. The modified estimator $\hat{M}_{\mathrm{MEM}}$ shrinks the singular values of Y stronger than the original estimator $\hat{M}_{\mathrm{EM}}$:

$$\sigma_i(\hat{M}_{\mathrm{MEM}}) = \left(1 - \frac{n - m - 1}{\sigma_i(Y)^2} - \frac{m^2 + m - 2}{\sum_{j=1}^{m} \sigma_j(Y)^2} \right) \sigma_i(Y) \quad (i = 1, \ldots, m). \tag{3}$$

In other words, $\hat{M}_{\mathrm{MEM}}$ adds scalar shrinkage to $\hat{M}_{\mathrm{EM}}$. Tsukuma (2008) provided a general method for improving matrix mean estimators by adding scalar shrinkage. We note that $\hat{M}_{\mathrm{EM}}$ and $\hat{M}_{\mathrm{MEM}}$ are not generalized Bayes estimators.

Recently, Matsuda and Komaki (2015) developed a singular value shrinkage prior

$$\pi_{\mathrm{SVS}}(M) = \det(M^\top M)^{-(n-m-1)/2} \tag{4}$$

and proved that it is superharmonic. This prior is a natural generalization of the Stein prior (Stein, 1974). The generalized Bayes estimator with respect to π_{SVS} is minimax and has similar properties to $\hat{M}_{\mathrm{EM}}$. This is an extension of the relationship between the James–Stein estimator and the Stein prior. Matsuda and Komaki

(2015) also discussed the application of the singular value shrinkage prior to multivariate linear regression. In multivariate linear regression, the regression coefficient matrix often has low rank and it is called the reduced-rank regression (Reinsel and Velu, 1998). Since the rank of a matrix is equal to the number of nonzero singular values, singular value shrinkage priors work effectively in such reduced-rank case.

Since $\hat{M}_{\mathrm{EM}}$ is dominated by $\hat{M}_{\mathrm{MEM}}$, it is expected that some generalized Bayes estimators have similar properties to $\hat{M}_{\mathrm{MEM}}$ and dominate the generalized Bayes estimator with respect to π_{SVS}. We show that the generalized Bayes estimator with respect to the prior

$$\pi_{\mathrm{MSVS1}}(M) = \pi_{\mathrm{SVS}}(M) \| M \|_{\mathrm{F}}^{-\gamma}$$

asymptotically dominates that with respect to π_{SVS} if $0 < \gamma < 2(m^2+m-2)$ (Theorem 1). Since π_{MSVS1} is a product of the singular value shrinkage prior and a prior shrinking to the zero matrix, the generalized Bayes estimator with respect to π_{MSVS1} adds scalar shrinkage to that with respect to π_{SVS}. In other words, the generalized Bayes estimator with respect to π_{MSVS1} not only shrinks singular values for each but also shrinks singular values overall like $\hat{M}_{\mathrm{MEM}}$. We also show that the generalized Bayes estimator with respect to the prior

$$\pi_{\mathrm{MSVS2}}(M) = \pi_{\mathrm{SVS}}(M) \prod_{j=1}^{m} \| M_{\cdot j} \|^{-\gamma_j}$$

asymptotically dominates that with respect to π_{SVS} if $0 < \gamma_j < 2m-2$ $(j = 1, \cdots, m)$ (Theorem 2). Here, $\| M_{\cdot j} \|$ denotes the norm of the j-th column vector of M. Since π_{MSVS2} is a product of the singular value shrinkage prior and a prior shrinking each column, the generalized Bayes estimator with respect to π_{MSVS2} adds column-wise shrinkage to that with respect to π_{SVS}. In particular, π_{MSVS2} attains response selection or predictor selection (Chen et al., 2012) when applied to multivariate linear regression.

In addition to the singular value shrinkage prior, we show that the block-wise Stein prior (Brown and Zhao, 2009) is also improved by additional shrinkage. Consider the problem of estimating $\theta \in \mathbb{R}^d$ from the observation $Y \sim \mathrm{N}_d(\theta, I_d)$. In many cases, the d-dimensional mean vector θ of a multivariate normal distribution is naturally split into several blocks: $\theta = (\theta^{(1)}, \cdots, \theta^{(B)})$ where the dimension of $\theta^{(b)}$ is d_b and $d = \sum_{b=1}^{B} d_b$. For example, when wavelet regression is reduced to the multivariate normal model, the mean vector has a block structure corresponding to the resolution of the wavelet basis (Clyde, Parmigiani and Vidakovic, 1998). In such case, the block-wise Stein prior is defined as

$$\pi_{\mathrm{BS}}(\theta) = \prod_{b=1}^{B} \| \theta^{(b)} \|^{R_b}, \quad R_b = -(d_b - 2)_+.$$

This prior puts the Stein prior on each block. The generalized Bayes estimator with respect to π_{BS} is minimax. Brown and Zhao (2009) proved that the generalized Bayes estimator $\hat{\theta}^{\pi_{\mathrm{BS}}}$ with respect to π_{BS} is dominated by estimators with additional shrinkage such as

$$\hat{\theta}(y) = \hat{\theta}^{\pi_{\mathrm{BS}}}(y) - \frac{R_{\#} + d - 2}{\|y\|^2} y,$$

where $R_{\#} = \sum_b R_b > 2 - d$. Namely, they added the James–Stein type shrinkage on the whole vector. However, their improved estimators are not generalized Bayes estimators. In Remark 3.2 of Brown and Zhao (2009), they conjectured that the block-wise Stein priors can be improved by multiplying Stein-type shrinkage priors. We show that their conjecture is true at least asymptotically. Namely, we prove that the generalized Bayes estimator with respect to the prior

$$\pi_{\mathrm{MBS}}(\theta) = \pi_{\mathrm{BS}}(\theta) \|\theta\|^{-\gamma}$$

asymptotically dominates that with respect to π_{BS} if $0 < \gamma < 2(R_{\#} + d - 2)$ (Theorem 3).

Recently, an interesting parallel has been found (George, Liang and Xu, 2012) between the point estimation of θ from $y \sim \mathrm{N}_d(\theta, I_d)$ under the quadratic loss and the predictive density estimation of $\widetilde{y} \sim \mathrm{N}_d(\theta, I_d)$ based on $y \sim \mathrm{N}_d(\theta, I_d)$ under the Kullback–Leibler loss

$$D(\widetilde{p}(\cdot \mid \theta), \hat{p}(\cdot \mid y)) = \int \widetilde{p}(\widetilde{y} \mid \theta) \log \frac{\widetilde{p}(\widetilde{y} \mid \theta)}{\hat{p}(\widetilde{y} \mid y)} \mathrm{d}\widetilde{y}.$$

The Bayesian predictive density based on a prior $\pi(\theta)$ is defined as

$$\hat{p}_{\pi}(\widetilde{y} \mid y) = \int \widetilde{p}(\widetilde{y} \mid \theta) \pi(\theta \mid y) \mathrm{d}\theta,$$

where $\pi(\theta \mid y)$ is the posterior distribution on θ given y. Komaki (2001) showed that the Bayesian predictive density based on the Stein prior $\pi(\theta) = \|\theta\|^{-(d-2)}$ dominates that based on the uniform prior, which is minimax. George, Liang and Xu (2006) extended this result and proved that Bayesian predictive densities based on superharmonic priors dominate those based on the uniform prior. Since matrix-variate normal distributions are special cases of vector-variate normal distributions, these results hold also in matrix-variate normal distributions. In particular, Bayesian predictive densities based on the singular value shrinkage priors dominate those based on the uniform prior (Matsuda and Komaki, 2015). We show that the proposed priors in this paper provide asymptotic improvement even in prediction. Here, asymptotic expansion of the Kullback–Leibler risk given by Komaki (2006) and Komaki (2015) is employed.

In section 2, formulas of the asymptotic expansion of risk in estimation and prediction are prepared. In section 3, priors that asymptotically dominate the singular value shrinkage prior are developed. Application to multivariate linear regression is also discussed. In section 4, priors that asymptotically dominate the block-wise Stein prior are developed.

2 Asymptotic expansion of risk

In this section, we prepare formulas of the asymptotic expansion of risk in estimation and prediction for vector-variate normal distributions. Although we consider vector-variate normal distributions for simplicity, the results in this section hold also in matrix-variate normal distributions.

2.1 Estimation

Consider the problem of estimating θ from the observation $Y^{(N)} \sim \mathrm{N}_d\left(\theta, N^{-1} I_d\right)$ under the quadratic loss $l(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2$. The generalized Bayes estimator $\hat{\theta}^{\pi}$ with respect to a prior $\pi(\theta)$ is expressed as

$$\hat{\theta}^{\pi}(y^{(N)}) = y^{(N)} + \frac{1}{N} \nabla \log m_{\pi}(y^{(N)}),$$

where

$$m_{\pi}(y^{(N)}) = \int p(y^{(N)} \mid \theta) \pi(\theta) \mathrm{d}\theta.$$

The difference of the quadratic risk between two generalized Bayes estimators is obtained as follows.

Lemma 1. *The difference of the quadratic risk between two generalized Bayes estimators $\hat{\theta}^{\pi_1}$ and $\hat{\theta}^{\pi_1 \pi_2}$ is expanded as*

$$\begin{aligned} &\mathrm{E}_{\theta}[\|\hat{\theta}^{\pi_1} - \theta\|^2] - \mathrm{E}_{\theta}[\|\hat{\theta}^{\pi_1\pi_2} - \theta\|^2] \\ =& -\frac{1}{N^2} \Big(2(\nabla \log \pi_1(\theta))^{\top} (\nabla \log \pi_2(\theta)) + \|\nabla \log \pi_2(\theta)\|^2 + 2\Delta \log \pi_2(\theta) \Big) + o(N^{-2}) \end{aligned} \tag{5}$$

Proof. From Stein's lemma, the quadratic risk is

$$\begin{aligned} &\mathrm{E}_{\theta}[\|\hat{\theta}^{\pi}(y^{(N)}) - \theta\|^2] \\ =& E_{\theta}[\|y^{(N)} - \theta\|^2] + 2E_{\theta}[(y^{(N)} - \theta)^{\top} \nabla \log m_{\pi}(y^{(N)})] + E_{\theta}[\|\nabla \log m_{\pi}(y^{(N)})\|^2] \\ =& \frac{d}{N} + \frac{1}{N^2} \mathrm{E}_{\theta} \left[\|\nabla \log m_{\pi}(y^{(N)})\|^2 + 2\Delta \log m_{\pi}(y^{(N)}) \right] \\ =& \frac{d}{N} + \frac{1}{N^2} \left(\|\nabla \log \pi(\theta)\|^2 + 2\Delta \log \pi(\theta) \right) + o(N^{-2}). \end{aligned} \tag{6}$$

Substituting $\pi = \pi_1$ and $\pi = \pi_1 \pi_2$ into (6) and taking difference, we obtain (5). □

2.2 Prediction

Consider the problem of predicting $\widetilde{Y} \sim \mathrm{N}_d(\theta, \widetilde{\Sigma})$ based on $Y^{(N)} \sim \mathrm{N}_d(\theta, N^{-1}\Sigma)$ by a predictive density $\hat{p}(\widetilde{y} \mid y^{(N)})$. The Bayesian predictive density based on a prior $\pi(\theta)$ is defined as

$$\hat{p}_\pi(\widetilde{y} \mid y^{(N)}) = \int \widetilde{p}(\widetilde{y} \mid \theta)\pi(\theta \mid y^{(N)})\mathrm{d}\theta,$$

where $\pi(\theta \mid y^{(N)})$ is the posterior distribution on θ given $y^{(N)}$.

2.2.1 Proportional covariance case

First, suppose that $\widetilde{\Sigma}$ is proportional to Σ (Komaki, 2001; George, Liang and Xu, 2006). Without loss of generality, we assume $\Sigma = \widetilde{\Sigma} = I_d$. From the results of Komaki (2006) on the asymptotic expansion of the Kullback-Leibler risk of Bayesian predictive densities, the difference of the Kullback–Leibler risk between two Bayesian predictive densities is obtained as follows.

Lemma 2. *The difference of the Kullback–Leibler risk between two Bayesian predictive densities $p_{\pi_1}(\widetilde{y} \mid y^{(N)})$ and $p_{\pi_1\pi_2}(\widetilde{y} \mid y^{(N)})$ is expanded as*

$$\begin{aligned} &\mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1}(\widetilde{y} \mid y^{(N)}))] - \mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1\pi_2}(\widetilde{y} \mid y^{(N)}))] \\ &= -\frac{1}{2N^2}\left(2(\nabla \log \pi_1(\theta))^\top(\nabla \log \pi_2(\theta)) + \|\nabla \log \pi_2(\theta)\|^2 + 2\Delta \log \pi_2(\theta)\right) + o(N^{-2}). \end{aligned} \tag{7}$$

Proof. For the normal model with known covariance, the information geometrical quantities (Amari, 1985) are

$$g_{ij} = g^{ij} = \delta_{ij}, \quad \Gamma^k_{ij} = 0, \quad T_{ijk} = 0.$$

Also, the Jeffreys prior coincides with the uniform prior $\pi_{\mathrm{J}}(\theta) \equiv 1$. Therefore, from equation (3) of Komaki (2006), the Kullback–Leibler risk of the Bayesian predictive density $p_\pi(\widetilde{y} \mid y^{(N)})$ based on a prior $\pi(\theta)$ is

$$\begin{aligned} &\mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_\pi(\widetilde{y} \mid y^{(N)}))] \\ =&\frac{d}{2N} + \frac{1}{2N^2}\|\nabla \log \pi(\theta)\|^2 + \frac{1}{N^2}\Delta \log \pi(\theta) + g(\theta) + o(N^{-2}), \end{aligned} \tag{8}$$

where $g(\theta)$ is a function independent of $\pi(\theta)$. Substituting $\pi = \pi_1$ and $\pi = \pi_1\pi_2$ into (8) and taking difference, we obtain (7). □

Since the second-order term in (7) is exactly half of the second-order term in (5), asymptotic improvement in estimation

$$\lim_{N\to\infty} N^2 \left(\mathrm{E}_\theta[\|\hat{\theta}^{\pi_1\pi_2} - \theta\|^2] - \mathrm{E}_\theta[\|\hat{\theta}^{\pi_1} - \theta\|^2] \right) \leq 0$$

immediately implies asymptotic improvement in prediction

$$\lim_{N\to\infty} N^2 \left(\mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1\pi_2}(\widetilde{y} \mid y^{(N)}))] - \mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1}(\widetilde{y} \mid y^{(N)}))] \right) \leq 0$$

for every θ.

2.2.2 General covariance case

Next, we consider the general setting where Σ and $\widetilde{\Sigma}$ are not necessarily proportional (Kobayashi and Komaki, 2008; George and Xu, 2008). Komaki (2015) obtained the asymptotic expansion of the Kullback-Leibler risk of Bayesian predictive densities when the observation and target to be predicted have different distributions with common parameters. He introduced a new metric on the manifold of parametric models, which is called the predictive metric. For the present problem, the coefficients of the predictive metric g° do not depend on θ and $g^\circ = N^2\Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1}$ (Example 1 in Komaki (2015)). The nabla operator ∇° and the Laplacian form Δ° are defined with respect to the predictive metric g° based on the usual framework of Riemannian geometry. Using these geometrical quantities, the difference of the Kullback–Leibler risk between two Bayesian predictive densities is obtained as follows.

Lemma 3. *The difference of the Kullback–Leibler risk between two Bayesian predictive densities* $p_{\pi_1}(\widetilde{y} \mid y^{(N)})$ *and* $p_{\pi_1\pi_2}(\widetilde{y} \mid y^{(N)})$ *is expanded as*

$$\begin{aligned} &\mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1}(\widetilde{y} \mid y^{(N)}))] - \mathrm{E}_\theta[D(p(\widetilde{y} \mid \theta), p_{\pi_1\pi_2}(\widetilde{y} \mid y^{(N)}))] \\ &= -\frac{1}{2N^2} \left(2(\nabla^\circ \log \pi_1(\theta))^\top (\nabla^\circ \log \pi_2(\theta)) + \|\nabla^\circ \log \pi_2(\theta)\|^2 + 2\Delta^\circ \log \pi_2(\theta) \right) + o(N^{-2}). \end{aligned}$$

Proof. See Theorem 1 in Komaki (2015). □

3 Improving on singular value shrinkage priors

In this section, we develop priors that asymptotically dominate the singular value shrinkage prior (4) in estimation and prediction. Two types of priors are proposed by introducing additional shrinkage: scalar shrinkage and column-wise shrinkage. In section 3.1 and section 3.2, we consider the estimation of M and prediction of $\widetilde{Y} \sim \mathrm{N}_{n,m}(M, I_n, I_m)$ based on $Y^{(N)} \sim \mathrm{N}_{n,m}(M, N^{-1}I_n, I_m)$. In section 3.3, we consider the general setting of multivariate linear regression: estimation of B and prediction of $\widetilde{Y} \sim \mathrm{N}_{m,q}(\widetilde{X}B, I_m, \Sigma)$ based on $Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$.

3.1 Addition of scalar shrinkage

The generalized Bayes estimator with respect to the singular value shrinkage prior π_{SVS} in (4) has similar properties to the Efron–Morris estimator $\hat{M}_{\mathrm{EM}}$ in (1). However, the estimator $\hat{M}_{\mathrm{EM}}$ is dominated by the estimator $\hat{M}_{\mathrm{MEM}}$ in (2). Therefore, we can reasonably expect that some generalized Bayes estimators have similar properties to $\hat{M}_{\mathrm{MEM}}$ and dominate the generalized Bayes estimator with respect to π_{SVS}. From the singular value decomposition form of $\hat{M}_{\mathrm{MEM}}$ in (3), we construct priors by adding scalar shrinkage to π_{SVS}:

$$\pi_{\mathrm{MSVS1}}(M) = \pi_{\mathrm{SVS}}(M) \| M \|_{\mathrm{F}}^{-\gamma}. \tag{9}$$

The following Theorem proves that π_{MSVS1} asymptotically dominates π_{SVS} in estimation and prediction.

Theorem 1. *(i) If* $0 < \gamma < 2(m^2+m-2)$*, then the generalized Bayes estimator with respect to* π_{MSVS1} *in* (9) *asymptotically dominates the generalized Bayes estimator with respect to* π_{SVS} *under the quadratic risk.*

(ii) If $0 < \gamma < 2(m^2+m-2)$*, then the Bayesian predictive density based on* π_{MSVS1} *in* (9) *asymptotically dominates the Bayesian predictive density based on* π_{SVS} *under the Kullback–Leibler risk.*

Proof. Let $K = M^\top M$. From

$$\frac{\partial K_{bc}}{\partial M_{ia}} = \delta_{ac} M_{ib} + \delta_{ab} M_{ic} \tag{10}$$

and

$$\frac{\partial}{\partial K_{ab}} \det K = K^{ab} \det K,$$

we have

$$\frac{\partial}{\partial M_{ia}} \det K = \sum_{b,c} \frac{\partial K_{bc}}{\partial M_{ia}} \frac{\partial}{\partial K_{bc}} \det K = 2 \sum_b M_{ib} K^{ab} \det K,$$

where K^{ab} is the (a,b)th entry of the inverse matrix of K^{-1}. Therefore,

$$\frac{\partial}{\partial M_{ia}} \log \pi_{\mathrm{SVS}}(M) = -(n-m-1) \sum_b M_{ib} K^{ab}. \tag{11}$$

Let

$$\pi_{\mathrm{S}}(M) = \| M \|_{\mathrm{F}}^{-\gamma} = (\mathrm{tr} K)^{-\gamma/2}.$$

From (10),

$$\frac{\partial}{\partial M_{ia}} \mathrm{tr} K = 2 M_{ia}.$$

Therefore,

$$\frac{\partial}{\partial M_{ia}} \log \pi_{\mathrm{S}}(M) = -\gamma M_{ia} (\mathrm{tr} K)^{-1}, \tag{12}$$

$$\frac{\partial^2}{\partial M_{ia}^2} \log \pi_{\mathrm{S}}(M) = -\gamma (\mathrm{tr} K - 2M_{ia}^2)(\mathrm{tr} K)^{-2}. \tag{13}$$

From (11), (12), and (13), we obtain

$$(\nabla \log \pi_{\mathrm{SVS}}(M))^\top (\nabla \log \pi_{\mathrm{S}}(M)) = \gamma m(n-m-1)(\mathrm{tr} K)^{-1},$$

$$(\nabla \log \pi_{\mathrm{S}}(M))^\top (\nabla \log \pi_{\mathrm{S}}(M)) = \gamma^2 (\mathrm{tr} K)^{-1},$$

$$\Delta \log \pi_{\mathrm{S}}(M) = -\gamma(nm-2)(\mathrm{tr} K)^{-1}.$$

Then, the difference of the quadratic risk (5) between two generalized Bayes estimators with respect to π_{SVS} and π_{MSVS1} is

$$\begin{aligned} & \mathrm{E}_M[\|\hat{M}^{\pi_{\mathrm{SVS}}} - M\|_{\mathrm{F}}^2] - \mathrm{E}_M[\|\hat{M}^{\pi_{\mathrm{MSVS1}}} - M\|_{\mathrm{F}}^2] \\ = & -\frac{1}{N^2}\gamma(\gamma - 2(m^2+m-2)) + o(N^{-2}). \end{aligned} \tag{14}$$

Also, the second-oder term in the difference of the Kullback–Leibler risk (7) between two Bayesian predictive densities with respect to π_{SVS} and π_{MSVS1} is exactly half of that in (14). Since (14) is positive when $0 < \gamma < 2(m^2+m-2)$, we obtain the Theorem. □

From (14), the choice $\gamma = m^2+m-2$ is optimal. This choice corresponds to $\hat{M}_{\mathrm{MEM}}$. Fig. 1 presents the Kullback–Leibler risk functions of Bayesian predictive densities when $m=3$, $n=5$, $\sigma_1 = 10$, $\sigma_3 = 0$ and $N=1$. Fig. 2 presents the Kullback–Leibler risk functions of Bayesian predictive densities when $m=3$, $n=10$, $\sigma_2 = \sigma_3 = 0$ and $N=1$. Here, the Stein prior is the prior $\pi(M) = \|M\|_{\mathrm{F}}^{2-mn}$. These figures imply that π_{MSVS1} with $\gamma = m^2+m-2$ dominates π_{SVS} even in finite samples.

3.2 Addition of column-wise shrinkage

In Theorem 1, we added scalar shrinkage to π_{SVS}. We can also construct priors by adding column-wise shrinkage to π_{SVS}:

$$\pi_{\mathrm{MSVS2}}(M) = \pi_{\mathrm{SVS}}(M) \prod_{j=1}^m \|M_{\cdot j}\|^{-\gamma_j}, \tag{15}$$

where $\|M_{\cdot j}\|$ denotes the norm of the j-th column vector of M. The following Theorem proves that π_{MSVS2} also asymptotically dominates π_{SVS} in estimation and prediction.

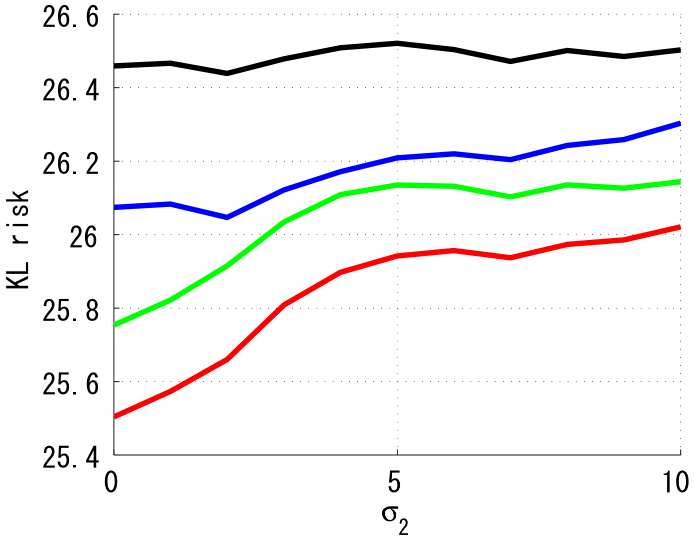

Figure 1: Risk functions of Bayesian predictive densities when $m = 3$, $n = 5$, $\sigma_1 = 10$, $\sigma_3 = 0$, and $N = 1$. black: uniform prior, blue: the Stein prior, green: the singular value shrinkage prior π_{SVS}, red: the prior π_{MSVS1} with $\gamma = m^2 + m - 2$.

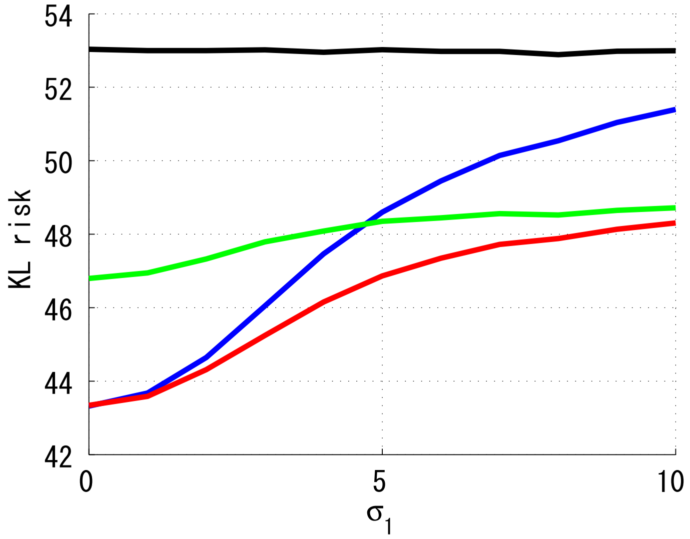

Figure 2: Risk functions of Bayesian predictive densities when $m = 3$, $n = 10$, $\sigma_2 = \sigma_3 = 0$, and $N = 1$. black: uniform prior, blue: the Stein prior, green: the singular value shrinkage prior π_{SVS}, red: the prior π_{MSVS1} with $\gamma = m^2 + m - 2$.

Theorem 2. *(i) If $0 < \gamma_j < 2m - 2$ $(j = 1, \cdots, m)$, then the generalized Bayes estimator with respect to π_{MSVS2} in* (15) *asymptotically dominates the generalized Bayes estimator with respect to π_{SVS} under the quadratic risk.*

(ii) If $0 < \gamma_j < 2m - 2$ $(j = 1, \cdots, m)$, then the Bayesian predictive density based on π_{MSVS2} in (15) *asymptotically dominates the Bayesian predictive density based on π_{SVS} under the Kullback–Leibler risk.*

Proof. Let

$$\pi_{\mathrm{CS}}(M) = \prod_{j=1}^{m} \|M_{\cdot j}\|^{-\gamma_j}.$$

Then,

$$\frac{\partial}{\partial M_{ia}} \log \pi_{\mathrm{CS}}(M) = -\gamma_a M_{ia} \|M_{\cdot a}\|^{-2}, \tag{16}$$

$$\frac{\partial^2}{\partial M_{ia}^2} \log \pi_{\mathrm{CS}}(M) = -\gamma_a \left(\|M_{\cdot a}\|^2 - 2M_{ia}^2 \right) \|M_{\cdot a}\|^{-4}. \tag{17}$$

From (11), (16), and (17), we obtain

$$(\nabla \log \pi_{\mathrm{SVS}}(M))^\top (\nabla \log \pi_{\mathrm{CS}}(M)) = (n - m - 1) \sum_a \gamma_a \|M_{\cdot a}\|^{-2},$$

$$(\nabla \log \pi_{\mathrm{CS}}(M))^\top (\nabla \log \pi_{\mathrm{CS}}(M)) = \sum_a \gamma_a^2 \|M_{\cdot a}\|^{-2},$$

$$\Delta \log \pi_{\mathrm{CS}}(M) = -(n-2) \sum_a \gamma_a \|M_{\cdot a}\|^{-2}.$$

Then, the difference of the quadratic risk (5) between two generalized Bayes estimators with respect to π_{SVS} and π_{MSVS2} is

$$\begin{aligned} &\mathrm{E}_M[\|\hat{M}^{\pi_{\mathrm{SVS}}} - M\|_{\mathrm{F}}^2] - \mathrm{E}_M[\|\hat{M}^{\pi_{\mathrm{MSVS2}}} - M\|_{\mathrm{F}}^2] \\ &= -\frac{1}{N^2} \sum_a \gamma_a (\gamma_a - 2(m-1)) \|M_{\cdot a}\|^{-2} + o(N^{-2}). \end{aligned} \tag{18}$$

Also, the second-order term in the difference of the Kullback–Leibler risk (7) between two Bayesian predictive densities with respect to π_{SVS} and π_{MSVS2} is exactly half of that in (18). Since (18) is positive when $0 < \gamma_a < 2m - 2$ $(a = 1, \cdots, m)$, we obtain the Theorem. □

From (18), the choice $\gamma_1 = \cdots = \gamma_m = m - 1$ is optimal. Fig. 3 presents the Kullback–Leibler risk functions of Bayesian predictive densities when $m = 3$, $n = 5$,

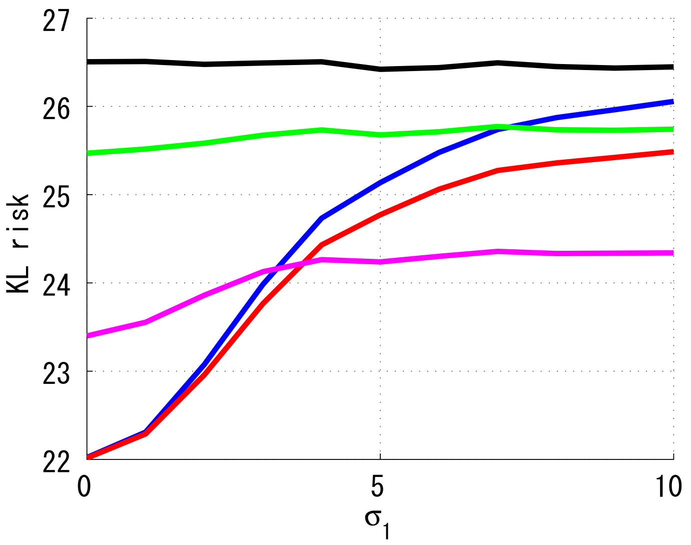

Figure 3: Risk functions of Bayesian predictive densities for mean matrices of the form (19) when $N = 1$. black: uniform prior, blue: the Stein prior, green: the singular value shrinkage prior π_{SVS}, red: the prior π_{MSVS1} with $\gamma = m^2 + m - 2$, magenta: the prior π_{MSVS2} with $\gamma_1 = \cdots = \gamma_m = m - 1$.

and $N = 1$. Here, we consider mean matrices of the form

$$M = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{19}$$

where $\sigma_2 = \sigma_3 = 0$. Although π_{MSVS1} is better than π_{SVS} when σ_1 is small, the risk of π_{MSVS1} becomes almost the same as that of π_{SVS} as σ_1 increases. On the other hand, π_{MSVS2} performs better than π_{SVS} regardless of the value of σ_1. In particular, π_{MSVS2} provides larger risk reduction than π_{MSVS1} when σ_1 is large. This is because π_{MSVS2} shrinks each column vector separately while π_{MSVS1} shrinks all the column vectors as a whole. Fig. 3 implies that π_{MSVS2} with $\gamma_1 = \cdots = \gamma_m = m - 1$ dominates π_{SVS} even in finite samples.

3.3 Application to multivariate linear regression

Now, we consider the general setting of multivariate linear regression. We use notations from Gupta and Nagar (2000). The size of a matrix $A \in \mathbb{R}^{p \times q}$ is indicated by writing $A(p \times q)$. The vectorization of $A(p \times q)$ is the $pq \times 1$ vector defined by

$$\mathrm{vec}(A) = (a_{11}, \dots, a_{p1}, a_{12}, \dots, a_{p2}, \dots, a_{1q}, \dots, a_{pq})^\top,$$

and the Kronecker product $A \otimes B$ of two matrices $A(p \times q) = (a_{ij})$ and $B(r \times s) = (b_{ij})$ is the $pr \times qs$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{pmatrix}.$$

3.3.1 Estimation

Consider the estimation of B from

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma),$$

where $X(n \times p)$ is a matrix of explanatory variables, $Y(n \times q)$ is a matrix of response variables, $B(p \times q)$ is a regression coefficient matrix, and Σ is a known covariance matrix. We assume $n \geq p$. By sufficiency reduction, the above model is reduced to

$$(X^\top X)^{-1} X^\top Y \sim \mathrm{N}_{p,q}(B, (X^\top X)^{-1}, \Sigma).$$

We adopt the invariant loss $l(B, \hat{B}) = \mathrm{tr}(\hat{B} - B)\Sigma^{-1}(\hat{B} - B)^\top (X^\top X)$.

Assume $p - q - 1 > 0$. From invariance, the generalized Bayes estimator with respect to $\pi_0(B) = \pi_{\mathrm{SVS}}((X^\top X)^{1/2} B \Sigma^{-1/2}) \propto \pi_{\mathrm{SVS}}((X^\top X)^{1/2} B)$ is minimax. Also, the generalized Bayes estimator with respect to $\pi_1(B) = \pi_{\mathrm{MSVS1}}((X^\top X)^{1/2} B \Sigma^{-1/2})$ and $\pi_2(B) = \pi_{\mathrm{MSVS2}}((X^\top X)^{1/2} B \Sigma^{-1/2})$ asymptotically dominate that with respect to $\pi_0(B)$. Here, asymptotics refer to the situation $\Sigma \otimes (X^\top X)^{-1} \to 0$. If the error is independent among q response variables, then we have $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_q^2)$ where $\sigma_1^2, \cdots, \sigma_q^2$ are known variances. In this case,

$$((X^\top X)^{1/2} B \Sigma^{-1/2})_{ij} = ((X^\top X)^{1/2} B)_{ij} \sigma_j^{-1}.$$

Thus,

$$\pi_2(B) \propto \pi_{\mathrm{SVS}}((X^\top X)^{1/2} B) \prod_{j=1}^{q} \| (X^\top X)^{1/2} B_{\cdot j} \|^{-\gamma_j}.$$

Therefore, $\pi_2(B)$ accomplishes column-wise shrinkage. In the context of multivariate linear regression, column-wise shrinkage on B corresponds to response selection (Chen et al., 2012). Namely, response variables that are not related to any explanatory variables are ignored by shrinking the corresponding regression coefficients.

Assume $q-p-1>0$. In this case, the generalized Bayes estimator with respect to $\pi_0(B) = \pi_{\mathrm{SVS}}(\Sigma^{-1/2}B^\top (X^\top X)^{1/2}) \propto \pi_{\mathrm{SVS}}(\Sigma^{-1/2}B^\top)$ is minimax and the generalized Bayes estimator with respect to $\pi_1(B) = \pi_{\mathrm{MSVS1}}(\Sigma^{-1/2}B^\top (X^\top X)^{1/2})$ and $\pi_2(B) = \pi_{\mathrm{MSVS2}}(\Sigma^{-1/2}B^\top (X^\top X)^{1/2})$ asymptotically dominate that with respect to $\pi_0(B)$. If $X^\top X$ is diagonal, then $\pi_2(B)$ accomplishes row-wise shrinkage. In the context of multivariate linear regression, row-wise shrinkage on B corresponds to predictor selection (Chen et al., 2012). Namely, explanatory variables that are not related to any response variables are ignored by shrinking the corresponding regression coefficients.

3.3.2 Prediction

Consider the prediction of

$$\widetilde{Y} \sim \mathrm{N}_{m,q}(\widetilde{X}B, I_m, \Sigma)$$

based on

$$Y \sim \mathrm{N}_{n,q}(XB, I_n, \Sigma)$$

by a predictive density $\hat{p}(\widetilde{Y} \mid Y)$, where $X(n \times p)$ and $\widetilde{X}(m \times p)$ are explanatory variables, $Y(n \times q)$ and $\widetilde{Y}(m \times q)$ are response variables, $B(p \times q)$ is a regression coefficient matrix, and Σ is a known covariance matrix. We assume $n \geq p$. Kobayashi and Komaki (2008) and George and Xu (2008) considered the same setting with $q=1$. By sufficiency reduction, this problem is reduced to the prediction of

$$(\widetilde{X}^\top \widetilde{X})^\dagger \widetilde{X}^\top \widetilde{Y} \sim \mathrm{N}_{p,q}(B, (\widetilde{X}^\top \widetilde{X})^\dagger, \Sigma) \tag{20}$$

based on

$$(X^\top X)^{-1} X^\top Y \sim \mathrm{N}_{p,q}(B, (X^\top X)^{-1}, \Sigma), \tag{21}$$

where $A^\dagger$ is the Moore-Penrose pseudo-inverse matrix of A.

Matsuda and Komaki (2015) investigated the prediction of $\widetilde{Y} \sim \mathrm{N}_{n,m}(M, \widetilde{C}, \widetilde{Q})$ based on $Y \sim \mathrm{N}_{n,m}(M, C, Q)$ where $n-m-1>0$. Let

$$Q_1 = \left\{ (Q \otimes C)^{-1} + (\widetilde{Q} \otimes \widetilde{C})^{-1} \right\}^{-1}, \quad Q_2 = Q \otimes C$$

and write the diagonalization of $Q_1^{1/2} Q_2^{-1} Q_1^{1/2}$ as $Q_1^{1/2} Q_2^{-1} Q_1^{1/2} = U^\top \Lambda U$, where U is an orthogonal matrix and Λ is a diagonal matrix. Let $A^* = Q_1^{1/2} U^\top (\Lambda^{-1} - I_m)^{1/2}$.

Lemma 4. *(Matsuda and Komaki, 2015) If $\pi\left[\mathrm{vec}^{-1}\left\{A^* \mathrm{vec}(M)\right\}\right]$ is superharmonic as a function of M, the Bayesian predictive density based on $\pi(M)$ dominates that based on the uniform prior.*

For example, the prior

$$\pi(M) = \pi_{\mathrm{SVS}}\left[\mathrm{vec}^{-1}\left\{(A^*)^{-1}\mathrm{vec}(M)\right\}\right] \tag{22}$$

satisfies the condition of Lemma 4.

Now, our present problem (20) and (21) with $p-q-1>0$ corresponds to $C = (X^\top X)^{-1}$, $\widetilde{C} = (\widetilde{X}^\top \widetilde{X})^{-1}$ and $Q = \widetilde{Q} = \Sigma$. Since $Q = \widetilde{Q}$, we have $Q_1 = Q \otimes (C^{-1} + \widetilde{C}^{-1})^{-1}$ and therefore $Q_1^{1/2} Q_2^{-1} Q_1^{1/2} = I_q \otimes (C^{-1} + \widetilde{C}^{-1})^{-1/2} C^{-1} (C^{-1} + \widetilde{C}^{-1})^{-1/2}$. Thus, letting $(C^{-1} + \widetilde{C}^{-1})^{-1/2} C^{-1} (C^{-1} + \widetilde{C}^{-1})^{-1/2} = V^\top K V$ be the diagonalization, we obtain $U = I_q \otimes V$, $\Lambda = I_q \otimes K$, and $A^* = Q^{1/2} \otimes (C^{-1} + \widetilde{C}^{-1})^{-1/2} V^\top (K^{-1} - I_p)^{1/2}$. Here, we used $(A \otimes B)(C \otimes D) = AC \otimes BD$, $(A \otimes B)^\top = A^\top \otimes B^\top$, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (Gupta and Nagar, 2000). Therefore, the prior (4) becomes

$$\begin{aligned}\pi_0(B) &= \pi_{\mathrm{SVS}}((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B \Sigma^{-1/2}) \\ &\propto \pi_{\mathrm{SVS}}((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B).\end{aligned}$$

From Lemma 4, the Bayesian predictive density based on $\pi_0(B)$ is minimax. By introducing additional shrinkage, we construct two priors:

$$\pi_1(B) = \pi_{\mathrm{MSVS1}}((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B \Sigma^{-1/2})$$

and

$$\pi_2(B) = \pi_{\mathrm{MSVS2}}((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B \Sigma^{-1/2}).$$

From similar arguments to Theorem 1 and Theorem 2, the Bayesian predictive densities based on $\pi_1(B)$ and $\pi_2(B)$ asymptotically dominate that based on $\pi_0(B)$. Here, asymptotics refer to the situation $\Sigma \otimes (X^\top X)^{-1} \to 0$. If the error is independent among q response variables, then we have $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_q^2)$ where $\sigma_1^2, \cdots, \sigma_q^2$ are known variances. In this case, the prior $\pi_2(B)$ is simplified as

$$\begin{aligned}\pi_2(B) \propto\ & \pi_{\mathrm{SVS}}((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B) \\ & \times \prod_{j=1}^q \|((K^{-1} - I_p)^{-1/2} V (X^\top X + \widetilde{X}^\top \widetilde{X})^{1/2} B)_{\cdot j}\|^{-\gamma_j}.\end{aligned}$$

Therefore, the prior $\pi_2(B)$ accomplishes column-wise shrinkage, which corresponds to response selection.

4 Improving on the block-wise Stein priors

In this section, we develop priors that asymptotically dominate the block-wise Stein prior in estimation and prediction. We consider the estimation of θ and prediction of $\widetilde{Y} \sim \mathrm{N}_d(\theta, I_d)$ based on the observation $Y^{(N)} \sim \mathrm{N}_d(\theta, N^{-1}I_d)$. Suppose that the d-dimensional mean vector θ is naturally split into B blocks $\theta^{(1)}, \cdots, \theta^{(B)}$ with size $d_1, \cdots, d_B$, where $d = \sum_b d_b$. Then, the block-wise Stein prior is defined as

$$\pi_{\mathrm{BS}}(\theta) = \prod_{b=1}^{B} \|\theta^{(b)}\|^{R_b}, \quad R_b = -(d_b - 2)_+.$$

and it is superharmonic. This prior puts the Stein prior on each block. The generalized Bayes estimator with respect to π_{BS} is minimax. We put $R_{\#} = \sum_b R_b > 2 - d$.

Brown and Zhao (2009) studied the admissibility and quasi-admissibility properties of block-wise shrinkage estimators. They showed that the generalized Bayes estimator $\hat{\theta}^{\pi_{\mathrm{BS}}}$ with respect to π_{BS} is dominated by estimators with additional James-Stein type shrinkage such as

$$\hat{\theta}(y) = \hat{\theta}^{\pi_{\mathrm{BS}}}(y) - \frac{R_{\#} + d - 2}{\|y\|^2} y.$$

From this result, in Remark 3.2, they conjectured that the block-wise Stein priors can be improved by multiplying Stein-type shrinkage priors. Following their conjecture, we construct priors by adding scalar shrinkage to the block-wise Stein priors:

$$\pi_{\mathrm{MBS}}(\theta) = \pi_{\mathrm{BS}}(\theta)\|\theta\|^{-\gamma}. \tag{23}$$

The following Theorem proves that π_{MBS} asymptotically dominates π_{BS} in estimation and prediction.

Theorem 3. *(i) If $0 < \gamma < 2(R_{\#} + d - 2)$, then the generalized Bayes estimator with respect to the prior π_{MBS} in (23) asymptotically dominates the generalized Bayes estimator with respect to the block-wise Stein prior π_{BS} under the quadratic risk.*

(ii) If $0 < \gamma < 2(R_{\#} + d - 2)$, then the Bayesian predictive density based on the prior π_{MBS} in (23) asymptotically dominates the Bayesian predictive density based on the block-wise Stein prior π_{BS} under the Kullback–Leibler risk.

Proof. Let

$$\pi_{\mathrm{S}}(\theta) = \|\theta\|^{-\gamma}.$$

From the definition, we obtain

$$(\nabla \log \pi_{\mathrm{BS}}(\theta))^\top (\nabla \log \pi_{\mathrm{S}}(\theta)) = -\gamma R_{\#} \|\theta\|^{-2},$$

$$(\nabla \log \pi_{\mathrm{S}}(\theta))^\top (\nabla \log \pi_{\mathrm{S}}(\theta)) = \gamma^2 \|\theta\|^{-2},$$

$$\Delta \log \pi_{\mathrm{S}}(\theta) = -\gamma(d-2)\|\theta\|^{-2}.$$

Then, the difference of the quadratic risk (5) between two generalized Bayes estimators with respect to π_{BS} and π_{MBS} is

$$\begin{aligned} &\mathrm{E}_\theta[\|\hat{\theta}^{\pi_{\mathrm{BS}}} - \theta\|^2] - \mathrm{E}_\theta[\|\hat{\theta}^{\pi_{\mathrm{MBS}}} - \theta\|^2] \\ =& -\frac{1}{N^2}\gamma(\gamma - 2(R_\# + d - 2)) + o(N^{-2}). \end{aligned} \tag{24}$$

Also, the second-order term in the difference of the Kullback–Leibler risk (7) between two Bayesian predictive densities with respect to π_{BS} and π_{MBS} is exactly half of that in (24). Since (24) is positive when $0 < \gamma < 2(R_\# + d - 2)$, we obtain the Theorem. □

From (24), the choice $\gamma = R_\# + d - 2$ is optimal. Fig. 4 presents the Kullback–Leibler risk functions of Bayesian predictive densities when $d = 9$, $d_1 = d_2 = d_3 = 3$, and $N = 1$. Here, we consider mean parameters of the form $\theta = (t, t, \cdots, t)^\top$ with $0 \le t \le 3$. Fig. 5 presents the Kullback–Leibler risk functions of Bayesian predictive densities when $d = 9$, $d_1 = d_2 = d_3 = 3$, and $N = 1$. Here, we consider mean parameters of the form $\theta = t(0, 0, 0, 1, 1, 1, 2, 2, 2)$ with $0 \le t \le 3$. These figures imply that the prior π_{MBS} with $\gamma = R_\# + d - 2$ dominates the block-wise Stein prior π_{BS} even in finite samples. The risk reduction by the prior π_{MBS} becomes larger when the true mean is closer to the origin.

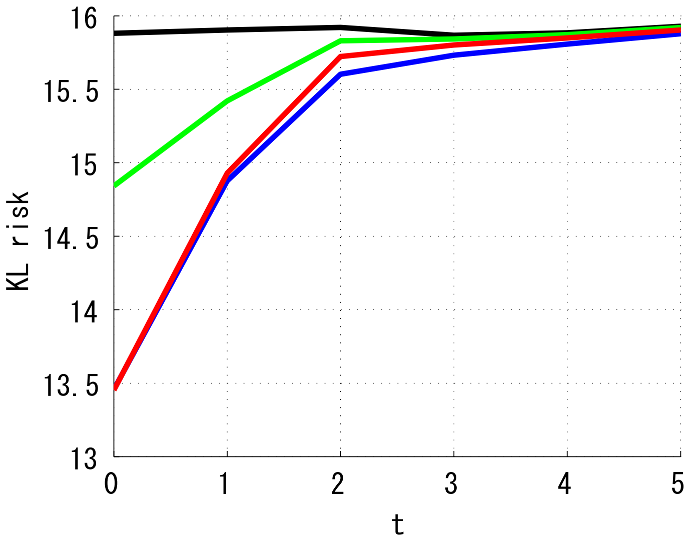

Figure 4: Risk functions of Bayesian predictive densities at $\theta = (t, t, \cdots, t)^\top$ when $d = 9$, $d_1 = d_2 = d_3 = 3$, and $N = 1$. black: uniform prior, blue: the Stein prior, green: the block-wise Stein prior π_{BS}, red: the prior π_{MBS} with $\gamma = R_\# + d - 2$.

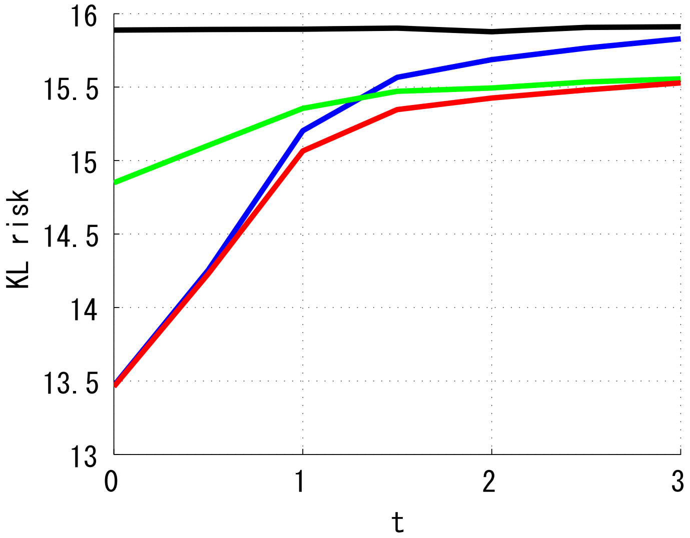

Figure 5: Risk functions of Bayesian predictive densities at $\theta = {}^t(0,0,0,1,1,1,2,2,2)$ when $d = 9$, $d_1 = d_2 = d_3 = 3$, and $N = 1$. black: uniform prior, blue: the Stein prior, green: the block-wise Stein prior π_{BS}, red: the prior π_{MBS} with $\gamma = R_{\#} + d - 2$.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 14J09148 and 26280005.

References

Amari, S. (1985). *Differential-Geometrical Methods in Statistics.* New York: Springer.

Brown, L. D. & Zhao, L. H. (2009). Estimators for Gaussian models having a block-wise structure. *Stat. Sin.* **19**, 885–903.

Chen, K., Chan, K-S. & Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *J. Roy. Statist. Soc. Ser. B* **74**, 203–221.

Clyde, B., Parmigiani, G. & Vidakovic, B (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–401.

DAWID, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.

EFRON, B. & MORRIS, C. (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* **59**, 335–347.

EFRON, B. & MORRIS, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4**, 22–32.

GEORGE, E. I., LIANG, F. & XU, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.* **34**, 78–91.

GEORGE, E. I. & XU, X. (2008). Predictive density estimation for multiple regression. *Econ. Theory* **24**, 528–44.

GEORGE, E. I., LIANG, F. & XU, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Stat. Sci.* **27**, 82–94.

GUPTA, A. K. & NAGAR, D. K. (2000). *Matrix Variate Distributions.* New York: Chapman & Hall.

KOBAYASHI, K. & KOMAKI, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate. Anal.* **99**, 1888–1905.

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

KOMAKI, F. (2006). Shrinkage priors for Bayesian prediction. *Ann. Statist.* **34**, 808–819.

KOMAKI, F. (2015). Asymptotic properties of Bayesian predictive densities when the distributions of data and target variables are different. *Bayesian Anal.* **10**, 31–51.

MATSUDA, T. & KOMAKI, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrika* **102**, 843–854.

REINSEL, G. C. & VELU, R. P. (1998). *Multivariate Reduced-Rank Regression.* New York: Springer-Verlag.

STEIN, C. (1974). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymptotic Statistics* **2**, 345–381.

TSUKUMA, H. (2008). Admissibility and minimaxity of generalized Bayes estimators for a normal mean matrix. *J. Multivariate. Anal.* **99**, 2251–2264.