A Review of Discriminant Analysis by Regression Approach

Yasunori Fujikoshi* and Tamio Kan**

*Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi Hiroshima 739-8626, Japan
**I-Stat Institute of Statistical Analysis
Suginami-ku, Tokyo 166-0011, Japan

Abstract

This paper reviews discriminant methods by regression approach. After a brief review of linear discriminant analysis, we explain an optimum scoring method based on Hastie et al. (1994), which introduces a formal multivariate regression model useful in deriving various discriminant methods. Based on the multivariate regression model, we discuss with linear discriminant functions, tests for discriminant functions, information criteria for selection of variables, sparse discriminant methods. It is examined that most of them are essentially the same as the ones based on direct approach for discriminant analysis. However, it is noted that some sparse methods in two approaches may be not the same.

AMS 2000 Subject Classification: primary 62H30; secondary 62H12

Key Words and Phrases: Additional information hypothesis, discriminant analysis, Information criteria, LR tests, selection of variables, regression approach, sparse discriminant methods.

Abbreviated title: Discriminant Analysis by Regression Approach.

1. Introduction

This paper is concerned with multiple discriminant analysis by regression approach. In two-group discriminant analysis of p variables, if one sets an appropriate dummy variable Y, a formal regression vector is proportional to the coefficient vector of the linear discriminant function. This result has been long, and can be seen in the book by Anderson (1958). An extension to multiple group was done by Hastie et al. (1994). This formal multivariate regression model is expected to be useful in finding various discriminant methods. In fact, through the formal multivariate regression model, we can provide various methods in discriminant analysis. Most of them are essentially equivalent to the ones based on direct approach in discriminant analysis. However, when the optimal scaling will be sequentially decided as in sparse methods due to Clemmensenen et al. (2011), the resultant methods may be different from the ones based on direct approaches.

This paper first reviews linear discriminant methods. Then, we explain the optimal scaling method which induces a formal multivariate regression model connecting discriminant analysis. The linear discriminant functions can be obtained an the estimator of regression coefficients. Tests on significance of regression coefficients are shown to be the tests of no additional information hypothesis of a subset of variables in discriminant analysis. We also examine an equivalence of information criteria for selection of variables in discriminant model and in multivariate regression model. The result is extended to the one in the dimensionality problem. We also give a brief review on sparse discriminant analysis based on regression analysis. For the case when each of the p variables are categorical, see Kan (2009) and Kan and Fujikoshi (2011).

2. Linear Discriminant Analysis

We consider the case of q-group discriminant analysis with p variables $x_1, \ldots, x_p$. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$, and let the mean and the covariance matrix of $\boldsymbol{x}$ in the i-th population $\Pi^{(i)}$ be $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$, $i = 1, \ldots, q$. Suppose that there are n_i samples $\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_{n_i}^{(i)}$ from $\Pi^{(i)}$ such that

$$\boldsymbol{x}_j^{(i)} \sim \mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}), \quad j = 1, \ldots, n_i;\ i = 1, \ldots, q. \tag{2.1}$$

Let the sample mean vectors and the sample covariance matrices by

$$\bar{\boldsymbol{x}}^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_j^{(i)}, \quad \mathbf{S}^{(i)} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})'.$$

All the samples are denoted by the $n \times p$ matrix

$$\mathbf{X} = (\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}, \ldots, \boldsymbol{x}_1^{(q)}, \ldots, \boldsymbol{x}_{n_q}^{(q)})', \tag{2.2}$$

where $n = n_1 + \cdots + n_q$. The notation $\mathbf{X}$ is also used for the centralized data matrix

$$\mathbf{X} = (\boldsymbol{x}_1^{(1)} - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_{n_1}^{(1)} - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_1^{(q)} - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_{n_q}^{(q)} - \bar{\boldsymbol{x}})', \tag{2.3}$$

where $\bar{\boldsymbol{x}}$ is the overall mean vector defined by $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{q} n_i \bar{\boldsymbol{x}}^{(i)}$.

For testing

$$H_0 : \ \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}, \tag{2.4}$$

we have two basic statistics given by

$$\mathbf{B} = \sum_{i=1}^{q} n_i (\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})', \quad \mathbf{W} = \sum_{i=1}^{q} (n_i - 1)\mathbf{S}^{(i)}. \tag{2.5}$$

The matrices $\mathbf{B}$ and $\mathbf{W}$ are called the matrices of sums of squares and products due to between-groups and within-groups, respectively. The matrix $\mathbf{T} = \mathbf{B} + \mathbf{W}$ is called the matrix of sums of squares and products due to the total variation.

Suppose that $n - q \geq p$. Then, $\mathbf{W}$ is nonsingular. Let $\ell_1 > \ldots > \ell_m > \ell_{m+1} = \cdots = \ell_p = 0$ be the non-zero eogenvalues of $\mathbf{B}\mathbf{W}^{-1}$, where

$m = \min(p, q-1)$. Then, the coeeficient vectors $\boldsymbol{h}_i, i = 1, \ldots, m$ of linear (or Fisher's) discriminant functions are defined as the characteristic vectors of $\mathbf{BW}^{-1}$ or $\mathbf{BT}^{-1}$ with normalizations, for example, $\boldsymbol{h}_i'\mathbf{W}\boldsymbol{h}_i = n$, that is, the solutions of the characteristic root problem

$$\mathbf{B}\boldsymbol{h}_i = \ell_i \mathbf{W}\boldsymbol{h}_i, \quad \boldsymbol{h}_i'\mathbf{W}\boldsymbol{h}_j = n\delta_{ij}, \tag{2.6}$$

where δ_{ij} is the Kroneker's delta, i.e., $\delta_{ii} = 1$, and for $i \neq j$, $\delta_{ij} = 0$. The coeficient vectors may be defined as the solutions of the characteristic equations

$$\mathbf{B}\boldsymbol{h}_i = d_i \mathbf{T}\boldsymbol{h}_i, \quad \boldsymbol{h}_i'\mathbf{T}\boldsymbol{h}_j = n(1-d_i)^{-1}\delta_{ij}, \tag{2.7}$$

where $d_1 > \cdots > d_m > d_{m+1} = \cdots = d_p = 0$ are the characteristic roots of $\mathbf{BT}^{-1}$. There are relations given by $d_i = \ell_i/(1+\ell_i),\ i = 1, \ldots, m$.

Note that the linear discriminant functions may be characterized as the solution of the following problem:

$$\begin{aligned} &\max_{\boldsymbol{h}_k} \{\boldsymbol{h}_k'\mathbf{B}\boldsymbol{h}_k\} \\ &\quad \text{subject to } \frac{1}{n}\boldsymbol{h}_k'\mathbf{W}\boldsymbol{h}_k = 1, \quad \boldsymbol{h}_k'\mathbf{W}\boldsymbol{h}_\ell = 0;\ \forall \ell < k. \end{aligned} \tag{2.8}$$

These discriminant fuctions are also used for a practical classification. In fact, let $\boldsymbol{x}_0$ be a new observation vector. Let

$$\boldsymbol{u}_0 = \mathbf{H}\boldsymbol{x}_0, \quad \mathbf{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k)'$$

and let $\bar{\boldsymbol{u}}^{(i)} = \mathbf{H}\bar{\boldsymbol{x}}^{(i)},\ i = 1, \ldots, q$. Then, the rule is to assign $\boldsymbol{x}_0$ to $\boldsymbol{\Pi}^{(i)}$ if

$$\min\{\|\boldsymbol{u}_0 - \bar{\boldsymbol{u}}^{(1)}\|, \ldots, \|\boldsymbol{u}_0 - \bar{\boldsymbol{u}}^{(q)}\|\} = \|\boldsymbol{u}_0 - \bar{\boldsymbol{u}}^{(i)}\|.$$

3. Regression Approach

In two-group discriminant analysis, it is known that if one sets an appropriate dummy variable y, a formal regression coefficient vector of y to $\boldsymbol{x}$ is proportional to the coefficient vector of the linear discriminant function. In fact, suppose the values of y are

$$y_j^{(1)} = \sqrt{n_2/n_1},\ j = 1, \ldots, n_1, \quad y_j^{(2)} = -\sqrt{n_1/n_2},\ j = 1, \ldots, n_2.$$

Note that the mean of y is 0. Consider the regression of y to $\boldsymbol{x}$, and minimize

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} [y_j^{(i)} - \boldsymbol{\beta}'(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}})]^2.$$

Then the optimum solution of $\boldsymbol{\beta}$ is proportional to $\mathbf{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})$, where $\mathbf{S}$ is the pooled sample covariance matrix, i.e., $\mathbf{S} = (n-2)^{-1}\mathbf{W}$.

Now we consider the case of q-group discriminant analysis along Hastie et al.(1994). Let $m = \min(p, q-1)$. Consider an m-dimensional variate $\boldsymbol{y} = (y_1, \ldots, y_m)'$. We denote the matrix value of $\boldsymbol{y}$ for n subjects by

$$\mathbf{Y} = (\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_{n_1}^{(1)}, \ldots, \boldsymbol{y}_1^{(q)}, \ldots, \boldsymbol{y}_{n_q}^{(q)})'.$$

Consider a scoring of $\mathbf{Y}$ which is of the form

$$\begin{aligned} \mathbf{Y} &= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \\ &= \mathbf{Z}\boldsymbol{\Theta} = \mathbf{Z}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m), \end{aligned}$$

where

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_q} \end{pmatrix},$$

and $\mathbf{1}_n$ is the n-dimensional vector whose elements are all one. Here, $\mathbf{Z} = (z_{ij})$ may be defined as follows: $z_{ij} = 1$ if the ith sample belongs to $\Pi^{(j)}$, and 0 otherwise. Hasite et al. (1994) considered the case that m may be any one less than $\min(p, q-1)$.

In the following, without loss of information we assume $\boldsymbol{x} = \mathbf{0}$, that is, the observations of $\boldsymbol{x}$ have been centralized with respect to the mean. The regression of $\mathbf{Y}$ to $\mathbf{X}$ is $\mathbf{X}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \mathbf{X}\mathcal{B}$. The optimal scoring problem involves finding the coefficients $\boldsymbol{\beta}_j (j = 1, \ldots, m)$ and the parameter $\boldsymbol{\Theta}$ that minimize the following average squared residual (ASR):

$$\begin{aligned} \mathrm{ASR}(\boldsymbol{\Theta}, \mathcal{B}) &= \frac{1}{n} \sum_{j=1}^{m} \|\boldsymbol{y}_j - \mathbf{X}\boldsymbol{\beta}_j\|^2 \\ &= \frac{1}{n} \sum_{j=1}^{m} \|\mathbf{Z}\boldsymbol{\theta}_j - \mathbf{X}\boldsymbol{\beta}_j\|^2 \\ &= \frac{1}{n} \mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta} - \mathbf{X}\mathcal{B})'(\mathbf{Z}\boldsymbol{\Theta} - \mathbf{X}\mathcal{B}), \end{aligned} \tag{3.1}$$

under the restriction

$$\mathbf{Y}'\mathbf{Y} = \boldsymbol{\Theta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Theta} = n\mathbf{I}_m. \tag{3.2}$$

When $\boldsymbol{\Theta}$ is fixed, ASR is minimized at

$$\widehat{\mathcal{B}}_0 = \widehat{\mathcal{B}}_0(\boldsymbol{\Theta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Theta},$$

and

$$\mathrm{ASR}(\boldsymbol{\Theta}, \widehat{\mathcal{B}}_0) = \frac{1}{n}\{\mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta})'\mathbf{Z}\boldsymbol{\Theta} - \mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta})'\mathbf{P}_{\mathbf{X}}\mathbf{Z}\boldsymbol{\Theta}\}, \tag{3.3}$$

where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection operator to the space spanned by the column vectos of $\mathbf{X}$. The first term of (3.3) is m, since the restriction (3.2) is satisfied. The optimal problem with respect to $\boldsymbol{\Theta}$ is to find $\boldsymbol{\Theta}$ such that $(1/n)\mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta})'\mathbf{P}_{\mathbf{X}}\mathbf{Z}\boldsymbol{\Theta}$ is maximized under $\boldsymbol{\Theta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Theta} = n\mathbf{I}_m$. Such $\boldsymbol{\Theta}$ is given as the characteristic vectors of $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Z}$, that is the solution of the following general characteristic root problem:

$$\mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Z}\widehat{\boldsymbol{\Theta}} = \mathbf{Z}'\mathbf{Z}\widehat{\boldsymbol{\Theta}}\mathbf{D}, \quad \widehat{\boldsymbol{\Theta}}'\mathbf{Z}'\mathbf{Z}\widehat{\boldsymbol{\Theta}} = n\mathbf{I}_m, \tag{3.4}$$

where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m)$, $d_1 > \cdots > d_m$ and d_i's are the the characteristic roots of $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Z}$. Noting that $\mathbf{X}$ has been centralized, we have $\mathbf{T} = \mathbf{X}'\mathbf{X}$ and $\mathbf{B} = \mathbf{X}'\mathrm{P}_{\mathbf{Z}}\mathbf{X}$. So, $d_i (i = 1, \ldots, m)$ are also the nonzero characteristic roots of $\mathbf{B}\mathbf{T}^{-1}$. Let us define

$$\widehat{\mathcal{B}} = \widehat{\mathcal{B}}_0(\widehat{\boldsymbol{\Theta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{3.5}$$

where $\widehat{\boldsymbol{\Theta}}$ is defined by (3.4).

Now we define a scoring for $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\Theta}}, \tag{3.6}$$

where $\hat{\boldsymbol{\Theta}}$ is defined by (3.4). Note that $\widehat{\mathcal{B}}$ can be regarded as an estimator in a formal multivariate regression model $\mathbf{Y}(= \mathbf{Z}\widehat{\boldsymbol{\Theta}})$:

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathbf{V}, \tag{3.7}$$

where $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)'$. Further, let us assume a usual assumption as in multivariate regression model such that

$$\begin{aligned} &(1)\ \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \sim i.i.d. \\ &(2)\ \mathrm{E}(\boldsymbol{v}_i) = \mathbf{0},\ \mathrm{Var}(\boldsymbol{v}_i) = \widetilde{\boldsymbol{\Sigma}}. \end{aligned}$$

The least squares estimator $\widehat{\mathcal{B}}$ is given by (3.5). Examine a relationship of $\mathbf{H}$ in (2.6) and $\mathcal{B}$ in (3.5). The first equation in (3.4) can be rewritten as

$$\mathbf{B}\hat{\mathcal{B}} = \mathbf{T}\hat{\mathcal{B}}\mathbf{D}. \tag{3.8}$$

Further, the second equation can be rewritten as

$$\hat{\mathcal{B}}'\mathbf{T}\hat{\mathcal{B}} = \mathbf{D}. \tag{3.9}$$

These imply that

$$\mathbf{H} = \widehat{\mathcal{B}}\mathbf{D}^{-1/2}.$$

except the signs of the column vectors. This implies that the coefficient vectors of the linear discriminant functions can be obtained by a multivariate regression approah.

Note that an optimum value of $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\mathcal{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ may be characterized as follows:

$$\begin{aligned} &\min_{\boldsymbol{\beta}_k, \boldsymbol{\theta}_k} \left\{ \|\mathbf{Z}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k\|^2 \right\} \\ &\quad \text{subject to } \frac{1}{n}\boldsymbol{\theta}_k'\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k'\mathbf{Z}'\mathbf{Z}\boldsymbol{\theta}_\ell = 0;\ \forall \ell < k. \end{aligned} \tag{3.10}$$

The formal model (3.7) is useful in finding some discriminant methods. However, their properties should be examined under the discriminant model (2.1).

Using the regression approach, it is possible to consider the discriminant analysis based on a transformed variate:

$$\boldsymbol{t} : \boldsymbol{x} \to \boldsymbol{t}(\boldsymbol{x}) = (t_1(\boldsymbol{x}), \ldots, t_s(\boldsymbol{x}))'$$

instead of $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Such discriminant method is called flexible discriminant analysis (Hastie et al. (1994)).

The discriminant method based on the idea of ridge regression is to use

$$\hat{\mathcal{B}}_* = \{\mathbf{X}'\mathbf{X} + \boldsymbol{\Omega}\}^{-1}\mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Theta}} \tag{3.11}$$

instead of the coefficient vectors of the discriminant functions in (3.5). Here $\boldsymbol{\Omega} = \lambda\mathbf{I}_p$, and λ is the ridge parameter. This is based on the ridge regression-based optimal scoring method (Friedman (1989), Hastie et al. (1994, 1995), Ghosh (2003)) based on the following objective function:

$$\mathrm{ASR}_*(\boldsymbol{\Theta}, \mathcal{B}) = \frac{1}{n}\left\{\mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta} - \mathbf{X}\mathcal{B})'(\mathbf{Z}\boldsymbol{\Theta} - \mathbf{X}\mathcal{B}) + \mathrm{tr}\mathcal{B}'\boldsymbol{\Omega}\mathcal{B}\right\}. \tag{3.12}$$

The optimal scoring is obtained under the restriction $\boldsymbol{\Theta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Theta} = n\mathbf{I}_p$. Then, similarly, the optimum solution of $\mathcal{B}$ when $\boldsymbol{\Theta}$ is fixed is given by (3.11) , and

$$\mathrm{ASR}_*(\boldsymbol{\Theta}, \hat{B}_*) = \frac{1}{n}\{\mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta})'\mathbf{Z}\boldsymbol{\Theta} - \mathrm{tr}(\mathbf{Z}\boldsymbol{\Theta})'\mathbf{X}(\mathbf{X}'\mathbf{X} + \boldsymbol{\Omega})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Theta}\}.$$

Therefore, the optimum $\boldsymbol{\Theta}$ is given by the solution

$$\mathbf{Z}'\mathbf{X}\{\mathbf{X}'\mathbf{X} + \boldsymbol{\Omega}\}^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Theta} = \mathbf{Z}'\mathbf{Z}\boldsymbol{\Theta}\mathbf{D}_\alpha, \quad \boldsymbol{\Theta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Theta} = n\mathbf{I}_m, \tag{3.13}$$

where $\mathbf{D}_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$ and $\alpha_1 > \cdots > \alpha_m$ are the characteristic roots of $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\{\mathbf{X}'\mathbf{X} + \boldsymbol{\Omega}\}^{-1}\mathbf{X}'\mathbf{Z}$.

4. Tests for Additional Information

In discriminant analysis it is important to consider whether a set of variables has no additional information, in the presence of remainder variables. For a notational simplicity, consider the case that $\boldsymbol{x}_2 = (x_{k+1}, \ldots, x_p)'$ has no additional information, in the presence of $\boldsymbol{x}_1 = (x_1, \ldots, x_k)'$. Such notion may be called sufficiency of $\boldsymbol{x}_1$, or redundancy of $\boldsymbol{x}_2$. Related to the partition of $\boldsymbol{x}$, let $\boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}$ partition as

$$\boldsymbol{\mu}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(j)} \\ \boldsymbol{\mu}_2^{(j)} \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

The sufficiency hypothesis of $\boldsymbol{x}_1$ was introduced (see, e.g., Rao (1970, 1973)) as

$$H_{2\cdot 1} : \ \boldsymbol{\mu}_{2\cdot 1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot 1}^{(q)}, \tag{4.1}$$

where $\boldsymbol{\mu}_{2\cdot 1}^{(j)} = \boldsymbol{\mu}_2^{(j)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1^{(j)}$, $j = 1, \ldots, q$. For two equivalent conditions, see Fujikoshi (1982). Let $\mathbf{W}$ and $\mathbf{T}$ partition in the same manner as the partition of $\boldsymbol{x}$:

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

Then, the likelihood ratio (LR) criterion λ for the hypothesis $H_{2\cdot 1}$ in (4.1) is given by

$$\lambda^{2/n} = \Lambda_{2\cdot 1} = \frac{|\mathbf{W}_{22\cdot 1}|}{|\mathbf{T}_{22\cdot 1}|},$$

where $\mathbf{W}_{22\cdot 1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{21}$ and $\mathbf{T}_{22\cdot 1} = \mathbf{T}_{22} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{21}$. Note that $\Lambda_{2\cdot 1}$ can be written as

$$\Lambda_{2\cdot 1} = \frac{|\mathbf{W}|}{|\mathbf{W}_{11}|} \cdot \frac{|\mathbf{T}|}{|\mathbf{T}_{11}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|} \cdot \left\{ \frac{|\mathbf{W}_{11}|}{|\mathbf{T}_{11}|} \right\}^{-1}. \tag{4.2}$$

We use the notations,

$$\Lambda(x_1, \cdots, x_p) = \frac{|\mathbf{W}|}{|\mathbf{T}|}, \ \Lambda(x_1, \cdots, x_k) = \frac{|\mathbf{W}_{11}|}{|\mathbf{T}_{11}|},$$

and

$$\begin{aligned}\Lambda_{2\cdot 1} &= \frac{\Lambda(x_1,\ldots,x_p)}{\Lambda(x_1,\ldots,x_k)} \\ &\equiv \Lambda(x_{k+1},\ldots,x_p|x_1,\ldots,x_k)\end{aligned}$$

In general, let $\mathbf{U}$ and $\mathbf{V}$ be independently distributed as a Wishart distribution $\mathrm{W}_p(s,\mathbf{\Sigma})$ and a Wishart distribution $\mathrm{W}_p(t,\mathbf{\Sigma})$, respectively, with $t \geq p$. Then the distribution of

$$\Lambda = \frac{|\mathbf{V}|}{|\mathbf{U}+\mathbf{V}|}$$

is called (se, e.g., Anderson (2003)) as a Λ-distribution with the degrees of freedom p, s, t, whose distribution is denoted as $\Lambda_p(s,t)$. Under the hypothesis $H_{2\cdot 1}$, it is known (see, e.g. Fujikoshi et al. (2010)) that

$$\Lambda(x_{k+1},\ldots,x_p|x_1,\ldots,x_k) \sim \Lambda_{p-k}(q, n-q-k).$$

For an $n \times m$ matrix $\mathbf{Y} = \mathbf{Z}\widehat{\mathbf{\Theta}}$, where $\widehat{\mathbf{\Theta}}$ is defined as a solution of (3.4), consider a formal multivariate regression model

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\mathcal{B} + \mathbf{V} \\ &= \mathbf{X}_1\mathcal{B}_1 + \mathbf{X}_2\mathcal{B}_2 + \mathbf{V},\end{aligned} \tag{4.3}$$

as in (3.7), where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, $\mathbf{X}_1 : n \times k$ and $\mathcal{B} = (\mathcal{B}_1', \mathcal{B}_2')'$. The additional information hypothesis (4.1) in discriminant analysis corresponds to

$$\widetilde{H}_{2\cdot 1} : \ \mathcal{B}_2 = \mathbf{O}, \tag{4.4}$$

in (4.3). Let $\mathbf{S}_e$ and $\mathbf{S}_h$ be the matrices of sums of squares and products due to the errors under model (4.3) and the hypothesis $\tilde{H}_{2\cdot 1}$. Then they are given by

$$\mathbf{S}_e = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}, \quad \mathbf{S}_h = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}.$$

The likelihood ratio criterion under normality is based on

$$\widetilde{\Lambda}_{2\cdot 1} = \frac{|\mathbf{S}_e|}{|\mathbf{S}_h + \mathbf{S}_e|}. \tag{4.5}$$

Interestingly, it holds that

$$\Lambda_{2\cdot 1} = \widetilde{\Lambda}_{2\cdot 1}, \tag{4.6}$$

whose proof is given in Appendix.

5. Information Criteria for Selection of Variables

First we consider information criteria for selection of variables in discriminant model (2.1). It is natural to select $\boldsymbol{x}_1$ if the sufficiency hypothesis $H_{2\cdot 1}$ of $\boldsymbol{x}_1$ is true, or the no additional information hypothesis of $\boldsymbol{x}_2$ is true. It is known that the hypothesis $H_{2\cdot 1}$ can be expressed in terms of the population discriminant functions as follows. Let ρ_i be the prior probability of populations $\Pi^{(i)}$, where *prior* means the a priori probability that an individual selected at random belongs to $\Pi^{(i)}$. The population between groups matrix is defined by

$$\boldsymbol{\Psi} = \sum_{i=1}^{q} \rho_i (\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}})'(\boldsymbol{\mu}^{(i)} - \bar{\boldsymbol{\mu}}), \tag{5.1}$$

and $\bar{\boldsymbol{\mu}} = \rho_1 \boldsymbol{\mu}^{(1)} + \cdots + \rho_q \boldsymbol{\mu}^{(q)}$. Let $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_p = 0$ be the characteristic roots of $\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}$. Then, the corresponding characteristic vectors are denoted by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ with normalizations $\boldsymbol{\gamma}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_i = 1$, $i = 1, \ldots, p$. They are the solutions of

$$\boldsymbol{\Psi}\boldsymbol{\gamma}_i = \lambda_i \boldsymbol{\Sigma}\boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_j = \delta_{ij}. \tag{5.2}$$

Then, $\boldsymbol{\gamma}_i$ is the coefficient vector of the ith population linear discriminant function. Let $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Psi}$ be partitioned as

$$\boldsymbol{\gamma}_j = \begin{pmatrix} \boldsymbol{\gamma}_{1j} \\ \boldsymbol{\gamma}_{2j} \end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix}$$

in the same way as $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$. Then the statement $H_{2\cdot 1}$ is equivalent to one

of the following statments:

$$(1) \quad \boldsymbol{\gamma}_{2j} = 0, \quad j = 1, \ldots, r, \quad r = \mathrm{rank}(\boldsymbol{\Psi}).$$
$$(2) \quad \mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi} = \mathrm{tr}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Psi}_{11}.$$

Related to a seletion of variables, we may consider a slight modification of $H_{2\cdot1}$, (1) or (2). The modification of (1) is defined by

$$M_{2\cdot1}: \ \boldsymbol{\gamma}_{2j} = \mathbf{0}, \text{and } \gamma_{ij} \neq 0, \ i = 1, \ldots, k, \ \text{for } j = 1, \ldots, r. \tag{5.3}$$

Then, AIC for $M_{2\cdot1}$ is given (see, e.g., Fujikoshi et al. (2010)) as

$$\begin{aligned} \mathrm{A}_D = & -n \log\{|\mathbf{W}_{22\cdot1}|/|\mathbf{T}_{22\cdot1}|\} + n \log |n^{-1}\mathbf{W}| \\ & + np(1 + \log 2\pi) + 2\left\{k(q-1) + p + \frac{1}{2}p(p+1)\right\}. \end{aligned} \tag{5.4}$$

Next, let us consider AIC in a formal multivariate regression model (4.3). Let $\mathcal{B} = (\mathcal{B}_1', \mathcal{B}_2')' = (\beta_{ij})$. Related to $\widetilde{H}_{2\cdot1}$, consider

$$\widetilde{M}_{2\cdot1} : \mathcal{B}_2 = \mathbf{O}, \ \text{and } \beta_{ij} \neq 0, \ \text{for } i = 1, \ldots, k, \ j = 1, \ldots, m. \tag{5.5}$$

Under $\widetilde{M}_{2\cdot1}$, we can write the minimum of $-2 \log f(Y; \mathcal{B}, \widetilde{\boldsymbol{\Sigma}})$ as

$$\begin{aligned} & n \log |n^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}| + mn(\log(2\pi) + 1) \\ & = -n \log\{|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}|/|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|\} \\ & + n \log\{n^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}| + mn(\log(2\pi) + 1). \end{aligned}$$

The AIC for $\widetilde{M}_{2\cdot1}$ is given by

$$\begin{aligned} \mathrm{A}_R = & -n \log\{|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}|/|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|\} \\ & + n \log |\frac{1}{n}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}| + mn(\log(2\pi) + 1) \\ & + 2\left\{km + \frac{1}{2}m(m+1)\right\}. \end{aligned} \tag{5.6}$$

We have seen (see (4.5)) that

$$-n \log \frac{|\mathbf{W}_{22\cdot1}|}{|\mathbf{T}_{22\cdot1}|} = -n \log \frac{|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}|}{|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|}.$$

When $p \geq q-1$, $m = q-1$, and then, the model selection criterion A_D is equivalent to A_R.

Similarly, it is seen that BIC for $\widetilde{H}_{2\cdot 1}$ in discriminant model (2.1) is equivalent to BIC for $\widetilde{M}_{2\cdot 1}$ in regression model (4.3).

6. Estimation of Dimensionality

The dimensionality in discriminant model is defined by the number of nonzero characteristic roots λ_i of $\boldsymbol{\Psi}$, or equivalently the rank of $\boldsymbol{\Psi}$, which is the number of meaningful discriminant functions. Let $\mathrm{rank}(\boldsymbol{\Psi}) = k$. This is equivalent to $\mathrm{rank}(\boldsymbol{\Xi}) = k$, where

$$\boldsymbol{\Xi} = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(q)}, \ldots, \boldsymbol{\mu}^{(q-1)} - \boldsymbol{\mu}^{(q)}).$$

In this section we consider to estimate the dimensionality based on selection of models $\mathcal{D}_k, k = 0, 1, \ldots, m$, where

$$\mathcal{D}_k : \ \mathrm{rank}(\boldsymbol{\Xi}) = k. \tag{6.1}$$

It is known (see, e.g., Fujikoshi et al. (2010)) that $-2\log\max$ likelihood is

$$n\log(1+\ell_{k+1})\cdots(1+\ell_m) + n\log|n^{-1}\mathbf{W}| + np(\log 2\pi + 1),$$

where $m = \min\{p, q-1\}$, $\ell_1 > \cdots > \ell_m > 0$ are the characteristic roots of $\mathbf{B}\mathbf{W}^{-1}$. Noting that the dimensionality of $\boldsymbol{\Xi}$ when $\mathrm{rank}(\boldsymbol{\Xi}) = k$ is $k(p+q-1-k)$, the number of independent parameters under $\mathcal{D}_k$ is

$$d(\mathcal{D}_k) = k(p+q-1-k) + p + \frac{1}{2}p(p+1).$$

Therefore, the AIC for $\mathcal{D}_k$ is

$$\begin{aligned}\mathrm{AIC}_k = {} & n\log(1+\ell_{k+1})\cdots(1+\ell_m) + n\log|n^{-1}\mathbf{W}| + np(\log 2\pi + 1) \\ & + 2\left\{k(p+q-1-k) + p + \frac{1}{2}p(p+1)\right\}. \end{aligned} \tag{6.2}$$

Based on AIC, if

$$\min\{\mathrm{AIC}_0, \mathrm{AIC}_1, \ldots, \mathrm{AIC}_m\} = \mathrm{AIC}_k,$$

we estimate the dimensionality as k. Instead of AIC, we may consider

$$\begin{aligned} \mathrm{A}_k &= \mathrm{AIC}_k - \mathrm{AIC}_m \\ &= n \log \prod_{j=k+1}^{m} (1 + \ell_j) - 2(p-k)(q-1-k), \ k = 0, \ldots, m. \end{aligned} \tag{6.3}$$

Here $\mathrm{A}_m = 0$. Then the estimation method is equivalent to that of

$$\min\{\mathrm{A}_0, \mathrm{A}_1, \ldots, \mathrm{A}_m\} = \mathrm{A}_k$$

and we estimate the dimensionality as k.

Next we consider to estimate the rank of $\mathcal{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ in the formal multivariate regression model (3.7). We have seen that the least squares estimator $\widehat{\mathcal{B}}$ is the coefficient vectors of the linear discriminant functions. So, the rank of $\mathcal{B}$ means the number of meaningful discriminant functions. We consider to estimate the rank of $\mathcal{B}$ based on selection of models $\widetilde{\mathcal{D}}_k, k = 0, 1, \ldots, m$, where

$$\widetilde{\mathcal{D}}_k : \ \mathrm{rank}(\mathcal{B}) = k. \tag{6.4}$$

Let $f(\mathbf{Y}; \mathcal{B}, \widetilde{\boldsymbol{\Sigma}})$ be the density function of $\mathbf{Y}$ in a formal multivariate regression model (3.7). Then,

$$\begin{aligned} &(-2) \log f(\mathbf{Y}; \mathcal{B}, \widetilde{\boldsymbol{\Sigma}}) \\ &= n \log |\widetilde{\boldsymbol{\Sigma}}| + \mathrm{tr} \widetilde{\boldsymbol{\Sigma}}^{-1} (\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B}) + mn \log(2\pi) \\ &\geq n \log |n^{-1}\mathbf{Q}(\mathcal{B})| + mn\{\log(2\pi) + 1\} \end{aligned}$$

where $\mathbf{Q}(\mathcal{B}) = (\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})$. In the above inequality, the equality holds when $\widetilde{\boldsymbol{\Sigma}} = n^{-1}(\mathbf{Y} - \mathbf{X}\mathcal{B})'(\mathbf{Y} - \mathbf{X}\mathcal{B})$. Note that

$$\begin{aligned} \mathbf{Q}(\mathcal{B}) &= \mathbf{A} + (\widehat{\mathcal{B}} - \mathcal{B})'(\mathbf{X}'\mathbf{X})(\widehat{\mathcal{B}} - \mathcal{B}) \\ &= \mathbf{A}^{1/2}\{\mathbf{I}_m + (\mathbf{V} - \boldsymbol{\Delta})'(\mathbf{V} - \boldsymbol{\Delta})\}\mathbf{A}^{1/2}, \end{aligned}$$

where

$$\begin{aligned} &\mathbf{A} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y}, \ \widehat{\mathcal{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\ &\mathbf{V} = (\mathbf{X}'\mathbf{X})^{1/2}\widehat{\mathcal{B}}\mathbf{A}^{-1/2}, \ \boldsymbol{\Delta} = (\mathbf{X}'\mathbf{X})^{1/2}\mathcal{B}\mathbf{A}^{-1/2}. \end{aligned}$$

Using (3.4), we have

$$\begin{aligned}\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} &= \widehat{\mathbf{\Theta}}' \cdot \mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Z}\widehat{\mathbf{\Theta}} \\ &= \widehat{\mathbf{\Theta}}'\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Theta}}\mathbf{D} = n\mathbf{D}.\end{aligned}$$

Further,

$$\begin{aligned}\mathbf{A} &= n(\mathbf{I}_m - \mathbf{D}). \\ \mathbf{V}'\mathbf{V} &= \mathbf{A}^{-1/2}\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}\mathbf{A}^{-1/2} = (\mathbf{I}_m - \mathbf{D})^{-1}\mathbf{D} = \mathbf{L}.\end{aligned}$$

Therefore,

$$\min_{\text{rank}(\mathcal{B})=k} |\mathbf{Q}(\mathcal{B})| = |\mathbf{I} - \mathbf{D}| \prod_{i=k+1}^{m} (1 + \ell_i).$$

Let $\widetilde{\text{AIC}}_k$ be AIC for $\widetilde{\mathcal{D}}_k$. Instead of $\widetilde{\text{AIC}}$, we may consider

$$\begin{aligned}\widetilde{\text{A}}_k &= \widetilde{\text{AIC}}_k - \widetilde{\text{AIC}}_m \\ &= n \log \prod_{j=k+1}^{m} (1 + \ell_j) - 2(p-k)(m-k), \ k = 0, \ldots, m. \end{aligned} \tag{6.5}$$

Here $\widetilde{\text{A}}_m = 0$.

When $p \geq q - 1$, the two estimation criteria A_k and $\widetilde{\text{A}}_k$ are eqivalent. For high-dimensional consistency properties of AIC_k, see Fujikoshi and Sakurai (2016).

7. Sparse Discriminant Analysis

When $p > n - q$, the matrix $\mathbf{W}$ of sums of squares and products due to within-groups becomes singular. The linear discriminant problem can be not performed. In order to overcome this problem, regularized and sparse methods have been proposed, based on direct discriminant approach and regression approach. However, their details are not given here, and we explain only a few methods.

One approach is to use a regularized estimate of $\mathbf{W}$ in the linear discriminant problem. For instance,

$$\max_{\boldsymbol{h}_k} \{\boldsymbol{h}_k' \mathbf{B} \boldsymbol{h}_k\} \tag{7.1}$$
$$\text{subject to } \frac{1}{n}\boldsymbol{h}_k'(\mathbf{W} + n\boldsymbol{\Omega})\boldsymbol{h}_k = 1, \quad \boldsymbol{h}_k'(\mathbf{W} + n\boldsymbol{\Omega})\boldsymbol{h}_\ell = 0; \ \forall \ell < k,$$

where $\boldsymbol{\Omega}$ is a positive definite marix. Witten and Tibshirani (2011) proposed the following sparse method based on an ℓ_1 penalty. The method is defined sequentially as follows:

$$\max_{\boldsymbol{h}_k} \{\boldsymbol{h}_k' \mathbf{B} \boldsymbol{h}_k - n\gamma \|\boldsymbol{h}_k\|_1\} \tag{7.2}$$
$$\text{subject to } \frac{1}{n}\boldsymbol{h}_k'(\mathbf{W} + n\boldsymbol{\Omega})\boldsymbol{h}_k = 1, \quad \boldsymbol{h}_k'(\mathbf{W} + n\boldsymbol{\Omega})\boldsymbol{h}_\ell = 0, \ \forall \ell < k.$$

As sparse discriminant analysis based on regression approach, there are two cases: (i) the case that $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta}$ has been defined as in (3.6), and (ii) the case that $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta}$ has been sequentially defined. For the first case, we can apply sparse method in multivariate regression model. For example, related to the problem of selection of $\boldsymbol{x}$, we can apply the penalized regression with a grouped lasso penalty (Yuan and Lin (2006)) based on the following optimization problem:

$$\min_{\mathcal{B}} \left\{ \|\mathbf{Y} - \mathbf{X}\mathcal{B}\|^2 + \sum_{i=1}^{p} \lambda_i \|\boldsymbol{\beta}_{(i)}\| \right\}, \tag{7.3}$$

where $\mathcal{B} = (\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(m)})'$, and $\lambda_i > 0$ are penalty parameters. For simultaneous dimension reduction and variable selection, we can use a sparse method based on Chen and Huang (2012). For the second case, Clemmensen et al. (2011) proposed a sparse method based on the elastic net due to Zou and Hastie (2005). The kth sparse discriminant analysis solution pair $\{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k\}$ are defined as the solutions of the problem:

$$\max_{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k} \left\{ \|\mathbf{Z}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k\|^2 + \gamma \boldsymbol{\beta}_k' \boldsymbol{\Omega} \boldsymbol{\beta}_k + \lambda \|\boldsymbol{\beta}_k\|_1 \right\} \tag{7.4}$$
$$\text{subject to } \frac{1}{n}\boldsymbol{\theta}_k' \mathbf{Z}'\mathbf{Z}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k' \mathbf{Z}'\mathbf{Z}\boldsymbol{\theta}_\ell = 0, \ \forall \ell < k,$$

where $\gamma > 0$ and $\lambda > 0$ are penalty parameters. It will be important to compare these methods with penalized and sparse methods based on direct discriminant approach.

Appendix: Proof of $\Lambda_{2\cdot 1} = \widetilde{\Lambda}_{2\cdot 1}$ in (4.6)

Let the nonzero roots of $\mathbf{B}\mathbf{T}^{-1}$ be $d_1 > \cdots > d_m$. Then,

$$\frac{|\mathbf{W}|}{|\mathbf{T}|} = (1-d_1)\cdots(1-d_m).$$

Similarly, let the nonzero roots of $\mathbf{B}_{11}\mathbf{T}_{11}^{-1}$ be $\widetilde{d}_1 > \cdots > \widetilde{d}_{\widetilde{m}}$, where $\widetilde{m} = \min(p_1, q-1)$. Then,

$$\frac{|\mathbf{W}_{11}|}{|\mathbf{T}_{11}|} = (1-\widetilde{d}_1)\cdots(1-\widetilde{d}_{\widetilde{m}}).$$

Therefore, from (4.2) we have

$$\Lambda_{2\cdot 1} = \frac{(1-d_1)\cdots(1-d_m)}{(1-\widetilde{d}_1)\cdots(1-\widetilde{d}_{\widetilde{m}})}. \tag{A.1}$$

Next, consider $\widetilde{\Lambda}_{2\cdot 1}$. We have seen in Section 6 that $\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = n\mathbf{D}$. This implies that $\mathbf{S}_e = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = n(\mathbf{I}_m - \mathbf{D})$, and

$$|\mathbf{S}_e| = n^m(1-d_1)\cdots(1-d_m).$$

Next we consider $\mathbf{S}_h + \mathbf{S}_e = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}$. We construct an extended version of $\mathbf{Y}$ such that

$$\begin{aligned}\mathbf{Y}_e = (\mathbf{Y}, \mathbf{Y}_0) &= \mathbf{Z}(\widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Theta}}_0) \\ &= \mathbf{Z}\widehat{\mathbf{\Theta}}_e,\end{aligned}$$

where $\widehat{\mathbf{\Theta}}_0$ is a $q \times (q-m)$ and the columns of $\widehat{\mathbf{\Theta}}_0$ are the characteristic vectors corresponding to the characteristic root 0 in the characteristic root problem (3.4). The lenghs of $\widehat{\mathbf{\Theta}}_0$ are determined by the restriction $\mathbf{Y}_e'\mathbf{Y}_e = n\mathbf{I}_q$. The characteristic root problem (3.4) can be extended as follows:

$$\mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Z}\widehat{\mathbf{\Theta}}_e = \mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Theta}}\mathbf{D}_e, \quad \widehat{\mathbf{\Theta}}_e'\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Theta}}_e = n\mathbf{I}_q, \tag{A.2}$$

where $\mathbf{D}_e = \mathrm{diag}(d_1, \ldots, d_m, 0, \ldots, 0)$. Note that $\mathbf{P}_{\mathbf{X}_1}\mathbf{Y}_0 = \mathbf{O}$. In fact, from (A.2), $\mathbf{Z}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}_0 = \mathbf{O}$. Multipling $\widehat{\mathbf{\Theta}}_0'$ from left, $\mathbf{Y}_0'\mathbf{P}_{\mathbf{X}}\mathbf{Y}_0 = \mathbf{O}$. Since $\mathbf{P}_{\mathbf{X}}$ is idenpotent, $\mathrm{P}_{\mathbf{X}}\mathbf{Y}_0 = \mathbf{O}$. Further, since $\mathrm{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}_1} = \mathbf{P}_{\mathbf{X}_1}$, we get $\mathrm{P}_{\mathbf{X}_1}\mathbf{Y}_0 = \mathbf{O}$. This implies that

$$\mathbf{Y}_e'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}_e = \begin{pmatrix} \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} & \mathbf{O} \\ \mathbf{O} & n\mathbf{I}_{q-m} \end{pmatrix}. \tag{A.3}$$

Let $\widetilde{\mathbf{\Theta}}_e = (1/\sqrt{n})(\mathbf{Z}'\mathbf{Z})^{1/2}\mathbf{\Theta}_e$. Then, we have

$$(1/n)\mathbf{Y}_e'\mathbf{P}_{X_1}\mathbf{Y}_e = \widetilde{\mathbf{\Theta}}_e'\mathbf{G}\widetilde{\mathbf{\Theta}}_e,$$

where $\mathbf{G} = (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$. Since $\widetilde{\mathbf{\Theta}}_e$ is an orthogonal matrix, the nonzero characteristic roots of $\mathbf{G}$ are the same as $\mathbf{B}_{11}\mathbf{T}_{11}^{-1}$. Using these properties, the determinant of the leftside in (A.3) is expressed as

$$\begin{aligned} |\mathbf{Y}_e'(\mathbf{I}_q - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}_e| &= n^q|\mathbf{I}_q - \widetilde{\mathbf{\Theta}}_e'\mathbf{G}\widetilde{\mathbf{\Theta}}_e| \\ &= n^q(1 - \widetilde{d}_1)\cdots(1 - \widetilde{d}_{\widetilde{m}}). \end{aligned}$$

On the other hand, the determinant of the right side in (A.3) is

$$n^q|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}|,$$

and hence

$$|\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y}| = (1 - \widetilde{d}_1)\cdots(1 - \widetilde{d}_{\widetilde{m}}).$$

This implies (4.6).

Acknowledgements

The first author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), 16K00047, 2016-2018.

References

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. John Wiley & Sons, New York.

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

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

[4] Clemmensen, L. Hastie, T. Witten, D. and Ersboll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406-413.

[5] Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165–175.

[6] Fujikoshi, Y. (1982), A test for additional information in canonical correlation analysis. *Ann. Inst. Statist. Math.*, **34**, 137–144.

[7] Fujikoshi, Y. and Sakurai, T. (2016). High-dimensional consistency of rank estimation criteria in multivariate linear Model. *J. Multivariate Anal.*, **149**, 199-212.

[8] Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, **59**, 992-1000.

[9] Gupta, A. K., Xu, J. and Fujikoshi, Y. (2006). An asymptotic expansion of the distribution of Rao's U-statistic under a general condition. *J. Multivariate Anal.*, **97**, 492-513.

[10] Hastie, T. Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.*, **23**, 73-102.

[11] Hastie, T. Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.*, **89**, 1255–1275.

[12] Kan, T. (2009). Test for additional information and variable selection in quantification method type II. *Japanese J. Appl. Statist.*, **38**, 1-18, (in Japanese).

[13] Kan, T. and Fujikoshi, Y. (2011). *Discriminant Analysis for Qualtative Data; Quantification Method Type II.* Modern Mathematics, INC. (in Japanese).

[14] Rao, C. R. (1970). Inference on discriminant function coefficients. In *Essays in Prob. and Statist.* (R. C. Bose, ed.), 587–602.

[15] Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. Roy. Stat. Soc. Ser. B*, **73**, 753–772.

[16] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. Ser. B*, **68**, 49–67.

[17] Zhou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B*, **67**, 301–320.