High-Dimensional Consistency of Estimation Criteria for the Rank in Multivariate Linear Model

Yasunori Fujikoshi* and Tetsuro Sakurai**

*Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

**Center of General Education, Tokyo University of Science, Suwa, 5000-1 Toyohira, Chino, Nagano 391-0292, Japan

Abstract

The model selection criteria AIC, BIC and C_p have been proposed for estimation of the rank of coefficient matrix in multivariate linear model. In general, it is known that under a large-sample asymptotic framework AIC and C_p is not consistent, but BIC is consistent. However, we note that these criteria have consistency when the number p of the response variables and the sample size n are large under a high-dimensional asymptotic framework such that the ratio p/n tends to a constant c $(0 \leq c < 1)$ as p and n are large. The consistency properties are also shown for extended criteria with a tuning parameter. Further, we propose the ridge-type criteria whose justifications are given under a large-sample asymptotic framework. Their consistencies are shown in a high-dimensional asymptotic framework. Through a Monte Carlo simulation experiment our results are checked numerically, and the estimation criteria are compared.

AMS 2000 subject classification: primary 62H12; secondary 62H30

Key Words and Phrases: AIC, BIC, C_p, Consistency property, Dimensionality, Discriminant analysis, High-dimensional framework, Multivariate regression model, Multivariate linear model, Rank, Ridge type criteria, Tuning parameter.

1 Introduction

The paper concentrates the problem of estimating the rank (or the dimensionality) of coefficient matrix in multivariate linear model. The model with a reduced rank in multivariate regression model is called multivariate reduced-rank regression model (see, Izenman (1975), Reinsel and Velu (1998)). The problem includes also the one of estimating the number of meaningful discriminant functions in discriminant analysis.

One of the estimation methods is based on sequential test procedures. The tests on each steps are to use the likelihood ratio tests which were first obtained by Anderson (1951, 2003). There are the methods based on the use of model selection criteria which are discussed in this paper. The cross-validation method is also used. Yuan, Ekici, Lu and Monteiro (2007) proposed a method based on penalization. Chen and Huang (2012) have proposed a simultaneous method for selecting the rank and the response variables by using a penalized regression with a group lasso penalty. For related works, see Bunea, She and Wegkamp (2011, 2012).

We consider multivariate linear model with p response variables $y_1, \ldots, y_p$, k explanatory variables $x_1, \ldots, x_k$, and coefficient matrix $\mathbf{\Theta}$ of size $k \times p$. Our interest is to estimate the rank of $\mathbf{C\Theta}$, where $\mathbf{C}$ is a $q \times k$ given matrix with rank q. We are concerned with the estimation methods by use of the model selection criteria AIC (Akaike (1973)), C_p (Mallows (1973)) and BIC (Schwarz (1978)). The criteria based on the first two model selection criteria were proposed by Fujikoshi and Veitch (1979) in multivariate linear model and canonical correlation analysis. The criterion based on BIC was considered by Gunderson and Muirhead (1997) in canonical correlation analysis.

It is known (Fujikoshi (1985)) that under a large-sample asymptotic framework such that

$$n \to \infty, \quad p,\ q,\ k;\ \text{fixed}, \tag{1}$$

the criteria A and C based on AIC and C_p are not consistent, but the criterion B based on BIC is consistent (Gunderson and Muirhead (1997)), with the additional assumption that the order of nonsentraly matrix $\tilde{\mathbf{\Omega}}$ (see (6)) is

$O(n)$. Here, for a matrix $\mathbf{A}$, $\mathrm{O}(a)$ means that each elements of $\mathbf{A}$ is $\mathrm{O}(a)$.

In this paper we examine consistencies of the criteria A, B and C in a high-dimensional asymptotic framework such that

$$p \to \infty, \quad n \to \infty, \quad p/n \to c \in [0, 1). \tag{2}$$

The true rank (or dimension) j_* and the possibly maximum rank $q(\geq j_*)$ are fixed. Further, we assume two types of assumptions on largeness of the noncentrality parameter matrix $\mathbf{\Omega}$ (see (16)) given by (i) $\mathbf{\Omega} = \mathrm{O}(n)$ and (ii) $\mathbf{\Omega} = \mathrm{O}(np)$, respectively. Then, it is shown that the criteria A and C are consistent under some additional assumptions depending on (i) and (ii). On the other hand, we note that the criterion B is consistent under $\mathbf{\Omega} = \mathrm{O}(np)$, but is not consistent under $\mathbf{\Omega} = \mathrm{O}(n)$. These results are shown by deriving sufficient conditions for the extended criteria IC_ν and $\mathrm{C}_{p\nu}$ to be consistent, where $\mathrm{IC}_2 = \mathrm{A}$, $\mathrm{IC}_{\log n} = \mathrm{B}$ and $\mathrm{C}_{p2} = \mathrm{C}$. Note that ν may be regarded as a tuning parameter. Similar high-dimensional consistency properties have been derived Yanagihara, Wakaki and Fujikoshi (2012) and Fujikoshi, Sakurai and Yanagihara (2013) for the AIC and/or in selection of the explanatory variables in multivariate linear model.

In this paper we also propose ridge-type criteria of A, B and C denoted by $\tilde{\mathrm{A}}_\lambda$, $\tilde{\mathrm{B}}_\lambda$ and $\tilde{\mathrm{C}}_\lambda$, respectively. The ridge-type criteria are defined by using a ridge estimator of the covariance matrix. The criteria are used also for the case $p > n - k$. Some justifications of these criteria are given under a large-sample asymptotic framework. We show their consistency properties under a high-dimensional asymptotic framework such that $p/n \to c \in [0, 1)$. For $p > n - k$, we point some tendency through a Monte Carlo simulation experiment.

Our methods are related to multivariate regression model and discrinant analysis which are very often used for analysis of multivariate data. The data with a relatively large number of response variables are appeared in many fields like medical field, and the data with a large number of response variables are appeared in stock data, genome data, etc. A high-dimensional

asymptotic framework will be also applicable for a data set based on a selection from a large number of response variables.

The present paper is organized as follows. In section 2, we present the multivariate linear model with a reduced rank, and two special cases are explained. Then we prepare three criteria and their extensions with a tuning parameter. In Section 3, we give sufficient conditions for the criteria IC_ν and $\mathrm{C}_{p\nu}$ to be consistent. As a special case, we give consistency properties of the criteria A, B and C. We check our theoretical results by conducting a Monte Carlo simulation experiment, and compare with the selection probabilities of the two criteria. In Section 4, we propose ridge-type criteria whose consistencies are theoretically and numerically studied. In Section 5, we discuss our conclusions. The proofs of our results are given in Appendix.

2 Rank Estimation Criteria

2.1 Multivariate Linear Model with Reduced Rank

We consider a multivariate linear model of p response variables $y_1, \ldots, y_p$ on a subset of k explanatory variables $x_1, \ldots, x_k$. Suppose that there are the n observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ on $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_k)'$, respectively, and let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be the $n \times p$ and $n \times k$ observation matrices of $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. The multivariate normal linear model is written as

$$\mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}\mathbf{\Theta}, \mathbf{\Sigma} \otimes \mathbf{I}_n), \tag{3}$$

where $\mathbf{\Theta}$ is a $k \times p$ unknown matrix of coefficients, $\mathbf{\Sigma}$ is a $p \times p$ unknown covariance matrix, and $\mathbf{I}_n$ is the identity matrix of order n. The notation $\mathrm{N}_{n\times p}(\cdot, \cdot)$ means the matrix normal distribution such that the mean of $\mathbf{Y}$ is $\mathbf{X}\mathbf{\Theta}$ and the covariance matrix of $\mathrm{vec}\,(\mathbf{Y})$ is $\mathbf{\Sigma}\otimes\mathbf{I}_n$, where $\mathrm{vec}\,(\mathbf{Y})$ is the $np\times 1$ vector formed by stacking the columns of $\mathbf{Y}$ under each other. We assume that $n - k > p$ and $\mathrm{rank}(\mathbf{X}) = k$. When $\boldsymbol{x}$ has a set of dummy variables, $\mathbf{X}$ may not be the full rank. However, as is well known, there are some linear restrictions on the parameters, and we can make $\mathbf{X}$ a full rank matrix.

Let $\mathbf{C}$ be a given $q \times k$ matrix with $\mathrm{rank}(\mathbf{C}) = q$, consider reduced rank models including H

$$M_j : \mathrm{rank}(\mathbf{C}\boldsymbol{\Theta}) = j, \quad j = 0, 1, \ldots, m, \quad m = \min(p, q). \tag{4}$$

For testing $M_j : \mathrm{rank}(\mathbf{C}\boldsymbol{\Theta}) = j$, we have an LR statistic given by

$$\Lambda_{(j)} = \{(1 + \ell_{j+1}) \cdots (1 + \ell_m)\}^{-1}, \tag{5}$$

where $\ell_1 > \cdots > \ell_m > 0$ are the non-zero characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$,

$$\mathbf{S}_e = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}, \quad \mathbf{S}_h = (\mathbf{C}\hat{\boldsymbol{\Theta}})'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}\mathbf{C}\hat{\boldsymbol{\Theta}},$$

and $\hat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Here, without loss of generality we may assume that $\ell_1 > \cdots > \ell_m > 0$, since the probability for ℓ_i 's to be equal is 0. It is well known (see, e.g. Anderson(2003)) that $\mathbf{S}_e$ and $\mathbf{S}_h$ are independently distributed as a Wishart distribution $\mathrm{W}_p(n-k, \boldsymbol{\Sigma})$ and a noncentral Wishart distribution $\mathrm{W}_p(q, \boldsymbol{\Sigma}; \boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{1/2})$, respectively, where

$$\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{C}\boldsymbol{\Theta})'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}\mathbf{C}\boldsymbol{\Theta}\boldsymbol{\Sigma}^{-1/2}. \tag{6}$$

Our problem involves the following two special cases. First we consider a multivariate reduced-rank regression model which is given by (3) with (4) and $\mathbf{C} = \mathbf{I}_k$. From the rank constraint (4) the regression matrix $\boldsymbol{\Theta}$ can be expressed as a product of two rank j matrices as follows:

$$\boldsymbol{\Theta} = \mathbf{G}\boldsymbol{\Xi},$$

where $\mathbf{G}$ is of dimension $k \times j$ and $\boldsymbol{\Xi}$ is of dimension $j \times p$. Then

$$\mathrm{E}[\mathbf{Y}] = (\mathbf{X}\mathbf{G}) \cdot \boldsymbol{\Xi} = \mathbf{Z}\boldsymbol{\Xi}, \quad \mathbf{Z} = \mathbf{X}\mathbf{G}. \tag{7}$$

The model means that the j linear combinations $\boldsymbol{z} = \mathbf{G}'\boldsymbol{x}$ of the k explanatory variables $\boldsymbol{x}$ are sufficient to model the validation in the p response variables $\boldsymbol{y}$. The j-variate $\boldsymbol{z}$ may be regarded as a factor or latent variate. In practice, the dimension j is unknown, and we need to estimate it.

Next we consider the reduced-rank problem in discriminant analysis, based on $(q+1)$ p-variate normal populations with common covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\mu}_i$ be the mean vector of the ith population. Suppose that a sample of size n_i is available from the ith population, and let $\boldsymbol{y}_{ij}$ be the jth observation from the ith population. Let us denote the between-group and within-group sums of squares and products matrices by

$$\mathbf{S}_b = \sum_{i=1}^{q+1} n_i(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})', \quad \mathbf{S}_w = \sum_{i=1}^{q+1}(n_i - 1)\mathbf{S}_i,$$

respectively, where $\bar{\boldsymbol{y}}_i$ and $\mathbf{S}_i$ are the mean vector and sample covariance matrix of the ith population, and $\bar{\boldsymbol{y}}$ is the total mean vector defined by $(1/n)\sum_{i=1}^{q+1} n_i\bar{\boldsymbol{y}}_i$, and $n = \sum_{i=1}^{q+1} n_i$. In general, $\mathbf{S}_w$ and $\mathbf{S}_b$ are independently distributed as a Wishart distribution $\mathrm{W}_p(n-q-1, \boldsymbol{\Sigma})$ and a noncentral Wishart distribution $\mathrm{W}_p(q, \boldsymbol{\Sigma}; \boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{1/2})$, respectively, where

$$\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Sigma}^{-1/2}\sum_{i=1}^{q+1} n_i(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1/2}, \quad \bar{\boldsymbol{\mu}} = (1/n)\sum_{i=1}^{q+1} n_i\boldsymbol{\mu}_i. \tag{8}$$

The coefficient vector $\boldsymbol{\beta}_i$ of the i-th population discriminant function is defined as the characteristic vector satisfying

$$\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_i = \omega_i\boldsymbol{\Sigma}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i'\boldsymbol{\Sigma}\boldsymbol{\beta}_j = \delta_{ij}, \quad i,j = 1,\ldots,m = \min(p,q),$$

where δ_{ij} denotes the Kroneker delta. Here, $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_m \geq 0$ are the possible non-zero characteristic roots of $\tilde{\boldsymbol{\Omega}}$. The between-groups variation of the i-th discriminant function $\boldsymbol{\beta}_i'\boldsymbol{X}$ is ω_i. Therefore, if ω_i is zero, the i-th discriminant function $\boldsymbol{\beta}_i'\boldsymbol{X}$ is not meaningful. The dimensionality in discriminant analysis may be defined (see Kishisagar (1972), Fujikoshi, Ulyanov and Shimizu (2010), etc.) as the number of non-zero characteristic roots of $\tilde{\boldsymbol{\Omega}}$ which is the number of meaningful population discriminant functions. The model that the dimension is $j(0 \leq j \leq m)$ may be expressed as

$$\begin{aligned} M_j : &\ \mathrm{rank}(\tilde{\boldsymbol{\Omega}}) = j, \\ &\Leftrightarrow \omega_1 \geq \cdots \geq \omega_j > \omega_{j+1} = \cdots = \omega_q = 0, \\ &\Leftrightarrow \mathrm{rank}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{q+1}, \ldots, \boldsymbol{\mu}_q - \boldsymbol{\mu}_{q+1}) = j. \end{aligned} \tag{9}$$

Note that the model M_j in (4) involves the M_j in discriminant analysis as a special case. This is easily seen by taking $k = q + 1$ and choosing $\mathbf{Y}$, $\mathbf{C}$, $\mathbf{X}$ and $\mathbf{\Theta}$ as follows.

$$\mathbf{Y} = (\boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1n_1}, \ldots, \boldsymbol{y}_{q+1,1}, \ldots, \boldsymbol{y}_{q+1,n_{q+1}})', \quad \mathbf{C} = (\mathbf{I}_q, \ -\mathbf{1}_q)$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{q+1}} \end{pmatrix}, \quad \mathbf{\Theta} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \\ \vdots \\ \boldsymbol{\mu}_{q+1}' \end{pmatrix},$$

where $\mathbf{1}_n$ is an $n \times 1$ vector whose elements are all one. Then, $\mathbf{S}_h = \mathbf{S}_b$ and $\mathbf{S}_e = \mathbf{S}_w$.

2.2 AIC, BIC, C_p and Their Extensions

In general, AIC for a model M is defined (Akaike (1973)) as

$$\mathrm{AIC} = -2 \log \hat{L} + d,$$

where $\hat{L}$ is the maximum likelihood under M, and d is the number of independent parameters under M. The AIC for M_j is expressed as

$$\begin{aligned} \mathrm{AIC}_j =& n \log(1 + \ell_{j+1}) \cdots (1 + \ell_m) + n \log |(1/n)\mathbf{S}_e| \\ &+ np(\log 2\pi + 1) + 2 \left\{ j(p + q - j) + (k - q)p + \frac{1}{2} p(p + 1) \right\}, \quad (10) \end{aligned}$$

The result has been obtained by Fujikoshi and Veitch (1979) as an asymptotic unbiased estimator of the risk function based on Kullback-Leibler distance.

Based on AIC_j, if $\min\{\mathrm{AIC}_0, \mathrm{AIC}_1, \ldots, \mathrm{AIC}_m\} = \mathrm{AIC}_j$, we estimate the rank as j. Instead of AIC_j, we may use

$$\begin{aligned} \mathrm{A}_j =& \mathrm{AIC}_j - \mathrm{AIC}_m \\ =& n \log \prod_{i=j+1}^{m} (1 + \ell_i) - 2(p - j)(q - j), \ j = 0, \ldots, m. \quad (11) \end{aligned}$$

Here $\mathrm{A}_m = 0$. Then the estimation method is equivalent to estimate the rank as j if $\min\{\mathrm{A}_0, \mathrm{A}_1, \ldots, \mathrm{A}_m\} = \mathrm{A}_j$.

Similarly, the B_j and C_j based on BIC and C_p are given as follows.

$$\begin{aligned}\mathrm{B}_j =&\mathrm{BIC}_j - \mathrm{BIC}_m \\ =&n \log \prod_{i=j+1}^{m} (1+\ell_i) - (\log n)(p-j)(q-j),\ j=0,\ldots,m. \qquad (12)\\ \mathrm{C}_j =&\mathrm{C}_{p,j} - \mathrm{C}_{p,m} \\ =&n \sum_{i=j+1}^{m} \ell_i - 2(p-j)(q-j), \quad j=0,\ldots,m. \qquad (13)\end{aligned}$$

Here, $\mathrm{B}_m = 0$, $\mathrm{C}_m = 0$.

Using a tuning parameter ν, we consider the following two extended criteria:

$$\mathrm{IC}_{\nu;j} = n \log \prod_{i=j+1}^{m} (1+\ell_i) - \nu(p-j)(q-j),\ j=0,\ldots,m. \qquad (14)$$

$$\mathrm{C}_{p\nu;j} = n \sum_{i=j+1}^{m} \ell_i - \nu(p-j)(q-j), \quad j=0,\ldots,m. \qquad (15)$$

Here $\mathrm{IC}_{\nu;m} = 0$ and $\mathrm{C}_{p\nu;m} = 0$. Then $\mathrm{IC}_{2;j} = \mathrm{A}_j$, $\mathrm{IC}_{\log n;j} = \mathrm{B}_j$, $\mathrm{C}_{p2;j} = \mathrm{C}_j$.

3 High-Dimensional Consistency

3.1 Theoretical Results

In the following, we are concerned with asymptotic behaviors of the criteria when p and n are large and q is fixed. So, without loss of generality, we assume that $p \geq q$, then $m = \min(p,q) = q$. Denoting M_j by j simply, the set of all the models is $\mathcal{F} = \{0,1,\ldots,q\}$. It is assumed that the true model is the model (3) with $\boldsymbol{\Theta} = \boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*$. However, we often write $\boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma}_*$ as $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ simply. Further, we assume that the minimum reduced rank model including M_* is j_*, where $0 \leq j_* \leq q$. The j_* denotes also the true rank or dimension. We separate $\mathcal{F}$ into two sets, one is a set of overspecified models, i.e., $\mathcal{F}_+ = \{j_*, j_*+1, \ldots, q\}$ and the other is a set of underspecified models, i.e., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F} = \{0,1,\ldots j_*-1\}$. Further, we

denote the set of models deleting the true model from $\mathcal{F}_+$ by $\mathcal{F}_+\backslash\{j_*\}$, i.e., $\mathcal{F}_+\backslash\{j_*\} = \{j_*+1, \ldots, q\}$.

The estimation methods based on A_j, B_j and C_j are expressed as

$$\hat{j}_{\mathrm{A}} = \arg\min_{j\in\mathcal{F}} \mathrm{A}_j,\ \hat{j}_{\mathrm{B}} = \arg\min_{j\in\mathcal{F}} \mathrm{B}_j,\ \text{and}\ \ \hat{j}_{\mathrm{C}} = \arg\min_{j\in\mathcal{F}} \mathrm{C}_j,$$

respectively. Similarly, the estimation methods based on $\mathrm{IC}_{\nu;j}$ and $\mathrm{C}_{p\nu;j}$ are expressed as

$$\hat{j}_{\mathrm{IC}_\nu} = \arg\min_{j\in\mathcal{F}} \mathrm{IC}_{\nu;j},\ \text{and}\ \ \hat{j}_{\mathrm{C}_{p\nu}} = \arg\min_{j\in\mathcal{F}} \mathrm{C}_{p\nu;j},$$

respectively. In this paper we assume that

A1 (The true model): $j_* \in \mathcal{F}$.

A2 (The asymptotic framework): q is fixed, $p\to\infty$, $n\to\infty$, $p/n\to c\in[0,1)$.

Further, we make two types of assumptions on the order of $\tilde{\boldsymbol{\Omega}}$ in (6). Since $\mathrm{rank}(\tilde{\boldsymbol{\Omega}}) \le q$, we can write $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma}$ is a $p\times q$ matrix. Let

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}, \tag{16}$$

which is a $q\times q$ matrix. In discriminant analysis with $q=1$,

$$\tilde{\boldsymbol{\Omega}} = (n_1n_2/n)\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1/2},$$

and

$$\boldsymbol{\Omega} = \omega = \frac{n_1n_2}{n}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2),$$

which is (n_1n_2/n) times the squared Mahalanobis distance between two normal populations $\mathrm{N}_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ and $\mathrm{N}_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$. When $n_i/n\to d_i>0$, $\omega=\mathrm{O}(n)$ and also $\omega=\mathrm{O}(np)$, depending on whether the squared Mahalanobis distance is $\mathrm{O}(1)$ or $\mathrm{O}(p)$, where $\mathrm{O}(\cdot)$ is the usual order under a high-dimensional framework (2). Note that, when we consider the distributions of our criteria, without loss of generality we may assume

$$\boldsymbol{\Omega} = \mathrm{diag}(\omega_1,\ldots,\omega_q), \tag{17}$$

where $\omega_1 \geq \cdots \geq \omega_q$ are the characteristic roots of $\mathbf{\Omega}$ or the non-zero characteristic roots of $\tilde{\mathbf{\Omega}}$. Based on these considerations we take up the following two types of assumptions on the noncentrality matrix.

A3 (The noncentrality matrix-1): For any $j(0 \leq j \leq j_*)$,
$$\omega_j = n\delta_j = \mathrm{O}(n), \quad \lim_{p/n \to c} \delta_j = \delta_j^* > 0.$$

A4 (The noncentrality matrix-2): For any $j(0 \leq j \leq j_*)$,
$$\omega_j = np\xi_j = \mathrm{O}(np), \quad \lim_{p/n \to c} \xi_j = \xi_j^* > 0.$$

In the following, we give sufficient conditions for IC_ν and $\mathrm{C}_{p\nu}$ to be consistent. Here, the consistency of e.g., IC_ν means that the probability that IC_ν selects the true model j_* tends asymptotically to 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{IC}_\nu} = j_*) = 1.$$

Theorem 1 *Suppose that the assumptions* A1 *is satisfied.*

(1) IC_ν *is consistent if the assumptions* A2, A3 *and the inequality "* $-c^{-1}\log(1-c) < \nu < -c^{-1}\log(1-c) + c^{-1}\log(1+\delta_{j_*}^*)$ *" are satisfied.*

(2) IC_ν *is consistent if the assumptions* A2, A4 *and the inequality "* $-c^{-1}\log(1-c) < \nu$ *" are satisfied.*

Theorem 1 implies the following results except for Corollary 2(1).

Corollary 1 *Suppose that the assumptions* A1 *is satisfied. Further, assume that* $c \in [0, c_{\mathrm{a}})$*, where* c_{a} (≈ 0.797) *is the larger constant satisfying* $\log(1-c_{\mathrm{a}}) + 2c_{\mathrm{a}} = 0$.

(1) A *is consistent if the assumptions* A2, A3 *and the inequality "* $\log(1+\delta_{j_*}^*) > (j_* - j)\{2c + \log(1-c)\}$ *are satisfied.*

(2) A *is consistent if the assumptions* A2 *and* A4 *are satisfied.*

Corollary 2 *Suppose that the assumptions* A1 *is satisfied.*

(1) B *is not consistent if the assumptions* A2 *with* $c > 0$ *and* A3 *are satisfied.*

(2) B *is consistent if the assumptions* A2 *and* A4 *are satisfied.*

Similar results are obtained for the criteria $\mathrm{C}_{p\nu}$ and C_p.

Theorem 2 *Suppose that the assumptions* A1 *is satisfied.*

(1) $\mathrm{C}_{p\nu}$ *is consistent if the assumptions* A2, A3 *and the inequality "* $-(1-c)^{-1} < \nu < (1-c)^{-1} + \{c(1-c)\}^{-1}\delta^*_{j_*}$ *" are satisfied.*

(2) $\mathrm{C}_{p\nu}$ *is consistent if the assumptions* A2, A4 *and the inequality "* $-(1-c)^{-1} < \nu$ *" are satisfied.*

Corollary 3 *Suppose that the assumptions* A1 *is satisfied. Further, assume that* $c \in [0, 0.5)$.

(1) C *is consistent if the assumptions* A2, A3 *and the inequality "* $\delta^*_{j_*} > (j_* - j)c(1-2c)$ *" are satisfied.*

(2) C *is consistent if the assumptions* A2 *and* A4 *are satisfied.*

Theorems 1 and 2 are helpful for selection of tuning parameters. Our sufficient conditions for consistency are derived as follows. Let T_j be a general criterion for M_j, $j \in \mathcal{F}$. Then, we attempt to show that

$$^{\forall} j \neq j_* \in \mathcal{F}, \quad \frac{1}{h_{j,j_*}}(T_j - T_{j_*}) \geq D_{j,j_*} \xrightarrow{p} \alpha_{j,j_*} > 0 \tag{18}$$

where D_{j,j_*} is some quantity, and h_{j,j_*} is some positive constant depending on models. It is easy to see that (18) implies $P(\hat{j}_T = j_*) \to 1$.

For example, under A1, A2 and A3 we shall show in Appendix A·1 that

for $j > j_*$;

$$\frac{1}{n}\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\} \xrightarrow{p} (j - j_*)\{\log(1-c) + \nu c\}.$$

for $j < j_*$;

$$\begin{aligned}\frac{1}{n}\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\} &\xrightarrow{p} \log(1+\delta^*_{j+1})\cdots(1+\delta^*_{j_*}) - (j_* - j)\{\log(1-c) + \nu c\} \\ &\geq (j_* - j)[\log(1+\delta^*_{j_*}) - \{\log(1-c) + \nu c\}].\end{aligned}$$

Therefore, by obtaining the ranges of ν such that the above right-hand sides are positive, we can obtain Theorem 1. Generally, we can say that for the consistency, it needs that the penalty term tends to infinity. When $\nu = \log n$, we have that for $j < j_*$

$$\frac{1}{n\log n}(\mathrm{B}_j - \mathrm{B}_{j_*}) \xrightarrow{p} -(j_* - j)c,$$

which implies $P(\mathrm{B}_j > \mathrm{B}_{j_*}) \to 0$. Therefore

$$\begin{aligned}P(\hat{j}_{\mathrm{B}} = j_*) &= P(\mathrm{B}_j > \mathrm{B}_{j_*}, \quad \text{for all } j \neq j_*) \\ &\leq P(\mathrm{B}_j > \mathrm{B}_{j_*}), \quad \text{for } j < j_* \\ &\to 0, \quad \text{for } j < j_*,\end{aligned}$$

which implies Corollary 2(1).

We note that the consistency properties in Theorems 1 and 2 hold case where the set of candidate models is a subfamily $\mathcal{G} \subset \mathcal{F}$. In fact, for example, the estimation method based on AIC is expressed as

$$\hat{j}_{\mathrm{A};\mathcal{G}} = \arg\min_{j\in\mathcal{G}} \mathrm{A}_j.$$

Then, the consistency of $\hat{j}_{\mathrm{A};\mathcal{G}}$ is given as Theorem 1 with the following modifications; $j_* \in \mathcal{F} \to j_* \in \mathcal{G}$, and "any j $(0 \leq j \leq j_*)$" in A2 and A3 $\to$ "any $j(0 \leq j \leq j_*) \in \mathcal{G}$".

3.2 Numerical Study

In this section, we numerically examine the validity of our claims and tendencies for the ranks estimated by A, B, C and their extensions. Our numerical results are given for the estimation problem of dimensionality in discriminant analysis with $q+1$ groups based on the total sample size n of p response variables. Assume that $p \geq q$.

The criteria are based on the nonzero characteristic roots $\ell_1 > \cdots > \ell_q$ of $\mathbf{S}_b\mathbf{S}_w^{-1}$. Without loss of generality we may assume that $\mathbf{S}_w$ and $\mathbf{S}_b$ are independently distributed as $\mathrm{W}_p(n-q-1, \mathbf{I}_p)$ and $\mathrm{W}_p(q, \mathbf{I}_p; \mathbf{\Omega}_p)$, respectively. Here, $\mathbf{\Omega}_p = \mathrm{diag}(\omega_1, \cdots, \omega_q, 0, \ldots, 0)$ and $\omega_1, \cdots, \omega_q$ are the possible nonzero characteristic roots of the noncentrality matrix $\mathbf{\Omega}$ defined by (8). Further, the sample roots $\ell_1 > \cdots > \ell_q$ may be regarded (see Lemma A1) as the ones of $\mathbf{B}\mathbf{W}^{-1}$, where $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as $\mathrm{W}_q(n-p-1, \mathbf{I}_q)$ and $\mathrm{W}_q(p, \mathbf{I}_q; \mathbf{\Omega})$, respectively and $\mathbf{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q)$.

Suppose that $q = 5$, and so we have six candidate models $M_0, M_1, \ldots, M_5$. It is assumed that the minimum model including the true model is M_3 and so $j_* = 3$. The two types of characteristic roots $\omega_i, i = 1, \ldots, 5$ are defined as follows:

$$\text{(a)} : \omega_1 = 2\omega_3, \quad \omega_2 = \frac{3}{2}\omega_3, \quad \omega_3 = n, \quad \omega_4 = \omega_5 = 0,$$
$$\text{(b)} : \omega_1 = 2\omega_3, \quad \omega_2 = \frac{3}{2}\omega_3, \quad \omega_3 = np, \quad \omega_4 = \omega_5 = 0.$$

These (a) and (b) are corresponding to the noncentrality matrix-1 and -2 on the assumptions A3 and A4. Several different values of n and $p = cn$ were prepared for Monte Carlo simulations with 10^4 repetitions. Tables 4.1 and 4.2 show simulation results for

$$(n, p) = (30, 5), \quad (60, 10), \quad (120, 20), \quad (210, 35), \quad (300, 50), \quad (480, 80), \quad (600, 100).$$

In these cases, the values of p/n are all $1/6$, and the assumptions A3 and A4 are satisfied.

Table 4.1. Selection probabilities of $\hat{j}_{\mathrm{A}}$, $\hat{j}_{\mathrm{B}}$ and $\hat{j}_{\mathrm{C}}$ under (a)

	$\hat{j}_{\mathrm{A}}$			$\hat{j}_{\mathrm{B}}$			$\hat{j}_{\mathrm{C}}$		
(n, p)	under	true	over	under	true	over	under	true	over
(30, 5)	0.01	0.76	0.23	0.13	0.82	0.05	0.00	0.72	0.28
(60, 10)	0.00	0.86	0.14	0.18	0.82	0.00	0.00	0.76	0.24
(120, 20)	0.00	0.95	0.05	0.35	0.65	0.00	0.00	0.86	0.15
(210, 35)	0.00	0.99	0.01	0.65	0.35	0.00	0.00	0.95	0.06
(300, 50)	0.00	1.00	0.00	0.87	0.13	0.00	0.00	0.97	0.03
(480, 80)	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.99	0.01
(600, 100)	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00

Table 4.2. Selection probabilities of $\hat{j}_{\mathrm{A}}$, $\hat{j}_{\mathrm{B}}$ and $\hat{j}_{\mathrm{C}}$ under (b)

	$\hat{j}_{\mathrm{A}}$			$\hat{j}_{\mathrm{B}}$			$\hat{j}_{\mathrm{C}}$		
(n, p)	under	true	over	under	true	over	under	true	over
(30,5)	0.00	0.76	0.24	0.00	0.95	0.05	0.00	0.70	0.30
(60, 10)	0.00	0.80	0.20	0.00	1.00	0.00	0.00	0.71	0.29
(120, 20)	0.00	0.93	0.07	0.00	1.00	0.00	0.00	0.83	0.17
(210, 35)	0.00	0.99	0.01	0.00	1.00	0.00	0.00	0.95	0.05
(300, 50)	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.98	0.02
(480, 80)	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
(600, 100)	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00

In Tables 4.1 and 4.2, the values for "under", "true" and "over" denote the probabilities of selecting the underspecified models, the true model and the overspecified models, respectively. From the tables we can see the following tendencies.

- Under the case (a), the selection probabilities of the true model by $\hat{j}_{\mathrm{A}}$ and $\hat{j}_{\mathrm{C}}$ are increasing when (n, p) is increasing , and tend to 1.
- Under the case (a), the selection probabilities of the true model by $\hat{j}_{\mathrm{B}}$ do not increase even when (n, p) is increasing
- Under the case (b), the selection probabilities of the true model by by $\hat{j}_{\mathrm{A}}$, $\hat{j}_{\mathrm{B}}$ and $\hat{j}_{\mathrm{C}}$ are increasing when (n, p) is increasing, and tend to 1.

Next we examined the selection probabilities when n is fixed and p increases as follows:

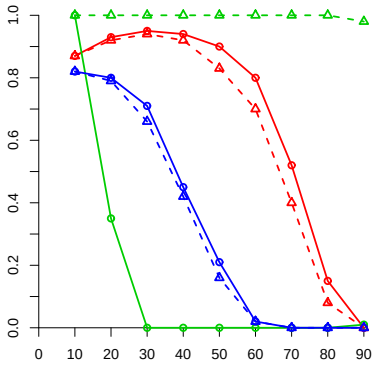
$$n = 100, \quad p = 10, \quad 20, \quad 30, \quad 40, \quad 50, \quad 60, \quad 70, \quad 80, \quad 90.$$

In this case, a range of p/n is $0.1 \sim 0.9$. Table 4.3 gives the selection probabilities of the true model by $\hat{j}_{\mathrm{A}}$, $\hat{j}_{\mathrm{B}}$, and $\hat{j}_{\mathrm{C}}$, and their graph displays. Here the red, the green and the blue denote the probabilities of $\hat{j}_{\mathrm{A}}$, $\hat{j}_{\mathrm{B}}$, and $\hat{j}_{\mathrm{C}}$, respectively. In Figure 4.1, the solid lines and the dotted lines are corresponding to case (a) and case (b), respectively.

Table 4.3. Selection probabilities of the true model

	under (a)			under (b)		
p	$\hat{j}_{\mathrm{A}}$	$\hat{j}_{\mathrm{B}}$	$\hat{j}_{\mathrm{C}}$	$\hat{j}_{\mathrm{A}}$	$\hat{j}_{\mathrm{B}}$	$\hat{j}_{\mathrm{C}}$
10	0.87	1.00	0.82	0.87	1.00	0.82
20	0.93	0.35	0.80	0.92	1.00	0.79
30	0.95	0.00	0.71	0.94	1.00	0.66
40	0.94	0.00	0.45	0.92	1.00	0.42
50	0.90	0.00	0.21	0.83	1.00	0.16
60	0.80	0.00	0.02	0.70	1.00	0.02
70	0.52	0.00	0.00	0.40	1.00	0.00
80	0.15	0.00	0.00	0.08	1.00	0.00
90	0.00	0.01	0.00	0.00	0.98	0.00

Figure 4.1. Graph displays of Table 4.3



From Table 4.3 we can see the following tendencies:

- The probabilities of $\hat{j}_{\mathrm{A}}$ in both case (a) and case (b) are increasing for $10 \leq p \leq 30$ and decreasing for $30 \leq p \leq 90$, taking the maximum at $p = 30$.
- The probabilities of $\hat{j}_{\mathrm{C}}$ in both case (a) and case (b) take the maximum at $p = 10$ and then are decreasing.
- For case (a) and case (b), the probabilities of $\hat{j}_{\mathrm{A}}$ and $\hat{j}_{\mathrm{C}}$ are near 0 for $p > 0.797 (\approx c_a)$ and $p > 0.5$, respectively.
- For case (a), the probabilities of $\hat{j}_{\mathrm{B}}$ are 1 whem $p = 10$, but zero when $30 \leq p \leq 90$.

- For case (b), the probabilities of $\hat{j}_{\mathrm{B}}$ are 1 for all $p(10 \leq p \leq 90)$.

Next we examined a range of tuning parameters ν such $\hat{j}_{\mathrm{IC}_\nu}$ and $\hat{j}_{\mathrm{C}_{p\nu}}$ are consistent. The experiment was done for $p/n = 0.9$, $n = 1000$ and $p = 900$.

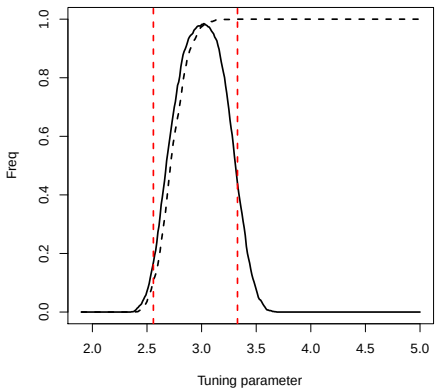


Figure 4.2. Selection probabilities of $\hat{j}_{\mathrm{IC}_\nu}$

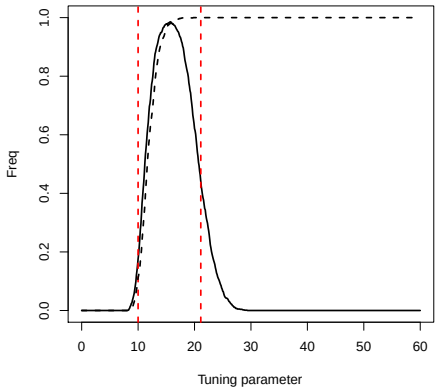


Figure 4.3. Selection probabilities of $\hat{j}_{\mathrm{C}_{p\nu}}$

The numerical results are given in Figures 4.2, 4.3 whose horizontal axis and vertical axis show the values of ν and the selection probabilities of the true model, respectively. The solid lines and the dotted lines correspond to case (a) and case (b). In Figure 4.2, the left dotted vertical line denotes $\nu = -\frac{1}{c}\log(1-c)$, and the right dotted vertical line denotes $-\frac{1}{c}\log(1-c) + \frac{1}{c}\log(1+\delta^*_{j_*})$. In Figure 4.3, the left dotted vertical line denotes $\nu = \frac{1}{1-c}$, and the right dotted vertical line denotes $\frac{1}{1-c} + \frac{1}{c(1-c)}\delta^*_{j_*}$. From these figures we can see the following tendencies:

- For case (a), $\hat{j}_{\mathrm{IC}_\nu}$ shall be consistent when
$$-\frac{1}{c}\log(1-c) < \nu < -\frac{1}{c}\log(1-c) + \log(1+\delta^*_{j_*}).$$
- For case (b), $\hat{j}_{\mathrm{IC}_\nu}$ shall be consistent when $-\frac{1}{c}\log(1-c) < \nu$.
- For case (a), $\hat{j}_{\mathrm{C}_{p\nu}}$ shall be consistent when
$$\frac{1}{1-c} < \nu < \frac{1}{1-c} + \frac{1}{c(1-c)}\delta^*_{j_*}.$$

- For case (b), $\hat{j}_{\mathrm{C}_{p\nu}}$ shall be consistent when $\frac{1}{1-c} < \nu$.

4 Ridge-type Criteria and Their Properties

4.1 Ridge-Type Criteria

When $p > n - k$, $\mathbf{S}_e$ becomes singular, and so we can not use the criteria AIC, BIC and C_p. One way to overcome such an issue is to use a ridge-type estimator as an estimator of $\mathbf{\Sigma}$ defined by

$$\tilde{\mathbf{\Sigma}}_\lambda = \frac{1}{n}(\mathbf{S}_e + \lambda \mathbf{I}_p) = \frac{1}{n}\mathbf{S}_{e,\lambda}, \tag{19}$$

or its modifications. Here, λ is an estimator of ridge parameter λ_0, and in this paper we use $\lambda = \{1/(np)\}\mathrm{tr}\mathbf{S}_e$. For a discussion on the use of λ, see Kubokawa and Srivastava (2012). Let $\tilde{\ell}_1 > \cdots > \tilde{\ell}_q$ be the non-zero characteristic roots of $\mathbf{S}_h\mathbf{S}_{e,\lambda}^{-1}$. Then, we propose the following modifications of A, B and C:
For $j = 0, \ldots, q$,

$$\begin{aligned}
\tilde{\mathrm{A}}_{\lambda,j} =& n \log \prod_{i=j+1}^{q} (1 + \tilde{\ell}_i) - 2(p-j)(q-j), \\
\tilde{\mathrm{B}}_{\lambda,j} =& n \log \prod_{i=j+1}^{q} (1 + \tilde{\ell}_i) - (\log n)(p-j)(q-j), \\
\tilde{\mathrm{C}}_{\lambda,j} =& n \sum_{i=j+1}^{q} \tilde{\ell}_i - 2(p-j)(q-j).
\end{aligned} \tag{20}$$

Here, $\tilde{\mathrm{A}}_{\lambda,q} = 0$, $\tilde{\mathrm{B}}_{\lambda,q} = 0$ and $\tilde{\mathrm{C}}_{\lambda,q} = 0$. The criteria $\tilde{\mathrm{A}}_\lambda$, $\tilde{\mathrm{B}}_\lambda$ and $\tilde{\mathrm{C}}_\lambda$ are obtained from A, B and C by substituting $\tilde{\ell}_j$ to ℓ_j.

4.2 Justifications of Ridge-Type Criteria

In this subsection we give justifications for $\tilde{\mathrm{A}}_\lambda$ and $\tilde{\mathrm{C}}_\lambda$ by deriving asymptotic unbiased estimators of the AIC-type or the C_p-type risks based on ridge-type estimators. For a notational simplicity, we may start a canonical form given

as follows. Let $\mathbf{Y}$ be an $n \times p$ random matrix whose rows are independently as $\mathrm{N}_p(\cdot, \mathbf{\Sigma})$ and

$$\mathrm{E}(\mathbf{Y}) = \mathrm{E}\{\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{pmatrix}\} = \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \mathbf{O} \end{pmatrix}, \quad \mathbf{Y}_1; q \times p,\ \mathbf{Y}_2; r \times p,\ \mathbf{Y}_3; (n-k) \times p, \tag{21}$$

where $\boldsymbol{\eta}_1; q \times p,\ \boldsymbol{\eta}_2; r \times p,\ \boldsymbol{\eta}_3; (n-k) \times p$ and $r = k - q$. Then the model M_j is expressed as $\mathrm{rank}(\boldsymbol{\eta}_1) = j$. It holds that

$$\mathbf{Y}_1'\mathbf{Y}_1 = \mathbf{S}_h, \quad \mathbf{Y}_3'\mathbf{Y}_3 = \mathbf{S}_e, \quad \boldsymbol{\eta}_1'\boldsymbol{\eta}_1 = \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2},$$

where $\tilde{\mathbf{\Omega}}$ is the noncentrality matrix defined in (6). Let the density function of $\mathbf{Y}$ denote by $f(\mathbf{Y}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \mathbf{\Sigma})$ which is expressed as

$$\begin{aligned} -2\log f(\mathbf{Y}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \mathbf{\Sigma}) &= n\log|\mathbf{\Sigma}| + np\log(2\pi) \\ &\quad + \mathrm{tr}\mathbf{\Sigma}^{-1}\left\{(\mathbf{Y}_1 - \boldsymbol{\eta}_1)'(\mathbf{Y}_1 - \boldsymbol{\eta}_1) + (\mathbf{Y}_2 - \boldsymbol{\eta}_2)'(\mathbf{Y}_2 - \boldsymbol{\eta}_2) + \mathbf{Y}_3'\mathbf{Y}_3\right\}. \end{aligned}$$

The maximum likelihood estimators of $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ and $\mathbf{\Sigma}$ are obtained by first maximizing with respect to $\mathbf{\Sigma}$, and then maximizing with respect to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. For the second maximization, we use a singular value decomposition of $\mathbf{S}_e^{-1/2}\mathbf{Y}_1'$ denoted by

$$\mathbf{S}_e^{-1/2}\mathbf{Y}_1' = \sum_{i=1}^{q} \sqrt{\ell_i}\boldsymbol{a}_i\boldsymbol{b}_i', \tag{22}$$

where $\boldsymbol{a}_i's$ are the orthonormal characteristic vectors of $\mathbf{S}_e^{-1/2}\mathbf{S}_h\mathbf{S}_e^{-1/2}$ and $\boldsymbol{b}_i's$ are the orthonormal characteristic vectors of $\mathbf{Y}_1\mathbf{S}_e^{-1/2}\mathbf{Y}_1'$. Then, the maximum likelihood estimators are expressed as

$$\begin{aligned} \hat{\boldsymbol{\eta}}_1 &= \left(\sum_{i=1}^{j} \sqrt{\ell_i}\boldsymbol{b}_i\boldsymbol{a}_i'\right)\mathbf{S}_e^{1/2}, \quad \hat{\boldsymbol{\eta}}_2 = \mathbf{Y}_2, \\ n\hat{\mathbf{\Sigma}} &= \mathbf{S}_e^{1/2}\left(\mathbf{I}_p + \sum_{i=j+1}^{q} \ell_i\boldsymbol{a}_i\boldsymbol{a}_i'\right)\mathbf{S}_e^{1/2}. \end{aligned}$$

Now we consider the ridge-type estimators $\hat{\boldsymbol{\eta}}_{1,\lambda}$, $\hat{\boldsymbol{\eta}}_{2,\lambda}$ and $\hat{\mathbf{\Sigma}}_\lambda$ defined from $\hat{\boldsymbol{\eta}}_1$, $\hat{\boldsymbol{\eta}}_2$ and $\hat{\mathbf{\Sigma}}$ by substituting $\mathbf{S}_{e,\lambda}$ to $\mathbf{S}_e$. The AIC-type risk of a candidate

model M_j based on the ridge-type estimator is

$$R_{A,\lambda} = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Z}}\left\{-2\log f(\mathbf{Z};\hat{\boldsymbol{\eta}}_{1,\lambda},\hat{\boldsymbol{\eta}}_{2,\lambda},\hat{\boldsymbol{\Sigma}}_{\lambda})\right\}, \tag{23}$$

which is based on Kullback-Leibler information. Here $\mathbf{Z}; n \times p$ may be regarded a future random matrix that has the same distribution as $\mathbf{Y}$ and is independent of $\mathbf{Y}$, and E^* denotes the expectation to the true model M_*. When we estimate $R_{A,\lambda}$ by

$$\begin{aligned}-2\log f(\mathbf{Y};&\hat{\boldsymbol{\eta}}_{1,\lambda},\hat{\boldsymbol{\eta}}_{2,\lambda},\hat{\boldsymbol{\Sigma}}_{\lambda})\\ &= n\log\prod_{i=j+1}^{q}(1+\tilde{\ell}_i) + n\log|(1/n)\mathbf{S}_{e,\lambda}| + np\{1+\log(2\pi)\},\end{aligned} \tag{24}$$

where $\tilde{\ell}_1 > \cdots > \tilde{\ell}_q > 0$ are the non-zero characteristic roots of $\mathbf{S}_h\mathbf{S}^{-1}_{e,\lambda}$, then, the bias term is expressed as "$-b_{A,\lambda}$", where

$$\begin{aligned}b_{A,\lambda} &= \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Z}}\left\{-2\log f(\mathbf{Z};\hat{\boldsymbol{\eta}}_{1,\lambda},\hat{\boldsymbol{\eta}}_{2,\lambda},\hat{\boldsymbol{\Sigma}}_{\lambda}) + \log f(\mathbf{Y};\hat{\boldsymbol{\eta}}_{1,\lambda},\hat{\boldsymbol{\eta}}_{2,\lambda},\hat{\boldsymbol{\Sigma}}_{\lambda})\right\}\\ &= \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Z}}\left\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}_{\lambda}(\mathbf{Z}-\boldsymbol{\eta})'(\mathbf{Z}-\boldsymbol{\eta})\right\} - np.\end{aligned} \tag{25}$$

Here, for a notational simplicity, we express the true parameters as the ones in (21). Taking the expectations with respect to $\mathbf{Z}$ and $\mathbf{Y}_2$, the bias term can be written as

$$b_{A,\lambda} = (n+k-q)b^{(1)}_{A,\lambda} + b^{(2)}_{A,\lambda} - np, \tag{26}$$

where

$$b^{(1)}_{A,\lambda} = \mathrm{E}^*_{\mathbf{Y}}\left(\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}_{\lambda}\boldsymbol{\Sigma}\right), \quad b^{(2)}_{A,\lambda} = \mathrm{E}^*_{\mathbf{Y}}\left\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}_{\lambda}(\boldsymbol{\eta}_1-\hat{\boldsymbol{\eta}}_1)'(\boldsymbol{\eta}_1-\hat{\boldsymbol{\eta}}_1)\right\}. \tag{27}$$

The terms $b^{(1)}_{A,\lambda}$ and $b^{(2)}_{A,\lambda}$ are asymptotically evaluated in Appendix under a large-sample asymptotic framework and $\tilde{\boldsymbol{\Omega}} = \mathrm{O}(n)$. We have

$$b_{A,\lambda} = 2\left\{j(p+q-j) + (k-q)p + \frac{1}{2}p(p+1) - \frac{1}{2p}\mathrm{tr}\boldsymbol{\Sigma}\right\} + \mathrm{o}(1). \tag{28}$$

This suggests that $\tilde{\mathrm{A}}_{\lambda}$ in (20) is an asymptotic unbiased estimator for $R_{A,\lambda}$.

Next we give justifications for $\tilde{\mathrm{C}}_\lambda$. Estimating $\mathbf{\Sigma}$ by a ridge-type estimator $\tilde{\mathbf{\Sigma}}_\lambda = (1/n)\mathbf{S}_{e,\lambda}$, we have the likelihood $f(\mathbf{Y}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \tilde{\mathbf{\Sigma}}_\lambda)$, which is maximized at $\boldsymbol{\eta}_1 = \hat{\boldsymbol{\eta}}_{1,\lambda}$ and $\boldsymbol{\eta}_2 = \hat{\boldsymbol{\eta}}_{2,\lambda}$. This implies that

$$\begin{aligned} -2\log f(\mathbf{Y};&\hat{\boldsymbol{\eta}}_{1,\lambda}, \hat{\boldsymbol{\eta}}_{2,\lambda}, \tilde{\mathbf{\Sigma}}_\lambda) \\ &= n\sum_{i=j+1}^{q} \tilde{\ell}_i + \mathrm{tr}\tilde{\mathbf{\Sigma}}_\lambda^{-1}\mathbf{S}_e + n\log|\tilde{\mathbf{\Sigma}}_\lambda| + np\{1+\log(2\pi)\}. \end{aligned} \tag{29}$$

Consider a different AIC-type risk $\tilde{R}_{A,\lambda}$ obtained from $R_{A,\lambda}$ by substituting $\tilde{\mathbf{\Sigma}}_\lambda$ to $\hat{\mathbf{\Sigma}}_\lambda$. Then, similarly it is shown that $\tilde{\mathrm{C}}_\lambda$ is an asymptotic unbiased estimator of $\tilde{R}_{A,\lambda}$. Another justification cab be obtained by considering a ridge-type C_p risk defined by

$$R_{C,\lambda} = \mathrm{E}^*_{\mathbf{Y}}\mathrm{E}^*_{\mathbf{Z}}\left\{\mathrm{tr}\tilde{\mathbf{\Sigma}}_\lambda^{-1}(\mathbf{Z}-\hat{\boldsymbol{\eta}})'(\mathbf{Z}-\hat{\boldsymbol{\eta}})\right\}, \tag{30}$$

and by deriving its asymptotic unbiased estimator under a large-sample framework and $\tilde{\mathbf{\Omega}} = \mathrm{O}(n)$.

4.3 Consistency of Ridge-Type Criteria

In this section we examine consistency of ridge-type criteria when $n-k > p$. More precisely, it is shown that the criteria $\tilde{\mathrm{A}}_\lambda$, $\tilde{\mathrm{B}}_\lambda$ and $\tilde{\mathrm{C}}_\lambda$ have the same consistency properties as the citeria A, B and C, respectively. The results are stated as follows:

Theorem 3 *Suppose that the assumptions* A1 *is satisfied, and* $(1/p)\mathrm{tr}\mathbf{\Sigma} \to \alpha_0$. *Then, we have the following results.*

(1) $\tilde{\mathrm{A}}_\lambda$ *is consistent if* $c \in [0, c_{\mathrm{a}})$, *the assumptions* A2, A3 *and the inequality* "$\log(1+\delta^*_{j_*}) > (j_*-j)\{2c+\log(1-c)\}$ *are satisfied.*

(2) $\tilde{\mathrm{A}}_\lambda$ *is consistent if* $c \in [0, c_{\mathrm{a}})$ *and the assumptions* A2 *and* A4.

(3) $\tilde{\mathrm{B}}_\lambda$ *is not consistent if the assumptions* A2 *and* A3 *are satisfied.*

(4) $\tilde{\mathrm{B}}_\lambda$ *is consistent if the assumptions* A2 *and* A4 *are satisfied.*

(5) $\tilde{\mathrm{C}}_{\lambda,p}$ *is consistent if* $c \in [0, 0.5)$*, the assumptions* A2 *and* A3*, and the inequality "* $\delta^*_{j_*} > (j_* - j)c(1-2c)$ *" are satisfied*

(6) $\tilde{\mathrm{C}}_{\lambda,p}$ *is consistent if* $c \in [0, 0.5)$*, the assumptions* A2 *and* A4 *are satisfied.*

Theorem 3 is shown by noting that the limiting values of $\tilde{\ell}_i, i = 1, \ldots, q$ are the same as the ones of $\ell_i, i = 1, \ldots, q$. For the proof, see Lemma A2 in Appendix. It is possible to generalize the Theorem for a generalized criterion with a tuning parameter.

4.4 Numerical Study

In this section we consider the selection probabilities of ridge-type criteria $\tilde{\mathrm{A}}$, $\tilde{B}$ and $\tilde{\mathrm{C}}$ under case (a) and case (b) considered in Section 3.2. First, simulations were done $\mathbf{\Sigma} = (0.8^{|i-j|})$, $n = 100$ and $p = 10, 20, \ldots, 90$. The results are given in Figures 4.4 and 4.5. The horizontal axis and the vertical axis show the values of p and the selection probabilities of the true model, respectively. Those color-codes are the same as before.

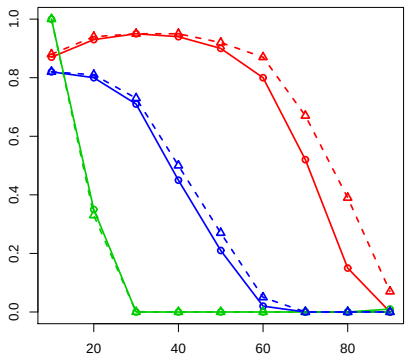


Figure 4.4. Selection probabilities of the true rank by $\tilde{\mathrm{A}}$, A, $\tilde{\mathrm{B}}$, B, $\tilde{\mathrm{C}}$ and C for case (a)

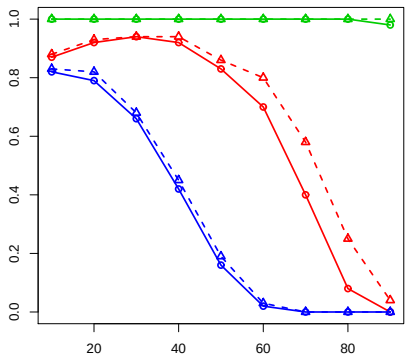


Figure 4.5. Selection probabilities of the true rank by $\tilde{\mathrm{A}}$, A, $\tilde{\mathrm{B}}$, B, $\tilde{\mathrm{C}}$ and C for case (b)

Based on Figures 4.4 and 4.5, it may be pointed that when $n-q-1 \geq p$, the selection probabilities of the true model by A, B and C are similar with the ones by their ridge-type criteria $\tilde{\mathrm{A}}$, $\tilde{\mathrm{B}}$ and $\tilde{\mathrm{C}}$, respectively.

Next, in order to examine behaviors of ridge-type criteria $\tilde{\mathrm{A}}$, $\tilde{\mathrm{B}}$ and $\tilde{\mathrm{C}}$ when $n-q-1 < p$, we did simulation experiments for $\mathbf{\Sigma} = (0.8^{|i-j|})$, $n = 100$ and $p = 100, 200, \ldots, 1000$. For case (a), the selection probabilities of the true model were 0 for all p and all the criteria. The numerical results for case (b) are given in Table 4.4, which implies the following tendencies.

- The selection probabilities of the true model by $\hat{j}_{\mathrm{A}}$ become 1 wwhen $p = 300, 400, 500$.
- $\hat{j}_{\mathrm{A}}$ chooses overspecified models when p is near n (less than 200), and chooses underspecified models when p is larger than 600.
- $\hat{j}_{\mathrm{B}}$ chooses the true model when p is near n, and chooses underspecified models as n is large.
- $\hat{j}_{\mathrm{C}}$ chooses overspecified models for all p such that $n-q-1 < p$.

Table 4.4. Selection probabilities of the true model under (b)

	$\hat{j}_{\mathrm{A}}$			$\hat{j}_{\mathrm{B}}$			$\hat{j}_{\mathrm{C}}$		
p	under	true	over	under	true	over	under	true	over
100	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
200	0.00	0.00	1.00	0.14	0.86	0.00	0.00	0.00	1.00
300	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
400	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
500	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
600	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
700	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
800	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
900	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
1000	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00

5 Concluding Remarks

In general, it is known that under the large-sample asymptotic framework (1), AIC and C_p are not consistent, but BIC is consistent, in the sense that the probabilities of selecting the true model do not approach to one. However, in this paper, we demonstrated that the AIC and C_p for estimating the rank (dimensionality) in multivariate linear model have a consistency property, under a high-dimensional asymptotic framework (2). For the consistency, it is required to satisfy some additional assumptions. For AIC, it needs that $c \in [0, c_a)$, where $c_a \approx 0.797$. For C_p, it needs that $c \in [0, 0.5)$. More precisely, the consistency was considered under two types of assumptions on the largeness of the characteristic roots of the noncentrality matrix $\mathbf{\Omega}$ in (16). For BIC, we note that it is consistent when $\mathbf{\Omega} = O(np)$, but it is not consistent $\mathbf{\Omega} = O(n)$. These results were extended for the criteria IC_ν and $C_{p\nu}$ with a tuning parameter in (14) and (15). We gave sufficient conditions for IC_ν and $C_{p\nu}$ to be consistent. The sufficient conditions are useful in selection of the tuning parameter ν. We proposed ridge-type criteria $\tilde{A}_\lambda$, $\tilde{B}_\lambda$ and $\tilde{C}_\lambda$ in (20) which are also defined for the case where $p > n - k$. It was shown that $\tilde{A}_\lambda$ and $\tilde{C}_\lambda$ are asymptotic unbiased estimators of AIC-type and C_p-type, respectively, under a large-sample framework. Further, these ridge-type criteria have the same consistency properties as A, B and C, respectively.

In discriminant analysis the number of groups may be not large. However, in multivariate regression model the number k of explanatory variables may be large. For such cases, it will occur that n, p and k are large. In Appendix A·1, we give the limiting behavior of the characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$ under a high-dimensional asymptotic framework (A1). It is lest to study asymptotic properties of the methods based on AIC and C_p, etc. under (A1) or a generalization of (A1).

Recently the methods based on the penalization technique have been proposed by Yuan, Ekici, Lu and Monteiro (2007), Bunea, She and Wegkamp (2011), etc., Chen and Hung (2012) and Bunea, She and Wegkamp (2012)

also consider simultaneous methods for dimension reduction and variable selection. It is hoped to combine the penalization techniques and the model selection methods.

Appendix

A.1 The proofs of Theorems 1 and 2

First we prepare a lemma on the limiting behavior of the characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$ in a high-dimensional case.

Lemma 1 *Let $\mathbf{S}_e$ and $\mathbf{S}_h$ be independently distributed as a Wishart distribution $\mathrm{W}_p(n-k,\mathbf{\Sigma})$ and a noncentral Wishart distribution $\mathrm{W}_p(q,\mathbf{\Sigma};\mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively. Here it is assumed that $n-k \geq p$. Let $\ell_1 > \cdots > \ell_q$ and $\omega_1 \geq \cdots \geq \omega_q$ be the possible nonzero characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$ and $\tilde{\mathbf{\Omega}}$, respectively. We assume that $\mathrm{rank}(\tilde{\mathbf{\Omega}}) = a$, and hence $\omega_1 \geq \cdots \geq \omega_a > \omega_{a+1} = \cdots = \omega_q = 0$. For the limiting behavior of $\ell_1 > \cdots > \ell_q$ under a high-dimensional asymptotic framework*

$$p \to \infty, \quad n \to \infty, \quad k \to \infty, \quad p/n \to c \in [0,1), \quad k/n \to 0. \tag{31}$$

we have the following results:

(1) *Suppose that for any $j(0 \leq j \leq a)$, $\omega_j = n\delta_j = \mathrm{O}(n)$ and $\lim_{p/n\to c} \delta_j = \delta_j^* > 0$. Then*

$$\ell_j \xrightarrow{p} \frac{c}{1-c} + \frac{1}{1-c}\delta_j^*, \quad j = 1,\ldots,a, \text{ and } \ell_j \xrightarrow{p} \frac{c}{1-c}, \quad j = a+1,\ldots,q.$$

(2) *Suppose that for any $j(0 \leq j \leq a)$, $\omega_j = np\xi_j = \mathrm{O}(np)$, and $\lim_{p/n\to c} \xi_j = \xi_j^* > 0$. Then*

$$\frac{1}{p}\ell_j \xrightarrow{p} \frac{1}{1-c}\xi_j^*, \quad j = 1,\ldots,a, \text{ and } \ell_j \xrightarrow{p} \frac{c}{1-c}, \quad j = a+1,\ldots,q.$$

Proof 1 *It is known (see Fujikoshi et al. (2010)) that the nonzero characteristic roots* $\ell_1 > \cdots > \ell_q > 0$ *of* $\mathbf{S}_h\mathbf{S}_e^{-1}$ *may be regarded as the ones of* $\mathbf{B}\mathbf{W}^{-1}$*, where* $\mathbf{W}$ *and* $\mathbf{B}$ *are independently distributed as a central Wishart distribution* $\mathrm{W}_q(m, \mathbf{I}_q)$ *and a noncentral Wishart distribution* $\mathrm{W}_q(p, \mathbf{I}_q; \mathbf{\Omega})$*, respectively. Here,* $m = n - k - p + q$*,* $\mathbf{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q)$*, and*

$$\omega_1 \geq \cdots \geq \omega_a > \omega_{a+1} = \cdots = \omega_q = 0.$$

In general, letting

$$\mathbf{U} = \frac{1}{\sqrt{p}}(\mathbf{B} - p\mathbf{I}_q - \mathbf{\Omega}), \quad \mathbf{V} = \frac{1}{\sqrt{m}}(\mathbf{W} - m\mathbf{I}_q),$$

the limiting distributions of $\mathbf{U}$ *and* $\mathbf{V}$ *are normal. When* $\mathbf{\Omega} = n\mathbf{\Delta} = \mathrm{O}(n)$ *and* $\mathbf{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_a, 0, \cdots, 0)$*, we have*

$$\frac{1}{p}\mathbf{B} = \mathbf{I}_q + \frac{n}{p}\mathbf{\Delta} + \mathbf{I}_q + \frac{1}{\sqrt{p}}\mathbf{U}, \quad \frac{1}{m}\mathbf{W} = \mathbf{I}_q + \frac{1}{\sqrt{m}}\mathbf{V}.$$

This implies that the characteristic roots of $\mathbf{B}\mathbf{W}^{-1}$ *are the same as the ones of*

$$\begin{aligned}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} &= \frac{p}{m}\left(\frac{1}{m}\mathbf{W}\right)^{-1/2}\left(\frac{1}{p}\mathbf{B}\right)\left(\frac{1}{m}\mathbf{W}\right)^{-1/2} \\ &\xrightarrow{p} \left(\mathbf{I}_q + \frac{1}{c}\mathbf{\Delta}^*\right)\frac{c}{1-c},\end{aligned}$$

where $\lim \mathbf{\Delta} = \mathbf{\Delta}^*$*, and* $\mathbf{\Delta}^* = \mathrm{diag}(\delta_1^*, \ldots, \delta_a^*, 0, \ldots, 0)$*. This shows the first result (1).*

Next we consider the case $\omega_j = \mathrm{O}(np) = np\xi_j, j = 1, \ldots, a$*. We have*

$$\begin{aligned}\frac{1}{np}\mathbf{B} &= \mathbf{\Xi} + \frac{1}{n}\mathbf{I}_q + \frac{1}{n\sqrt{p}}\mathbf{U}, \\ \left(\frac{1}{m}\mathbf{W}\right)^{-1/2} &= \mathbf{I}_q - \frac{1}{2\sqrt{m}}\mathbf{V} + \frac{3}{8m}\mathbf{V}^2 + \mathrm{O}(m^{-3/2}),\end{aligned}$$

where $\mathbf{\Xi} = \mathrm{diag}(\xi_1, \ldots, \xi_a, 0, \ldots, 0)$*. Therefore*

$$\begin{aligned}\frac{m}{np}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} &= \left(\frac{1}{m}\mathbf{W}\right)^{-1/2}\left(\frac{1}{np}\mathbf{B}\right)\left(\frac{1}{m}\mathbf{W}\right)^{-1/2} \\ &= \begin{pmatrix} \mathbf{\Xi}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} + \frac{1}{\sqrt{m}}\begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}, \qquad (32)\end{aligned}$$

where $\mathbf{\Xi}_1 = \mathrm{diag}(\xi_1, \ldots, \xi_a)$,

$$\mathbf{Q}_{11} = -\frac{1}{2}(\mathbf{V}_{11}\mathbf{\Xi}_1 + \mathbf{\Xi}_1\mathbf{V}_{11}) + \mathrm{O}(m^{-1/2}),$$
$$\mathbf{Q}_{12} = \mathbf{Q}_{21}' = -\frac{1}{2}\mathbf{\Xi}_1\mathbf{V}_{11} + \mathrm{O}(m^{-1/2}),$$
$$\mathbf{Q}_{22} = \frac{\sqrt{m}}{n}\mathbf{I}_{q-a} + \frac{1}{4\sqrt{m}}\mathbf{V}_{21}\mathbf{\Xi}_1\mathbf{V}_{12} + \mathrm{O}(m^{-1}),$$

and

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}, \quad \mathbf{V}_{12}; a \times (q-a).$$

From (32) it is easy to see that

$$\frac{1}{p}(\ell_1, \ldots, \ell_a) \xrightarrow{p} \frac{1}{1-c}(\xi_1^*, \ldots, \xi_a^*).$$

Further, applying Lawley (1959) to (32), the larst $q-a$ *characteristic roots* $\{m/(np)\}(\ell_{a+1}, \ldots, \ell_q)$ *are the same as the ones of*

$$\frac{1}{\sqrt{m}}\mathbf{Q}_{22} - \frac{1}{m}\mathbf{Q}_{21}\mathbf{\Xi}_1\mathbf{Q}_{12} + \mathrm{O}(m^{-/2}) = \frac{1}{n}\mathbf{I}_{q-a} + \mathrm{O}(m^{-/2}).$$

This shows that $\ell_j \xrightarrow{p} c/(1-c)$ *for* $j = a+1, \ldots, q$.

The Proof of Theorem 1

In the proof of Theorem 1 it is assumed that the true dimensionality is j_*. Since the number of possible models is finite, it is sufficient to show that the values of $\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}$ converges to positive values.

Note that for $j > j_*$

$$\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*} = -n\log\{(1+\ell_{j_*+1})\cdots(1+\ell_j)\} + \nu(j-j_*)(p+q-j-j_*),$$

and for $j < j_*$

$$\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*} = n\log\{(1+\ell_{j+1})\cdots(1+\ell_{j_*})\} + \nu(j-j_*)(p+q-j-j_*).$$

Suppose that $\mathbf{\Omega} = \mathrm{O}(n)$. Then, using Lemma A1 (1), we have that for $j > j_*$

$$\frac{1}{n}\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\} \xrightarrow{p} (j-j_*)\{\log(1-c) + \nu c\}.$$

The limiting value is positive when $-\frac{1}{c}\log(1-c) < \nu$. Next suppose that $j < j_*$. Then, using Lemma A1 (1), we have

$$\begin{aligned}\frac{1}{n}\left\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\right\} &\xrightarrow{p} \log(1+\delta^*_{j+1})\cdots(1+\delta^*_{j_*}) - (j_*-j)\{\log(1-c)+\nu c\} \\ &\geq \log(1+\delta^*_{j_*})\cdots(1+\delta^*_{j_*}) - (j_*-j)\{\log(1-c)+\nu c\} \\ &= (j_*-j)[\log(1+\delta^*_{j_*}) - \{\log(1-c)+\nu c\}].\end{aligned}$$

The limiting value is positive when

$$\nu < -\frac{1}{c}\log(1-c) + \frac{1}{c}\log(1+\delta^*_{j_*}).$$

Now we shall prove the result (2). For $j > j_*$, the limiting behavior of ℓ_j under $\omega_j = \mathrm{O}(np)$ is the same as the one under $\omega_j = \mathrm{O}(n)$. Therefore, the limiting value of $(1/n)\left\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\right\}$ is positive when $-\frac{1}{c}\log(1-c) < \nu$, from Lemma A1 (2) we have

$$\frac{1}{np}\left\{\mathrm{IC}_{\nu;j} - \mathrm{IC}_{\nu;j_*}\right\} \xrightarrow{p} j_* - j.$$

This proves Theorem 1 (2).

The Proof of Theorem 2

Theorem 2 is proved by the same way as in Theorem 1, due to Lemma A1. In the following we give an outline of the proof. We have that for $j > j_*$

$$\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*} = -n(\ell_{j_*+1} + \cdots + \ell_j) + \nu(j-j_*)(p+q-j-j_*),$$

and for $j < j_*$

$$\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*} = n(\ell_{j+1} + \cdots + \ell_{j_*}) + \nu(j-j_*)(p+q-j-j_*).$$

Suppose that $\boldsymbol{\Omega} = \mathrm{O}(n)$. Then, using Lemma 1 (1), we have that for $j > j_*$

$$\frac{1}{n}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} \xrightarrow{p} (j-j_*)c\left\{-\frac{1}{1-c} + \nu\right\}.$$

The limiting value is positive when $-\frac{1}{1-c} < \nu$. Next suppose that $j < j_*$. Then, using Lemma 1 (1), we have

$$\begin{aligned}\frac{1}{n}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} &\xrightarrow{p} \frac{c}{1-c}(j_* - j) + \frac{1}{1-c}(\delta^*_{j+1} + \cdots + \delta^*_{j_*}) + \nu(j - j_*)c \\ &\geq \frac{c}{1-c}(j_* - j) + \frac{(j_* - j)}{1-c}\delta^*_{j_*} + \nu(j - j_*)c \\ &= (j_* - j)\left\{\frac{c}{1-c} + \frac{1}{1-c}\delta^*_{j_*} - \nu c\right\}.\end{aligned}$$

The limiting value is positive when

$$\nu < \frac{1}{1-c} + \frac{1}{c(1-c)}\delta^*_{j_*}.$$

Now we shall prove the result (2). For $j > j_*$, the limiting behavior of ℓ_j under $\omega_j = \mathrm{O}(np)$ is the same as the one under $\omega_j = \mathrm{O}(n)$. Therefore, the limiting value of $(1/n)\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\}$ is positive when $-\frac{1}{1-c} < \nu$, from Lemma 1 (2) we have

$$\frac{1}{np}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} \xrightarrow{p} j_* - j.$$

This proves Theorem 2 (2).

Noting that for $j > j_*$, $\ell_j \to c/(1-c)$ under both cases $\mathbf{\Omega} = \mathrm{O}(n)$ and $\mathbf{\Omega} = \mathrm{O}(np)$, it holds that

$$\frac{1}{p}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} \xrightarrow{p} (j - j_*)\frac{1-2c}{1-c}$$

whose limiting value is positive when $c \in [0, 1/2)$. When $j < j_*$ and $\mathbf{\Omega} = \mathrm{O}(n)$,

$$\begin{aligned}\frac{1}{n}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} &\xrightarrow{p} \frac{1}{1-c}(\delta_{j+1} + \cdots + \delta_{j_*}) + (j_* - j)\left\{\frac{c}{1-c} - 2c\right\} \\ &\geq \frac{(j - j_*)}{1-c}\left\{\delta_{j_*} - c(1-2c)\right\}.\end{aligned}$$

Further, when $j < j_*$ and $\mathbf{\Omega} = \mathrm{O}(np)$,

$$\frac{1}{n}\left\{\mathrm{C}_{p\nu;j} - \mathrm{C}_{p\nu;j_*}\right\} \xrightarrow{p} \frac{1}{1-c}(\xi^*_{j+1} + \cdots + \xi^*_{j_*}) > 0.$$

These imply Theorem 2.

A.2 Bias Terms and Consistency for Ridge-Type Criteria

In deriving the bias terms $b^{(1)}_{A,\lambda}$ and $b^{(2)}_{A,\lambda}$ in (30), without loss of generality we may assume that

$$\boldsymbol{\eta} = \left(\mathrm{diag}(\sqrt{\omega_1}, \ldots, \sqrt{\omega_q}) \, , \mathbf{O}\right), \quad \boldsymbol{\Sigma} = \mathbf{I}_p, \quad \lambda = \frac{1}{np}\mathrm{tr}\boldsymbol{\Sigma}\mathbf{S}_e,$$

where $\mathbf{S}_e$ and $\mathbf{S}_h$ are independently distributed as $\mathrm{W}_p(n-k, \mathbf{I}_p)$ and $\mathrm{W}_p(q, \mathbf{I}_q; \tilde{\boldsymbol{\Omega}})$, repectively, and $\tilde{\boldsymbol{\Omega}} = \mathrm{diag}(\omega_1, \ldots, \omega_q, 0, \ldots, 0)$. Our derivation is done under a large-sample framework in (1), assuming that $\omega_i = \mathrm{O}(n) = n\delta_i$, and $\delta_1 \geq \cdots \geq \delta_j > \delta_{j+1} = \cdots = \delta_q = 0$. Note that

$$\begin{aligned}
(1/n)\mathbf{S}_e =& \mathbf{I}_p + \mathrm{O}(n^{-1/2}), \\
(1/n)\mathbf{S}_{e,\lambda} =& (1/n)\mathbf{S}_e + (\alpha/n)\mathbf{I}_p + \mathrm{o}(1), \quad \alpha = (1/p)\mathrm{tr}\boldsymbol{\Sigma}, \\
\sum_{i=j+1}^{q} \tilde{\ell}_i \tilde{\boldsymbol{a}}_i \tilde{\boldsymbol{a}}_i' =& \sum_{i=j+1}^{q} \ell_i \tilde{\boldsymbol{a}}_i \boldsymbol{a}_i' + \mathrm{o}(n^{-1}),
\end{aligned}$$

where $\boldsymbol{a}_i's$ are the orthonormal characteristic vectors of $\mathbf{S}_{e,\lambda}^{-1/2}\mathbf{S}_h\mathbf{S}_{e,\lambda}^{-1/2}$. Further, for $i = j+1, \ldots, q$, $n\ell_i = \mathrm{O}(1)$ and the limiting distribution of $n(\ell_{j+1} + \cdots + \ell_q)$ is a chi-square distribution with $(p-j)(q-j)$ degrees of freedom (see, e.g., Muirhead (1982), Fujikoshi et al. (2010)). Therefore, we have

$$\begin{aligned}
b^{(1)}_{A,\lambda} =& \mathrm{E}\left\{ \mathrm{tr}\left(\frac{1}{n}\mathbf{S}_e\right)^{-1} - (\ell_{j+1} + \cdots + \ell_q) \right\} - \frac{p\alpha}{n} + \mathrm{o}(n^{-1}) \\
=& \frac{np}{n-k-p-1} - \frac{1}{n}(p-j)(q-j) - \frac{p\alpha}{n} + \mathrm{o}(n^{-1}).
\end{aligned}$$

On the other hand

$$\begin{aligned}
b^{(2)}_{A,\lambda} =& \mathrm{E}\left\{ \mathrm{tr}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{\eta}_1 - \mathbf{Y}_1)'(\boldsymbol{\eta}_1 - \mathbf{Y}_1) \right\} + \mathrm{E}\left\{ \mathrm{tr}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{\eta}_1 - \hat{\boldsymbol{\eta}}_1)'(\boldsymbol{\eta}_1 - \hat{\boldsymbol{\eta}}_1) \right\} \\
& + 2\mathrm{E}\left\{ \mathrm{tr}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}(\boldsymbol{\eta}_1 - \mathbf{Y}_1)'(\mathbf{Y}_1 - \hat{\boldsymbol{\eta}}_1) \right\} \\
=& (1) + (2) + (3).
\end{aligned}$$

We use

$$\mathbf{Y}_1 = \mathbf{S}_{e,\lambda}^{1/2}\left(\sum_{i=1}^{q}\sqrt{\tilde{\ell}_i}\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{a}}_i'\right), \quad \mathbf{Y}_1 - \hat{\boldsymbol{\eta}}_{1,\lambda} = \sum_{i=j+1}^{q}\sqrt{\tilde{\ell}_i}\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{a}}_i'\mathbf{S}_{e,\lambda}^{1/2}. \tag{33}$$

Then

$$(1) = \mathrm{E}\left\{\mathrm{tr}\left(\frac{1}{n}\mathbf{S}_e\right)^{-1}(\boldsymbol{\eta}_1 - \mathbf{Y}_1)'(\boldsymbol{\eta}_1 - \mathbf{Y}_1)\right\} + \mathrm{o}(1) = pq + \mathrm{o}(1).$$

Using (33), we have

$$\begin{aligned}(2) =& \mathrm{E}\left\{\mathrm{tr}\left(\frac{1}{n}\mathbf{S}_{e,\lambda}\right)^{1/2}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\left(\frac{1}{n}\mathbf{S}_{e,\lambda}\right)^{1/2}\left(\sum_{i=j+1}^{q}n\tilde{\ell}_i\tilde{\boldsymbol{a}}_i\tilde{\boldsymbol{a}}_i'\right)\right\}\\ =& \mathrm{E}\left(\sum_{i=j+1}^{q}n\ell_i\right) + \mathrm{o}(1) = (p-j)(q-j) + \mathrm{o}(1).\end{aligned}$$

(3) is decomposed

$$\begin{aligned}(3) =& 2\mathrm{E}\left\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{\eta}_1'\left(\sum_{i=j+1}^{q}\sqrt{\tilde{\ell}_i}\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{a}}_i'\right)\mathbf{S}_{e,\lambda}^{1/2}\right\} - 2\mathrm{E}\left\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{S}_{e,\lambda}^{1/2}\left(\sum_{i=j+1}^{q}\tilde{\ell}_i\tilde{\boldsymbol{a}}_i\tilde{\boldsymbol{a}}_i'\right)\mathbf{S}_{e,\lambda}^{1/2}\right\}\\ =& (3\mathrm{a}) - (3\mathrm{b})\end{aligned}$$

We can see that (3a) is o(1), by perturbation expansion method based on $\mathbf{V} = \sqrt{n-k}\{1/(n-k)\mathbf{S}_e - \mathbf{I}_p\}$ and $\mathbf{U} = \mathbf{Y}_1 - \boldsymbol{\eta}_1$. (3b) is evaluated as

$$\begin{aligned}(3\mathrm{b}) =& 2\mathrm{E}\left\{\mathrm{tr}\left(\sum_{i=j+1}^{q}n\ell_i\boldsymbol{a}_i\boldsymbol{a}_i'\right)\right\} + \mathrm{o}(1)\\ =& 2\mathrm{E}\left\{n(\ell_{j+1} + \cdots + \ell_q)\right\} + \mathrm{o}(1) = 2(p-j)(q-j) + \mathrm{o}(1).\end{aligned}$$

These give

$$b_{A,\lambda}^{(2)} = pq - (p-j)(q-j) + \mathrm{o}(1),$$

and hence (28).

Lemma 2 *Let $\mathbf{S}_e$ and $\mathbf{S}_h$ be independently distributed as a Wishart distribution $\mathrm{W}_p(n-k, \mathbf{I}_p)$ and a noncentral Wishart distribution $\mathrm{W}_p(q, \mathbf{I}_p; \tilde{\boldsymbol{\Omega}})$, respectively. Here, we assume that $n-k \geq p \geq q$ and $\tilde{\boldsymbol{\Omega}} = \mathrm{diag}(\omega_1, \ldots, \omega_p)$. Let*

$\mathbf{S}_{e,\lambda} = \mathbf{S}_e + \lambda \mathbf{I}_p$, *where* $\lambda = (np)^{-1}\mathrm{tr}\boldsymbol{\Sigma}\mathbf{S}_e$. *Let* $\ell_1 > \cdots > \ell_q$ *and* $\tilde{\ell}_1 > \cdots > \tilde{\ell}_q$ *be the possible nonzero characteristic roots of* $\mathbf{S}_h\mathbf{S}_e^{-1}$ *and* $\mathbf{S}_h\mathbf{S}_{e,\lambda}^{-1}$, *respectively. We assume that* (i) *a high-dimensional asymptotic framework given by* β1), (ii) *when* $p/n \to c \in [0,1), (1/p)\mathrm{tr}\boldsymbol{\Sigma} \to \alpha_0$, *and* (iii) $\omega_1 \geq \cdots \geq \omega_a > \omega_{a+1} = \cdots = 0$. *Then, the characteristic roots* $\tilde{\ell}_1 > \cdots > \tilde{\ell}_q$ *have the same limiting values as the ones of the characteristic roots* $\ell_1 > \cdots > \ell_q$ *given in Lemma* A1.

Proof 2 *In general, it holds that* $\ell_i \geq \tilde{\ell}_i, i = 1, \ldots, q$. *Noting that*

$$\begin{aligned}(\mathbf{S}_e + \lambda\mathbf{I}_p)^{-1} =& \mathbf{S}_e^{-1}\left(\mathbf{I}_p + \lambda\mathbf{S}_e^{-1}\right)^{-1} \\ =& \mathbf{S}_e^{-1} - \lambda\mathbf{S}_e^{-2}\left(\mathbf{I}_p + \lambda\mathbf{S}_e^{-1}\right)^{-1},\end{aligned}$$

the following decomposition is obtained

$$\mathbf{S}_h\mathbf{S}_e^{-1} = \mathbf{S}_h\mathbf{S}_e^{-1}\left(\mathbf{I}_p + \lambda\mathbf{S}_e^{-1}\right)^{-1} + \lambda\mathbf{S}_h\mathbf{S}_e^{-2}\left(\mathbf{I}_p + \lambda\mathbf{S}_e^{-1}\right)^{-1}.$$

Using Weyl's Theorem (see Seber (2008, p.117)), we have

$$\ell_i \leq \tilde{\ell}_i + \lambda\mathrm{tr}\mathbf{S}_h\mathbf{S}_e^{-2}\left(\mathbf{I}_p + \lambda\mathbf{S}_e^{-1}\right)^{-1} \leq \tilde{\ell}_i + \lambda\mathrm{tr}\left(\mathbf{S}_h\mathbf{S}_e^{-1}\right)\left(\mathbf{S}_e^{-1}\right).$$

Note that $2\mathrm{tr}\mathbf{A}\mathbf{B} \leq \mathrm{tr}\mathbf{A}^2 + \mathrm{tr}\mathbf{B}^2$ *for any square matrices* $\mathbf{A}$ *and* $\mathbf{B}$. *Therefore,*

$$\begin{aligned}2\mathrm{tr}\left(\mathbf{S}_h\mathbf{S}_e^{-1}\right)\left(\mathbf{S}_e^{-1}\right) \leq& \mathrm{tr}\left(n^{-\gamma/2}\mathbf{S}_h\mathbf{S}_e^{-1}\right)^2 + \mathrm{tr}\left(n^{\gamma/2}\mathbf{S}_e^{-1}\right)^2 \\ =& n^{-\gamma}(\ell_1^2 + \cdots + \ell_q^2) + n^{\gamma}(n-k)^{-2}\mathrm{tr}\left\{(n-k)^{-1}\mathbf{S}_e\right\}^{-2}.\end{aligned}$$

where γ *is a positive constant. By Marčenko-Pastur law (see Bai and Silverstein (2010)), it is known (Bai, Chen and Fujikoshi (2015)) that*

$$\lim \frac{1}{n}\mathrm{tr}\left\{(n-k)^{-1}\mathbf{S}_e\right\}^{-2} = \frac{c}{(1-c)^3}.$$

When $\omega_j = n\delta_j = \mathrm{O}(n)$ *and* $\lim \delta_j = \delta_j^* > 0$, *from Lemma A1 we have*

$$\lim(\ell_1^2 + \cdots + \ell_q^2) = \sum_{i=1}^{a}\left(\frac{c}{1-c} + \frac{1}{1-c}\delta_i^*\right)^2 + (q-a)\left(\frac{c}{1-c}\right)^2.$$

Noting that $\lambda \to \alpha$, *and taking* $0 < \gamma < 1$, *we get* $\lim \ell_i \leq \lim \tilde{\ell}_i$, *and hence* $\lim \tilde{\ell}_i = \lim \ell_i$. *When* $\omega_j = np\xi_j = O(np)$ *and* $\lim \xi_j = \lim \xi_j^*$, *we have seen that* $(1/p)\ell_j \to (1-c)^{-1}\xi_j^*$, $j = 1, \ldots, a$ *and* $\ell_j \to c(1-c)^{-1}$, $j = a+1, \ldots, q$. *By a similar discussion as in the case of* $\omega_j = O(n)$, *we can see that* $\tilde{\ell}_j$ *has the same limiting value as* ℓ_j, *for* $j = 1, \ldots, q$.

References

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

[2] Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.*, **22**, 327–351.

[3] Anderson, T.W. (2003). *Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, N. J.

[4] Bai, Z. W. and Silverstein, J. W. (2010). *Spectral Analysisof Large Dimensional Random Matrices*, 2nd ed. Springer.

[5] Bai, Z. W., Chen, K. P. and Fujikoshi, Y. (2015). Limiting behavior of eigenvalues in MANOVA with high-dimension by RMT, In Manuscript.

[6] Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, **39**, 1282–1309.

[7] Bunea, F., She, Y. and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, **40**, 2359–2388.

[8] Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journ. Amer. Statist. Assoc.*, **107**, 1533–1545.

[9] Fujikoshi, Y. and Veitch, L. G. (1979). Estimation of dimensionality in canonical correlation analysis. *Biometrika*, **66**, 345–351.

[10] Fujikoshi, Y. (1985). Two methods for estimation of dimensionality in canonical correlation analysis and the multivariate linear model. In *Statistical Theory and Data Analysis* (K. Matsushita, Ed.), 233–240. Elsevier Science, Amsterdam.

[11] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, Hoboken, N.J.

[12] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2013). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. To appear in Journal of Multivariate Analysis.

[13] Gunderson, B. K. and Muirhead, R. J. (1997). On estimating the dimensionality in canonical correlation analysis. *Journal of Multivariate Analysis*, **62**, 121–136.

[14] Izenman, A. J. (1975). Reduced-Rank Regression for the multivariate linear model, *J. Multivariate Anal.*, **5**, 248–262.

[15] Kshirsagar, A. M. (1972). *Multivariate Analysis.* Marcel Dekker, New York.

[16] Lawley, D.N. (1959). Tests of significance in canonical analysis. *Biometrika*, **46**, 59–66.

[17] Kubokawa, T. and Srivastava, M. S. (2012). Selection of variables in multivariate regression models for large dimensions. *Communication in Statistics-Theory and Methods*, **41**, 2465–2489.

[18] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[19] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, New York.

[20] Reisel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression*. Lecure Notes in Statistics **136**, Springer, New York.

[21] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[22] Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians*, Wiley.

[23] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2012). A consistency property of AIC for multivariate linear model when the dimension and the sample size are large. Submitted for publication. TR No. 12-08, *Statistical Research Group, Hiroshima University*.

[24] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear model. *J. R. Statist. Soc. B*, bf 69, 329–346.