High-Dimensional AICs for Selection of Redundancy Models in Discriminant Analysis

Tetsuro Sakurai*, Takeshi Nakada** and Yasunori Fujikoshi*

*Faculty of Science and Engineering,
Chuo University, Kasuga, Bunkyo-ku, 112-8551, Japan

**Financial Systems Development Dept. No. 4,
ITOCHU Techno-Solutions Corporation, 1-2-2, Osaki, Shinagawa-ku,
141-8522, Japan

Abstract

This paper is concerned with high-dimensional modifications of Akaike information criterion (AIC) for redundancy (no additional information) models in discriminant analysis. The AIC has been proposed as an asymptotically unbiased estimator of the risk function when the dimension is fixed and the sample size tends to infinity. On the other hand, Fujikoshi (2002) attempted to modify the AIC in two-groups discriminant analysis when the dimension and the sample size tend to infinity. However, its modification was obtained under a restrictive assumption, and furthermore, it was difficult to extend the method to multiple-groups case. In this paper, by a new approach we propose HAIC which is an asymptotically unbiased estimator of the risk function in multiple-groups discriminant analysis when both the dimension and the sample size tend to infinity, for a general class of candidated models. By simulation experiments it is shown that HAIC is more useful than the usual AIC and CAIC.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: AIC, Bias correction, Discriminant analysis, HAIC, High dimensional case, Multiple-groups case, Selection of variables. .

1. Introduction

This paper is concerned with selection of variables in discriminant analysis. One way of selecting variables is to formulate as a problem of selecting redundancy models based on no additional information hypothesis due to Rao (1948, 1973). Then, we apply the idea of a model selection criterion AIC to the models.

The selection criterion AIC is proposed as an approximately unbiased estimator (AIC) of the AIC type of risk defined by the expected log-predictive likelihood. The AIC for redundancy models has been proposed under large-sample framework, i.e., the dimension is fixed and the sample size tends to infinity.

On the other hand, in discriminant analysis we encounter a high-dimensional case, i.e., the case when the dimension is relatively large. There are some works on asymptotic approximations for the expected probabilities of misclassification in the two-groups discriminant analysis under a high dimensional framework. For these works, see, e.g., Raudys (1972), Wyman et al. (1990), and Fujikoshi and Seo (1998), in which they point a goodness of such approximations.

In this paper, we consider the problem of estimating the AIC type of risk when both the dimension and sample size are large, in multiple-groups discriminant analysis. More precisely, we attempt to reduce for the bias term when we estimate the AIC type of risk by $-2\log$ likelihood, in a high dimensional case. An attempt has been done by Fujikoshi (2002) in two-groups discriminant analysis. However, a modification was obtained under a restrictive assumption (see Section 3.1), and furthermore, it was difficult to extend the method to multiple-groups case. In this paper, by a new approach we obtain an asymptotically unbiased estimator of the risk function in multiple-groups discriminant analysis when both the dimension and the sample size tend to infinity, for a general class of candidated models. which

is not necessary to include the true model. Such an estimator is called high-dimensional AIC, which is denoted by HAIC. Furthermore, it is pointed that HAIC has smaller biases than the AIC in a large-sample framework in a wide range of dimensions and sample sizes, through simulation experiments. It is also shown that HIC provides better model selections than does AIC or CAIC.

2. Preliminaries

Let $\boldsymbol{x} = (x_1, ..., x_p)'$ be a p-dimensional random vector measurable on the individuals of each of $q+1$ populations $\Pi_1, \ldots, \Pi_{q+1}$. Let $\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_{N_i}^{(i)}$ be a sample from Π_i, and denote all the observations by the $N \times p$ matrix,

$$X = \left(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{N_1}^{(1)}, \ldots, \boldsymbol{x}_1^{(q+1)}, \ldots, \boldsymbol{x}_{N_{q+1}}^{(q+1)}\right), \tag{2.1}$$

where $N = N_1 + \cdots + N_{q+1}$. It is assumed that

$$\mathrm{E}(\boldsymbol{x}|\Pi_i) = \boldsymbol{\mu}^{(i)}, \quad \mathrm{Var}(\boldsymbol{x}|\Pi_i) = \Sigma. \tag{2.2}$$

We consider AIC and its high-dimensional modifications for a redundancy model of a given subset of $\{x_1, \ldots, x_p\}$. Without loss of generality we treat a candidate model M_k, which means that the first k variate $\boldsymbol{x}_1 = (x_1, \ldots, x_k)'$ is sufficient, or the remainder variate $\boldsymbol{x}_2 = (x_{k+1}, \ldots, x_p)$ is redundant, i.e., the remainder $p-k$ variate $\boldsymbol{x}_2$ has no additional information in canonical discriminant analysis, in presence of $\boldsymbol{x}_1$. In order to write the model M_k in terms of unknown parameters, let us consider the partitions

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(i)} \\ \boldsymbol{\mu}_2^{(i)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{2.3}$$

and let

$$\boldsymbol{\mu}_{2\cdot1}^{(i)} = \boldsymbol{\mu}_2^{(i)} - \Gamma\boldsymbol{\mu}_1^{(i)}, \quad i = 1, \ldots, q+1; \quad \Gamma = \Sigma_{21}\Sigma_{11}^{-1}.$$

Then, M_k is defined by

$$M_k: \quad \boldsymbol{\mu}_{2\cdot1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot1}^{(q+1)} \left(= \boldsymbol{\mu}_{2\cdot1}^{(i)}\right). \tag{2.4}$$

Let $f(X;\Theta)$ be the density function of X under M_k with $\Theta = \{\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(q+1)}, \Sigma\}$. Further, let $g(X)$ be the density function of X under the true model M^*. Then, we can write the AIC type of risk defined by the expected log-predictive likelihood for a model M_k as

$$R_k = \mathrm{E}_X^* \mathrm{E}_Z^* [-2 \log f(Z; \hat{\Theta}_k)], \tag{2.5}$$

where Z is an $N \times p$ random matrix that has the same distribution as X and is independent of X, and $\hat{\Theta}_k$ is the maximum likelihood estimator of Θ under M_k. Here E^* and Var^* denote the expectation under the true model. Note that the risk R_k can be expressed as

$$R_k = \mathrm{E}_X^* [-2 \log f(X; \hat{\Theta}_k)] + b_k, \tag{2.6}$$

where

$$b_k = \mathrm{E}_X^* \mathrm{E}_Z^* [-2 \log f(Z; \hat{\Theta}_k)] - \mathrm{E}_X [-2 \log f(X; \hat{\Theta}_k)]. \tag{2.7}$$

This means that a naive estimator of R_k is $-2 \log f(X; \hat{\Theta}_k)$, and b_k is its bias term in the estimation of R_k. In this paper we assume that

$$g(X) = f(X; \Theta_0), \text{ for some } \Theta_0. \tag{2.8}$$

In the following we write Θ_0 as Θ simply.

Let $\bar{\boldsymbol{x}}^{(i)}$ and $\bar{\boldsymbol{x}}$ be the sample mean vectors of the observations of the ith groups and all the groups, respectively, i.e.,

$$\bar{\boldsymbol{x}}^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{x}_j^{(i)}, \quad i = 1, \ldots, q+1; \quad \bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{q+1} \sum_{j=1}^{N_i} \boldsymbol{x}_j^{(i)}.$$

Further, let W and B be the matrices of sums of squares and products due to within-groups and between-groups, respectively, i.e.,

$$W = \sum_{i=1}^{q+1} \sum_{j=1}^{N_i} (\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})', \quad B = \sum_{i=1}^{q+1} N_i (\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})'.$$

Put $T = W + B$ and partition W, B and T in the same way as in (2.3). Then we can write the MLE $\hat{\Theta}_k$ of Θ under M_k (see, e.g., Fujikoshi (1985))

as

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}_1^{(i)} = \bar{\boldsymbol{x}}_1^{(i)}, \quad i = 1, \ldots, q; \quad \hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{x}}_2 - \hat{\Gamma}\bar{\boldsymbol{x}}_1, \\
&\hat{\Gamma} = T_{21}T_{11}^{-1}, \quad N\hat{\Sigma}_{11} = W_{11}, \\
&N\hat{\Sigma}_{22\cdot 1} = T_{22\cdot 1} = T_{22} - T_{21}T_{11}^{-1}T_{12},
\end{aligned} \tag{2.9}$$

and hence, putting $\ell_k(W,T) = -2\log f(X;\hat{\Theta}_k)$ we have

$$\begin{aligned}
\ell_k(W,T) &= N\log|N^{-1}W_{11}| + N\log|N^{-1}T_{22\cdot 1}| + Np(1+\log 2\pi) \\
&= -N\log\frac{|W_{22\cdot 1}|}{|T_{22\cdot 1}|} + N\log|N^{-1}W| + Np(1+\log 2\pi). \quad (2.10)
\end{aligned}$$

Our main concern is to evaluate the bias term b_k under the assumption of normality. Note that in the evaluation, it is not assumed that M_k includes the true model. Put

$$\bar{\boldsymbol{\mu}} = \sum_{i=1}^{q+1}\frac{N_i}{N}\boldsymbol{\mu}^{(i)}, \quad \Xi = \sum_{i=1}^{q+1}\frac{N_i}{N}(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})'.$$

Corresponding to a partition of Σ, we partition Ξ as

$$\Xi = \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{pmatrix}.$$

Then, we can write b_k (see, e.g., Fujikoshi (2002)) as

$$b_k = -Np + \frac{Nk(N+q+1)}{N-k-q-2} + N^2\tilde{b}_k, \tag{2.11}$$

where

$$\tilde{b}_k = \mathrm{E}\left[\mathrm{tr}T^{-1}\left\{(1+N^{-1})\Sigma + \Xi\right\} - \mathrm{tr}T_{11}^{-1}\left\{(1+N^{-1})\Sigma_{11} + \Xi_{11}\right\}\right]. \tag{2.12}$$

Here the expectation E means the one under general normal populations. Note that T and T_{11} are distributed as noncentral Wishart distributions $W_p(n,\Sigma;N\Xi)$ and $W_k(n,\Sigma_{11};N\Xi_{11})$, respectively, where $n = N-1$.

Our interest is to examine the problem of evaluating the bias term b_k and estimating it. In general, the bias correction problem has been studied under a usual large sample framework,

$$\mathrm{LS}: \ p, q, k; \ \text{fix}, \ N \to \infty.$$

Fujikoshi (1985) derived an asymptotic unbiased estimator for b_k under LS without assuming that M_k includes the true model, which is given as

$$b_{LS} = b_{LS}^{(0)} + b_{LS}^{(1)}, \tag{2.13}$$

where

$$b_{LS}^{(0)} = 2\left\{k(q+1) + p - k + \frac{1}{2}p(p+1)\right\},$$
$$b_{LS}^{(1)} = -2\left(\mathrm{tr}C + \frac{1}{2}\mathrm{tr}C^2 + \frac{1}{2}(\mathrm{tr}C)^2\right) + 2\left(\mathrm{tr}C_k + \frac{1}{2}\mathrm{tr}C_k^2 + \frac{1}{2}(\mathrm{tr}C_k)^2\right).$$

Here

$$C = \Sigma^{-1}\Xi(I_p + \Sigma^{-1}\Xi)^{-1}, \quad C_k = \Sigma_k^{-1}\Xi_k(I_k + \Sigma_k^{-1}\Xi_k)^{-1},$$

and Σ_k and Ξ_k are the first $k \times k$ submatrices of Σ and Ξ, respectively. When M_k includes the true model, $b_{LS}^{(1)} = 0$, and hence $b_{LS} = b_{LS}^{(0)}$, which corresponds to $2 \times$ (the number of independent parameters under M_k). The usual AIC and its corrected version have been proposed as

$$\mathrm{AIC} = \ell_k(W, T) + b_{LS}^{(0)}, \tag{2.14}$$

and

$$\mathrm{CAIC} = \ell_k(W, T) + \hat{b}_{LS}, \tag{2.15}$$

respectively. Here, $\hat{b}_{LS}$ is the one obtained from b_{LS} by substituting the sample quantities to C and C_k.

However, the result will not work well as the dimension p increases. In order to overcome this weakness, we study asymptotic unbiased estimator for b_k under two high-dimensional frameworks such that

$$\mathrm{HD1}: q, k; \mathrm{fix},\ N \to \infty,\ p \to \infty,\ N - p \to \infty,\ c = p/N \to c_0 \in (0, 1).$$
$$\mathrm{HD2}: q; \mathrm{fix},\ N \to \infty,\ p \to \infty,\ N - p \to \infty,\ c = p/N \to c_0 \in (0, 1),$$
$$k \to \infty,\ N - k \to \infty,\ d = k/N \to d_0 \in (0, 1).$$

Our aim is to construct $\hat{b}_{HD}$ such that

$$\mathrm{E}[\hat{b}_{HD}] = b_{HD} + O_{1/2}(N^{-1}, p^{-1}, k^{-1}),$$

where $O_j(N^{-1}, p^{-1}, k^{-1})$ denotes the term of the jth order with respect to (N^{-1}, p^{-1}, k^{-1}). The AIC with such a bias term is denoted as

$$\text{HAIC} = \ell_k(W, T) + \hat{b}_{HD}. \tag{2.16}$$

3. Asymptotic Evaluation of the Bias Term

In this section we obtain asymptotic expressions for b_k in (2.12) under high-dimensional frameworks. In the derivation we are not necessary to assume that M_k includes the true model M^*, but we assume only that the true model M^* satisfies (2.8), i.e., that Π_i's are normal populations. It is easy to see that the $\tilde{b}_k$ in (2.12) can be expressed as

$$\tilde{b}_k = \mathrm{E}\left[\mathrm{tr}T^{-1}\left\{(1+N^{-1})I_p + \Omega_p\right\} - \mathrm{tr}T_k^{-1}\left\{(1+N^{-1})I_k + \Psi_k\right\}\right]. \tag{3.1}$$

Here

$$\Omega_p = \mathrm{diag}(\omega_1 \ldots, \omega_q, 0 \ldots, 0), \quad \Psi_k = \mathrm{diag}(\psi_1 \ldots, \psi_q, 0 \ldots, 0), \tag{3.2}$$

where $\omega_1 \geq \cdots \geq \omega_q \geq 0$ and $\psi_1 \geq \cdots \geq \psi_q \geq 0$ are possibly non-zero roots of $\Sigma^{-1}\Xi$ and $\Sigma_{11}^{-1}\Xi_{11}$, respectively. Further, T and T_k are distributed as noncentral Wishart distributions $W_p(n, I_p; N\Omega_p)$ and $W_k(n, I_k; N\Psi_k)$, respectively. Such reductions are obtained by considering appropriate transformations. For example, let H be an orthogonal matrix H such that

$$\tilde{\Xi} = H'\Sigma^{-1/2}\Xi\Sigma^{-1/2}H = \Omega_p.$$

Then

$$\mathrm{tr}T^{-1}\left\{(1+N^{-1})\Sigma + \Xi\right\} = \mathrm{tr}(H'\Sigma^{-1/2}T\Sigma^{-1/2}H)^{-1}\left\{(1+N^{-1})I_p + \Omega_p\right\}.$$

The reduction of the first expectation is obtained by noting that

$$H'\Sigma^{-1/2}T\Sigma^{-1/2}H \sim W_p(n, I_p; N\Omega_p).$$

Similarly the expectation of the second term is obtained.

3.1 Two-Groups Case

For two-groups case, Fujikoshi (2002) used the following expression for the bias term b_k:

$$\begin{aligned} b_k &= -Np + \frac{Nk(N+2)}{N-k-3} \\ &\quad +N^2\{Q(N,p,\tau^2) - Q(N,k,\tau_1^2)\}, \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} Q(N,p,\tau^2) &= \mathrm{E}[\mathrm{tr}(S+\boldsymbol{u}\boldsymbol{u}')^{-1}\{(1+N^{-1})I_p + \boldsymbol{\nu}\boldsymbol{\nu}'\}], \\ Q(N,k,\tau_1^2) &= \mathrm{E}[\mathrm{tr}(S_{11}+\boldsymbol{u}_1\boldsymbol{u}_1')^{-1}\{(1+N^{-1})I_k + \boldsymbol{\nu}_1\boldsymbol{\nu}_1'\}], \end{aligned}$$

$\boldsymbol{\nu} = (\sqrt{N_1N_2}/N)\Sigma^{-1/2}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$, $\tau^2 = \boldsymbol{\nu}'\boldsymbol{\nu} = \{(N_1N_2)/N^2\}\Delta^2$ and

$$\Delta^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}).$$

Here $\boldsymbol{u}$ and S are independently distributed as a normal distribution $N_p(\sqrt{N}\boldsymbol{\nu}, I_p)$ and a Wishart distribution $W_p(N-2, I_p)$, respectively. Similarly $S_{11}, \boldsymbol{u}_1$ and $\boldsymbol{\nu}_1$ denote the first k parts of $S, \boldsymbol{u}$ and $\boldsymbol{\nu}$, respectively, and $\tau_1^2 = \boldsymbol{\nu}_1'\boldsymbol{\nu}_1$. Using that

$$(S+\boldsymbol{u}\boldsymbol{u}')^{-1} = S^{-1} - (1+\boldsymbol{u}'S^{-1}\boldsymbol{u})^{-1}S^{-1}\boldsymbol{u}\boldsymbol{u}'S^{-1},$$

Fujikoshi (2002) derived an asymptotic expansion of $Q(N,p,\tau^2)$ such that

$$Q(N,p,\tau^2) = Q_1(N,p,\tau^2) + O_{5/2}(N^{-1},p^{-1}), \tag{3.4}$$

assuming that $\boldsymbol{u} \sim N_p(\boldsymbol{\nu}, I_p)$. However, the mean of $\boldsymbol{u}$ is not $\boldsymbol{\nu}$, but $\sqrt{N}\boldsymbol{\nu}$. For a general case, the Q_1 in Fujikoshi (2002) should be corrected as follows:

$$\begin{aligned} N^2Q_1(N,p,\tau^2) &= \frac{N(Np+p-4)}{N-p-2} \\ &\quad +2\left(\frac{N-3}{N-p-2}\right)\left\{3(1+\tau^2)^{-1} - (1+\tau^2)^{-2}\right\}. \end{aligned} \tag{3.5}$$

These imply that the bias term can be expressed under HD2 as

$$\begin{aligned} b_k &= -Np + \frac{Nk(N+2)}{N-k-3} \\ &\quad +N^2\{Q_1(N,p,\tau^2) - Q_1(N,k,\tau_1^2)\} + O_{1/2}(N^{-1},p^{-1},k^{-1}). \end{aligned} \tag{3.6}$$

We note that the result (3.6) is the same as a special case of the result (see Theorem 3.3) in the present paper. It is difficult to extend the above method to the case $q \geq 2$. For the case $q \geq 2$, we give an alternative method which is more powerful.

3.2 Some Basic Results

For our evaluation of the $\tilde{b}_k$ in (3.1), we use the following theorem.

Theorem 3.1. *Let $T \sim W_p(N-1, I_p; N\Omega)$, and partition T and Ω as*

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

where T_{ij} and Ω_{ij} are $p_i \times p_j$. If Ω_{12}, Ω_{21} and Ω_{22} are zero matrices, then

$$\begin{aligned} \mathrm{E}\left[\mathrm{tr}T^{-1}\left\{\left(1+\frac{1}{N}\right)I_p + \Omega\right\}\right] &= \left(1+\frac{1}{N}\right)\frac{p-p_1}{N-p-2} \\ &+ \left(\frac{N-p_1-2}{N-p-2}\right)\mathrm{E}\left[\mathrm{tr}T_{11}^{-1}\left\{\left(1+\frac{1}{N}\right)I_{p_1} + \Omega_{11}\right\}\right]. \end{aligned}$$

The proof of Theorem 3.1 is given in Appendix. Moreover, we use the following lemma which is obtained by modifying the derivation in Fujikoshi (1985).

Lemma 3.1. *Suppose that $T \sim W_p(N-1, \Sigma; m\Omega)$, and let $L; p \times p$ be a constant matrix. Then, if p is fixed, $m = O(N)$ and $\Omega = O(1)$, then*

$$\begin{aligned} \mathrm{E}[N^2\mathrm{tr}T^{-1}L] &= N\mathrm{tr}M\Delta + (p+3)\mathrm{tr}M\Delta^2 + \mathrm{tr}\Delta\mathrm{tr}M\Delta \\ &\quad - \mathrm{tr}M\Delta^3 - \mathrm{tr}\Delta\mathrm{tr}M\Delta^2 + O(N^{-1/2}), \end{aligned} \tag{3.7}$$

where $M = \Sigma^{-1}L$ and $\Delta = (I_p + (m/N)\Sigma^{-1}\Omega)^{-1}$. In particular, if $m = N$,

$$\mathrm{E}\left[N^2\mathrm{tr}T^{-1}\left\{\left(1+\frac{1}{N}\right)I_p + \Omega\right\}\right] = Np + 2(p+2)\eta_1 - \eta_2 - \eta_1^2 + O(N^{-1/2}),$$

where $\eta_i = \mathrm{tr}(I_p + \Omega)^{-i}$.

Combining Theorem 3.1 and Lemma 3.1, we obtain the following theorem.

Theorem 3.2. *Suppose that T and Ω are the same ones as in Theorem 3.1. Then, if p_1 is fixed,*

$$\mathrm{E}\left[N^2\mathrm{tr}T^{-1}\left\{\left(1+\frac{1}{N}\right)I_p+\Omega\right\}\right]=\frac{N\{Np+p-p_1^2-3p_1\}}{N-p-2}$$
$$+\left(\frac{N-p_1-2}{N-p-2}\right)\left\{K+O(N^{-1/2})\right\}, \tag{3.8}$$

where

$$K=2(p_1+2)\eta_1-\eta_2-\eta_1^2, \qquad \eta_i=\mathrm{tr}(I_{p_1}+\Omega_{11})^{-i}.$$

It may be noted that the $O(N^{-1/2})$ in (3.8) is not the same one as in the usual large-sample case where p is fixed. The result (3.8) holds also under the assumption that p is not fixed, for example, under HD1.

3.3 Evaluations of (3.1)

In this section we evaluate each of the expectations in (3.1). Using Theorem 3.2 the expectation of the first term in (3.1) is expressed as

$$\mathrm{E}\left[N^2\mathrm{tr}T^{-1}\left\{\left(1+\frac{1}{N}\right)I_p+\Omega_p\right\}\right]$$
$$=\frac{N(Np+p-q^2-3q)}{N-p-2}+\left(\frac{N-q-2}{N-p-2}\right)\{K_\omega+O(N^{-1/2})\}, \tag{3.9}$$

where

$$K_\omega=2(q+2)\eta_{1\omega}-\eta_{2\omega}-\eta_{1\omega}^2, \quad \eta_{i\omega}=\mathrm{tr}(I_q+\Omega_q)^{-i}, \tag{3.10}$$

and $\Omega_q=\mathrm{diag}(\omega_1,\ldots,\omega_q)$. We note that the result holds under the asymptotic framework HD1, and also HD2 since the expectation does not depend on k.

Next we show that

$$\mathrm{E}\left[N^2\mathrm{tr}T_k^{-1}\left\{\left(1+\frac{1}{N}\right)I_k+\Psi_k\right\}\right]$$
$$=\frac{N(Nk+k-q^2-3q)}{N-k-2}+\left(\frac{N-q-2}{N-k-2}\right)\left\{K_\psi+O(N^{-1/2})\right\}, \tag{3.11}$$

where

$$K_\psi = 2(q+2)\eta_{1\psi} - \eta_{2\psi} - \eta_{1\psi}^2, \quad \eta_{i\psi} = \mathrm{tr}(I_q + \Psi_q)^{-i}, \tag{3.12}$$

and $\Psi_q = \mathrm{diag}(\psi_1, \ldots, \psi_q)$. It is shown that the result holds for $k > q$, $k = q$ and k, q. In fact, when $k > q$, from Theorem 3.2 we have

$$\begin{aligned}
&\mathrm{E}\left[N^2 \mathrm{tr} T_k^{-1}\left\{\left(1+\frac{1}{N}\right) I_k + \Psi_k\right\}\right] \\
&= \frac{N(N+1)(k-q))}{N-k-2} + \left(\frac{N-q-2}{N-k-2}\right) \mathrm{E}\left[\mathrm{tr} T_q^{-1}\left\{\left(1+\frac{1}{N}\right) I_q + \Psi_q\right\}\right] \\
&= \frac{N(Nk+k-q^2-3q)}{(N-k-2)} + \left(\frac{N-q-2}{N-k-2}\right)\left\{K_\psi + O(N^{-1/2})\right\}.
\end{aligned}$$

When $k = q$, then $T_k = T_q$, $\Psi_k = \Psi_q$, and hece we have

$$\begin{aligned}
&\mathrm{E}\left[N^2 \mathrm{tr} T_k^{-1}\left\{\left(1+\frac{1}{N}\right) I_k + \Psi_k\right\}\right] \\
&= \mathrm{E}\left[N^2 \mathrm{tr} T_q^{-1}\left\{\left(1+\frac{1}{N}\right) I_q + \Psi_q\right\}\right] = Nq + K_\psi + O(N^{-1/2}).
\end{aligned}$$

When $k < q$, we may regard $T_k \sim W_k(n, I_k; \Psi_k)$ as a submatrix of $T_q \sim W_q(n, I_q; \Psi_q)$. Moreprecisely, we may write

$$T_q = \begin{pmatrix} T_k & * \\ * & * \end{pmatrix}, \quad \Psi_q = \begin{pmatrix} \Psi_k & O \\ O & O \end{pmatrix}, \quad T_q, \Psi_q; q \times q, \quad T_k, \Psi_k; k \times k.$$

Using Theorem 3.1 we have

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{tr} T_q^{-1}\left\{\left(1+\frac{1}{N}\right) I_q + \Psi_q\right\}\right] = \left(1+\frac{1}{N}\right)\frac{q-k}{N-q-2} \\
&+ \left(\frac{N-k-2}{N-q-2}\right) \mathrm{E}\left[\mathrm{tr} T_k^{-1}\left\{\left(1+\frac{1}{N}\right) I_k + \Psi_k\right\}\right].
\end{aligned}$$

This gives us

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{tr} T_k^{-1}\left\{\left(1+\frac{1}{N}\right) I_k + \Psi_k\right\}\right] = \left(1+\frac{1}{N}\right)\frac{k-q}{N-k-2} \\
&+ \left(\frac{N-q-2}{N-k-2}\right) \mathrm{E}\left[\mathrm{tr} T_q^{-1}\left\{\left(1+\frac{1}{N}\right) I_q + \Psi_q\right\}\right].
\end{aligned}$$

In this case $\Psi_q = \mathrm{diag}(\psi_1, \ldots, \psi_k, 0, \ldots, 0)$. Again, using Lemma 3.1 we obtain (3.11).

Theorem 3.3. *Let R_k be the risk function of the model M_k defined by* (2.6)*, and let b_k be the bias term defined by* (2.12) *when we estimate R_k by $\ell_k(W,T) = -2\log f(X;\hat{\Theta}_k)$. Then the bias term can be expressed as*

$$\begin{aligned} b_k &= b_{HD} \\ &\quad + \left(\frac{N-q-2}{N-p-2}\right) \times O(N^{-1/2}) + \left(\frac{N-q-2}{N-k-2}\right) \times O(N^{-1/2}), \end{aligned} \tag{3.13}$$

where $b_{HD} = b_{HD}^{(0)} + b_{HD}^{(1)}$,

$$\begin{aligned} b_{HD}^{(0)} &= -Np + \frac{Nk(N+q+1)}{N-k-q-2} \\ &\quad + \frac{N(Np+p-q^2-3q)}{N-p-2} - \frac{N(Nk+k-q^2-3q)}{N-k-2}, \end{aligned} \tag{3.14}$$

$$b_{HD}^{(1)} = \left(\frac{N-q-2}{N-p-2}\right) K_\omega - \left(\frac{N-q-2}{N-k-2}\right) K_\psi, \tag{3.15}$$

and

$$\begin{aligned} K_\omega &= 2(q+2)\eta_{1\omega} - \eta_{2\omega} - \eta_{1\omega}^2, \quad \eta_{ip} = \mathrm{tr}(I_q + \Omega_q)^{-i}, \\ K_\psi &= 2(q+2)\eta_{1\psi} - \eta_{2\psi} - \eta_{1\psi}^2, \quad \eta_{i\psi} = \mathrm{tr}(I_q + \Psi_q)^{-i}, \end{aligned}$$

$\Omega_q = \mathrm{diag}(\omega_1, \ldots, \omega_q)$, $\Psi_q = \mathrm{diag}(\psi_1, \ldots, \psi_q)$, and $\omega_1 \geq \cdots \geq \omega_q \geq 0$ and $\psi_1 \geq \cdots \geq \psi_q \geq 0$ are possibly non-zero roots of $\Sigma^{-1}\Xi$ and $\Sigma_{11}^{-1}\Xi_{11}$, respectively.

In general, the bias term b_{HD} includes divergent terms under HD2. However, it is possible to make it a convergent term under some additional assumptions. In the following we examine the limiting value of b_{HD} when $p/N \to 0$. Note that the $p/N \to 0$ implies that $k/N \to 0$ and $q/N \to 0$. The constant term $b_{HD}^{(0)}$ can be expressed as

$$\begin{aligned} b_{HD}^{(0)} &= 2\left\{k(q+1) + p - k + \frac{1}{2}p(p+1)\right\} + \frac{k(k+q+2)(k+2q+3)}{N-k-q-2} \\ &\quad + \frac{(p+2)(p-q)(p+q+3)}{N-p-2} - \frac{(k+2)(k-q)(k+q+3)}{N-k-2}. \end{aligned} \tag{3.16}$$

Therefore, we have

$$\lim_{p/N\to 0} b_{HD}^{(0)} = b_{LS}^{(0)}.$$

Similarly, it is pointed that the constant term has the following property

$$b_{HD}^{(0)} = b_{LS}^{(0)} + O_{LS}(N^{-1/2}),$$

where $O_{LS}(N^{-1/2})$ denotes the order under the large-sample framework, i.e., under the case when p is fixed.

In order to get the relationship between $b_{LS}^{(1)}$ and $b_{HD}^{(1)}$, we use that

$$\begin{aligned}
\mathrm{tr}C &= \mathrm{tr}\Sigma^{-1}\Xi(I_p + \Sigma^{-1}\Xi)^{-1} = \sum_{i=1}^{q} \frac{\omega_i}{1+\omega_i} \\
&= \sum_{i=1}^{q} \frac{(1+\omega_i)-1}{1+\omega_i} = q - \eta_{1\omega}, \\
\mathrm{tr}C^2 &= \mathrm{tr}\left\{\Sigma^{-1}\Xi(I_p + \Sigma^{-1}\Xi)^{-1}\right\}^2 = \sum_{i=1}^{q} \left(\frac{\omega_i^2}{1+\omega_i}\right)^2 \\
&= \sum_{i=1}^{q} \frac{(1+\omega_i)^2 - 2(1+\omega_i) + 1}{(1+\omega_i)^2} = q - 2\eta_{1\omega} + \eta_{2\omega}.
\end{aligned}$$

Furthermore,

$$2\left(\mathrm{tr}C + \frac{1}{2}\mathrm{tr}C^2 + \frac{1}{2}(\mathrm{tr}C)^2\right) = q^2 + 3q - K_\omega,$$

and similary

$$2\left(\mathrm{tr}C_k + \frac{1}{2}\mathrm{tr}C_k^2 + \frac{1}{2}(\mathrm{tr}C_k)^2\right) = q^2 + 3q - K_\psi.$$

These imply that

$$\lim_{p/N\to 0} b_{HD}^{(1)} = b_{LS}^{(1)}, \text{ and hence } \lim_{p/N\to 0} b_{HD} = b_{LS}. \tag{3.17}$$

It is also pointed that

$$b_{HD}^{(1)} = b_{LS}^{(1)} + O_{LS}(N^{-1/2}), \text{ and hence } b_{HD} = b_{LS} + O_{LS}(N^{-1/2}). \tag{3.18}$$

4. High-Dimensional AIC Criteria

In order to get an asymptotic unbiased estimator of b_k or R_k it is enough to get asymptotic unbiased estimators for K_ω and K_ψ. Let W and B be the matrices of sums of squares and products due to within-groups and between-groups, respectively. Then, we know that W and B are independently distributed as $W_p(N-q-1,\Sigma)$ and $W_p(q,\Sigma;N\Xi)$. Further, let W_k and B_k be the first $k\times k$ submatrices of W and B, respectively. Then, W_k and B_k are independently distributed as $W_k(N-q-1,\Sigma_k)$ and $W_k(q,\Sigma_k;N\Xi_k)$, respectively. When k is fixed, based on large-sample theory it is easy to see that

$$\begin{aligned}\mathrm{E}\left[\mathrm{tr}(I_k+W_k^{-1}B_k)^{-i}\right] &= \mathrm{tr}(I_k+\Sigma_k^{-1}\Xi_k)^{-i}+\tilde{O}(N^{-1/2})\\ &= \eta_{i\psi}+k-q+\tilde{O}(N^{-1/2}),\end{aligned}$$

where $\tilde{O}_(N^{-j})$ denotes the term of the jth order with respect to N^{-1} when k is fixed. This suggests an asymptotic unbiased estimator of $\eta_{i\psi}$ defined by

$$\tilde{\eta}_{i\psi}=\mathrm{tr}(I_k+W_k^{-1}B_k)^{-i}-(k-q). \tag{4.1}$$

Therefore, when k is fixed, we get an asymptotic unbiased estimator of K_ψ defined by

$$\tilde{K}_\psi=2(q+2)\tilde{\eta}_{1\psi}-\tilde{\eta}_{2\psi}-\tilde{\eta}_{1\psi}^2. \tag{4.2}$$

Next we consider to estimate $\eta_{i\omega}$ and $\eta_{i\psi}$ under HD1 or HD2. We use high-dimensional estimators of $\eta_{i\omega}$ and $\eta_{i\psi}$ defined by

$$\hat{\eta}_{i\omega}=(1-c)^{-i}\mathrm{tr}\left(I_p+W^{-1}B\right)^{-i}-(1-c)^{-i}(p-q), \tag{4.3}$$

$$\hat{\eta}_{i\psi}=(1-d)^{-i}\mathrm{tr}\left(I_k+W_k^{-1}B_k\right)^{-i}-(1-d)^{-i}(k-q). \tag{4.4}$$

These estimators are asymptotically unbiased estimators as in Theorem 4.1 whose proof is given in Appendix.

Theorem 4.1. *Let $\hat{\eta}_{i\omega}$ be the estimator in* (4.3) *of $\eta_{i\omega}$ in* (3.10)*. Suppose that $c = p/N \to c_0 \in (0,1)$. Then,*

$$\mathrm{E}(\hat{\eta}_{i\omega}) = \eta_{i\omega} + O_{1/2}(N^{-1}, p^{-1}).$$

Similarly, let $\hat{\eta}_{i\psi}$ be the estimator in (4.4) *of $\eta_{i\psi}$ in* (3.12)*. Suppose that $d = k/N \to d_0 \in (0,1)$. Then,*

$$\mathrm{E}(\hat{\eta}_{i\psi}) = \eta_{i\psi} + O_{1/2}(N^{-1}, k^{-1}).$$

Using (4.3) and (4.4) we have the following high-dimensional asymptotic estimators of K_ω and K_ψ:

$$\hat{K}_\omega = 2(q+2)\hat{\eta}_{1\omega} - \hat{\eta}_{2\omega} - \hat{\eta}_{1\omega}^2, \quad \hat{K}_\psi = 2(q+2)\hat{\eta}_{1\psi} - \hat{\eta}_{2\psi} - \hat{\eta}_{1\psi}^2. \tag{4.5}$$

When d tends to zero, we can see that $\hat{K}_\psi$ tends to $\tilde{K}_\psi$. This means that the estimator $\hat{K}_\psi$ may be used even for the case when k is small. Therefore, we propose the following high-dimensional estimator of b_{HD}:

$$\hat{b}_{HD} = b_{HD}^{(0)} + \hat{b}_{HD}^{(1)}, \tag{4.6}$$

which may be used when k is small, where

$$\hat{b}_{HD}^{(1)} = \left(\frac{N-q-2}{N-p-2}\right)\hat{K}_\omega - \left(\frac{N-q-2}{N-k-2}\right)\hat{K}_\psi.$$

Theorem 4.2. *Under the high-dimensional asymptotic framework* HD2 *it holds that*

$$\mathrm{E}(\hat{b}_{HD}) = b_k + O_{1/2}(N^{-1}, p^{-1}, k^{-1}) \tag{4.7}$$

Proof. From Theorem 4.1 we have

$$\mathrm{E}(\hat{b}_{HD}) = b_{HD} + O_{1/2}(N^{-1}, p^{-1}, k^{-1}).$$

The final result is obtained by using Theorem 3.3. □

Finally we propose the following high-dimensional AIC:

$$\mathrm{HAIC} = \ell_k(W,T) + b_{HD}^{(0)} + \hat{b}_{HD}^{(1)}. \tag{4.8}$$

We have seen that HAIC is an asymptotic unbiased estimator of R_k under HD2. We have pointed some relationships between b_{LS} and b_{HD}. So we can say that such relationships can be expected also for $\hat{b}_{LS}$ and $\hat{b}_{HD}$.

When $q = 1$, we have

$$B = \frac{N_1 N_2}{N}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})', \quad W = (N-2)S,$$

where S is the usual pooled sample covariance matrix. Then, it is well known (see, e.g., Fujikoshi et al. (2010)) that

$$\begin{aligned} \ell_k(W,T) &= N\log\left\{1 + \frac{a(D^2 - D_1^2)}{N - 2 + aD_1^2}\right\} + N\log|\{(N-2)/N\}S| \\ &\quad + Np(1 + \log 2\pi), \end{aligned}$$

where D and D_1 are the Mahalanobis distances based on $\boldsymbol{x}$ and $\boldsymbol{x}_1$, respectively, and $a = (N_1 N_2)/N$. The quantity b_k in (3.6) involves the unknown parameters τ^2 and τ_1^2, or equivalently Δ^2 and Δ_1^2. In a practical use it is suggested to replace these by the following unbiased estimators (see, e.g., Fujikoshi et al. (2010)):

$$\tilde{\Delta}^2 = \frac{N-p-3}{N-2}D^2 - \frac{Np}{N_1 N_2}, \quad \tilde{\Delta}_1^2 = \frac{N-k-3}{N-2}D_1^2 - \frac{Nk}{N_1 N_2}.$$

Therefore, it is suggested to use

$$\begin{aligned} \tilde{\mathrm{H}}\mathrm{AIC} &= N\log\left\{1 + \frac{a(D^2 - D_1^2)}{N - 2 + aD_1^2}\right\} + N\log|\{(N-2)/N\}S| \\ &\quad + Np(1 + \log 2\pi) - Np + \frac{Nk(N+2)}{N-k-3} \\ &\quad + N^2\{Q_1(N,p,\tilde{\tau}^2) - Q_1(N,k,\tilde{\tau}_1^2)\}, \end{aligned} \tag{4.9}$$

where $\tilde{\tau}^2 = \{(N_1 N_2)/N^2\}\tilde{\Delta}^2$ and $\tilde{\tau}_1^2 = \{(N_1 N_2)/N^2\}\tilde{\Delta}_1^2$. Here, we note that $\tilde{\mathrm{H}}\mathrm{AIC}$ is not exactly the same as HAIC with $q = 1$, but they are asymptotically equivalent in a high-dimensional sense. In fact, we have

$$\begin{aligned} \frac{1}{1+\tilde{\tau}^2} &= \left\{1 + \frac{N_1 N_2}{N^2}\frac{N-p-3}{N-2}D^2 - \frac{p}{N}\right\}^{-1} \\ &\approx (1-c)^{-1}\left\{1 + \frac{N_1 N_2}{N}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})' W^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})\right\}^{-1}, \end{aligned}$$

where "≈" means an asymptotially equivalence in a high-dimensional framework HD1. Further, using that $(I_p + \boldsymbol{u}\boldsymbol{v}')^{-1} = I_p - (1 + \boldsymbol{u}'\boldsymbol{v})^{-1}\boldsymbol{u}\boldsymbol{v}'$,

$$\begin{aligned}
&\left\{1 + \frac{N_1 N_2}{N}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})' W^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})\right\}^{-1} \\
&\quad = \mathrm{tr}\left\{I_p + \frac{N_1 N_2}{N} W^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'\right\}^{-1} - (p-1) \\
&\quad = \mathrm{tr}(I_p + W^{-1}B)^{-1} - (p-1).
\end{aligned}$$

Similarly

$$\frac{1}{(1+\tilde{\tau}^2)^2} \approx (1-c)^{-2}\mathrm{tr}(I_p + W^{-1}B)^{-2} - (1-c)^{-2}(p-1).$$

These imply that $\tilde{\mathrm{H}}$AIC is asymptoticall the same as HAIC with $q = 1$.

5. Simulation Results

In this section, we attempt to give an impression of the relative performances of AIC, CAIC and HAIC as an estimator of R_k through simulation experiments. Furthermore, we also examine relative frequencies selected by the criteria. The risk function R_k is expressed as

$$R_k = E_X^*[-2\log f(X; \hat{\Theta}_k)] + b_k,$$

where b_k is defined by (2.7). The three information criteria are given as

$$\begin{aligned}
\mathrm{AIC} &= \ell_k(W,T) + b_{LS}^{(0)}, \\
\mathrm{CAIC} &= \ell_k(W,T) + b_{LS}^{(0)} + \hat{b}_{LS}^{(1)}, \\
\mathrm{HAIC} &= \ell_k(W,T) + b_{HD}^{(0)} + \hat{b}_{HD}^{(1)},
\end{aligned}$$

where $\ell_k(W,T) = -2\log f(X; \hat{\Theta}_k)$.

In our simulation experiments, we assume that the true model is

$$M_{k^*}: \ \boldsymbol{x}|\Pi_i \sim N(\boldsymbol{\mu}^{(i)}, \Sigma), \ (i = 1, \ldots, q+1), \ \boldsymbol{\mu}_{2\cdot 1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot 1}^{(q+1)}; (p-k^*) \times 1.$$

The covariance matrix was assumed to be $\Sigma = I_p$. We considered the following three cases of (p, k^*, q):

$P_1;\ (p, k^*, q) = (5, 4, 2),\ P_2;\ (p, k^*, q) = (10, 8, 2),\ P_3;\ (p, k^*, q) = (20, 10, 2).$

In our setting the elements of $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, \ldots, \mu_p^{(i)})'$ were determined as follows:

$P_1;\ p = 5,\ k^* = 4,\ q = 2$

j	1	2	3	4	5
$\mu_j^{(1)}$	3.67	-3.67	0.00	0.00	0.00
$\mu_j^{(2)}$	2.12	2.12	-4.24	0.00	0.00
$\mu_j^{(3)}$	0.50	0.50	0.50	-1.50	0.00

$P_2;\ p = 10,\ k^* = 8,\ q = 2$

j	1	2	3	4	5	6	7	8	9	10
$\mu_j^{(1)}$	0.95	0.95	0.95	0.95	0.95	-4.74	0.00	0.00	0.00	0.00
$\mu_j^{(2)}$	0.80	0.80	0.80	0.80	0.80	0.80	-4.81	0.00	0.00	0.00
$\mu_j^{(3)}$	0.23	0.23	0.23	0.23	0.23	0.23	0.23	-1.62	0.00	0.00

$P_3;\ p = 20,\ k^* = 10,\ q = 2$

j	1	2	3	4	5	6	7	8	9	10	11	$\cdots$	20
$\mu_j^{(1)}$	0.69	0.69	0.69	0.69	0.69	0.69	0.69	-4.86	0.00	0.00	0.00	$\cdots$	0.00
$\mu_j^{(2)}$	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	-4.90	0.00	0.00	$\cdots$	0.00
$\mu_j^{(3)}$	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	-1.64	0.00	$\cdots$	0.00

The candidate models considered are considered $M_k, k = 1, \ldots, p$. Let Σ_{kk} and Ξ_{kk} be the first $k \times k$ matrices of Σ and Ξ, respectively. Let $\omega_{k1} \geq \omega_{k2} \geq 0$ be the possible non-zero roots of $\Sigma^{-1}\Xi$. In our setting the roots were determined as follows:

$P_1;\ p = 5,\ k^* = 4,\ q = 2$

k	1	2	3	4	5
ω_{k1}	1.68	6.87	9.00	9.00	9.00
ω_{k2}	NA	0.76	3.17	3.67	3.67

$P_2;\ p = 10,\ k^* = 8,\ q = 2$

k	1	2	3	4	5	6	7	8	9	10
ω_{k1}	0.10	0.19	0.29	0.38	0.48	6.37	9.00	9.00	9.00	9.00
ω_{k2}	NA	0.00	0.00	0.00	0.00	0.31	3.08	3.67	3.67	3.67

P_3; $p = 20$, $k^* = 10$, $q = 2$

k	1	2	3	4	5	6	7	8	9	10	11	$\cdots$	20
ω_{k1}	0.10	0.19	0.29	0.38	0.48	6.37	9.00	9.00	9.00	9.00	9.00	$\cdots$	9.00
ω_{k2}	NA	0.00	0.00	0.00	0.00	0.31	3.08	3.67	3.67	3.67	3.67	$\cdots$	3.67

(NA means that the value is not defined)

For the sample sizes (N_1, N_2, N_3) the following three cases were considered:

$$N_a;\ (N_1, N_2, N_3) = (30, 30, 30), \quad N_b;\ (N_1, N_2, N_3) = (50, 50, 50),$$
$$N_c;\ (N_1, N_2, N_3) = (100, 100, 100).$$

Then, first the averages of R_k, AIC, CAIC and HAIC were computed with 10,000 replications. The results are given in Table 1. In Table 2 we give differences between R_k and each of AIC, CAIC and HAIC. In Table 3 we give selected percentages of AIC, CAIC and HAIC.

From Table 1 and Table 2, the large sample approximations AIC and CAIC perform well for P_1. In the case of P_1, we can see that CAIC is better than AIC in the estimation of R_k. In contrast, the large sample approximations AIC and CAIC are poor for P_2 and P_3. When p is large, we can see that HAIC is better than AIC and CAIC in the estimation of R_k in particular that sample size is small. In particular, the approximation HAIC is the best of these approximations for all cases.

For model selections of AIC, CAIC and HAIC, from Table 3 the probabilities of selecting the true model are increasing as the sample-sizes are increasing. HAIC selects the true model with higher probabilities than does AIC or CAIC.

Table 1. Risks and the averages of AIC, CAIC and HAIC

The case: (P_1, N_a)

M_k	Risk	AIC	CAIC	HAIC
1	1555.9	1558.5	1552.8	1555.9
2	1412.0	1410.6	1408.2	1411.9
3	1316.8	1312.4	1312.2	1316.7
4	1309.6	1304.0	1304.0	1309.5
5	1312.5	1305.9	1305.9	1312.5

The case; (P_1, N_b)

M_k	Risk	AIC	CAIC	HAIC
1	2577.9	2581.7	2576.1	2577.9
2	2334.9	2335.2	2332.8	2334.9
3	2173.7	2171.3	2171.1	2173.7
4	2159.3	2156.2	2156.2	2159.3
5	2162.0	2158.1	2158.1	2162.0

The case ; (P_1, N_c)

M_k	Risk	AIC	CAIC	HAIC
1	5134.1	5139.4	5133.8	5134.7
2	4644.4	4646.2	4643.8	4644.8
3	4318.0	4317.3	4317.1	4318.4
4	4286.9	4285.2	4285.2	4286.8
5	4289.0	4287.2	4287.2	4289.0

The case; (P_2, N_a)

M_k	Risk	AIC	CAIC	HAIC
1	2971.4	2958.9	2951.2	2969.7
2	2966.1	2953.1	2945.7	2964.6
3	2961.7	2948.0	2940.8	2960.3
4	2958.0	2943.3	2936.3	2956.8
5	2955.0	2939.0	2932.3	2953.8
6	2791.2	2770.4	2767.0	2790.0
7	2666.5	2640.5	2640.2	2665.2
8	2657.9	2630.2	2630.2	2657.0
9	2661.2	2631.9	2631.9	2660.8
10	2664.9	2633.6	2633.6	2664.9

The case; (P_2, N_b)

M_k	Risk	AIC	CAIC	HAIC
1	4892.0	4888.1	4880.5	4891.0
2	4881.6	4877.5	4870.2	4880.9
3	4872.4	4867.9	4860.7	4871.8
4	4864.0	4859.0	4852.1	4863.6
5	4856.5	4850.7	4844.0	4856.1
6	4579.7	4569.6	4566.2	4579.2
7	4367.3	4353.2	4352.9	4367.0
8	4350.5	4335.0	4335.0	4350.1
9	4353.3	4336.9	4336.8	4353.1
10	4356.2	4338.7	4338.7	4356.2

The case; (P_2, N_c)

M_k	Risk	AIC	CAIC	HAIC
1	9707.2	9709.0	9701.5	9706.5
2	9684.3	9685.8	9678.4	9683.6
3	9663.2	9664.3	9657.2	9662.5
4	9644.1	9644.6	9637.6	9643.2
5	9626.0	9626.4	9619.7	9625.5
6	9067.7	9064.5	9061.1	9067.3
7	8638.1	8631.4	8631.1	8637.9
8	8600.6	8593.1	8593.1	8600.3
9	8602.6	8595.0	8595.0	8602.8
10	8605.3	8596.9	8596.9	8605.3

The case; (P_3, N_a)

k	R_k	AIC	CAIC	HAIC
1	5803.5	5665.0	5657.0	5798.6
2	5802.0	5662.6	5654.8	5796.8
3	5800.5	5660.4	5652.7	5795.4
4	5799.7	5658.4	5650.8	5794.4
5	5798.7	5656.5	5649.1	5793.7
6	5798.0	5654.6	5647.4	5793.3
7	5797.9	5652.9	5645.9	5793.3
8	5632.6	5482.3	5478.6	5628.0
9	5502.4	5346.6	5346.1	5498.2
10	5494.2	5335.8	5335.6	5490.0
11	5498.5	5337.5	5337.2	5494.4
12	5502.8	5339.1	5338.9	5499.0
13	5507.3	5340.7	5340.5	5503.9
14	5511.8	5342.3	5342.1	5509.1
15	5517.0	5343.8	5343.7	5514.6
16	5522.6	5345.4	5345.3	5520.4
17	5528.2	5346.8	5346.7	5526.6
18	5534.4	5348.2	5348.2	5533.1
19	5540.3	5349.6	5349.6	5540.0
20	5547.3	5350.9	5350.9	5547.3

The case (P_3, N_b)

k	R_k	AIC	CAIC	HAIC
1	9377.0	9307.3	9299.5	9374.8
2	9372.4	9302.3	9294.6	9370.1
3	9368.3	9297.4	9289.9	9365.8
4	9364.4	9292.8	9285.4	9361.8
5	9360.7	9288.4	9281.2	9358.2
6	9357.3	9284.4	9277.2	9354.9
7	9354.4	9280.5	9273.5	9352.1
8	9074.6	8996.6	8992.8	9072.5
9	8853.4	8770.3	8770.0	8851.0
10	8835.7	8751.4	8751.2	8833.6
11	8839.1	8753.2	8753.1	8836.8
12	8842.2	8755.0	8754.9	8840.2
13	8845.6	8756.8	8756.7	8843.6
14	8848.7	8758.5	8758.4	8847.2
15	8852.2	8760.2	8760.1	8850.8
16	8856.0	8762.0	8761.9	8854.7
17	8859.8	8763.6	8763.6	8858.7
18	8863.8	8765.3	8765.3	8862.8
19	8867.4	8767.0	8767.0	8867.1
20	8871.5	8768.7	8768.7	8871.5

The case; (P_3, N_a)

k	R_k	AIC	CAIC	HAIC
1	18423.7	18395.3	18387.6	18422.3
2	18411.7	18383.1	18375.6	18410.4
3	18401.3	18371.6	18364.1	18399.1
4	18390.5	18360.7	18353.3	18388.6
5	18380.8	18350.3	18343.1	18378.6
6	18371.0	18340.4	18333.3	18369.2
7	18361.7	18331.0	18324.0	18360.3
8	17797.4	17762.9	17759.2	17795.9
9	17349.0	17310.4	17310.1	17347.5
10	17310.5	17270.8	17270.7	17308.8
11	17313.1	17272.8	17272.7	17311.3
12	17315.6	17274.7	17274.6	17314.0
13	17318.8	17276.6	17276.5	17316.6
14	17321.0	17278.5	17278.4	17319.3
15	17323.6	17280.4	17280.3	17322.1
16	17325.5	17282.2	17282.2	17324.9
17	17328.5	17284.1	17284.1	17327.7
18	17331.5	17285.9	17285.9	17330.6
19	17333.8	17287.8	17287.7	17333.5
20	17336.6	17289.6	17289.6	17336.6

Table 2. Differences of Risks and each of AIC, CAIC and HAIC

The case; P_1

	AIC			CAIC			HAIC		
M_k	N_a	N_b	N_c	N_a	N_b	N_c	N_a	N_b	N_c
1	-2.6	-3.8	-5.3	3.1	1.8	0.3	0.0	0.0	-0.6
2	1.4	-0.3	-1.8	3.8	2.1	0.6	0.1	0.0	-0.5
3	4.4	2.5	0.8	4.6	2.7	0.9	0.2	0.1	-0.3
4	5.6	3.1	1.7	5.6	3.1	1.7	0.1	-0.1	0.1
5	6.7	3.8	1.9	6.7	3.8	1.9	0.0	0.0	0.0

The case; P_2

	AIC			CAIC			HAIC		
M_k	N_a	N_b	N_c	N_a	N_b	N_c	N_a	N_b	N_c
1	12.5	3.9	-1.8	20.1	11.4	5.7	1.7	1.0	0.7
2	13.0	4.0	-1.5	20.5	11.4	5.8	1.6	0.7	0.7
3	13.7	4.5	-1.1	20.9	11.7	6.0	1.4	0.6	0.7
4	14.7	4.9	-0.4	21.6	11.9	6.5	1.2	0.3	0.9
5	16.0	5.8	-0.4	22.7	12.6	6.4	1.2	0.4	0.5
6	20.8	10.1	3.2	24.3	13.5	6.7	1.3	0.5	0.4
7	26.0	14.2	6.8	26.2	14.4	7.0	1.3	0.3	0.3
8	27.7	15.5	7.5	27.8	15.5	7.5	0.9	0.4	0.3
9	29.3	16.4	7.6	29.3	16.5	7.6	0.4	0.2	-0.2
10	31.3	17.5	8.3	31.3	17.5	8.3	0.0	0.0	0.0

The case; P_3

	AIC			CAIC			HAIC		
M_k	N_a	N_b	N_c	N_a	N_b	N_c	N_a	N_b	N_c
1	138.5	69.6	28.4	146.5	77.5	36.1	4.9	2.2	1.4
2	139.4	70.1	28.6	147.2	77.8	36.2	5.2	2.2	1.3
3	140.1	70.9	29.7	147.8	78.5	37.1	5.1	2.5	2.1
4	141.3	71.6	29.8	148.9	79.0	37.1	5.3	2.5	1.9
5	142.3	72.3	30.5	149.7	79.6	37.7	5.0	2.5	2.2
6	143.4	72.9	30.5	150.7	80.1	37.7	4.7	2.4	1.8
7	145.0	73.8	30.6	152.1	80.9	37.6	4.7	2.3	1.4
8	150.2	78.0	34.5	154.0	81.7	38.2	4.5	2.1	1.5
9	155.8	83.1	38.6	156.3	83.4	38.9	4.3	2.4	1.5
10	158.4	84.3	39.6	158.6	84.4	39.7	4.1	2.1	1.7
11	161.0	85.9	40.4	161.3	86.0	40.4	4.1	2.2	1.8
12	163.7	87.2	41.0	163.9	87.3	41.0	3.8	2.0	1.7
13	166.6	88.9	42.2	166.7	89.0	42.3	3.4	2.0	2.2
14	169.6	90.2	42.5	169.7	90.3	42.5	2.8	1.5	1.7
15	173.2	92.0	43.2	173.3	92.1	43.2	2.4	1.4	1.5
16	177.2	94.0	43.3	177.3	94.1	43.3	2.1	1.2	0.6
17	181.4	96.2	44.4	181.4	96.2	44.5	1.6	1.1	0.8
18	186.1	98.5	45.6	186.2	98.5	45.6	1.2	1.0	0.9
19	190.7	100.4	46.0	190.8	100.4	46.0	0.3	0.3	0.2
20	196.4	102.9	47.0	196.4	102.9	47.0	0.0	0.0	0.0

Table 3. Selected percentages of AIC, CAIC and HAIC

The case; P_1

	N_a			N_b			N_c		
M_k	AIC	CAIC	HAIC	AIC	CAIC	HAIC	AIC	CAIC	HAIC
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
3	7.1%	7.3%	11.7%	0.7%	0.8%	1.2%	0.0%	0.0%	0.0%
4	78.3%	78.5%	80.5%	84.9%	85.0%	88.6%	85.6%	85.7%	87.7%
5	14.6%	14.2%	7.9%	14.4%	14.2%	10.2%	14.4%	14.3%	12.3%

The case; P_2

	N_a			N_b			N_c		
M_k	AIC	CAIC	HAIC	AIC	CAIC	HAIC	AIC	CAIC	HAIC
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7	4.2%	4.3%	10.6%	0.4%	0.4%	0.8%	0.0%	0.0%	0.0%
8	72.8%	73.0%	81.4%	79.3%	79.6%	88.4%	81.1%	81.2%	86.0%
9	14.7%	14.5%	6.5%	13.6%	13.4%	8.2%	12.5%	12.4%	10.0%
10	8.4%	8.2%	1.5%	6.7%	6.6%	2.6%	6.5%	6.4%	4.0%

The case; P_3

	N_a			N_b			N_c		
M_k	AIC	CAIC	HAIC	AIC	CAIC	HAIC	AIC	CAIC	HAIC
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	3.7%	3.9%	12.0%	0.2%	0.2%	0.7%	0.0%	0.0%	0.0%
10	65.2%	65.6%	81.6%	73.5%	73.8%	88.4%	77.7%	78.0%	85.7%
11	12.2%	12.1%	5.1%	11.5%	11.5%	7.3%	10.9%	10.8%	8.7%
12	5.8%	5.6%	1.0%	5.5%	5.4%	2.4%	4.6%	4.6%	3.1%
13	3.7%	3.6%	0.2%	3.1%	3.0%	0.7%	2.3%	2.2%	1.1%
14	2.4%	2.4%	0.1%	1.9%	1.9%	0.3%	1.7%	1.7%	0.8%
15	1.7%	1.6%	0.0%	1.6%	1.5%	0.1%	0.9%	0.9%	0.3%
16	1.5%	1.5%	0.0%	0.9%	0.9%	0.0%	0.6%	0.6%	0.2%
17	1.1%	1.1%	0.0%	0.7%	0.6%	0.0%	0.5%	0.5%	0.1%
18	1.0%	1.0%	0.0%	0.6%	0.5%	0.0%	0.4%	0.4%	0.1%
19	0.8%	0.7%	0.0%	0.4%	0.4%	0.0%	0.2%	0.2%	0.0%
20	1.0%	0.9%	0.0%	0.3%	0.3%	0.0%	0.2%	0.2%	0.0%

We have seen that

$$b_{HD} = b_{LS} + O_{LS}(N^{-1/2}).$$

This means that as N is large, HIC approaches to CAIC. In order to reconfirm this property, we tried the following additional experiments on P_1:

$$N_d;\ (N_1, N_2, N_3) = (150, 150, 150), \quad N_e;\ (N_1, N_2, N_3) = (200, 200, 200).$$

The results are given in Tables 4 and 5. We can see that HAIC approaches to CAIC as N is large.

Table 4. Risks and the averages of AIC, CAIC and HAIC

The case: (P_1, N_d)

M_k	Risk	AIC	CAIC	HAIC
1	7690.2	7695.3	7689.7	7690.3
2	6953.4	6955.3	6952.9	6953.6
3	6462.9	6462.7	6462.5	6463.3
4	6414.2	6413.6	6413.6	6414.6
5	6416.8	6415.6	6415.6	6416.8

The case; (P_1, N_e)

M_k	Risk	AIC	CAIC	HAIC
1	10248.1	10253.3	10247.7	10248.1
2	9264.7	9266.2	9263.9	9264.4
3	8608.8	8608.1	8607.9	8608.6
4	8542.6	8542.2	8542.2	8543.0
5	8545.1	8544.2	8544.2	8545.1

Table 5. Differences of CAIC and HAIC

M_k	N_a	N_b	N_c	N_d	N_e
1	-3.2	-1.8	-0.9	-0.6	-0.4
2	-3.7	-2.1	-1.0	-0.7	-0.5
3	-4.5	-2.6	-1.3	-0.8	-0.6
4	-5.5	-3.2	-1.5	-1.0	-0.8
5	-6.7	-3.8	-1.9	-1.2	-0.9

6. Conclusive Remarks

In this paper, first we derived asymptotic formulas for the bias term in the problem of estimating the AIC type of risk R_k in the multiple-groups

case in the situation where the dimensions p and k may be large. The results were obtained without assuming that the true model is included in the model M_k. It was shown that the high-dimensional approximations are more useful than the large-sample approximations in a sense that the large-sample approximations can be obtained from the high-dimensional approximations by considering their large-sample approximations. Based on the high-dimensional results we proposed HAIC, which is a high-dimensional asymptotic unbiased estimator of R_k under the high-dimensional framework HD2. By simulation experiments, it was shown that for estimating the risk function HAIC is better than the large sample AIC and its corrected CAIC in a wide range. Further, it was pointed that HIC provides better model selections than does AIC or CAIC. It was also noted that HAIC approaches to CAIC as the sample size N is large.

A. Proofs of Theorems 3.1 and 4.1

Proof of Theorem 3.1.

For our derivation, we use the following properties (see Kabe (1964)) on Wishart matrix: Let $T \sim W_p(n, I_p; N\Omega)$ with $n = N - 1$, and T and Ω be partitioned as

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{pmatrix},$$

respectively, where $T_{ij} : p_i \times p_j$ and $\Omega_{ij} : p_i \times p_j$. If Ω_{12}, Ω_{21} and Ω_{22} are zero matrices, then

(1) $T_{22\cdot1} \sim W_{p_2}(n - p_1, I_{p_2})$.

(2) $T_{21}T_{11}^{-1/2} \sim N_{p_2 \times p_1}(\mathrm{O}, I_{p_2} \otimes I_{p_1})$.

(3) $T_{11} \sim W_{p_2}(n, I_{p_1} : \Omega_{11})$.

(4) $T_{22\cdot1}$, $T_{21}T_{11}^{-1/2}$ and T_{11} are independent.

It is easy to see that

$$\begin{aligned}
&\mathrm{tr}T^{-1}\left\{\left(1+\frac{1}{N}\right)I_p+\Omega\right\}-\mathrm{tr}T_{11}^{-1}\left\{\left(1+\frac{1}{N}\right)I_{p_1}+\Omega_{11}\right\}\\
&\quad=\mathrm{tr}T_{22\cdot1}^{-1}\begin{pmatrix}-T_{21}T_{11}^{-1} & I_{p-p_1}\end{pmatrix}\left\{\left(1+\frac{1}{N}\right)I_p+\Omega\right\}\begin{pmatrix}-T_{11}^{-1}T_{12}\\ I_{p-p_1}\end{pmatrix}\\
&\quad=\left(1+\frac{1}{N}\right)\left(\mathrm{tr}W^{-1}ZT_{11}^{-1}Z'+\mathrm{tr}W^{-1}\right)+\mathrm{tr}W^{-1}ZT_{11}^{-1/2}\Omega_{11}T_{11}^{-1/2}Z',
\end{aligned}$$

where $W=T_{22\cdot1}=T_{22}-T_{21}T_{11}^{-1}T_{12}$ and $Z=T_{21}T_{11}^{-1/2}$. From the above properties,

$$W\sim W_{p_2}(n-p_1,I_{p_2}),\quad Z\sim N_{p_2\times p_1}(\mathrm{O},I_{p_2}\otimes I_{p_1}),\quad T_{11}\sim W_{p_1}(n,I_{p_1}:\Omega_{11}),$$

and W, Z and T_{11} are independent. Therefore, the required result is obtained by computing these expectations.

Proof of Theorem 4.1.

We may assume that B and W are independently distributed as $W_p(q,\Sigma;N\Xi)$ and $W_p(N-q-1,\Sigma)$, respectively. Note that the noncentrality matrix $N\Xi$ can be expressed as MM' with a $p\times q$ matrix M. Therefore, B and W can be written as

$$W=\Sigma^{1/2}A\Sigma^{1/2},\quad B=\Sigma^{1/2}ZZ'\Sigma^{1/2},$$

where $Z\sim N_{p\times q}(\Sigma^{-1/2}M,I_p\otimes I_q)$, $A\sim W_p(N-q-1),I_p)$, and Z and A are independent. Let

$$B_q=Z'Z \text{ and } W_q=B^{1/2}(ZA^{-1}Z')^{-1}B^{1/2}.$$

Then, it is known (Wakaki, Fujikoshi and Ulyanov (2002) or Fujikoshi, Ulyanov and Shimizu (2010)) that B_q and W_q are independently distributed as $W_q(p,I_q;N\Gamma)$ and $W_q(m,I_q)$, respectively, where $\Gamma=M'\Sigma^{-1}M$ and $m=N-p-1$. Further, the non-zero characteristic roots of BW^{-1} are equal to the ones of $B_qW_q^{-1}$, and hence we have

$$\mathrm{tr}(I_p+W^{-1}B)^{-i}=\mathrm{tr}(I_q+W_q^{-1}B_q)^{-i}+(p-q).$$

Let U and V be defined by

$$\frac{1}{p}B_q = I_q + \frac{1}{c}\Theta + \frac{1}{\sqrt{p}}U, \quad \frac{1}{m}W_q = I_q + \frac{1}{\sqrt{m}}V,$$

with $c = p/N$. Then, under the asymptotic framework HD2 the limiting distribution of (U, V) is normal. We can see that

$$(1-c)(I_q + W_q^{-1}B_q) = I_q + \Gamma + O_{1/2}(N^{-1}, p^{-1}).$$

Therefore,

$$\begin{aligned}
&(1-c)^{-i}\mathrm{tr}(I_p + W^{-1}B)^{-i} \\
&\quad = (1-c)^{-i}\mathrm{tr}(I_q + W_q^{-1}B_q)^{-i} + (1-c)^{-i}(p-q) \\
&\quad = \mathrm{tr}(I_q + \Gamma)^{-i} + (1-c)^{-i}(p-q) + O_{1/2}(N^{-1}, p^{-1}).
\end{aligned}$$

The first result is obtained by noting that $\mathrm{tr}(I_q + \Gamma)^{-i} = \eta_{i\omega}$. Similarly the second result is proved.

Acknowledgement

The authors would like to thank the editors and a referee for valuable comments.

References

[1] Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices: a review. *Statist. Sinica*, **9**, 611–677.

[2] Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In *Multivariate Analysis-VI*, Ed. P.R. Krishnaian, pp. 219–236, Elsevier Science Publishers B.V.

[3] Fujikoshi, Y. (2002). Selection of variables for discriminant analysis in a high-dimensional case. *Sankhya Ser. A*, **64**, 256–257.

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

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

[6] Ledoit,, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.*, **30**(4), 1081–1102.

[7] Kabe, D. G. (1964). A note on the Bartlett decomposition of a Wishart matrix. *J. Roy. Statist. Soc. Ser. B*, **26**, 270–273.

[8] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, N. Y.

[9] Raudys, S. and Young, D. M. (2004). Results in statistical discriminant analysis: a review of the former Soviet Union literature. *J. Multivariate Anal.*, **89**, 1–35.

[10] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Ohio.

[11] Wakaki, H., Fujikoshi, Y. and Ulyanov, V. (2002). Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large, *Hiroshima Statistical Group Technical Report 10*, **97**, 1–10.