Journal of Multivariate Analysis 149 (2016) 199–212

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

High-dimensional consistency of rank estimation criteria in multivariate linear model

Yasunori Fujikoshi [a], Tetsuro Sakurai [b,*]

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

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

ARTICLE INFO

Article history:
Received 13 April 2015
Available online 6 May 2016

AMS 2000 subject classifications:
primary 62H12
secondary 62H10

Keywords:
AIC
BIC
C_p
Consistency property
Dimensionality
Discriminant analysis
High-dimensional framework
Multivariate regression model
Multivariate linear model
Rank
Ridge-type criterion
Tuning parameter

ABSTRACT

This paper is concerned with consistency properties of rank estimation criteria in a multivariate linear model, based on the model selection criteria AIC, BIC and C_p. The consistency properties of these criteria are studied under a high-dimensional framework with two different assumptions on the noncentrality matrix such that the number of response variables and the sample size tend to infinity. In general, it is known that under a large-sample asymptotic framework, the criteria based on AIC and C_p are not consistent, but the criterion based on BIC is consistent. However, we note that there are cases that the criteria based on AIC and C_p are consistent, but the criterion based on BIC is not consistent. Such consistency properties are also shown for the generalized criteria with a tuning parameter. Further, the modified criteria with a ridge-type estimator are also examined. Through a Monte Carlo simulation experiment, our results are checked numerically, and the estimation criteria are compared.

1. Introduction

This paper is concerned with the problem of estimating the rank (or dimensionality) of a regression coefficient matrix in a multivariate linear model. A multivariate regression model with a reduced rank is called a multivariate reduced-rank regression model (see Izenman [15], Reisel and Velu [19]). Our problem includes also that of estimating the number of meaningful discriminant functions in discriminant analysis.

One commonly used rank estimation method is based on sequential test procedures. In this method, the test in each step is a likelihood ratio test which was first proposed by Anderson [2], [3]. Recently, related to sparse reduced-rank regression, some penalized methods have been studied. Yuan et al. [23] used the Ky-Fan norm penalty for factor selection and shrinkage. Chen and Huang [8] proposed a simultaneous method for selecting the rank and the explanatory variables by using penalized regression with a group lasso penalty. For related works, see Bunea et al. [6], [7]. These penalized methods all involve finding the multivariate least squares under a restriction.

* Corresponding author.
E-mail addresses: fujikoshi_y@yahoo.co.jp (Y. Fujikoshi), sakurai@rs.tus.ac.jp (T. Sakurai).

http://dx.doi.org/10.1016/j.jmva.2016.04.005
0047-259X/

There are other methods based on the use of model selection criteria as follows. In a general multivariate regression, we have n observations $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ on p response variables $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ and q explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_k)^\top$, and

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \boldsymbol{\mathcal{E}}, \tag{1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ regression coefficient matrix, and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$ is the error matrix, with the $\boldsymbol{\epsilon}_i$'s independently distributed as $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$.

Our interest is to estimate the rank of $\mathbf{C}\boldsymbol{\Theta}$, where $\mathbf{C}$ is a given $q \times k$ matrix with rank q, and hence $q \leq k$. As the model selection approach, we consider the models $M_j, j = 0, \ldots, m = \min(p, q)$, where M_j is the model (1) with $\mathrm{rank}(\mathbf{C}\boldsymbol{\Theta}) = j$. If M_j is selected by a model selection criterion, we estimate the rank as j. We are concerned with rank estimation methods that use the following model selection criteria: AIC (Akaike [1]), C_p (Mallows [18]) and BIC (Schwarz [20]). The rank estimation criteria based on AIC and C_p were proposed by Fujikoshi and Veitch [13] in multivariate linear model and canonical correlation analysis. In this context, Gunderson and Muirhead [14] consider a rank estimation criterion based on BIC. These rank estimation methods are denoted herein as AIC, BIC and C_p, for simplicity.

In general, a selection criterion is said to be consistent if the probability of selecting the true model tends to 1 under an asymptotic framework. Our rank estimation criteria depend on p, q, k, n and $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q)$, where $\omega_1 \geq \cdots \geq \omega_q \geq 0$ are the possible nonzero characteristic roots of the (standardized) noncentrality matrix $\tilde{\boldsymbol{\Omega}}$ in (7). It is known (Fujikoshi [9]) that, under the large-sample asymptotic framework

$$n \to \infty; \quad p,\ q,\ k : \text{fixed}, \tag{2}$$

with $\Omega = \mathrm{O}(n)$, AIC and C_p are not consistent. Here, for a matrix $\mathbf{A}$, $\mathbf{A} = \mathrm{O}(n)$ means that each element of $\mathbf{A}$ is order n. Gunderson and Muirhead [14] showed the consistency of BIC in canonical correlation analysis.

One of our purposes is to study consistency properties of AIC, BIC and C_p when p/n tends to c. More precisely, we consider the high-dimensional asymptotic framework given by

$$q : \text{fixed}, \quad k \to \infty,\ k/n \to 0,\ p \to \infty,\ n \to \infty,\ p/n \to c \in [0, 1), \tag{3}$$

with each for (i) $\boldsymbol{\Omega} = \mathrm{O}(n)$ and (ii) $\boldsymbol{\Omega} = \mathrm{O}(np)$. In the present study, it is shown that the criteria AIC and C_p are consistent under some additional assumptions depending on (i) and (ii). Further, we find that the criterion BIC is consistent under $\boldsymbol{\Omega} = \mathrm{O}(np)$, but that it is not consistent under $\boldsymbol{\Omega} = \mathrm{O}(n)$.

Instead of the criteria AIC, BIC and C_p, we may use the criteria A, B and C defined by

$$\mathrm{A}_j = \mathrm{AIC}_j - \mathrm{AIC}_p, \qquad \mathrm{B}_j = \mathrm{BIC}_j - \mathrm{BIC}_m, \qquad \mathrm{C}_j = \mathrm{C}_{p,j} - \mathrm{C}_{p,m}.$$

For the details of these criteria, refer to (12), (13) and (14) in Section 3. In this paper, we propose generalizations and modifications of A, B and C. One such generalization is the criteria IC_ν and $\mathrm{C}_{p,\nu}$ with a tuning parameter ν, where $\mathrm{IC}_2 = \mathrm{A}$, $\mathrm{IC}_{\ln n} = \mathrm{B}$ and $\mathrm{C}_{p,2} = \mathrm{C}$. The other is the ridge-type criteria of A, B and C denoted by A_λ, B_λ and C_λ, respectively. The ridge-type criteria are defined by using the estimator of the covariance matrix with a ridge parameter. The criteria are defined also for the case $p > n - k$. Our notation, for example, A_j, A_ν, A_λ, might cause confusion. However, note that we use i and j for possible ranks, ν for a tuning parameter, and λ for the ridge parameter. We derive sufficient conditions for the generalized criteria IC_ν and $\mathrm{C}_{p,\nu}$ to be consistent under the high-dimensional asymptotic framework (3). Such sufficient conditions will be useful in selecting the tuning parameter ν. For $n - k > p$, we note that the ridge-type criteria A_λ, B_λ and C_λ have the same consistency properties as those of A, B and C under a high-dimensional asymptotic framework. We check our theoretical results through a Monte Carlo simulation experiment and compare numerically the selection probabilities of the criteria A, B and C with their modifications. For $p > n - k$, some tendencies in the ridge-type criteria are clarified through numerical experiments.

Our results may be applied to data sets in multivariate regression model and discriminant analyses with a relatively large number of response variables, which are carried out in many fields, including medicine. Examples of data with a very large number of response variables include stock data and genomic data. We may also apply our results to the analysis of data obtained by choosing fewer response variables less than the sample size. Similar high-dimensional consistency properties have been derived by Fujikoshi et al. [11] and Yanagihara et al. [22] for AIC, BIC and C_p in selection of the explanatory variables in a multivariate linear model.

The present paper is organized as follows. In Section 2, we present a reduced-rank multivariate linear model, and two important special cases are explained. Then we prepare three criteria and their modifications 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 consistency properties are then theoretically and numerically examined. In Section 5, we discuss our conclusions. The proofs of our results are given in the Appendix.

2. Rank estimation criteria

2.1. Reduced-rank multivariate linear model

The multivariate regression model in (1) may be called a multivariate linear model when the x_i's are any fixed variables, including explanatory variables or dummy variables. The model may be expressed as

$$\mathbf{Y} \sim \mathcal{N}_{n\times p}(\mathbf{X\Theta},\ \mathbf{\Sigma}\otimes\mathbf{I}_n), \tag{4}$$

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 $\mathcal{N}_{n\times p}(\cdot,\cdot)$ means the matrix normal distribution such that the mean of $\mathbf{Y}$ is $\mathbf{X\Theta}$ 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 $\mathrm{rank}(\mathbf{X})=k$ and in most case, $n-k\geq p$. However, the restriction $n-k\geq p$ is not necessarily assumed when we treat the ridge-type criterion.

Let $\mathbf{C}$ be a given $q\times k$ matrix with $\mathrm{rank}(\mathbf{C})=q$, and consider reduced-rank models M_j; (1) or (4) with $\mathrm{rank}(\mathbf{C\Theta})=j,\ j=0,\ldots,m=\min(p,q)$. For simplicity, we write M_j as

$$M_j:\ \mathrm{rank}(\mathbf{C\Theta})=j,\quad j=0,\ldots,m. \tag{5}$$

Our interest is to estimate $\mathrm{rank}(\mathbf{C\Theta})$ by the criteria based on model selection criteria AIC, BIC and C_p. An LR statistic for $M_j:\ \mathrm{rank}(\mathbf{C\Theta})=j$ is given (see, e.g., Anderson [3]) by

$$\Lambda_{(j)}=\{(1+\ell_{j+1})\cdots(1+\ell_m)\}^{-1}, \tag{6}$$

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

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

and $\hat{\mathbf{\Theta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. Here, without loss of generality we may assume that $\ell_1>\cdots>\ell_m>0$, since the probability of any two ℓ_is being equal is 0. It is well known (see, e.g., Anderson [3]) that $\mathbf{S}_e$ and $\mathbf{S}_h$ are independently distributed as a Wishart distribution $\mathcal{N}_p(n-k,\ \mathbf{\Sigma})$ and a noncentral Wishart distribution $\mathcal{N}_p(q,\ \mathbf{\Sigma};\ \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively, where

$$\tilde{\mathbf{\Omega}}=\mathbf{\Sigma}^{-1/2}(\mathbf{C\Theta})^\top\{\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top\}^{-1}\mathbf{C\Theta}\mathbf{\Sigma}^{-1/2}. \tag{7}$$

The reduced-rank model includes the following two important special cases. One is a multivariate reduced-rank regression model which is given by (4) and (5) with $\mathbf{C}=\mathbf{I}_k$. From the rank constraint (5), the regression matrix $\mathbf{\Theta}$ can be expressed as a product of two rank j matrices as follows:

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

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

$$\mathrm{E}[\mathbf{Y}]=(\mathbf{XG})\cdot\mathbf{\Xi}=\mathbf{Z\Xi},\quad \mathbf{Z}=\mathbf{XG}. \tag{8}$$

The model means that the j linear combinations $\boldsymbol{z}=\mathbf{G}^\top\boldsymbol{x}$ of the k explanatory variables $\boldsymbol{x}$ are sufficient to model the variation 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.

The other special case is a reduced-rank problem in discriminant analysis, based on $(q+1)$ p-variate normal populations with common covariance matrix $\mathbf{\Sigma}$. 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. Then, the model of $\mathbf{Y}=(\boldsymbol{y}_{11},\ldots,\boldsymbol{y}_{1n_1},\ldots,\boldsymbol{y}_{q+1,1},\ldots,\boldsymbol{y}_{q+1,n_{q+1}})^\top$ is a model (4) with $k=q+1$ and

$$\mathbf{X}=\begin{pmatrix}\mathbf{1}_{n_1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{1}_{n_{q+1}}\end{pmatrix},\qquad \mathbf{\Theta}=\begin{pmatrix}\boldsymbol{\mu}_1^\top\\ \vdots\\ \boldsymbol{\mu}_{q+1}^\top\end{pmatrix}.$$

Now we consider the reduced-rank model with $\mathbf{C}=\begin{pmatrix}\mathbf{I}_q, & -\mathbf{1}_q\end{pmatrix}$, and $\mathbf{C\Theta}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{q+1},\ldots,\boldsymbol{\mu}_q-\boldsymbol{\mu}_{q+1})$. Then the matrices $\mathbf{S}_e$ and $\mathbf{S}_h$ are expressed as

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

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 $(1/n)\sum_{i=1}^{q+1}n_i\bar{\boldsymbol{y}}_i$, in which $n=\sum_{i=1}^{q+1}n_i$. Then, $\mathbf{S}_b$ and $\mathbf{S}_w$ are called the between-group and the within-group sums of squares and products matrices. In general, $\mathbf{S}_w$ and $\mathbf{S}_b$ are independently distributed as a Wishart distribution $\mathcal{N}_p(n-q-1,\ \mathbf{\Sigma})$ and a noncentral Wishart distribution $\mathcal{N}_p(q,\ \mathbf{\Sigma};\ \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively, where

$$\tilde{\mathbf{\Omega}}=\mathbf{\Sigma}^{-1/2}\sum_{i=1}^{q+1}n_i(\boldsymbol{\mu}_i-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\bar{\boldsymbol{\mu}})^\top\mathbf{\Sigma}^{-1/2},\quad \bar{\boldsymbol{\mu}}=(1/n)\sum_{i=1}^{q+1}n_i\boldsymbol{\mu}_i. \tag{9}$$

It is known (see, e.g., Kshirsagar [16]) that

$$M_j : \mathrm{rank}(\mathbf{C\Theta}) = j \;\Leftrightarrow\; \omega_1 \geq \cdots \geq \omega_j > \omega_{j+1} = \cdots = \omega_q = 0, \tag{10}$$

where $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_m \geq 0$ are the possible non-zero characteristic roots of $\tilde{\mathbf{\Omega}}$.

2.2. Criteria based on AIC, BIC, C_p and their modifications

In general, AIC for a model M is defined (Akaike [1]) as

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

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

$$\begin{aligned}\mathrm{AIC}_j &= n\ln(1+\ell_{j+1})\cdots(1+\ell_m) + n\ln|(1/n)\mathbf{S}_e| + np(\ln 2\pi + 1)\\ &\quad + 2\left\{j(p+q-j) + (k-q)p + \frac{1}{2}p(p+1)\right\}.\end{aligned} \tag{11}$$

The expression (11) was obtained by Fujikoshi and Veitch [13] as an asymptotic unbiased estimator of the risk function based on Kullback–Leibler distance.

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

$$\begin{aligned}\mathrm{A}_j &= \mathrm{AIC}_j - \mathrm{AIC}_m\\ &= n\ln\prod_{i=j+1}^{m}(1+\ell_i) - 2(p-j)(q-j), \quad j = 0, \ldots, m.\end{aligned} \tag{12}$$

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

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

$$\begin{aligned}\mathrm{B}_j &= \mathrm{BIC}_j - \mathrm{BIC}_m\\ &= n\ln\prod_{i=j+1}^{m}(1+\ell_i) - (\ln n)(p-j)(q-j), \quad j = 0, \ldots, m.\end{aligned} \tag{13}$$

$$\begin{aligned}\mathrm{C}_j &= \mathrm{C}_{p,j} - \mathrm{C}_{p,m}\\ &= n\sum_{i=j+1}^{m}\ell_i - 2(p-j)(q-j), \quad j = 0, \ldots, m.\end{aligned} \tag{14}$$

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

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

$$\mathrm{IC}_{\nu,j} = n\ln\prod_{i=j+1}^{m}(1+\ell_i) - \nu(p-j)(q-j), \quad j = 0, \ldots, m. \tag{15}$$

$$\mathrm{C}_{p,\nu,j} = n\sum_{i=j+1}^{m}\ell_i - \nu(p-j)(q-j), \quad j = 0, \ldots, m. \tag{16}$$

Here, $\mathrm{IC}_{\nu,m} = 0$ and $\mathrm{C}_{p,\nu,m} = 0$. Then,

$$\mathrm{IC}_{2,j} = \mathrm{A}_j, \qquad \mathrm{IC}_{\ln n,j} = \mathrm{B}_j, \qquad \mathrm{C}_{p,2,j} = \mathrm{C}_j.$$

3. High-dimensional consistency

3.1. Theoretical results

In this subsection, we are concerned with asymptotic behaviors of the criteria when p and n are large and q is fixed. Without loss of generality, we assume that $p \geq q$, so that $m = \min(p, q) = q$. Let the set of all possible ranks be denoted by $\mathcal{F} = \{0, \ldots, q\}$. It is assumed that the true model is the model (4) with $\mathbf{\Theta} = \mathbf{\Theta}_*$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_*$, which is denoted by M_*. The true rank is defined by $\mathrm{rank}(\mathbf{C\Theta}_*) = j_*$. Then, the model M_{j_*} is the minimum reduced-rank model including M_*. For notational simplicity, we often write $\mathbf{\Theta}_*$ and $\mathbf{\Sigma}_*$ as simply $\mathbf{\Theta}$ and $\mathbf{\Sigma}$.

We separate $\mathcal{F}$ into two sets, $\mathcal{F}_+ = \{j_*, \ldots, q\}$, the set of ranks of overspecified models, and $\mathcal{F}_- = \mathcal{F}_+^c \cap \mathcal{F} = \{0, \ldots, j_* - 1\}$, the set of ranks of underspecified models. Further, we denote the set of ranks after deleting the true model from $\mathcal{F}_+$ by $\mathcal{F}_+ \setminus \{j_*\}$, i.e., $\mathcal{F}_+ \setminus \{j_*\} = \{j_* + 1, \ldots, q\}$.

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

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

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

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

respectively. In the following, we summarize some assumptions.

A1 (The true model M_* and rank j_*): The true model is (1) or (4) with $\boldsymbol{\Theta} = \boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*$. The true rank is $\mathrm{rank}(\mathbf{C}\boldsymbol{\Theta}_*) = j_*$.
A2 (The asymptotic framework): q is fixed, $p \to \infty,\ n \to \infty,\ k \to \infty,\ p/n \to c \in [0, 1),\ k/n \to 0$.

Further, we make two types of assumptions on the order of the noncentrality matrix $\tilde{\boldsymbol{\Omega}}$ in (7). Since $\mathrm{rank}(\tilde{\boldsymbol{\Omega}}) \le q$, we can write $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$, where $\boldsymbol{\Gamma}$ is a $p \times q$ matrix. Let

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}, \tag{17}$$

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

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

and

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

which is $(n_1 n_2/n)$ times the squared Mahalanobis distance between two normal populations $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. 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 (3). Note that, when we consider the distributions of our criteria, without loss of generality we may assume

$$\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q), \tag{18}$$

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

A3 (The noncentrality matrix-1): For any $i (1 \le i \le j_*)$,

$$\omega_i = n\delta_i = \mathrm{O}(n), \qquad \lim_{p/n\to c} \delta_i = \delta_i^* > 0.$$

A4 (The noncentrality matrix-2): For any $i (1 \le i \le j_*)$,

$$\omega_i = np\xi_i = \mathrm{O}(np), \qquad \lim_{p/n\to c} \xi_i = \xi_i^* > 0.$$

In A3 and A4, it is assumed that the multiplicities of the ω_i's do not depend on p and n. 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 rank j_* tends asymptotically to 1, i.e.,

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

Here, the notation $\lim_{p/n\to c}$ is used as an abbreviation for the asymptotic framework A2 or (3).

Theorem 3.1. *Suppose that assumption* A1 *is satisfied.*

(1) IC_ν *is consistent if assumptions* A2 *and* A3 *and the inequality* $-c^{-1}\ln(1-c) < \nu < -c^{-1}\ln(1-c) + c^{-1}\ln(1+\delta_{j_*}^*)$ *are satisfied.*
(2) IC_ν *is consistent if assumptions* A2 *and* A4 *and the inequality* $-c^{-1}\ln(1-c) < \nu$ *are satisfied.*

When $c = 0$ in Theorem 3.1, "$c^{-1}\ln(1-c)$" should be read as $\lim_{c\to"0+} c^{-1}\ln(1-c) = -1$", and "$c^{-1}\ln(1+\delta_{j_*}^*)$" should be read as "$\infty$". Thus, when $c = 0$, "$-c^{-1}\ln(1-c) < \nu < -c^{-1}\ln(1-c) + c^{-1}\ln(1+\delta_{j_*}^*)$" should be read as "$-c^{-1}\ln(1-c) < \nu$". These conventions will be used elsewhere herein. Theorem 3.1 implies the following Corollaries 3.1 and 3.2 except for Corollary 3.2(1).

Table 3.1
Boundary values of ν for consistency of IC_ν and $\mathrm{C}_{p,\nu}$.

c	$b_1(c) = -c^{-1}\ln(1-c)$	$b_2(c) = (1-c)^{-1}$
0.1	1.05	1.11
0.2	1.12	1.25
0.3	1.19	1.43
0.4	1.28	1.67
0.5	1.39	2.00
0.6	1.53	2.50
0.7	1.72	3.33
0.8	2.01	5.00
0.9	2.56	10.00

Corollary 3.1. *Suppose that assumption* A1 *is satisfied. Further, assume that* $c \in [0, c_a)$, *where* c_a (≈ 0.797) *is the larger constant satisfying* $\ln(1-c_a) + 2c_a = 0$.

(1) AIC *is consistent if assumptions* A2 *and* A3 *and the inequality* $\ln(1+\delta^*_{j_*}) > (j_* - j)\{2c + \ln(1-c)\}$ *are satisfied.*
(2) AIC *is consistent if assumptions* A2 *and* A4 *are satisfied.*

Corollary 3.2. *Suppose that assumption* A1 *is satisfied.*

(1) BIC *is not consistent if assumptions* A2 *with* $c > 0$ *and* A3 *are satisfied.*
(2) BIC *is consistent if assumptions* A2 *and* A4 *are satisfied.*

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

Theorem 3.2. *Suppose that assumption* A1 *is satisfied.*

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

When $c = 0$, "$(1-c)^{-1} < \nu < (1-c)^{-1} + \{c(1-c)\}^{-1}\delta^*_{j_*}$" should be read as "$1 < \nu$".

Corollary 3.3. *Suppose that assumption* A1 *is satisfied. Further, assume that* $c \in [0, 0.5)$.

(1) C_p *is consistent if assumptions* A2 *and* A3 *and the inequality* $\delta^*_{j_*} > (j_* - j)c(1-2c)$ *are satisfied.*
(2) C_p *is consistent if assumptions* A2 *and* A4 *are satisfied.*

Theorems 3.1 and 3.2 are helpful for the selection of tuning parameters. For consistency of IC_ν, it is necessary that $b_1(c) = -c^{-1}\ln(1-c) < \nu$. Similarly, for consistency of $\mathrm{C}_{p,\nu}$, it is necessary that $b_2(c) = (1-c)^{-1} < \nu$. The boundary values for $c = 0.1, 0.2, \ldots, 0.9$ are given as in Table 3.1.

The boundary values of $b_1(c)$ are relatively steady for a change of c. However, those of $b_2(c)$ become large as c is increased beyond 0.5.

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

$$^\forall j \neq j_* \in \mathcal{F}, \quad \frac{1}{h_{j,j_*}}(T_j - T_{j_*}) \geq D_{j,j_*} \xrightarrow{p} \alpha_{j,j_*} > 0 \tag{19}$$

where D_{j,j_*} is some quantity, and h_{j,j_*} is some positive constant depending on the models. It is easy to see that (19) implies $P(\hat{j}_T = j_*) \to 1$. For example, under assumptions A1, A2 and A3, we shall show in the Appendix that
for $j > j_*$;

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

for $j < j_*$;

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

Therefore, by obtaining the ranges of ν such that the above right-hand sides are positive, we can obtain Theorem 3.1(1). Generally, we can say that for the consistency, it is necessary that the penalty term tends to infinity. When $\nu = \ln n$, we have that for $j < j_*$,

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

which implies $\Pr(\mathrm{B}_j > \mathrm{B}_{j_*}) \to 0$ for $c > 0$. Therefore,

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

which implies Corollary 3.2(1).

We note that the consistency properties in Theorems 3.1 and 3.2 hold with slight modifications for the case where the set of candidate ranks is a subfamily $\mathcal{G} \subset \mathcal{F}$. In fact, for example, the estimation method based on AIC can be 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 3.1 with the following modifications: add "$j_* \in \mathcal{G}$" to A1, and replace "any i $(1 \leq i \leq j_*)$" in A2 and A3 by "any $i (1 \leq i \leq j_*) \in \mathcal{G}$".

3.2. Differences between high-dimensional and large-sample results

From the proofs of Theorems 3.1 and 3.2, we can see that the AIC and C_p criteria on the dimensionality in a multivariate linear model satisfy the following:

$$\text{(i) if } c \in [0, c_a), \qquad \lim_{p/n \to c} P(\hat{j}_{\mathrm{A}} \in \mathcal{F}_+ \setminus \{j_*\}) = 0, \tag{20}$$

$$\text{(ii) if } c \in [0, 0.5), \qquad \lim_{p/n \to c} P(\hat{j}_{\mathrm{C}} \in \mathcal{F}_+ \setminus \{j_*\}) = 0, \tag{21}$$

under $\boldsymbol{\Omega} = \mathrm{O}(n)$ and $\boldsymbol{\Omega} = \mathrm{O}(np)$, where c_a is the constant given Theorem 3.1. Note that the properties (20) and (21) are different from those under a large-sample framework. In fact, under the large-sample framework (2) and assumption A4, it is known (Fujikoshi [9]) that

$$\lim_{n \to \infty} \Pr(\hat{j}_{\mathrm{A}} = j) = \Pr(\hat{j}_{\mathrm{C}} = j) = h(j|j_0), \tag{22}$$

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

$$h(j|j_*) = \Pr\left(\sum_{i=k+1}^{j-j_0} z_i > 2p_{k,j-j_*};\ k = 0, \ldots, j - j_* - 1,\ \text{and},\ \sum_{i=j-j_*+1}^{k} z_i > 2p_{j-j_*,k};\ k = j - j_* + 1, \ldots, q\right). \tag{23}$$

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

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

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

For BIC, we have seen that under a high-dimensional asymptotic framework it is not consistent under $\boldsymbol{\Omega} = \mathrm{O}(n)$, but it is consistent under $\boldsymbol{\Omega} = \mathrm{O}(np)$. Under a large-sample framework, it is consistent when $\boldsymbol{\Omega} = \mathrm{O}(n)$.

3.3. Numerical study

In this subsection, we numerically examine the validity of our claims and tendencies for the ranks estimated by the criteria A, B, C, or equivalently the criteria AIC, BIC, C_p, and their extensions. Our numerical results are given for the rank or dimensionality estimation criteria 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 $\mathcal{N}_p(n - q - 1, \mathbf{I}_p)$ and $\mathcal{N}_p(q, \mathbf{I}_p; \boldsymbol{\Omega}_p)$, respectively. Here, $\boldsymbol{\Omega}_p = \mathrm{diag}(\omega_1, \ldots, \omega_q, 0, \ldots, 0)$ and $\omega_1, \ldots, \omega_q$ are the possible nonzero characteristic roots of the noncentrality matrix $\boldsymbol{\Omega}$ defined by (9). Further, the sample roots $\ell_1 > \cdots > \ell_q$ may be regarded (see Lemma A.1) as those of $\mathbf{B}\mathbf{W}^{-1}$, where $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as $\mathcal{N}_q(n - p - 1, \mathbf{I}_q)$ and $\mathcal{N}_q(p, \mathbf{I}_q; \boldsymbol{\Omega})$, respectively, and $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q)$.

If we suppose that $q = 5$, then 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 ω_i, $i = 1, \ldots, 5$, are defined as follows:

$$\begin{aligned}&\text{(a)}:\ \omega_1 = 2\omega_3, \quad \omega_2 = 1.5\omega_3, \quad \omega_3 = n, \quad \omega_4 = \omega_5 = 0,\\ &\text{(b)}:\ \omega_1 = 2\omega_3, \quad \omega_2 = 1.5\omega_3, \quad \omega_3 = np, \quad \omega_4 = \omega_5 = 0.\end{aligned}$$

Table 3.2
Selection probabilities of the true rank by A, B and C when $p/n = 1/6$.

(n, p)	Case (a)			Case (b)		
	A	B	C	A	B	C
(30, 5)	0.76	0.82	0.72	0.76	0.95	0.70
(120, 20)	0.95	0.65	0.86	0.93	1.00	0.83
(300, 50)	1.00	0.13	0.97	1.00	1.00	0.98
(480, 80)	1.00	0.00	0.99	1.00	1.00	1.00
(600, 100)	1.00	0.00	1.00	1.00	1.00	1.00

Table 3.3
Selection probabilities of the true rank by A, B and C when $n = 100$.

p	Case (a)			Case (b)		
	A	B	C	A	B	C
10	0.87	1.00	0.82	0.87	1.00	0.82
30	0.95	0.00	0.71	0.94	1.00	0.66
50	0.90	0.00	0.21	0.83	1.00	0.16
70	0.52	0.00	0.00	0.40	1.00	0.00
90	0.00	0.01	0.00	0.00	0.98	0.00

These (a) and (b) correspond to the noncentrality matrix-1 and -2 under assumptions A3 and A4. Several different values of n and $p = cn$ were prepared for Monte Carlo simulations with 10^4 repetitions. Table 3.2 shows simulation results for

$$(n, p) = (30, 5),\ \ (120, 20),\ \ (300, 50),\ \ (480, 80),\ \ (600, 100).$$

In these cases, the values of p/n are all $1/6$, and assumptions A3 and A4 are satisfied. We use the selection probabilities as the relative frequencies. From Table 3.2, we can identify the following tendencies.

- For case (a), the selection probabilities of the true rank by A and C are increasing when (n, p) with $p/n = 1/6$ is increasing, and tend to 1.
- For case (a), the selection probabilities of the true rank by B do not increase even when (n, p) with $p/n = 1/6$ is increasing
- For case (b), the selection probabilities of the true rank by A, B and C are increasing when (n, p) with $p/n = 1/6$ is increasing, and tend to 1.

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

$$n = 100, \qquad p = 10,\ \ 30,\ \ 50,\ \ 70,\ \ 90.$$

In this case, the p/n values are 0.1, 0.3, 0.5, 0.7, 0.9. Table 3.3 gives the selection probabilities of the true model by A, B and C.

From Table 3.3, we can identify the following tendencies in the selection probabilities:

- The selection probabilities of the true rank by A for both case (a) and case (b) are increasing for $10 \le p \le 30$ and decreasing for $30 \le p \le 90$, taking the maximum at $p = 30$.
- The selection probabilities of the true rank by C for both case (a) and case (b) take the maximum at $p = 10$ and then are decreasing.
- For case (a) and case (b), the selection probabilities of the true rank by A and C are near 0 for $p > 0.797 (\approx c_a)$ and $p > 0.5$, respectively.
- For case (a), the selection probabilities of the true rank by B are 1 when $p = 10$, but zero when $30 \le p \le 90$.
- For case (b), the selection probabilities of the true rank by B are 1 for all $p (10 \le p \le 90)$.

Next we examined a range of tuning parameters ν such that IC_ν and $\mathrm{C}_{p,\nu}$ are consistent. The experiment was performed for $p/n = 0.9$, $n = 1000$ and $p = 900$. Note that in this case we cannot expect consistency of the usual AIC and C_p. Further, the usual BIC is not consistent for case (a).

The numerical results are given in Figs. 3.1 and 3.2 whose horizontal axis and vertical axis show the values of ν and the selection probabilities of the true rank, respectively. The solid lines and the dotted lines correspond to case (a) and case (b). In Fig. 3.1, the left dotted vertical line denotes $\nu = -\frac{1}{c}\ln(1-c)$, and the right dotted vertical line denotes $-\frac{1}{c}\ln(1-c) + \frac{1}{c}\ln(1+\delta_{j*}^*)$. In Fig. 3.2, 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 Figs. 3.1 and 3.2, we can identify the following tendencies:

- For case (a), IC_ν shall be consistent when $-c^{-1}\ln(1-c) < \nu < -c^{-1}\ln(1-c) + \ln(1+\delta_{j*}^*)$.
- For case (b), IC_ν shall be consistent when $-c^{-1}\ln(1-c) < \nu$.
- For case (a), $\mathrm{C}_{p,\nu}$ shall be consistent when $(1-c)^{-1} < \nu < (1-c)^{-1} + c(1-c)^{-1}\delta_{j*}^*$.
- For case (b), $\mathrm{C}_{p,\nu}$ shall be consistent when $(1-c)^{-1} < \nu$.

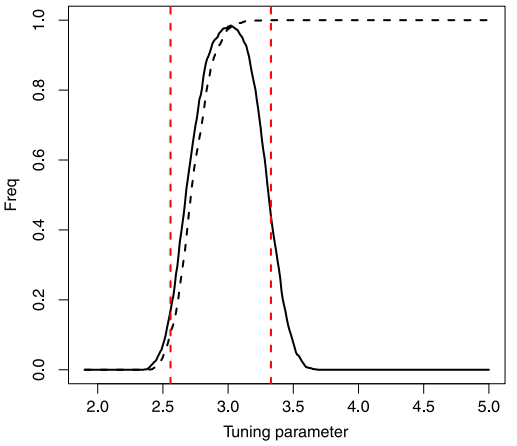

Fig. 3.1. Selection probabilities by IC_v.

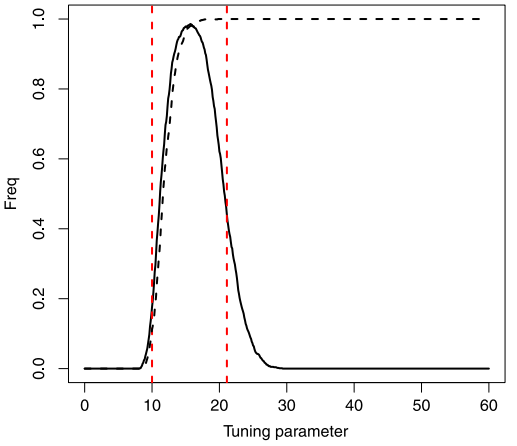

Fig. 3.2. Selection probabilities by $\mathrm{C}_{p,v}$.

4. Ridge-type criteria and their properties

4.1. Ridge-type criteria

When $p > n - k$, $\mathbf{S}_e$ becomes singular, and so we cannot use the criteria AIC, BIC and C_p. One way to overcome this problem is to use the ridge-type estimator of $\mathbf{\Sigma}$ defined by

$$\hat{\mathbf{\Sigma}}_\lambda = \frac{1}{n}(\mathbf{S}_e + \lambda \mathbf{I}_p) = \frac{1}{n}\mathbf{S}_{e,\lambda}, \tag{24}$$

or its modifications. Here, $\lambda = \{1/(np)\}\mathrm{tr}\mathbf{S}_e$. For a discussion on the use of λ, see Kubokawa and Srivastava [17]. Let $\ell_{\lambda,1} > \cdots > \ell_{\lambda,q}$ 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:

$$\begin{aligned}
\mathrm{A}_{\lambda,j} &= n \ln \prod_{i=j+1}^{q} (1 + \ell_{\lambda,i}) - 2(p-j)(q-j), \quad j = 0, \ldots, q, \\
\mathrm{B}_{\lambda,j} &= n \ln \prod_{i=j+1}^{q} (1 + \ell_{\lambda,i}) - (\ln n)(p-j)(q-j), \quad j = 0, \ldots, q, \\
\mathrm{C}_{\lambda,j} &= n \sum_{i=j+1}^{q} \ell_{\lambda,i} - 2(p-j)(q-j), \quad j = 0, \ldots, q.
\end{aligned} \tag{25}$$

Here, $\mathrm{A}_{\lambda,q} = 0$, $\mathrm{B}_{\lambda,q} = 0$ and $\mathrm{C}_{\lambda,q} = 0$. The criteria A_λ, B_λ and C_λ are obtained from the criteria A, B and C by substituting $\ell_{\lambda,i}$ to ℓ_i.

The criteria A_λ and C_λ can be justified by deriving asymptotic unbiased estimators of the AIC-type or the C_p-type risks based on ridge-type estimators. In the following, we shall give an outline of the derivations. Let $f(\mathbf{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$ be the density function of $\mathbf{Y}$, and let $\hat{\boldsymbol{\Theta}}$ be the maximum likelihood estimator of $\boldsymbol{\Theta}$ when $n - k > p$. We consider the ridge-type estimator $\hat{\boldsymbol{\Theta}}_\lambda$ defined from $\hat{\boldsymbol{\Theta}}$ by substituting $\mathbf{S}_{e,\lambda}$ for $\mathbf{S}_e$. The AIC-type risk of a candidate model M_j based on the ridge-type estimator is

$$R_{A,\lambda} = \mathrm{E}_{\mathbf{Y}}^* \mathrm{E}_{\mathbf{Z}}^* \left\{ -2 \ln f(\mathbf{Z}; \hat{\boldsymbol{\Theta}}_\lambda, \hat{\boldsymbol{\Sigma}}_\lambda) \right\}, \tag{26}$$

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

$$-2 \ln f(\mathbf{Y}; \hat{\boldsymbol{\Theta}}_\lambda, \hat{\boldsymbol{\Sigma}}_\lambda) = n \ln \prod_{i=j+1}^{q} (1 + \ell_{\lambda,i}) + n \ln |(1/n)\mathbf{S}_{e,\lambda}| + np\{1 + \ln(2\pi)\}, \tag{27}$$

where $\ell_{\lambda,1} > \cdots > \ell_{\lambda,q} > 0$ are the non-zero characteristic roots of $\mathbf{S}_h \mathbf{S}_{e,\lambda}^{-1}$. Then, the bias term is expressed as $-b_{A,\lambda}$, where

$$b_{A,\lambda} = \mathrm{E}_{\mathbf{Y}}^* \mathrm{E}_{\mathbf{Z}}^* \left\{ \mathrm{tr} \hat{\boldsymbol{\Sigma}}_\lambda^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\Theta})^\top (\mathbf{Z} - \mathbf{X}\boldsymbol{\Theta}) \right\} - np. \tag{28}$$

Here, for a notational simplicity, we express the true parameters as those in (4). Under a large-sample asymptotic framework and $\tilde{\boldsymbol{\Omega}} = \mathrm{O}(n)$, the expectation in (28) can be evaluated by the same method as in Fujikoshi and Veitch [13] as follows.

$$b_{A,\lambda} = 2 \left\{ j(p + q - j) + (k - q)p + \frac{1}{2}p(p + 1) - \frac{1}{2p}\mathrm{tr}\boldsymbol{\Sigma} \right\} + \mathrm{o}(1). \tag{29}$$

This suggests that A_λ in (25) is an asymptotic unbiased estimator for $R_{A,\lambda}$. Estimating $\boldsymbol{\Sigma}$ by a ridge-type estimator $\hat{\boldsymbol{\Sigma}}_\lambda = (1/n)\mathbf{S}_{e,\lambda}$, we can justify A_λ as an estimator of $R_{A,\lambda}$. The justification C_λ can be obtained by considering a ridge-type C_p risk defined by

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

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

4.2. Consistency of ridge-type criteria

In this subsection, we examine the consistency of ridge-type criteria when $n - k > p$. More precisely, it is shown that the criteria A_λ, B_λ and C_λ have the same consistency properties as the criteria A, B and C, respectively, under an additional assumption. The results are stated as follows:

Theorem 4.1. *Suppose that assumption* A1 *is satisfied,* $n - k > p$ *and* $(1/p)tr\boldsymbol{\Sigma} \to \alpha_0$. *Then, we have the following results.*

(1) A_λ *is consistent if* $c \in [0, c_a)$ *and assumptions* A2 *and* A3 *and the inequality* $\ln(1 + \delta_{j_*}^*) > (j_* - j)\{2c + \ln(1 - c)\}$ *are satisfied.*
(2) A_λ *is consistent if* $c \in [0, c_a)$ *and assumptions* A2 *and* A4 *are satisfied.*
(3) B_λ *is not consistent if assumptions* A2 *and* A3 *are satisfied.*
(4) B_λ *is consistent if assumptions* A2 *and* A4 *are satisfied.*
(5) C_λ *is consistent if* $c \in [0, 0.5)$ *and assumptions* A2 *and* A3 *and the inequality* $\delta_{j_*}^* > (j_* - j)c(1 - 2c)$ *are satisfied*
(6) C_λ *is consistent if* $c \in [0, 0.5)$ *and assumptions* A2 *and* A4 *are satisfied.*

Theorem 4.1 is shown by noting that the limiting values of $\ell_{\lambda,i}$, $i = 1, \ldots, q$ are the same as those of ℓ_i, $i = 1, \ldots, q$. For the proof, see Lemma A.2 in the Appendix. It is possible to generalize Theorem 4.1 for a generalized criterion with a tuning parameter. The consistency properties of ridge-type criteria for $p > n - k$ are studied numerically in the next section.

4.3. Numerical study

The case of $n - q - 1 \geq p$

In this section, we consider the selection probabilities of ridge-type criteria A_λ, B_λ and C_λ for case (a) and case (b) considered in Section 3.2. First, simulations were performed for $\boldsymbol{\Sigma} = (0.8^{|i-j|})$, $n = 100$ and $p = 10, 20, \ldots, 90$. The results are given in Figs. 4.1 and 4.2. The horizontal axis and the vertical axis show the values of p and the selection probabilities of the true model, respectively. Here, (α), (β) and (γ) denote the probabilities of selecting the true rank by (A_λ, A), (B_λ, B) and (C_λ, C), respectively. The selection probabilities by A, B and C are denoted by solid lines. Their ridge-type ones are denoted

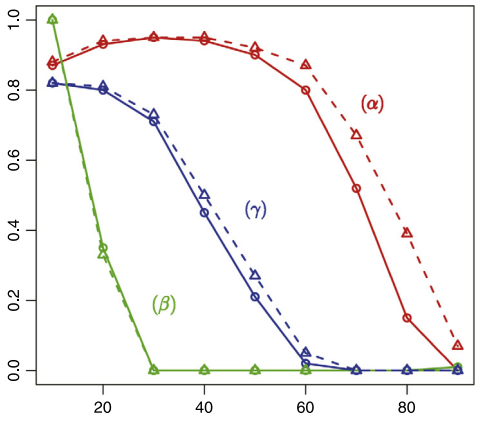

Fig. 4.1. Selection probabilities of the true rank by (A_λ, A), (B_λ, B) and (C_λ, C) for case (a).

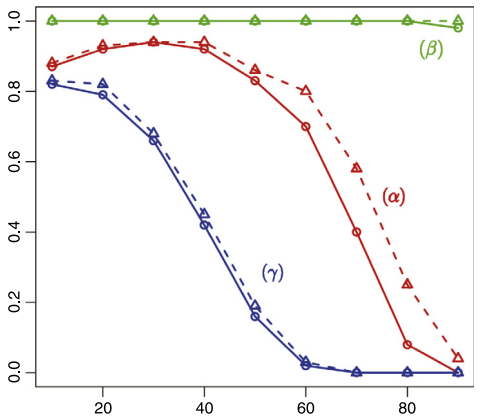

Fig. 4.2. Selection probabilities of the true rank by (A_λ, A), (B_λ, B) and (C_λ, C) for case (b).

by dotted lines. Based on Figs. 4.1 and 4.2, it is clear that when $n - q - 1 \geq p$, the selection probabilities of the true model by A_λ, B_λ and C_λ are similar to the ones by their ridge-type criteria A, B and C, respectively.

The case of $n - q - 1 < p$

Next, in order to examine behaviors of ridge-type criteria A_λ, B_λ and C_λ when $n-q-1 < p$, we did simulation experiments for $\mathbf{\Sigma} = (0.8^{|i-j|})$, $n = 30, 50, 100$ and $c = p/n = 1, 1.5, 2, 2.5, 3$ for case (b). The numerical results are given in Table 4.1, where the values for "Under", "True" and "Over" denote the probabilities of selecting the underspecified models, the true model and the overspecified models, respectively.

From Table 4.1, the following tendencies are clear for $1 \leq c = p/n \leq 3$ and $30 \leq n \leq 100$.

- For each n, the selection probabilities of the true model by A_λ become large as c becomes large. The probabilities are 1 when $c \geq 2$ and $n = 30, 50$, but for $n = 100$, the probabilities are 1 when $c \geq 2.5$.
- The selection probabilities of the true model by B_λ are 1 when $c \leq 1.5$. However, for $c \geq 2$, B_λ underestimates the true model when p is near n, and chooses underspecified models when n is large.
- The criterion C_λ chooses overspecified ranks for all most cases. The cause seems to be that the smallest roots $\ell_{\lambda,4}$ and $\ell_{\lambda,5}$ are relatively large, and this gives large effects to C_p, λ in the comparison with A_λ.

5. Concluding remarks

In general, it is known that under the large-sample asymptotic framework (2), AIC and C_p are not consistent, but BIC is consistent. However, in this paper, we demonstrated that the criteria based on AIC and C_p for estimating the rank (dimensionality) in a multivariate linear model can have consistency, under the high-dimensional asymptotic framework (3). For consistency, some additional assumptions are required. For AIC, it is necessary that $c \in [0, c_a)$, where $c_a \approx 0.797$. For C_p, it is necessary that $c \in [0, 0.5)$. More precisely, consistency was considered under two types of assumptions on the largeness of the characteristic roots of the noncentrality matrix $\mathbf{\Omega}$ in (17). For the criterion based on BIC, we note that it is consistent when $\mathbf{\Omega} = O(np)$, but it is not consistent when $\mathbf{\Omega} = O(n)$. These results were extended for the general criteria IC_ν and $C_{p,\nu}$ with a tuning parameter in (15) and (16). We gave sufficient conditions for IC_ν and $C_{p,\nu}$ to be consistent. The

Table 4.1
Selection probabilities by A_λ, B_λ and C_λ when $c = p/n > 1$.

c	$n = 30$			$n = 50$			$n = 100$		
	Under	True	Over	Under	True	Over	Under	True	Over
Selection probabilities by A_λ									
1	0.00	0.81	0.19	0.00	0.79	0.21	0.00	0.55	0.45
1.5	0.00	0.93	0.07	0.00	0.87	0.13	0.00	0.24	0.76
2	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.89	0.11
2.5	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
3	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Selection probabilities by B_λ									
1	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
1.5	0.13	0.87	0.00	0.06	0.94	0.00	0.00	1.00	0.00
2	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
2.5	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
3	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Selection probabilities by C_λ									
1	0.00	0.01	0.99	0.00	0.00	1.00	0.00	0.00	1.00
1.5	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
2	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
2.5	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
3	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00

sufficient conditions are useful in the selection of the tuning parameter ν. When q is large, we showed that the probability of selecting ranks $j > j_*$ by AIC tends to 0. It is left as a future problem to examine consistency properties of AIC and the other criteria.

We proposed the ridge-type criteria A_λ, B_λ and C_λ in (25) which are also defined for the case where $p > n - k$. It was shown that A_λ and C_λ 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 the criteria A, B and C, respectively, when $n - k > p$. For the case $n - k < p$, we studied the consistency properties numerically in the range $1 \le p/n \le 3$. It was found that A_λ becomes consistent as $c = p/n$ becomes large, B_λ is consistent if n is close to p, and $C_{p,\lambda}$ estimates overspecified ranks. The theoretical justifications for these observations are left as future research.

Recently, sparse estimation methods under a reduced rank based on penalization techniques have been proposed by Yuan et al. [23] and Bunea et al. [6]. Chen and Hung [8] and Bunea et al. [7] also considered simultaneous methods for dimension reduction and variable selection. These methods are based on least squares methods, and so it seems that the correlations of the response variables have been ignored. In our approach, the correlations of the response variables have been taken into consideration, but in most case, it is assumed that $n - k > p$, except for the ridge-type criteria. In both the penalized methods and the generalized criteria A_ν, B_ν and $C_{p,\nu}$, there are problems in how to decide the tuning parameter.

Acknowledgments

We wish to thank the editor and two reviewers for their critical and helpful comments on the first version, which allowed us to greatly improve our manuscript. The first author's research is partially supported by a Grant-in-Aid for Scientific Research (C), #25330038, 2013–2015, from the Ministry of Education, Science, Sports and Culture, Japan.

Appendix. Proofs of theorems and lemmas

Lemma A.1

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 $\mathcal{N}_p(n-k, \mathbf{\Sigma})$ and a noncentral Wishart distribution $\mathcal{N}_p(q, \mathbf{\Sigma}; \mathbf{\Sigma}^{1/2}\tilde{\mathbf{\Omega}}\mathbf{\Sigma}^{1/2})$, respectively. Here, it is assumed that $n-k \ge p$. Let $\ell_1 > \cdots > \ell_q$ and $\omega_1 \ge \cdots \ge \omega_q$ be the possible nonzero characteristic roots of $\mathbf{S}_h\mathbf{S}_e^{-1}$ and $\tilde{\mathbf{\Omega}}$, respectively. We assume that $rank(\tilde{\mathbf{\Omega}}) = a$, and hence $\omega_1 \ge \cdots \ge \omega_a > \omega_{a+1} = \cdots = \omega_q = 0$. Further, we assume that the multiplicities of the ω_i's do not depend on p and n. 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 $i (1 \le i \le a)$, $\omega_i = n\delta_i = O(n)$ and $\lim_{p/n\to c} \delta_i = \delta_i^* > 0$. Then*

$$\ell_i \xrightarrow{p} \frac{c}{1-c} + \frac{1}{1-c}\delta_i^*, \quad i = 1, \ldots, a, \quad \text{and} \quad \ell_i \xrightarrow{p} \frac{c}{1-c}, \quad i = a+1, \ldots, q.$$

(2) *Suppose that for any* $i(1 \le i \le a)$, $\omega_j = np\xi_i = O(np)$ *and* $\lim_{p/n\to c} \xi_i = \xi_i^* > 0$. *Then*

$$\frac{1}{p}\ell_i \xrightarrow{p} \frac{1}{1-c}\xi_i^*, \quad i = 1, \ldots, a, \quad \text{and} \quad \ell_i \xrightarrow{p} \frac{c}{1-c}, \quad i = a+1, \ldots, q.$$

It is known (see Fujikoshi, Ulyanov and Shimizu [12]) that the nonzero characteristic roots $\ell_1 > \cdots > \ell_q > 0$ of $\mathbf{S}_h\mathbf{S}_e^{-1}$ may be regarded as those of $\mathbf{B}\mathbf{W}^{-1}$, where $\mathbf{W}$ and $\mathbf{B}$ are independently distributed as a central Wishart distribution $\mathcal{N}_q(m, \mathbf{I}_q)$ and a noncentral Wishart distribution $\mathcal{N}_q(p, \mathbf{I}_q; \mathbf{\Omega})$, respectively. Here, $m = n - k - p + q$, $\mathbf{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_q)$, and

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

Using this property, Fujikoshi [10] has derived limiting distributions of the ℓ_i's under a general multiplicity of the δ_i's and ξ_i's. Lemma A.1 follows from Fujikoshi [10].

The Proof of Theorem 3.1. In the proof of Theorem 3.1, it is assumed that the true rank or dimensionality is j_*. Since the number of possible models is finite, it is sufficient to show that the values of $\mathrm{IC}_{v;j} - \mathrm{IC}_{v;j_*}$ converge to positive values.

Note that for $j > j_*$,

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

and for $j < j_*$

$$\mathrm{IC}_{v,j} - \mathrm{IC}_{v,j_*} = n\ln\{(1+\ell_{j+1})\cdots(1+\ell_{j_*})\} + v(j-j_*)(p+q-j-j_*).$$

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

$$\frac{1}{n}\left(\mathrm{IC}_{v,j} - \mathrm{IC}_{v,j_*}\right) \xrightarrow{p} (j-j_*)\{\ln(1-c) + vc\}.$$

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

$$\begin{aligned}\frac{1}{n}\left(\mathrm{IC}_{v,j} - \mathrm{IC}_{v,j_*}\right) &\xrightarrow{p} \ln(1+\delta_{j+1}^*)\cdots(1+\delta_{j_*}^*) - (j_*-j)\{\ln(1-c) + vc\} \\ &\ge \ln(1+\delta_{j_*}^*)\cdots(1+\delta_{j_*}^*) - (j_*-j)\{\ln(1-c) + vc\} \\ &= (j_*-j)[\ln(1+\delta_{j_*}^*) - \{\ln(1-c) + vc\}].\end{aligned}$$

The limiting value is positive when

$$v < -\frac{1}{c}\ln(1-c) + \frac{1}{c}\ln(1+\delta_{j_*}^*).$$

For the proof for case $c = 0$, from the above result, it holds that for $0 < c < 1$,

$$\lim \frac{1}{p}\left(\mathrm{IC}_{v,j} - \mathrm{IC}_{v,j_*}\right) \ge (j_*-j)\left[\frac{1}{c}\ln(1+\delta_{j_*}^*) - \left\{\frac{1}{c}\ln(1-c) + v\right\}\right].$$

The result is obtained by considering $\lim_{c+0}$.

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 that under $\omega_j = \mathrm{O}(n)$. Therefore, the limiting value of $(1/n)\left\{\mathrm{IC}_{v,j} - \mathrm{IC}_{v,j_*}\right\}$ is positive when $-c^{-1}\ln(1-c) < v$. From Lemma A.1(2), we have

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

This proves Theorem 3.1(2).

The Proof of Theorem 3.2. Note that for $j > j_*$,

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

and for $j < j_*$,

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

Therefore, Theorem 3.2 can be proved in the same way as was Theorem 3.1, with the help of Lemma A.1.

Lemma A.2. *Let* $\mathbf{S}_e$ *and* $\mathbf{S}_h$ *be independently distributed as a Wishart distribution* $\mathcal{N}_p(n-k, \mathbf{I}_p)$ *and a noncentral Wishart distribution* $\mathcal{N}_p(q, \mathbf{I}_p; \tilde{\mathbf{\Omega}})$, *respectively. Here, we assume that* $n - k \ge p \ge q$ *and* $\tilde{\mathbf{\Omega}} = diag(\omega_1, \ldots, \omega_p)$. *Let* $\mathbf{S}_{e,\lambda} = \mathbf{S}_e + \lambda\mathbf{I}_p$, *where* $\lambda = (np)^{-1}tr\mathbf{\Sigma}\mathbf{S}_e$. *Let* $\ell_1 > \cdots > \ell_q$ *and* $\ell_{\lambda,1} > \cdots > \ell_{\lambda,q}$ *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) *we are under the high-dimensional asymptotic framework given by* (A.1), (ii) *when* $p/n \to c \in [0, 1)$, $(1/p)tr\mathbf{\Sigma} \to \alpha_0$, *and* (iii) $\omega_1 \ge \cdots \ge \omega_a > \omega_{a+1} = \cdots = 0$. *Then, the characteristic roots* $\ell_{\lambda,1} > \cdots > \ell_{\lambda,q}$ *have the same limiting values as those of the characteristic roots* $\ell_1 > \cdots > \ell_q$ *given in Lemma* A.1.

In general, it holds that $\ell_i \geq \ell_{\lambda,i}$, $i = 1, \ldots, q$. Noting that

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

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 [21, p. 117]), we have

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

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 the Marčenko–Pastur law (see Bai and Silverstein [5]), it is known (Bai et al. [4]) 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 A.1 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 $\lim \ell_i \leq \lim \ell_{\lambda,i}$, and hence $\lim \ell_{\lambda,i} = \lim \ell_i$. 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 argument as that in the case of $\omega_j = O(n)$, we can show that $\tilde{\ell}_j$ has the same limiting value as ℓ_j, for $j = 1, \ldots, q$.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki (Eds.), 2nd. International Symposium on Information Theory, Akadémiai Kiadó, Budapest, 1973, pp. 267–281.
[2] T.W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann. Math. Statist. 22 (1951) 327–351.
[3] T.W. Anderson, Introduction to Multivariate Statistical Analysis, third ed., Wiley, Hoboken, NJ, 2003.
[4] Z.W. Bai, K.P. Choi, Y. Fujikoshi, Limiting behavior of eigenvalues in MANOVA with high-dimension by RMT, 2016 (submitted for publication).
[5] Z.W. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, second ed., Springer, 2010.
[6] F. Bunea, Y. She, M.H. Wegkamp, Optimal selection of reduced rank estimators of high-dimensional matrices, Ann. Statist. 39 (2011) 1282–1309.
[7] F. Bunea, Y. She, M.H. Wegkamp, Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, Ann. Statist. 40 (2012) 2359–2388.
[8] L. Chen, J.Z. Huang, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, J. Amer. Statist. Assoc. 107 (2012) 1533–1545.
[9] Y. Fujikoshi, Two methods for estimation of dimensionality in canonical correlation analysis and the multivariate linear model, in: K. Matsushita (Ed.), Statistical Theory and Data Analysis, Elsevier Science, Amsterdam, 1985, pp. 233–240.
[10] Y. Fujikoshi, High-dimensional asymptotic distributions of characteristic roots in multivariate linear model and canonical correltion analysis, 2016 (submitted for publication).
[11] Y. Fujikoshi, T. Sakurai, H. Yanagihara, Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression, J. Multivariate Anal. 123 (2014) 184–200.
[12] Y. Fujikoshi, V.V. Ulyanov, R. Shimizu, Multivariate Statistics: High-Dimensional and Large-Sample Approximations, Wiley, Hoboken, NJ, 2010.
[13] Y. Fujikoshi, L.G. Veitch, Estimation of dimensionality in canonical correlation analysis, Biometrika 66 (1979) 345–351.
[14] B.K. Gunderson, R.J. Muirhead, On estimating the dimensionality in canonical correlation analysis, J. Multivariate Anal. 62 (1997) 121–136.
[15] A.J. Izenman, Reduced-Rank Regression for the multivariate linear model, J. Multivariate Anal. 5 (1975) 248–262.
[16] A.M. Kshirsagar, Multivariate Analysis, Marcel Dekker, New York, 1972.
[17] T. Kubokawa, M.S. Srivastava, Selection of variables in multivariate regression models for large dimensions, Commun. Stat. - Theory Methods 41 (2012) 2465–2489.
[18] C.L. Mallows, Some comments on C_p, Technometrics 15 (1973) 661–675.
[19] G.C. Reisel, R.P. Velu, Multivariate Reduced-Rank Regression, in: Lecure Notes in Statistics, vol. 136, Springer, New York, 1998.
[20] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[21] G.A.F. Seber, A Matrix Handbook for Statisticians, Wiley, Hoboken, NJ, 2008.
[22] H. Yanagihara, H. Wakaki, Y. Fujikoshi, A consistency property of AIC for multivariate linear model when the dimension and the sample size are large, Electron. J. Stat. 9 (2015) 59–81.
[23] M. Yuan, A. Ekici, Z. Lu, R. Monteiro, Dimension reduction and coefficient estimation in multivariate linear model, J. R. Stat. Soc. B 69 (2007) 329–346.