CIRJE-F-472

Akaike Information Criterion for Selecting Components of the Mean Vector in High Dimensional Data with Fewer Observations

Muni S. Srivastava
University of Toronto

Tatsuya Kubokawa
University of Tokyo

February 2007; Revised in November 2007

Akaike Information Criterion for Selecting Components of the Mean Vector in High Dimensional Data with Fewer Observations

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

November 9, 2007

Abstract

The Akaike information criterion (AIC) has been successfully used in the literature in model selection for small number of parameters p and large number of observations N. The cases when p is large and close to N or when $p > N$ have not been considered in the literature. In fact, when p is large and close to N, the available AIC does not perform well at all. We consider these cases in the context of finding the number of components of the mean vector that may be different from zero in one-sample multivariate analysis. In fact, we consider this problem in more generality by considering it as a growth curve model introduced in Rao (1959) and Potthoff and Roy (1964). Using simulation, it has been shown that the proposed AIC procedures perform well.

Key words and phrases: Akaike information criterion, high correlation, high dimensional model, ridge estimator, selection of means.

1 Introduction

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be p-dimensional random vectors, independently and identically distributed (hereafter, i.i.d.) as multivariate normal with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$, which is assumed to be positive definite (hereafter, p.d., or simply > 0). We usually wish to test the global hypothesis $H : \boldsymbol{\theta} = \mathbf{0}$ against the alternative $A : \boldsymbol{\theta} \neq \mathbf{0}$. The global hypothesis H can also be written as $H = \bigcap_{i=1}^{p} H_i$, where $H_i : \theta_i = 0$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$. When the global hypothesis H is rejected, it is often desired to find out which component or components θ_i may have caused the rejection of the hypothesis H. Often, it is accomplished by considering the confidence intervals for θ_i by Bonferroni inequality method or Roy's (1953) method. The confidence intervals that do not include zero are the ones that may have caused the rejection of the hypothesis H. The above two methods provide satisfactory solution for small $p < 10$. However, when $p \geq 10$, the above two methods fail to provide satisfactory solution, and either the False Discovery Rate (FDR) method of Benjamini and Hochberg (1995) or k-FWER method of Hoffman and Hommel (1988), and Lehmann and Romano (2005) are used. The FDR method, however, requires that the test statistics that are used for testing the hypotheses H_i are either independently distributed or positively related, see Benjamini and Yekutieli (2001). Similarly, in the k-FWER method, it is not known how to choose 'k' .

As an alternative to FDR and k-FWER procedures, which have limitations as pointed out above, we consider the Akaike information criterion (1973) to determine the number of components

*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.jpFaculty

that may have caused the rejection. Essentially the problem is that for some $r \times 1$ vector $\boldsymbol{\eta}$, $r \leq p$,

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta} \\ \mathbf{0} \end{pmatrix} = \boldsymbol{B}\boldsymbol{\eta},$$

where $\boldsymbol{B}$ is a $p \times r$ matrix given by $(\boldsymbol{I}_r, \mathbf{0}')'$. For a general known $p \times r$ matrix $\boldsymbol{B}$, this problem is called the growth curve model introduced by Rao (1959). A model for the mean matrix was introduced by Potthoff and Roy (1964). For a general discussion of these models, see Srivastava and Khatri (1979), Srivastava (2002) and Kollo and von Rosen (2005).

The aim of this article is to use the Akaike information criterion to choose r, the number of components of $\boldsymbol{\theta}$ that are different from zero. We consider the case when $N > p$ as well as the case when $N \leq p$. In Section 2, we define the Akaike information criterion as well as obtain its exact expression in the growth curve model when $N \geq p + 2$. The AIC is recognized to be a useful method for selecting models when N is large, but it does not perform well when p is large and close to N, because the inverse of the sample covariance matrix is unstable. When $p \geq N$, no information criteria have been considered in the literature. In Section 3, we derive the AIC variants based on the ridge-type estimators of the precision matrix. The case of $N > p$ is treated in Section 3.1, and the ridge information criterion AIC_λ is obtained for large N. The case of $p \geq N$ is handled in Section 3.2, and the ridge information criterion AIC^*_λ is derived for large p. Section 3.3 presents numerical investigation of the proposed information criteria and shows that the AIC variants based on the ridge-type estimators of the precision matrix have nice behaviors, especially in the high dimensional cases and/or high correlation cases. In Section 4, we extend the results to the two-sample problem. All the analytical proofs of the results are given in the appendix.

2 Akaike Information Criterion for Growth Curve Model

2.1 Akaike information criterion and its variant

For model selection and its evaluation, Akaike (1973, 74) developed an information criterion, known in the literature as AIC. It is based on the Kullback and Leibler (1951) information of the true model with respect to the fitted model. Let f be the true but unknown density of the data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, a $p \times N$ matrix of observation vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. And let $g_{\boldsymbol{\theta}} \in \mathcal{G} = \{g(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the density of the approximating model, where $\boldsymbol{\theta} \in \boldsymbol{R}^k$. It will be assumed that $f \in \mathcal{G}$. Since $\boldsymbol{\theta}$ is unknown, it can be estimated by an efficient estimator such as maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\theta}}$. Thus, for a future $p \times N$ observation matrix $\boldsymbol{Z}$, its predictive density can be approximated by $g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{z})$. For model selection, Akaike (1973) proposed to choose that $g \in \mathcal{G}$ for which the average quantity

$$E_{f(\boldsymbol{X})}[I\{f(\boldsymbol{z}); g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{z})\}] = E_{f(\boldsymbol{z})}[\log f(\boldsymbol{Z})] - E_{f(\boldsymbol{X})}[E_{f(\boldsymbol{z})}[\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{Z})]], \tag{2.1}$$

is small. The first term on the right-side of (2.1) does not depend on the model. The Akaike information AI is defined by the second term in (2.1), namely,

$$AI = -2E_{f(\boldsymbol{X})}[E_{f(\boldsymbol{z})}[\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{Z})]].$$

The AIC is an estimator of AI. When $f \in \mathcal{G}$ and $\widehat{\boldsymbol{\theta}}$ is MLE, it is given by

$$AIC_0 = -2 \log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) + 2d \tag{2.2}$$

where d is the number of free parameters in the model $\mathcal{G}$.

Let Δ^* be the bias in estimating AI by $-2\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})$, namely,

$$\Delta^* = E[-2\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X})] - AI.$$

Akaike (1973) showed that $\Delta^* = -2d + o(1)$ as $N \to \infty$ when $f \in \mathcal{G}$ and $\widehat{\boldsymbol{\theta}}$ is MLE. Thus $2d$ in AIC_0 is interpreted as an approximated value of the bias correction term. An exact value of Δ^* can be derived for a specific model, and if Δ^* is free of parameters, then the corrected version of AIC is given by

$$AIC_C = -2\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) - \Delta^*,$$

which was introduced by Sugiura (1978) and studied by Hurvich and Tsai (1989).

When the MLE of $\boldsymbol{\theta}$ is unstable or inefficient, we use a stable or efficient estimator. In this case, the bias Δ^* may depend on unknown parameters. We use an estimator $\widehat{\Delta}^*$, then the AIC-variant based on the estimator is given by

$$AIC_G = -2\log g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{X}) - \widehat{\Delta}^*.$$

For the generalization and recent development of AIC, see Konishi and Kitagawa (1996), Konishi, Ando and Imoto (2004) and Fujikoshi and Satoh (1997). In this paper, we shall derive the AIC variants AIC_G for the growth curve model in various situations.

2.2 AIC for the growth curve model

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ be the $p \times N$ observation matrix, where $\boldsymbol{x}_i$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, $p < N$. The true model density f is also normal with mean vector $\boldsymbol{\theta}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$ except that

$$\boldsymbol{\theta}^* = \boldsymbol{B}\boldsymbol{\eta}^*$$

where $\boldsymbol{B} : p \times r$ and $\boldsymbol{\eta}^* \in \boldsymbol{R}^r$. The model that we wish to fit to the data (hereafter called candidate model), namely $g_{\boldsymbol{\theta},\boldsymbol{\Sigma}}$ is also normal with mean vector $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma} > 0$. Thus the class of candidate models includes the true model. For simplicity of notation we shall write

$$AI = E_{f(\boldsymbol{X})}E_{f(\boldsymbol{Z})}\left[-2\log g_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Sigma}}}(\boldsymbol{Z})\right] = E^*_{\boldsymbol{X}}E^*_{\boldsymbol{Z}}\left[-2\log g(\boldsymbol{Z}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Sigma}})\right], \tag{2.3}$$

where $\overline{\boldsymbol{x}} = N^{-1}\sum_{i=1}^N \boldsymbol{x}_i$, $\boldsymbol{V} = \sum_{i=1}^N (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$,

$$\begin{aligned}\widehat{\boldsymbol{\theta}} =& \boldsymbol{B}\widehat{\boldsymbol{\eta}} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}^{-1}\overline{\boldsymbol{x}},\\ N\widehat{\boldsymbol{\Sigma}} =& \boldsymbol{V} + N\left(\overline{\boldsymbol{x}} - \widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{x}} - \widehat{\boldsymbol{\theta}}\right)'.\end{aligned}$$

As seen in Srivastava and Khatri (1979, pp120), $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the MLE of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ respectively for the candidate model. Note that

$$\begin{aligned}-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Sigma}}) =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}| + \sum_{i=1}^N \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\theta}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\theta}})']\\ =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}| + Np.\end{aligned} \tag{2.4}$$

When the Akaike information AI is estimated by the estimator $-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Sigma}})$, the resulting bias is denoted by Δ^*, given by

$$\Delta^* = E^*_{\boldsymbol{X}}\left[-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Sigma}})\right] - AI. \tag{2.5}$$

The following proposition gives the value of Δ^*, where all the proofs of Propositions will be given in the Appendix.

Proposition 2.1 *For $N = n + 1 > p + 2$, the bias Δ^* in estimating AI by (2.4) is given by*

$$\Delta^* = -\frac{Np(p+3)}{n-p-1} + \frac{N(p-r)(2n-p+r+1)}{(n-p+r)(n-p+r-1)}. \tag{2.6}$$

Thus, from Proposition 2.1, we get the following Corollary.

Corollary 2.1 *The so-called corrected AIC is given by*

$$AIC_C = -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}) - \Delta^*, \tag{2.7}$$

and the uncorrected AIC is given by

$$AIC_0 = -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}) + p(p+1) + 2r. \tag{2.8}$$

It may be noted that Δ^* given in (2.6) can be approximated by

$$\begin{aligned}\Delta_A^* = &-[p(p+1)+2r] - \frac{1}{n}[p(p+2)(p+3) - (p-r)(3(p-r)+5)] \\ &- \frac{1}{n^2}\{p(p+1)(p+2)(p+3) - 2(p-r)(2(p-r)+1)(p-r+2)\}\end{aligned} \tag{2.9}$$

Since the results in the remainder of the paper are asymptotic, and it is easier to handle Δ_A^*, we will use Δ_A^* given by (2.9).

3 Ridge Information Criterion

When $N > p$ and p is large and close to N, the sample matrix $\boldsymbol{V}$ is very unstable, because of many small eigenvalues, and the available AIC does not perform well. And, in the case of $p \geq N$, no information criterion has been considered in the literature. In this section, we obtain information criteria based on a ridge-type estimator of the precision matrix and show numerically that the proposed information criteria perform well in both the cases. The usefulness of the ridge-type estimators has been recognized recently. For example, Srivastava and Kubokawa (2007) and Kubokawa and Srivastava (2005) showed that discriminant procedures based on the ridge estimators yield high correct classification rates in multivariate classification problems.

3.1 Case of $N > p$

Consider the case when $N > p$. In the situation that $\boldsymbol{V}$ is not stable, we consider the ridge-type estimator for $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{V}_\lambda = \boldsymbol{V} + \hat{\lambda}\boldsymbol{I}_p, \tag{3.1}$$

where $\hat{\lambda}$ is a positive function of $\boldsymbol{V}$. Thus we consider the estimator of $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\eta}$ given by

$$\widehat{\boldsymbol{\theta}}_\lambda = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\overline{\boldsymbol{x}}, \tag{3.2}$$

and the corresponding estimator of $\boldsymbol{\Sigma}$ by

$$N\widehat{\boldsymbol{\Sigma}}_\lambda = \boldsymbol{V}_\lambda + N[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}]\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}'[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}]' \tag{3.3}$$

Define the Akaike information AI_λ by

$$AI_\lambda = E_{\boldsymbol{X}}^* E_{\boldsymbol{Z}}^* \left[-2\log g(\boldsymbol{Z}|\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\boldsymbol{\Sigma}}_\lambda)\right], \tag{3.4}$$

and it is estimated by

$$\begin{aligned}-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda,\widehat{\boldsymbol{\Sigma}}_\lambda) =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \sum_{i=1}^{N}\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{x}_i-\widehat{\boldsymbol{\theta}}_\lambda)(\boldsymbol{x}_i-\widehat{\boldsymbol{\theta}}_\lambda)']\\ =& Np\log(2\pi) + Np + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| - \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1},\end{aligned}$$

since $\sum_{i=1}^{N}(\boldsymbol{x}_i-\widehat{\boldsymbol{\theta}}_\lambda)(\boldsymbol{x}_i-\widehat{\boldsymbol{\theta}}_\lambda)' = N\widehat{\boldsymbol{\Sigma}}_\lambda - \hat{\lambda}\boldsymbol{I}_p$. Then the bias is given by

$$\Delta_\lambda^* = E_{\boldsymbol{X}}^*\left[-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda,\widehat{\boldsymbol{\Sigma}}_\lambda)\right] - AI_\lambda. \tag{3.5}$$

Proposition 3.1 *Let Δ_A^* be given by* (2.9)*. Then for larger n and $\hat{\lambda}$ satisfying $\hat{\lambda} = O_p(\sqrt{n})$, the bias Δ_λ^* can be approximated as*

$$\Delta_\lambda^* = \Delta^* + NE_{\boldsymbol{X}}^*\left[\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1}\right] - E_{\boldsymbol{X}}^*\left[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\right] + o(1), \tag{3.6}$$

where Δ^* is given in (2.6)*.*

We choose

$$\hat{\lambda} = \sqrt{np}\hat{a}_1, \quad \text{for} \quad \hat{a}_1 = \mathrm{tr}\,\boldsymbol{V}/(np). \tag{3.7}$$

It is noted that $\hat{\lambda}$ is of the order $O_p(\sqrt{n})$ for fixed p, and for bounded $a_1 = \mathrm{tr}\,\boldsymbol{\Sigma}/p > 0$ for all p, $\hat{a}_1$ converges to a_1 as $n \to \infty$, see Srivastava (2005). Thus $\hat{\lambda}$ increases as p gets large in the order of $\sqrt{p}$. Thus, we get the following corollary for the corrected AIC.

Corollary 3.1 *For $N > p$, and Δ_A^* defined in* (2.9)*, the corrected AIC using $\widehat{\boldsymbol{\theta}}_\lambda$ and $\widehat{\boldsymbol{\Sigma}}_\lambda$ given in* (3.2) *and* (3.3)*, respectively, as estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ can be approximated by*

$$\begin{aligned}AIC_\lambda =& -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda,\widehat{\boldsymbol{\Sigma}}_\lambda) - \Delta_A^* - N\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1} + \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\\ =& Np\log 2\pi + Np + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| - \Delta_A^* - N\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1},\end{aligned} \tag{3.8}$$

where Δ_A^* is given in (2.9)*.*

Our numerical evaluation shows that AIC_λ behaves well in our model selection when p is close to n and $n > p$, see Table 1.

3.2 Case of $p \geq N$

Consider the case when $N \leq p$. In this case, $\boldsymbol{V} = \sum_{i=1}^{N}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})'$ is a singular matrix and its inverse does not exist. Thus, while $n^{-1}\boldsymbol{V}$, $n = N-1$, is an unbiased estimator of $\boldsymbol{\Sigma}^*$, we need an estimator of the precision matrix $\boldsymbol{\Sigma}^{*-1}$. Two types of estimators have been proposed in the literature by Srivastava (2007), Srivastava and Kubokawa (2007) and Kubokawa and Srivastava (2005). One is based on the Moore-Penrose inverse of $\boldsymbol{V}$ such as $a_{n,p}\boldsymbol{V}^+ = a_{n,p}\boldsymbol{H}\boldsymbol{L}^{-1}\boldsymbol{H}'$, where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ and $\boldsymbol{L} = \mathrm{diag}\,(\ell_1,\ldots,\ell_n)$ is the diagonal matrix of the non-zero eigenvalues of $\boldsymbol{V}$, $a_{n,p}$ is a constant depending on n and p. The other is a ridge-type estimator given by

$$\boldsymbol{V}_\lambda = \boldsymbol{V} + \hat{\lambda}\boldsymbol{I}_p,$$

as employed in (3.1). However, since n is usually much smaller than p, we will use

$$\hat{\lambda} = \sqrt{p}\hat{a}_1, \tag{3.9}$$

instead of the one given in (3.7). Let $a_i = \operatorname{tr} \boldsymbol{\Sigma}^i/p$ for $i = 1, 2, 3, 4$. We shall assume the following conditions:

(C.1) $0 < \lim_{p\to\infty} a_i \equiv a_{i0} < \infty$,

(C.2) $n = O(p^\delta)$ for $0 < \delta \leq 1/2$,

(C.3) the maximum eigenvalue of $\boldsymbol{\Sigma}^*$ is bounded in large p.

Then from Lemma A.3, it can be observed that $E[\operatorname{tr} \boldsymbol{V}/(np)] = a_1$ and $\lim_{p\to\infty} \operatorname{tr} \boldsymbol{V}/(np) = a_{10}$ in probability. Hence, the ridge function $\hat{\lambda}$ goes to infinity as $p \to \infty$, and $\hat{\lambda}/n$ goes to zero if $n = O(p^\delta)$ for $1/2 < \delta \leq 1$. However, from the assumption (C.2), it remains constant. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are estimated by the ridge-estimators (3.2) and (3.3) for the ridge function (3.9). Although the MLEs of $\boldsymbol{\theta}^*$ and $\boldsymbol{\Sigma}^*$ do not exist, we define the Akaike information AI_λ as in (3.4) with the estimators $\widehat{\boldsymbol{\theta}}_\lambda$ and $\widehat{\boldsymbol{\Sigma}}_\lambda$ in place of MLE. This gives Δ^*_λ defined by (3.5) instead of Δ^* given in (2.5). When the dimension p tends to infinity, a second-order approximation of Δ^*_λ is given by the following proposition.

Proposition 3.2 *Let $a_{ic} = \operatorname{tr}(\boldsymbol{C}'\boldsymbol{\Sigma}\boldsymbol{C})^i/q$ for $q = p-r$ and $i = 1, 2$, where $\boldsymbol{C}$ is a $p\times(p-r)$ matrix such that $\boldsymbol{C}'\boldsymbol{B} = \boldsymbol{0}$ and $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_{p-r}$. Also let $\boldsymbol{a} = (a_1, a_2, a_{1c}, a_{2c})$. Then under the assumptions* (C.1) - (C.3) *and $\hat{\lambda} = \sqrt{p}\hat{a}_1$, Δ^*_λ can be approximated as*

$$\Delta^*_\lambda(\boldsymbol{a}) = Np - E^*_{\boldsymbol{X}}\left[\hat{\lambda}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\right] - h(\boldsymbol{a}) + o(p^{-1/2}), \tag{3.10}$$

where

$$\begin{aligned} h(\boldsymbol{a}) =& \sqrt{p}N\left(N+1-\frac{qa_{1c}}{pa_1}\right)\left(1+\frac{2a_2}{npa_1^2}\right) - \frac{N(N+1)}{1+1/\sqrt{p}}\frac{na_2}{\sqrt{p}a_1^2} \\ & - \sqrt{p}N\frac{qa_{2c}}{pa_1}\frac{1}{\sqrt{p}a_1+qa_{1c}}. \end{aligned} \tag{3.11}$$

The bias Δ^*_λ includes the unknown values a_i and a_{ic} for $i = 1, 2$, which are estimated by the consistent estimators

$$\hat{a}_1 = \frac{\operatorname{tr}\boldsymbol{V}}{np}, \qquad \hat{a}_2 = \frac{1}{(n-1)(n+2)p}\left[\operatorname{tr}\boldsymbol{V}^2 - (\operatorname{tr}\boldsymbol{V})^2/n\right], \tag{3.12}$$

$$\hat{a}_{1c} = \frac{\operatorname{tr}\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C}}{nq}, \quad \hat{a}_{2c} = \frac{1}{(n-1)(n+2)q}\left[\operatorname{tr}(\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C})^2 - \left(\operatorname{tr}\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C}\right)^2/n\right], \tag{3.13}$$

for $i = 1, 2$. Replacing the unknown values with their estimators yields an estimator of $\Delta^*_\lambda(\boldsymbol{a})$, denoted by $\Delta^*_\lambda(\hat{\boldsymbol{a}})$, where $\hat{\boldsymbol{a}} = (\hat{a}_1, \hat{a}_2, \hat{a}_{1c}, \hat{a}_{2c})$.

Corollary 3.2 *The AIC^*_λ can be approximated by*

$$\begin{aligned} AIC^*_\lambda =& -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\boldsymbol{\Sigma}}_\lambda) - \Delta^*_\lambda(\hat{\boldsymbol{a}}) \\ =& Np\log(2\pi) + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + h(\hat{\boldsymbol{a}}), \end{aligned} \tag{3.14}$$

for the function $h(\cdot)$ given in (3.11).

It is noted that the term $\widetilde{\Delta}^*_\lambda$ depends on the data, namely, it may be affected by random fluctuation. Another choice is to use the rough approximations such that $a_i = 1$ and $a_{ic} = 1$ for $i = 1, 2$, namely $\boldsymbol{a} = (1, 1, 1, 1)$, and the resulting information criterion is given in the following corollary.

Corollary 3.3

$$AIC^*_A = Np\log(2\pi) + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + h(\boldsymbol{1}). \tag{3.15}$$

In the estimation of $\mathbf{\Sigma}^{-1}$, it may be important how to estimate λ. As seen from Lemma A.3, $\hat{\lambda}$ given in (3.9) goes to infinity as $p \to \infty$, namely $\hat{\lambda} = O_p(p^{1/2})$, and it is interesting to consider another estimate of λ with the order $O_p(1)$. We here consider another estimator of the form

$$\hat{\lambda}^{\#} = \sqrt{n}\hat{a}_1. \tag{3.16}$$

Using the same arguments as in Proposition 3.2, we get the following proposition.

Proposition 3.3 *Assume* (C.1) - (C.3) *and* $\hat{\lambda} = \sqrt{n}\hat{a}_1$. *Then the bias* $\Delta_\lambda^{\#}$ *corresponding to* Δ_λ^* *can be approximated as*

$$\Delta_\lambda^{\#} = Np - \hat{\lambda}\mathrm{tr}\,\widehat{\mathbf{\Sigma}}_\lambda^{-1} - h^{\#}(\boldsymbol{a}) + o(1), \tag{3.17}$$

where

$$\begin{aligned} h^{\#}(\boldsymbol{a}) =& \frac{Np}{\sqrt{n}}\left(N+1-\frac{qa_{1c}}{pa_1}\right)\left(1+\frac{2a_2}{npa_1^2}\right) - \frac{N(N+1)}{1+\sqrt{n}/p}\frac{\sqrt{n}a_2}{a_1^2} \\ & - N\frac{qa_{2c}}{\sqrt{n}a_1}\frac{1}{\sqrt{n}a_1+qa_{1c}}. \end{aligned} \tag{3.18}$$

Corollary 3.4 *The* $AIC_\lambda^{\#}$ *for* $\hat{\lambda}^{\#} = \sqrt{n}\hat{a}_1$ *can be approximated by*

$$\begin{aligned} AIC_\lambda^{\#} =& -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\mathbf{\Sigma}}_\lambda) - \Delta_\lambda^{\#}(\hat{\boldsymbol{a}}) \\ =& Np\log(2\pi) + N\log|\widehat{\mathbf{\Sigma}}_\lambda| + h^{\#}(\hat{\boldsymbol{a}}). \end{aligned} \tag{3.19}$$

Srivastava and Kubokawa (2007) proposed the ridge-type empirical Bayes estimator of λ given by $\hat{\lambda}^{\dagger} = \mathrm{tr}\,\boldsymbol{V}/n = p\hat{a}_1$. Using this estimate, we can also other estimates of $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$ based on (3.2) and (3.3). Although intuitive, we here inverstigate the performance of the following criterion such that the bias term corresponds to that of the conventional AIC, namely 2×(the number of unknown parameters).

$$AIC_R^{\dagger} = -2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\mathbf{\Sigma}}_\lambda) + 2\times\{p(p+1)+r\}. \tag{3.20}$$

It is noted that $AIC_R^{\dagger}$ is motivate from the conventional AIC, but no justification can be guranteed in the asymptotics of $p \to \infty$. The performances of $AIC_R^{\dagger}$, $AIC_\lambda^{\#}$, AIC_λ^* and AIC_A^* are investigated in the following section.

3.3 Simulation experiments

We now compare numerical performances of the proposed selection criteria through simulation experiments. As the true model, we consider the model that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are i.i.d. $\sim \mathcal{N}_p(\boldsymbol{\theta}^*, \mathbf{\Sigma}^*)$ where

$$\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*, 0, \ldots, 0)', \quad \theta_i^* = (-1)^i(1+u_i),\ i = 1, \ldots, k,$$

for random variable u_i from a uniform distribution on the interval $[0,1]$, and

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

for a constant ρ on the interval $(-1,1)$ and $\sigma_i = 2 + (p-i+1)/p$. Let (r) be the set $\{0,1,\ldots,r\}$, and we write the model using the first r nonnegative components by M_r or simply (r), namely, the model (r) means that $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ are i.i.d. $\sim \mathcal{N}_p(\boldsymbol{\theta}^{(r)},\boldsymbol{\Sigma})$ where $\boldsymbol{\theta}^{(r)} = (\theta_1,\ldots,\theta_r,0,\ldots,0)'$. For this model, $\boldsymbol{B}$ corresponds to $(\boldsymbol{I}_r,\boldsymbol{0})'$. In our experiments, the true model is M_k or (k), and we consider the set $\{M_r; r = 0,1,\ldots,7\}$ as candidate models.

When $N > p$, we investigate the performances of the information criteria AIC_0, AIC_C and AIC_λ defined in Section 3.1. The following two cases are examined: (A) $N = 50, 100$, $p = 40$, $k = 4$ and models $(0) \sim (7)$; (B) $N = 50, 100$, $p = 45$, $k = 10$ and models $(0) \sim (13)$. The frequencies of models (r), selected by the three criteria are reported in Table 1 based on 100 samples for $\rho = 0.3$. From Table 1, we can find some properties and features about the three criteria. (1) When $N = 50$ and $p = 40, 45$, the conventional AIC and the corrected AIC, denoted by AIC_0 and AIC_C, do not work well, while AIC_λ based on the ridge-type estimator behaves very well. (2) For the large sample case of $N = 100$, AIC_C and AIC_λ perform reasonably well. Thus, it is found that AIC_λ has higher frequencies than AIC_0 and AIC_C. From these observations, we can recommend the use of AIC_λ, which performs well especially when p is close to N.

When $p \geq N$, we carried out similar simulation experiments for the information criteria $AIC_R^\dagger$, $AIC_\lambda^\#$, AIC_λ^* and AIC_A^* which are defined by (3.20), (3.19), (3.14) and (3.15), respectively. Table 2 reports the frequencies of models (r) selected by the three criteria in the three cases: (A) $N = 10$, $p = 100$, $k = 4$ and models $(0) \sim (7)$; (B) $N = 20$, $p = 100$, $k = 10$ and models $(0) \sim (13)$, where for the value of ρ, we handle the two cases $\rho = 0.2$ and $\rho = 0.8$. From the table, it is seen that AIC_A^* performs well except the case of $\rho = 0.8$ and $N = 10$, where $AIC_R^\dagger$ and $AIC_\lambda^\#$ are not very good and AIC_λ^* is superior. For larger N, the preformance of $AIC_R^\dagger$ and $AIC_\lambda^\#$ get better. Thus, the criterion AIC_A^* is recommended.

4 Two-Sample Problem

In this section, we extend the results of the previous section to the two sample problem which may be useful in a practical situation.

4.1 Extension to the two-sample model

Let $\boldsymbol{X}_1 = (\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{1N_1})$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{21},\ldots,\boldsymbol{x}_{2N_2})$ be the two $p \times N_1$ and $p \times N_2$ observation matrices independently distributed in which $\boldsymbol{x}_{1i}$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}_1,\boldsymbol{\Sigma})$ and $\boldsymbol{x}_{2i}$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}_2,\boldsymbol{\Sigma})$. We wish to investigate which components of $\boldsymbol{\theta}_1$ are different from $\boldsymbol{\theta}_2$. Thus, in this model,

$$\boldsymbol{\theta}_1 = \begin{pmatrix} \boldsymbol{\theta}_{11} \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\theta}_2 = \begin{pmatrix} \boldsymbol{\theta}_{21} \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

where $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{21}$ are the r-vectors and $\boldsymbol{\mu}_2$ is a $(p-r)$ vector. That is,

$$\boldsymbol{\delta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = \begin{pmatrix} \boldsymbol{\theta}_{11} - \boldsymbol{\theta}_{21} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta},$$

where $\boldsymbol{B} = (\boldsymbol{I}_r, \boldsymbol{0})'$, and $\boldsymbol{\eta}$ is an r-vector of unknown parameters.

Since all the information from the two observation matrices are contained in the sufficient statistics $\overline{\boldsymbol{x}}_1 = N_1^{-1}\sum_{i=1}^{N_1}\boldsymbol{x}_{1i}$, $\overline{\boldsymbol{x}}_2 = N_2^{-1}\sum_{i=1}^{N_2}\boldsymbol{x}_{2i}$ and

$$\boldsymbol{V}_x = \sum_{i=1}^{N_1}(\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_1)(\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_1)' + \sum_{i=1}^{N_2}(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2)(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2)'$$

Table 1: Frequencies of models selected by the three criteria based on 100 samples for $N > p$ and $\rho = 0.3$

	$\boldsymbol{N = 50}$			$\boldsymbol{N = 100}$		
M_k	AIC_0	AIC_C	AIC_λ	AIC_0	AIC_C	AIC_λ
	$\boldsymbol{p = 40}$		True model: (4)			
(0)	0	100	0	0	0	0
(1)	0	0	0	0	0	0
(2)	0	0	0	0	0	0
(3)	0	0	0	0	0	0
(4)	30	0	100	62	98	98
(5)	21	0	0	16	2	1
(6)	22	0	0	11	0	0
(7)	27	0	0	11	0	1
	$\boldsymbol{p = 45}$		True model: (10)			
(0)	0	100	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(9)	0	0	0	0	0	0
(10)	27	0	100	50	96	97
(11)	21	0	0	19	3	2
(12)	20	0	0	15	1	1
(13)	32	0	0	16	0	0

Table 2: Frequencies of models selected by the four criteria based on 100 samples for $p \geq N$

	$\boldsymbol{\rho = 0.2}$				$\boldsymbol{\rho = 0.8}$			
M_k	$AIC_R^\dagger$	$AIC_\lambda^\#$	AIC_λ^*	AIC_A^*	$AIC_R^\dagger$	$AIC_\lambda^\#$	AIC_λ^*	AIC_A^*
	$\boldsymbol{N = 10}$		$\boldsymbol{p = 100}$		True model: (4)			
(0)	0	0	0	0	99	0	0	0
(1)	0	0	0	0	0	28	0	0
(2)	0	0	0	0	1	25	0	0
(3)	5	0	0	0	0	31	10	18
(4)	95	95	99	100	0	16	89	82
(5)	0	0	1	0	0	1	1	0
(6)	0	0	1	0	0	0	0	0
(7)	0	0	0	0	0	0	0	0
	$\boldsymbol{N = 20}$		$\boldsymbol{p = 100}$		True model: (10)			
(0)	0	0	0	0	0	0	0	0
(1)	0	0	0	0	0	1	0	0
(2)	0	0	0	0	0	3	0	0
(9)	0	0	0	0	0	0	0	0
(10)	100	100	99	100	100	96	98	100
(11)	0	0	1	0	0	0	2	0
(12)	0	0	0	0	0	0	0	0
(13)	0	0	0	0	0	0	0	0

for the parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\Sigma})$, we will consider these sufficient statistics instead of the entire observation matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Let

$$\boldsymbol{d}_x = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2, \quad \text{and} \quad \boldsymbol{u}_x = (N_1\overline{\boldsymbol{x}}_1 + N_2\overline{\boldsymbol{x}}_2)/N,$$

where $N = N_1 + N_2$. Then $\boldsymbol{d}_x$, $\boldsymbol{u}_x$ and $\boldsymbol{V}_x$ are independently distributed and since $(\boldsymbol{d}_x, \boldsymbol{u}_x)$ are one-to-one transformation from $(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2)$, it contains the same amount of information as $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$. Let $\boldsymbol{\nu} = (N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2)/N$ and $k = N_1N_2/N$. Then, $\boldsymbol{d}_x \sim \mathcal{N}_p(\boldsymbol{\delta}, k^{-1}\boldsymbol{\Sigma})$ and $\boldsymbol{u}_x \sim \mathcal{N}_p(\boldsymbol{\nu}, N^{-1}\boldsymbol{\Sigma})$, where

$$\boldsymbol{\delta} = \boldsymbol{B}\boldsymbol{\eta}.$$

Let $\boldsymbol{V}_\lambda = \boldsymbol{V}_x + \hat{\lambda}\boldsymbol{I}_p$, where $\hat{\lambda}$ is a function of $\boldsymbol{V}_x$ and will be specified later. We shall also consider the case when $\hat{\lambda} = 0$. Let

$$\boldsymbol{A}_V = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}.$$

Then we estimate $\boldsymbol{\delta}$, $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\delta}}^* = \boldsymbol{A}_V\boldsymbol{d}_x$, $\widehat{\boldsymbol{\nu}}^* = \boldsymbol{u}_x$ and

$$N\widehat{\boldsymbol{\Sigma}}_\lambda = \boldsymbol{V}_\lambda + k(\boldsymbol{I} - \boldsymbol{A}_V)\boldsymbol{d}_x\boldsymbol{d}_x'(\boldsymbol{I} - \boldsymbol{A}_V) = \boldsymbol{V}_\lambda + k\boldsymbol{V}_\lambda\boldsymbol{G}\boldsymbol{d}_x\boldsymbol{d}_x'\boldsymbol{G}\boldsymbol{V}_\lambda,$$

where $\boldsymbol{G} = \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'$ and $\boldsymbol{C}' = (\boldsymbol{0}, \boldsymbol{I}_{p-r})$: $(p-r)\times p$, that is the first r colums of $\boldsymbol{C}'$ are zeros. In general for $p\times r$ matrix $\boldsymbol{B}$ of rank r, $\boldsymbol{C}$ is a $p\times(p-r)$ matrix such that $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_{p-r}$, and $\boldsymbol{C}'\boldsymbol{B} = \boldsymbol{0}$. Let $\hat{g}$ be the approximatiing model with estimates $\widehat{\boldsymbol{\delta}}^*$, $\widehat{\boldsymbol{\nu}}^*$ and $\widehat{\boldsymbol{\Sigma}}_\lambda$ used for the unknown parameters in the normal model where $\boldsymbol{V}_x = \boldsymbol{Y}\boldsymbol{Y}'$ with $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{y}_i$ i.i.d. $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here $n = N_1 + N_2 - 2$. Thus,

$$\begin{aligned}-2\log g(\boldsymbol{d}_x, \boldsymbol{u}_x, \boldsymbol{V}_x|\widehat{\boldsymbol{\delta}}_x, \widehat{\boldsymbol{\nu}}_x, \widehat{\boldsymbol{\Sigma}}_\lambda) =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{V} + k\boldsymbol{V}_\lambda\boldsymbol{G}\boldsymbol{d}_x\boldsymbol{d}_x'\boldsymbol{G}\boldsymbol{V}_\lambda)] \\ =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + Np - \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1},\end{aligned}$$

with $\boldsymbol{G} = \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'$ and $\boldsymbol{d}_x = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2$.

Under the true model, the parameters are $\boldsymbol{\delta}^*$, $\boldsymbol{\nu}^*$ and $\boldsymbol{\Sigma}^*$, while the random vectors are still normally distributed. The futurre observation matrix is $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ which is independently distributed of $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$. The true model for $\boldsymbol{Z}$ is the same as that for $\boldsymbol{X}$. That is normal with parameters $(\boldsymbol{\delta}^*, \boldsymbol{\nu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\delta}^* = \boldsymbol{B}\boldsymbol{\eta}^*$. The bias is given by

$$\Delta_{2,\lambda}^* = E_{\boldsymbol{X}}^*[-2\log g(\boldsymbol{d}_x, \boldsymbol{u}_x, \boldsymbol{V}_x|\widehat{\boldsymbol{\delta}}_x, \widehat{\boldsymbol{\nu}}_x, \widehat{\boldsymbol{\Sigma}}_\lambda)] - E_{\boldsymbol{X}}^*[E_{\boldsymbol{Z}}^*[-2\log g(\boldsymbol{d}_z, \boldsymbol{u}_z, \boldsymbol{V}_z|\widehat{\boldsymbol{\delta}}_x, \widehat{\boldsymbol{\nu}}_x, \widehat{\boldsymbol{\Sigma}}_\lambda)]]. \tag{4.1}$$

Then we obtain the following results corresponding to Propositions 2.1, 3.1 and 3.2.

Proposition 4.1 *Assume that $\hat{\lambda} = 0$ and $n > p+1$ for $n = N-2$. Then the exact value of the bias $\Delta_{2,\lambda}^*$ given by (A.19), denoted by Δ_2^* for $\hat{\lambda} = 0$, is*

$$\Delta_2^* = -\frac{Np(p+5)}{n-p-1} + \frac{N(p-r)(2n-p+r+3)}{(n-p+r)(n-p+r-1)}. \tag{4.2}$$

Also, Δ_2^* is approximated as $\Delta_2^* = \Delta_{2,A}^* + o(n^{-2})$, where

$$\begin{aligned}\Delta_{2,A}^* =& -[p(p+1) + 2(p+r)] - \frac{1}{n}\{p(p+3)(p+5) - 3(p-r)(p-r+3)\} \\ & - \frac{1}{n^2}\{2p(p+1)(p+3)(p+5) - (p-r)(4(p-r)+5)(p-r+3)\}.\end{aligned}$$

Proposition 4.2 *Assume that $n > p+1$ and $\hat{\lambda}$ satisfies the property $\hat{\lambda} = O_p(\sqrt{n})$ such as in* (3.7). *Then the bias $\Delta^*_{2,\lambda}$ can be approximated as*

$$\Delta^*_{2,\lambda} = \Delta^*_2 + NE^*_{\boldsymbol{X}}\left[\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1}\right] - E^*_{\boldsymbol{X}}\left[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\right] + o(1). \tag{4.3}$$

as $n \to \infty$, where Δ^*_2 is given by (4.2).

Proposition 4.3 *Assume that $p \geq N$ and $\hat{\lambda}$ has the form $\hat{\lambda} = \sqrt{p}\hat{a}_1$ for $\hat{a}_1 = \mathrm{tr}\,\boldsymbol{V}/(np)$. Then under the assumptions* (C.1) *and* (C.2), *$\Delta^*_{2,\lambda}$ can be approximated as $\Delta^*_{2,\lambda}(\boldsymbol{a}) = Np - E^*_{\boldsymbol{X}}\left[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\right] - h_2(\boldsymbol{a}) + o(p^{-1/2})$, where*

$$\begin{aligned} h_2(\boldsymbol{a}) =& \sqrt{p}N\left(N+2-\frac{qa_{1c}}{pa_1}\right)\left(1+\frac{2a_2}{npa_1^2}\right) \\ &- \frac{N(N+2)}{1+1/\sqrt{p}}\frac{na_2}{\sqrt{p}a_1^2} - 3N\frac{qa_{2c}}{\sqrt{p}a_1}\frac{1}{\sqrt{p}a_1+qa_{1c}}. \end{aligned}$$

Taking the simulation results in Section 3.3 into account, we suggest the following ridge information criteria from Propositions 2.1, 3.1 and 3.2. When $N > p$, let $\hat{\lambda} = \sqrt{np}\hat{a}_1$ and consider the ridge information criterion

$$\begin{aligned} AIC_\lambda =& -2\log g(\boldsymbol{X},\boldsymbol{Y}|\widehat{\boldsymbol{\theta}}_{1\lambda},\widehat{\boldsymbol{\theta}}_{2\lambda},\widehat{\boldsymbol{\Sigma}}_\lambda) - \Delta^*_{2,A} - N\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1} + \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} \\ =& Np\log 2\pi + Np + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| - \Delta^*_{2,A} - N\hat{\lambda}\mathrm{tr}\,\boldsymbol{V}_\lambda^{-1}, \end{aligned}$$

where $\Delta^*_{2,A}$ is given in Proposition 4.1. When $p \geq N$, we can propose the criteria corresponding to (3.14) and (3.15). Let $\hat{\lambda} = \sqrt{p}\hat{a}_1$ and consider

$$\begin{aligned} AIC^*_\lambda =& -2\log g(\boldsymbol{X},\boldsymbol{Y}|\widehat{\boldsymbol{\theta}}_{1\lambda},\widehat{\boldsymbol{\theta}}_{2\lambda},\widehat{\boldsymbol{\Sigma}}_\lambda) - \Delta^*_{2,\lambda}(\hat{\boldsymbol{a}}) \\ =& Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + h_2(\hat{\boldsymbol{a}}), \end{aligned}$$

and the AIC corresponding to (3.15) is given by

$$AIC^*_A = Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + h_2(\mathbf{1}).$$

Since $AIC^\dagger_R$ and $AIC^\#_\lambda$ do not perform well as examined in Section 3.3, we do not investigate them in the comparison.

4.2 Numerical studies

We briefly state the numerical results of the information criteria proposed in the previous subsection through the simulation and empirical studies when $p \geq N$ for $N = N_1 + N_2$.

For the simulation study, we carried out similar experiments to Section 3.3 where the mean vectors of the true model are given by

$$\begin{aligned} \boldsymbol{\theta}^*_1 =& (\theta^*_{11},\ldots,\theta^*_{1k},0,\ldots,0)', \quad \theta^*_{1i} = 1.5\times u_{1i},\ i = 1,\ldots,k, \\ \boldsymbol{\theta}^*_2 =& (\theta^*_{21},\ldots,\theta^*_{2k},0,\ldots,0)', \quad \theta^*_{2i} = -1.5\times u_{2i},\ i = 1,\ldots,k, \end{aligned}$$

for random variables u_{1i} and u_{2i} from a uniform distribution on the interval $[0,1]$, and the covarinace matrix $\boldsymbol{\Sigma}^*$ of the true model has the same structure as used there. The performances of the criteria AIC^*_λ and AIC^*_A are examined in the three cases: (A) $N_1 = 10$, $N_2 = 10$, $k = 4$ and models

Table 3: Frequencies of models selected by the two criteria based on 100 samples in the two sample problem for $p = 100$

	$\boldsymbol{\rho = 0.2}$		$\boldsymbol{\rho = 0.6}$	
M_k	AIC^*_λ	AIC^*_A	AIC^*_λ	AIC^*_A
$\boldsymbol{N_1 = 10}$	$\boldsymbol{N_2 = 10}$		True model: (4)	
(0)	0	0	0	0
(1)	0	0	0	0
(2)	0	0	0	0
(3)	2	5	27	54
(4)	94	94	69	46
(5)	4	1	3	0
(6)	0	0	1	0
(7)	0	0	0	0
$\boldsymbol{N_1 = 10}$	$\boldsymbol{N_2 = 30}$		True model: (10)	
(0)	0	0	0	0
(8)	0	0	0	0
(9)	0	0	1	1
(10)	97	99	95	98
(11)	3	1	4	1
(12)	0	0	0	0
(13)	0	0	0	0

(0) $\sim$ (7); (B) $N_1 = 10$, $N_2 = 30$, $k = 10$ and models (0) $\sim$ (13). Table 3 reports the frequencies of models (r) selected by the three criteria based on 100 samples for $p = 100$ and $\rho = 0.2, 0.6$. Table 3 shows that both AIC^*_λ and AIC^*_A have good performances except for the case of small sample sizes $(N_1, N_2) = (10, 10)$ and the high-correlation $\rho = 0.6$. In this case, AIC^*_λ is slightly better.

We next apply the information criterion to the real datasets of microarray referred to as Leukemia. This dataset contains gene expression levels of 72 patients either suffering from acute lymphoblastic leukemia($N_1 = 47$ cases) or acute myeloid leukemia($N_2 = 25$ cases) for 3571 genes. These data are publicly available at

"http://www-genome.wi.mit.edu/cancer".

The description of the above datasets and preprocessing are due to Dettling and Buhlmann (2002), except that we do not process the datasets such that each tissue sample has zero mean and unit variance across genes, which is not explainable in our framework.

We carried out the simple experiments of using the first $p = 150$ dimensional data. We here use the criterion AIC^*_A. The value of the information criterion AIC^*_A under the model of $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ is denoted by $AIC^*_A(0)$, which takes the value $945.951 + C_0$ where C_0 is a constant. We first consider the case of $r = 1$. Let $AIC^*_A(1)_j$ be the value under the model of $\theta_{1j} \neq \theta_{2j}$ and $\theta_{1i} = \theta_{2i}$ for all $i(\neq j)$. Computing $AIC^*_A(1)_j$ for all j from $j = 1$ to $j = 150$, we see that the location which gives the minimum value of $AIC^*_A(1)_j$ is $j = 61$ with $AIC^*_A(1)_{61} = 940.954 + C_0$. We next consider the case of $r = 2$. Let $AIC^*_A(2)_{61,j}$ be the value under the model of $\theta_{1,61} \neq \theta_{2,61}$, $\theta_{1j} \neq \theta_{2j}$ and $\theta_{1i} = \theta_{2i}$ for all $i(\neq 61, j)$. Then the location minimizing $AIC^*_A(2)_{61,j}$ is $j = 118$ with $AIC^*_A(2)_{61,118} = 939.578 + C_0$. We can further consider the case of $r = 3$, and search for locations giving smaller values of $AIC^*_A(3)_{61,118,j}$ which is defined similarly. The possible locations are $j = 110$ and $j = 131$ with $AIC^*_A(3)_{61,118,110} = 940.893 + C_0$ and $AIC^*_A(3)_{61,118,131} = 940.517 + C_0$. It is

observed that the values of AIC_A^* for $r \geq 4$ are larger than $942 + C_0$. Hence the model minimizing the information criterion AIC_A^* is the model with the location $(i, j) = (61, 118)$ for $r = 2$, namely, $\theta_{1,61} \neq \theta_{2,61}$, $\theta_{1,118} \neq \theta_{2,118}$ and $\theta_{1i} = \theta_{2i}$ for all $i(\neq 61, 118)$. Also we can suggest the models with the locations $(i, j, k) = (61, 110, 118)$, $(61, 110, 131)$ for $r = 3$.

5 Concluding remarks

The Akaike information criterion has been very successfully used in model selection. But so far the focus has been for small p (dimension or parameters) and large sample size N. For large p and or when p is close to N, the estimators of the parameters are unstable. However nothing has been known about the performance of AIC. In this article we have modified AIC using the ridge estimator of the precision matrix and evaluated its performance not only for the case when $p < N$ and close to N but have also considered the case when $p \geq N$. We have proposed AIC_λ given in (3.8) for the case when $N > p$, and AIC_λ^* and AIC_A^* given by (3.14) and (3.15) for the case when $p \geq N$. Finally, we have extended the results to the two sample problem.

A Appendix

Before proving Propositions, we provide a unified expression of the bias Δ_λ^* given by (3.5), where the ridge-type estimators $\widehat{\boldsymbol{\theta}}_\lambda$ and $\widehat{\boldsymbol{\Sigma}}_\lambda$ are given by (3.2) and (3.3). To evaluate the bias, we need the following two lemmas which are referred to Srivastava and Khatri (1979, Corollary 1.9.2 and Theorem 1.4.1).

Lemma A.1 *Let $\boldsymbol{B}$ be a $p \times r$ matrix of rank $r \leq p$, and $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. Then there exists a matrix $\boldsymbol{C} : p \times (p - r)$ such that $\boldsymbol{C}'\boldsymbol{B} = \boldsymbol{0}$, $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_{p-r}$, and*

$$\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}^{-1} + \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C})^{-1}\boldsymbol{C}'.$$

Lemma A.2 *Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be nonsingular matrices of proper orders. Then, if $\boldsymbol{Q} = \boldsymbol{P} + \boldsymbol{U}\boldsymbol{V}$,*

$$\boldsymbol{Q}^{-1} = \boldsymbol{P}^{-1} - \boldsymbol{P}^{-1}\boldsymbol{U}(\boldsymbol{I} + \boldsymbol{V}\boldsymbol{P}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{P}^{-1}.$$

From Lemma A.1, it follows that

$$\begin{aligned} N\widehat{\boldsymbol{\Sigma}}_\lambda =& \boldsymbol{V}_\lambda + N[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}]\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}'[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}]' \\ =& \boldsymbol{V}_\lambda + N\boldsymbol{V}_\lambda\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{V}_\lambda. \end{aligned}$$

From Lemma A.2, it is seen that

$$\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} = N\left\{\boldsymbol{V}_\lambda^{-1} - N\frac{\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'}{1 + N\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}}\right\}. \tag{A.1}$$

For $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$, the Akaike information AI_λ of (3.4) is written as

$$\begin{aligned} AI_\lambda =& E_{\boldsymbol{X}}^* E_{\boldsymbol{Z}}^* \left[Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \sum_{i=1}^N \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{z}_i - \widehat{\boldsymbol{\theta}}_\lambda)(\boldsymbol{z}_i - \widehat{\boldsymbol{\theta}}_\lambda)']\right] \\ =& E_{\boldsymbol{X}}^* \left[Np\log 2\pi + N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + N\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\{\boldsymbol{\Sigma}^* + (\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)'\}]\right]. \end{aligned}$$

Then the bias (3.5) is expressed as

$$\begin{aligned}\Delta_\lambda^* =& E\left[-2\log g(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\boldsymbol{\Sigma}}_\lambda)\right] - AI_\lambda \\ =& E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\sum_{i=1}^N(\boldsymbol{x}_i - \widehat{\boldsymbol{\theta}}_\lambda)(\boldsymbol{x}_i - \widehat{\boldsymbol{\theta}}_\lambda)'] - N\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*] - N\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)']\right] \\ =& Np - E_{\boldsymbol{X}}^*[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}] - NE_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*]\right] - NE_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)']\right]. \qquad \text{(A.2)}\end{aligned}$$

To calculate Δ_λ^* in (A.2), we need to evaluate the two terms $E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*]]$ and $E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)']]$, where $\boldsymbol{\theta}^* = \boldsymbol{B}\boldsymbol{\eta}^*$. Then from (A.1),

$$E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*]\right] = NE_{\boldsymbol{X}}^*\left[\mathrm{tr}\,\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1} - N\frac{\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}}{1 + N\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}}\right].$$

Noting that $\boldsymbol{C}'\boldsymbol{B}\boldsymbol{\theta}^* = \boldsymbol{0}$, we get $\boldsymbol{u} = \sqrt{N}(\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}\boldsymbol{C}'\overline{\boldsymbol{x}} \sim \mathcal{N}_{p-r}(\boldsymbol{0}, \boldsymbol{I})$. Let

$$\boldsymbol{W}_\lambda = (\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})(\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}; \qquad \text{(A.3)}$$

$\boldsymbol{W}$ is $\boldsymbol{W}_\lambda$ with $\hat{\lambda} = 0$. Then, it is observed that

$$E_{\boldsymbol{X}}^*\left[N\frac{\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}}{1 + N\overline{\boldsymbol{x}}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\overline{\boldsymbol{x}}}\right] = E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1 + \boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right],$$

which yields that

$$E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*]\right] = NE_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{V}_\lambda^{-1}\boldsymbol{\Sigma}^*]\right] - NE_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1 + \boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right]. \qquad \text{(A.4)}$$

Let $\boldsymbol{A}_V = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}$. Noting that $\boldsymbol{B}'\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{B} = N\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B}$ since $\boldsymbol{C}'\boldsymbol{B} = \boldsymbol{0}$, we see that

$$\begin{aligned}&E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)']\right] \\ &\quad = E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{A}_V(\overline{\boldsymbol{x}} - \boldsymbol{B}\boldsymbol{\eta}^*)(\overline{\boldsymbol{x}} - \boldsymbol{B}\boldsymbol{\eta}^*)'\boldsymbol{A}_V']\right] \\ &\quad = \frac{1}{N}E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{A}_V\boldsymbol{\Sigma}^*\boldsymbol{A}_V']\right],\end{aligned}$$

which is equal to

$$\begin{aligned}&E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{B})^{-1}\boldsymbol{V}_\lambda^{-1}]\right] \\ &\quad = E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\{\boldsymbol{V}_\lambda^{-1} - \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\boldsymbol{C}'\}]\right] \\ &\quad = E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1}] - \mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}\right]. \qquad \text{(A.5)}\end{aligned}$$

Combining (A.2), (A.4) and (A.5), Δ_λ^* given by (A.2) is expressed as

$$\begin{aligned}\Delta_\lambda^* =& Np - N(N+1)E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1}]\right] - E_{\boldsymbol{X}}^*[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}] \\ &+ NE_{\boldsymbol{X}}^*\left[\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}\right] + N^2E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1 + \boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right], \qquad \text{(A.6)}\end{aligned}$$

Propositions 2.1, 3.1 and 3.2 can be proved using the expression (A.6).

A.1 Proof of Proposition 2.1

For this proof, let $\hat{\lambda} = 0$, and denote $\boldsymbol{V}_\lambda$ and $\boldsymbol{W}_\lambda$ for $\hat{\lambda} = 0$ by $\boldsymbol{V}$ and $\boldsymbol{W}$, respectively. In the expression (A.6), it can be easily shown that $E^*_{\boldsymbol{X}}[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\boldsymbol{V}^{-1}]] = p/(n-p-1)$ and $E^*_{\boldsymbol{X}}[\mathrm{tr}\,\boldsymbol{W}^{-1}] = (p-r)/(n-(p-r)-1)$ for $n = N-1$. To evaluate the second term $E^*_{\boldsymbol{X}}\left[\boldsymbol{u}'\boldsymbol{W}^{-2}\boldsymbol{u}/(1+\boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{u})\right]$, the arguments as in Srivastava (1995) are useful. It is noted that under the true model $\boldsymbol{u} \sim \mathcal{N}_{p-r}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{W} \sim \mathcal{W}_{p-r}(\boldsymbol{I}, n)$ are independently distributed. Let $\boldsymbol{\Gamma}$ be an orthogonal matrix with the last row as $\boldsymbol{u}'/\|\boldsymbol{u}\|$, where $\|\boldsymbol{u}\| = (\boldsymbol{u}'\boldsymbol{u})^{1/2}$. Then making the transformation $\widetilde{\boldsymbol{W}} = \boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}'$, we find that $\widetilde{\boldsymbol{W}}$ is still distributed as Wishart, $\mathcal{W}_{p-r}(\boldsymbol{I}, n)$ and hence is independent of $\boldsymbol{u}$. Let $\widetilde{\boldsymbol{W}} = \boldsymbol{T}\boldsymbol{T}'$, where

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

is the unique triangular factorization of $\widetilde{\boldsymbol{W}}$ for $m = p-r$. Then, $\boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{u}/t^2_{mm}$, and

$$\begin{aligned}
\boldsymbol{u}'\boldsymbol{W}^{-2}\boldsymbol{u} =& (\boldsymbol{u}'\boldsymbol{u})(\boldsymbol{0}', 1)\widetilde{\boldsymbol{W}}^{-2}(\boldsymbol{0}'1)' \\
=& (\boldsymbol{u}'\boldsymbol{u})\left[(\tilde{\boldsymbol{w}}^{12})'\tilde{\boldsymbol{w}}^{12} + (\tilde{w}^{mm})^2\right] \\
=& (\boldsymbol{u}'\boldsymbol{u})\left[t^{-4}_{mm} + t^{-4}_{mm}\boldsymbol{t}'_{12}(\boldsymbol{T}'_1\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}\right],
\end{aligned}$$

where $\boldsymbol{0}'$ is an $m-1$ row vector of zeros and

$$\widetilde{\boldsymbol{W}}^{-1} = \begin{pmatrix} \widetilde{\boldsymbol{W}}^{11} & \tilde{\boldsymbol{w}}^{12} \\ (\tilde{\boldsymbol{w}}^{12})' & \tilde{w}^{mm} \end{pmatrix} = (\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}')^{-1}.$$

Hence,

$$\frac{\boldsymbol{u}'\boldsymbol{W}^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{u}} = \frac{\boldsymbol{u}'\boldsymbol{u}}{t^2_{mm}}\frac{1+\boldsymbol{t}'_{12}(\boldsymbol{T}'_1\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}}{t^2_{mm}+\boldsymbol{u}'\boldsymbol{u}},$$

where $\boldsymbol{u}'\boldsymbol{u} \sim \chi^2_m$, $t^2_{mm} \sim \chi^2_{n-m+1}$ and $[1+\boldsymbol{t}'_{12}(\boldsymbol{T}'_1\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}]$ are independently distributed. And from Basu's theorem $\boldsymbol{u}'\boldsymbol{u}/t^2_{mm}$ and $t^2_{mm}+\boldsymbol{u}'\boldsymbol{u}$ are independently distributed. Since $\boldsymbol{t}_{12}$ is independently distributed of $\boldsymbol{T}_1$ and $\boldsymbol{t}_{12} \sim \mathcal{N}_{m-1}(\boldsymbol{0}, \boldsymbol{I})$, it follows that

$$\begin{aligned}
E^*_{\boldsymbol{X}}\left[1+\boldsymbol{t}'_{12}(\boldsymbol{T}'_1\boldsymbol{T}_1)^{-1}\boldsymbol{t}_{12}\right] =& 1 + E^*_{\boldsymbol{X}}\left[\mathrm{tr}\,(\boldsymbol{T}'_1\boldsymbol{T}_1)^{-1}\right] = 1 + E^*_{\boldsymbol{X}}\left[\mathrm{tr}\,(\boldsymbol{T}_1\boldsymbol{T}'_1)^{-1}\right] \\
=& 1 + (m-1)/(n-m) = (n-1)/(n-p+r).
\end{aligned}$$

Note that $E^*_{\boldsymbol{X}}\left[(t^2_{mm}+\boldsymbol{u}'\boldsymbol{u})^{-1}\right] = E[(\chi^2_{n+1})^{-1}] = (n-1)^{-1}$ and $E^*_{\boldsymbol{X}}\left[\boldsymbol{u}'\boldsymbol{u}/t^2_{mm}\right] = m/(n-m-1) = (p-r)/(n-p+r-1)$. Hence,

$$E^*_{\boldsymbol{X}}\left[\frac{\boldsymbol{u}'\boldsymbol{W}^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{u}}\right] = \frac{p-r}{(n-p+r)(n-p+r-1)}. \tag{A.7}$$

Combining (A.6) and (A.7), we get

$$\Delta^* = Np - N(N+1)\frac{p}{n-p-1} + N\frac{p-r}{n-p+r-1} + N^2\frac{p-r}{(n-p+r)(n-p+r-1)},$$

which is equal to the expression (2.6).

A.2 Proof of Proposition 3.1

We shall evaluate each terms in (A.6). It is noted that $\hat{\lambda}/n = O_p(1/\sqrt{n})$ and that $n\boldsymbol{V}_\lambda^{-1} = \{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1} = n\boldsymbol{V}^{-1} + o_p(1)$. Then it can be shown that

$$\begin{aligned} NE_{\boldsymbol{X}}^*\left[\operatorname{tr}\boldsymbol{W}_\lambda^{-1}\right] =& NE_{\boldsymbol{X}}^*\left[\operatorname{tr}\boldsymbol{W}^{-1}\right] + o(1), \\ N^2E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right] =& N^2E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{u}}\right] + o(1), \end{aligned}$$

for $\boldsymbol{W}$ defined below (A.3). Since $\operatorname{tr}[\boldsymbol{\Sigma}^*(\boldsymbol{V}/n)^{-1}] - \operatorname{tr}[\boldsymbol{\Sigma}^*\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}] = (\hat{\lambda}/n)\operatorname{tr}[\boldsymbol{\Sigma}^*\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}(\boldsymbol{V}/n)^{-1}]$, it is observed that

$$\Delta_\lambda^* = \Delta^* + \frac{N}{n}(N+1)E\left[(\hat{\lambda}/n)\operatorname{tr}[\boldsymbol{\Sigma}^*\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}(\boldsymbol{V}/n)^{-1}]\right] + o(1),$$

so that we need to evaluate the second term in the r.h.s. of the equality. Note that $\boldsymbol{V}/n - \boldsymbol{\Sigma}^* = O_p(1/\sqrt{n})$ and $\hat{\lambda}/n = O_p(1/\sqrt{n})$. From the Taylor expansion,

$$(\boldsymbol{V}/n)^{-1} = \boldsymbol{\Sigma}^{*-1} - \boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^{*-1} + O_p(1/n).$$

Substituting this expansion in the second expression on the r.h.s. of Δ_λ^*, we can see that

$$\begin{aligned} &\frac{N}{n}(N+1)E\left[(\hat{\lambda}/n)\operatorname{tr}[\boldsymbol{\Sigma}^*\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}(\boldsymbol{V}/n)^{-1}]\right] \\ &\quad = N(1+2/n)E\left[(\hat{\lambda}/n)\operatorname{tr}\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}\right] \\ &\qquad - N(1+2/n)E\left[(\hat{\lambda}/n)\operatorname{tr}[\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)]\right] + O(1/\sqrt{n}). \end{aligned} \tag{A.8}$$

The first term in the r.h.s. of (A.8) can be written as $N(1+2/n)E[(\hat{\lambda}/n)\operatorname{tr}\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}] = NE[\hat{\lambda}\operatorname{tr}(\boldsymbol{V} + \hat{\lambda}\boldsymbol{I})^{-1}] + O(1/\sqrt{n})$. The second term can be expressed as

$$\begin{aligned} &\sqrt{n}E[\operatorname{tr}(\boldsymbol{V}/n)\operatorname{tr}[\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)]] \\ &= \sqrt{n}E[\operatorname{tr}\boldsymbol{\Sigma}^*\operatorname{tr}[\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)]] \\ &\quad + \sqrt{n}E[\operatorname{tr}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)\operatorname{tr}[\{\boldsymbol{V}/n + (\hat{\lambda}/n)\boldsymbol{I}\}^{-1}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)]] \\ &= \sqrt{n}E[\operatorname{tr}\boldsymbol{\Sigma}^*\operatorname{tr}[(\boldsymbol{V}/n)^{-1}\boldsymbol{\Sigma}^{*-1}(\boldsymbol{V}/n - \boldsymbol{\Sigma}^*)]] + O(1/\sqrt{n}), \end{aligned}$$

which can be seen to be $O(1/\sqrt{n})$ by substituting the Taylor expansion of $(\boldsymbol{V}/n)^{-1}$. Therefore, the proof of Proposition 3.1 is complete.

A.3 Proof of Proposition 3.2

To calculate Δ_λ^* given by (A.6), we need some preliminary results. The following lemmas due to Srivastava (2005, 2007) are useful for evaluating expectations based on $\hat{a}_1$ and $\hat{a}_2$ given by (3.12).

Lemma A.3 (Srivastava (2005)) *Let $\boldsymbol{V} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$. Then,*
(i) *$E[\hat{a}_i] = a_i$ for $i = 1, 2$.*
(ii) *$\lim_{p\to\infty}\hat{a}_i = a_{i0}$ in probability for $i = 1, 2$ if the conditions* (C.1) *and* (C.2) *are satisfied.*
(iii) *$Var(\hat{a}_1) = 2a_2/(np)$.*

Corollary A.1 *Under the conditions* (C.1) *and* (C.2)*, $\hat{a}_i$ is a consistent estimator of a_i for $i = 1, 2$, if $p \to \infty$, or $(n, p) \to \infty$.*

Lemma A.4 (Srivastava (2007)) *Let* $\boldsymbol{V} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$, $n < p$, *and* $\boldsymbol{V} = \boldsymbol{H}_1 \boldsymbol{L} \boldsymbol{H}_1'$, *where* $\boldsymbol{H}_1' \boldsymbol{H}_1 = \boldsymbol{I}_n$ *and* $\boldsymbol{L} = (\ell_1, \ldots, \ell_n)$, *an* $n \times n$ *diagonal matrix which are the non-zero eigenvalues of* $\boldsymbol{V}$. *Then,*
(i) $\lim_{p\to\infty} \boldsymbol{L}/p = a_{10}\boldsymbol{I}_n$ *in probability.*
(ii) $\lim_{p\to\infty} \boldsymbol{H}_1' \boldsymbol{\Sigma} \boldsymbol{H}_1 = (a_{20}/a_{10})\boldsymbol{I}_n$ *in probability.*

For the proofs, see Srivastava (2005, 2007). For the estimators $\hat{a}_{1c}$ and $\hat{a}_{2c}$ given by (3.13), similar results hold.

To prove Proposition 3.2, we need to evaluate each terms in (A.6). We first evaluate the term $E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{V}_\lambda^{-1}]\right]$. Let $\boldsymbol{H} : p \times p$ be an orthogonal matrix $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_p$ such that

$$\boldsymbol{V} = \boldsymbol{H} \begin{pmatrix} \boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{H}',$$

where $\boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)$, $\boldsymbol{H}_1 : p \times n$, $\boldsymbol{L} : n \times n$ diagonal matrix defined above, and $\boldsymbol{H}_2 \boldsymbol{H}_2' = \boldsymbol{I}_p - \boldsymbol{H}_1 \boldsymbol{H}_1'$. Then we get from Lemma A.4,

$$\begin{aligned} E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{V}_\lambda^{-1}]] =& E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{H} \begin{pmatrix} (\boldsymbol{L} + \hat{\lambda}\boldsymbol{I})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\lambda}^{-1}\boldsymbol{I}_{p-n} \end{pmatrix} \boldsymbol{H}']] \\ =& E_{\boldsymbol{X}}^*[\mathrm{tr}\,\boldsymbol{\Sigma}^* \{\boldsymbol{H}_1 (\boldsymbol{L} + \hat{\lambda}\boldsymbol{I})^{-1} \boldsymbol{H}_1' + \hat{\lambda}^{-1} \boldsymbol{H}_2 \boldsymbol{H}_2'\}] \\ =& E_{\boldsymbol{X}}^*[p\hat{\lambda}^{-1} \mathrm{tr}\,\boldsymbol{\Sigma}^*/p - \hat{\lambda}^{-1} \mathrm{tr}\,[(\boldsymbol{I}_n + \hat{\lambda}\boldsymbol{L}^{-1})^{-1} \boldsymbol{H}_1' \boldsymbol{\Sigma}^* \boldsymbol{H}_1]]. \end{aligned}$$

Note that $\hat{\lambda} = \sqrt{p}\hat{a}_1$. Then, $E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{V}_\lambda^{-1}]\right]$ is expressed as

$$E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{V}_\lambda^{-1}]\right] = E_{\boldsymbol{X}}^*\left[\sqrt{p}\frac{a_1}{\hat{a}_1} - \frac{1}{\sqrt{p}\hat{a}_1} \mathrm{tr}\,[(\boldsymbol{I}_n + \sqrt{p}\hat{a}_1 \boldsymbol{L}^{-1})^{-1} \boldsymbol{H}_1' \boldsymbol{\Sigma}^* \boldsymbol{H}_1]\right].$$

It is here noted that

$$\frac{a_1}{\hat{a}_1} = \frac{1}{1 + (\hat{a}_1 - a_1)/a_1} = 1 - \frac{\hat{a}_1 - a_1}{a_1} + \frac{(\hat{a}_1 - a_1)^2}{a_1^2} + o_p(p^{-1}),$$

which gives from Lemma A.3

$$E_{\boldsymbol{X}}^*\left[\frac{a_1}{\hat{a}_1}\right] = 1 + \frac{Var(\hat{a}_1)}{a_1^2} + o(p^{-1}) = 1 + \frac{2a_2}{npa_1^2} + o(p^{-1}).$$

Also from Lemmas A.3 and A.4,

$$\begin{aligned} & E_{\boldsymbol{X}}^*\left[\frac{1}{\sqrt{p}\hat{a}_1} \mathrm{tr}\,[(\boldsymbol{I}_n + \sqrt{p}\hat{a}_1 \boldsymbol{L}^{-1})^{-1} \boldsymbol{H}_1' \boldsymbol{\Sigma}^* \boldsymbol{H}_1]\right] \\ &= \frac{1}{\sqrt{p}a_1} \mathrm{tr}\,[(\boldsymbol{I}_n + \boldsymbol{I}_n/\sqrt{p})^{-1} \frac{a_2}{a_1} \boldsymbol{I}_n] + o(p^{-1/2}) \\ &= \frac{na_2}{(1 + 1/\sqrt{p})\sqrt{p}a_1^2} + o(p^{-1/2}). \end{aligned}$$

Combining these evaluations, we get

$$E_{\boldsymbol{X}}^*\left[\mathrm{tr}\,[\boldsymbol{\Sigma}^* \boldsymbol{V}_\lambda^{-1}]\right] = \sqrt{p}\left(1 + \frac{2a_2}{npa_1^2} - \frac{na_2}{(1 + 1/\sqrt{p})pa_1^2}\right) + o(p^{-1/2}). \tag{A.9}$$

We next evaluate the term $E^*_{\boldsymbol{X}}\left[\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}\right]$. Let $q=p-r$, and let $\boldsymbol{\Sigma}^*_c=\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C}$ and $\boldsymbol{V}_c=\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C}$. Let $\boldsymbol{H}_c: q\times q$ be an orthogonal matrix such that $\boldsymbol{H}_c\boldsymbol{H}_c'=\boldsymbol{I}_q$ and

$$\boldsymbol{C}'\boldsymbol{V}\boldsymbol{C}=\boldsymbol{H}_c\begin{pmatrix}\boldsymbol{L}_c & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{H}_c',$$

where $\boldsymbol{H}_c=(\boldsymbol{H}_{1c},\boldsymbol{H}_{2c})$, $\boldsymbol{H}_{1c}: q\times n$, $\boldsymbol{L}_c: n\times n$ diagonal matrix, and $\boldsymbol{H}_{2c}\boldsymbol{H}_{2c}'=\boldsymbol{I}_q-\boldsymbol{H}_{1c}\boldsymbol{H}_{1c}'$. Then, we get

$$\begin{aligned}
\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}=&\mathrm{tr}\,[\boldsymbol{\Sigma}^*_c(\boldsymbol{V}_c+\hat\lambda\boldsymbol{I}_q)^{-1}]\\
=&\mathrm{tr}\,[\boldsymbol{\Sigma}^*_c\boldsymbol{H}_c\begin{pmatrix}(\boldsymbol{L}_c+\hat\lambda\boldsymbol{I}_n)^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \hat\lambda^{-1}\boldsymbol{I}_{q-n}\end{pmatrix}\boldsymbol{H}_c']\\
=&\mathrm{tr}\,[\boldsymbol{\Sigma}^*_c\{\boldsymbol{H}_{1c}(\boldsymbol{L}_c+\hat\lambda\boldsymbol{I}_n)^{-1}\boldsymbol{H}_{1c}'+\hat\lambda^{-1}\boldsymbol{H}_{2c}\boldsymbol{H}_{2c}'\}]\\
=&\mathrm{tr}\,\boldsymbol{\Sigma}^*_c/\hat\lambda-\hat\lambda^{-1}\mathrm{tr}\,[(\boldsymbol{I}_n+\hat\lambda\boldsymbol{L}_c^{-1})^{-1}\boldsymbol{H}_{1c}'\boldsymbol{\Sigma}^*_c\boldsymbol{H}_{1c}]\\
=&\frac{qa_{1c}}{\sqrt{p}\hat a_1}-\frac{1}{\sqrt{p}\hat a_1}\mathrm{tr}\,[(\boldsymbol{I}_n+(\sqrt{p}\hat a_1/q)q\boldsymbol{L}_c^{-1})^{-1}\boldsymbol{H}_{1c}'\boldsymbol{\Sigma}_c\boldsymbol{H}_{1c}],
\end{aligned}\tag{A.10}$$

where $\hat\lambda=\sqrt{p}\hat a_1$. Using similar arguments as in (A.9), we can see that

$$\begin{aligned}
&E^*_{\boldsymbol{X}}[\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}]\\
=&\frac{qa_{1c}}{\sqrt{p}a_1}\left\{1+\frac{1}{a_1^2}Var(\hat a_1)\right\}-\frac{1}{\sqrt{p}a_1}\mathrm{tr}\left(\boldsymbol{I}_n+\frac{\sqrt{p}a_1}{qa_{1c}}\boldsymbol{I}_n\right)^{-1}\frac{a_{2c}}{a_{1c}}+o(p^{-1/2})\\
=&\frac{qa_{1c}}{\sqrt{p}a_1}\left\{1+\frac{2a_2}{npa_1^2}\right\}-\frac{na_{2c}}{\sqrt{p}a_1a_{1c}}\frac{qa_{1c}}{\sqrt{p}a_1+qa_{1c}}+o(p^{-1/2}).
\end{aligned}\tag{A.11}$$

Finally, we shall show that $E^*_{\boldsymbol{X}}\left[\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}/(1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u})\right]$ can be evaluated as

$$E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\Big|\boldsymbol{V}\right]=\frac{\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-2})}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}+o_p(p^{-1/2}).\tag{A.12}$$

To this end, it is noted that

$$\begin{aligned}
&E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{(1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u})}\Big|\boldsymbol{V}\right]-\frac{\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-2})}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}\\
&=E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{(1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u})}\Big|\boldsymbol{V}\right]-E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}\Big|\boldsymbol{V}\right]\\
&=-E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}(\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1}))}{(1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u})(1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1}))}\Big|\boldsymbol{V}\right],
\end{aligned}$$

the absolute value of which, from the Cauchy-Schwartz inequality, is less than or equal to

$$\begin{aligned}
&\left\{E\Big[\Big\{\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\Big\}^2\Big|\boldsymbol{V}\Big]\times E\Big[\Big\{\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}\Big\}^2\Big|\boldsymbol{V}\Big]\right\}^{1/2}\\
&\le\left\{E\Big[\Big(\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\Big)^2\Big|\boldsymbol{V}\Big]\right\}^{1/2}\times\frac{\left\{E\Big[\big(\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big)^2\Big|\boldsymbol{V}\Big]\right\}^{1/2}}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}.
\end{aligned}$$

Hence, it is sufficient to show that

$$\lim_{p\to\infty} pE\Big[\big(\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\big)^2\Big|\boldsymbol{V}\Big] \times \frac{E\Big[\big(\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big)^2\Big|\boldsymbol{V}\Big]}{\big\{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big\}^2}=0, \tag{A.13}$$

for the proof of (A.12). It can be verified that for $q\times q$ matrices $\boldsymbol{G}$ and $\boldsymbol{Q}$,

$$E\left[\boldsymbol{u}'\boldsymbol{G}\boldsymbol{u}\times\boldsymbol{u}'\boldsymbol{Q}\boldsymbol{u}\right]=\mathrm{tr}\,(\boldsymbol{G})\mathrm{tr}\,(\boldsymbol{Q})+2\mathrm{tr}\,(\boldsymbol{G}\boldsymbol{Q}),$$

which is used to get that

$$\begin{aligned}E\Big[\big(\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big)^2\Big|\boldsymbol{V}\Big]&=(\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1})^2+2\mathrm{tr}\,\boldsymbol{W}_\lambda^{-2}-2(\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1})^2+(\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1})^2\\&=2\mathrm{tr}\,\boldsymbol{W}_\lambda^{-2}.\end{aligned} \tag{A.14}$$

Using the same arguments as in (A.10), we can show that

$$\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}=\frac{qa_{1c}}{\sqrt{p}a_1}+o_p(p^{1/2}), \tag{A.15}$$

$$\begin{aligned}\mathrm{tr}\,\boldsymbol{W}_\lambda^{-2}=&\frac{1}{p\hat{a}_1^2}\mathrm{tr}\left[\boldsymbol{\Sigma}_c^*(\boldsymbol{I}_q-\boldsymbol{H}_{1c}(\boldsymbol{I}_n+\sqrt{p}\hat{a}_1\boldsymbol{L}_c^{-1})^{-1}\boldsymbol{H}_{1c}')\right.\\&\left.\times\boldsymbol{\Sigma}_c^*(\boldsymbol{I}_q-\boldsymbol{H}_{1c}(\boldsymbol{I}_n+\sqrt{p}\hat{a}_1\boldsymbol{L}_c^{-1})^{-1}\boldsymbol{H}_{1c}')\right]\\=&\frac{qa_{2c}}{pa_1^2}+o_p(1).\end{aligned} \tag{A.16}$$

Hence, it is observed that

$$\begin{aligned}&\lim_{p\to\infty}p\frac{E\Big[\big(\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}-\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big)^2\Big|\boldsymbol{V}\Big]}{\big\{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big\}^2}\\&\quad=2\lim_{p\to\infty}p\frac{\mathrm{tr}\,\boldsymbol{W}_\lambda^{-2}}{\big\{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})\big\}^2}=2\lim_{p\to\infty}\frac{qa_{2c}/(pa_1^2)}{\big\{1/\sqrt{p}+a_{1c}/a_1\big\}^2},\end{aligned}$$

which is bounded. On the other hand, it is noted that

$$\begin{aligned}\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\le&\sup_{\boldsymbol{u}}\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\le\mathrm{ch}_{max}(\boldsymbol{\Sigma}_c^*)\times\mathrm{ch}_{max}\{(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})^{-1}\}\\\le&\mathrm{ch}_{max}(\boldsymbol{\Sigma}_c^*)\times\frac{1}{\hat{\lambda}}=\frac{\mathrm{ch}_{max}(\boldsymbol{\Sigma}_c^*)}{\sqrt{p}\hat{a}_1},\end{aligned}$$

where $\mathrm{ch}_{max}(\boldsymbol{A})$ denotes the maximum eigenvalue of a matrix $\boldsymbol{A}$. Since $\mathrm{ch}_{max}(\boldsymbol{\Sigma}_c^*)$ is bounded from the condition (C.2), it is seen that $\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}/\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}=O_p(1/\sqrt{p})$. Hence, the approximation (A.12) is proved, so that

$$E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right]=E\left[\frac{\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-2})}{1+\mathrm{tr}\,(\boldsymbol{W}_\lambda^{-1})}\right]+o(p^{-1/2}).$$

Using (A.15) and (A.16) again, we obtain that

$$\begin{aligned}E\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right]=&\frac{qa_{2c}/(pa_1^2)}{1+qa_{1c}/(\sqrt{p}a_1)}+o(p^{-1/2})\\=&\frac{qa_{2c}}{\sqrt{p}a_1}\frac{1}{\sqrt{p}a_1+qa_{1c}}+o(p^{-1/2}).\end{aligned} \tag{A.17}$$

Combining (A.6), (A.9), (A.11) and (A.17), we get the second order approximation given by

$$\Delta_\lambda = Np - \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} - h(\boldsymbol{a}) + o(p^{-1/2}), \tag{A.18}$$

where

$$\begin{aligned} h(\boldsymbol{a}) =& N(N+1)\sqrt{p}\left\{1 + \frac{2a_2}{npa_1^2} - \frac{na_2}{pa_1^2}\frac{1}{1+1/\sqrt{p}}\right\} \\ & - N\sqrt{p}\left\{\frac{qa_{1c}}{pa_1}\left(1 + \frac{2a_2}{npa_1^2}\right) - \frac{na_{2c}}{pa_1a_{1c}}\frac{1}{1+\sqrt{p}a_1/(qa_{1c})}\right\} - N^2\frac{qa_{2c}}{pa_1^2}\frac{1}{1+qa_{1c}/(\sqrt{p}a_1)}. \end{aligned}$$

which can be expressed as in (3.10), and the proof of Proposition 3.2 is complete.

A.4 Proof of Proposition 3.3

This can be shown by using the same arguments as in the proof of Proposition 3.2, and we observe that

$$\Delta_\lambda^* = Np - \hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} - h^{\#}(\boldsymbol{a}) + o(1),$$

where

$$\begin{aligned} h^{\#}(\boldsymbol{a}) =& N(N+1)\frac{p}{\sqrt{n}}\left\{1 + \frac{2a_2}{npa_1^2} - \frac{na_2}{pa_1^2}\frac{1}{1+\sqrt{n}/p}\right\} \\ & - \frac{Np}{\sqrt{n}}\left\{\frac{qa_{1c}}{pa_1}\left(1 + \frac{2a_2}{npa_1^2}\right) - \frac{na_{2c}}{pa_1a_{1c}}\frac{1}{1+\sqrt{n}a_1/(qa_{1c})}\right\} - N^2\frac{qa_{2c}}{na_1^2}\frac{1}{1+qa_{1c}/(\sqrt{n}a_1)}, \end{aligned}$$

which can be expressed as in (3.17), and the proof of Proposition 3.3 is complete.

A.5 Proof of Propositions 4.1, 4.2 and 4.3

We shall evaluate the bias (4.1). Let $R^* = E_{\boldsymbol{Z}}^*[-2\log g(\boldsymbol{d}_z, \boldsymbol{u}_z, \boldsymbol{V}_z|\widehat{\boldsymbol{\delta}}_x, \boldsymbol{\nu}_x, \widehat{\boldsymbol{\Sigma}}_\lambda) - Np\log 2\pi]$. Then,

$$\begin{aligned} R^* =& N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + E_{\boldsymbol{Z}}^*[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\{\boldsymbol{V}_z + k(\boldsymbol{d}_z - \widehat{\boldsymbol{\delta}}_x)(\boldsymbol{d}_z - \widehat{\boldsymbol{\delta}}_x)' + N(\boldsymbol{u}_z - \widehat{\boldsymbol{\nu}}_x)(\boldsymbol{u}_z - \widehat{\boldsymbol{\nu}}_x)'\}]] \\ =& N\log|\widehat{\boldsymbol{\Sigma}}_\lambda| + \mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}[N\boldsymbol{\Sigma}^* + k(\widehat{\boldsymbol{\delta}}_x - \boldsymbol{\delta}^*)(\boldsymbol{\delta}_x - \boldsymbol{\delta}^*)' + N(\widehat{\boldsymbol{\nu}}_x - \boldsymbol{\nu}^*)(\widehat{\boldsymbol{\nu}}_x - \boldsymbol{\nu}^*)']], \end{aligned}$$

and thus using the results from (A.3) and (A.4), we observe that

$$\begin{aligned} E_{\boldsymbol{X}}^*[R^*] =& NE_{\boldsymbol{X}}^*[\log|\widehat{\boldsymbol{\Sigma}}_\lambda|] + E_{\boldsymbol{X}}^*[(N+1)\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\Sigma}^*]] + E_{\boldsymbol{X}}^*[k\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\delta}}_x - \boldsymbol{\delta}^*)(\widehat{\boldsymbol{\delta}}_x - \boldsymbol{\delta}^*)']] \\ =& NE_{\boldsymbol{X}}^*[\log|\widehat{\boldsymbol{\Sigma}}_\lambda|] + (N+1)E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\boldsymbol{\Sigma}^*\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}]] + E_{\boldsymbol{X}}^*[\mathrm{tr}\,[\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{A}_V\boldsymbol{\Sigma}^*\boldsymbol{A}_V']] \\ =& NE_{\boldsymbol{X}}^*[\log|\widehat{\boldsymbol{\Sigma}}_\lambda|] + N(N+2)E_{\boldsymbol{X}}^*[\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1}] - NE_{\boldsymbol{X}}^*[\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}] \\ & - N(N+1)E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right], \end{aligned}$$

where $\boldsymbol{u} = \sqrt{k}(\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}\boldsymbol{C}'\boldsymbol{u}_x \sim \mathcal{N}_{p-r}(\boldsymbol{0}, \boldsymbol{I})$ and

$$\boldsymbol{W}_\lambda = (\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}(\boldsymbol{C}'\boldsymbol{V}_\lambda\boldsymbol{C})(\boldsymbol{C}'\boldsymbol{\Sigma}^*\boldsymbol{C})^{-1/2}.$$

Hence, the bias is given by

$$\begin{aligned} \Delta_{2,\lambda}^* =& E_{\boldsymbol{X}}^*[-2\log g(\boldsymbol{d}_x, \boldsymbol{u}_x, \boldsymbol{V}_x|\widehat{\boldsymbol{\delta}}_x, \widehat{\boldsymbol{\nu}}_x, \widehat{\boldsymbol{\Sigma}}_\lambda)] - E_{\boldsymbol{X}}^*[E_{\boldsymbol{Z}}^*[-2\log g(\boldsymbol{d}_z, \boldsymbol{u}_z, \boldsymbol{V}_z|\widehat{\boldsymbol{\delta}}_x, \widehat{\boldsymbol{\nu}}_x, \widehat{\boldsymbol{\Sigma}}_\lambda)]] \\ =& Np - E_{\boldsymbol{X}}[\hat{\lambda}\mathrm{tr}\,\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}] - N(N+2)E_{\boldsymbol{X}}^*[\boldsymbol{\Sigma}^*\boldsymbol{V}_\lambda^{-1}] + NE_{\boldsymbol{X}}^*[\mathrm{tr}\,\boldsymbol{W}_\lambda^{-1}] \\ & - N(N+1)E_{\boldsymbol{X}}^*\left[\frac{\boldsymbol{u}'\boldsymbol{W}_\lambda^{-2}\boldsymbol{u}}{1+\boldsymbol{u}'\boldsymbol{W}_\lambda^{-1}\boldsymbol{u}}\right]. \end{aligned} \tag{A.19}$$

Acknowledgments. The research of the first author was supported by NSERC. The research of the second author was supported in part by a grant from the Ministry of Education, Japan, No. 16500172 and in part by a grant from the 21st Century COE Program at Faculty of Economics, University of Tokyo.

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] Benjamini, Y. and Hockberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.*, **B 57**, 289-300.

[4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165-1181.

[5] Dettling, M. and Buhlmann, P. (2002). Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**, 1061-1069.

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

[7] Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In *Multiple Hypotheses Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.), 154–161. Springer, Heidelberg.

[8] Hurvich, C. and Tsai C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

[9] Kollo, T. and von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht.

[10] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875-890.

[11] Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27-43.

[12] Kubokawa, T. and Srivastava, M.S. (2005). Estimation of the Precision Matrix of a Singular Wishart Distribution and its Application in High Dimensional Data. Discussion paper CIRJE-F-362, Faculty of Economics, University of Tokyo.

[13] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

[14] Lehmann, E.L. and Romano, J.P. (2005). Generalizations of the familywise error rate. *Ann. Statist.*, **33**, 1138-1154.

[15] Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-.

[16] Rao, C.R. (1959). Some problems involving linear hypothesis in multivariate analysis. *Biometrika*, **46**, 49-58.

[17] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.*, **24**, 220-238.

[18] Srivastava, M.S. (1995). Comparison of the inverse and classical estimators in multi-univariate linear calibration. *Comm. Statist.-Theory Methods*, **24**, 2753-2767.

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

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

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

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

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

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