High-Dimensional Properties of AIC and C_p for Estimation of Dimensionality in Multivariate Models

Yasunori Fujikoshi

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

Abstract

The AIC and C_p have been proposed for estimation of the dimensionality in some multivariate models. In this paper we consider high-dimensional properties of the criteria in multivariate linear model and canonical correlation analysis. First we show consistency properties of the criteria when the number p of variables and the sample size n are large under a high-dimensional asymptotic framework such that $p/n \to c \in [0, 1)$. The consistency properties are shown, under two types of assumptions on the order of the noncentrality parameter matrix, but the true dimension j_0 and the possible maximum dimension $q(\geq j_0)$ are fixed. When j_0 and q are also large with $q/n \to 0$, we give a sufficient condition such that the probabilities of estimating the overspecified dimensions tend to zero, without any assumption on the order of the noncentrality parameter matrix. Through a Monte Carlo simulation experiment we see that our results are checked numerically. Further, we compare with the dimensionalities estimated by *AIC* and C_p.

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

Key Words and Phrases: AIC, Canonical correlation analysis, C_p, Consistency property, Dimensionality, Discriminant analysis, High-dimensional framework, Multivariate regression model.

1. Introduction

The dimensionality problem have been studied in some multivariate models. The dimensionality in multivariate linear model is defined by a rank condition on the mean parameter matrix. The likelihood ratio tests on dimensionality were first obtained by Anderson (1951, 2003) in multivariate linear model including discriminant analysis. Izenman (1985) treated the case when the set of explanatory variables is a random vector as well as a fixed vector. The model with a rank condition on the regression matrix is called multivariate reduced-rank regression (RRR) model. In general, we call these models reduced-rank models. Reinsel and Velu (1998) have reviewd various applications and inferential problems in the RRR model.

One of the important problems in multivariate reduced-rank models is concerned with the estimation of dimensionality. There are some approaches for estimating the dimensionality. One is based on sequential test procedures. One of the other approaches is based on the use of model selection criteria. The cross-validiation method is also used. Yuan and Ekici(2007) proposed a method based on penalized least squares estimate. Recently some other methods have been proposed by Bunea, She and Wegkamp (2011, 2012) and Chen and Hung (2012).

In this paper we are concerned with the estimation method by use of the model selection criteria AIC (Akaike (1973)) and C_p (Mallows (1973)), which were proposed by Fujikoshi and Veitch (1979) in multivariate linear model with p response variables and k explanatory variables and canonical correlation analysis with two random vectors of p and $q(q \leq p)$ components, based on the sample size n. The large sample properties of the criteria were studied by Fujikoshi (1985), Gunderson and Muirhead (1997). It is known that the criteria have no consistency property in a large-sample asymptotic framework.

The AIC, C_p and their modifications have been proposed for selection of

k predictive variables in multivariate linear model with p response variables and the sample size n. The criteria have not a consistency property under a large-sample framework such that n tends to infinity, but p, q and k are fixed. However, recently it is known that there is a situation such that these criteria have a consistency property when the number p of response variables and the sample size n are large under a high-dimensional framework such that $p/n \to c \in [0, 1)$. These results can be found in Fujikoshi, Sakurai and Yanagihara (2013) and Yanagihara, Wakaki and Fujikoshi (2014).

In this paper we first consider asymptotic properties of AIC and C_p for estimation of dimensionality in a high-dimensional asymptotic framework such that

$$p \to \infty, \quad n \to \infty, \quad p/n \to c \in [0, 1). \tag{1.1}$$

It is shown that the AIC and C_p for estimation of dimensionality have consistency properties under two types of assumptions on largeness of the noncentrality parameter matrix when the true dimension j_0 and the maximum dimension $q(\geq j_0)$ are fixed. The noncentrality matrix essentially expresses the true discriminant power in discriminant analysis and the true canonical correlations in canonical correlation analysis. Next we consider a situation where j_0, q and the number k of explanatory variables are also large. More precisely we assume an extended high-dimensional asymptotic framework such that

$$p \to \infty, \quad n \to \infty, \quad k \to \infty, \quad p/n \to c \in [0, 1), \quad k/n \to 0, \tag{1.2}$$

which implies $j_0/n \to 0$ and $q/n \to 0$. Here, in canonical correlation analysis, k should be understood as q. Then, we give a sufficient condition such that the probabilities of estimating the overspecified dimensions tend to zero, without any assumption on the order of the noncentrality parameter matrix. It may be noted that these properties are different from the ones in a large-sample case, since in general these criteria have a positive probability of selecting each of the overspecified models. Our results are checked numerically by conducting a Monte Carlo simulation experiment. Further, we compare with the dimensionalities estimated by AIC and C_p.

In Section 2, we treat the criteria for estimating the dimensionality in multivariate linear model including multivariate RRR model and discriminant analysis. High-dimensional properties of the criteria are given. In Section 3 we treat the criteria for estimating the dimensionality in canonical correlation analysis. In Section 4 we check our theoretical results by conducting a Monte Carlo simulation experiment, and compare with the selection probabilities of the two criteria. The proofs of our results are given in Appendix.

2. Multivariate Linear Model

2.1. Preliminaries

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 n observations on $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_k)'$, and let $\mathbf{Y} : n \times p$ and $\mathbf{X} : n \times k$ be the observation matrices of $\boldsymbol{y}$ and $\boldsymbol{x}$ with the sample size n, respectively. The multivariate normal linear model on $\boldsymbol{y}$ and $\boldsymbol{x}$ is written as

$$\mathbf{Y} \sim \mathrm{N}_{n\times p}(\mathbf{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n), \tag{2.1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ unknown matrix of regression coefficients, $\boldsymbol{\Sigma}$ 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 $\mathbf{X}\boldsymbol{\Theta}$ and the covariance matrix of $\mathrm{vec}\,(\mathbf{Y})$ is $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$, where $\mathrm{vec}\,(\mathbf{Y})$ is the $np \times 1$ vector formed by stacking the columns of $\boldsymbol{Y}$ under each other. We assume that $n - k > 0$ 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.

Consider a testing problem

$$H: \mathbf{C\Theta} = \mathbf{O} \quad \text{vs.} \quad K: \mathbf{C\Theta} \neq \mathbf{O}, \tag{2.2}$$

where $\mathbf{C}$ is a given $q \times k$ matrix with $\text{rank}(\mathbf{C}) = q$. Then, we have an LR statistic given

$$\Lambda = \frac{|\mathbf{S}_e|}{|\mathbf{S}_e + \mathbf{S}_h|}, \tag{2.3}$$

where

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

where $\hat{\mathbf{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is well known that $\mathbf{S}_e$ and $\mathbf{S}_h$ are independently distributed as a Wishart distribution $\text{W}_p(n-k, \mathbf{\Sigma})$ and a noncentral Wishart distribution $\text{W}_p(q, \mathbf{\Sigma}; \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively where

$$\tilde{\mathbf{\Omega}} = \mathbf{\Sigma}^{-1/2}(\mathbf{C\Theta})'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}\mathbf{C\Theta}\mathbf{\Sigma}^{-1/2}. \tag{2.4}$$

One of the interested problems when the hypothesis is rejected is to consider a reduced-rank condition defined

$$M_j : \text{rank}(\mathbf{C\Theta}) = j. \tag{2.5}$$

Relating to estimating the rank of $\mathbf{C\Theta}$, we consider a family of models expressed as $\{M_0, M_1, \ldots, M_q\}$.

Some applications have been discussed in the case $\mathbf{C} = \mathbf{I}_k$ by Izenman (2008), Reinsel and Velu (1998). An important application when $\mathbf{C} \neq \mathbf{I}_k$ appears in discriminant analysis. These two special cases are explained in the next subsection.

2.2. Two Special Cases

First we consider a multivariate RRR model (Izenman (2008), Reinsel and Velu (1998)) which is given by (2.1) with (2.5) and $\mathbf{C} = \mathbf{I}_k$. From the rank

constraint the regression matrix $\mathbf{\Theta}$ can be expressed as a product of two rank j matrices as follows:

$$\mathbf{\Theta} = \mathbf{FG},$$

where $\mathbf{F}$ is of dimension $k \times j$ and $\mathbf{G}$ is of dimension $j \times p$. Then, the model can be written as

$$\mathbf{Y} \sim \mathrm{N}_{n \times p}((\mathbf{XF}) \cdot \mathbf{G}, \mathbf{\Sigma} \otimes \mathbf{I}_n). \tag{2.6}$$

The model means that the j linear combinations $\mathsf{F}\boldsymbol{x}$ of the k explanatory variavles $\boldsymbol{x}$ are sufficient to model the valiation in the p response variables $\boldsymbol{y}$. For a more detailed discussion, see Reinsel and Velu (1986). In practice, the dimension j is unknown, and we need to estimate it. Recently Chen and Huang (2012) has proposed to select relevant variables for reduced-rank regression by using a spacity-inducing penality.

Next we consider $(q+1)$ p-variate normal populations with common covariance matrix $\mathbf{\Sigma}$ and the ith population having mean vector $\boldsymbol{\mu}_i$. 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

$$\begin{aligned}
\mathsf{S}_b &= n_1(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})' + \cdots + n_{q+1}(\bar{\boldsymbol{y}}_{q+1} - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_{q+1} - \bar{\boldsymbol{y}})', \\
\mathsf{S}_w &= (n_1 - 1)\mathsf{S}_1 + \cdots + (n_{q+1} - 1)\mathsf{S}_{q+1},
\end{aligned}$$

respectively, where $\bar{\boldsymbol{y}}_i$ ans 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, S_w and S_b are independently distributed as a Wishart distribution $\mathrm{W}_p(n - q - 1, \mathbf{\Sigma})$ and a noncentral Wishart distribution $\mathrm{W}_p(q, \mathbf{\Sigma}; \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively where

$$\tilde{\mathbf{\Omega}} = \mathbf{\Sigma}^{-1/2} \sum_{i=1}^{q+1} n_i (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}) \mathbf{\Sigma}^{-1/2}, \tag{2.7}$$

where $\bar{\boldsymbol{\mu}} = (1/n)\sum_{i=1}^{q+1} n_i \boldsymbol{\mu}_i$.

In this paper, since we are interested in asymptotic properties in a high-dimensional situation, we assume that $p \geq q$. The coefficient vector $\boldsymbol{\beta}_i$ of the

i-th population discriminant function is defined as the characteristic vector satisfying

$$\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_i = \omega_i\boldsymbol{\Sigma}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i'\boldsymbol{\Sigma}\boldsymbol{\beta}_j = \delta_{ij},$$

where δ_{ij} denotes the Kroneker delta. Here, $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_q \geq 0$ 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 et al. (2010), etc.) as the number of non-zero characteristic roots of $\tilde{\Omega}$ which is the number of meaningful population discriminant functions. The model that the dimension is j may be expressed as

$$\begin{aligned} M_j : &\operatorname{rank}(\tilde{\boldsymbol{\Omega}}) = j, \\ &\Leftrightarrow \omega_1 \geq \cdots \geq \omega_j > \omega_{j+1} = \cdots = \omega_q = 0, \\ &\Leftrightarrow \operatorname{rank}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{q+1}, \ldots, \boldsymbol{\mu}_{q-1} - \boldsymbol{\mu}_{q+1}) = j. \end{aligned} \tag{2.8}$$

Note that the model M_j in (2.5) is a generalization of M_j in discriminant analysis. This is easily seen by taking $k = q + 1$ and choosing $\mathbf{Y}$, $\mathbf{C}$, $\mathbf{X}$ and $\boldsymbol{\Theta}$ as follows.

$$\mathbf{Y} = (\boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1N_1}, \ldots, \boldsymbol{y}_{q+1,1}, \ldots, \boldsymbol{y}_{q+1,N_{q+1}})', \quad \mathbf{C} = (\mathbf{I}_q, \ -\mathbf{1}_q)$$

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

where $\mathbf{1}_n$ is an $n \times 1$ vector whose elements are all one.

2.3. AIC and C_p

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 (Fujikoshi and Veitch (1979)) as

$$\begin{aligned}\mathrm{AIC}_j =& n\log(1+\ell_{j+1})\cdots(1+\ell_q) + n\log|(1/n)\mathbf{S}_e| \\ &+ np(\log 2\pi + 1) + 2\{j(p+q-j) + (k-q)p + \frac{1}{2}p(p+1)\}. \quad (2.9)\end{aligned}$$

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

$$\begin{aligned}\mathrm{A}_j =& \mathrm{AIC}_j - \mathrm{AIC}_q \\ =& n\log\prod_{i=j+1}^{q}(1+\ell_i) - 2(p-j)(q-j),\ j=0,\ldots,q. \quad (2.10)\end{aligned}$$

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

The $\mathrm{C}_{p,j}$ corresponding to A_j is given (Fujikoshi and Veitch (1979)) by

$$\mathrm{C}_{p,j} = n\sum_{i=j+1}^{q}\ell_i - 2(p-j)(q-j),\ j=0,\ldots,q. \quad (2.11)$$

2.4. High-Dimensional Properties

We denote M_j by j simply. Then, the set of all the models is $\mathcal{F} = \{0, 1, \ldots, q\}$. It is assumed that the true dimension is j_0, where $0 \le j_0 \le q$. We also denote the true model by j_0, and also denote the minimum model including the true model by j_0. We separate $\mathcal{F}$ into two sets, one is a set of overspecified models, i.e., $\mathcal{F}_+ = \{j_0, j_0+1, \ldots, q\}$ and the other is a set of underspecified models, i.e., $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F} = \{0, 1, \ldots j_0 - 1\}$. Further, we denote the set of models deleting the true model from $\mathcal{F}_+$ by $\mathcal{F}_+\backslash\{j_0\}$, i.e., $\mathcal{F}_+\backslash\{j_0\} = \{j_0+1, \ldots, q\}$.

The estimation methods are expressed as

$$\hat{j}_{\mathrm{A}} = \arg\min_{j\in\mathcal{F}} \mathrm{A}_j, \quad \hat{j}_{\mathrm{C}} = \arg\min_{j\in\mathcal{F}} \mathrm{C}_{p,j}.$$

We make two types of assumptions on the order of $\tilde{\mathbf{\Omega}}$ in (2.4). Since $\mathrm{rank}(\tilde{\mathbf{\Omega}}) \leq q$, we can write $\tilde{\mathbf{\Omega}} = \mathbf{\Gamma}_1 \mathbf{\Gamma}_1^{'}$, where Γ_1 is a $p \times q$ matrix. Let

$$\mathbf{\Omega} = \mathbf{\Gamma}_1' \mathbf{\Gamma}_1, \tag{2.12}$$

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

$$\tilde{\mathbf{\Omega}} = (n_1 n_2/n) \mathbf{\Sigma}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Sigma}^{-1/2},$$

and

$$\mathbf{\Omega} = \frac{n_1 n_2}{n} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

which is $(n_1 n_2/n)$ times the squared Mahalanobis distance between two normal populations $\mathrm{N}_p(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ and $\mathrm{N}_p(\boldsymbol{\mu}_2, \mathbf{\Sigma})$. It may be natural for assuming that $\mathbf{\Omega} = \mathrm{O}(n)$ and also $\mathbf{\Omega} = \mathrm{O}(np)$, depending on the largeness of $\mathbf{\Omega}$, where $\mathrm{O}(n)$ is the usual order in a high-dimensional framework (1.1) or (1.2). Note that, when we consider the distributions of AIC and C_p, without loss of information we may assume

$$\mathbf{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q), \tag{2.13}$$

where $\omega_1 \geq \cdots \geq \omega_q$ are the characteristic roots of $\mathbf{\Omega}$ or the non-zero characteristic roots of $\tilde{\mathbf{\Omega}}$.

Here we list our main assumptions whose parts are used, depending on Theorems:

A1 (The true model): $j_0 \in \mathcal{F}$.

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

A3 (The asymptotic framework-2): $j_0 \to \infty,\ \ p \to \infty,\ n \to \infty,$
$q/n \to 0, \quad p/n \to c \in [0, 1)$.

A4 (The noncentrality matrix-1): For any $j (0 \leq j < j_0)$,
$\omega_j = n\delta_j = \mathrm{O}(n), \quad \lim_{p/n \to c} \delta_j = \delta_{j0} > 0.$

A5 (The noncentrality matrix-2): For any $j(0 \leq j < j_0)$,

$$\omega_j = np\xi_j = \mathrm{O}(np), \quad \lim_{p/n \to c} \xi_j = \xi_{j0} > 0.$$

The assumption A3 (the asymptotic framework-2) means that q and j_0 may tend to infinity, under the restriction that $q/n \to 0$ and hence $j_0/n \to 0$. First we show that the asymptotic probabilities of selecting the true model by $\hat{j}_{\mathrm{A}}$ and $\hat{j}_{\mathrm{C}}$ go to 1 when the sample size and the dimension of response variables are approaching to ∞ as in (1.1), and q is fixed. Next we consider the case when j_0 and q may tend to infinity.

Theorem 2.1. *Suppose that the assumptions* A1 *and* A2 *are satisfied. Further, assume that* $c \in [0, c_{\mathrm{a}})$, *where* c_{a} (≈ 0.797) *is the constant satisfying* $\log(1 - c_{\mathrm{a}}) + 2c_{\mathrm{a}} = 0$.

(1) *Suppose that the assumption* A4 *is satisfied, and*

A6: $\log(1 + \delta_{j_0 0}) > 2c + \log(1 - c)$.

Then, the asymptotic probability of selecting the true model k_0 *by* AIC *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} = j_0) = 1.$$

(2) *Suppose that the assumption* A5 *is satisfied. Then, the asymptotic probability of selecting the true model* j_0 *by* AIC *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} = j_0) = 1.$$

Similar results are obtained for $\hat{j}_c$ as in the next theorem.

Theorem 2.2. *Suppose that the assumptions* A1 *and* A2 *are satisfied. Further, assume that* $c \in [0, 1/2)$.

(1) *Suppose that the assumption* A4 *is satisfied, and*

A7: $\delta_{j_0 0} > c(1 - 2c)$.

Then, the asymptotic probability of selecting the true model j_0 by C_p tends to 1, i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{C}} = j_0) = 1.$$

(2) *Suppose that the assumption* A4 *is satisfied. Then, the asymptotic probability of selecting the true model j_0 by C_p tends to* 1, i.e.

$$\lim_{p/n\to c} P(\hat{j}_{\mathrm{C}} = j_0) = 1.$$

From the proofs of Theorems 2.1 and 2.2 we can see that the AIC and C_p criteria on the dimensionality in multivariate linear model satisfy the followings:

$$(\mathrm{i};\mathrm{MLM}) \ \text{ if } c \in [0, c_0), \ \lim_{p/n\to c} P(\hat{j}_{\mathrm{A}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0, \tag{2.14}$$

$$(\mathrm{ii};\mathrm{MLM}) \ \text{ if } c \in [0, 0.5), \ \lim_{p/n\to c} P(\hat{j}_{\mathrm{C}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0, \tag{2.15}$$

without the assumptions on the order of $\mathbf{\Omega}$. The first result can be extended to the case when q is large as follows.

Theorem 2.3. *Suppose that the assumptions* A1 *and* A3 *are satisfied. If there exists a positive number δ such that $2c(q - j_0)^{-1} + \log(1 - c) > \delta$, then*

$$\lim_{p/n\to c, q/n\to 0} P(\hat{j}_{\mathrm{A}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0. \tag{2.16}$$

When $j_0 + 1 = q$, the sufficient condition in Theorem 2.3 is expressed as $c \in [0, c_a)$, where c_a is the constant given Theorem 2.1.

It may be noted that these properties (2.14), (2.14) and (2.16) are different from the ones in large-sample framework. In fact, under a large-sample framework

$$p, q, k; \ \text{fixed}, \quad n \to \infty, \tag{2.17}$$

and A4, it is known (Fujikoshi (1985)) that

$$\lim_{n\to\infty} P(\hat{j}_{\mathrm{A}} = j) = P(\hat{j}_{\mathrm{C}} = j) = h(j|j_0), \tag{2.18}$$

where for $j = 0, 1, \ldots, j_0 - 1$, $h(j|j_0) = 0$, and for $j = j_0, \ldots, q$, $h(j|j_0)$'s are positive and are expressed in terms of the characteristic roots $z_1 > \cdots > z_s$ of a $s \times s$ Wishart matrix $\mathbf{W}$ whose distribution is distributed as $\mathrm{W}_s(t, \mathbf{I}_s)$ as

$$h(j|j_0) = P\Big(\sum_{i=k+1}^{j-j_0} z_i > 2p_{k,j-j_0};\ k = 0, 1, \ldots, j - j_0 - 1,\ \text{and},$$
$$\sum_{i=j-j_0+1}^{k} z_i > 2p_{j-j_0,k};\ k = j - j_0 + 1, j - j_0 + 2, \ldots, q\Big). \quad (2.19)$$

Here $p_{ij} = (p - j_0 - i)(q - j_0 - i) - (p - j_0 - j)(q - j_0 - j)$, and the density function of $z_1, \ldots, z_s$ is expressed as

$$\frac{\pi^{s^2/2}}{2^{st/2}\Gamma_s\big(\frac{1}{2}s\big)\Gamma_s\big(\frac{1}{2}t\big)} \exp\Big(-\frac{1}{2}\sum_{i=1}^{s} z_i\Big) \prod_{i=1}^{s} z_i^{(t-s-1)/2} \prod_{i<j}^{s} (z_i - z_j),$$

where $\Gamma_a(b) = \pi^{a(a-1)/4} \prod_{i=1}^{a} \Gamma[b - (i-1)/2]$.

3. Canonical Correlation Analysis

3.1. Preliminaries

Let

$$\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)', \quad \mathbf{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)'$$

be a sample of size $N = n + 1$ of $(\boldsymbol{x}',\ \boldsymbol{y}')'$ from $(p+q)$-dimensional normal distribution $\mathrm{N}_{q+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{x} : p \times 1$ and $\boldsymbol{y} : q \times 1$. Let $\mathbf{S}$ be the sample covariance matrix formed from the sample. In this section we assume that $q \leq p$. Corresponding to a partition $(\boldsymbol{x}', \boldsymbol{y}')$, we partition $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{S}$ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

Let $\rho_1 \geq \cdots \geq \rho_q \geq 0$ and $r_1 > \cdots > r_q > 0$ be the population and sample canonical correlations between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then $\rho_1^2 \geq \cdots \geq \rho_q^2 \geq 0$

and $r_1^2 > \cdots > r_q^2 > 0$ are the characteristic roots of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ and $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, respectively.

As is well known, by considering the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ we can regard the canonical correlation model as a multivariate linear model, i.e.

$$\mathbf{Y}|\mathbf{X} \sim \mathrm{N}_{N\times q}(\mathbf{1}_N\boldsymbol{\mu}_{2\cdot 1}' + \mathbf{X}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{22\cdot 1}\otimes \mathbf{I}_N), \tag{3.1}$$

where $\boldsymbol{\mu}_{2\cdot 1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{22\cdot 1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

We are interested in the number of nonzero canonical correlations, which is called the dimensionality in canonical correlation analysis. Related to the estimation of the dimensionality we consider a dimensionality model:

$$\begin{aligned} M_j: \quad & \mathrm{rank}(\Sigma_{12}) = j, \\ & \Leftrightarrow \ \rho_j > \rho_{j+1} = \cdots = \rho_q = 0 \end{aligned} \tag{3.2}$$

If M_j is true, we can explain the correlation structure between $\boldsymbol{x}$ and $\boldsymbol{y}$ by the first j canonical correlation variables, since the remeining canonical variables have no power of prediction.

Based on the likelihood of $\mathbf{S}$, it is known (Fujikoshi and Veitch (1979)) that AIC for M_j is given by

$$\begin{aligned} \mathrm{AIC}_j = & -\sum_{i=j+1}^{q} n\log(1-r_i^2) + n(p+q) + (p+q+1)\log|\mathbf{S}| + K \\ & + 2\left\{ j(p+q-j) + \frac{1}{2}p(p+1) + \frac{1}{2}q(q+1) \right\}, \end{aligned} \tag{3.3}$$

where $K = 2\log\left\{\boldsymbol{\Gamma}_{p+q}(\frac{1}{2}/(\frac{1}{2}n)^{(1/2)n(p+q)}\right\}$. Instead of AIC, we may use

$$\begin{aligned} \mathrm{A}_j =& \mathrm{AIC}_j - \mathrm{AIC}_q \\ =& -\sum_{i=j+1}^{q} n\log(1-r_i^2) - 2(p-j)(q-j),\ j = 0,\ldots,q. \end{aligned} \tag{3.4}$$

Here $\mathrm{A}_q = 0$. The $\mathrm{C}_{p,j}$ corresponding to A_j is given by

$$\mathrm{C}_{p,j} = n\sum_{i=j+1}^{q}\frac{r_i^2}{1-r_i^2} - 2(p-j)(q-j),\ j = 0,\ldots,q. \tag{3.5}$$

Note that the A_j and $\mathrm{C}_{p,j}$ based on the likelihood of $\mathbf{X}$ and $\mathbf{Y}$ can be expressed as the ones in (3.4) replaced n by N .

3.2. High-Dimensional Properties

We use the same notation for the models of dimensionality as in multivariate linear model. For examples, the model M_j is denoted by j simply. Then, the set of all the models is $\mathcal{F} = \{0, 1, \ldots, q\}$. It is also assumed that the true dimension is j_0, where $0 \leq j_0 \leq q$. The notation j_0 is also used for the true model and the minimum model including the true model.

When we treat the distributions of the canonical correlations themselves or their function, without loss of generality we may assume that

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_p & \mathbf{\mathcal{R}}' \\ \mathbf{\mathcal{R}} & \mathbf{I}_q \end{pmatrix}, \tag{3.6}$$

$\mathbf{\mathcal{R}} = (\mathbf{\mathcal{R}}_1 \ \mathbf{O})'$, $\mathbf{\mathcal{R}}_1 = \mathrm{diag}(\rho_1, \ldots, \rho_q)$. The number of possible nonzero canonical correlations is q. We will consider the transformed population and sample canonical correlations defined by

$$\gamma_j = \frac{\rho_j}{(1-\rho_j^2)^{1/2}}, \quad d_j = \frac{r_j}{(1-r_j^2)^{1/2}}, \quad j = 1, \ldots, q. \tag{3.7}$$

We use the assumptions similar to the ones in multivariate linear model. The following assumptions are used, depending on Theorems:

B1 (The true model): $j_0 \in \mathcal{F}$.

B2 (The asymptotic framework-1): q is fixed,
$p \to \infty,\ n \to \infty,\ p/n \to c \in [0, 1)$.

B3 (The asymptotic framework-2): $j_0 \to \infty,\ \ p \to \infty,\ n \to \infty,$
$q/n \to 0, \quad p/n \to c \in [0, 1)$.

B4 (The canonical correlations-1): For any $j (0 \leq j < j_0)$,
$\rho_j^2 = \mathrm{O}(1)$ and $\lim_{p/n \to c} \rho_j^2 = \rho_{j0}^2$.

B5 (The canonical correlations-2): For any $j(0 \leq j < j_0)$,
$\gamma_j^2 = p\theta_j^2 = \mathrm{O}(p)$ and $\lim_{p/n \to c} \theta_j^2 = \theta_{j0}^2$.

Theorem 3.1. *Suppose that the assumptions* B1 *and* B2 *are satisfied. Further, assume that* $c \in [0, c_{\mathrm{a}})$, *where* c_{a} (≈ 0.797) *is the constant satisfying* $\log(1 - c_{\mathrm{a}}) + 2c_{\mathrm{a}} = 0$.

(1) *Suppose that the assumption* B4 *is satisfied, and*

B6: $-\log(1 - \rho_{j_0 0}^2) > 2c + \log(1 - c)$.

Then, the asymptotic probability of selecting the true model j_0 *by* AIC *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} = j_0) = 1.$$

(2) *Suppose that the assumption* B5 *is satisfied. Then, the asymptotic probability of selecting the true model* j_0 *by* AIC *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} = j_0) = 1.$$

Theorem 3.2. *Suppose that the assumptions* B1 *and* B2 *are satisfied. Further, assume that* $c \in [0, 1/2)$.

(1) *Suppose that the assumption is satisfied, and*

B6: $\delta_{j_0 0} > c(1 - 2c)$.

Then, the asymptotic probability of selecting the true model j_0 *by* C_p *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{C}} = j_0) = 1.$$

(2) *Suppose that the assumption* B5 *is satisfied. Then, the asymptotic probability of selecting the true model* j_0 *by* C_p *tends to* 1, i.e.

$$\lim_{p/n \to c} P(\hat{j}_{\mathrm{C}} = j_0) = 1.$$

From the proofs of Theorems 3.1 and 3.2 we can also see that the AIC and C_p criteria on the dimensionality in canonical correlation analysis satisfy the followings:

$$(\mathrm{i};\mathrm{CCA}) \quad \text{if } c \in [0, c_0), \ \lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0, \tag{3.8}$$

$$(\mathrm{ii};\mathrm{CCA}) \quad \text{if } c \in [0, 0.5), \ \lim_{p/n \to c} P(\hat{j}_{\mathrm{C}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0. \tag{3.9}$$

These results hold without the assumptions on the order of population canonical correlations.

Theorem 3.3. *Suppose that the assumptions* B1 *and* B3 *are satisfied. Further, if there exists a positive number* δ *such that* $2c(q-j_0)^{-1}+\log(1-c) > \delta$, *then*

$$\lim_{p/n \to c, q/n \to 0} P(\hat{j}_{\mathrm{A}} \in \mathcal{F}_+ \backslash \{j_0\}) = 0. \tag{3.10}$$

Under a large-sample framework (2.17) with $k = q$ and B4 it is known (Fujikoshi (1985)) that

$$\lim_{n \to \infty} P(\hat{j}_{\mathrm{A}} = j) = \lim_{n \to \infty} P(\hat{j}_{\mathrm{C}} = j) = h(j|j_0), \tag{3.11}$$

where for $j = 0, 1, \ldots, j_0 - 1$, $h(j|j_0) = 0$, and for $j = j_0, \ldots, q$, $h(j|j_0)$'s are positive and expressed as (2.19). Gunderson and Muirhead (1997) gave a different expression for $h(j|j_0)$ and extended the result to the case of an elliptical distribution.

4. Numerical Study

In this section, we numerically examine the validity of our claims, and point some tendencies for the dimensionalities estimated by AIC and C_p.

4.1. DA(Discriminant Analysis)

The nonzero characteristic roots $\ell_1 > \cdots > \ell_q$ in DA are the ones of $\mathbf{S}_b\mathbf{S}_w^{-1}$. When we consider distributions of AIC and C_p, without loss of generality we may assume that $\mathbf{S}_w$ and $\mathbf{S}_b$ are independently distributed as $\mathrm{W}_p(n-q-1, \mathbf{I}_q)$ and $\mathrm{W}_p(q, \mathbf{I}_p; \mathbf{\Omega}_p)$, respectively. Here, $\mathbf{\Omega} = \mathrm{diag}(\omega_1, \cdots, \omega_q, 0, \ldots, 0)$ and $\omega_1, \cdots, \omega_q$ are the possible nonzero characteristic roots of the noncentrality matrix $\mathbf{\Omega}$ defined by (2.7).

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 $j_0 = 3$. The two types of characteristic roots $\omega_i, i = 1, \ldots, 5$ are defined as follows:

$$\text{(a)} : \omega_1 = 2\omega_3, \quad \omega_2 = \frac{3}{2}\omega_3, \quad \omega_3 = 0.2n, \quad \omega_4 = \omega_5 = 0,$$
$$\text{(b)} : \omega_1 = 2\omega_3, \quad \omega_2 = \frac{3}{2}\omega_3, \quad \omega_3 = 0.2np, \quad \omega_4 = \omega_5 = 0.$$

These are corresponding to the two types of noncentrality matrix-1 and -2 in A4 and A5. Several different values of n and $p = cn$, were prepared for Monte Carlo simulations. We give simulation results with 10^4 repetitions for

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

The values of p/n in these cases are all $1/6$. The assumptions A6 and A7 are sasified.

Table 4.1. Selection probabilities of the true model by AIC in DA under (a)

n	p	A_0	A_1	A_2	A_3	A_4	A_5
30	5	0.05	0.29	0.41	0.21	0.04	0.00
60	10	0.01	0.17	0.48	0.32	0.03	0.00
120	20	0.00	0.07	0.50	0.42	0.01	0.00
210	35	0.00	0.02	0.46	0.52	0.00	0.00
300	50	0.00	0.01	0.41	0.58	0.00	0.00
480	80	0.00	0.00	0.34	0.66	0.00	0.00
600	100	0.00	0.00	0.30	0.70	0.00	0.00

Tabe 4.2. Selection probabilities of the true model by C_p in DA under (a)

n	p	C_0	C_1	C_2	C_3	C_4	C_5
30	5	0.01	0.17	0.45	0.31	0.06	0.01
60	10	0.00	0.06	0.40	0.47	0.06	0.00
120	20	0.00	0.01	0.29	0.65	0.05	0.00
210	35	0.00	0.00	0.18	0.80	0.02	0.00
300	50	0.00	0.00	0.12	0.88	0.01	0.00
480	80	0.00	0.00	0.05	0.94	0.00	0.00
600	100	0.00	0.00	0.03	0.97	0.00	0.00

Table 4.3. Selection probabilities of the true model by AIC in DA under (b)

n	p	A_0	A_1	A_2	A_3	A_4	A_5
30	5	0.00	0.00	0.01	0.79	0.18	0.03
60	10	0.00	0.00	0.00	0.83	0.16	0.01
120	20	0.00	0.00	0.00	0.94	0.06	0.00
210	35	0.00	0.00	0.00	0.99	0.01	0.00
300	50	0.00	0.00	0.00	1.00	0.00	0.00
480	80	0.00	0.00	0.00	1.00	0.00	0.00
600	100	0.00	0.00	0.00	1.00	0.00	0.00

Table 4.4. Selection probabilities of the true model by C_p

in DA under (b)

n	p	C_0	C_1	C_2	C_3	C_4	C_5
30	5	0.00	0.00	0.00	0.75	0.21	0.03
60	10	0.00	0.00	0.00	0.75	0.24	0.01
120	20	0.00	0.00	0.00	0.85	0.15	0.00
210	35	0.00	0.00	0.00	0.94	0.06	0.00
300	50	0.00	0.00	0.00	0.97	0.03	0.00
480	80	0.00	0.00	0.00	1.00	0.00	0.00
600	100	0.00	0.00	0.00	1.00	0.00	0.00

The simulation results under the noncentrality matrix-1 are given in Tables 4.1 and 4.2. The simulation results under the noncentrality matrix-2 are given in Tables 4.3 and 4.4. From Tables 4.1 $\sim$ 4.4 we can see that AIC and C_p are consistent for the estimation of dimensionality in the high-dimensional settings considered. On the velocity of convergence to the true dimension, the case of noncetrality matrix-2 is faster than the one of noncetrality matrix-1. Further, in the case of noncetrality matrix-1 C_p is faster than AIC. Tables 4.4 and 4.5 shows that there are some possibilities for the two criteria to overestimate the dimension when p is small, but n is large. On the other hand, when p becomes increasing, but n is not so large, there are some possibilities for the two criteria to underestimate the dimension.

4.2. CCA(Canonical Correlation Analysis)

We consider the selection probabilities of the true model by AIC and C_p in CCA. The setting similar to the one as in DA is considered. It is assumed that $q = 5$ and the true dimension is $j_0 = 3$. The two types of population canonical correlations are defined as follows:

$$\text{(c) } \rho_1 = 2\rho, \quad \rho_2 = \frac{3}{2}\rho, \quad \rho_3 = \rho, \quad \rho_4 = \rho_5 = 0,$$

$$\text{(d) } \rho_1 = \tilde{\rho}, \quad \rho_2 = \frac{3}{4}\tilde{\rho}, \quad \rho_3 = \frac{1}{2}\tilde{\rho}, \quad \rho_4 = \rho_5 = 0.$$

where

$$\rho = \sqrt{\frac{(4p)/(21)}{p+1+(4p)/(21)}}, \quad \tilde{\rho} = \sqrt{\frac{p}{p+1}}\sqrt{\frac{(4p)/(21)}{1+(4p)/(21)}}.$$

These are corresponding to the two types of canonical correlations-1 and -2 in B4 and B5. The values of n and $p = cn$, were chosen by the same way in MLM. Note that the assumptions A9 and A10 are satisfied. Our simulation results are given in Tables 4.5-4.8.

Table 4.5. Selection probabilities of the true model by AIC in CCA under (c)

n	p	A_0	A_1	A_2	A_3	A_4	A_5
30	5	0.00	0.14	0.48	0.31	0.05	0.01
60	10	0.00	0.03	0.47	0.45	0.05	0.00
120	20	0.00	0.00	0.43	0.55	0.02	0.00
210	35	0.00	0.00	0.39	0.61	0.00	0.00
300	50	0.00	0.00	0.36	0.64	0.00	0.00
480	80	0.00	0.00	0.31	0.69	0.00	0.00
600	100	0.00	0.00	0.30	0.70	0.00	0.00

Table 4.6. Selection probabilities of the true model by C_p in CCA under (c)

n	p	C_0	C_1	C_2	C_3	C_4	C_5
30	5	0.00	0.06	0.43	0.41	0.08	0.01
60	10	0.00	0.01	0.31	0.58	0.11	0.00
120	20	0.00	0.00	0.21	0.73	0.07	0.00
210	35	0.00	0.00	0.14	0.83	0.03	0.00
300	50	0.00	0.00	0.11	0.88	0.01	0.00
480	80	0.00	0.00	0.05	0.94	0.00	0.00
600	100	0.00	0.00	0.04	0.96	0.00	0.00

Table 4.7. Selection probabilities of the true model by AIC

in CCA under (d)

n	p	A_0	A_1	A_2	A_3	A_4	A_5
30	5	0.01	0.20	0.49	0.26	0.04	0.01
60	10	0.00	0.01	0.40	0.52	0.06	0.00
120	20	0.00	0.00	0.21	0.76	0.03	0.00
210	35	0.00	0.00	0.09	0.91	0.01	0.00
300	50	0.00	0.00	0.04	0.96	0.00	0.00
480	80	0.00	0.00	0.01	0.99	0.00	0.00
600	100	0.00	0.00	0.00	1.00	0.00	0.00

Table 4.8. Selection probabilities of the true model by C_p in CCA under (d)

n	p	C_0	C_1	C_2	C_3	C_4	C_5
30	5	0.00	0.09	0.47	0.36	0.07	0.01
60	10	0.00	0.00	0.25	0.62	0.12	0.01
120	20	0.00	0.00	0.08	0.82	0.09	0.00
210	35	0.00	0.00	0.02	0.94	0.04	0.00
300	50	0.00	0.00	0.01	0.98	0.02	0.00
480	80	0.00	0.00	0.00	1.00	0.00	0.00
600	100	0.00	0.00	0.00	1.00	0.00	0.00

The high-dimensional settings considered are corresponding to the ones in discriminant analysis, though there are some differences for largeness of the characteristics of the noncentrality matrix and the population canonical correlations. The simulation results are given in Tables 4.5 $\sim$ 4.8. It is pointed that there are similar tendencies in discriminant analysis.

From Tables 4.1 $\sim$ 4.4 we can see that AIC and C_p are consistent for the estimation of dimensionality in the high-dimensional settings considered. On the velocity of convergence to the true dimension, the case of noncetrality matrix-2 is faster than the one of noncetrality matrix-1. Further, In the case of noncetrality matrix-1 C_p is faster than AIC. Tables 4.4 and 4.5 shows

that there are some possibilities of underestimating the dimensionality when (n, p) is not so large.

5. Concluding Remarks

In general, it is known that under the large sample asymptotic framework (2.17) AIC and C_p have no consistency prorperty, 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 dimensionalities in multivariate linear model and canonical correlation analysis have a consistency property, under a high-dimensional framework (1.1). 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 $c \in [0, 1/2]$. Further, the consitency was considered under two types of assumptions on the largeness of the characteristic roots of the noncentrality matrix and the population canonical correlations.

In discriminant analysis the number of groups may be assumed to be finite. However, in multivariate regression model and canonical correlation analysis the number k of explanatory variables and the number q of $\boldsymbol{x}$ may be large. For such cases, we note that the probability of estimating overspecified dimensions tends to zero under some condition. These results are expected to be extended in the future. In a very high-dimensional case, it will occur that $p > n$. In this case, AIC and C_p should be modified by using, for example, ridge estimators. On the other hand, recently some other methods have been proposed by Yuan and Ekici(2007), Bunea, She and Wegkamp (2011), etc. Bunea, She and Wegkamp (2012) and Chen and Hung (2012) also consider simultaneous methods for dimension reduction and variable selection.

Appendix

A. The proofs of Theorems 2.1, 2.2 and 2.3

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 A.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*

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

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 $\lim_{p/n\to c}\delta_j=\delta_{j0}>0$. Then*

$$\ell_j\xrightarrow{p}\frac{c}{1-c}+\frac{1}{1-c}\delta_{j0},\quad j=1,\ldots,a,$$
$$\ell_j\xrightarrow{p}\frac{c}{1-c},\quad j=a+1,\ldots,q.$$

(2) *Suppose that for any $j(0\leq j\leq a)$, $\omega_j=np\xi_j=\mathrm{O}(np)$, and $\lim_{p/n\to c}\xi_j=\xi_{j0}>0$. Then*

$$\frac{1}{p}\ell_j\xrightarrow{p}\frac{1}{1-c}\xi_{j0},\quad j=1,\ldots,a,$$
$$\ell_j\xrightarrow{p}\frac{c}{1-c},\quad j=a+1,\ldots,q.$$

Proof. It is known (see Fujikoshi et al. (2013)) 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 $\mathrm{W}_q(m, \mathbf{I}_q; \mathbf{\Omega})$, 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 $\mathbf{\Delta} = n\mathrm{diag}(\delta_1, \ldots, \delta_a, 0, \cdots, 0)$, 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

$$\begin{aligned}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} &= \frac{p}{m}(\frac{1}{m}\mathbf{W})^{-1/2}(\frac{1}{p}\mathbf{B})(\frac{1}{m}\mathbf{W})^{-1/2} \\ &\xrightarrow{p} \left(\mathbf{I}_q + \frac{1}{c}\Delta_0\right)\frac{c}{1-c},\end{aligned}$$

where $\lim \mathbf{\Delta} = \mathbf{\Delta}_0$, and $\mathbf{\Delta}_0 = \mathrm{diag}(\delta_{10}, \ldots, \delta_{a0}, 0, \ldots, 0)$. 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

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

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 (A.2) it is easy to see that

$$\frac{1}{p}(\ell_1, \ldots, \ell_a) \xrightarrow{p} \frac{1}{1-c}(\xi_{10}, \ldots, \xi_{a0}).$$

Further, applying Lawley (1959) to (A.2), 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 2.1

In the proof of Theorem 2.1 it is assumed that the true dimensionality is j_0. Since the number of possible models is finite, it is sufficient to show that the values of $\mathrm{AIC}_j - \mathrm{AIC}_{j_0}$ converges to positive values.

Note that for $j > j_0$

$$\mathrm{AIC}_j - \mathrm{AIC}_{j_0} = -n\log(1+\ell_{j_0+1})\cdots(1+\ell_j) + 2(j-j_0)(p+q-j-j_0),$$

and for $j < j_0$

$$\mathrm{AIC}_j - \mathrm{AIC}_{j_0} = n\log(1+\ell_{j+1})\cdots(1+\ell_{j_0}) + 2(j-j_0)(p+q-j-j_0).$$

Suppose that $\mathbf{\Omega} = \mathrm{O}(n)$. Then, using Lemma A.1 (1), we have that for $j > j_0$

$$\frac{1}{n}\left\{\mathrm{AIC}_j - \mathrm{AIC}_{j_0}\right\} \xrightarrow{p} (j - j_0)\left\{\log(1-c) + 2c\right\}.$$

The limiting value is positive when $c \in (0, c_a)$. The proof in the case $c = 0$ may be modified as follows. For $0 < c < 1$, we have

$$\begin{aligned}\frac{1}{p}\{\mathrm{AIC}(j) - \mathrm{AIC}(j_0)\} &\xrightarrow{p} \frac{1}{c}(j - j_0)\left\{\log(1-c) + 2c\right\} \\ &= (j - j_0)\left\{\frac{1}{c}\log(1-c) + 2\right\}.\end{aligned}$$

Noting that $\lim_{c\to 0+} c^{-1}\log(1-c) = -1$, it holds that the limiting value is positive. Next suppose that $j < j_0$. Then, using Lemma A.1 (1), we have

$$\begin{aligned}&\frac{1}{n}\left\{\mathrm{AIC}_j - \mathrm{AIC}_{j_0}\right\} \\ &\quad\to \log(1+\delta_{j+1})\cdots(1+\delta_{j_0}) - (j_0 - j)\left\{\log(1-c) + 2c\right\} \\ &\quad\geq (j_0 - j)\left[\log(1+\delta_{j_0}) - \left\{\log(1-c) + 2c\right\}\right]\end{aligned}$$

The lower bound is positive by the assumption A6. The case $c = 0$ is similarly proved as in the case $j > j_0$.

Now we shall prove the result (2). For $j > j_0$, 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 $(1/n)\left\{\mathrm{AIC}_j - \mathrm{AIC}_{j_0}\right\}$ is positive when $c \in [0, c_a)$. For $j < j_0$, from Lemma A.1 (2) we have

$$\frac{1}{np}\{\mathrm{AIC}(j) - \mathrm{AIC}(j_0)\} \xrightarrow{p} j_0 - j.$$

This proves Theorem 2.1 (2).

The Proof of Theorem 2.2

Theorem 2.2 is proved by the same way as in Theorem 2.1, due to Lemma A.1. In the following we give an outline of the proof. We have that for $j > j_0$

$$\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0} = -n(\ell_{j_0+1} + \cdots + \ell_j) + 2(j - j_0)(p + q - j - j_0),$$

and for $j < j_0$

$$\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0} = n(\ell_{j+1} + \cdots + \ell_{j_0}) + 2(j - j_0)(p + q - j - j_0).$$

Noting that for $j > j_0$, $\ell_j \to c/(1-c)$ under both cases $\mathbf{\Omega} = \mathrm{O}(n)$ and $\mathbf{\Omega} = \mathrm{O}(np)$, it holds that

$$\frac{1}{p}\left\{\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0}\right\} \xrightarrow{p} (j - j_0)\frac{1-2c}{1-c}$$

whose limiting value is positive when $c \in [0, 1/2)$. When $j < j_0$ and $\mathbf{\Omega} = \mathrm{O}(n)$,

$$\begin{aligned}\frac{1}{n}\left\{\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0}\right\} &\xrightarrow{p} \frac{1}{1-c}(\delta_{j+1} + \cdots + \delta_{j_0}) + +(j_0 - j)\left\{\frac{c}{1-c} - 2c\right\} \\ &\geq \frac{(j - j_0)}{1-c}\left\{\delta_{j_0} - c(1-2c)\right\}.\end{aligned}$$

Further, when $j < J_0$ and $\mathbf{\Omega} = \mathrm{O}(np)$,

$$\frac{1}{n}\left\{\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0}\right\} \xrightarrow{p} \frac{1}{1-c}(\xi_{j+1,0} + \cdots + \xi_{j_0 0}) > 0.$$

These imply Theorem 2.2.

The Proof of Theorem 2.3

In general, it holds that

$$\begin{aligned}&P(\hat{j}_A = j) \leq P(\mathrm{AIC}_j - \mathrm{AIC}_{j_0} < 0), \\ &\sum_{j=j_0+1}^{q} P(\hat{j}_A = j) \leq \sum_{j=j_0+1}^{q} P(\mathrm{AIC}_j - \mathrm{AIC}_{j_0} < 0).\end{aligned}$$

So, it is sufficient to show

$$\lim_{q\to\infty} \sum_{j=j_0+1}^{q} P(\mathrm{AIC}_j - \mathrm{AIC}_{j_0} < 0) \to 0.$$

For any $j(\leq q)$, we have

$$\begin{aligned}&P(\mathrm{AIC}_j - \mathrm{AIC}_{j_0} < 0) \\ &= P(n\log(1+\ell_{j_0+1})\cdots(1+\ell_j) > 2(j-j_0)(p+q-j-j_0)) \\ &\leq P(n\log(1+\ell_{j_0+1})\cdots(1+\ell_q) > 2(j-j_0)(p+q-j-j_0)).\end{aligned}$$

Here, $\ell_1 > \cdots > \ell_q$ are the characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$. We have noted that $\ell_1 > \cdots > \ell_q$ may be regarded as the characteristic roots of $\mathbf{B}\mathbf{W}^{-1}$, where $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as a central Wishart distribution $\mathrm{W}_q(m, \mathbf{I}_q)$ with $m = n - k - p + q$ and a noncentral Wishart distribution $\mathrm{W}_q(p, \mathbf{I}_q; \mathbf{\Omega})$, respectively. Note that $\mathbf{\Omega} = \mathrm{diag}(\omega_1, \cdots, \omega_{j_0}, 0, \ldots, 0)$. Consider a $q \times (q - j_0)$ matrix $\mathbf{H}$ such that $\mathbf{H}'\mathbf{\Omega} = \mathrm{O}$ and $\mathbf{H}'\mathbf{H} = \mathbf{I}_{q-j_0}$. Let $\tilde{\mathbf{B}} = \mathbf{H}'\mathbf{B}\mathbf{H}$ and $\tilde{\mathbf{W}} = \mathbf{H}'\mathbf{W}\mathbf{H}$. Then, $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{W}}$ are independently distributed as a Wishart distribution $\mathrm{W}_{q-j_0}(p, \mathbf{I}_q)$ and a Wishart distribution $\mathrm{W}_{q-j_0}(m, \mathbf{I}_q)$, respectively. Let the characteristic roots of $\tilde{\mathbf{B}}\tilde{\mathbf{W}}^{-1}$ denote by $\tilde{\ell}_1 > \cdots > \tilde{\ell}_{q-j_0}$. Then, using Saw (1974) (see Fujikoshi and Isogai (1976), Olkin and Tosky (1981), Schott (1984)), we have

$$\tilde{\ell}_i \geq \ell_{j_0+i}, \quad i = 1, \ldots, q - j_0,$$

which implies

$$\begin{aligned} & P(n\log(1+\ell_{j_0+1})\cdots(1+\ell_q) > 2(j-j_0)(p+q-j-j_0)) \\ & \leq P(n\log(1+\tilde{\ell}_1)\cdots(1+\tilde{\ell}_{q-j_0}) > 2(j-j_0)(p+q-j-j_0)) \\ & = P(-n\log\Lambda > 2(j-j_0)(p+q-j-j_0)). \end{aligned} \tag{A.3}$$

Here, Λ is defined by

$$\Lambda = \frac{|\tilde{\mathbf{W}}|}{|\tilde{\mathbf{W}} + \tilde{\mathbf{B}}|}$$

which is distributed as a Lambda distribution $\Lambda_{q-j_0}(p, m)$. Now, using the bellow Lemma A.3 we have that the last expression of (A.3) is $\mathrm{O}((pq)^{-\ell})$ for any positive number, under the assumption of Theorem 2.3. This proves Theorem 2.3.

The following Lemma is obtained from Lemma A.1 in Yanagihara, Wakaki and Fujikoshi (2014).

Lemma A.2. *Let $T = -n(pq)^{-1}\log\Lambda$, where Λ is distributed to Lambda distribution $\Lambda_q(p, n+q)$. Let $c_{p,n} = p/n$ and $\hat{\kappa} = c_{p,n}^{-1}\log(1 + c_{p,\ell})$. Assume*

the asymptotic framework given by

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

Let m be a constant depending on n, p and q. If there exists a positive constant δ such that $m - \hat{\kappa} > \delta$ for large n, then

$$P(T > m) = \mathrm{O}((pq)^{-\ell}),$$

for any positive number ℓ.

B. The proofs of Theorems 3.1, 3.2 and 3.3

First we consider the limiting behavior of the squares $r_1^2 > \cdots > r_q^2$ of the canonical correlations under a high-dimensional framework.

Lemma A.3. *Let $r_1^2 > \cdots > r_q^2$ and $\rho_1^2 \geq \cdots \geq \rho_q^2$ be the squares of the sample and population canonical correlations between $\boldsymbol{x}; p \times 1$ and $\boldsymbol{y}; q \times 1$ with $p \geq q$, based on a sample of size $N = n + 1$ from $\mathrm{N}_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $d_j^2 = r_j^2/(1 - r_j^2)$, $\gamma_j^2 = \rho_j^2/(1 - \rho_j^2)$, $j = 1, \ldots, q$. We assume that the number of nonzero population canonical correlations is a, and hence $\rho_1 \geq \cdots \geq \rho_a > \rho_{a+1} = \cdots \rho_q = 0$. For the limiting behaviors of $r_1^2 > \cdots > r_q^2$ and $d_1^2 > \cdots > d_q^2$ under a high-dimensional asymptotic framework*

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

we have the following results:

(1) *Suppose that for any $j (0 \leq j \leq a)$, $\rho_j^2 = \mathrm{O}(1)$ and $\lim_{p/n \to c} \rho_j^2 = \rho_{j0}^2 > 0$. Then*

$$r_j^2 \xrightarrow{p} \rho_j^2 + c(1 - \rho_j^2), \quad d_j^2 \xrightarrow{p} \frac{c}{1-c} + \frac{1}{1-c}\gamma_{j0}^2; \quad j = 1, \ldots, a,$$
$$r_j^2 \xrightarrow{p} c, \quad d_j^2 \xrightarrow{p} \frac{c}{1-c}; \quad j = a+1, \ldots, q.$$

(2) *Suppose that for any* $j(0 \le j \le a)$, $\gamma_j^2 = p\theta_j^2 = \mathrm{O}(p)$, *and* $\lim_{p/n \to c} \theta_j = \theta_{j0} > 0$. *Then*

$$\frac{1}{p}\frac{r_j^2}{1-r_j^2} \xrightarrow{p} \frac{1}{1-c}\theta_{j0}^2, \quad \frac{1}{p}d_j^2 \xrightarrow{p} \frac{1}{1-c}\theta_{j0}^2; \quad j = 1, \ldots, a,$$
$$r_j^2 \xrightarrow{p} c, \quad d_j^2 \xrightarrow{p} \frac{c}{1-c}; \quad j = a+1, \ldots, q. j = a+1, \ldots, q.$$

Proof. Let $\mathbf{A} = n\mathbf{S}$, which is distributed as a Wishart distribution $\mathrm{W}_{q+p}(n, \Sigma)$ with $\mathbf{\Sigma}$ given by (3.6). We partition $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

corresponding to a partition of $\mathbf{S}$. In the following we consider the limiting behavior of $d_1^2 > \cdots > d_q^2$ defined by $d_j^2 = r_j^2/(1-r_j^2)$, which are the characteristic roots of $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot1}^{-1}$, instead of instead of $r_1^2 > \cdots > r_q^2$. Here $\mathbf{A}_{22\cdot1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. It is known (see Fujikoshi and Sakurai (2010)) that (i) $\mathbf{A}_{22\cdot1} \sim \mathrm{W}_q(m, \mathbf{I}_q)$, where $m = n - p$. (ii) $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \sim \mathrm{W}_q(p, \mathbf{I}_p; \mathbf{\Gamma}\mathbf{G}\mathbf{\Gamma})$, when the first $q \times q$ matrix $\mathbf{G}$ of $\mathbf{A}_{11}$ is given. Here, $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_q)$. Further, $\mathbf{G} \sim \mathrm{W}(n, \mathbf{I}_q)$. (iii) $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\mathbf{A}_{22\cdot1}$ are independent. Let

$$\mathbf{U} = \frac{1}{\sqrt{p}}(\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - p\mathbf{I}_q - \mathbf{\Omega}), \quad \mathbf{\Omega} = \mathbf{\Gamma}\mathbf{G}\mathbf{\Gamma},$$
$$\mathbf{V} = \frac{1}{\sqrt{m}}(\mathbf{A}_{22\cdot1} - m\mathbf{I}_q), \quad \mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{G} - n\mathbf{I}_q).$$

Then the limiting distributions of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{Z}$ are normal. When $\mathbf{\Gamma} = \mathrm{O}(1)$, we have

$$\begin{aligned}(\mathbf{A}_{22\cdot1})^{-1/2}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22\cdot1})^{-1/2} &= \frac{p}{m}\left(\mathbf{I}_q + \frac{1}{\sqrt{m}}\mathbf{V}\right)^{-1/2} \\ &\times \left(\mathbf{I}_q + \frac{n}{p}\mathbf{\Gamma}^2 + \frac{n}{p\sqrt{n}}\mathbf{Z} + \frac{1}{\sqrt{n}}\mathbf{U}\right)\left(\mathbf{I}_q + \frac{1}{\sqrt{m}}\mathbf{V}\right)^{-1/2} \\ &\xrightarrow{p} \frac{c}{1-c}\left(\mathbf{I}_q + \frac{1}{c}\mathbf{\Gamma}_0^2\right),\end{aligned}$$

where $\lim \mathbf{\Gamma} = \mathbf{\Gamma}_0$, and $\mathbf{\Gamma}_0 = \mathrm{diag}(\gamma_{10}, \ldots, \gamma_{a0}, 0, \ldots, 0)$. From the result we can see the first result (1).

Next we consider the case $\mathbf{\Gamma} = \sqrt{p}\mathbf{\Theta}$. We have

$$\frac{1}{np}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \mathbf{\Theta}^2 + \frac{1}{\sqrt{n}}\mathbf{\Theta}\mathbf{Z}\mathbf{\Theta} + \frac{1}{n}\mathbf{I}_q + \frac{1}{n\sqrt{p}}\mathbf{U},$$
$$\left(\frac{1}{m}\mathbf{A}_{22\cdot 1}\right)^{-1/2} = \mathbf{I}_q - \frac{1}{2\sqrt{m}}\mathbf{V} + \frac{3}{8m}\mathbf{V}^2 + \mathrm{O}(m^{-3/2}).$$

Therefore

$$\frac{m}{np}\mathbf{A}_{22\cdot 1}^{-1/2}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1/2} = \begin{pmatrix} \mathbf{\Theta}_1^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} + \frac{1}{\sqrt{m}}\begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{pmatrix}, \qquad \text{(B.4)}$$

where $\mathbf{\Theta}_1 = \mathrm{diag}(\theta_1, \ldots, \theta_a)$,

$$\mathbf{T}_{11} = -\frac{1}{2}(\mathbf{V}_{11}\mathbf{\Theta}_1^2 + \mathbf{\Theta}_1^2\mathbf{V}_{11}) + \mathbf{\Theta}_1\mathbf{Z}\mathbf{\Theta}_1 + \mathrm{O}(m^{-1/2}),$$
$$\mathbf{T}_{12} = \mathbf{T}_{21}' = -\frac{1}{2}\mathbf{\Theta}_1^2\mathbf{V}_{11} + \mathrm{O}(m^{-1/2}),$$
$$\mathbf{T}_{22} = \frac{\sqrt{m}}{n}\mathbf{I}_{q-a} + \frac{1}{4\sqrt{m}}\mathbf{V}_{21}\mathbf{\Theta}_1^2\mathbf{V}_{12} + \mathrm{O}(m^{-1}).$$

From (B.4) it is easy to see that

$$\frac{1}{p}(d_1^2, \ldots, d_a^2) \xrightarrow{p} \frac{1}{1-c}(\theta_{10}^2, \ldots, \theta_{a0}^2).$$

Further, applying Lawley (1959) to (B.4), the larst $q-a$ characteristic roots of $\{m/(np)\}(d_{a+1}^2, \ldots, \ell_q^2)$ are the same as the ones of

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

This shows that $d_j^2 \xrightarrow{p} c/(1-c)$ for $j = a+1, \ldots, q$. □

The proofs of Theorems 3.1 and 3.2

We can prove by the same way as in the proof of Theorems 2.1 and 2.2. In fact, noting that $1 - r_j^2 = (1 + d_j^2)^{-1}$, the difference between AIC_j and AIC_{j_0} is expressed in terms of $d_1^2, \ldots, d_q^2$ as, for $j > j_0$

$$\mathrm{AIC}_j - \mathrm{AIC}_{j_0} = -n \log(1 + d_{j_0+1}^2) \cdots (1 + d_j^2) + 2(j - j_0)(p + q - j - j_0),$$

and for $j < j_0$

$$\mathrm{AIC}_j - \mathrm{AIC}_{j_0} = n \log(1 + r_{j+1}^2) \cdots (1 + d_{j_0}^2) + 2(j - j_0)(p + q - j - j_0).$$

Similarly the difference between $\mathrm{C}_{p,j}$ and C_{p,j_0} is expressed in terms of $d_1^2, \ldots, d_q^2$ as, for $j > j_0$

$$\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0} = -n(d_{j_0+1}^2 + \cdots + d_j^2) + 2(j - j_0)(p + q - j - j_0),$$

and for $j < j_0$

$$\mathrm{C}_{p,j} - \mathrm{C}_{p,j_0} = n(d_{j+1}^2 + \cdots + d_{j_0}^2) + 2(j - j_0)(p + q - j - j_0).$$

The limiting behaivior of $(d_1^2, \ldots, d_q^2)$ is given in Lemma A.3, which are corresponding to the one of $(\ell_1, \ldots, \ell_q)$ in multivariate linear model.

The proof of Theorem 3.3

We can prove by a parallel argument as in the proof of Theorem 2.3, using the conditional set-up stated in the proof of Lemma A.3 and a probability inequality in Fujikoshi and Isogai (1976). Its details are omitted.

Acknowledgements

We wish to thank Dr.Tetsuro Sakurai for his help of simulation study in Section 4. The author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #25330038, 2013-2015.

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

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

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

[7] Fujikoshi, Y. and Isogai, T. (1976). Lower bounds for the distributions of certain multivariate test statistics. *J. Multivariate Anal.*, **6**, 250–255.

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

[9] 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.

[10] Fujikoshi, Y. and Sakurai, T. (2008). High-dimensional asymptotic expansions for the distributions of canonical correlations. *J. Multivariate Anal.*, **100**, 231–241.

[11] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, Hobeken, 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. (2008). *Modern Multivariate Statistical Techniques*. Springer, New York.

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

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

[19] Olkin, I. and Tomskey, J. L. (1981). A new class of multivariate tests based on the union-intersection principle. *Ann. Statist.*, **9**, 792–802.

[20] Schott, J. R. (1984). Optimal bounds for the distributions of sme tests criteria for tetsts of dimensionality. *Biometrika*, **71**, 561–567.

[21] Saw, G. (1974). A lower bound for the distribution of a partial product of latent roots. *Commun. Statist.*, **3**, 665–669.

[22] 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.

[23] 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.