HIROSHIMA MATH. J.
18 (1988), 451–462

Strong consistency of the information criterion for model selection in multivariate analysis

Ryuei NISHII[†], Zhidong BAI[††] and Paruchuri R. KRISHNAIAH[††]

(Received September 21, 1987)

1. Introduction

In the area of model selection, various procedures have been proposed in the literature and their some properties have been examined. In this paper we consider a generalized information criterion (GIC) obtained by the information theoretic approach. According to this procedure, we find the model which minimizes

$$\text{GIC} = -2\log L(\hat{\theta}) + pc_N \tag{1.1}$$

where $L(\hat{\theta})$ is the maximized likelihood and p is the number of parameters. Akaike [1] proposed to take $c_N \equiv 2$, and Rissanen [12] and Schwarz [13] proposed $c_N = \log N$ where N denotes the sample size (see also [2], [8]). Recently Zhao, Krishnaiah and Bai [14] considered the GIC such that

$$\lim_{N\to\infty} c_N/N = 0 \quad \text{and} \quad \lim_{N\to\infty} c_N/\log\log N = +\infty. \tag{1.2}$$

The above criterion is sometimes referred to as the efficient detection (ED) criterion. They used the criterion for the determination of the number of signals under a signal processing model.

In the present paper, we propose to use the ED criterion for certain problems of multivariate analysis. Sometimes the statistician is expected to predict the explanatory variables using some of the response variables under the multivariate regression model. This problem is treated in Section 2 by using the ED criterion, and its consistency is established. Here we may note that Nishii [10] pointed out the inconsistency of Akaike's AIC in calibration. In Section 3 we discuss the selection of variables in discriminant analysis. Our interest is to find the variables which contribute for discrimination between the populations. Section 4 is concerned with canonical correlation analysis, i.e., among two sets of variables we want to find which subsets are important for studying the association between the two sets. The investigations for the above cases will be carried out under a mild condition on the underlying distribution.

2. Multivariate calibration

Let q explanatory variables $\boldsymbol{x} \equiv (x_1, \ldots, x_q)'$ and p response variables $\boldsymbol{y} \equiv (y_1, \ldots, y_p)'$ have the linear relation:

$$\boldsymbol{y} = \boldsymbol{\alpha} + \beta' \boldsymbol{x} + \boldsymbol{e} \tag{2.1}$$

where the error vector $\boldsymbol{e}$ follows $N_p(\boldsymbol{0}, \Sigma)$, and $\boldsymbol{\alpha}$: $p \times 1$, β: $q \times p$ and Σ: $p \times p$ are parameters such that Σ is positive definite. Note that we do not assume that q is less than or equal to p, which is the usual assumption in calibration. Suppose we are interested in estimating $\boldsymbol{x}$ by using observed $\boldsymbol{y}$. If all parameters are known, the maximum likelihood estimate of the unknown explanatory variables $\boldsymbol{x}$ will be obtained by

$$\hat{\boldsymbol{x}} = (\beta \Sigma^{-1} \beta')^{-} \beta \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\alpha}),$$

where $(\beta \Sigma^{-1} \beta')^{-}$ denotes a generalized inverse matrix of $\beta \Sigma^{-1} \beta'$. However, if the last column of $\beta \Sigma^{-1}$ is the zero vector, the response variable y_p would supply no additional information on $\boldsymbol{x}$ (see 8c.4 of [11]). Hence, we want to obtain the best subset of response variables such that each variable has some information. For this problem, criteria based on information theory can be used. For a review of the literature on multivariate calibration, the reader is referred to [4].

Let J be a subset of indices of response variables $\{1, \ldots, p\}$, and J^c be its complement. We say that "the assumed model is J" when y_j provides information on $\boldsymbol{x}$ for any index j in J, whereas y_j does not for any index j' in J^c. We assume the existence of the true but unknown model $\{1, \ldots, p_t\} = J_t$ for $p_t \leq p$. This is equivalent to the following two conditions:

$$\beta_J \Sigma_{JJ}^{-1} \beta_J' = \beta_t \Sigma_{tt}^{-1} \beta_t' \quad \text{if } J \text{ includes the true model } J_t, \tag{2.2}$$

$$\beta_J \Sigma_{JJ}^{-1} \beta_J' \leq \beta_t \Sigma_{tt}^{-1} \beta_t' \quad \text{and} \quad \operatorname{tr} \beta_J \Sigma_{JJ}^{-1} \beta_J' < \operatorname{tr} \beta_t \Sigma_{tt}^{-1} \beta_t', \quad \text{if } J \text{ does not include } J_t, \tag{2.3}$$

where β_J: $q \times \#J$ and Σ_{JJ}: $\#J \times \#J$ are respectively submatrices of β: $q \times p$ and Σ: $p \times p$ corresponding to the subset J of indices, β_t: $q \times p_t$ and Σ_{tt}: $p_t \times p_t$ are similarly defined corresponding to J_t, and $\#J$ denotes the number of indices of J (see [6], [9] for the definition of the model).

When all parameters are unknown, but N independent observations $\boldsymbol{y}_i$ at $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$ with the relationship (2.1) are available, we use the estimates of $\boldsymbol{\alpha}$, β and $N\Sigma$ as

$$\boldsymbol{a} = \bar{\boldsymbol{y}} - B' \bar{\boldsymbol{x}}\text{: } q \times 1,\ B = S_{xx}^{-1} S_{xy}\text{: } q \times p \text{ and } S = S_{yy} - B' S_{xx} B\text{: } p \times p, \tag{2.4}$$

where

$$(2.5)\quad \begin{pmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{y}} \end{pmatrix} = N^{-1}\sum_{i=1}^{N}\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix},\quad \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} = \sum_{i=1}^{N}\begin{pmatrix} \boldsymbol{x}_i-\bar{\boldsymbol{x}} \\ \boldsymbol{y}_i-\bar{\boldsymbol{y}} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_i-\bar{\boldsymbol{x}} \\ \boldsymbol{y}_i-\bar{\boldsymbol{y}} \end{pmatrix}'.$$

Note that S and $B'S_{xx}B$ follow the Wishart distribution $\mathrm{W}_p(N-q-1, \Sigma)$ and the noncentral Wishart distribution $\mathrm{W}_p(q, \Sigma; \beta'S_{xx}\beta)$ respectively. The likelihood ratio for the model J against the full model $J_f \equiv \{1,\ldots, p\}$, $\Lambda(J_f; J)$, is expressed by [7]. Instead of minimizing GIC of (1.1), we minimize the difference:

$$(2.6)\quad G_N(J) \equiv \mathrm{GIC}(J) - \mathrm{GIC}(J_f) = \Lambda(J_f; J) - q(p-\sharp J)c_N,\quad G_N(J_f) = 0,$$

where $\{c_N\}$ satisfies (1.2) and

$$(2.7)\qquad \Lambda(J_f; J) = N\log\{|S_{JJ}|\,|S+B'S_{xx}B|/|S|\,|S_{JJ}+B_J'S_{xx}B_J|\},$$

S_{JJ} and B_J are submatrices of S and B corresponding to J. Define the selected model $\hat{J}_N$ based on N calibration samples as

$$(2.8)\qquad \hat{J}_N \quad \text{minimizes} \quad G_N(J) \quad \text{of} \quad (2.6)$$

among the models under consideration.

Recall the criterion function (2.6) based on the log-likelihood ratio (2.7) is derived when $\boldsymbol{y}_i$ are normally distributed. However, we apply this procedure when we relax the assumption of normality. Nishii [10] studied the asymptotic behavior of the AIC for the case $c_N \equiv 2$ in (2.6) under a weak assumption and he showed that the AIC is not consistent in the multivariate calibration problem. We will show that the ED criterion defined in (1.1) with (1.2) is strongly consistent under the following mild conditions:

ASSUMPTION. (i) *The error vectors* $\boldsymbol{e}_i$ $(i=1,\ldots, N,\ldots)$ *are independently and identically distributed (i.i.d.) with*

$$(2.9)\qquad \mathrm{E}\boldsymbol{e}_1 = \boldsymbol{0},\ \mathrm{E}\boldsymbol{e}_1\boldsymbol{e}_1' = \Sigma \text{ and } \mathrm{E}(\boldsymbol{e}_1'\boldsymbol{e}_1)^{\gamma/2} < \infty \text{ for some } \gamma \in [2, 3].$$

(ii) *The sequence of the vectors of explanatory variables* $\{\boldsymbol{x}_i=(x_{i1},\ldots, x_{iq})' \mid i=1,\ldots, N,\ldots\}$ *satisfies*

$$(2.10)\qquad 0 < m\mathrm{I}_q \le N^{-1}S_{xx} = N^{-1}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_N)(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_N)' \le M\mathrm{I}_q,$$

$$(2.11)\qquad \sum_{i=1}^{N}|x_{ik}-\bar{x}_{Nk}|^3 \le \begin{cases} \Gamma N^{\gamma/2}(\log\log N)^{3/2}, & \text{if } 2 \le \gamma < 3, \\ \Gamma N^{3/2}/\log N, & \text{if } \gamma = 3, \end{cases}$$

$$(1 \le k \le q)$$

where $\bar{\boldsymbol{x}}_N = N^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i = (\bar{x}_{N1},\ldots, \bar{x}_{Nq})'$, $m < M$ *and* Γ *are positive constants.*

The proof of the following lemma is given in Section 5.

LEMMA 2.1. *Under Assumption, it holds that*

$$T_N = \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)\boldsymbol{e}_i' : q \times p = O(\sqrt{(N \log\log N)}), \quad a.s. \tag{2.12}$$

THEOREM 2.1. *Under Assumption, the model selection procedure based on the* ED *criterion is strongly consistent to the true model in the multivariate calibration problem, i.e.,* $\lim_{N\to\infty} \hat{J}_N = J_t$, *a.s.*

PROOF. From (ii) of Assumption, $S_{xx} = O(N)$. Using Lemma 2.1 we have

$$\begin{aligned} N^{-1}B'S_{xx}B &= N^{-1}\beta' S_{xx}\beta + T_N'\beta + \beta' T_N + T_N' S_{xx}^{-1} T_N \\ &= N^{-1}\beta' S_{xx}\beta + O(\sqrt{(N^{-1}\log\log N)}), \quad \text{a.s.}, \end{aligned} \tag{2.13}$$

$$\begin{aligned} N^{-1}S &= N^{-1}(S_{yy} - B'S_{xx}B) \\ &= N^{-1}\sum_{i=1}^{N} (\boldsymbol{e}_i - \bar{\boldsymbol{e}}_N)(\boldsymbol{e}_i - \bar{\boldsymbol{e}}_N)' - N^{-1} T_N S_{xx}^{-1} T_N \\ &= \Sigma + O(\sqrt{(N^{-1}\log\log N)}), \quad \text{a.s.} \end{aligned} \tag{2.14}$$

where $T_N : q \times p$ is defined in (2.12) and $\bar{\boldsymbol{e}}_N = N^{-1}\sum_{i=1}^{N} \boldsymbol{e}_i$. If J does not include the unknown true model J_t, by (2.6–7) and (2.13–14) we have almost surely

$$G_N(J) = \operatorname{tr}\{(\beta\Sigma^{-1}\beta' - \beta_J \Sigma_{JJ}^{-1}\beta_J')S_{xx}\} - q(p - \sharp J)c_N + O(\sqrt{\log\log N}). \tag{2.15}$$

Case 1: J does not include the unknown true model J_t. The first term of the right hand side of (2.15) is positive by (2.3) and it increases with the order N by (2.10), which together with $\lim_{N\to\infty} N^{-1}c_N = 0$ implies

$$G_N(J) > 0 \text{ for large } N, \text{ a.s. if } J \text{ does not include } J_t. \tag{2.16}$$

On the other hand $G_N(J_f) \equiv 0$ by the definition (2.6) of G_N, and we want to minimize $G_N(J)$ for given N. This yields that the ED criterion asymptotically prefers the full model J_f to J if J does not include the true model J_t. If the full model J_t coincides with the full model J_f, the consistency is established.

Case 2: J properly includes J_t. First we examine which model will be chosen among two models J_t and J_f. Define $S = \begin{pmatrix} S_{tt} & S_{t1} \\ S_{1t} & S_{11} \end{pmatrix} : p \times p$, $S_{tt} : p_t \times p_t$, $B = [B_t, B_1] : q \times p$, $B_t : q \times p_t$. Let $S_{11\cdot t} = S_{11} - S_{1t}S_{tt}^{-1}S_{t1}$, and define $(S + B'S_{xx}B)_{11\cdot t}$ and $\Sigma_{11\cdot t}$ in the similar way. Put $U = S_{xx}^{1/2}B = [U_t, U_1] : q \times p$ and $U_t : q \times p_t$. From [6] we know that

$$\begin{aligned} (S + B'S_{xx}B)_{11\cdot t} - S_{11\cdot t} &= (S + U'U)_{11\cdot t} - S_{11\cdot t} \\ &= (U_1 - U_t S_{tt}^{-1}S_{t1})'(\mathrm{I}_q + U_t S_{tt}^{-1}U_t')^{-1}(U_1 - U_t S_{tt}^{-1}S_{t1}). \end{aligned}$$

By the law of iterated logarithm and Lemma 2.1, we have

$$N^{-1}S_{11\cdot t} = \Sigma_{11\cdot t} + O(\sqrt{(N^{-1}\log\log N)}), \quad \text{a.s.},$$
$$U_t S_{tt}^{-1} U_t' = N^{-1} S_{xx}^{1/2} \beta_t \Sigma_{tt}^{-1} \beta_t' S_{xx}^{1/2} + O(\sqrt{(N^{-1}\log\log N)})$$
$$= O(1), \quad \text{a.s.}$$
$$U_1 - U_t S_{tt}^{-1} S_{t1} = S_{xx}^{1/2}\beta_1 - S_{xx}^{1/2}\beta_t \Sigma_{tt}^{-1}\Sigma_{t1} + O(\sqrt{\log\log N})$$
$$= O(\sqrt{\log\log N}), \quad \text{a.s.}$$

The last equality follows from the relation $\beta_1 = \beta_t \Sigma_{tt}^{-1}\Sigma_{t1}$ which is obtained by (2.3). Hence

$$\begin{aligned} G_N(J_t) &= \Lambda(J_f, J_t) - q(p-p_t)c_N \\ &= N\log\{|(S+U'U)_{11\cdot t}|/|S_{11\cdot t}|\} - q(p-p_t)c_N \\ &= N\log|I_{p-q} + S_{11\cdot t}^{-1}\{(S+U'U)_{11\cdot t} - S_{11\cdot t}\}| - q(p-p_t)c_N \\ &= O(\log\log N) - q(p-p_t)c_N \longrightarrow -\infty, \quad (N\to\infty), \quad \text{a.s.} \end{aligned} \tag{2.17}$$

because $p-p_t>0$ and $\lim_{N\to\infty} c_N/\log\log N = +\infty$. Thus $G_N(J_t)$ takes negative values for sufficiently large N, which implies that the ED criterion asymptotically prefers the true model J_t to the full model J_f. Second when J properly includes the true model, the smilar lines lead to that the log-likelihood ratio (2.7) is:

$$\Lambda(J_f, J) = O(\log\log N), \quad \text{a.s.}$$

Hence, by (1.2) and $\#J > p_t$,

$$\begin{aligned} G_N(J_t) - G_N(J) &= \Lambda(J_f, J_t) - \Lambda(J_f, J) + q(p_t - \#J)c_N \\ &= O(\log\log N) - q(\#J - p_t)c_N \longrightarrow -\infty, \quad \text{a.s.} \end{aligned}$$

This implies that the ED criterion asymptotically prefers J_t to J, completing the proof.

However, we must calculate 2^p-1 $G_N(\cdot)$'s to obtain $\hat{J}_N$ of (2.8). When p is large, this would involve extensive computation. To overcome this problem, we propose an alternate procedure, which is also based on the ED criterion and which is essentially due to [14]. Let J_{-i} be the subset of the full model omitting the index i $(1\le i\le p)$. Choose the model:

$$\tilde{J}_N = \{i\in J_f \mid G_N(J_{-i}) > 0 = G_N(J_f)\}. \tag{2.18}$$

This subset is obtained by calculating only p $G_N(\cdot)$'s, but this is still a strongly consistent estimate of the true model J_t.

THEOREM 2.2. *Under Assumption, the set of the indices selected as* (2.18)

is also a strongly consistent estimate of that of the unknown true model, i.e., $\lim_{N\to\infty} \tilde{J}_N = J_t$, *a.s.*

PROOF. If i lies in the true model J_t, then J_{-i} does not include J_t. By (2.16), $G_N(J_{-i})>0$ or $i\in\tilde{J}_N$ for large N, a.s. If j does not lie in J_t, then J_{-j} includes J_t. By the similar discussion as (2.17), we know that $G_N(J_{-j})<0$ or $\tilde{J}_N$ will not contain j for large N, a.s., and this completes the proof.

3. Discriminant analysis

The discussion on multivariate calibration can be applied to the variable selection in multiple discriminant analysis. Consider $q+1$ p-variate normal populations π_α with mean vector $\boldsymbol{\mu}_\alpha$ and common covariance matrix Σ ($\alpha=1,\ldots, q+1$). Assume N_α samples $\boldsymbol{x}_{\alpha 1},\ldots, \boldsymbol{x}_{\alpha N_\alpha}$ are drawn from π_α. We are interested in interpreting the differences among the $q+1$ populations in terms of only a few canonical discriminant variates.

Let Ω be the population between-groups covariance matrix as

$$\Omega = N^{-1}\sum_{\alpha=1}^{q+1} N_\alpha(\boldsymbol{\mu}_\alpha-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_\alpha-\bar{\boldsymbol{\mu}})' : p\times p,$$

where $\bar{\boldsymbol{\mu}}=N^{-1}\sum_{\alpha=1}^{q+1} N_\alpha\boldsymbol{\mu}_\alpha$ and $N=\sum_{\alpha=1}^{q+1} N_\alpha$. Let J be a subset of the full model $\{1,\ldots, p\}\equiv J_f$. We say that the model is J when unknown parameters satisfy

$$\text{tr}\,\Sigma^{-1}\Omega = \text{tr}\,\Sigma_{JJ}^{-1}\Omega_{JJ} > \text{tr}\,\Sigma_{J'J'}^{-1}\Omega_{J'J'} \quad \text{if } J' \text{ does not include } J, \tag{3.1}$$

where Ω_{JJ} and Σ_{JJ} are submatrices of Ω and Σ of size $\#J\times\#J$ respectively. We assume that the true but unknown model exists and denote it by $J_t=\{1,\ldots, p_t\}$. The maximum likelihood function under the model J is known (see [6]). Hence, we have

$$\begin{aligned} G_N(J) &= \text{GIC}(J) - \text{GIC}(J_f) \\ &= N\log\{|W_{JJ}|\,|W+V|/|W|\,|W_{JJ}+V_{JJ}|\} - q(p-\#J)c_N \end{aligned} \tag{3.2}$$

where

$$W = \sum_{\alpha=1}^{q+1}\sum_{i=1}^{N_\alpha}(z_{\alpha i}-\bar{z}_\alpha)(z_{\alpha i}-\bar{z}_\alpha)' : p\times p, \tag{3.3}$$

$$V = \sum_{\alpha=1}^{q+1} N_\alpha(\bar{z}_\alpha-\bar{z})(\bar{z}_\alpha-\bar{z})' : p\times p, \tag{3.4}$$

$$\bar{z}_\alpha = N_\alpha^{-1}\sum_{i=1}^{N_\alpha} z_{\alpha i},\quad \bar{z} = N^{-1}\sum_{q=1}^{q+1} N_\alpha\bar{z}_\alpha .$$

Here W and V denote the matrices of sums of squares and products due to within groups and due to between groups, respectively. Note that $W\sim W_p(N-q-1, \Sigma)$ and $V\sim W_p(q, \Sigma; N\Omega)$, and recall that $S\sim W_p(N-q-1, \Sigma)$ and $B'S_{xx}B\sim W_p(q, \Sigma; \beta'S_{xx}\beta)$ in (2.5). Let $\{S_{xx}=S_{xx}^{(N)}\}$ be a sequence of matrices satisfying Assumption

with $\gamma=2$. Then we can find $\beta=\beta_N$: $q\times p$ such that $\beta' S_{xx}\beta=N\Omega$ since rank $\Omega\leq \min(p, q)$. Put $S=W$ and $B'S_{xx}B=U$ in (2.5). This gives the complete correspondence between (2.5) and (3.2) except that β depends on N.

Let $\tilde{J}_N$ be a subset of J_f minimizing (3.2) and let $\hat{J}_N$ be a subset of J_f defined by (2.18) in this situation.

THEOREM 3.1. *Let* $z_{\alpha i}-\mu_\alpha$ $(i=1,\ldots, N_\alpha;\ \alpha=1,\ldots, q+1)$ *be i.i.d. with* $\mathrm{E}z_{\alpha i}=\mu_\alpha$ *and* $\mathrm{E}(z_{\alpha i}-\mu_\alpha)(z_{\alpha i}-\mu_\alpha)'=\Sigma$. *Assume that the data increases with satisfying the condition*

$$0<m'<N^{-1}N_\alpha<1 \quad (\alpha=1,\ldots, q+1), \quad N=\sum_{\alpha=1}^{q+1}N_\alpha$$

where m' *is a positive constant. Then both* $\hat{J}_N$ *and* $\tilde{J}_N$ *are strongly consistent estimators of the unknown true model* J_t.

4. Canonical correlation analysis

In this section we treat the variable selection problem in canonical correlation analysis. Let $z=(x', y')'$ follow $\mathrm{N}_{p+q}(\mu, \Sigma)$ where x: $q\times 1$, y: $p\times 1$, $\mu=(\mu_x', \mu_y')'$: $(p+q)\times 1$, μ_x: $q\times 1$, $\Sigma=\begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}$: $(p+q)\times(p+q)$ and Σ_{xx}: $q\times q$. Suppose we are interested in summarizing the relationship between x and y by using a small number of variables. Let $I_f=\{1,\ldots, q\}$ and $J_f=\{1,\ldots, p\}$ be the sets of the indices of x and y respectively. Consider subsets $I\subseteq I_f$ and $J\subseteq J_f$. We say that the model is (I, J) when we suppose that using submatrix Σ_{IJ} of Σ_{xy} and so on,

$$\text{(4.1)} \qquad \operatorname{tr}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}=\operatorname{tr}\Sigma_{JI}\Sigma_{II}^{-1}\Sigma_{IJ}\Sigma_{JJ}^{-1}.$$

Further we suppose the existence of the true but unknown model (I_t, J_t) which consists of the smallest number of parameters satisfying (4.1) when $I_t=\{1,\ldots, q_t\}$ and $J_t=\{1,\ldots, p_t\}$. Also, let (x_i', y_i') be N independent observations of z' and put

$$S=\begin{pmatrix}S_{xx} & S_{xy}\\ S_{yx} & S_{yy}\end{pmatrix}=\sum_{i=1}^{N}\begin{pmatrix}x_i-\bar{x}\\ y_i-\bar{y}\end{pmatrix}\begin{pmatrix}x_i-\bar{x}\\ y_i-\bar{y}\end{pmatrix}': (p+q)\times(p+q).$$

Consider the model (I, J) where $I=\{1,\ldots, q_1\}$ and $J=\{1,\ldots, p_1\}$. Corresponding to I and J, we partition S into 16 submatrices (S_{ij}); $i, j=1,\ldots, 4$ as $S_{xx}=\begin{pmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{pmatrix}$: $q\times q$, $S_{xy}=\begin{pmatrix}S_{13} & S_{14}\\ S_{23} & S_{24}\end{pmatrix}$: $q\times p$, $S_{yy}=\begin{pmatrix}S_{33} & S_{34}\\ S_{43} & S_{44}\end{pmatrix}$: $p\times p$, S_{11}: $q_1\times q_1$, S_{13}: $q_1\times p_1$, S_{33}: $p_1\times p_1$ and $S_{ij}=S_{ji}'$. (In the similar way, the submatrices of Σ are defined.) Then the log-likelihood ratio of the model (I, J) and the full model is given by [5] as

$$(4.2)\qquad \Lambda(I_f, J_f; I, J) = N \log \left\{ |S_{22\cdot 1}|\,|S_{44\cdot 3}| \Big/ \begin{vmatrix} S_{22\cdot 13} & S_{24\cdot 13} \\ S_{42\cdot 13} & S_{44\cdot 13} \end{vmatrix} \right\},$$

where

$$S_{ij\cdot 13} = S_{ij\cdot 1} - S_{i3\cdot 1}S_{33\cdot 1}^{-1}S_{3j\cdot 1} = S_{ij\cdot 3} - S_{i1\cdot 3}S_{11\cdot 3}^{-1}S_{1j\cdot 3},$$
$$S_{ij\cdot k} = S_{ij} - S_{ik}S_{kk}^{-1}S_{kj}.$$

Define $\Sigma_{ij\cdot 13}$ and $\Sigma_{ij\cdot k}$ in the similar way.

If $I \supseteq I_t$ and $J \supseteq J_t$ or $q_1 \geq q_t$ and $p_1 \geq p_t$ then (4.1) is true, which yields $(\Sigma_{41\cdot 3}, \Sigma_{42\cdot 3}) = 0$ and $(\Sigma_{23\cdot 1}, \Sigma_{24\cdot 1}) = 0$. Hence, by the law of iterated logarithm, using $\ell_N = \sqrt{(N^{-1} \log\log N)}$,

$$N^{-1}S_{22\cdot 1} = \Sigma_{22\cdot 1} + O(\ell_N), \quad \text{a.s.}, \quad N^{-1}S_{44\cdot 3} = \Sigma_{44\cdot 3} + O(\ell_N), \quad \text{a.s.},$$

$$N^{-1}\begin{pmatrix} S_{22\cdot 13} & S_{24\cdot 13} \\ S_{42\cdot 13} & S_{44\cdot 13} \end{pmatrix} = \begin{pmatrix} \Sigma_{22\cdot 13} & \Sigma_{24\cdot 13} \\ \Sigma_{42\cdot 13} & \Sigma_{44\cdot 13} \end{pmatrix} + O(\ell_N)$$
$$= \begin{pmatrix} \Sigma_{22\cdot 1} & 0 \\ 0 & \Sigma_{44\cdot 3} \end{pmatrix} + O(\ell_N), \quad \text{a.s.},$$

and

$$(4.3)\quad \Lambda(I_f, J_f; I, J) = N \log \left\{ |\Sigma_{22\cdot 1}|\,|\Sigma_{44\cdot 3}| \Big/ \begin{vmatrix} \Sigma_{22\cdot 1} & 0 \\ 0 & \Sigma_{44\cdot 3} \end{vmatrix} + O(\ell_N^2) \right\}$$
$$= O(\log\log N), \quad \text{a.s.}$$

If $q_1 < q_t$ or $p_1 < p_t$ (which implies $I \not\supseteq I_t$ or $J \not\supseteq J_t$), then $(\Sigma_{23\cdot 1}, \Sigma_{24\cdot 1}) \neq 0$ or $(\Sigma_{41\cdot 3}, \Sigma_{42\cdot 3}) \neq 0$. Hence, $|\Sigma_{22\cdot 1}||\Sigma_{44\cdot 3}| > |\Sigma_{22\cdot 31}||\Sigma_{44\cdot 13}|$. Therefore,

$$(4.4)\quad \Lambda(I_f, J_f; I, J) = N \log\{|\Sigma_{22\cdot 1}||\Sigma_{44\cdot 3}|/|\Sigma_{22\cdot 13}||\Sigma_{44\cdot 13}|\} + O(\log\log N)$$
$$= O(N) \quad \text{and} \quad \longrightarrow +\infty, \quad (N \to \infty), \quad \text{a.s.}$$

This discussion is applicable in the general case of $I \not\supseteq I_t$ or $J \not\supseteq J_t$. In this case let $I_f^* = I \cup I_t$ and $J_f^* = J \cup J_t$. When we restrict the variables of $\boldsymbol{x}$ and $\boldsymbol{y}$ as x_i $(i \in J_f^*)$, the true model remains (I_t, J_t). Recalling the definition (4.2) and using (4.4), we get

$$\Lambda(I_f, J_f; I_f^*, J_f^*) = O(\log\log N), \quad \text{a.s.},$$
$$\Lambda(I_f^*, J_f^*; I, J) = O(N) \quad \text{and} \quad \Lambda(I_f^*, J_f^*; I, J) \longrightarrow +\infty, \ (N \to \infty), \ \text{a.s.}$$

Hence, if $I \not\supseteq I_t$ or $J \not\supseteq J_t$,

(4.5) $\Lambda(I_f, J_f; I, J) = \Lambda(I_f, J_f; I_f^*, J_f^*) + \Lambda(I_f^*, J_f^*; I, J) \longrightarrow +\infty$, a.s.

We hold the relations (4.3) and (4.5) under the assumption that $z'=(x', y')$ has the finite variance-covariance matrix.

Now we define $(\hat{I}_N, \hat{J}_N)$, the selected model, which minimizes

$$G_N(I, J) = \Lambda(I_f, J_f; I, J) - (pq - \#I\#J)c_N,$$

and we propose an another procedure to select the model $(\tilde{I}_N, \tilde{J}_N)$:

$$\tilde{I}_N = \{i \in I_f \mid G_N(I_{-i}, J_f) > 0\}, \quad \tilde{J}_N = \{j \in J_f \mid G_N(I_f, J_{-j}) > 0\}$$

where $I_{-i}=I_f-\{i\}$ and $J_{-j}=J_f-\{j\}$.

Combining (4.3) and (4.5), we obtain

THEOREM 4.1. *Let $\{z_i=(x_i', y_i')' \mid i=1,\ldots, N,\ldots\}$ be i.i.d. with mean vector $(\mu_x', \mu_y')'$ and variance-covariance matrix Σ. Then $(\hat{I}_N, \hat{J}_N)$ and $(\tilde{I}_N, \tilde{J}_N)$ are strongly consistent estimators of the true model (I_t, J_t).*

5. Proof of Lemma 2.1

To prove Lemma 2.1 it is sufficient to show that

(5.1) $$\sum_{i=1}^N (x_i - \bar{x}_N)e_i = O(\sqrt{(N \log\log N)}), \quad \text{a.s.}$$

if (i) random variables $e_1,\ldots, e_N,\ldots$ are i.i.d. with $Ee_1=0$, $Ee_1^2=1$, (ii) $x_1,\ldots, x_N,\ldots$ satisfy $m \le N^{-1}\sum_{i=1}^N (x_i-\bar{x}_N)^2 \le M$ for any $N \ge 2$ where m and M are positive constants and $\bar{x}_N = N^{-1}\sum_{i=1}^N x_i$, and (iii) for some $\gamma \in [2, 3]$, (2.9) and (2.11) are satisfied.

To show (5.1) we shall prove

(5.2) $$\sum_{k=1}^\infty P[\cup_k^\dagger \{(e_1,\ldots, e_N) \mid \sum_{i=1}^N (x_i-\bar{x}_N)e_i > C\sqrt{(N \log\log N)}\}] < \infty$$

for a constant C such that $C > \sqrt{M}/\sqrt{\log 2} + \sqrt{(2m)}$, where $\cup_k^\dagger$ denotes unions with repsect to N running through $2^{k-1}+1$ to 2^k, i.e.,

$$\cup_k^\dagger = \cup_{2^{k-1} < N \le 2^k}$$

and M appeared in (ii). If $2^{k-1}+1 \le N \le 2^k$,

$$|\bar{x}_k^* - \bar{x}_N| = |N^{-1}\sum_{i=1}^N (\bar{x}_k^* - x_i)| \le \{N^{-1}\sum_{i=1}^N (\bar{x}_k^* - x_i)^2\}^{1/2} < \sqrt{(2M)},$$

where $\bar{x}_k^* = 2^{-k}\sum_{i=1}^{2^k} x_i$. Hence $(\bar{x}_k^* - \bar{x}_N)$ is bounded and by the law of iterated logarithm, $\sum_{i=1}^N e_i = O(\sqrt{(N \log\log N)})$, a.s. Thus

$$(\bar{x}_k^* - \bar{x}_N)\sum_{i=1}^N e_i = O(\sqrt{(N \log\log N)}), \quad \text{a.s.}$$

Using the following trivial relation:

$$\sum_{i=1}^{N}(x_i-\bar{x}_N)e_i = \sum_{i=1}^{N}(x_i-\bar{x}_k^*)e_i + (\bar{x}_k^*-\bar{x}_N)\sum_{i=1}^{N}e_i,$$

it is necessary to show

(5.3) $\sum_{i=1}^{N}(x_i-\bar{x}_k^*)e_i = O(\sqrt{(N\log\log N)})$ or $\sum_{k=1}^{\infty}P(E_k) < \infty$,

where E_k is the event $\cup_k^{\dagger}\{\sum_{i=1}^{N}(x_i-\bar{x}_k^*)e_i > C2^{k/2}\sqrt{\log k}\}$.
Define

$$e'_{ik} = \begin{cases} e_i, & \text{if } |e_i| < 2^{k/2}, \\ 0, & \text{otherwise,} \end{cases} \qquad E''_k = \cup_{i=1}^{2^k}\{e_i \neq e'_{ik}\},$$

and

$$E'_k = \cup_k^{\dagger}\{\sum_{i=1}^{N}(x_i-\bar{x}_k^*)e'_{ik} \geq C2^{k/2}\sqrt{\log k}\}.$$

Then the event E_k is a subset of the event $E'_k \cup E''_k$, which implies

(5.4) $$P(E_k) \leq P(E'_k) + P(E''_k).$$

So,

(5.5) $$\begin{aligned} \sum_{k=1}^{\infty}P(E''_k) &= \sum_{k=1}^{\infty}2^kP(e_1 \neq e'_{1k}) = \sum_{k=1}^{\infty}2^kP(|e_1| \geq 2^{k/2}) \\ &= \sum_{k=1}^{\infty}2^k\sum_{\ell=k}^{\infty}P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}] \\ &= \sum_{\ell=1}^{\infty}P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}]\sum_{k=1}^{\ell}2^k \\ &\leq \sum_{\ell=1}^{\infty}2^{\ell+1}P[2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}] \leq 2\sum_{\ell=1}^{\infty}E(e_1^2\chi_\ell) \leq 2Ee_1^2 = 2, \end{aligned}$$

where χ_ℓ denotes the indicator function of the event $\{2^{\ell/2} \leq |e_1| < 2^{(\ell+1)/2}\}$. Using the assumption $Ee_1 = 0$, we have

$$|Ee'_{1k}| = |E(e'_{1k}-e_1)| \leq E(|e_1|\tilde{\chi}_k) \leq 2^{-k/2}Ee_1^2 = 2^{-k/2},$$

where $\tilde{\chi}_k$ denotes the indicator function of the event $\{|e_1| \geq 2^{k/2}\}$. This relation yields

$$\sum_{i=1}^{N}|x_i - \bar{x}_k^*|\,|Ee'_{ik}| \leq \{N\sum_{i=1}^{N}(x_i-\bar{x}_k^*)^2\}^{1/2}2^{-k/2} \leq 2^{k/2}\sqrt{M},$$

for $2^{(k-1)/2} < N \leq 2^{k/2}$. Putting

$$e_{ik} = e'_{ik} - Ee'_{ik} \quad \text{and} \quad t_N = \sum_{i=1}^{N}(x_i-\bar{x}_k^*)e_{ik}$$

we obtain that for $k \geq 2$

$$\begin{aligned} P(E'_k) &\leq P[\cup_k^{\dagger}\{t_N \geq C2^{k/2}\sqrt{\log k} - \sum_{i=1}^{N}|x_i-\bar{x}_k^*|\,|Ee'_{ik}|\}] \\ &\leq P[\cup_k^{\dagger}\{t_N \geq C2^{k/2}\sqrt{\log k} - 2^{k/2}\sqrt{M}\}]. \end{aligned}$$

If we take C' such that $C-\sqrt{M/\log 2}>C'>\sqrt{(2m)}$, the last formula is domainted by

$$\mathrm{P}[\cup_k^{\dagger}\{t_N \geq C'2^{k/2}\sqrt{\log k}\}] \leq \mathrm{P}[\textstyle\sum_{i=1}^{2^k}(x_i-\bar{x}_k^*)e_{ik} \geq C'2^{k/2}\sqrt{\log k}].$$

Further the last formula is evaluated by the inequality due to [3] using the relation $m2^k \leq \sum_{i=1}^{2^k}(x_i-\bar{x}_k^*)^2 \leq M2^k$, we have

$$\mathrm{P}(E_k') \leq 2\{1-\Phi(C''\sqrt{\log k})\} + C_0 R_k \tag{5.6}$$

where $\Phi(x)$ is the standard normal distribution function, C_0 is a constant, $C''=C'/\sqrt{m}>\sqrt{2}$, $C'''=C'/\sqrt{M}>0$ and

$$R_k = \textstyle\sum_{i=1}^{2^k}|x_i-\bar{x}_k^*|^3\mathrm{E}|e_{1k}^3|2^{-3k/2}(1+C'''\sqrt{\log k})^{-3}.$$

Now employing the inequality $1-\Phi(x)\leq(2\pi)^{-1/2}\exp(-x^2/2)$ $(x\geq 1)$, we get

$$\textstyle\sum_{k=3}^{\infty}\{1-\Phi(C''\sqrt{\log k})\} \leq (2\pi)^{-1/2}\sum_{k=3}^{\infty}k^{-C''^2/2} < \infty \tag{5.7}$$

because $C''>\sqrt{2}$. On the other hand, if $\gamma=3$, i.e., $\mathrm{E}|e_1^3|<\infty$, then

$$R_k \leq \Gamma\mathrm{E}|e_1^3|/[(\log 2)k\{1+C'''\sqrt{\log k}\}^3], \tag{5.8}$$

$$\textstyle\sum_{k=2}^{\infty}R_k \leq C_1\sum_{k=2}^{\infty}1/(k\log k) < \infty,$$

where $C_1>0$ is a constant. If $2\leq\gamma<3$, i.e., $\mathrm{E}|e_1|^{\gamma}<\infty$, then

$$\begin{aligned} R_k &\leq \Gamma 2^{(\gamma-3)k/2}\mathrm{E}|e_{1k}^3|(\log k+\log\log 2)^{3/2}/(1+C'''\sqrt{\log k})^3 \\ &\leq C_2 2^{-(3-\gamma)k/2}\mathrm{E}|e_{1k}^3|, \end{aligned}$$

$$\begin{aligned} \textstyle\sum_{k=1}^{\infty}R_k &\leq C_2\textstyle\sum_{k=1}^{\infty}2^{-(3-\gamma)k/2}\mathrm{E}|e_{1k}^3| \\ &\leq C_2\textstyle\sum_{k=1}^{\infty}2^{-(3-\gamma)k/2}\{\sum_{\ell=1}^{k}\mathrm{E}(|e_1^3|\chi_{\ell-1})+1\} \\ &\leq C_3\textstyle\sum_{\ell=1}^{\infty}2^{-(3-\gamma)\ell/2}\{\mathrm{E}(|e_1^3|\chi_{\ell-1})+1\} \\ &\leq C_3\textstyle\sum_{\ell=1}^{\infty}\mathrm{E}(|e_1|^{\gamma}\chi_{\ell-1}) + C_4 \leq C_3\mathrm{E}|e_1|^{\gamma}+C_4 < \infty, \end{aligned} \tag{5.9}$$

where the indicator function $\chi_{\ell-1}$ is used in (5.5) and C_2, C_3, C_4 are positive constants and Γ is used in (2.11). Thus (5.7–9) yield that $\sum_{k=1}^{\infty}\mathrm{P}(E_k')<\infty$. This and (5.4–5) yield (5.2). Hence the proof is completed.

Acknowledgment

Part of this work was completed while Nishii visited the Center for Multivariate Analysis, University of Pittsburgh. Bai and Krishnaiah were supported by the U.S. Air Force Office of Scientific Research under Contract F49620–85–C–0008.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, 2nd International Symposium on Information Theory (1973, B. N. Petrov and F. Czáki, Eds.), Akadémia Kiado, Budapest, 267–281.

[2] H. Akaike, A Bayesian analysis of the minimum AIC procedure, Ann. Inst. Statist. Math., **30** (1978), 9–14.

[3] A. Bikjalis, Estimates of the remainder term in the central limit theorem, Litovsk. Mat. Sb., **6** (1966), 323–346, (in Russian).

[4] P. J. Brown, Multivariate calibration, J. R. Statist. Soc. B, **44** (1982), 287–321.

[5] Y. Fujikoshi, A test for additional information in canonical correlation analysis, Ann. Inst. Statist. Math., **34** (1982), 523–530.

[6] Y. Fujikoshi, A criterion for variable selection in multiple discriminant analysis, Hiroshima Math. J., **13** (1983), 203–214.

[7] Y. Fujikoshi and R. Nishii, Selection for variables in a multivariate inverse regression problem, Hiroshima Math. J., **16** (1986), 269–277.

[8] E. J. Hannan and B. G. Quinn, The determination of the order of an autoregression, J. Roy. Statist. Soc., B, **41** (1979), 190–195.

[9] R. J. McKay, Simultaneous procedures for variable selection in multiple discriminant analysis, Biometrika, **64** (1977), 283–290.

[10] R. Nishii, Criteria for selection of response variables and the asymptotic properties in a multivariate calibration, Ann. Inst. Statist. Math., **38** (1986), 319–329.

[11] C. R. Rao, Linear statistical inference and its applications, Wiley, New York, 1973.

[12] J. Rissanen, Modeling by shortest data description, Automatica, **14** (1978), 465–471.

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

[14] L. C. Zhao, P. R. Krishnaiah and Z. D. Bai, On detection of the number of signals in presence of white noise, J. Multivariate Anal., **20** (1986), 1–25.

†*Division of Information and Behavioral Sciences,*
Faculty of Integrated Arts and Sciences,
Hiroshima University
††*Center for Multivariate Analysis,*
University of Pittsburgh