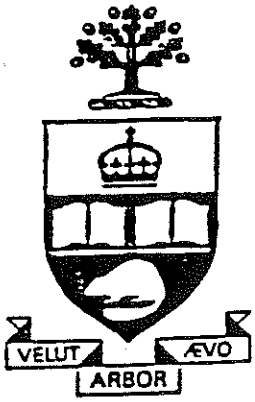

Reduced Rank Discrimination

by

M.S. Srivastava
Department of Statistics
University of Toronto

Technical Report No. 9510, June 28, (1995)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

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 μ_i and common known covariance matrix Σ. 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 (1979) 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

$i = 0, 1, \cdots, k$ and $V = nS \sim W_p(\Sigma, n)$. If the matrix μ 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

$$\mu = FG = F(g_0, \cdots, g_k),$$

where g_i's are r-vectors. Let

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

Then μ can be rewritten as

$$\begin{aligned} \mu &= F\bar{g}1' + F(g_0 - \bar{g}, \cdots, g_k - \bar{g}) \\ &= \theta_0 1' + F\xi, \qquad \sum_{i=0}^{k} N_i \xi_i = 0, \end{aligned} \tag{2.1}$$

where $\xi = (\xi_0, \cdots, \xi_k) = (g_0 - \bar{g}, \cdots, g_k - \bar{g})$, and $\theta_0 = F\bar{g}$. The problem of testing that μ is of rank $r < p$ started from Fisher (1938, 1939) and a solution when Σ is known is given in Rao (1973, pp. 556-559). When the covariance matrix Σ is unknown, the problem of estimating F, ξ, Σ 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

$$P^t F = 0, \tag{2.2}$$

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

$$P^t \Sigma P = I_{p-r}, \tag{2.3}$$

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, Σ), these estimates must also satisfy (2.2) and (2.3). As will be apparent later (see (3.4)), this implies that

$$\hat{P}^t (V + B)\hat{P} = N I_{p-r}, \quad \hat{P}^t \hat{F} = 0, \quad \hat{P}^t \hat{\Sigma} \hat{P} = I_{p-r} \tag{2.4}$$

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

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

Differentiating it partially with respect to θ_0 and equating it to zero gives the MLE of θ_0,

$$\begin{aligned} \hat{\theta}_0 &= N^{-1}\Sigma N_i \bar{x}_i - F(\Sigma N_i \xi_i / N) \\ &\equiv \bar{x}, \qquad \text{since} \quad \Sigma N_i \xi_i = 0. \end{aligned} \tag{3.1}$$

Let

$$y_i = \bar{x}_i - \bar{x}, \qquad i = 0, 1, \cdots, k. \tag{3.2}$$

Then differentiating the log-likelihood partially with respect to ξ_i and equating it to zero gives the MLE of ξ_i,

$$\hat{\xi}_i = (F^t \hat{\Sigma}^{-1} F)^{-1} F^t \hat{\Sigma}^{-1} y_i, \qquad i = 0, 1, \cdots, k.$$

Similarly, the MLE of Σ 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

$$\begin{aligned} B &= \sum_{i=0}^{k} N_i y_i y_i^t \\ &= \sum_{i=0}^{k} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t \end{aligned}$$

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

$$NF^t V^{-1} = F^t \hat{\Sigma}^{-1}, \qquad \text{and} \qquad NF^t V^{-1} F = F^t \hat{\Sigma}^{-1} F.$$

Hence,

$$\hat{\xi}_i = (F^t V^{-1} F)^{-1} F^t V^{-1} y_i, \tag{3.3}$$

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),

$$\begin{aligned} V\hat{A}\hat{A}^t B \hat{A}\hat{A}^t V &= \left(I - V\hat{C}\hat{C}^t \right) B \hat{A}\hat{A}^t V \\ &= B\hat{A}\hat{A}^t V. \end{aligned}$$

Thus, the MLE of Σ is given by

$$\begin{aligned} N\hat{\Sigma} &= V + V\hat{A}\hat{A}^t B \hat{A}\hat{A}^t V \\ &= V + B\hat{A}\hat{A}^t V \\ &= \left[I + B\hat{A}\hat{A}^t \right] V. \end{aligned} \tag{3.8}$$

Since

$$\begin{aligned} \hat{F}\hat{\xi}_i &= VV^{-1}\hat{F}(\hat{F}^t V^{-1}\hat{F})^{-1}\hat{F}^t V^{-1}(\bar{x}_i - \bar{x}) \\ &= V[V^{-1} - \hat{P}(\hat{P}^t V \hat{P}^t)^{-1}\hat{P}](\bar{x}_i - \bar{x}) \\ &= [I - V\hat{A}\hat{A}^t](\bar{x}_i - \bar{x}) \\ &= V\hat{C}\hat{C}^t (\bar{x}_i - \bar{x}), \end{aligned}$$

$$\begin{aligned} \hat{\mu}_i &= \bar{x} + V\hat{C}\hat{C}^t(\bar{x}_i - \bar{x}) \\ &= \bar{x} + [I - V\hat{A}\hat{A}^t](\bar{x}_i - \bar{x}) \end{aligned} \tag{3.9}$$

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

$$U_{r+1} = [n - \frac{1}{2}(k - p - 1) \sum_{i=1}^{r} l_i^{-2}][\sum_{i=r+1}^{p} \log(1 + l_i)], \tag{3.10}$$

Letting

$$y_i = x_i - \bar{x}, \qquad i = 0, 1, \cdots, k,$$

the MLEs of $\xi_1, \xi_2, \cdots, \xi_k$ under model H_1, are given by

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

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

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

$$\begin{aligned} & N_0 (y_0 - F\hat{\xi}_1)^t \hat{\Sigma}^{-1} (y_0 - F\hat{\xi}_1) \\ = \; & N_0 (y_0 - Ly_0)^t \hat{\Sigma}^{-1} (y_0 - Ly_0) \\ & + \frac{2N_0 N_1}{N_0 + N_1} (y_0 - Ly_0)^t \hat{\Sigma}^{-1} L (y_0 - y_1) + \frac{N_0 N_1^2}{(N_0 + N_1)^2} (y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L (y_0 - y_1) \end{aligned}$$

and

$$\begin{aligned} & N_1 (y_1 - F\hat{\xi}_1)^t \hat{\Sigma}^{-1} (y_1 - F\hat{\xi}_1) \\ = \; & N_1 (y_1 - Ly_1)^t \hat{\Sigma}^{-1} (y_1 - Ly_1) \\ & - \frac{2N_0 N_1}{N_0 - N_1} (y_1 - Ly_1)^t \hat{\Sigma}^{-1} L (y_0 - y_1) + \frac{N_0^2 N_1}{(N_0 + N_1)^2} (y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L (y_0 - y_1) \end{aligned}$$

Adding the two terms, we get

$$\begin{aligned} & \sum_{i=0}^{1} N_i (y_i - F\hat{\xi}_1)^t \hat{\Sigma}^{-1} (y_i - F\hat{\xi}_1) \\ = \; & \sum_{i=0}^{1} N_i (y_i - Ly_i)^t \hat{\Sigma}^{-1} (y_i - Ly_i) + \frac{N_0 N_1}{N_0 + N_1} (y_0 - y_1)^t L^t \hat{\Sigma}^{-1} L (y_0 - y_1), \end{aligned}$$

Since $(I-L_1)^t W_1^{-1}(I-L_1) = W_1^{-1}(I-L_1)$, we need to minimize

$$
\begin{aligned}
& |W_1 + (I-L_1)(B, -z_1 z_1^t) \begin{pmatrix} (I-L_1)^t \\ I \end{pmatrix}| \\
= \; & |W_1||I + (B, -z_1 z_1^t) \begin{pmatrix} (I-L_1)^t \\ I \end{pmatrix} W_1^{-1}(I-L_1)| \\
= \; & |W_1||I + (B, -z_1 z_1^t) \begin{pmatrix} W_1^{-1}(I-L_1) \\ W_1^{-1}(I-L_1) \end{pmatrix}| \\
= \; & |W_1||I + (B - z_1 z_1^t) W_1^{-1}(I-L_1)| \\
= \; & |W_1||I + (B - z_1 z_1^t) P (P^t W_1 P)^{-1} P^t| \\
= \; & |W_1||P^t W_1 P|^{-1} |P^t W_1 P + P^t (B - z_1 z_1^t) P|,
\end{aligned}
$$

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,

$$
\begin{aligned}
& |W_1||P^t W_1 P|^{-1} |P^t V P + P^t B P| \\
= \; & |W_1||(P^t W_1 P)^{-1} P^t (V+B) P| \\
= \; & |W_1||(P^t W_1 P)^{-1/2} P^t (V+B) P (P^t W_1 P)^{-1/2}| \\
= \; & |W_1||A_1^t (V+B) A_1|,
\end{aligned}
$$

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 $W_1^{\frac{1}{2}} A_1$ 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

$$
c N^{-Np/2} (1 + z_1^t V^{-1} z_1) |V| \prod_{j=r+1}^{p} \lambda_j^{(1)}
$$

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

$$
\begin{aligned}
&= (\bar{x}_i - \bar{x}_0)^t \hat{C}\hat{C}^t \left[I - B\hat{A}\left(I + \hat{A}^t B \hat{A} \right)^{-1} \hat{A}^t \right] V \hat{C}\hat{C}^t (\bar{x}_i - \bar{x}_0) \\
&= (\bar{x}_i - \bar{x}_0)^t \hat{C}\hat{C}^t V \hat{C}\hat{C}^t (\bar{x}_i - \bar{x}_0) \\
&= (\bar{x}_i - \bar{x}_0)^t \hat{C}\hat{C}^t (\bar{x}_i - \bar{x}_0), \quad \text{since } \hat{C}^t V \hat{C} = I_r,
\end{aligned}
$$

Let

$$
\begin{aligned}
z_i &= \hat{C}^t (\bar{x}_i - \bar{x}), \quad i = 0, 1, \ldots, k \\
a_i &= (\bar{x}_i - \bar{x})^t \hat{C}\hat{C}^t (\bar{x}_i - \bar{x})
\end{aligned}
$$

Then

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

Hence, we classify Π_0 into Π_i if

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

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

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

for discrimination. Thus, we would classify Π_0 into Π_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 $y_i = \hat{C}^t \hat{\mu}_i$, $i = 0, 1, \ldots, k$, and $\hat{C} = (c_1, \ldots, c_r)$. Then the j^{th} component of y_i is given by $c_j^t \hat{\mu}_i$, 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

1. Anderson,T.W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* 22, 327-351.

2. Bartlett,M.S. (1947) Test of significance in canonical analysis. *Biometrika* 46, 59-66.

3. Cambell,N.A. (1984) Canonical variable analysis - a general model formulation. *Australian J. of Statist.* 26, 86-96.

4. Fisher,R.A. (1936) The statistical utilization of multiple measurements. *Ann. Eugenics.* 8, 376-386.

5. Fisher,R.A. (1939) The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugenics.* 9, 238-249.

6. Fujikoshi,Y. (1974) The likelihood ratio test for the dimensionality of regression coefficients. *J. Multi. Anal.* 4, 327-340.

7. Fujikoshi,Y. (1977) Asymptotic expansions for the distribution of some multivariate tests. *Multivariate Analysis* IV (P.R. Krishnaiah, Ed.) pp 55-70. North-Holland, Amsterdam.

8. Hastie,T. and Tibshirani,R. (1994) Discriminant analysis by Gaussian Mixtures. *Tech. Report. No. 9401, University of Toronto.*

9. Rao,C.R. (1973) *Linear statistical inference and its applications.* 2nd. Edition. Wiley, New York.

10. Rao,C.R. (1985) Tests for dimensionality and interactions of mean vectors under general and reducible covariance structures. *J. Multi. Anal.* 16, 173-184.

11. Srivastava,M.S. (1967) Classification into multivariate normal populations when the population means are linearly restricted. *Ann. Inst. Statist. Math.* 19, 473-478.

12. Srivastava,M.S. and Khatri,C.G. (1979) *An introduction to multivariate statistics.* North-Holland, New York.

13. Srivastava,M.S. and Carter, E.M. (1983) *An introduction to applied multivariate statistics.* North-Holland, New York.