CIRJE-F-709

Selection of Variables in Multivariate Regression Models for Large Dimensions

Muni S. Srivastava
University of Toronto

Tatsuya Kubokawa
University of Tokyo

January 2010

Selection of Variables in Multivariate Regression Models for Large Dimensions

Muni S. Srivastava*and Tatsuya Kubokawa†
University of Toronto and University of Tokyo

January 13, 2010

Abstract

The Akaike information criterion, AIC, and Mallows' C_p statistic have been proposed for selecting a smaller number of regressor variables in the multivariate regression models with fully unknown covariance matrix. All these criteria are, however, based on the implicit assumption that the sample size is substantially larger than the dimension of the covariance matrix. To obtain a stable estimator of the covariance matrix, it is required that the dimension of the covariance matrix be much smaller than the sample size. When the dimension is close to the sample size, it is necessary to use ridge type of estimators for the covariance matrix. In this paper, we use a ridge type of estimators for the covariance matrix and obtain the modified AIC and modified C_p statistic under the asymptotic theory that both the sample size and the dimension go to infinity. It is numerically shown that these modified procedures perform very well in the sense of selecting the true model in large dimensional cases.

Key words and phrases: Akaike information criterion, Mallows' C_p, large dimension, multivariate linear regression model, selection of variables.

1 Introduction

Consider a multivariate linear regression model in which p response variables $y_1, \dots, y_p$ are regressed on k explanatory variables $x_{(1)}, \dots, x_{(K)}$, when n observations are available on $y_1, \dots, y_p$ and $x_{(1)}, \dots, x_{(K)}$. Let $\boldsymbol{Y}$ denotes the $n \times p$ observation matrix on the response variable, and $\widetilde{\boldsymbol{X}}$ denotes the $n \times K$ observation matrix on the K explanatory variables. Then the multivariate regression model is given by

$$\text{Full Model :} \qquad \boldsymbol{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F + \boldsymbol{E}, \tag{1}$$

*Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, E-Mail: srivasta@utstat.toronto.edu

†Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

where the n rows of $\boldsymbol{E}$ are independent and identically distributed (iid) as multivariate normal with mean vector zero and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, that is, $\boldsymbol{e}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{E} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)'$ and $\boldsymbol{\beta}_F$ is a $K \times p$ matrix of unknown parameters. The $n \times p$ matrix $\boldsymbol{Y}$ is given by $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)'$ and the $n \times K$ matrix $\widetilde{\boldsymbol{X}}$ is given by $\widetilde{\boldsymbol{X}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)' = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(K)})$ where $\boldsymbol{y}_i$'s are random p-vector and $\boldsymbol{x}_i$'s and $\boldsymbol{x}_{(i)}$ are, respectively, K and n-vectors considered known or fixed.

In this paper, we consider the model (1) as a full model and we want to address the problem of selecting the explanatory variables $x_{(1)}, \dots, x_{(K)}$ when n and p are large. When k variables $x_{(\gamma_1)}, \dots, x_{(\gamma_k)}$ are selected from $\{x_{(1)}, \dots, x_{(K)}\}$, the candidate model is written as

$$\text{Candidate Model :} \qquad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{X} = (\boldsymbol{x}_{(\gamma_1)}, \dots, \boldsymbol{x}_{(\gamma_k)})$, and $\boldsymbol{\beta}$ is a $k \times p$ matrix of unknown parameters. For simplicity, we hereafter write $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(k)})$ without any loss of generality. The above model is written as

$$\boldsymbol{Y} \sim \mathcal{N}_{n,p}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n, \boldsymbol{\Sigma}). \tag{3}$$

The *Akaike Information Criterion* (AIC) proposed by Akaike (1973, 1974) is recognized as a useful tool for selecting variables in linear regression models. For obtaining an expression for the AIC, we shall assume that the model given in (2) is an overspecified model, and the true model is given by

$$\text{True Model :} \qquad \boldsymbol{Y} \sim \mathcal{N}_{n,p}(\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{I}_n, \boldsymbol{\Sigma}^*). \tag{4}$$

It will be assumed that the true model belongs to the overspecified model (2). Let $f(\boldsymbol{Y}; \boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)$ denote the pdf of the true model, namely,

$$f(\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*) = (2\pi)^{-pn/2}|\boldsymbol{\Sigma}^*|^{-n/2}\mathrm{etr}\left[-\frac{1}{2}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)\right].$$

Let $\widehat{\boldsymbol{\beta}}(\boldsymbol{Y})$ and $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{Y})$ be estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ based on the candidate model. When the true model is predicted based on the candidate model, the prediction error relative to the the Kullback-Leibler information is given by

$$R_{KL}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}) = E^*_{\boldsymbol{Y}}[E^*_{\boldsymbol{Z}}[\log\{f(\boldsymbol{Z}|\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)/f(\boldsymbol{Z}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}), \widehat{\boldsymbol{\Sigma}}(\boldsymbol{Y})\}]], \tag{5}$$

where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independently distributed but having the same distribution as $f(\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)$ and $f(\boldsymbol{Z}|\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)$. Let us define the *Akaike Information* (AI) by

$$AI = -2E^*_{\boldsymbol{Y}}\left[E^*_{\boldsymbol{Z}}[\log f(\boldsymbol{Z}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}), \widehat{\boldsymbol{\Sigma}}(\boldsymbol{Y}))]\right], \tag{6}$$

which is a model-related part of the prediction error $R_{KL}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$. Then the AIC is generally defined as an asymptotically unbiased estimator of AI, where $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are estimated by the maximum likelihood estimators (MLE), given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}} =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},\\ \widehat{\boldsymbol{\Sigma}}_0 =& \boldsymbol{S}/n = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'/n,\end{aligned}$$

where $\boldsymbol{S} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'$. For more accounts on the AIC and the related selection criteria, see Sugiura (1978) and Konishi and Kitagawa (2007).

The AIC and the modified criterion in the multivariate linear regression model were derived by Fujikoshi and Satoh (1997) when $n \to \infty$ and p is bounded. In the large dimensional case, namely the case that $p \to \infty$, the MLE $\widehat{\boldsymbol{\Sigma}}_0$ and the inverse matrix $\widehat{\boldsymbol{\Sigma}}_0^{-1}$ must be instable or nonexistent, which means that the AIC based on the MLE $\widehat{\boldsymbol{\Sigma}}_0$ is not appropriate. Srivastava and Kubokawa (2008) considered the ridge type estimator $\widehat{\boldsymbol{\Sigma}}_\lambda = (\boldsymbol{S} + \hat{\lambda}\boldsymbol{I}_p)/n$ for a function $\hat{\lambda} = \hat{\lambda}(\boldsymbol{S})$ instead of the MLE, and derived the AIC when $p > n$, $p \to \infty$ and n is bounded based on the theory given in Srivastava (2007). Recently, Yamamura, Yanagihara and Srivastava (2009) obtained the AIC when $p > n$, $n - k = O(p^\delta)$ and $(n, p) \to \infty$ for $0 < \delta < 1/3$.

In this paper, we consider the case that

$$\nu_k \equiv n - k - p - 3 > 0 \text{ and } (n, p) \to \infty \text{ such that } p/n = c \text{ for } 0 < c < 1, \tag{7}$$

where the condition of $n - k - p - 3 > 0$ is required for the existence of the moment $E[\mathrm{tr}\,[\boldsymbol{S}^{-2}]]$. Since $n - k > p + 1$, there exists the inverse matrix of the MLE $\widehat{\boldsymbol{\Sigma}}_0$. Thus, the AIC based on the MLE are available, but not appropriate for large dimension p, because the MLE is very unstable when p is large, see, e.g.,Johnston (2001). In this case, the ridge-type estimator $\widehat{\boldsymbol{\Sigma}}_\lambda$ should be used instead of the MLE, and we obtain the AIC based on $\widehat{\boldsymbol{\Sigma}}_\lambda$.

When a squared error loss function is employed instead of the Kullback-Leibler information, the prediction error is given by

$$R_{PE}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}}) = E_{\boldsymbol{Y}}^*[E_{\boldsymbol{Z}}^*[\mathrm{tr}\,[(\boldsymbol{Z} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}))']]]. \tag{8}$$

Corresponding to the derivation of the AIC, we can suggest an unbiased estimator of $R_{PE}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}})$ for the model selection. The unbiased estimator is related to the *Mallows' C_p statistic* proposed by Mallows (1973), and we here call it the C_p-type statistic. In this paper, we also obtain the C_p-type statistic based on the ridge-type estimator $\widehat{\boldsymbol{\Sigma}}_\lambda$.

The AIC and C_p-type statistics based on the ridge-type estimator $\widehat{\boldsymbol{\Sigma}}_\lambda$ are given in Section 2. We also propose the double ridge AIC and C_p-type statistics based on $\widehat{\boldsymbol{\Sigma}}_\lambda$ and the ridge regression estimator of $\boldsymbol{\beta}$, which can be expected to work well in the multicolinearity case of $\boldsymbol{X}$. The proofs of their derivation is given in Section 3. A simulation experiment is carried out in Section 4 to compare the AIC and C_p criteria for different value of λ including $\lambda = 0$, and it is shown that the usual AIC and C_p based on the MLE do not work in the large dimensional case, but the AIC and C_p statistics based on the ridge-type estimator perform very well in all the cases. We conclude in Section 5.

2 Ridge-type variable selection procedures

2.1 Ridge-type AIC

In this section, we derive ridge-type AIC and C_p statistic based on the ridge-type estimator of $\boldsymbol{\Sigma}$. The ridge-type estimator we want to propose for $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}}_\lambda = n^{-1}(\boldsymbol{S} + \hat{\lambda}\boldsymbol{I}_p), \tag{9}$$

where

$$\hat{\lambda} = c_n(\mathrm{tr}\, \boldsymbol{S}/np), \quad \text{for} \quad c_n = O(n^{-\delta}),\ \delta \geq 0. \tag{10}$$

Let us define the Akaike information based on the ridge-type estimator by

$$AI_\lambda = -2E^*_{\boldsymbol{Y}}\left[E^*_{\boldsymbol{Z}}[\log f(\boldsymbol{Z}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{Y}), \widehat{\boldsymbol{\Sigma}}_\lambda(\boldsymbol{Y}))]\right].$$

The Akaike information criterion is an asymptotically unbiased estimator of AI_λ based on $-2\log f(\boldsymbol{Y}; \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_\lambda)$, where the bias is given by

$$\Delta_\lambda = \Delta_\lambda(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_\lambda) = AI_\lambda - E^*_{\boldsymbol{Y}}[-2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_\lambda)]. \tag{11}$$

When Δ_λ is estimated by Δ^*_λ, the AIC is provided by

$$AIC_\lambda = -2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}_\lambda) + \Delta^*_\lambda, \tag{12}$$

We shall assume that $\lim_{p\to\infty} \mathrm{tr}\, \boldsymbol{\Sigma}/p \in (0, \infty)$. Under this assumption, it follows from Srivastava (2005) that $\mathrm{tr}\, \boldsymbol{S}/np \to \mathrm{tr}\, \boldsymbol{\Sigma}/p$ in probability as $(n, p) \to \infty$. We shall consider the case when $c_n = n/p$, other choices of c_n can also be considered. We obtain an asymptotic expression for the bias in calculating the AIC under (7) and the assumption

$$\lim_{p\to\infty} \mathrm{tr}\, [\boldsymbol{\Sigma}]/p < \infty. \tag{13}$$

Theorem 2.1 *Assume the conditions* (7) *and* (13). *Then,* Δ_λ *given in* (11) *is approximated as*

$$\begin{aligned}\Delta_\lambda =& \frac{np(p+1+2k)}{n-k-p-1} \\ &+ \frac{c_n(n-k)}{p(n-k-p-1)}\left\{\frac{(n+k)(n-k)}{(n-p)^2} - 1\right\}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}).\end{aligned} \tag{14}$$

The unknown quantity $\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]$ *is estimated based on the equality*

$$E^*_{\boldsymbol{Y}}\left[\hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}]\right] = \frac{c_n(n-k)}{p(n-k-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}). \tag{15}$$

Combining these approximations yields AIC_λ *given by*

$$\begin{aligned}AIC_\lambda =& np\log 2\pi + n\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}] \\ &+ \frac{np(p+1+2k)}{n-k-p-1} + \left\{\frac{(n+k)(n-k)}{(n-p)^2} - 1\right\}\hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}],\end{aligned} \tag{16}$$

From Theorem 2.1, the AIC based on the MLE $\widehat{\boldsymbol{\Sigma}}_0$ is derived by putting $\hat{\lambda} = 0$ in (16) as

$$AIC_0 = np \log 2\pi + n \log |\widehat{\boldsymbol{\Sigma}}_0| + np + \frac{np(p+1+2k)}{n-k-p-1}. \tag{17}$$

2.2 Ridge-type C_p

As explained above, the AIC is an asymptotically unbiased estimator of a part of the prediction error based on the Kullback-Leiber information. We here use the squared error loss function instead of the Kullback-Leibler information, and consider to estimate the prediction error given by

$$R_{PE}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}}) = E^*_{\boldsymbol{Y}}[E^*_{\boldsymbol{Z}}[\mathrm{tr}\,[(\boldsymbol{Z} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})']]].$$

This is rewritten as $R_{PE}(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \widehat{\boldsymbol{\beta}}) = np + PE$, where

$$PE = E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]]. \tag{18}$$

Since an unbiased estimator of PE is related to the C_p statistic, we here call it the C_p-type statistic. According to the same arguments as in the derivation of the Mallows' C_p statistic, we estimate the covaraince matrix $\boldsymbol{\Sigma}$ based on the full model (1). Let $\widetilde{\boldsymbol{S}} = n^{-1}(\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}})'(\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}})$ for $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{Y}$. Then $\boldsymbol{\Sigma}$ is estimated by

$$\widetilde{\boldsymbol{\Sigma}}_\lambda = n^{-1}(\widetilde{\boldsymbol{S}} + \widetilde{\lambda}\boldsymbol{I}_p),$$

where

$$\widetilde{\lambda} = c_n(\mathrm{tr}\,\widetilde{\boldsymbol{S}}/np), \quad \text{for} \quad c_n = O(n^{-\delta}),\ \delta \geq 0.$$

When PE is estimated based on the statistic $\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})]$, the bias is $\Delta_{PE} = PE - E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})]]$. Then, the C_p statistic based on the ridge-type estimator $\widetilde{\boldsymbol{\Sigma}}_\lambda$ is given by

$$C_{p,\lambda} = \mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})] + \Delta^*_{PE},$$

where Δ^*_{PE} is an estimator of Δ_{PE}.

Theorem 2.2 *Assume the conditions* (7) *and* (13). *Then, Δ_{PE} is evaluated as*

$$\Delta_{PE} = \frac{np(n-k-p-1)}{n-K-p-1} - \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}). \tag{19}$$

Let us define C_λ by

$$C_\lambda = \mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}] - \frac{np(n-k-p-1)}{n-K-p-1} + pk + \widetilde{\lambda}\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}^{-1}]. \tag{20}$$

Then, C_λ is an asymptotically unbiased estimator of PE given in (18), *namely, $E[C_\lambda] = PE + O(n^{-\delta})$.*

When $\widetilde{\lambda} = 0$, from Theorem 2.2, we get Mallows' C_p statistic based on the MLE, given by

$$C_0 = n\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{S}] - \frac{np(n-k-p-1)}{n-K-p-1} + pk. \tag{21}$$

2.3 An extension to double ridge criteria for selection

We are often faced with the multicolinearity cases, where the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K$ are highly correlated for $\widetilde{\boldsymbol{X}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)$. In this case, the inverse matrix of $\boldsymbol{X}'\boldsymbol{X}$ is not stable, and it is known that the least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ does not behave well. An alternative procedure is the ridge regression estimator

$$\widehat{\boldsymbol{\beta}}_\tau = (\boldsymbol{X}'\boldsymbol{X} + \tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

where τ is a nonnegative constant. It is certain that $\widehat{\boldsymbol{\beta}}_\tau$ must be stable for an aprropriate constant τ, which results in a good predictor based on $\widehat{\boldsymbol{\beta}}_\tau$. However, it may be important how to determine τ. A possible method in the framework of variable selection is that τ and the variable in $\boldsymbol{X}$ can be chosen based on AIC or C_p. We thus extend the results given in the previous subsections to the criteria based on the ridge regression estimator $\widehat{\boldsymbol{\beta}}_\tau$ instead of $\widehat{\boldsymbol{\beta}}$, which we call here the double ridge criteria.

Let us define the Akaike information based on the double ridge-type estimators by

$$AI_{\lambda,\tau} = -2E_{\boldsymbol{Y}}^*\left[E_{\boldsymbol{Z}}^*[\log f(\boldsymbol{Z}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau(\boldsymbol{Y}), \widehat{\boldsymbol{\Sigma}}_\lambda(\boldsymbol{Y}))]\right].$$

The Akaike information criterion is an asymptotically unbiased estimator of $AI_{\lambda,\tau}$ based on $-2\log f(\boldsymbol{Y}; \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau, \widehat{\boldsymbol{\Sigma}}_\lambda)$, where the bias is given by

$$\Delta_{\lambda,\tau} = \Delta_{\lambda,\tau}(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*, \widehat{\boldsymbol{\beta}}_\tau, \widehat{\boldsymbol{\Sigma}}_\lambda) = AI_{\lambda,\tau} - E_{\boldsymbol{Y}}^*[-2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau, \widehat{\boldsymbol{\Sigma}}_\lambda)]. \tag{22}$$

Theorem 2.3 *Assume the conditions* (7) *and* (13). *Then,* $\Delta_{\lambda,\tau}$ *given in* (22) *is approximated as*

$$\begin{aligned}\Delta_{\lambda,\tau} =& \frac{np\{p+1+k+(1-\tau^2)\rho_\tau\}}{n-k-p-1} \\ &+ \frac{c_n(n-k)}{p(n-k-p-1)}\left\{\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k)}{(n-p)^2} - 1\right\}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}),\end{aligned} \tag{23}$$

where $\rho_\tau = \mathrm{tr}\,[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}]^2$. *The double ridge Akaike information criterion is given by*

$$\begin{aligned}AIC_{\lambda,\tau} =& np\log 2\pi + n\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)] \\ &+ \frac{np\{p+1+k+(1-\tau^2)\rho_\tau\}}{n-k-p-1} + \left\{\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k)}{(n-p)^2} - 1\right\}\hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}],\end{aligned} \tag{24}$$

When τ takes a value in the range of $[0, \tau_0]$ for a fixed τ_0, the optimal ridge parameter τ and the optimal variables can be simultaneously and numerically selected so as to minimize the double ridge criterion $AIC_{\lambda,\tau}$.

The C_p statistic can be similarly extended to the case of the double ridge criterion. Since the prediction error based on the ridge estimator $\widehat{\boldsymbol{\beta}}_\tau$ is written as

$$\begin{aligned} R_{PE}(\boldsymbol{\beta},\boldsymbol{\Sigma};\widehat{\boldsymbol{\beta}}_\tau) =& E^*_{\boldsymbol{Y}}[E^*_{\boldsymbol{Z}}[\mathrm{tr}\,[(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)']]] \\ =& np + PE_\tau, \end{aligned}$$

where $PE_\tau = E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]]$. When PE_τ is estimated based on the statistic $\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]$, the bias is $\Delta_{PE,\tau} = PE_\tau - E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]]$. To evaluate $\Delta_{PE,\tau}$, we assume the condition

$$\lim_{p\to\infty} \boldsymbol{\beta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'/p < \infty. \tag{25}$$

Theorem 2.4 *Assume the conditions* (7), (13) *and* (25). *Then,* $\Delta_{PE,\tau}$ *is evaluated as*

$$\begin{aligned} \Delta_{PE,\tau} =& p\rho_\tau - \frac{np(n-k-p-1+\tau^2\rho_\tau)}{n-K-p-1} \\ &+ \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(1). \end{aligned} \tag{26}$$

Let us define $C_{\lambda,\tau}$ *by*

$$\begin{aligned} C_{\lambda,\tau} =& \mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)] \\ &- \frac{np(n-k-p-1+\tau^2\rho_\tau)}{n-K-p-1} + p\rho_\tau + \widetilde{\lambda}\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}^{-1}]. \end{aligned} \tag{27}$$

Then, $C_{\lambda,\tau}$ *is an asymptotically unbiased estimator of* PE_τ*, namely,* $E[C_{\lambda,\tau}] = PE_\tau + O(1)$.

Similarly to $AIC_{\lambda,\tau}$, the optimal ridge parameter τ and the optimal variables can be simultaneously and numerically selected so as to minimize $C_{\lambda,\tau}$ for $0 \le \tau \le \tau_0$.

3 Proofs of the main results

3.1 Proofs of Theorems 2.1 and 2.3.

Since Theorem 2.1 is a spacial case of Theorem 2.3, we here prove Theorem 2.3. For large p we consider the ridge-type estimator of the $p\times p$ covariance matrix $\boldsymbol{\Sigma}$ given in (9). In order to obtain AIC_λ defined in (17), we need to first evaluate Δ_λ under the true model and, if it depends on some of the unknown parameters, we may need to provide an estimated value of Δ_λ. To prove Theorem 2.1, we note that $-2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau,\widehat{\boldsymbol{\Sigma}}_\lambda)$ is given by

$$-2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau,\widehat{\boldsymbol{\Sigma}}_\lambda) = np\log(2\pi) + n\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]$$

and $AI_{\lambda,\tau}$ is written as

$$AI_{\lambda,\tau} = E^*_{\boldsymbol{Y}}[E^*_{\boldsymbol{Z}}[np\log(2\pi) + n\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]]].$$

Taking the expectation with respect to $\boldsymbol{Z}$ yields that

$$\begin{aligned}&E^*_{\boldsymbol{Z}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Z}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]]\\ &=n\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]+\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})],\end{aligned}$$

so that the bias is written as

$$\begin{aligned}\Delta_{\lambda,\tau}=&AI_{\lambda,\tau}-E^*_{\boldsymbol{Y}}[-2\log f(\boldsymbol{Y}|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau,\widehat{\boldsymbol{\Sigma}}_\lambda)]\\ =&E^*_{\boldsymbol{Y}}[n\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]+\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]\\ &-\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]].\end{aligned}\tag{28}$$

It is here observed that

$$\begin{aligned}&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]]\\ &=E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]]-2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]]\\ &\quad+E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]],\end{aligned}\tag{29}$$

where $\boldsymbol{S}=(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$. Note that $\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}}=-\tau(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=-\tau(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\widehat{\boldsymbol{\beta}}$ and that $\boldsymbol{Y}-X\widehat{\boldsymbol{\beta}}$ is independent of $\boldsymbol{\beta}$. Since $\boldsymbol{S}$ is invariant under the sign change of $\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, it can be seen that

$$\begin{aligned}&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]]\\ &=\mathrm{tr}\,[E^*_{\boldsymbol{Y}}[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)']\,E^*_{\boldsymbol{Y}}[\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]]=0.\end{aligned}\tag{30}$$

Also, note that $\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}\sim\mathcal{N}(-\tau(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta},(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1},\boldsymbol{\Sigma})$ and $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{0},(\boldsymbol{X}'\boldsymbol{X})^{-1},\boldsymbol{\Sigma})$. Then,

$$\begin{aligned}&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta})]]\\ &\quad=E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]\rho_\tau+\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]],\end{aligned}\tag{31}$$

$$\begin{aligned}&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]]\\ &\quad=\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\widehat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\widehat{\boldsymbol{\beta}}]]\\ &\quad=\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]\rho_\tau+\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]].\end{aligned}\tag{32}$$

Combining these observations, from (28) we can express the bias as

$$\Delta_{\lambda,\tau}=E^*_{\boldsymbol{Y}}[\{n+(1-\tau^2)\rho_\tau\}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]-\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]].\tag{33}$$

From the equation given in problem 1.6 (i) of Srivastava and Khatri (1979, pp33), it is noted that

$$(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}=\boldsymbol{I}-\hat{\lambda}\boldsymbol{S}^{-1}+\hat{\lambda}^2\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}.\tag{34}$$

Hence, Δ_λ is rewritten as

$$\begin{aligned}\Delta_{\lambda,\tau} =& n\{n+(1-\tau^2)\rho_\tau\}E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\{\boldsymbol{I}-\hat{\lambda}\boldsymbol{S}^{-1}+\hat{\lambda}^2\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}\}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^*]]\\ &-nE^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\boldsymbol{I}-\hat{\lambda}\boldsymbol{S}^{-1}+\hat{\lambda}^2\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}]]\\ =& nE^*_{\boldsymbol{Y}}\big[\{n+(1-\tau^2)\rho_\tau\}\mathrm{tr}\,[\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^*]-p\big]-nE^*_{\boldsymbol{Y}}\big[\{n+(1-\tau^2)\rho_\tau\}\hat{\lambda}\mathrm{tr}\,[\boldsymbol{S}^{-2}\boldsymbol{\Sigma}^*]-\hat{\lambda}\mathrm{tr}\,[\boldsymbol{S}^{-1}]\big]\\ &+nE^*_{\boldsymbol{Y}}\big[\{n+(1-\tau^2)\rho_\tau\}\hat{\lambda}^2\mathrm{tr}\,[\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^*]-\hat{\lambda}^2\mathrm{tr}\,[\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}]\big]\\ =& I_1-I_2+I_3. \quad \text{(say)}\end{aligned}$$

We first evaluate I_3. Since $p/n\to c$, $0<c<1$, it follows from Bai and Yin (1993) that $\boldsymbol{S}/n$ is almost surely bounded by a constant matrix. Also, we have assumed that $\lim_{p\to\infty}\mathrm{tr}\,[\boldsymbol{\Sigma}]/p<\infty$. Hence, it can be seen that

$$\begin{aligned}&n\{n+(1-\tau^2)\rho_\tau\}\mathrm{tr}\,[\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^*]\\ &\qquad\le n\{n+(1-\tau^2)\rho_\tau\}\mathrm{tr}\,[\boldsymbol{S}^{-3}\boldsymbol{\Sigma}^*]=\frac{n+k}{n}\frac{p}{n}\frac{\mathrm{tr}\,[(\boldsymbol{S}/n)^{-3}\boldsymbol{\Sigma}^*]}{p}=O_p(1),\end{aligned}$$

and

$$n\mathrm{tr}\,[\boldsymbol{S}^{-2}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}]\le n\mathrm{tr}\,[\boldsymbol{S}^{-2}]=\frac{p}{n}\frac{\mathrm{tr}\,[(\boldsymbol{S}/n)^{-2}]}{p}=O_p(1). \tag{35}$$

Also, note that $\hat{\lambda}=c_n\mathrm{tr}\,[\boldsymbol{S}]/(np)=c_n\mathrm{tr}\,[\boldsymbol{S}/n]/p=O_p(n^{-\delta})$ since $c_n=O(n^{-\delta})$, $\delta\ge 0$. These evaluations mean that $I_3=O(n^{-2\delta})$. Since $E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^*]]=p/(n-k-p-1)$, it is easy to see that

$$I_1=\frac{np(p+1+2k)}{n-k-p-1},$$

which is of order $O(n^2)$. To estimate I_2, we can express I_2 as

$$I_2=\frac{c_n}{p}E^*_{\boldsymbol{Y}}\big[\{n+(1-\tau^2)\rho_\tau\}\mathrm{tr}\,[\boldsymbol{S}]\mathrm{tr}\,[\boldsymbol{S}^{-2}\boldsymbol{\Sigma}^*]-\mathrm{tr}\,[\boldsymbol{S}]\mathrm{tr}\,[\boldsymbol{S}^{-1}]\big].$$

Thus, from Lemmas A.1 and A.2, it follows that

$$\begin{aligned}I_2 =& \frac{c_n}{p(n-k-p-1)}\Big\{\Big[\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k-1)(n-k+1)}{(n-k-p+1)(n-k-p-3)}-(n-k)\Big]\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]\\ &+2p-\frac{\{n+(1-\tau^2)\rho_\tau\}p}{n-k-p-3}\Big[\frac{(n-k)^2-1}{n-k-p+1}-\frac{(n-k)^2-5(n-k)+2p+2}{n-k-p}\Big]\Big\}.\end{aligned}$$

Since $\{n+(1-\tau^2)\rho_\tau\}(n-k-1)(n-k+1)/\{(n-k-p+1)(n-k-p-3)\}=\{n+(1-\tau^2)\rho_\tau\}(n-k)(n-k)/(n-p)^2+O(1)$, I_2 can be approximated as

$$I_2=\frac{c_n(n-k)}{p(n-k-p-1)}\Big\{\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k)}{(n-p)^2}-1\Big\}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]+O(n^{-\delta}).$$

Combining the above evaluations, we get

$$\begin{aligned}\Delta_{\lambda,\tau} =& \frac{np\{p+1+k+(1-\tau^2)\rho_\tau\}}{n-k-p-1}\\ &+\frac{c_n(n-k)}{p(n-k-p-1)}\Big\{\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k)}{(n-p)^2}-1\Big\}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]+O(n^{-\delta}).\end{aligned} \tag{36}$$

Since $\Delta_{\lambda,\tau}$ involves unknown quantity, we need to estimate $\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]$. Using (34) and (35), we can observe that

$$\begin{aligned} E[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]] =& np - \frac{c_n}{p}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\,[\boldsymbol{S}]\mathrm{tr}\,[\boldsymbol{S}^{-1}]\right] + O(n^{-2\delta}) \\ =& np - \frac{c_n(n-k)}{p(n-k-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}), \end{aligned}$$

where Lemma A.1 is used to show the second equality. Since $np - \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}] = \hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}]$, it follows that

$$E_{\boldsymbol{Y}}^*\left[\hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}]\right] = \frac{c_n(n-k)}{p(n-k-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}), \tag{37}$$

which is substituted into (36) to get the expression

$$\begin{aligned} \Delta_{\lambda,\tau} =& \frac{np\{p+1+k+(1-\tau^2)\rho_\tau\}}{n-k-p-1} \\ &+ \left\{\frac{\{n+(1-\tau^2)\rho_\tau\}(n-k)}{(n-p)^2} - 1\right\}E_{\boldsymbol{Y}}^*\left[\hat{\lambda}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}]\right] + O(n^{-\delta}). \end{aligned} \tag{38}$$

Hence, the approximated value of AIC_λ stated in Theorem 2.1 is obtained. ■

3.2 Proofs of Theorems 2.2 and 2.4

Since Theorem 2.2 is a special case of Theorem 2.4, we here prove Theorem 2.4. Letting $PE_\tau = E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta})]$, we can see that

$$PE_\tau = p\rho_\tau + \tau^2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}.$$

We shall obtain an asymptotic unbised estimator of PE_τ based on $\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]$. The expectation can be evaluated as

$$\begin{aligned} &E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]] \\ &=E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]] - 2E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]] \\ &\quad + E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]]. \end{aligned}$$

It is noted that $\boldsymbol{Y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are mutually independent for $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{Y}$. Since $\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}) + (\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$, the product term can be evaluated as

$$\begin{aligned} &E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]] \\ &=E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\{(\boldsymbol{Y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}) + (\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]] \\ &=\mathrm{tr}\,[E_{\boldsymbol{Y}}^*[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}})']E_{\boldsymbol{Y}}^*[\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]] \\ &=0, \end{aligned}$$

where the same arguments as in (30) have been used to show the last equality. Similar to (32), it can be observed that

$$\begin{aligned}&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\tau-\widehat{\boldsymbol{\beta}})]]\\ &=\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]\rho_\tau+\tau^2E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]].\end{aligned}$$

Hence, the bias can be evaluated as

$$\begin{aligned}\Delta_{PE,\tau}=&PE_\tau-E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\tau)]]\\ =&p\rho_\tau+\tau^2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]\\ &-E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]-\tau^2\rho_\tau E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]\\ &-\tau^2\mathrm{tr}\,[E^*_{\boldsymbol{Y}}[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}]\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}].\end{aligned}\tag{39}$$

Since $\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'/p$ is bounded for large p from the condition (25), it is seen that

$$\begin{aligned}&\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]\\ &\quad\le\mathrm{tr}\,[\boldsymbol{\beta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}]=O(1).\end{aligned}$$

Similarly,

$$\begin{aligned}&\mathrm{tr}\,[E^*_{\boldsymbol{Y}}[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}]\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}\boldsymbol{\beta}]\\ &\quad\le\mathrm{tr}\,[\boldsymbol{\beta}E^*_{\boldsymbol{Y}}[n\widetilde{\boldsymbol{S}}^{-1}]\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\tau\boldsymbol{I})^{-1}]=O(1).\end{aligned}$$

Thus,

$$\Delta_{PE,\tau}=p\rho_\tau-E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]-\tau^2\rho_\tau E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]+O(1).\tag{40}$$

We first evaluate $E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]$. Noting that $\boldsymbol{S}=\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}$ and $\widetilde{\boldsymbol{S}}=\boldsymbol{Y}'(\boldsymbol{I}-\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}')\boldsymbol{Y}$, there exists a $p\times(K-k)$ random matrix $\boldsymbol{U}$ such that $\boldsymbol{S}=\widetilde{\boldsymbol{S}}+\boldsymbol{U}\boldsymbol{U}'$ and $\boldsymbol{U}$ is distributed as $\boldsymbol{U}'\sim\mathcal{N}_{K-k}(\boldsymbol{0},\boldsymbol{I}_N,\boldsymbol{\Sigma}^*)$, independent of $\widetilde{\boldsymbol{S}}$. Since $\widetilde{\boldsymbol{\Sigma}}_\lambda$ is a function of $\widetilde{\boldsymbol{S}}$, it is seen that

$$\begin{aligned}E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]]=&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}(\widetilde{\boldsymbol{S}}+\boldsymbol{U}\boldsymbol{U}')]]\\ =&E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\widetilde{\boldsymbol{S}}]+(K-k)\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]\\ =&E^*_{\boldsymbol{Y}}[n\mathrm{tr}\,[(\widetilde{\boldsymbol{S}}+\widetilde{\lambda}\boldsymbol{I})^{-1}\widetilde{\boldsymbol{S}}]+n(K-k)\mathrm{tr}\,[(\widetilde{\boldsymbol{S}}+\widetilde{\lambda}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}]]\\ =&E^*_{\boldsymbol{Y}}[n\mathrm{tr}\,[(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}]+n(K-k)\mathrm{tr}\,[(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{\Sigma}]].\end{aligned}$$

From (34) and the fact that $(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}=\boldsymbol{I}-\hat{\lambda}\boldsymbol{S}^{-1}(\boldsymbol{I}+\hat{\lambda}\boldsymbol{S}^{-1})^{-1}$, it follows that

$$\begin{aligned}E^*_{\boldsymbol{Y}}[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]]=&E^*_{\boldsymbol{Y}}[n\mathrm{tr}\,[\boldsymbol{I}-\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1}+\widetilde{\lambda}^2\widetilde{\boldsymbol{\Sigma}}^{-2}(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}]\\ &+n(K-k)\mathrm{tr}\,[\{\boldsymbol{I}-\widetilde{\lambda}\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}\}\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{\Sigma}]].\end{aligned}$$

Using the arguments as in (35), we can see that

$$\begin{aligned}E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}]] =& np - E_{\boldsymbol{Y}}^*[n\widetilde{\lambda}\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}] + n(K-k)\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{\Sigma}]] + O(n^{-2\delta})\\ =& np - \frac{c_n}{p}E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{S}}]\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}]] + n(K-k)\frac{p}{n-K-p-1} + O(n^{-2\delta})\\ =& \frac{np(n-k-p-1)}{n-K-p-1} - \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}). \quad (41)\end{aligned}$$

where Lemma A.1 is used at the third equality.

Using a similar argument, we next evaluate $E_{\boldsymbol{Y}}^*[\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}]]$. Since $(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1} = \boldsymbol{I} - \widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}$, it can be seen that

$$\begin{aligned}\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}] =& n\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{\Sigma}] - n\widetilde{\lambda}\mathrm{tr}\,[\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{\Sigma}\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{I}+\widetilde{\lambda}\widetilde{\boldsymbol{S}}^{-1})^{-1}]\\ =& \frac{np}{n-K-p-1} + O(n^{-\delta}).\end{aligned}$$

Hence from (40) and (41), we can see that

$$\begin{aligned}\Delta_{PE,\tau} =& p\rho_\tau - \frac{np(n-k-p-1)}{n-K-p-1} + \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}]\\ &- \tau^2\rho_\tau\frac{np}{n-K-p-1} + O(1)\\ =& p\rho_\tau - \frac{np(n-k-p-1+\tau^2\rho_\tau)}{n-K-p-1} + \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(1).\end{aligned} \quad (42)$$

From (15), it follows that

$$E_{\boldsymbol{Y}}^*\left[\widetilde{\lambda}\mathrm{tr}\,[\widetilde{\boldsymbol{\Sigma}}_\lambda^{-1}]\right] = \frac{c_n(n-K)}{p(n-K-p-1)}\mathrm{tr}\,[\boldsymbol{\Sigma}^*]\mathrm{tr}\,[\boldsymbol{\Sigma}^{*-1}] + O(n^{-\delta}),$$

where K is the rank of $\widetilde{\boldsymbol{X}}$. Hence, we get the $C_{p,\tau}$ type criterion given in (27). ■

4 Simulation and empirical studies

4.1 Simulation experiments

We now investigate the numerical performances of the ridge-type and double ridge-type AICs and C_p statistics derived in Section 2 through simulation and compare them in terms of the frequencies of selecting the true model.

As the true model, we consider the model that $\boldsymbol{Y} \sim \mathcal{N}_{n,p}(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}^*, \boldsymbol{I}_n, \boldsymbol{\Sigma}^*)$, where $\widetilde{\boldsymbol{X}} = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(K)})$ is a matrix of regressor variables in a full model given in (1),

$$\boldsymbol{\beta}^* = ((\boldsymbol{\beta}_1^*)', \dots, (\boldsymbol{\beta}_{k^*}^*)', \boldsymbol{0}, \dots, \boldsymbol{0})', \ \beta_{ij}^* = 2(-1)^i(u_{ij}+i), \ i = 1, \dots, k^*, j = 1, \dots, p,$$

for random variable u_{ij} from a uniform distribution on the interval $[0,1]$, and

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix} \begin{pmatrix} \rho^{|1-1|^{\frac{1}{7}}} & \rho^{|1-2|^{\frac{1}{7}}} & \dots & \rho^{|1-p|^{\frac{1}{7}}} \\ \rho^{|2-1|^{\frac{1}{7}}} & \rho^{|2-2|^{\frac{1}{7}}} & \dots & \rho^{|2-p|^{\frac{1}{7}}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{|p-1|^{\frac{1}{7}}} & \rho^{|p-2|^{\frac{1}{7}}} & \dots & \rho^{|p-p|^{\frac{1}{7}}} \end{pmatrix} \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix}.$$

for a constant ρ on the interval $(-1,1)$ and $\sigma_i = 2 + (p-i+1)/p$.

The simulation experiments have been carried out for $n = 76$, $K = 7$, $\rho = 0.7$, $p = 10, 20, 30, 40, 50, 60$. For the $n \times K$ matrix $\widetilde{\boldsymbol{X}}$ of the regressor variables in the full model (1), the row vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ for $\boldsymbol{X}' = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ are generated as mutually independent random variables distributed as $\mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma}_x)$ where $\boldsymbol{\Sigma}_x = (1-\rho_x)\boldsymbol{I}_K + \rho_x \boldsymbol{J}_K$ for $\rho_x = 0.7$, where $\boldsymbol{J}_K = \boldsymbol{j}_K \boldsymbol{j}_K'$ for $\boldsymbol{j}_K = (1, \dots, 1)'$, a K-vector of ones. The above true model is expressed as

$$M_{k^*} \qquad \boldsymbol{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

where $1 \le k^* \le 7$, $\boldsymbol{\beta}^* = ((\boldsymbol{\beta}_1)', \dots, (\boldsymbol{\beta}_{k^*})', \boldsymbol{0}, \dots, \boldsymbol{0})'$, and $\boldsymbol{\epsilon}$ is a random variable having $\boldsymbol{\epsilon} \sim \mathcal{N}_{n,p}(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}^*, \boldsymbol{I}_n, \boldsymbol{\Sigma}^*)$. Let us write the model using the first m regressor variables $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m$ by M_m. Then, the full model is M_7 and the true model is M_{k^*}. As candidate models, we consider the nested subsets $M_1, \dots, M_7$, namely,

$$M_m \qquad \boldsymbol{y} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m)} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}^{(m)} = (\beta_1, \dots, \beta_m, 0, \dots, 0)'$.

In the simulation experiments, 20 observations of the regressor variables $\widetilde{\boldsymbol{X}}$ are generated, and for each observation of $\widetilde{\boldsymbol{X}}$, 50 observations of the response variable $\boldsymbol{y}$ are generated from the true model M_{k^*} for $k^* = 4$. Thus, we have $20 \times 50 (= 1,000)$ total data sets. For each data set, we calculate the values of AIC_0, AIC_λ, C_0 and C_λ with $c_n = n/p$ given in (17), (16), (21) and (20), respectively, for the seven candidate models $M_1, \dots, M_7$, and we select the models minimizing the values of the selection procedures. For each criterion and each candidate model M_m, the number of selecting the model M_m is counted for 1,000 data set. We thus obtain the frequencies of the model M_m selected by the criteria by dividing the number by 1,000.

Table 1 reports the frequencies in the cases of $p = 10, 20, 30, 40, 50, 60$ under the true model M_4, namely $k^* = 4$. From this table, it is seen that all the criteria perform well for small p in the sense of selecting the true model. For larger p, AIC_0 and C_0 based on the MLE of $\boldsymbol{\Sigma}$ perform much worse, while AIC_λ and C_λ based on the ridge-type estimator of $\boldsymbol{\Sigma}$ perform quite well. Table 2 handles the extreme cases of $p = 65$, namely $\nu_K = n - K - p - 3 = 1$ for $k^* = 2, 3, 4, 5, 6, 7$. For the extreme cases reported in this table, AIC_λ and C_λ work still well.

It is interesting to investigate how the double ridge criteria $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ work in multicolinearity cases, where $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ are given in (24) and (27). To clarify the difference between the ridge-type and the double ridge-type criteria, we consider the extreme case of $n = 22$, $K = 7$, $p = 10$ and $\nu_K = n - K - p - 3 = 2$. For the $n \times K$ matrix $\widetilde{\boldsymbol{X}} = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(7)})$ of the regressor variables in the full model (1), it is supposed that $\boldsymbol{x}_{(3)}$, $\boldsymbol{x}_{(5)}$ and $\boldsymbol{x}_{(7)}$ are generated as

Table 1: Frequencies selected by the four criteria AIC_0, AIC_λ, C_0 and C_λ in 1,000 replications for $n = 76$, $K = 7$, $p = 10, 20, 30, 40, 50, 60$ and $\nu_K = n - K - p - 3$ under the true model M_4, namely $k^* = 4$

M_k	AIC_0	AIC_λ	C_0	C_λ	AIC_0	AIC_λ	C_0	C_λ
	$p = 10$, $\nu_K = 56$				$p = 20$, $\nu_K = 46$			
M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	99.7	99.6	90.3	99.0	100.0	100.0	91.6	99.2
M_5	0.3	0.4	7.2	0.9	0.0	0.0	6.7	0.7
M_6	0.0	0.0	2.0	0.1	0.0	0.0	1.4	0.0
M_7	0.0	0.0	0.5	0.0	0.0	0.0	0.3	0.1
	$p = 30$, $\nu_K = 36$				$p = 40$, $\nu_K = 26$			
M_1	1.8	0.0	0.0	0.0	100.0	0.0	0.0	0.0
M_2	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	97.6	100.0	89.4	99.3	0.0	100.0	83.3	99.2
M_5	0.0	0.0	6.9	0.7	0.0	0.0	10.8	0.8
M_6	0.0	0.0	2.5	0.0	0.0	0.0	3.4	0.0
M_7	0.0	0.0	1.2	0.0	0.0	0.0	2.5	0.0
	$p = 50$, $\nu_K = 16$				$p = 60$, $\nu_K = 6$			
M_1	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	0.0	100.0	69.3	99.0	0.0	100.0	53.1	100.0
M_5	0.0	0.0	14.4	1.0	0.0	0.0	17.9	0.0
M_6	0.0	0.0	8.3	0.0	0.0	0.0	13.1	0.0
M_7	0.0	0.0	8.0	0.0	0.0	0.0	15.9	0.0

Table 2: Frequencies selected by the four criteria AIC_0, AIC_λ, C_0 and C_λ in 1,000 replications for the extreme case of $n = 76$, $K = 7$, $p = 65$, namely $\nu_K = n - K - p - 3 = 1$

M_k	AIC_0	AIC_λ	C_0	C_λ	AIC_0	AIC_λ	C_0	C_λ
		$k^* = 2$				$k^* = 3$		
M_1	100.0	0.0	0.0	0.7	100.0	0.0	0.0	0.2
M_2	0.0	100.0	50.7	99.3	0.0	0.0	0.0	0.7
M_3	0.0	0.0	9.7	0.0	0.0	100.0	52.0	99.1
M_4	0.0	0.0	8.3	0.0	0.0	0.0	14.7	0.0
M_5	0.3	0.1	8.1	0.0	0.0	0.0	7.9	0.0
M_6	0.0	0.0	8.8	0.0	0.0	0.0	9.3	0.0
M_7	0.0	0.0	14.4	0.0	0.0	0.0	16.1	0.0
		$k^* = 4$				$k^* = 5$		
M_1	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
M_4	0.0	100.0	54.8	99.8	0.0	0.0	0.0	0.0
M_5	0.0	0.0	15.2	0.0	0.0	100.0	63.3	100.0
M_6	0.0	0.0	14.3	0.0	0.0	0.0	17.3	0.0
M_7	0.0	0.0	15.7	0.0	0.0	0.0	19.4	0.0
		$k^* = 6$				$k^* = 7$		
M_1	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_6	0.0	100.0	71.6	100.0	0.0	0.0	0.0	0.0
M_7	0.0	0.0	28.4	0.0	0.0	100.0	100.0	100.0

follows:

$$\begin{aligned}
\boldsymbol{x}_{(3)} =& 0.3\boldsymbol{x}_{(1)} + 0.7\boldsymbol{x}_{(2)} + \varepsilon Z_1,\\
\boldsymbol{x}_{(5)} =& 0.5\boldsymbol{x}_{(3)} + 0.5\boldsymbol{x}_{(4)} + \varepsilon Z_2,\\
\boldsymbol{x}_{(7)} =& 0.7\boldsymbol{x}_{(5)} + 0.3\boldsymbol{x}_{(6)} + \varepsilon Z_3,
\end{aligned}$$

where ε is a positive constant and Z_1, Z_2 and Z_3 are mutually independently distributed as a standard normal distribution. For smaller ε, $\widetilde{\boldsymbol{X}}$ is closer to the multicolinierity case. In this experiment, we treat the two cases: $\varepsilon = 1$ and $\varepsilon = 0.0001$, which correspond to the non-multicolinearity and the multicolinearity cases, respectively. In the multicolinerity case, the ridge parameter τ in the ridge regression estimator $\widehat{\boldsymbol{\beta}}_\tau$ should be large since $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is instable. Define $L(\widetilde{\boldsymbol{X}})$ by

$$L(\widetilde{\boldsymbol{X}}) = \{|\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/\mathrm{tr}\,[\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/K]| - \log(|\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/\mathrm{tr}\,[\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/K]|) - 1\}/K,$$

which measures the discrepancy between the two matrices $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ and $\mathrm{tr}\,[\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}/K]\boldsymbol{I}_K$. $L(\widetilde{\boldsymbol{X}})$ takes a large value when $\widetilde{\boldsymbol{X}}$ is close to the multicolinearity. For the double ridge AIC, we select the regressor variables and the ridge parameter τ so as to minimize $AIC_{\lambda,\tau}$ for $0 \le \tau \le L(\widetilde{\boldsymbol{X}})/5$. For the double ridge C_p, the regressor variables and the ridge parameter τ are selected to minimize $C_{\lambda,\tau}$ for $0 \le \tau \le L(\widetilde{\boldsymbol{X}})/10$. The frequencies selected by AIC_λ, $AIC_{\lambda,\tau}$, C_λ and $C_{\lambda,\tau}$ in this experiment are reported in Table 3. For $\varepsilon = 1$, the non-multicolinearity case, there are little difference between $(AIC_{\lambda,\tau}, C_{\lambda,\tau})$ and (AIC_λ, C_λ). For $\varepsilon = 0.0001$, which is close to the multicolinearity case, the double ridge criteria $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ are slightly better than AIC_λ and C_λ. When the true model is M_6, we can observe that $AIC_{\lambda,\tau}$ performs well while AIC_λ does not work.

4.2 An application to posted land price data

We here treat the posted land price data along the Keikyu train line which connects the suburbs in Kanagawa prefecture to the Tokyo metropolitan area. Those who live in the suburbs take this line to work or study in Tokyo every weekday. Thus, it is expected that the land price depends on the distance from Tokyo. We use the selection procedures AIC_0, AIC_λ, C_0 and C_λ to search for the covariates which affect the land price.

The posted land price data for fifteen years from 1987 to 2001 are available for 47 sites along the Keikyu train line. Each site is indexed by i, namely, $i = 1, \dots, n$ for $n = 47$. The values which are transformed by logarithm from the posted land price (Yen) per m^2 of the i-th site for the fifteen years are described by $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})$ for $T = 15$. For each y_{it}, we consider the following five explanatory variables: T_{1i} is the time to take from the nearby station to the Tokyo station around 8:30 in the morning, T_{2i} is the time to take on foot from the site i to the nearby station and FAR_i and ACR_i denote, respectively, the floor-area ratio and the acreage of the site i. Also, TKY_i is the dummy variable indicating whether the site i is in Tokyo or in Kanagawa prefecture, namely $TKY_i = 0$ if the site i is in Tokyo, otherwise $TKY_i = 1$. As the full model, we consider the mixed linear model

$$y_{it} = \beta_{0t} + T_{1i}\beta_{1t} + (T_{1i}^2)\beta_{2t} + T_{2i}\beta_{3t} + FAR_i\beta_{4t} + TKY_i\beta_{5t} + ACR_i\beta_{6t} + e_{it}.$$

For simplicity, the regressor variables 1, T_{1i}, T_{1i}^2, T_{2i}, FAR_i, TKY_i and ACR_i are denoted by x_{0i}, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i} and x_{6i}. Let $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_6) = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(N)})'$, which

Table 3: Frequencies selected by the four criteria AIC_λ, $AIC_{\lambda,\tau}$, C_λ and $C_{\lambda,\tau}$ in 1,000 replications for $n = 22$, $K = 7$, $p = 10$ and $\nu_K = n - K - p - 3 = 2$ in the case of multicolinearity under the true models M_2, M_4, M_6

	$\varepsilon = 1$, non-multicolinearity				$\varepsilon = 0.0001$, multicolinearity			
M_k	AIC_λ	$AIC_{\lambda,\tau}$	C_λ	$C_{\lambda,\tau}$	AIC_λ	$AIC_{\lambda,\tau}$	C_λ	$C_{\lambda,\tau}$
				M_2 : the true model				
M_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_2	100.0	100.0	95.9	95.9	100.0	100.0	95.9	99.8
M_3	0.0	0.0	3.7	3.7	0.0	0.0	3.7	0.0
M_4	0.0	0.0	0.3	0.3	0.0	0.0	0.3	0.2
M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_6	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.0
M_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
				M_4 : the true model				
M_1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	100.0	100.0	97.0	96.9	99.9	100.0	97.0	99.3
M_5	0.0	0.0	2.4	2.5	0.0	0.0	2.4	0.0
M_6	0.0	0.0	0.3	0.3	0.0	0.0	0.3	0.7
M_7	0.0	0.0	0.3	0.3	0.0	0.0	0.3	0.0
				M_6 : the true model				
M_1	4.8	4.2	0.0	0.0	83.5	0.0	0.0	0.0
M_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_6	95.2	95.8	97.0	97.0	16.5	100.0	97.0	100.0
M_7	0.0	0.0	3.0	3.0	0.0	0.0	3.0	0.0

is an $n \times 7$ matrix, for $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jN})'$ and $\boldsymbol{x}'_{(i)} = (x_{0i}, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})$. Also let $\boldsymbol{Y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n)'$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$, and $\boldsymbol{E}$ is similarly defined. Then, the model is expressed as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(T)})$, which is a $7 \times T$ matrix, for $\boldsymbol{\beta}_{(t)} = (\beta_{0t}, \ldots, \beta_{6t})'$.

Table 4 reports values of AIC_0, AIC_λ, C_0 and C_λ for several candidate models, where the regressor variable which minimizes AIC_λ is added to the model based on the forward selection rule. Among these candidate models, the minimum value of AIC_λ is -670 and attained by the model with the regressor variables $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, while AIC_0 and C_λ select $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_5\}$ or $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_4, \boldsymbol{x}_5\}$. It is also observed that C_0 selects the full model, which shows that C_0 does not work well for this example. According to these observations based on AIC_0, AIC_λ and C_λ, we can recommend the model given by

$$y_{it} = \beta_{0t} + T_{1i}\beta_{1t} + FAR_i\beta_{4t} + TKY_i\beta_{5t} + e_{it}.$$

Although values of $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ are not reported in Table 4, it is noted that their values are very close to those of AIC_λ and C_λ, respectively.

We here investigate whether the selected model is endorsed by a testing procedure. The general linear hypothesis is expressed as a testing of hypothesis

$$H : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0} \quad \text{vs} \quad A : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{0}$$

where $\boldsymbol{C}$ is a known $m \times 7$ matrix of rank $m \leq 7$. The error sum of squares and products is given by the matrix

$$\boldsymbol{V} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y},$$

and the sum of squares and products due to regression under the hypotheses H is

$$\boldsymbol{W} = \widehat{\boldsymbol{\beta}}'\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}\boldsymbol{C}\widehat{\boldsymbol{\beta}}.$$

To test the hypothesis H we need to compare these matrices. The likelihood ratio test rejects the hypothesis H if

$$\frac{|\boldsymbol{V}|}{|\boldsymbol{V} + \boldsymbol{W}|} = U_{p,m,f} \leq U_{p,m,f,\alpha}$$

where $f = n - 7$ and $U_{p,m,f,\alpha}$ is the upper $100\alpha\%$ point of the distribution of $U_{p,m,f}$. The asymptotic approximation for $U_{p,m,f}$ is given by

$$\begin{aligned} &P[-\{f - (p - m + 1)/2\} \log U_{p,m,f} \geq z] \\ &\qquad = P[\chi^2_{pm} \geq z] + f^{-2}\gamma_2 \left\{P[\chi^2_{pm+4} \geq z] - P[\chi^2_{pm} \geq z]\right\} \end{aligned} \tag{43}$$

where $\gamma_2 = pm(p^2 + p - 5)/48$. See Srivastava (2002, p.282).

Let $\boldsymbol{\beta}$ be decomposed into $\boldsymbol{\beta} = (\boldsymbol{\beta}'_0, \boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_k)'$. When the null hypothesis $H : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_4 = \boldsymbol{\beta}_5 = \boldsymbol{0}$ is tested, the P-value given in (43) is 0.000, and the hypothesis is rejected strongly. When each hypothesis $H_i : \boldsymbol{\beta}_i = 0$ is tested for $i = 0, 1, \ldots, 6$, the P-values are given by $P_0 = P_1 = P_2 = P_4 = P_5 = 0.000$, $P_3 = 0.022$ and $P_6 = 0.008$, where P_i is the P-value given in (43) for testing H_i. When the P-values can be also obtained numerically based on simulation experiments, those values for testing the hypotheses H_3 and H_6 are 0.027 and 0.007. Thus, it may be plausible that the variables $\boldsymbol{x}_3$ and $\boldsymbol{x}_6$, namely T_{2i} and ACR_i are deleted from the regressor variables.

Table 4: Selection of regressor variables in the posted land price data

k	x_i	AIC_0	AIC_λ	C_0	C_λ
1	x_0	-3040	105	2051	286
2	x_0, x_1	-3106	-490	1279	68
3	x_0, x_1, x_2	-3108	-489	906	54
3	x_0, x_1, x_3	-3070	-438	1271	109
3	x_0, x_1, x_4	-3104	-619	1158	74
3	x_0, x_1, x_5	-3174	-521	353	-14
3	x_0, x_1, x_6	-3081	-445	1215	95
4	x_0, x_1, x_4, x_2	-3101	-594	777	58
4	x_0, x_1, x_4, x_3	-3064	-578	1150	113
4	x_0, x_1, x_4, x_5	-3167	-670	232	-9
4	x_0, x_1, x_4, x_6	-3067	-560	1136	108
5	x_0, x_1, x_4, x_5, x_2	-3145	-601	160	29
5	x_0, x_1, x_4, x_5, x_3	-3129	-634	211	30
5	x_0, x_1, x_4, x_5, x_6	-3128	-605	215	26
6	x_0, x_1, x_4, x_5, x_3, x_2	-3106	-566	127	68
6	x_0, x_1, x_4, x_5, x_3, x_6	-3085	-565	190	67
7	x_0, x_1, x_4, x_5, x_3, x_2, x_6	-3056	-490	105	105

5 Concluding remarks

The variable selection problem in the multivariate linear regression model is addressed under the asymptotic condition that both n and p tend to infinity subject to $n - k - p - 3 > 0$ and $\lim_{(n,p)\to\infty} p/n = c$ for $0 < c < 1$. In this paper, we have proposed the modified AIC and C_p statistic, denoted by AIC_λ and C_λ, based on the ridge-type estimator of $\mathbf{\Sigma}$ instead of the MLE, and proved their analytical justifications, namely, they are asymptotic unbiased estimators of the quantities related to the prediction errors. We also have extended the modified AIC and C_p statistic to the double ridge-type criteria which use the ridge regression estimator of $\boldsymbol{\beta}$ instead of the least squares estimator.

Through simulation studies reported in Tables 1 and 2, it is seen that AIC_0 and C_0 statistic, based on MLE of $\mathbf{\Sigma}$, perform well for small p and large n, as it should be. The performances of AIC_λ and C_λ are, however, equally good and somewhat better. In contrast for large p, the performance of AIC_0 and C_0 are rather poor in comparison to the performance of AIC_λ and C_λ. In the case close to the multicolinearity, the double ridge-type criteria $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ have been shown to work well. Thus we recommend the use of AIC_λ and C_λ, or $AIC_{\lambda,\tau}$ and $C_{\lambda,\tau}$ for all p so long as $n - k - p > 3$.

A Appendix

Lemma A.1 *Let* $\boldsymbol{S} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, m)$. *Then,*

$$E[(\operatorname{tr}\boldsymbol{S})(\operatorname{tr}\boldsymbol{S}^{-1})] = \frac{m}{m-p-1}\operatorname{tr}\boldsymbol{\Sigma}\operatorname{tr}\boldsymbol{\Sigma}^{-1} - \frac{2p}{m-p-1}.$$

Proof. Since $\operatorname{tr}\boldsymbol{S}$ and $\operatorname{tr}\boldsymbol{S}^{-1}$ are invariant under an orthogonal transformation, we may assume that $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements σ_i, $i = 1, \dots, p$, $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_p)$. Thus, with $\boldsymbol{W} \sim \mathcal{W}_p(\boldsymbol{I}, m)$, $\boldsymbol{W} = (w_{ij})$, $\boldsymbol{W}^{-1} = (w^{ij})$, we get

$$\operatorname{tr}\boldsymbol{S} = \operatorname{tr}\boldsymbol{\Sigma W} = \sum_i \sigma_i w_{ii}.$$

Hence,

$$(\operatorname{tr}\boldsymbol{S})(\operatorname{tr}\boldsymbol{S}^{-1}) = (\sum_{i=1}^p \sigma_i w_{ii})(\sum \sigma_i^{-1} w^{ii}) = \sum_{i=1}^p w_{ii}w^{ii} + \sum_{i\neq j} \sigma_i \sigma_j^{-1} w_{ii} w^{jj}.$$

Noting that $E[w_{ii}w^{ii}] = E[w_{pp}w^{pp}]$ for any i, and $E[w_{ii}w^{jj}] = E[w_{11}w^{pp}]$ for any $i \neq j$, we get

$$E[(\operatorname{tr}\boldsymbol{S})(\operatorname{tr}\boldsymbol{S}^{-1})] = pE[w_{pp}w^{pp}] + \sum_{i\neq j} \sigma_i \sigma_j^{-1} E[w_{11}w^{pp}].$$

Consider now the triangular factorization of $\boldsymbol{W} = \boldsymbol{TT}'$, where

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_1 & \boldsymbol{0} \\ \boldsymbol{t}_{12}' & t_{pp} \end{pmatrix}.$$

Then,

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{T}_1\boldsymbol{T}_1' & \boldsymbol{T}_1\boldsymbol{t}_{12} \\ \boldsymbol{t}_{12}'\boldsymbol{T}_1' & t_{pp}^2 + \boldsymbol{t}_{12}'\boldsymbol{t}_{12} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}' & w_{pp} \end{pmatrix}$$

and

$$w^{pp} = (w_{pp} - \boldsymbol{w}_{12}'\boldsymbol{W}_{11}^{-1}\boldsymbol{w}_{12})^{-1} = (t_{pp}^2)^{-1}.$$

Hence,

$$\begin{aligned} E[w_{pp}w^{pp}] =& E\left[\frac{t_{pp}^2 + \boldsymbol{t}_{12}'\boldsymbol{t}_{12}}{t_{pp}^2}\right] = 1 + E\left[\frac{\boldsymbol{t}_{12}'\boldsymbol{t}_{12}}{t_{pp}^2}\right] \\ =& 1 + E[\boldsymbol{t}_{12}'\boldsymbol{t}_{12}]E[t_{pp}^{-2}] = 1 + \frac{p-1}{m-p-1} = \frac{m-2}{m-p-+1}, \end{aligned}$$

since $\boldsymbol{t}_{12} \sim \mathcal{N}_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ is independently distributed of t_{pp}^2, and t_{pp}^2 is distributed as chisquare with $m-p+1$ degrees of freedom, see Srivastava and Khatri (1979, Lemma 3.2.1, pp 74). Similarly,

$$E[w_{11}w^{pp}] = E[t_{11}^2/t_{pp}^2] = E[t_{11}^2]E[1/t_{pp}^2] = \frac{m}{m-p-1}.$$

Hence,

$$
\begin{aligned}
E[(\mathrm{tr}\,\boldsymbol{S})(\mathrm{tr}\,\boldsymbol{S}^{-1})] =& \frac{(m-2)p}{m-p-1} + \frac{m}{m-p-1}\sum_{i\neq j}\sigma_i\sigma_j^{-1} \\
=& \frac{m}{m-p-1}[(\mathrm{tr}\,\boldsymbol{\Sigma})(\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}) - p] + \frac{(m-2)p}{m-p-1} \\
=& \frac{m}{m-p-1}(\mathrm{tr}\,\boldsymbol{\Sigma})(\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}) - \frac{2p}{m-p-1}.
\end{aligned}
$$

■

Lemma A.2 *Let* $\boldsymbol{S} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, m)$. *Then,*

$$
\begin{aligned}
E[(\mathrm{tr}\,\boldsymbol{S})\mathrm{tr}\,(\boldsymbol{\Sigma}\boldsymbol{S}^{-2})] =& \frac{(m-1)(m+1)}{(m-p+1)(m-p-1)(m-p-3)}(\mathrm{tr}\,\boldsymbol{\Sigma})(\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}) \\
& - \frac{p}{(m-p-1)(m-p-3)}\left(\frac{m^2-1}{m-p+1} - \frac{m^2-5m+2p+2}{m-p}\right).
\end{aligned}
$$

Proof. As explained above, we assume that $\boldsymbol{\Sigma} = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_p)$, and $\boldsymbol{W} \sim \mathcal{W}_p(\boldsymbol{I}, m)$,

$$
\begin{aligned}
E[(\mathrm{tr}\,\boldsymbol{S})(\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{S}^{-2})] =& E[(\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{W})(\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{-2})] \\
=& E[\left(\sum_i \sigma_i w_{ii}\right)\left(\sum_i \sigma_i^{-1}(\boldsymbol{W}^{-2})_{ii}\right)] \\
=& E[\sum_{i=1}^p w_{ii}(\boldsymbol{W}^{-2})_{ii} + \sum_{i\neq j}\sigma_i\sigma_j^{-1}w_{ii}(\boldsymbol{W}^{-2})_{jj}] \\
=& pE[w_{pp}(\boldsymbol{W}^{-2})_{pp} + w_{11}(\boldsymbol{W}^{-2})_{pp}\sum_{i\neq j}\sigma_i\sigma_j^{-1}].
\end{aligned}
$$

Note that

$$
(\boldsymbol{W}^{-2})_{pp} = \frac{1}{t_{pp}^4}[1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}],
$$

and $w_{pp} = t_{pp}^2 + \boldsymbol{t}_{12}'\boldsymbol{t}_{12}$. Thus,

$$
\begin{aligned}
E[w_{pp}(\boldsymbol{W}^{-2})_{pp}] =& E[(t_{pp}^2 + \boldsymbol{t}_{12}'\boldsymbol{t}_{12})(1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}/t_{pp}^4] \\
=& E[(t_{pp}^{-2} + \boldsymbol{t}_{12}'\boldsymbol{t}_{12}t_{pp}^{-4})(1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}] \\
=& E\left[(\frac{1}{m-p-1} + \frac{\boldsymbol{t}_{12}'\boldsymbol{t}_{12}}{(m-p-1)(m-p-3)})(1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12})\right] \\
=& \frac{1}{m-p-1}E[1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}] + E[\frac{\boldsymbol{t}_{12}'\boldsymbol{t}_{12}\boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12} + \boldsymbol{t}_{12}'\boldsymbol{t}_{12}}{(m-p-1)(m-p-3)}] \\
=& \frac{1}{m-p-1}\{1 + E[\mathrm{tr}\,(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}]\} + E[\frac{\boldsymbol{t}_{12}'\boldsymbol{t}_{12}\boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12} + \boldsymbol{t}_{12}'\boldsymbol{t}_{12}}{(m-p-1)(m-p-3)}].
\end{aligned}
$$

Note that $\boldsymbol{t}_{12}'\boldsymbol{t}_{12}\boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12} = \mathrm{tr}\,(\boldsymbol{t}_{12}\boldsymbol{t}_{12}')^2(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}$, where $\boldsymbol{t}_{12} \sim \mathcal{N}_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$ and $\boldsymbol{T}_1$ are independently distributed. Hence, $\boldsymbol{t}_{12}\boldsymbol{t}_{12}' \sim \boldsymbol{W}_{p-1}(\boldsymbol{I}, 1)$. From Srivastava and Khatri (1979, Problem 3.2, pp 97),

$$E[(\boldsymbol{t}_{12}\boldsymbol{t}_{12}')^2] = 2\boldsymbol{I}_{p-1} + (p-1)\boldsymbol{I}_{p-1} = (p+1)\boldsymbol{I}_{p-1}.$$

Hence,

$$\begin{aligned} E[\boldsymbol{t}_{12}'\boldsymbol{t}_{12}\boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}] =&(p+1)E[\mathrm{tr}\,(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}] \\ =&(p+1)E[\mathrm{tr}\,(\boldsymbol{T}_1\boldsymbol{T}_1')^{-1}] = \frac{(p+1)(p-1)}{m-p}, \end{aligned}$$

since $\boldsymbol{T}_1\boldsymbol{T}_1' \sim \mathcal{W}_{p-1}(\boldsymbol{I}_{p-1}, m)$, and $E[(\boldsymbol{T}_1\boldsymbol{T}_1')^{-1}] = (m-p)^{-1}\boldsymbol{I}_{p-1}$. Hence,

$$E[w_{pp}(\boldsymbol{W}^{-2})_{pp}] = \frac{1}{m-p-1}[1 + \frac{p-1}{m-p} + \frac{p-1}{m-p-3} + \frac{(p+1)(p-1)}{(m-p)(m-p-3)}].$$

We shall need to calculate

$$\begin{aligned} E[w_{11}(\boldsymbol{W}^{-2})_{pp}] =&E[\frac{t_{11}^2}{t_{pp}^4}(1 + \boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12})] \\ =&\frac{1}{(m-p-1)(m-p-3)}E[m + t_{11}^2\boldsymbol{t}_{12}'(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}]. \end{aligned}$$

It may be noted that t_{11} is the $(1,1)$st element of $\boldsymbol{T}_1$, so we need to write $\boldsymbol{T}_1$ as

$$\boldsymbol{T}_1 = \begin{pmatrix} t_{11} & \mathbf{0} \\ \boldsymbol{t}_{31} & \boldsymbol{T}_3 \end{pmatrix}, \quad \boldsymbol{T}_1^{-1} = \begin{pmatrix} t_{11}^{-1} & \mathbf{0} \\ -\boldsymbol{T}_3^{-1}t_{11}^{-1}\boldsymbol{t}_{31} & \boldsymbol{T}_3^{-1} \end{pmatrix}.$$

Thus,

$$\begin{aligned} E[t_{11}^2\mathrm{tr}\,(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}\boldsymbol{t}_{12}'] =&E[t_{11}^2\mathrm{tr}\,(\boldsymbol{T}_1'\boldsymbol{T}_1)^{-1}] \\ =&E[t_{11}^2\left\{t_{11}^{-2} + \frac{\mathrm{tr}\,\boldsymbol{T}_3^{-1}\boldsymbol{t}_{31}\boldsymbol{t}_{31}'(\boldsymbol{T}_3^{-1})'}{t_{11}^2} + \mathrm{tr}\,\boldsymbol{T}_3^{-1}(T_3^{-1})'\right\}] \\ =&E[1 + \mathrm{tr}\,\boldsymbol{t}_{31}\boldsymbol{t}_{31}'(\boldsymbol{T}_3\boldsymbol{T}_3')^{-1} + t_{11}^2\mathrm{tr}\,(\boldsymbol{T}_3\boldsymbol{T}_3')^{-1}] \\ =&1 + (m+1)E[\mathrm{tr}\,(\boldsymbol{T}_3\boldsymbol{T}_3')^{-1}] = 1 + \frac{(m+1)(p-2)}{m-p+1}. \end{aligned}$$

Combining all the above calculation, we get

$$E[w_{11}(\boldsymbol{W}^{-2})_{pp}] = \frac{(m-1)(m+1)}{(m-p+1)(m-p-1)(m-p-3)}.$$

Hence, after some simplification, we get

$$\begin{aligned} E[(\mathrm{tr}\,\boldsymbol{S})(\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{S}^{-2})] =&\frac{(m-1)(m+1)}{(m-p+1)(m-p-1)(m-p-3)}(\mathrm{tr}\,\boldsymbol{\Sigma})(\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}) \\ &- \frac{p}{(m-p-1)(m-p-3)}(\frac{m^2-1}{m-p-1} - \frac{m^2-5m+2p+2}{m-p}). \end{aligned}$$

■

Acknowledgments. The research was supported by NSERC. The research of the second author was supported in part by Grant-in-Aid for Scientific Research (19200020 and 21540114), Japan.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B.N. Petrov and Csaki, F, eds.), 267-281, Akademia Kiado, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Trans. Autom. Contr.*, **AC-19**, 716-723.

[3] Bai, Z.D., and Yin, Y.Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Prob.*, **21**, 1275-1294.

[4] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707-716.

[5] Johnston, I.M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.*, **29**, 295-327.

[6] Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling.* Springer.

[7] Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661-676.

[8] Srivastava, M.S. (2002). *Methods of Multivariate Statistics*, Wiley, New York.

[9] Srivastava, M.S. (2005). Some test concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, **35**, 251-272.

[10] Srivastava, M.S. (2007). Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc.*, **37**, 53-86.

[11] Srivastava, M.S., and Khatri, C.G. (1979). *An Introduction to Multivariate Statistics.* North-Holland, New York.

[12] Srivastava, M.S. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *J. Japan Statist. Soc.*, **37**, 123-134.

[13] Srivastava, M.S. and Kubokawa, T. (2008). Akaike information criterion for selecting components of the mean vector in high dimensional data with fewer observations. *J. Japan Statist. Soc.*, **38**, 259-283.

[14] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. - Theory Methods*, **1**, 13-26.

[15] Yamamura, M., Yanagihara, H. and Srivastava, M.S. (2009). Variable selection in multivariate linear regression models with fewer observations than the dimension. Unpublished manuscript.