Variable Selections and Their Applications for Data Analysis in Principal Canonical Correlation Analysis

Ogura, Toru
Chuo University, Industrial and Systems Engineering
1-13-27, Kasuga,
Bunkyo-ku, Tokyo, 112-8551, Japan
E-mail: ogura@indsys.chuo-u.ac.jp

Fujikoshi, Yasunori
Hiroshima University
1-3-1, Kagamiyama,
HigashiHiroshima-shi, Hiroshima, 739-8526, Japan
E-mail: fujikoshi_y@yahoo.co.jp

Sugiyama, Takakazu
Soka University
1-236, Tangimachi,
Hachiouji-shi, Tokyo, 192-8577, Japan
E-mail: stakakaz@gmail.com

1. Introduction

Canonical correlation analysis (CCA) is often used to analyze the correlation between two random vectors (for examples, see Anderson (2003), Siotani et al. (1985), Sugiura (1976)). However, sometimes interpretation of CCA results may be hard. As an attempt of addressing these difficulties, Sugiyama and Takeda (1999) proposed principal canonical correlation analysis (PCCA). PCCA is CCA between two sets of principal component (PC) scores. That is, each set of PC scores (components) is calculated from each random vector by principal component analysis (PCA). PCCA uses each set of PC scores instead of the original random vectors. PCA transforms a given data set of correlated variables into a new data set of uncorrelated variables, or PC scores. Each PC score is defined from the original variable set and retains a certain percentage of the inherent variability. Each PC score accounts for a decreasing proportion of the total variance inherent in the data. Therefore, it is assumed that PCCA has some merit.

Because PC scores descend in order of the amount of information that they contain, it is important to select useful PC scores in PCCA. By using only selected PC scores, it will be easier to interpret the CCA. Some procedures for selecting variables in CCA of random vectors have been proposed. Several authors (Fujikoshi (1985), Ichikawa and Konishi (1999), Fujikoshi and Kurata (2008), Konishi and Kitagawa (2008)) have proposed the use of a method based on Akaike's (1973) idea. Ogura (2010) formally investigated the same criterion in an application of PCCA, proposing a variable selection criterion for one set of PC scores in PCCA, and proposed some advantages of using this procedure. For example, the principal canonical correlation coefficients from selected PC scores provide almost the same information about the principal canonical correlation coefficients as do those from all PC scores. Furthermore, it is easier to interpret the canonical variables. The effectiveness of this procedure was demonstrated using an example.

In this paper we propose a variable selection criterion for two sets of PC scores in PCCA that is an extension of Ogura's (2010) approach, based on a reasonable derivation. Furthermore, we demonstrate the effectiveness of this criterion using a simulation and an example. We also compare variable selection for two sets of PC scores in PCCA with variable selection for one set of PC scores.

2. A General Criterion for the Selection of Covariance Structure

We use a general approach for selecting the best model from a set of covariance structure models. For the selection of variables in CCA, Fujikoshi (1985) and Fujikoshi and Kurata (2008) proposed a criterion based on Akaike's idea (Akaike (1973)). We summarize the original idea and the resulting criterion, as presented in their paper. Let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')' = (x_1, \ldots, x_p, y_1, \ldots, y_q)'$ be a $(p+q)$-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose that a sample $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)'$ of size $N = n+1$ is available. Let $\boldsymbol{S}$ be the usual unbiased estimator of $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Omega}$ be a subset of the set of symmetric positive definite matrices. Suppose that a covariance structure model M is defined by

$$M : \boldsymbol{\Sigma} \in \boldsymbol{\Omega}.$$

As a measure of the goodness-of-fit of $\boldsymbol{\Sigma}$ for a given sample covariance matrix $\boldsymbol{S}$, we use

$$D(\boldsymbol{S}, \boldsymbol{\Sigma}) = -n \log |\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| + n\{\mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{S} - (p+q)\}.$$

Note that, apart from a constant term, this measure is equal to "-2 log likelihood" of $\boldsymbol{\Sigma}$ based on $\boldsymbol{S}$ under the assumption of normality. More precisely, the measure may be regarded as essentially "-2 log likelihood" of $\boldsymbol{\Sigma}$ based on the sample $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)'$, after being maximized with respect to $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\Sigma}}$ be the minimum distance estimator under M, defined by

$$\min_M D(\boldsymbol{S}, \boldsymbol{\Sigma}) = D(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}).$$

Following Akaike's idea, we define the risk function of M as

$$R = \mathrm{E}_F \mathrm{E}_Z [D(\boldsymbol{S}_F, \hat{\boldsymbol{\Sigma}})],$$

where $\boldsymbol{S}_F$ is the sample covariance matrix for a future sample $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N)'$ that has the same distribution as $\boldsymbol{Z}$, and is independent of $\boldsymbol{Z}$. The notations E_F and E_Z denote the expectations with respect to the true model of $\boldsymbol{F}$ and $\boldsymbol{Z}$, respectively. A general model selection criterion is suggested by considering an estimator for R. We can write

$$R = \mathrm{E}_F \mathrm{E}_Z [D(\boldsymbol{S}_F, \hat{\boldsymbol{\Sigma}})] + B,$$

where

$$B = \mathrm{E}_F \mathrm{E}_Z [D(\boldsymbol{S}_F, \hat{\boldsymbol{\Sigma}}) - D(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}})].$$

The quantity B is the bias term when we estimate R by $D(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}})$. Following Akaike, we have a distance information criterion (DIC, see, e.g., Fujikoshi and Kurata (2008)), defined by

$$DIC = -n \log |\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}| + n\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S} - (p+q)\} + B_0, \quad (1)$$

where B_0 is the number of independent parameters under M. The first term is a likelihood ratio statistic for M. The second term becomes zero for the covariance structure model M considered in this paper. When $\boldsymbol{z}$ is normal, DIC is essentially the same as AIC, apart from the constant term. However, DIC can be used even when the distribution of $\boldsymbol{z}$ is non-normal. In general, it has been pointed out that AIC (and hence DIC) underestimates its risk function for overspecified models. To overcome this weakness, attempts have been made to evaluate B and obtain an estimator which is less biased. The bias term can be expressed as

$$\begin{aligned} B &= \mathrm{E}_F\mathrm{E}_Z[-n\log|\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_F| + n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_F - n(p+q) \\ &\quad -\{-n\log|\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}| + n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S} - n(p+q)\}] \\ &= \mathrm{E}_F\mathrm{E}_Z[-n\log|\boldsymbol{S}_F| + n\log|\boldsymbol{S}| + n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_F - n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}] \\ &= \mathrm{E}_Z[n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})] \\ &= \mathrm{E}_Z[n\mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}] - n(p+q), \ (\text{if } \mathrm{tr}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S} = p+q). \end{aligned} \tag{2}$$

3. The Variable Selection Model

We first summarize the notation. Let a random vector $\boldsymbol{z}$ of $(p+q)$ components with an unknown covariance matrix $\boldsymbol{\Psi}$, which is assumed to be symmetric and positive definite. We partition $\boldsymbol{z}$ into two subvectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, of p and q components as

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}.$$

Similarly, the covariance matrix of $\boldsymbol{z}$ is partitioned into p and q,

$$\mathrm{Cov}\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{xx} & \boldsymbol{\Psi}_{xy} \\ \boldsymbol{\Psi}_{yx} & \boldsymbol{\Psi}_{yy} \end{pmatrix}, \tag{3}$$

where $\boldsymbol{\Psi}_{xx}$ is $p \times p$, $\boldsymbol{\Psi}_{xy}$ is $p \times q$, $\boldsymbol{\Psi}_{yx}$ is $q \times p$, and $\boldsymbol{\Psi}_{yy}$ is $q \times q$. Let $\lambda_{1x} \geq \ldots \geq \lambda_{px}$ be the ordered latent roots of $\boldsymbol{\Psi}_{xx}$, and let $\gamma_{1x}, \ldots, \gamma_{px}$ be the corresponding latent vectors with $\gamma'_{ix}\gamma_{jx} = \delta_{ij}$, where δ_{ij} is the Kronecker delta, i.e., $\delta_{ii} = 1$, and $\delta_{ij} = 0$ for $i \neq j$. Similarly, let $\lambda_{1y} \geq \ldots \geq \lambda_{qy}$ be the ordered latent roots of $\boldsymbol{\Psi}_{yy}$ and $\gamma_{1y}, \ldots, \gamma_{qy}$ the corresponding latent vectors with $\gamma'_{iy}\gamma_{jy} = \delta_{ij}$. We can decompose $\boldsymbol{\Psi}_{xx}$ and $\boldsymbol{\Psi}_{yy}$ as:

$$\boldsymbol{\Gamma}'_x\boldsymbol{\Psi}_{xx}\boldsymbol{\Gamma}_x = \boldsymbol{\Sigma}_{uu} = \boldsymbol{\Lambda}_u, \quad \boldsymbol{\Gamma}'_y\boldsymbol{\Psi}_{yy}\boldsymbol{\Gamma}_y = \boldsymbol{\Sigma}_{vv} = \boldsymbol{\Lambda}_v,$$

where $\boldsymbol{\Lambda}_u = \mathrm{diag}(\lambda_{1x}, \ldots, \lambda_{px})$ and $\boldsymbol{\Lambda}_v = \mathrm{diag}(\lambda_{1y}, \ldots, \lambda_{qy})$ are diagonal matrices, $\boldsymbol{\Gamma}_x = (\gamma_{1x}, \ldots, \gamma_{px})$ and $\boldsymbol{\Gamma}_y = (\gamma_{1y}, \ldots, \gamma_{qy})$ are orthogonal matrices. PC scores of $\boldsymbol{x}$ and $\boldsymbol{y}$ are then defined by $\boldsymbol{u} = \boldsymbol{\Gamma}'_x\boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{\Gamma}'_y\boldsymbol{y}$, respectively. We denote the covariance matrix of $(\boldsymbol{u}', \ \boldsymbol{v}')'$ by

$$\mathrm{Cov}\left[\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix}\right] = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{vu} & \boldsymbol{\Sigma}_{vv} \end{pmatrix}. \tag{4}$$

Then, there is a relationship between $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ such that

$$\boldsymbol{\Gamma}'\boldsymbol{\Psi}\boldsymbol{\Gamma} = \boldsymbol{\Sigma}, \tag{5}$$

which is expressed as

$$\begin{pmatrix} \boldsymbol{\Lambda}_u & \boldsymbol{\Gamma}'_x\boldsymbol{\Psi}_{xy}\boldsymbol{\Gamma}_y \\ \boldsymbol{\Gamma}'_y\boldsymbol{\Psi}_{yx}\boldsymbol{\Gamma}_x & \boldsymbol{\Lambda}_v \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{vu} & \boldsymbol{\Sigma}_{vv} \end{pmatrix},$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_x & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_y \end{pmatrix}.$$

Now, we consider the covariance structure that corresponds to the variable selection model for two sets of PC scores. For this, we partition $\boldsymbol{u}$ into two subvectors $\boldsymbol{u}'_1 = (u_1, \ldots, u_{p_1})$ and $\boldsymbol{u}'_2 = (u_{p_1+1}, \ldots, u_p)$ of p_1 and $(p - p_1)$ components, and $\boldsymbol{v}$ into two subvectors $\boldsymbol{v}'_1 = (v_1, \ldots, v_{q_1})$ and

$\boldsymbol{v}_2' = (v_{q_1+1}, \ldots, v_q)$ of q_1 and $(q - q_1)$ components, such that $\boldsymbol{u} = (\boldsymbol{u}_1' \ \boldsymbol{u}_2')'$ and $\boldsymbol{v} = (\boldsymbol{v}_1' \ \boldsymbol{v}_2')'$. We use the following notations,

$$(6) \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{vu} & \boldsymbol{\Sigma}_{vv} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \boldsymbol{\Sigma}_{41} & \boldsymbol{\Sigma}_{42} & \boldsymbol{\Sigma}_{43} & \boldsymbol{\Sigma}_{44} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Lambda}_3 & \mathbf{0} \\ \boldsymbol{\Sigma}_{41} & \boldsymbol{\Sigma}_{42} & \mathbf{0} & \boldsymbol{\Lambda}_4 \end{pmatrix},$$

where $\boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_{1x}, \ldots, \lambda_{p_1 x})$, $\boldsymbol{\Lambda}_2 = \mathrm{diag}(\lambda_{(p_1+1)x}, \ldots, \lambda_{px})$, $\boldsymbol{\Lambda}_3 = \mathrm{diag}(\lambda_{1y}, \ldots, \lambda_{q_1 y})$, and $\boldsymbol{\Lambda}_4 = \mathrm{diag}(\lambda_{(q_1+1)y}, \ldots, \lambda_{qy})$. Note that there is a one-to-one correspondence between the parameter sets of $\boldsymbol{\Psi}$ and $\{\boldsymbol{\Sigma}, \boldsymbol{\Gamma}\}$. Because the covariance matrix of PC scores depends on $\boldsymbol{\Sigma}$, it is only natural that the variable selection model is introduced through $\boldsymbol{\Sigma}$. Let M_{p_1,q_1} be the variable selection model that represents a redundancy of $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ or a sufficiency of $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ in CCA between $(\boldsymbol{u}_1', \boldsymbol{u}_2')$ and $(\boldsymbol{v}_1', \boldsymbol{v}_2')$. For the following, we simply write M_{p_1,q_1} as M_r. The model may be defined (for an example, see Fujikoshi (1982)) as

$$(7) \qquad M_r : \mathrm{tr}\boldsymbol{\Sigma}_{uu}^{-1}\boldsymbol{\Sigma}_{uv}\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{vu} = \mathrm{tr}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{31}.$$

Note that the model M_r is related to graphical models (see Anderson (2003)). The condition (7) can be expressed as:

$$\mathrm{tr}\begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_3 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_4 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} \\ \boldsymbol{\Sigma}_{41} & \boldsymbol{\Sigma}_{42} \end{pmatrix}$$
$$= \mathrm{tr}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{31}$$
$$(8) \qquad \Leftrightarrow \quad \boldsymbol{\Sigma}_{14} = \mathbf{0}, \boldsymbol{\Sigma}_{23} = \mathbf{0}, \boldsymbol{\Sigma}_{24} = \mathbf{0}.$$

Then it is shown that the model M_r is equivalent to that $\boldsymbol{\Sigma}$ has the following structure:

$$(9) \qquad \boldsymbol{\Sigma}_r = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} & \boldsymbol{\Sigma}_{13} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Sigma}_{31} & \mathbf{0} & \boldsymbol{\Sigma}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{44} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \boldsymbol{\Sigma}_{13} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Sigma}_{31} & \mathbf{0} & \boldsymbol{\Lambda}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}_4 \end{pmatrix}.$$

4. An Information Criterion for a Variable Selection Model

Let $\boldsymbol{T}$ be the sample covariance matrix based on the sample of $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ of size $N = n + 1$, and partition $\boldsymbol{T}$ as

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{xx} & \boldsymbol{T}_{xy} \\ \boldsymbol{T}_{yx} & \boldsymbol{T}_{yy} \end{pmatrix},$$

in accordance with the partition of $\boldsymbol{z}$. Then, the latent roots $\lambda_{1x} \geq \ldots \geq \lambda_{px}$ of $\boldsymbol{\Psi}_{xx}$ and the corresponding latent vectors $\gamma_{1x}, \ldots, \gamma_{px}$ are estimated by the latent roots $l_{1x} \geq \ldots \geq l_{px}$ of $\boldsymbol{T}_{xx}$ and the corresponding latent vectors $\boldsymbol{h}_{1x}, \ldots \boldsymbol{h}_{px}$, respectively. Similarly, the latent roots $\lambda_{1y} \geq \ldots \geq \lambda_{qy}$ of $\boldsymbol{\Psi}_{yy}$ and the corresponding latent vectors $\gamma_{1y}, \ldots, \gamma_{qy}$ are estimated by the latent roots $l_{1y} \geq \ldots \geq l_{qy}$ of $\boldsymbol{T}_{yy}$ and the corresponding latent vectors $\boldsymbol{h}_{1y}, \ldots \boldsymbol{h}_{qy}$, respectively.

Consider to estimate $\boldsymbol{\Psi}$ by minimizing

$$D(\boldsymbol{T}, \boldsymbol{\Psi}) = -n \log |\boldsymbol{\Psi}^{-1}\boldsymbol{T}| + n\{\mathrm{tr}\,\boldsymbol{\Psi}^{-1}\boldsymbol{T} - (p+q)\},$$

under M_r. We have $\boldsymbol{\Psi} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}'$. Therefore, as a naive estimator of $\boldsymbol{\Psi}$ it is natural to consider an estimator in the form $\boldsymbol{\Psi}_{(1)} = \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}'$, where $\boldsymbol{H}_x = (\boldsymbol{h}_{1x}, \ldots, \boldsymbol{h}_{px})$, $\boldsymbol{H}_y = (\boldsymbol{h}_{1y}, \ldots, \boldsymbol{h}_{qy})$,

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_y \end{pmatrix}.$$

Then we have

$$\begin{aligned} D(\boldsymbol{T}, \boldsymbol{\Psi}_{(1)}) &= -n \log |\boldsymbol{\Psi}_{(1)}^{-1}\boldsymbol{T}| + n\{\mathrm{tr}\,\boldsymbol{\Psi}_{(1)}^{-1}\boldsymbol{T} - (p+q)\} \\ &= -n \log |\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| + n\{\mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{S} - (p+q)\} = D(\boldsymbol{S}, \boldsymbol{\Sigma}), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{S} &= \begin{pmatrix} \boldsymbol{H}_x'\boldsymbol{T}_{xx}\boldsymbol{H}_x & \boldsymbol{H}_x'\boldsymbol{T}_{xy}\boldsymbol{H}_y \\ \boldsymbol{H}_y'\boldsymbol{T}_{yx}\boldsymbol{H}_x & \boldsymbol{H}_y'\boldsymbol{T}_{yy}\boldsymbol{H}_y \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{0} & \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \\ \boldsymbol{0} & \boldsymbol{S}_{22} & \boldsymbol{S}_{23} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{31} & \boldsymbol{S}_{32} & \boldsymbol{S}_{33} & \boldsymbol{0} \\ \boldsymbol{S}_{41} & \boldsymbol{S}_{42} & \boldsymbol{0} & \boldsymbol{S}_{44} \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \\ \boldsymbol{0} & \boldsymbol{D}_2 & \boldsymbol{S}_{23} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{31} & \boldsymbol{S}_{32} & \boldsymbol{D}_3 & \boldsymbol{0} \\ \boldsymbol{S}_{41} & \boldsymbol{S}_{42} & \boldsymbol{0} & \boldsymbol{D}_4 \end{pmatrix}, \end{aligned} \tag{10}$$

Here, $\boldsymbol{D}_1 = \mathrm{diag}(l_{1x}, \ldots, l_{p_1x})$, $\boldsymbol{D}_2 = \mathrm{diag}(l_{(p_1+1)x}, \ldots, l_{px})$, $\boldsymbol{D}_3 = \mathrm{diag}(l_{1y}, \ldots, l_{q_1y})$, and $\boldsymbol{D}_4 = \mathrm{diag}(l_{(q_1+1)y}, \ldots, l_{qy})$.

Now we consider to derive DIC for M_r. Under M_r, it is assumed that $\boldsymbol{\Sigma}$ has the structure $\boldsymbol{\Sigma}_r$ in (9). Here, for the diagonal elements of $\boldsymbol{\Lambda}_i$, $i = 1, 2, 3, 4$ it is assumed that these are unknown parameters satisfying $\lambda_{1x} \geq \cdots \geq \lambda_{px} > 0$ and $\lambda_{1y} \geq \cdots \geq \lambda_{qy} > 0$. Our purpose is to derive

$$DIC = -n \log |\hat{\boldsymbol{\Sigma}}_r^{-1}\boldsymbol{S}| + n\{\mathrm{tr}\hat{\boldsymbol{\Sigma}}_r^{-1}\boldsymbol{S} - (p+q)\} + B_0,$$

where B_0 is the number of independent parameters in the set of $\boldsymbol{\Sigma}$ under M_r, and $\hat{\boldsymbol{\Sigma}}_r$ is the minimum distance estimator satisfying

$$\min_{M_r} D(\boldsymbol{S}, \boldsymbol{\Sigma}) = D(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}_r).$$

Then, we obtain the minimum distance estimator of $\boldsymbol{\Sigma}$ under M_r as follows:

$$\hat{\boldsymbol{\Sigma}}_{11} = \boldsymbol{S}_{11},\ \hat{\boldsymbol{\Sigma}}_{22} = \boldsymbol{S}_{22},\ \hat{\boldsymbol{\Sigma}}_{33} = \boldsymbol{S}_{33},\ \hat{\boldsymbol{\Sigma}}_{44} = \boldsymbol{S}_{44},\ \hat{\boldsymbol{\Sigma}}_{13} = \boldsymbol{S}_{13},$$

and hence $\hat{\boldsymbol{\Sigma}}_{(13)(13)} = \boldsymbol{S}_{(13)(13)}$. The minimum distance estimator can be expressed in a matrix form as:

$$\hat{\boldsymbol{\Sigma}}_r = \boldsymbol{S}_r = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{0} & \boldsymbol{S}_{13} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{22} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{S}_{31} & \boldsymbol{0} & \boldsymbol{S}_{33} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{S}_{44} \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{S}_{13} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{S}_{31} & \boldsymbol{0} & \boldsymbol{D}_3 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_4 \end{pmatrix}. \tag{11}$$

We now derive a closed form of DIC. It is easily seen that

$$|\hat{\boldsymbol{\Sigma}}_r^{-1}\boldsymbol{S}| = \frac{|\boldsymbol{S}|}{|\boldsymbol{S}_r|} = \frac{|\boldsymbol{S}_{(13)(13)}||\boldsymbol{S}_{(24)(24)\cdot 13}|}{|\boldsymbol{S}_{(13)(13)}||\boldsymbol{S}_{22}||\boldsymbol{S}_{44}|} = \frac{|\boldsymbol{S}_{(24)(24)\cdot 13}|}{|\boldsymbol{S}_{22}||\boldsymbol{S}_{44}|}, \tag{12}$$

and

$$\begin{aligned} \mathrm{tr}\hat{\boldsymbol{\Sigma}}_r^{-1}\boldsymbol{S} &= \mathrm{tr}\hat{\boldsymbol{\Sigma}}_{(13)(13)}^{-1}\boldsymbol{S}_{(13)(13)} + \mathrm{tr}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\boldsymbol{S}_{22} + \mathrm{tr}\hat{\boldsymbol{\Sigma}}_{44}^{-1}\boldsymbol{S}_{44} \\ &= p_1 + q_1 + p - p_1 + q - q_1 = p + q. \end{aligned} \tag{13}$$

Since $\boldsymbol{\Sigma}$ has a structure given in (9), the independent number of $\boldsymbol{\Sigma}$ under M_r is given as $p + q + p_1q_1$. Further, $\boldsymbol{\Gamma}_x$ and $\boldsymbol{\Gamma}_y$ are orthogonal matrices of orders p and q, respectively,

$$\begin{aligned} B_0 &= 2\left\{p + q + p_1q_1 + \frac{1}{2}p(p-1) + \frac{1}{2}q(q-1)\right\} \\ &= 2\left\{\frac{1}{2}(p+q)(p+q+1) - (p-p_1)(q-q_1) - p_1(q-q_1) - (p-p_1)q_1\right\}. \end{aligned} \tag{14}$$

Substituting (12), (13) and (14) into (1), we obtain:

$$DIC = -n\log\{|\boldsymbol{S}_{(24)(24)\cdot 13}|/(|\boldsymbol{S}_{22}||\boldsymbol{S}_{44}|)\} \tag{15}$$
$$+2\left\{\frac{1}{2}(p+q)(p+q+1)-(p-p_1)q-p_1(q-q_1)\right\}.$$

In the use of DIC, we calculate DIC of each subset of $\boldsymbol{u}$ and $\boldsymbol{v}$. The subset to which DIC is minimized is the best subset. In PCCA, as being pointed in Section 1, we may consider only the subsets such that

$$\{u_1,\ldots,u_i\}\cup\{v_1,\ldots,v_j\},\ i=1,\ldots,p,\ j=1,\ldots,q.$$

That is, in PCCA, it is sufficient to calculate DIC in times of $(p\times q)$ times. However, If we consider all the subsets of the original variables, we need to calculate DIC in times of $\{(2^p-1)\times(2^q-1)\}$ times. This is one of the advantages using PCCA instead of CCA.

5. A Numerical Example

We demonstrate the effectiveness of DIC by using a numeric example. The results are shown on the day of our representation.

REFERENCES

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

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, (3rd ed.). John Wiley & Sons, New York.

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

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

Fujikoshi, Y. and Kurata, H. (2008). Information criterion for some condition independence structures. to appear in *New Trends in Psychometrics* (K. Sigemasu et al. eds.), Universal Academy Press, Tokyo, 69–79.

Ichikawa, M. and Konishi, S. (1999). Model evaluation and information criteria in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, **52**, 285–302.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, Springer, New York.

Ogura, T. (2010). A variable selection method in principal canonical correlation analysis. *Computational Statistics and Data Analysis*, **54**, 1117–1123.

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

Sugiura, N. (1976). Asymptotic expansions of the distributions of the canonical correlations. *Commemoration Volume of Founding of Faculty of Integrated Arts and Sciences Hiroshima University*, 141–163.

Sugiyama, T. and Takeda, Y. (1999). On the characterization of principal canonical correlation analysis. *Bulletin of the International Statistical Institute 52nd Session, Finland 1999.*