Inadmissibility of the corrected Akaike information criterion

TAKERU MATSUDA[1,2,a]

[1]*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan,* [a]*matsuda@mist.i.u-tokyo.ac.jp*
[2]*Statistical Mathematics Unit, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan*

For the multivariate linear regression model with unknown covariance, the corrected Akaike information criterion is the minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy. In this study, based on the loss estimation framework, we show its inadmissibility as an estimator of the Kullback–Leibler discrepancy itself, instead of the expected Kullback–Leibler discrepancy. We provide improved estimators of the Kullback–Leibler discrepancy that work well in reduced-rank situations and examine their performance numerically.

Keywords: Admissibility; Akaike information criterion; corrected Akaike information criterion; Kullback–Leibler discrepancy; loss estimation

1. Introduction

We consider the multivariate linear regression model with p explanatory variables and q response variables:

$$y_i = B^\top x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}_q(0, \Sigma), \tag{1}$$

for $i = 1, \ldots, n$, where $n \geq p$, $B \in \mathbb{R}^{p \times q}$ is an unknown regression coefficient matrix, $\Sigma \in \mathbb{R}^{q \times q}$ is an unknown covariance matrix (positive definite) and $\varepsilon_1, \ldots, \varepsilon_n$ are independent. In the following, the probability density function of $Y = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^{n \times q}$ is denoted by $p(Y \mid B, \Sigma)$ and the expectation of $f(Y)$ under $p(Y \mid B, \Sigma)$ is written as $\mathrm{E}_{B,\Sigma}[f(Y)]$.

The maximum likelihood estimate for the model (1) is given by

$$\hat{B} = (X^\top X)^{-1} X^\top Y, \quad \hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^\top (Y - X\hat{B}).$$

The Akaike Information Criterion (AIC; Akaike, 1973) is an approximately unbiased estimator of the expected Kullback–Leibler discrepancy:

$$\mathrm{E}_{B,\Sigma}[\mathrm{AIC}] = \mathrm{E}_{B,\Sigma}[d((B,\Sigma),(\hat{B},\hat{\Sigma}))] + o(1)$$

as $n \to \infty$, where

$$d((B,\Sigma),(\hat{B},\hat{\Sigma})) = -2 \int p(\widetilde{Y} \mid B, \Sigma) \log p(\widetilde{Y} \mid \hat{B}, \hat{\Sigma}) \mathrm{d}\widetilde{Y} \tag{2}$$

$$= nq \log(2\pi) + n \log \det \hat{\Sigma} + n\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + \mathrm{tr}(\hat{\Sigma}^{-1}(\hat{B} - B)^\top X^\top X(\hat{B} - B))$$

is called the Kullback–Leibler discrepancy from $p(\widetilde{Y} \mid B, \Sigma)$ to $p(\widetilde{Y} \mid \hat{B}, \hat{\Sigma})$. The AIC is widely used for evaluation and selection of linear regression models (Burnham and Anderson, 2002, Konishi and Kitagawa, 2008).

The bias of the AIC is non-negligible when the sample size n is not sufficiently large. Thus, a corrected AIC (AICc) has been derived (Bedrick and Tsai, 1994, Hurvich and Tsai, 1989, Sugiura, 1978), which is exactly unbiased:

$$\mathrm{E}_{B,\Sigma}[\mathrm{AICc}] = \mathrm{E}_{B,\Sigma}[d((B,\Sigma),(\hat{B},\hat{\Sigma}))].$$

Cavanaugh (1997) provided a unified derivation of AIC and AICc and Davies, Neath and Cavanaugh (2006) showed that AICc is the minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy.

Both AIC and AICc were developed to unbiasedly estimate the expected Kullback–Leibler discrepancy. This idea dates back to Stein's unbiased risk estimate (SURE; Stein, 1974), which unbiasedly estimates the quadratic risk of estimators of a normal mean (Fourdrinier, Strawderman and Wells, 2018, Lehmann and Casella, 2006). In this context, Johnstone (1988) considered estimation of the quadratic loss itself, instead of its average (quadratic risk). Although SURE is still unbiased for this problem, Johnstone (1988) showed that it can be improved in terms of the mean squared error. In other words, SURE is inadmissible as an estimator of the quadratic loss. See Section 2 for details. This finding led to the development of a field called loss estimation (Fourdrinier and Wells, 2012).

In this study, we examine AIC and AICc from the loss estimation viewpoint and investigate their admissibility as estimators of the Kullback–Leibler discrepancy $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$, instead of its average $\mathrm{E}_{B,\Sigma}[d((B,\Sigma),(\hat{B},\hat{\Sigma}))]$ (expected Kullback–Leibler discrepancy). The former estimand is random (depends on Y as well as (B,Σ)) whereas the latter one is non-random (depends only on (B,Σ)). In this sense, it may be more correct to refer to the current problem as prediction rather than estimation[1]. The current setting to estimate (or predict) $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ is considered to reflect the practical usage of AIC (and AICc) as a model evaluation criterion more faithfully as follows. Given data at hand, we estimate (B,Σ) by $(\hat{B},\hat{\Sigma})$ and construct the plug-in predictive distribution $p(\widetilde{Y} \mid \hat{B},\hat{\Sigma})$ for a future observation. The disparity between this predictive distribution and the true data-generating distribution $p(\widetilde{Y} \mid B,\Sigma)$ is given by the Kullback–Leibler discrepancy $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$. Whereas the usual argument on AIC (and AICc) considers estimation of the average of $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ over the possible realizations of Y (expected Kullback–Leibler discrepancy), here we focus on estimation (or prediction) of $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ itself for the specific realization of Y at hand. Thus, the current setting provides direct (conditional) assessment of the performance of the predictive distribution obtained from the data at hand. Note that Matsuda and Strawderman (2016) studied the Pitman closeness property of predictive distributions in a similar spirit. We develop improved estimators of $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ and show that they attain better variable selection result than AIC and AICc in simulation. It demonstrates a practical advantage of introducing the current setting. See Fourdrinier and Wells (2012) for further discussion on motivation for considering loss estimation.

This paper is organized as follows. In Section 2, we briefly review the loss estimation framework and existing results for normal mean vector and matrix. We also derive an improved loss estimator for a normal mean matrix, which will be the basis of the main results of this paper. In Section 3, we introduce the general setting of loss estimation for a predictive distribution and study the properties of AIC and AICc as loss estimators in multivariate linear regression. For the multivariate linear regression model (1) with known covariance, AIC is shown to be inadmissible and an improved loss estimator is given. For the multivariate linear regression model (1) with unknown covariance, AIC is shown to be inadmissible and dominated by AICc. Then, in Section 4, we prove that AICc is still inadmissible and provide improved loss estimators that work well in reduced-rank situations. In Section 5, we present

---

[1]Similarly, Lehmann and Casella (2006) states that it is common to speak of prediction, rather than estimation, of random effects (Example 3.5.5). See also Sandved (1968)

numerical results to examine the performance of the improved estimators. The results demonstrate that the improved estimators often outperform the corrected AIC in variable selection. Finally, we provide concluding remarks in Section 6. Technical lemmas are given with proofs in the Appendix.

2. Loss estimation framework

2.1. General setting

Here, we briefly introduce the loss estimation framework. See Fourdrinier, Strawderman and Wells (2018), Fourdrinier and Wells (2012) for a comprehensive review of loss estimation. Recently, the idea of loss estimation has been applied to high-dimensional inference (Bellec and Zhang, 2021).

Suppose that we have an observation $Y \sim p(y \mid \theta)$, where θ is an unknown parameter. In usual setting of point estimation (Lehmann and Casella, 2006), we consider estimation of θ using an estimator $\hat{\theta} = \hat{\theta}(y)$. The discrepancy of an estimate $\hat{\theta}$ from the true value θ is quantified by a loss function $L(\theta, \hat{\theta})$. Then, estimators are compared by using the risk function $R(\theta, \hat{\theta}) = \mathrm{E}_\theta[L(\theta, \hat{\theta}(y))]$, which is the average of the loss. An estimator $\hat{\theta}_1$ is said to dominate another estimator $\hat{\theta}_2$ if $R(\theta, \hat{\theta}_1) \leq R(\theta, \hat{\theta}_2)$ holds for every θ, with strict inequality for at least one value of θ. An estimator $\hat{\theta}$ is said to be admissible if no estimator dominates $\hat{\theta}$. An estimator $\hat{\theta}$ is said to be inadmissible if it is not admissible (i.e. there exists an estimator that dominates $\hat{\theta}$).

In the setting above, loss estimation concerns estimation of the loss $L(\theta, \hat{\theta}(y))$, which depends not only on θ but also on y. The performance of a loss estimator $\lambda(y)$ is evaluated by squared error $(\lambda(y) - L(\theta, \hat{\theta}(y)))^2$. Thus, a loss estimator $\lambda_1(y)$ is said to dominate another loss estimator $\lambda_2(y)$ if

$$\mathrm{E}_\theta[(\lambda_1(y) - L(\theta, \hat{\theta}(y)))^2] \leq \mathrm{E}_\theta[(\lambda_2(y) - L(\theta, \hat{\theta}(y)))^2]$$

holds for every θ, with strict inequality for at least one value of θ. A loss estimator $\lambda(y)$ is said to be admissible if no loss estimator dominates $\lambda(y)$. A loss estimator $\lambda(y)$ is said to be inadmissible if it is not admissible (i.e. there exists a loss estimator that dominates $\lambda(y)$).

2.2. Loss estimation for a normal mean vector

Now, we focus on loss estimation for a normal mean vector (Johnstone, 1988). Suppose that we estimate $\theta \in \mathbb{R}^p$ from an observation $Y \sim \mathrm{N}_p(\theta, I_p)$ by an estimator $\hat{\theta}(y) = y + g(y)$ under the quadratic loss

$$L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2.$$

Stein (1974) showed that the quadratic risk $R(\theta, \hat{\theta}) = \mathrm{E}_\theta[L(\theta, \hat{\theta}(y))]$ satisfies

$$R(\theta, \hat{\theta}) = \mathrm{E}_\theta[\lambda^{\mathrm{U}}(y)],$$

where

$$\lambda^{\mathrm{U}}(y) = p + 2\nabla \cdot g(y) + \|g(y)\|^2$$

is called Stein's unbiased risk estimate (SURE). SURE plays a central role in the theory of shrinkage estimation (Fourdrinier, Strawderman and Wells, 2018) and also closely related to model selection criteria such as Mallows' C_p and AIC (Boisbunon et al., 2014). However, Johnstone (1988) showed that SURE is inadmissible for the maximum likelihood estimator when $p \geq 5$ as follows.
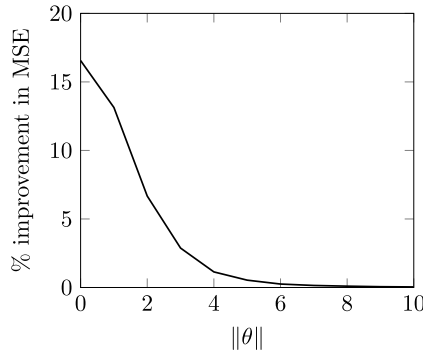
Figure 1. Percentage improvements in mean squared error of the loss estimator (3) over SURE for $p = 8$.

Proposition 2.1 (Johnstone, 1988). *In estimation of θ from $Y \sim \mathrm{N}_p(\theta, I_p)$ under the quadratic loss, consider the maximum likelihood estimator $\hat{\theta}(y) = y$. If $p \geq 5$, then SURE $\lambda^{\mathrm{U}}(y) = p$ is inadmissible and dominated by the loss estimator*

$$\lambda(y) = p - 2(p-4)\|y\|^{-2}. \tag{3}$$

Figure 1 plots the percentage improvements in mean sqaured error of the loss estimator (3) over SURE defined by

$$100 \frac{\mathrm{E}_\theta[(\lambda^{\mathrm{U}}(y) - \|\hat{\theta}(y) - \theta\|^2)^2] - \mathrm{E}_\theta[(\lambda(y) - \|\hat{\theta}(y) - \theta\|^2)^2]}{\mathrm{E}_\theta[(\lambda^{\mathrm{U}}(y) - \|\hat{\theta}(y) - \theta\|^2)^2]}.$$

The improvement is large when the true value of θ is close to the origin, which is qualitatively similar to the risk behavior of the James–Stein estimator. In addition to the maximum likelihood estimator, Johnstone (1988) also proved the inadmissibility of SURE for the James–Stein estimator and provided improved loss estimators. Based on these findings by Johnstone (1988), many studies have investigated loss estimation for a normal mean vector and single-response linear regression, such as (Boisbunon et al., 2014, Fourdrinier and Strawderman, 2003, Fourdrinier and Wells, 2012, Lu and Berger, 1989, Narayanan and Wells, 2015, Wan and Zou, 2004).

2.3. Loss estimation for a normal mean matrix

Recently, Matsuda and Strawderman (2019) generalized the results of Johnstone (1988) to matrices and developed loss estimators that dominate SURE. This is motivated from the Efron–Morris estimator, which is a matrix generalization of the James–Stein estimator that shrinks singular values towards zero (Efron and Morris, 1972). Specifically, suppose that we estimate $M \in \mathbb{R}^{p \times q}$ from an observation $Y \sim \mathrm{N}_{p,q}(M, I_p, I_q)$ by an estimator $\hat{M}(Y)$ under the Frobenius loss

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

We write the singular values of a matrix $Z \in \mathbb{R}^{p \times q}$ with $p \geq q$ by $\sigma_1(Z) \geq \cdots \geq \sigma_q(Z) \geq 0$.

Proposition 2.2 ([Matsuda and Strawderman, 2019](#)). *In estimation of M from $Y \sim \mathrm{N}_{p,q}(M, I_p, I_q)$ under the Frobenius loss, consider the maximum likelihood estimator $\hat{M}(Y) = Y$. If $p \geq 3q + 2$ and $q \geq 2$, then SURE $\lambda^{\mathrm{U}}(Y) = pq$ is inadmissible and dominated by the loss estimator*

$$\lambda(Y) = pq - \sum_{i=1}^{q} c_i \sigma_i(Y)^{-2}, \quad c_i = \frac{4(p - q - 2i - 1)}{q}. \tag{4}$$

Whereas Proposition [2.2](#) shows the inadmissibility of SURE, it excludes[2] the case $q = 1$. Here, we provide another loss estimator dominating SURE, which reduces to [(3)](#) in Proposition [2.1](#) when $q = 1$ and will be the basis of the main results of this paper.

Theorem 2.3. *In estimation of M from $Y \sim \mathrm{N}_{p,q}(M, I_p, I_q)$ under the Frobenius loss, consider the maximum likelihood estimator $\hat{M}(Y) = Y$. If $p \geq 2q + 3$, then SURE $\lambda^{\mathrm{U}}(Y) = pq$ is inadmissible and dominated by the loss estimator*

$$\lambda(Y) = pq - \frac{2(p - 2q - 2)}{q} \mathrm{tr}((Y^\top Y)^{-1}). \tag{5}$$

Proof. Let $h(Y) = -c \, \mathrm{tr}((Y^\top Y)^{-1})$ with $c = 2(p - 2q - 2)/q$ so that $\lambda(Y) = \lambda^{\mathrm{U}}(Y) + h(Y)$. Then, from Lemma 5 of [Matsuda and Strawderman](#) ([2019](#)),

$$\mathrm{E}_M[(\lambda(Y) - \|\hat{M}(Y) - M\|^2)^2] - \mathrm{E}_M[(\lambda^{\mathrm{U}}(Y) - \|\hat{M}(Y) - M\|^2)^2] = \mathrm{E}_M[-2\Delta h(Y) + h(Y)^2],$$

where

$$\Delta h(Y) = \sum_{i,j} \frac{\partial^2 h}{\partial Y_{ij}^2}(Y).$$

From Lemma [A.2](#) and Lemma [A.3](#),

$$\Delta h(Y) = -c \sum_{i,j} \frac{\partial^2}{\partial Y_{ij}^2} \mathrm{tr}((Y^\top Y)^{-1})$$

$$= 2c \sum_{i,j} \frac{\partial}{\partial Y_{ij}} (Y(Y^\top Y)^{-2})_{ij}$$

$$= 2c(p - q - 2) \mathrm{tr}((Y^\top Y)^{-2}) - 2c(\mathrm{tr}((Y^\top Y)^{-1}))^2.$$

Thus,

$$-2\Delta h(Y) + h(Y)^2 = -4c(p - q - 2) \mathrm{tr}((Y^\top Y)^{-2}) + (c^2 + 4c)(\mathrm{tr}((Y^\top Y)^{-1}))^2.$$

From the Cauchy–Schwarz inequality,

$$(\mathrm{tr}((Y^\top Y)^{-1}))^2 = \left(\sum_{i=1}^{q} \lambda_i((Y^\top Y)^{-1}) \right)^2 \leq q \sum_{i=1}^{q} \lambda_i((Y^\top Y)^{-1})^2 = q \, \mathrm{tr}((Y^\top Y)^{-2}), \tag{6}$$

---

[2]The condition $q \geq 2$ in Proposition [2.2](#) was not explicitly stated in the original paper ([Matsuda and Strawderman](#), [2019](#)).
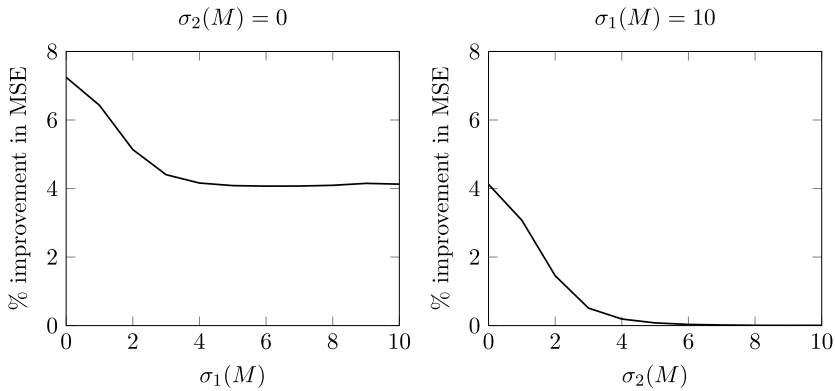
Figure 2. Percentage improvements in mean squared error of $\lambda(Y)$ in (5) over SURE for $p = 8$ and $q = 2$. Left: $\sigma_2(M) = 0$. Right: $\sigma_1(M) = 10$.

where $\lambda_i(A)$ denotes the i-th eigenvalue of a matrix A. Therefore, by substituting $c = 2(p - 2q - 2)/q$,

$$-2\Delta h(Y) + h(Y)^2 \le c(-4(p - q - 2) + q(c + 4))\text{tr}((Y^\top Y)^{-2})$$

$$= -\frac{4(p - 2q - 2)^2}{q}\text{tr}((Y^\top Y)^{-2})$$

$$< 0.$$

Hence,

$$E_M[(\lambda(Y) - \|\hat{M}(Y) - M\|^2)^2] < E_M[(\lambda^U(Y) - \|\hat{M}(Y) - M\|^2)^2]$$

for every M. $\qquad\qquad\square$

Figure 2 plots the percentage improvements in mean squared error of the loss estimator (5) over SURE like Figure 1. The improvement is large when some of the singular values of M are small. In particular, the left panel of Figure 2 indicates that the loss estimator (5) attains constant reduction of MSE as long as $\sigma_2(M) = 0$, even when $\sigma_1(M)$ is large. Thus, the loss estimator (5) works well when M is close to low-rank. Note that the loss estimator (4) has qualitatively the same property (Matsuda and Strawderman, 2019). These results are understood from the fact that both loss estimators (4) and (5) are based on the inverse square of the singular values of Y. The Efron–Morris estimator for a normal mean matrix has a similar risk property (Matsuda and Strawderman, 2022).

In addition to the maximum likelihood estimator, Matsuda and Strawderman (2019) also proved the inadmissibility of SURE for a general class of orthogonally invariant estimators, including the Efron–Morris estimator and reduced-rank estimators, and provided improved loss estimators.

3. Information criterion as loss estimator

3.1. Loss estimation for a predictive distribution

Suppose that we have an observation $Y \sim p(y \mid \theta)$, where θ is an unknown parameter. Then, we consider prediction of a future observation $\widetilde{Y} \sim p(\widetilde{y} \mid \theta)$ by using a predictive distribution $\hat{p}(\widetilde{y} \mid y)$. The

discrepancy of a predictive distribution $\hat{p}(\widetilde{y} \mid y)$ from the true distribution $p(\widetilde{y} \mid \theta)$ is evaluated by the Kullback–Leibler discrepancy

$$d(p(\widetilde{y} \mid \theta), \hat{p}(\widetilde{y} \mid y)) = -2 \int p(\widetilde{y} \mid \theta) \log \hat{p}(\widetilde{y} \mid y) \mathrm{d}\widetilde{y},$$

which is equivalent to twice the Kullback–Leibler divergence

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

up to an additive constant. The plug-in predictive distribution is defined by $p(\widetilde{y} \mid \hat{\theta}(y))$, where $\hat{\theta}(y)$ is the maximum likelihood estimate of θ from y. AIC (Akaike, 1973) is an approximately unbiased estimator of the Kullback–Leibler discrepancy for the plug-in predictive distribution:

$$\mathrm{E}_{\theta}[\mathrm{AIC}] \approx d(p(\widetilde{y} \mid \theta), p(\widetilde{y} \mid \hat{\theta}(y))).$$

See Burnham and Anderson (2002), Konishi and Kitagawa (2008) for details.

Similarly to point estimation in Section 2, we can formulate estimation of the Kullback–Leibler discrepancy as a loss estimation problem. Then, AIC can be viewed as a default loss estimator like SURE in estimation of a normal mean. From this viewpoint, it is of interest to determine whether AIC is admissible or not. In the following, we investigate this problem for the multivariate linear regression model (1).

3.2. Multivariate linear regression with known covariance

First, consider the multivariate linear regression model (1) with known covariance $\Sigma > O$. The maximum likelihood estimate is $\hat{B} = (X^{\top}X)^{-1}X^{\top}Y$. From (2), the Kullback–Leibler discrepancy for the plug-in predictive distribution is

$$d((B, \Sigma), (\hat{B}, \Sigma)) = nq \log(2\pi e) + n \log \det \Sigma + \mathrm{tr}(\Sigma^{-1}(\hat{B} - B)^{\top}X^{\top}X(\hat{B} - B)).$$

The AIC is

$$\begin{aligned}
\mathrm{AIC} &= -2 \log p(Y \mid \hat{B}, \Sigma) + 2pq \\
&= nq \log(2\pi) + n \log \det \Sigma + \mathrm{tr}(\Sigma^{-1}(Y - X\hat{B})^{\top}(Y - X\hat{B})) + 2pq.
\end{aligned}$$

Then, the inadmissibility of AIC is proved as follows, where MAIC is an abbreviation of "Modified AIC."

Theorem 3.1. *Consider the multivariate linear regression model* (1) *with known* $\Sigma > O$. *If* $p \geq 2q + 3$, *then AIC is inadmissible and dominated by*

$$\mathrm{MAIC} = \mathrm{AIC} - \frac{2(p - 2q - 2)}{q} \mathrm{tr}(\Sigma((X\hat{B})^{\top}(X\hat{B}))^{-1})$$

as an estimator of the Kullback–Leibler discrepancy.

Proof. Let $R = (Y - X\hat{B})^\top (Y - X\hat{B})$ be the residual. Then, from the standard theory of multivariate linear regression (Anderson, 2003), $\hat{B}$ and R are independent and distributed as $\hat{B} \sim \mathrm{N}_{p,q}(B, (X^\top X)^{-1}, \Sigma)$ and $R \sim W_q(n-p, \Sigma)$, respectively. Thus, $Z = (X^\top X)^{1/2}\hat{B}\Sigma^{-1/2}$ is independent from R and distributed as $Z \sim \mathrm{N}_{p,q}(\bar{Z}, I_p, I_q)$ where $\bar{Z} = (X^\top X)^{1/2}B\Sigma^{-1/2}$.

Let $h = -c\mathrm{tr}(\Sigma((X\hat{B})^\top(X\hat{B}))^{-1}) = -c\mathrm{tr}((Z^\top Z)^{-1})$ with $c = 2(p - 2q - 2)/q$ so that MAIC = AIC + $h(Y)$. Then,

$$\mathrm{E}_B[(\mathrm{MAIC} - d)^2] - \mathrm{E}_B[(\mathrm{AIC} - d)^2] = \mathrm{E}_B[h^2 + 2h(\mathrm{AIC} - d)],$$

where we write $d((B, \Sigma), (\hat{B}, \Sigma))$ by d for simplicity. Note that

$$\mathrm{AIC} - d = (2p - n)q + \mathrm{tr}(\Sigma^{-1}(Y - X\hat{B})^\top(Y - X\hat{B})) - \mathrm{tr}(\Sigma^{-1}(\hat{B} - B)^\top X^\top X(\hat{B} - B))$$

$$= (2p - n)q + \mathrm{tr}(\Sigma^{-1}R) - \mathrm{tr}((Z - \bar{Z})^\top(Z - \bar{Z})).$$

From $\mathrm{E}[R] = (n - p)\Sigma$ and the independence of Z and R,

$$\mathrm{E}_B[\mathrm{tr}((Z^\top Z)^{-1})\mathrm{tr}(\Sigma^{-1}R)] = (n - p)q \cdot \mathrm{E}_B[\mathrm{tr}((Z^\top Z)^{-1})].$$

Also, from Lemma B.8,

$$\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})\mathrm{tr}((Z - \bar{Z})^\top(Z - \bar{Z}))]$$

$$= pq\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] - 2(p - q - 2)\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})] + 2\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2].$$

Therefore,

$$\mathrm{E}_B[(\mathrm{MAIC} - d)^2] - \mathrm{E}_B[(\mathrm{AIC} - d)^2] = \mathrm{E}_B[(c^2 + 4c)(\mathrm{tr}((Z^\top Z)^{-1}))^2 - 4c(p - q - 2)\mathrm{tr}((Z^\top Z)^{-2})]$$

$$\leq qc\left(c - \frac{4(p - 2q - 2)}{q}\right)\mathrm{E}_B[\mathrm{tr}((Z^\top Z)^{-2})]$$

$$< 0$$

for every B, where we used $(\mathrm{tr}((Z^\top Z)^{-1}))^2 \leq q\mathrm{tr}((Z^\top Z)^{-2})$ from (6) and $c = 2(p - 2q - 2)/q$. $\qquad\square$

For the Gaussian linear regression model with known variance ($q = 1$), Boisbunon et al. (2014) discussed the equivalence between AIC and SURE. Such a correspondence holds in the current setting as well. Specifically, consider estimation of M from $Y \sim \mathrm{N}_{n,q}(M, I_n, \Sigma)$ under the loss $L(M, \hat{M}) = \mathrm{tr}(\Sigma^{-1}(\hat{M} - M)^\top(\hat{M} - M))$. Then, SURE for the estimator $\hat{M} = X\hat{B}$ is

$$\lambda^{\mathrm{U}}(Y) = \mathrm{tr}(\Sigma^{-1}(Y - X\hat{B})^\top(Y - X\hat{B})) + (2p - n)q$$

$$= \mathrm{AIC} - nq\log(2\pi e) - n\log\det\Sigma,$$

and the loss is related to the Kullback–Leibler discrepancy as

$$L(M, \hat{M}) = d((B, \Sigma), (\hat{B}, \Sigma)) - nq\log(2\pi e) - n\log\det\Sigma.$$

Thus, estimation of the loss $L(M, \hat{M})$ for the estimator $\hat{M} = X\hat{B}$ is equivalent to estimation of the Kullback–Leibler discrepancy for the plug-in predictive distribution, and both SURE and AIC are exactly unbiased. Under this correspondence, Proposition 3.1 is rewritten as follows.

Corollary 3.2. *For the multivariate linear regression model* (1) *with known* $\Sigma > O$, *consider the estimator* $\hat{M} = X\hat{B}$ *of* $M = XB$ *under the loss* $L(M,\hat{M}) = \text{tr}(\Sigma^{-1}(\hat{M}-M)^\top(\hat{M}-M))$. *If* $p \geq 2q+3$, *then SURE is inadmissible and dominated by the loss estimator*

$$\lambda(Y) = \lambda^{\mathrm{U}}(Y) - \frac{2(p-2q-2)}{q}\text{tr}(\Sigma((X\hat{B})^\top(X\hat{B}))^{-1}).$$

3.3. Multivariate linear regression with unknown covariance

Next, consider the multivariate linear regression model (1) with unknown covariance $\Sigma > O$. The maximum likelihood estimate is $\hat{B} = (X^\top X)^{-1}X^\top Y$ and $\hat{\Sigma} = (Y - X\hat{B})^\top(Y - X\hat{B})/n$. From (2), the Kullback–Leibler discrepancy is

$$d((B,\Sigma),(\hat{B},\hat{\Sigma})) = nq\log(2\pi) + n\log\det\hat{\Sigma} + n\text{tr}(\hat{\Sigma}^{-1}\Sigma) + \text{tr}(\hat{\Sigma}^{-1}(\hat{B}-B)^\top X^\top X(\hat{B}-B)).$$

The AIC is

$$\text{AIC} = nq\log(2\pi) + n\log\det\hat{\Sigma} + 2\left(pq + \frac{q(q+1)}{2}\right).$$

The corrected AIC is

$$\text{AICc} = nq\log(2\pi) + n\log\det\hat{\Sigma} + \frac{2n}{n-p-q-1}\left(pq + \frac{q(q+1)}{2}\right).$$

The corrected AIC is exactly unbiased while AIC is biased (Bedrick and Tsai, 1994, Hurvich and Tsai, 1989, Sugiura, 1978). Then, we obtain the following.

Theorem 3.3. *For the multivariate linear regression model* (1) *with unknown* $\Sigma > O$, *AIC is inadmissible and dominated by AICc as an estimator of the Kullback–Leibler discrepancy.*

Proof. For two random variables S and T, we have

$$\begin{aligned}
&\text{E}[(S-T)^2]\\
=&\text{E}[(S - \text{E}[S] + \text{E}[S] - \text{E}[T] + \text{E}[T] - T)^2]\\
=&\text{E}[(S - \text{E}[S])^2] + (\text{E}[S] - \text{E}[T])^2 + \text{E}[(T - \text{E}[T])^2] - 2\text{E}[(S - \text{E}[S])(T - \text{E}[T])]\\
=&\text{Var}[S] + (\text{E}[S] - \text{E}[T])^2 + \text{Var}[T] - 2\text{Cov}[S,T].
\end{aligned}$$

Hence, the mean squared error of AIC is given by

$$\text{E}[(\text{AIC} - d)^2] = \text{Var}[\text{AIC}] + (\text{E}[\text{AIC}] - \text{E}[d])^2 + \text{Var}[d] - 2\text{Cov}[\text{AIC},d], \tag{7}$$

where we write $d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ by d for simplicity. Similarly, the mean squared error of AICc is

$$\text{E}[(\text{AICc} - d)^2] = \text{Var}[\text{AICc}] + \text{Var}[d] - 2\text{Cov}[\text{AICc},d], \tag{8}$$

where we used $\text{E}[\text{AICc}] = \text{E}[d]$.

On the other hand, since the difference between AIC and AICc is constant,

$$\text{Var[AIC]} = \text{Var[AICc]}, \quad \text{Cov[AIC}, d] = \text{Cov[AICc}, d]. \tag{9}$$

From (7), (8) and (9),

$$\text{E}_{B,\Sigma}[(\text{AICc} - d)^2] \le \text{E}_{B,\Sigma}[(\text{AIC} - d)^2]$$

for every B and Σ. $\qquad\square$

In the next section, we show that the corrected AIC is still inadmissible and provide improved loss estimators. Note that Davies, Neath and Cavanaugh (2006) showed that the corrected AIC is the minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy.

4. Inadmissibility of the corrected AIC

For the multivariate linear regression model (1) with unknown covariance, the corrected AIC is the minimum variance unbiased estimator of the expected Kullback–Leibler discrepancy from the Lehmann–Scheffé theorem (Davies, Neath and Cavanaugh, 2006). Also, Theorem 3.3 showed that the AIC is dominated by the corrected AIC as an estimator of the Kullback–Leibler discrepancy. However, the corrected AIC is still inadmissible as follows, where MAICc is an abbreviation of "Modified AICc."

Theorem 4.1. *Consider the multivariate linear regression model* (1) *with unknown* $\Sigma > O$. *Let*

$$\bar{c} = \frac{4n^2}{(n-p)(q(n-p)+2)}\left(p - 2q - 2 - \frac{q^2 + q - 2}{n - p - q - 1}\right).$$

If $n - p - q - 1 > 0$ *and* $\bar{c} > 0$, *then for any* $c \in (0, \bar{c}]$, *AICc is inadmissible and dominated by*

$$\text{MAICc} = \text{AICc} - c\text{tr}(\hat{\Sigma}((X\hat{B})^\top(X\hat{B}))^{-1})$$

as an estimator of the Kullback–Leibler discrepancy.

Proof. From the standard theory of multivariate linear regression (Anderson, 2003), the maximum likelihood estimates $\hat{B} = (X^\top X)^{-1}X^\top Y$ and $\hat{\Sigma} = (Y - X\hat{B})^\top(Y - X\hat{B})/n$ for (1) are independently distributed as

$$\hat{B} \sim \text{N}_{p,q}(B, (X^\top X)^{-1}, \Sigma), \quad \hat{\Sigma} \sim W_q\left(n - p, \frac{1}{n}\Sigma\right).$$

Thus, $Z = (X^\top X)^{1/2}\hat{B}\Sigma^{-1/2}$ and $S = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$ are independetly distributed as

$$Z \sim \text{N}_{p,q}(\bar{Z}, I_p, I_q), \quad S \sim W_q\left(n - p, \frac{1}{n}I_q\right),$$

where $\bar{Z} = (X^\top X)^{1/2}B\Sigma^{-1/2}$.

Again, we write $d((B, \Sigma), (\hat{B}, \hat{\Sigma}))$ in (2) as d for simplicity. Let $d = d((B, \Sigma), (\hat{B}, \hat{\Sigma}))$ and

$$h = \text{MAICc} - \text{AICc} = -c\text{tr}(\hat{\Sigma}((X\hat{B})^\top(X\hat{B}))^{-1}).$$

Then,

$$\mathrm{E}_{B,\Sigma}[(\mathrm{MAICc} - d)^2] - \mathrm{E}_{B,\Sigma}[(\mathrm{AICc} - d)^2] = \mathrm{E}_{B,\Sigma}[h^2 + 2h(\mathrm{AICc} - d)]. \tag{10}$$

We evaluate each term. Note that $h = -c\mathrm{tr}(S(Z^\top Z)^{-1})$, since

$$\begin{aligned}
\mathrm{tr}(\hat{\Sigma}((X\hat{B})^\top(X\hat{B}))^{-1}) &= \mathrm{tr}(S\Sigma^{1/2}((X\hat{B})^\top(X\hat{B}))^{-1}\Sigma^{1/2}) \\
&= \mathrm{tr}(S(\Sigma^{-1/2}\hat{B}^\top X^\top X\hat{B}\Sigma^{-1/2})^{-1}) \\
&= \mathrm{tr}(S(Z^\top Z)^{-1}).
\end{aligned}$$

In the following, we write $\mathrm{E}_{B,\Sigma}$ as E for simplicity.

First, by using Lemma B.2,

$$\begin{aligned}
\mathrm{E}[h^2] &= c^2\mathrm{E}[\mathrm{tr}(S(Z^\top Z)^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})] \\
&= \frac{2(n-p)}{n^2}c^2\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})] + \frac{(n-p)^2}{n^2}c^2\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{11}
\end{aligned}$$

Next, from

$$\mathrm{AICc} = nq\log(2\pi) + n\log\det\hat{\Sigma} + \frac{nq(n+p)}{n-p-q-1},$$

$$\begin{aligned}
d &= nq\log(2\pi) + n\log\det\hat{\Sigma} + n\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + \mathrm{tr}(X(\hat{B} - B)\hat{\Sigma}^{-1}(\hat{B} - B)^\top X^\top) \\
&= nq\log(2\pi) + n\log\det\hat{\Sigma} + n\mathrm{tr}(S^{-1}) + \mathrm{tr}((Z - \bar{Z})^\top(Z - \bar{Z})S^{-1}),
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{E}[2h(\mathrm{AICc} - d)] = {}& -\frac{2n(n+p)q}{n-p-q-1}c\mathrm{E}[\mathrm{tr}(S(Z^\top Z)^{-1})] + 2nc\mathrm{E}[\mathrm{tr}(S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})] \\
&+ 2c\mathrm{E}[\mathrm{tr}((Z - \bar{Z})^\top(Z - \bar{Z})S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]. \tag{12}
\end{aligned}$$

Using the independence of S and Z,

$$\begin{aligned}
\mathrm{E}[\mathrm{tr}(S(Z^\top Z)^{-1})] &= \mathrm{tr}(\mathrm{E}[S] \cdot \mathrm{E}[(Z^\top Z)^{-1})]) \\
&= \mathrm{tr}\left(\frac{n-p}{n}I_q \cdot \mathrm{E}[(Z^\top Z)^{-1})]\right) \\
&= \frac{n-p}{n}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]. \tag{13}
\end{aligned}$$

Similarly, using the independence of S and Z and Lemma B.6,

$$\begin{aligned}
\mathrm{E}[\mathrm{tr}(S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})] &= \mathrm{tr}(\mathrm{E}[\mathrm{tr}(S^{-1})S] \cdot \mathrm{E}[(Z^\top Z)^{-1}]) \\
&= \mathrm{tr}\left(\frac{(n-p)q-2}{n-p-q-1}I_q \cdot \mathrm{E}[(Z^\top Z)^{-1}]\right) \\
&= \frac{(n-p)q-2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]. \tag{14}
\end{aligned}$$

Also, from Lemma B.9,

$$\mathrm{E}[\mathrm{tr}((Z - \bar{Z})^\top (Z - \bar{Z})S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]$$

$$= p\frac{(n-p)q-2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] - 2\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]$$

$$+ 2\frac{n-q-2}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{15}$$

Therefore, by substituting (13), (14) and (15) into (12),

$$\mathrm{E}[2h(\mathrm{AICc} - d)]$$

$$= \frac{-2(n-p)(n+p)q + 2n((n-p)q-2) + 2p((n-p)q-2)}{n-p-q-1}c\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]$$

$$- 4c\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})] + 4c\frac{n-q-2}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]$$

$$= -4c\frac{n+p}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]$$

$$- 4c\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})] + 4c\frac{n-q-2}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{16}$$

Hence, by substituting (11) and (16) into (10),

$$\mathrm{E}_{B,\Sigma}[(\mathrm{MAICc} - d)^2] - \mathrm{E}_{B,\Sigma}[(\mathrm{AICc} - d)^2]$$

$$= -4c\frac{n+p}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]$$

$$+ \left(\frac{2(n-p)}{n^2}c^2 - 4c\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\right)\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]$$

$$+ \left(\frac{(n-p)^2}{n^2}c^2 + 4c\frac{n-q-2}{n-p-q-1}\right)\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]$$

$$\leq -4c\frac{n+p}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]$$

$$+ c\left(\frac{(n-p)(q(n-p)+2)}{n^2}c - 4\frac{(n-p-1)(p-q-2) - q(n-q-2)+2}{n-p-q-1}\right)\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]$$

$$= -4c\frac{n+p}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] + \frac{(n-p)(q(n-p)+2)}{n^2}c(c - \bar{c})\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})],$$

where we used $(\mathrm{tr}((Z^\top Z)^{-1}))^2 \leq q\mathrm{tr}((Z^\top Z)^{-2})$ from (6) and $(n-p-1)(p-q-2) - q(n-q-2) = (n-p-q-1)(p-2q-2) - q^2 - q$. Therefore, if $0 < c \leq \bar{c}$, then

$$\mathrm{E}_{B,\Sigma}[(\mathrm{MAICc} - d)^2] \leq \mathrm{E}_{B,\Sigma}[(\mathrm{AICc} - d)^2]$$

for every B and Σ. □

Note that, when n is sufficiently large, the condition $\bar{c} > 0$ in Theorem 4.1 is reduced to $p \geq 2q + 3$, which is the same as in Theorem 2.3 and Theorem 3.1.

In the case of a single response ($q = 1$), Theorem 4.1 is reexpressed as follows. Here, we employ the usual notation for the linear regression with a single response:

$$y \sim N_n(X\beta, \sigma^2 I_n), \quad \hat{\beta} = (X^\top X)^{-1} X^\top y, \quad \hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/n. \tag{17}$$

Corollary 4.2. *Consider the linear regression model $y \sim N_n(X\beta, \sigma^2 I_n)$ with unknown σ^2. Let*

$$\bar{c} = \frac{4n^2(p-4)}{(n-p)(n-p+2)}.$$

If $n - p - 2 > 0$ and $\bar{c} > 0$, then for any $c \in (0, \bar{c}]$, AICc is inadmissible and dominated by

$$\text{MAICc} = \text{AICc} - c\hat{\sigma}^2 \|X\hat{\beta}\|^{-2}$$

as an estimator of the Kullback–Leibler discrepancy.

Note that the condition $\bar{c} > 0$ in Corollary 4.2 is equivalent to $p \geq 5$, as in Proposition 2.1.

5. Numerical results

5.1. Single response

First, we consider the case of single response (17), which corresponds to the model (1) with $q = 1$ and Corollary 4.2. We compare the mean squared errors of AICc and MAICc by Monte Carlo experiments with 10^6 repetitions. Each entry of X is generated from $N(0, 1)$ independently. We plot the percentage improvement in mean squared error (MSE) of MAICc over AICc:

$$100 \frac{E_{\beta, \sigma^2}[(\text{AICc} - d)^2] - E_{\beta, \sigma^2}[(\text{MAICc} - d)^2]}{E_{\beta, \sigma^2}[(\text{AICc} - d)^2]},$$

where $d = d((\beta, \sigma^2), (\hat{\beta}, \hat{\sigma}^2))$ is the Kullback–Leibler discrepancy (2).

Figure 3 compares MAICc with different values of c for $n = 30$, $p = 10$ and $\sigma^2 = 1$. The left panel indicates that MAICc with $c = \bar{c}$ dominates MAICc with $c = 0.5\bar{c}$, whereas the right panel shows that
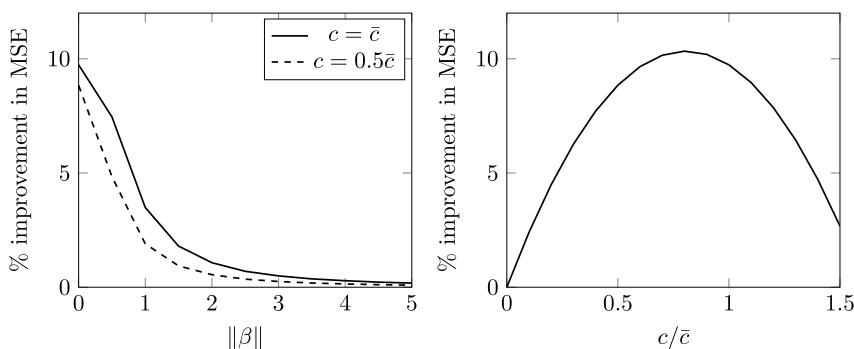

Figure 3. Percentage improvement in mean squared error of MAICc over AICc when $n = 30$, $p = 10$ and $\sigma^2 = 1$. Left: Plot with respect to $\|\beta\|$ for $c = \bar{c}$ and $c = 0.5\bar{c}$. Right: Plot with respect to $c/\bar{c}$ for $\beta = 0$.
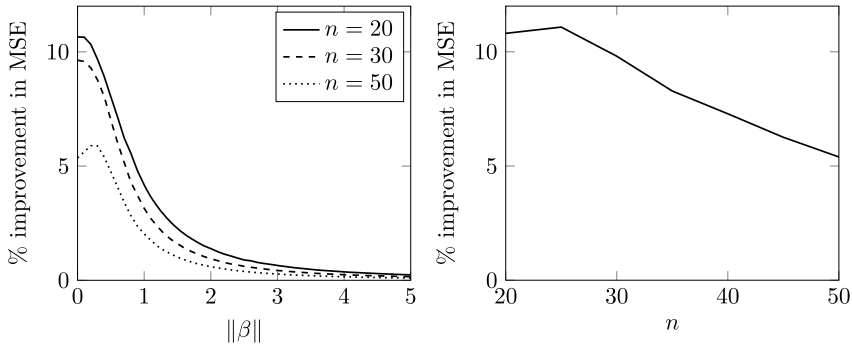
Figure 4. Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $p = 10$ and $\sigma^2 = 1$. Left: Plot with respect to $\|\beta\|$ for $n = 30, 50, 100$. Right: Plot with respect to n for $\beta = 0$.

the mean squared error of MAICc at $\beta = 0$ attains its minimum around $c = 0.8\bar{c}$. Overall, setting $c = \bar{c}$ in MAICc seems to be a reasonable choice. Thus, we adopt this value of c in the following experiments.

Figure 4 compares the performance of MAICc for different values of n when $p = 10$ and $\sigma^2 = 1$. It indicates that the percentage improvement in MSE is larger for smaller n. Figure 5 compares the performance of MAICc for different values of p when $n = 30$ and $\sigma^2 = 1$. It indicates that the percentage improvement in MSE is maximized around $p = 15$. Figure 6 compares the performance of MAICc for different values of σ^2 when $n = 30$ and $p = 10$. It indicates that the percentage improvement in MSE is larger for larger σ^2 at $\beta \neq 0$. Note that the percentage improvement in MSE at $\beta = 0$ does not depend on σ^2.

5.2. Multi-response

Now, consider the multi-response case (1). As in the previous subsection, we compare the mean squared error of AICc and MAICc by Monte Carlo experiments with 10^6 repetitions. Each entry of X is generated from N(0, 1) independently. We plot the percentage improvement in mean squared error (MSE) of
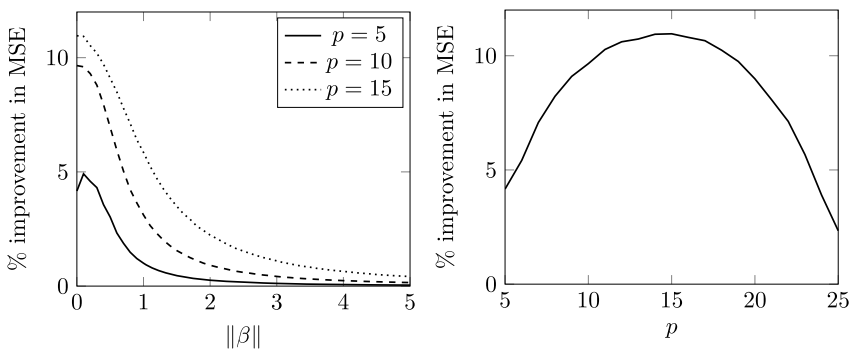

Figure 5. Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $n = 30$ and $\sigma^2 = 1$. Left: Plot with respect to $\|\beta\|$ for $p = 5, 10, 20$. Right: Plot with respect to p for $\beta = 0$.
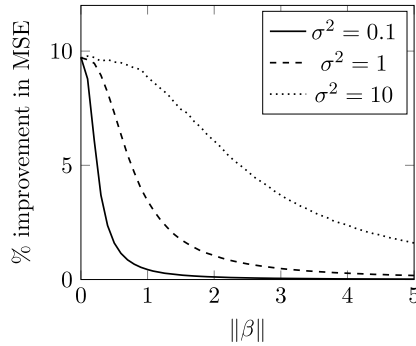
Figure 6. Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $n = 30$ and $p = 10$.

MAICc over AICc:

$$100\frac{\mathrm{E}_{B,\Sigma}[(\mathrm{AICc} - d)^2] - \mathrm{E}_{B,\Sigma}[(\mathrm{MAICc} - d)^2]}{\mathrm{E}_{B,\Sigma}[(\mathrm{AICc} - d)^2]},$$

where $d = d((B,\Sigma),(\hat{B},\hat{\Sigma}))$ is the Kullback–Leibler discrepancy (2).

Figure 7 compares MAICc with different values of c for $n = 30$, $p = 10$ and $q = 2$. The improvement is large when some of the singular values of M are small. Similarly to Figure 2, MAICc attains constant reduction of MSE as long as $\sigma_2(M) = 0$, even when $\sigma_1(M)$ is large. Thus, MAICc works well when B is close to low-rank, which corresponds to reduced-rank regression (Reinsel and Velu, 1998). We found that MAICc with $c = \bar{c}$ is numerically dominated by MAICc with $c = 2\bar{c}$, which implies that the upper bound for c in Theorem 4.1 may be improved. We adopt $c = \bar{c}$ in the following experiments.

Figure 8 compares the performance of MAICc for different values of n when $p = 10$, $q = 2$ and $\Sigma = I_2$. It indicates that the percentage improvement in MSE is maximized around $n = 40$. Figure 9 compares the performance of MAICc for different values of p when $n = 30$, $q = 2$ and $\Sigma = I_2$. It indicates that the percentage improvement in MSE is smaller for larger p. Figure 10 (left) compares the performance of MAICc for different values of $r = \Sigma_{12}$ when $n = 30$, $p = 10$, $q = 2$ and $\Sigma_{11} = \Sigma_{22} = 1$. It indicates that the percentage improvement in MSE is largest for $r = 0$ (no correlation).

Finally, Figure 10 (right) compares the performance of MAICc for different values of q when $n = 30$, $p = 10$, $\Sigma = I_q$. Compared to the single response case ($q = 1$) in the previous subsection, the percentage improvement in MSE of MAICc over AICc is not large. We expect that MAICc for $q \geq 2$ can be improved in several ways. For example, as shown in Figure 7, MAICc with $c = \bar{c}$ is numerically dominated by MAICc with larger value of c. Thus, improving the upper bound of c in Theorem 4.1 would be beneficial. Also, in analogy to the method of Efron and Morris (1976) for improving the Efron–Morris estimator by adding scalar shrinkage, MAICc is expected to be improved by adding a term of the form in Propostion 2.1 after vectorization. Another solution may be to use different coefficients for the singular values as in Proposition 2.2.

5.3. Variable selection

Here, we compare the variable selection performance of AIC, AICc, and MAICc with $c = \bar{c}$. The experimental setting is similar to that of Hurvich and Tsai (1989): $n = 20$, $p = 10$, $q = 1$, $\beta = (0.1, 0.2, 0.3, 0.4, 0.5, 0, 0, 0, 0, 0)^\top$ and $\sigma^2 = 1$. Each entry of X is generated from $N(0,1)$ independently
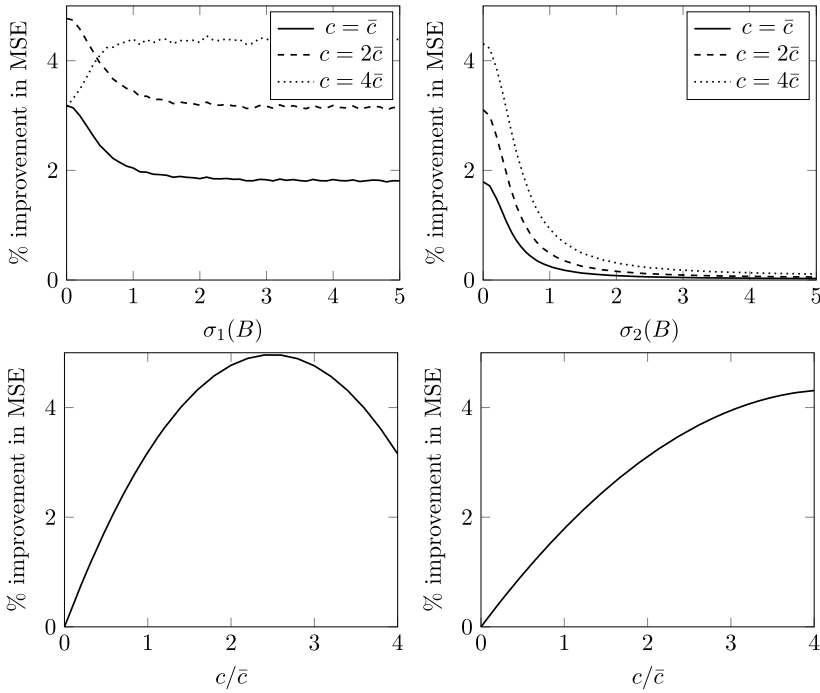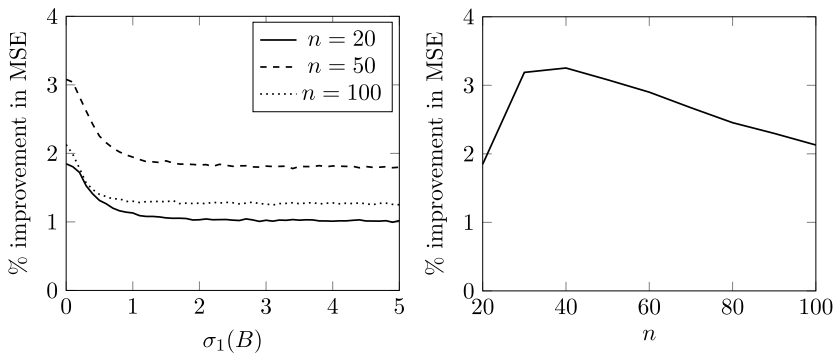
Figure 7. Percentage improvement in mean squared error of MAICc over AICc when $n = 30$, $p = 10$, $q = 2$ and $\Sigma = I_2$. Upper left: Plot with respect to $\sigma_1(B)$ for $c = \bar{c}$, $c = 2\bar{c}$ and $c = 4\bar{c}$ when $\sigma_2(B) = 0$. Upper right: Plot with respect to $\sigma_2(B)$ for $c = \bar{c}$, $c = 2\bar{c}$ and $c = 4\bar{c}$ when $\sigma_1(B) = 5$. Lower left: Plot with respect to $c/\bar{c}$ for $\sigma_1(B) = \sigma_2(B) = 0$ ($B = O$). Lower right: Plot with respect to $c/\bar{c}$ for $\sigma_1(B) = 5$ and $\sigma_2(B) = 0$.

and fixed for the whole experiment. Ten submodels were considered as candidate models, where the k-th model uses the first k columns of X as covariates ($k = 1, \ldots, 10$). Thus, the fifth model is the truth here. For each realization of y, we selected the model order k by minimizing AIC, AICc, or MAICc with $c = c^*$. Table 1 shows the frequency of the selected order in 1000 realizations. MAICc selects

Figure 8. Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $p = 10$, $q = 2$ and $\Sigma = I_2$. Left: Plot with respect to $\sigma_1(B)$ for $n = 20, 50, 100$ when $\sigma_2(B) = 0$. Right: Plot with respect to n for $B = O$.
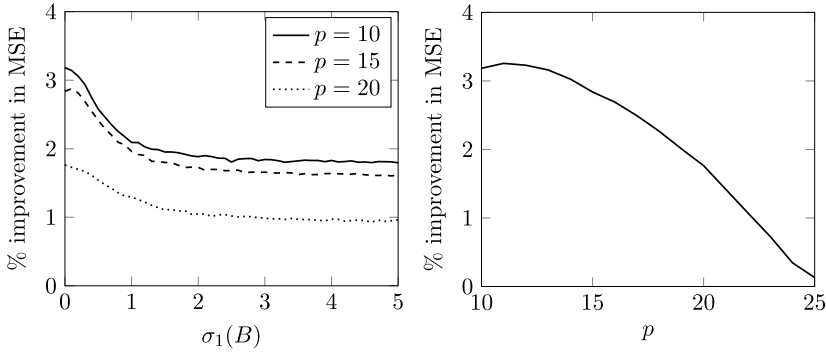
Figure 9. Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $n = 30$, $q = 2$ and $\Sigma = I_2$. Left: Plot with respect to $\sigma_1(B)$ for $p = 10, 15, 20$ when $\sigma_2(B) = 0$. Right: Plot with respect to p for $B = O$.

the true order $k = 5$ more frequently than AIC and AICc. Thus, MAICc attains better performance in variable selection than AIC and AICc. This result indicates a practical advantage of introducing the current loss estimation framework to investigate information criterion.

6. Conclusion

In this study, we showed that the corrected AIC is inadmissible as an estimator of the Kullback–Leibler discrepancy and provided improved loss estimators. To the best of our knowledge, such a loss estimation framework has not been employed in the study of information criteria, and there are several possible directions for future research. For example, generalizations of the current results to out-of-sample prediction (Rosset and Tibshirani, 2020), high-dimensional settings (Bellec and Zhang, 2021, Fujikoshi, Sakurai and Yanagihara, 2014, Yanagihara, Wakaki and Fujikoshi, 2015), and mis-specified cases (Fujikoshi and Satoh, 1997, Reschenhofer, 1999) may be interesting. Also, whereas we focused on the Gaussian linear regression model in this study, similar results may be obtained in general models
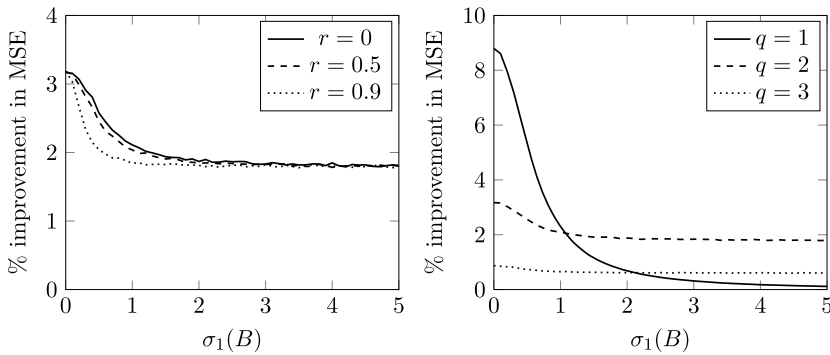

Figure 10. Left: Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $n = 30$, $p = 10$, $q = 2$, $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = r$. Plot with respect to r. Right: Percentage improvement in mean squared error of MAICc with $c = \bar{c}$ over AICc when $n = 30$, $p = 10$ and $\Sigma = I_q$. Plot with respect to $\sigma_1(B)$ for $q = 1, 2, 3$ when $\sigma_2(B) = \cdots = \sigma_q(B) = 0$.

Table 1. Frequency of order selected by three criteria in 1000 realizations

	1	2	3	4	5	6	7	8	9	10
AIC	89	8	15	29	352	129	76	76	81	145
AICc	277	147	37	16	460	44	15	4	0	0
MAICc	248	137	34	14	492	54	17	4	0	0

for AIC and other information criteria such as TIC and GIC (Konishi and Kitagawa, 2008) by asymptotic arguments. Improvement of model averaging criteria such as Mallows criterion (Hansen, 2007, Wan, Zhang and Zou, 2010) is another future problem. Finally, whereas we focused on the plug-in predictive distribution in this study, it would be interesting to study extension to the Bayesian predictive distribution, which minimizes the Bayes risk under the Kullback–Leibler loss (Aitchison, 1975). For the linear regression model, Kitagawa (1997) derived an information criterion for the Bayesian predictive distribution and Kobayashi and Komaki (2008) studied the problem of Bayesian out-of-sample prediction.

Appendix A: Matrix derivative formulas

Lemma A.1. *For* $Z \in \mathbb{R}^{p \times q}$,

$$\frac{\partial}{\partial Z_{ij}}((Z^\top Z)^{-1})_{kl} = -((Z^\top Z)^{-1})_{kj}(Z(Z^\top Z)^{-1})_{il} - ((Z^\top Z)^{-1}Z^\top)_{ki}((Z^\top Z)^{-1})_{jl}.$$

Proof. Let δ_{ab} be the Kronecker delta: $\delta_{ab} = 1$ if $a = b$ and $\delta_{ab} = 0$ if $a \neq b$. From

$$\frac{\partial}{\partial Z_{ij}}(Z^\top Z)_{ab} = \sum_c \frac{\partial}{\partial Z_{ij}}(Z_{ca}Z_{cb}) = \delta_{ja}Z_{ib} + \delta_{jb}Z_{ia}$$

and $d(A^{-1}) = -A^{-1}(dA)A^{-1}$, we obtain

$$
\begin{aligned}
\frac{\partial}{\partial Z_{ij}}((Z^\top Z)^{-1})_{kl} &= \left(-(Z^\top Z)^{-1}\left(\frac{\partial}{\partial Z_{ij}}(Z^\top Z)\right)(Z^\top Z)^{-1}\right)_{kl} \\
&= -\sum_{a,b}((Z^\top Z)^{-1})_{ka}\left(\frac{\partial}{\partial Z_{ij}}(Z^\top Z)_{ab}\right)((Z^\top Z)^{-1})_{bl} \\
&= -\sum_{a,b}((Z^\top Z)^{-1})_{ka}(\delta_{ja}Z_{ib} + \delta_{jb}Z_{ia})((Z^\top Z)^{-1})_{bl} \\
&= -\sum_b (Z^\top Z)^{-1})_{kj}Z_{ib}((Z^\top Z)^{-1})_{bl} - \sum_a ((Z^\top Z)^{-1})_{ka}Z_{ia}((Z^\top Z)^{-1})_{jl} \\
&= -((Z^\top Z)^{-1})_{kj}(Z(Z^\top Z)^{-1})_{il} - ((Z^\top Z)^{-1}Z^\top)_{ki}((Z^\top Z)^{-1})_{jl}. \qquad \square
\end{aligned}
$$

Lemma A.2. *For* $Z \in \mathbb{R}^{p \times q}$ *and* $S \in \mathbb{R}^{q \times q}$,

$$\frac{\partial}{\partial Z_{ij}}\mathrm{tr}(S(Z^\top Z)^{-1}) = -2(Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})_{ij}.$$

Proof. From Lemma A.1,

$$\frac{\partial}{\partial Z_{ij}}\text{tr}(S(Z^\top Z)^{-1}) = \sum_{k,l} S_{kl}\frac{\partial}{\partial Z_{ij}}((Z^\top Z)^{-1})_{kl}$$

$$= \sum_{k,l} S_{kl}(-((Z^\top Z)^{-1})_{kj}(Z(Z^\top Z)^{-1})_{il} - ((Z^\top Z)^{-1}Z^\top)_{ki}((Z^\top Z)^{-1})_{jl})$$

$$= -2(Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})_{ij}. \qquad \square$$

Lemma A.3. *For $Z \in \mathbb{R}^{p\times q}$ and $A, B \in \mathbb{R}^{q\times q}$,*

$$\sum_{ij}\frac{\partial}{\partial Z_{ij}}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{ij}$$

$$= (p - q - 2)\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - \text{tr}(A(Z^\top Z)^{-1})\text{tr}((Z^\top Z)^{-1}B).$$

Proof. From Lemma A.1,

$$\frac{\partial}{\partial Z_{ij}}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{ij}$$

$$= \sum_k\left(\frac{\partial}{\partial Z_{ij}}Z_{ik}\right)((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{kj} + \sum_{k,l}Z_{ik}\left(\frac{\partial}{\partial Z_{ij}}((Z^\top Z)^{-1})_{kl}\right)(A(Z^\top Z)^{-1}B)_{lj}$$

$$\qquad + \sum_{k,l}(Z(Z^\top Z)^{-1}A)_{ik}\left(\frac{\partial}{\partial Z_{ij}}((Z^\top Z)^{-1})_{kl}\right)B_{lj}$$

$$= \sum_k \delta_{jk}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{kj}$$

$$\qquad + \sum_{k,l}Z_{ik}(-((Z^\top Z)^{-1})_{kj}(Z(Z^\top Z)^{-1})_{il} - ((Z^\top Z)^{-1}Z^\top)_{ki}((Z^\top Z)^{-1})_{jl})(A(Z^\top Z)^{-1}B)_{lj}$$

$$\qquad + \sum_{k,l}(Z(Z^\top Z)^{-1}A)_{ik}(-((Z^\top Z)^{-1})_{kj}(Z(Z^\top Z)^{-1})_{il} - ((Z^\top Z)^{-1}Z^\top)_{ki}((Z^\top Z)^{-1})_{jl})B_{lj}$$

$$= ((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{jj} - (Z(Z^\top Z)^{-1})_{ij}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{ij}$$

$$\qquad - (Z(Z^\top Z)^{-1}Z^\top)_{ii}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{jj} - (Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1})_{ij}(Z(Z^\top Z)^{-1}B)_{ij}$$

$$\qquad - (Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}Z^\top)_{ii}((Z^\top Z)^{-1}B)_{jj}.$$

Therefore,

$$\sum_{ij}\frac{\partial}{\partial Z_{ij}}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)_{ij}$$

$$= p\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - \text{tr}((Z^\top Z)^{-1}Z^\top \cdot Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)$$

$$\qquad - \text{tr}(Z(Z^\top Z)^{-1}Z^\top)\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - \text{tr}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1} \cdot B^\top(Z^\top Z)^{-1}Z^\top)$$

$$\qquad - \text{tr}(Z(Z^\top Z)^{-1}A(Z^\top Z)^{-1}Z^\top)\text{tr}((Z^\top Z)^{-1}B)$$

$$=p\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - \text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - q\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B)$$
$$- \text{tr}(A(Z^\top Z)^{-1} \cdot B^\top(Z^\top Z)^{-1}) - \text{tr}(A(Z^\top Z)^{-1})\text{tr}((Z^\top Z)^{-1}B)$$
$$=(p - q - 2)\text{tr}((Z^\top Z)^{-1}A(Z^\top Z)^{-1}B) - \text{tr}(A(Z^\top Z)^{-1})\text{tr}((Z^\top Z)^{-1}B). \qquad \square$$

Appendix B: Expectation formulas

Lemma B.1 ([Stein, 1974]). *If $Z \sim N_{p,q}(O, I_p, I_q)$ and $g : \mathbb{R}^{p \times q} \to \mathbb{R}^{p \times q}$ is absolutely continuous, then*

$$E[\text{tr}(Z^\top g(Z))] = E\left[\sum_{i,j} \frac{\partial g_{ij}}{\partial Z_{ij}}(Z)\right].$$

Lemma B.2 ([Gupta and Nagar, 2000], **Theorem 3.3.15 (iv)**). *If $d - q \geq 0$, $S \sim W_q(d, \Sigma)$ and $A, B \in \mathbb{R}^{q \times q}$, then*

$$E[\text{tr}(AS)\text{tr}(BS)] = d\text{tr}(A\Sigma B\Sigma) + d\text{tr}(A^\top \Sigma B\Sigma) + d^2\text{tr}(A\Sigma)\text{tr}(B\Sigma).$$

Lemma B.3 ([Gupta and Nagar, 2000], **Theorem 3.3.16 (i)**). *If $d - q - 1 > 0$ and $S \sim W_q(d, \Sigma)$, then*

$$E[S^{-1}] = \frac{1}{d - q - 1}\Sigma^{-1}.$$

Lemma B.4 ([Gupta and Nagar, 2000], **Theorem 3.3.16 (iii)**). *If $d - q - 3 > 0$, $S \sim W_q(d, \Sigma)$ and $A \in \mathbb{R}^{q \times q}$ is positive semidefinite, then*

$$E[S^{-1}AS^{-1}] = \frac{\text{tr}(\Sigma^{-1}A)}{(d - q)(d - q - 1)(d - q - 3)}\Sigma^{-1} + \frac{1}{(d - q)(d - q - 3)}\Sigma^{-1}A\Sigma^{-1}.$$

In particular, when $\Sigma = A = I_q$,

$$E[S^{-2}] = \frac{d - 1}{(d - q)(d - q - 1)(d - q - 3)}I_q.$$

Lemma B.5 ([Gupta and Nagar, 2000], **Theorem 3.3.17 (ii)**). *If $d - q - 3 > 0$ and $S \sim W_q(d, \Sigma)$, then*

$$E[\text{tr}(S^{-1})S^{-1}] = \frac{d - q - 2}{(d - q)(d - q - 1)(d - q - 3)}\text{tr}(\Sigma^{-1})\Sigma^{-1} + \frac{2}{(d - q)(d - q - 1)(d - q - 3)}\Sigma^{-2}.$$

In particular, when $\Sigma = I_q$,

$$E[\text{tr}(S^{-1})S^{-1}] = \frac{q(d - q - 2) + 2}{(d - q)(d - q - 1)(d - q - 3)}I_q.$$

Lemma B.6 ([Gupta and Nagar, 2000], **Theorem 3.3.17 (iii)**). *If $d - q - 1 > 0$ and $S \sim W_q(d, \Sigma)$, then*

$$E[\text{tr}(S^{-1})S] = \frac{d}{d - q - 1}\text{tr}(\Sigma^{-1})\Sigma - \frac{2}{d - q - 1}I_q.$$

In particular, when $\Sigma = I_q$,

$$\mathrm{E}[\mathrm{tr}(S^{-1})S] = \frac{dq-2}{d-q-1}I_q.$$

Lemma B.7 (Styan (1989))**.** *If $d - q - 1 > 0$, $S \sim W_q(d, \Sigma)$ and $A \in \mathbb{R}^{q \times q}$, then*

$$\mathrm{E}[SAS^{-1}] = \frac{1}{d-q-1}(d\Sigma A \Sigma^{-1} - A^\top - \mathrm{tr}(A)I_q).$$

In particular, when $\Sigma = I_q$ and $A^\top = A$,

$$\mathrm{E}[SAS^{-1}] = \frac{1}{d-q-1}((d-1)A - \mathrm{tr}(A)I_q).$$

Lemma B.8. *If $n - p - q - 1 > 0$ and $Z \sim \mathrm{N}_{p,q}(\bar{Z}, I_p, I_q)$, then*

$$\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top(Z-\bar{Z}))\mathrm{tr}((Z^\top Z)^{-1})]$$
$$= pq\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] - 2(p-q-2)\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})] + 2\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{18}$$

Proof. From Lemma B.1 and Lemma A.2,

$$\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top(Z-\bar{Z}))\mathrm{tr}((Z^\top Z)^{-1})]$$
$$= \mathrm{E}\left[\sum_{i,j}\frac{\partial}{\partial Z_{ij}}((Z-\bar{Z}))_{ij}\mathrm{tr}((Z^\top Z)^{-1}))\right]$$
$$= \mathrm{E}\left[\sum_{i,j}(\mathrm{tr}((Z^\top Z)^{-1}) - 2((Z-\bar{Z}))_{ij}(Z(Z^\top Z)^{-1}(Z^\top Z)^{-1})_{ij})\right]$$
$$= pq\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] - 2\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top Z(Z^\top Z)^{-1}(Z^\top Z)^{-1})]. \tag{19}$$

From Lemma B.1 and Lemma A.3,

$$\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top Z(Z^\top Z)^{-1}(Z^\top Z)^{-1})] = \mathrm{E}\left[\sum_{i,j}\frac{\partial}{\partial Z_{ij}}(Z(Z^\top Z)^{-1}(Z^\top Z)^{-1})_{ij}\right]$$
$$= \mathrm{E}[(p-q-2)\mathrm{tr}((Z^\top Z)^{-2}) - (\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{20}$$

Substituting (20) into (19), we obtain (18). $\square$

Lemma B.9. *If $n - p - q - 1 > 0$, $Z \sim \mathrm{N}_{p,q}(\bar{Z}, I_p, I_q)$ and $S \sim W_q(n-p, (1/n)I_q)$, then*

$$\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top(Z-\bar{Z})S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]$$
$$= p\frac{(n-p)q-2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})] - 2\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]$$
$$+ 2\frac{n-q-2}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{21}$$

Proof. From Lemma B.1 and Lemma A.2,

$$
\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top(Z-\bar{Z})S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]
$$

$$
=\mathrm{E}\left[\sum_{i,j}\frac{\partial}{\partial Z_{ij}}((Z-\bar{Z})S^{-1})_{ij}\mathrm{tr}(S(Z^\top Z)^{-1}))\right]
$$

$$
=\mathrm{E}\left[\sum_{i,j}((S^{-1})_{jj}\mathrm{tr}(S(Z^\top Z)^{-1})-2((Z-\bar{Z})S^{-1})_{ij}(Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})_{ij})\right]
$$

$$
=p\mathrm{E}[\mathrm{tr}(S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]-2\mathrm{E}[\mathrm{tr}(S^{-1}(Z-\bar{Z})^\top Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})]. \tag{22}
$$

Then, from the linearity of expectation and Lemma B.6,

$$
\mathrm{E}[\mathrm{tr}(S^{-1})\mathrm{tr}(S(Z^\top Z)^{-1})]=\mathrm{tr}(\mathrm{E}[\mathrm{tr}(S^{-1})S\cdot(Z^\top Z)^{-1}])
$$

$$
=\mathrm{tr}(\mathrm{E}[\mathrm{tr}(S^{-1})S]\cdot\mathrm{E}[(Z^\top Z)^{-1}])
$$

$$
=\frac{(n-p)q-2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1})]. \tag{23}
$$

Also, from Lemma B.1 and Lemma A.3,

$$
\mathrm{E}[\mathrm{tr}(S^{-1}(Z-\bar{Z})^\top Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})]
$$

$$
=\mathrm{E}[\mathrm{tr}((Z-\bar{Z})^\top Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1}S^{-1})]
$$

$$
=\mathrm{E}\left[\sum_{i,j}\frac{\partial}{\partial Z_{ij}}(Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1}S^{-1})_{ij}\right]
$$

$$
=\mathrm{E}[(p-q-2)\mathrm{tr}((Z^\top Z)^{-1}S(Z^\top Z)^{-1}S^{-1})-\mathrm{tr}((Z^\top Z)^{-1}S)\mathrm{tr}((Z^\top Z)^{-1}S^{-1})]. \tag{24}
$$

Now, from Lemma B.7,

$$
\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1}S(Z^\top Z)^{-1}S^{-1})]
$$

$$
=\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1}\cdot\mathrm{E}[S(Z^\top Z)^{-1}S^{-1}\mid Z])]
$$

$$
=\mathrm{E}\left[\mathrm{tr}\left((Z^\top Z)^{-1}\cdot\frac{1}{n-p-q-1}((n-p-1)(Z^\top Z)^{-1}-\mathrm{tr}((Z^\top Z)^{-1})I_q)\right)\right]
$$

$$
=\frac{n-p-1}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]-\frac{1}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{25}
$$

Also, by putting $\tilde{S}=((Z^\top Z)^{1/2}S((Z^\top Z)^{1/2}\sim W_q(n-p,Z^\top Z)$ and using Lemma B.6,

$$
\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-1}S)\mathrm{tr}((Z^\top Z)^{-1}S^{-1})]
$$

$$
=\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2}\tilde{S})\mathrm{tr}(\tilde{S}^{-1})]
$$

$$
=\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2}\cdot\mathrm{E}[\mathrm{tr}(\tilde{S}^{-1})\tilde{S}\mid Z])]
$$

$$=\mathrm{E}\left[\mathrm{tr}((Z^\top Z)^{-2}\left(\frac{n-p}{n-p-q-1}\mathrm{tr}((Z^\top Z)^{-1})(Z^\top Z)-\frac{2}{n-p-q-1}I_q\right)\right]$$

$$=\frac{n-p}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]-\frac{2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]. \tag{26}$$

Thus, by substituting (25) and (26) into (24),

$$\mathrm{E}[\mathrm{tr}(S^{-1}(Z-\bar{Z})^\top Z(Z^\top Z)^{-1}S(Z^\top Z)^{-1})]$$

$$=(p-q-2)\left(\frac{n-p-1}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]-\frac{1}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]\right)$$

$$-\left(\frac{n-p}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]-\frac{2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]\right)$$

$$=\frac{(n-p-1)(p-q-2)+2}{n-p-q-1}\mathrm{E}[\mathrm{tr}((Z^\top Z)^{-2})]-\frac{n-q-2}{n-p-q-1}\mathrm{E}[(\mathrm{tr}((Z^\top Z)^{-1}))^2]. \tag{27}$$

Substituting (23) and (27) into (22), we obtain (21). □

Acknowledgments

Funding

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353 https://doi.org/10.1093/biomet/62.3.547

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. Budapest: Akadémiai Kiadó. MR0483125

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley. MR0091588

Bedrick, E.J. and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50** 226–231.

Bellec, P.C. and Zhang, C.-H. (2021). Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *Ann. Statist.* **49** 1864–1903. MR4319234 https://doi.org/10.1214/20-aos2005

Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W. and Wells, M.T. (2014). Akaike's information criterion, C_P and estimators of loss for elliptically symmetric distributions. *Int. Stat. Rev.* **82** 422–439. MR3280984 https://doi.org/10.1111/insr.12052

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer. MR1919620

Cavanaugh, J.E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statist. Probab. Lett.* **33** 201–208. MR1458291 https://doi.org/10.1016/S0167-7152(96)00128-9

Davies, S.L., Neath, A.A. and Cavanaugh, J.E. (2006). Estimation optimality of corrected AIC and modified Cp in linear regression. *Int. Stat. Rev.* **74** 161–168.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347. MR0334386 https://doi.org/10.1093/biomet/59.2.335

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960

Fourdrinier, D. and Strawderman, W.E. (2003). On Bayes and unbiased estimators of loss. *Ann. Inst. Statist. Math.* **55** 803–816. MR2028618 https://doi.org/10.1007/BF02523394

Fourdrinier, D., Strawderman, W.E. and Wells, M.T. (2018). *Shrinkage Estimation. Springer Series in Statistics*. Cham: Springer. MR3887633 https://doi.org/10.1007/978-3-030-02185-6

Fourdrinier, D. and Wells, M.T. (2012). On improved loss estimation for shrinkage estimators. *Statist. Sci.* **27** 61–81. MR2953496 https://doi.org/10.1214/11-STS380

Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_P-type criteria in multivariate linear regression. *J. Multivariate Anal.* **123** 184–200. MR3130429 https://doi.org/10.1016/j.jmva.2013.09.006

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_P in multivariate linear regression. *Biometrika* **84** 707–716. MR1603952 https://doi.org/10.1093/biomet/84.3.707

Gupta, A.K. and Nagar, D.K. (2000). *Matrix Variate Distributions. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics* **104**. Boca Raton, FL: CRC Press/CRC. MR1738933

Hansen, B.E. (2007). Least squares model averaging. *Econometrica* **75** 1175–1189. MR2333497 https://doi.org/10.1111/j.1468-0262.2007.00785.x

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020 https://doi.org/10.1093/biomet/76.2.297

Johnstone, I. (1988). On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics, IV, Vol. 1 (West Lafayette, Ind., 1986)* 361–379. New York: Springer. MR0927112

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Comm. Statist. Theory Methods* **26** 2223–2246. MR1484250 https://doi.org/10.1080/03610929708832043

Kobayashi, K. and Komaki, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.* **99** 1888–1905. MR2466542 https://doi.org/10.1016/j.jmva.2008.01.014

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling. Springer Series in Statistics*. New York: Springer. MR2367855 https://doi.org/10.1007/978-0-387-71887-3

Lehmann, E.L. and Casella, G. (2006). *Theory of Point Estimation*. New York: Springer. MR1639875

Lu, K.L. and Berger, J.O. (1989). Estimation of normal means: Frequentist estimation of loss. *Ann. Statist.* **17** 890–906. MR0994274 https://doi.org/10.1214/aos/1176347149

Matsuda, T. and Strawderman, W.E. (2016). Pitman closeness properties of Bayes shrinkage procedures in estimation and prediction. *Statist. Probab. Lett.* **119** 21–29. MR3555265 https://doi.org/10.1016/j.spl.2016.07.005

Matsuda, T. and Strawderman, W.E. (2019). Improved loss estimation for a normal mean matrix. *J. Multivariate Anal.* **169** 300–311. MR3875601 https://doi.org/10.1016/j.jmva.2018.10.001

Matsuda, T. and Strawderman, W.E. (2022). Estimation under matrix quadratic loss and matrix superharmonicity. *Biometrika* **109** 503–519. MR4430971 https://doi.org/10.1093/biomet/asab025

Narayanan, R. and Wells, M.T. (2015). Improved loss estimation for the Lasso: A variable selection tool. *Sankhya B* **77** 45–74. MR3337562 https://doi.org/10.1007/s13571-015-0095-1

Reinsel, G.C. and Velu, R.P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications. Lecture Notes in Statistics* **136**. New York: Springer. MR1719704 https://doi.org/10.1007/978-1-4757-2853-8

Reschenhofer, E. (1999). Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory* **15** 377–387. MR1704229 https://doi.org/10.1017/S0266466699153052

Rosset, S. and Tibshirani, R.J. (2020). From Fixed-X to Random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *J. Amer. Statist. Assoc.* **115** 138–151. MR4078450 https://doi.org/10.1080/01621459.2018.1424632

Sandved, E. (1968). Ancillary statistics and prediction of the loss in estimation problems. *Ann. Math. Stat.* **39** 1756–1758. MR0231481 https://doi.org/10.1214/aoms/1177698162

Stein, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II* 345–381. Prague: Charles Univ. MR0381062

Styan, G.P.H. (1989). Three useful expressions for expectations involving a Wishart matrix and its inverse. In *Statistical Data Analysis and Inference (Neuchâtel, 1989)* 283–296. Amsterdam: North-Holland. MR1089643 https://doi.org/10.1016/B978-0-444-88029-1.50032-0

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods* **7** 13–26.

Wan, A.T.K., Zhang, X. and Zou, G. (2010). Least squares model averaging by Mallows criterion. *J. Econometrics* **156** 277–283. MR2609932 https://doi.org/10.1016/j.jeconom.2009.10.030

Wan, A.T.K. and Zou, G. (2004). On unbiased and improved loss estimation for the mean of a multivariate normal distribution with unknown variance. *J. Statist. Plann. Inference* **119** 17–22. MR2018447 https://doi.org/10.1016/S0378-3758(02)00406-8

Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.* **9** 869–897. MR3338666 https://doi.org/10.1214/15-EJS1022