Reduced Rank Discrimination

by

M.S. Srivastava
Department of Statistics
University of Toronto

Reduced Rank Discrimination

M.S. Srivastava
University of Toronto
Toronto, Canada

Abstract

In this paper the problem of classifying an individual with \( p \) characteristics into 1 of \( k \) multivariate normal distributions with common unknown covariance matrix is considered when the matrix of \( (k + 1) \) means has a linear structural relationship, that is, it lies in an \( r \)-dimensional plane, where \( r < \min(p, k) \).

1 Introduction

The statistical discriminant analysis developed by Fisher (1936) to classify skeleton remains to a number of known groups has found wide applications in many areas of statistics, such as medicine, image analysis, etc. Initially, the problem was to classify an object with measurements on several of its characteristics to one of the many known groups which were assumed to be distributed as multivariate normal with known mean vectors \( \mu_i \) and common known covariance matrix \( \Sigma \). The problem was later generalized when the means and covariance matrix were not known, including the case where the covariances were the same for each group. See Srivastava and Khatri (1976) for many of these results.

Often in data analysis, observations are taken on many characteristics with the hope of doing a better job of classification even though the means of some of these

1
\[ i = 0, 1, \ldots, k \text{ and } V = nS \sim W_p(\Sigma, n). \] If the matrix \( \mu \) of \( k + 1 \) means has rank \( r \), then it can be shown (see, e.g., Srivastava and Khatri, 1979, p11) that there exists two matrices \( F \) and \( G \) of dimensions \( p \times r \) and \( r \times (k + 1) \) respectively, each of rank \( r \), such that

\[ \mu - FG = F(g_0, \ldots, g_k), \]

where \( g_i \)'s are \( r \)-vectors. Let

\[ \bar{g} = N^{-1} \sum_{i=0}^{k} N_i g_i, \quad \text{where} \quad N = \sum_{i=0}^{k} N_i. \]

Then \( \mu \) can be rewritten as

\[
\begin{align*}
\mu &= F\bar{g}1' + F(g_0 - \bar{g}, \ldots, g_k - \bar{g}) \\
&= \theta_0 1' + F\xi, \quad \sum_{i=0}^{k} N_i \xi_i = 0,
\end{align*}
\]

(2.1)

where \( \xi = (\xi_0, \ldots, \xi_k) = (g_0 - \bar{g}, \ldots, g_k - \bar{g}) \), and \( \theta_0 = F\bar{g} \). The problem of testing that \( \mu \) is of rank \( r < p \) started from Fisher (1938, 1939) and a solution when \( \Sigma \) is known is given in Rao (1973, pp. 556-559). When the covariance matrix \( \Sigma \) is unknown, the problem of estimating \( F, \xi, \Sigma \) and the problem of testing for the rank has been considered by Anderson (1951). Since \( F \) is of rank \( r \), there exists a \( p \times (p-r) \) matrix \( P \) of rank \( p-r \) such that

\[ P'F = 0, \]

(2.2)

and \( (P, F) \) is nonsingular. (See Srivastava and Khatri, 1979, p. 11.) Anderson (1951) assumes that \( P \) is such that

\[ P'\Sigma P = I_{p-r}, \]

(2.3)

and shows that this restriction has no effect on the maximum likelihood estimator. For any consistent estimates \((\hat{P}, \hat{F}, \hat{\Sigma})\) of \((P, F, \Sigma)\), these estimates must also satisfy (2.2) and (2.3). As will be apparent later (see (3.4)), this implies that

\[
\hat{P}'(V + B)\hat{P} = N I_{p-r}, \quad \hat{P}'F = 0, \quad \hat{P}'\Sigma \hat{P} = I_{p-r}
\]

(2.4)
The logarithm of the likelihood function, denoted by \( l(\theta_0, \xi, F, \Sigma) \) is given by

\[
l(\theta_0, \xi, F, \Sigma) = -\frac{1}{2} N \log |\Sigma| - \frac{1}{2} tr \Sigma^{-1} [V + \sum_{i=0}^{k} N_i (\bar{x}_i - \theta_0 - F\xi)(\bar{x}_i - \theta_0 - F\xi)]^t.
\]

Differentiating it partially with respect to \( \theta_0 \) and equating it to zero gives the MLE of \( \theta_0 \),

\[
\hat{\theta}_0 = N^{-1} \Sigma N_i \bar{x}_i - F(\Sigma N_i \xi_i / N)
\]

\[
= \bar{x}, \quad \text{since} \quad \Sigma N_i \xi_i = 0. \quad (3.1)
\]

Let

\[
y_i = \bar{x}_i - \bar{x}, \quad i = 0, 1, \ldots, k.
\]

Then differentiating the log-likelihood partially with respect to \( \xi_i \) and equating it to zero gives the MLE of \( \xi_i \),

\[
\hat{\xi}_i = (F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1} y_i, \quad i = 0, 1, \ldots, k.
\]

Similarly, the MLE of \( \Sigma \) is given by

\[
N \hat{\Sigma} = V + [I - F(F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1}] B[I - F(F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1}]^t
\]

where

\[
B = \sum_{i=0}^{k} N_i y_i y_i^t
\]

\[
= \sum_{i=0}^{k} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t
\]

is the sum of squares and products "between treatments" as defined in (2.5). Pre-multiplying by \( \hat{\Sigma}^{-1} \) and then post-multiplying by \( V^{-1} \), we get

\[
NF^t V^{-1} = F^t \hat{\Sigma}^{-1}, \quad \text{and} \quad NF^t V^{-1} F = F^t \hat{\Sigma}^{-1} F.
\]

Hence,

\[
\hat{\xi}_i = (F^t V^{-1} F)^{-1} F^t V^{-1} y_i, \quad (3.3)
\]
Thus, the determinant in (3.5) is minimized for $A = \hat{A}$ and the minimum value of the determinant is

$$|V| \prod_{i=r+1}^{p} (1 + l_i)$$

Thus, the MLE of $P$ is given by any $\hat{P}$ such that

$$\hat{A} = \hat{P} \left( \hat{P}^t V \hat{P} \right)^{-\frac{1}{2}}$$

Clearly, $\hat{A}$ can be post-multiplied by any $(p - r) \times (p - r)$ orthogonal matrix. Note that from (3.7),

$$V \hat{A} \hat{A}^t B \hat{A} \hat{A}^t V = \left( I - V \hat{C} \hat{C}^t \right) B \hat{A} \hat{A}^t V$$

$$= B \hat{A} \hat{A}^t V.$$ 

Thus, the MLE of $\Sigma$ is given by

$$N \hat{\Sigma} = V + V \hat{A} \hat{A}^t B \hat{A} \hat{A}^t V$$

$$= V + B \hat{A} \hat{A}^t V$$

$$= \left[ I + B \hat{A} \hat{A}^t \right] V.$$

Since

$$\hat{P} \xi_i = V V^{-1} \hat{P} \left( \hat{P}^t V^{-1} \hat{P} \right)^{-1} \hat{P}^t V^{-1} (\bar{x}_i - \bar{x})$$

$$= V [V^{-1} - \hat{P} \left( \hat{P}^t V \hat{P} \right)^{-1} \hat{P}] (\bar{x}_i - \bar{x})$$

$$= [I - V \hat{A} \hat{A}^t] (\bar{x}_i - \bar{x})$$

$$= V \hat{C} \hat{C}^t (\bar{x}_i - \bar{x}),$$

$$\hat{\mu}_i = \bar{x} + V \hat{C} \hat{C}^t (\bar{x}_i - \bar{x})$$

$$= \bar{x} + [I - V \hat{A} \hat{A}^t] (\bar{x}_i - \bar{x})$$

(3.9)

The modified likelihood ratio test for testing the hypothesis that the rank of $\mu$ is $r$, where $r$ is specified against the alternative that it is $p$ is given by,

$$U_{r+1} = [n - \frac{1}{2} (k - p - 1) \sum_{i=1}^{r} l_i^{-2}] \left[ \sum_{i=r+1}^{p} \log(1 + l_i) \right],$$

(3.10)
Letting

\[ y_i = x_i - \bar{x}, \quad i = 0, 1, \ldots, k, \]

the MLEs of \( \xi_1, \xi_2, \ldots, \xi_k \) under model \( H_1 \), are given by

\[ \hat{\xi}_1 = (F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1} (N_0 y_0 + N_1 y_1)/(N_0 + N_1) \]

\[ \hat{\xi}_i = (F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1} y_i, \quad i = 2, \ldots, k, \]

respectively. Let

\[ L = F(F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1}. \]

Then

\[ L^t \hat{\Sigma}^{-1} = \hat{\Sigma}^{-1} F(F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1} = \hat{\Sigma}^{-1} L, \]

and

\[ L^t \hat{\Sigma}^{-1} L = L^t \hat{\Sigma}^{-1} \hat{\Sigma}^{-1} L. \]

Hence, it can be shown that

\[ N_0(y_0 - F \hat{\xi}_1)^t \hat{\Sigma}^{-1}(y_0 - F \hat{\xi}_1) \]

\[ = N_0(y_0 - Ly_0)^t \hat{\Sigma}^{-1}(y_0 - Ly_0) \]

\[ + \frac{2N_0N_1}{N_0 + N_1}(y_0 - Ly_0)^t \hat{\Sigma}^{-1} L(y_0 - y_1) + \frac{N_0N_1^2}{(N_0 + N_1)^2}(y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L(y_0 - y_1) \]

and

\[ N_1(y_1 - F \hat{\xi}_1)^t \hat{\Sigma}^{-1}(y_1 - F \hat{\xi}_1) \]

\[ = N_1(y_1 - Ly_1)^t \hat{\Sigma}^{-1}(y_1 - Ly_1) \]

\[ - \frac{2N_0N_1}{N_0 + N_1}(y_1 - Ly_1)^t \hat{\Sigma}^{-1} L(y_0 - y_1) + \frac{N_0^2N_1}{(N_0 + N_1)^2}(y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L(y_0 - y_1) \]

Adding the two terms, we get

\[ \sum_{i=0}^{1} N_i(y_i - F \hat{\xi}_1)^t \hat{\Sigma}^{-1}(y_i - F \hat{\xi}_1) \]

\[ = \sum_{i=0}^{1} N_i(y_i - Ly_i)^t \hat{\Sigma}^{-1}(y_i - Ly_i) + \frac{N_0N_1}{N_0 + N_1}(y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L(y_0 - y_1), \]
Since \((I - L_1)^t W_1^{-1} (I - L_1) = W_1^{-1} (I - L_1)\), we need to minimize

\[
|W_1 + (I - L_1)(B, -z_1 z_1^t) \begin{pmatrix} (I - L_1)^t \\ I \end{pmatrix} |
\]

\[
= |W_1||I + (B, -z_1 z_1^t) \begin{pmatrix} (I - L_1)^t \\ I \end{pmatrix} W_1^{-1} (I - L_1)|
\]

\[
= |W_1||I + (B - z_1 z_1^t) W_1^{-1} (I - L_1)|
\]

\[
= |W_1||I + (B - z_1 z_1^t) P (P^t W_1 P)^{-1} P^t||
\]

\[
= |W_1||P^t W_1 P|^{-1} |P^t W_1 P + P^t (B - z_1 z_1^t) P|,
\]

where, as before, \(P\) is such that \((F, P)\) is a \(p \times p\) matrix of rank \(p\) and \(P^t F = 0\).

Thus, we need to minimize, with respect to \(P\),

\[
|W_1||P^t W_1 P|^{-1} |P^t V P + P^t B P|
\]

\[
= |W_1||P^t (P^t W_1 P)^{-1} P^t (V + B) P|
\]

\[
= |W_1||(P^t W_1 P)^{-1/2} P^t (V + B) P (P^t W_1 P)^{-1/2}|
\]

\[
= |W_1||A_1^t (V + B) A_1|,
\]

where \(A_1^t = (P^t W_1 P)^{-1/2} P^t\), and \(A_1^t W_1 A_1 = I_{p-r}\). Thus, by choosing \(W_1^{1/2} A_1\) as a matrix of \((p-r)\) orthonormal column vectors of the characteristic vectors corresponding to the \(p-r\) smallest characteristic roots of \((V + z_1 z_1^t)^{-1/2} (V + B) (V + z_1 z_1^t)^{-1/2}\), the maximum value of the likelihood function under model \(H_1\) is given by

\[
cN^{-Np/2} (1 + z_1^t V^{-1} z_1)|V| \prod_{j=r+1}^p \lambda_j^{(1)}
\]

where \(\lambda_1^{(1)} > \cdots > \lambda_p^{(1)}\) are the characteristic roots of \((V + z_1 z_1^t)^{-1} (V + B)\). Hence we have the following
\[\begin{align*}
&= (\bar{x}_i - \bar{x}_0)^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t \left[ I - B \hat{\mathbf{A}} \left( I + \hat{\mathbf{A}}^t B \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}^t \right] V \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}_0) \\
&= (\bar{x}_i - \bar{x}_0)^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t V \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}_0) \\
&= (\bar{x}_i - \bar{x}_0)^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}_0), \text{ since } \hat{\mathbf{C}}^t V \hat{\mathbf{C}} = I_r,
\end{align*}\]

Let

\[\begin{align*}
z_i &= \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}), \quad i = 0, 1, \ldots, k \\
a_i &= (\bar{x}_i - \bar{x})^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x})
\end{align*}\]

Then

\[\begin{align*}
&(\bar{x}_i - \bar{x}_0)^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}_0) \\
&= z_i^t z_0 - 2 \left[ z_i^t z_0 - \frac{1}{2} a_i \right].
\end{align*}\]

Hence, we classify \(\Pi_0\) into \(\Pi_i\) if

\[z_i^t z_0 - \frac{1}{2} a_i = \max_{1 \leq j \leq k} \left( z_j^t z_0 - \frac{1}{2} a_j \right),\]

giving a linear discriminant rule. See Srivastava and Carter (1983, p. 248) for an example.

The sample Mahalanobis squared distance may be weighted by sample sizes. In this case, we may use

\[d_i^2 = \frac{N_i + N_0}{N_i N_0} (\bar{x}_i - \bar{x}_0)^t \hat{\mathbf{C}} \hat{\mathbf{C}}^t (\bar{x}_i - \bar{x}_0).
\]

for discrimination. Thus, we would classify \(\Pi_0\) into \(\Pi_i\) if \(i\) is the smallest integer for which the minimum of \(d_j^2\) is attained over \(j = 1, 2, \ldots, k\).

4.2 Graphical Method

Let \(y_i = \hat{\mathbf{C}}^t \mu_i\), \(i = 0, 1, \ldots, k\), and \(\hat{\mathbf{C}} = (c_1, \ldots, c_r)\). Then the \(j^{th}\) component of \(y_i\) is given by \(c_j \mu_i\), and is called the \(j^{th}\) canonical coordinate. It is a linear combination of the estimated means of the \(i^{th}\) group with the property that it maximizes the
References


