Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 97 (2006) 2057–2070

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Minimum distance classification rules for high dimensional data

Muni S. Srivastava

Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3

Received 10 April 2005
Available online 10 July 2006

Abstract

In this article, the problem of classifying a new observation vector into one of the two known groups $\Pi_i, i = 1, 2$, distributed as multivariate normal with common covariance matrix is considered. The total number of observation vectors from the two groups is, however, less than the dimension of the observation vectors. A sample-squared distance between the two groups, using Moore–Penrose inverse, is introduced. A classification rule based on the minimum distance is proposed to classify an observation vector into two or several groups. An expression for the error of misclassification when there are only two groups is derived for large p and $n = O(p^{\delta}), 0 < \delta < 1$.

AMS 2000 subject classification: primary 62H30; secondary 62H10

Keywords: Fisher discriminant rule; Misclassification error; Moore–Penrose inverse; Multivariate normal; Singular Wishart

1. Introduction

In this article, we consider the problem of classifying a new observation vector $\boldsymbol{x}_0$ of dimension p into one of the two known groups Π_1 and Π_2. It is assumed that independent observation vectors $\boldsymbol{x}_{ij},\ j = 1, \dots, N_i,\ i = 1, 2$, are available from the two groups. We shall assume that $\boldsymbol{x}_{ij}$ are independently distributed as multivariate normal with mean vectors $\boldsymbol{\mu}_i,\ i = 1, 2$, and common $p \times p$ positive definite covariance matrix Σ. The mean vectors $\boldsymbol{\mu}_i,\ i = 1, 2$, and the covariance matrix Σ are assumed unknown and are estimated by the sample mean vectors $\bar{\boldsymbol{x}}_i$ and the pooled

E-mail address: srivasta@utstat.toronto.edu.

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.09.014

sample covariance matrix S given, respectively, by

$$\bar{\boldsymbol{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}, \quad i = 1, 2, \tag{1.1}$$

$$S = n^{-1} V = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)', \tag{1.2}$$

$$n = N_1 + N_2 - 2. \tag{1.3}$$

Beginning with the seminal work of Fisher [4] and Wald [15] in the known parameter case, this problem has been considered many times in the statistical literature with parameter known or unknown when $n > p$, see for example, Kiefer and Schwartz [6] for the admissibility of the maximum likelihood ratio (MLR) rule, Srivastava [11] for the admissibility of the MLR rule in linear models, and DasGupta [2] for the monotonicity of the errors of misclassification for many rules including the MLR rule. The MLR procedure when $n > p$ classifies $\boldsymbol{x}_0$ into the group Π_1, if

$$\begin{aligned}&(1 + N_1^{-1})^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1)' S^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1)\\ &\quad < (1 + N_2^{-1})^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)' S^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2).\end{aligned} \tag{1.4}$$

Otherwise, it is classified into the group Π_2. The probability of misclassifying $\boldsymbol{x}_0$ into group Π_2 when it actually belongs to Π_1 is called an error of misclassification and denoted by e_1. Similarly, the error in misclassifying $\boldsymbol{x}_0$ into group Π_1 when it actually belongs to Π_2 will be denoted by e_2. It is difficult to obtain an explicit expression for e_1 or e_2. But when the classification is carried out without the factor $(1 + N_1^{-1})^{-1}$ on the left side of (1.4) and $(1 + N_2^{-1})^{-1}$ on the right side of (1.4), Okamoto [9] gave an asymptotic expression for e_1 and e_2. These expressions are obtained when $n \to \infty$ and p is fixed $p < n$. Asymptotic expressions for e_i when n and p both go to infinity such that $\frac{p}{n} \to c$, $0 \leqslant c < 1$ has also been considered in the literature, see for example, Saranadasa [10] and Fujikoshi [5] among others.

For $n > p$, the classification rule (1.4) may also be considered as minimum distance rule. Properties such as invariance under a linear transformation by a $p \times p$ nonsingular matrix A holds. However, when $n < p$, there does not exist an invariant statistic as the nonsingular linear transformation group acts transitively on the sample space since the covariance matrix Σ is assumed positive definite, see Lehmann [7, p. 318, Problem 24 (ii)]. Thus, any classification rule that may be proposed for the case $n < p$ will not be invariant under nonsingular linear transformations. A rule that is invariant under a linear transformation by an orthogonal matrix has been proposed by Saranadasa [10]. According to this rule $\boldsymbol{x}_0$ is classified into the group Π_1, if

$$(1 + N_1^{-1})^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1)' (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1) < (1 + N_2^{-1})^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)' (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2). \tag{1.5}$$

Otherwise, it is classified into the group Π_2. Saranadasa derived an asymptotic expression for the errors of misclassification as $p \to \infty$, for the classification rule (1.5). However, the procedure in (1.5) ignores the information available from S.

The focus of this paper is to propose a classification procedure that utilizes the information available in S. In order to use the information available in the singular sample covariance matrix S, we define a sample distance between the observation vector $\boldsymbol{x}_0$ and the group Π_i. We use the Moore–Penrose inverse of S, where the Moore–Penrose inverse of a matrix A is defined by A^+ satisfying the following four properties: (i) $AA^+A = A$, (ii) $A^+AA^+ = A^+$, (iii) $(A^+A)' = A^+A$,

(iv) $(AA^+)' = AA^+$. The Moore–Penrose inverse is unique. The sample covariance matrix S can be written as

$$S = H'LH, \tag{1.6}$$

where $H : n \times p, \ HH' = I_n$, the $n \times n$ identity matrix and L is an $n \times n$ diagonal matrix with the diagonal elements as the n nonzero eigenvalues $l_1, \dots, l_n$ of the $p \times p$ matrix S. The Moore–Penrose inverse of S is defined by

$$S^+ = H'L^{-1}H. \tag{1.7}$$

We define the sample distance between $\boldsymbol{x}_0$ and the group Π_i by

$$D_i^{+2} = (1 + N_i^{-1})^{-1}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'S^+(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i), \quad i = 1, 2. \tag{1.8}$$

We propose the classification rule as classifying $\boldsymbol{x}_0$ into the group Π_1, if

$$D_1^{+2} < D_2^{+2}. \tag{1.9}$$

Otherwise, we classify $\boldsymbol{x}_0$ into the group Π_2.

It may be noted that the sample covariance matrix S has many small and near zero eigenvalues, when p is large, even when $n \geqslant p$. That is, even if the inverse of the sample covariance matrix S exists, at least theoretically, the classification rules such as MLR rule do not perform well due to some near zero eigenvalues, as shown in many examples by Dudoit et al. [3]. Thus, it is proposed to drop zero or near zero eigenvalues in both the cases when $n < p$ and when $n \geqslant p$. Thus, in practice, we may not use all the n column vectors of H', but only $r \leqslant n$ of them corresponding to the retained eigenvalues of S after deleting zero and near zero eigenvalues, and define

$$D_i^{+2} = (1 + N_i^{-1})^{-1}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'H_1'L_1^{-1}H_1(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i), \quad i = 1, 2, \tag{1.10}$$

where $H' = (H_1', H_2'), \ H_1' : p \times r, \ r \leqslant n, \ H_1H_1' = I_r$, and L_1 is an $r \times r$ diagonal matrix consisting of only the retained largest eigenvalues of S. An illustrative example is given in Section 2.2.

We generalize the above results for classifying a new observation vector $\boldsymbol{x}_0$ into several groups, when they have common covariance matrix as well as when they are different. This is done in Section 2. In Section 3, an asymptotic expansion of the errors of misclassification is given as p goes to infinity and $n = O(p^\delta), \ 0 < \delta < 1$, for the case of classifying a new observation vector $\boldsymbol{x}_0$ into one of the two groups.

2. Classifying an individual into several populations

In this section, we consider the problem of classifying an individual with observations on p characteristics into one of several populations. We first consider in Section 2.1, the case when all the populations have a common covariance matrix which is estimated by pooling the observations from all the populations. Then in Section 2.3, we consider the case when the population covariances are unequal.

2.1. Classification when the population covariances are equal

Let $\bar{\boldsymbol{x}}_i$ be the sample mean vector of the ith population Π_i from which N_i independent observations have been obtained, $i = 1, \dots, k$. Let S be the pooled estimate of the covariance matrix Σ based on $f = \sum_{i=1}^k (N_i - 1)$ degrees of freedom where $f < p$. We have an observation vector $\boldsymbol{x}_0$ on an individual from a population Π_0 and wish to classify the observation (i.e., the individual) into one of the k populations $\Pi_i,\ i=1,\dots,k$. The sample (squared) distance between Π_0 and Π_i is given by

$$\tilde{D}_i^{+2} = (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)' S^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i), \quad i = 1, \dots, k, \tag{2.1}$$

where S^+ is the Moore–Penrose inverse of S. Here the multiplying factor $(1+N_i^{-1})^{-1}$ has been dropped so that it can be connected with the canonical variables method defined later. Thus, according to the minimum distance rule, the observation vector $\boldsymbol{x}_0$ from Π_0 is classified into Π_i if and only if

$$\tilde{D}_i^{+2} = \min_{1 \leqslant j \leqslant k} \tilde{D}_j^{+2}. \tag{2.2}$$

The above classification rule is equivalent to the classification rule based on the canonical variables $\boldsymbol{a}_l'(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}),\ l = 1, \dots, m,\ m = \min(k-1, f)$, where $A = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_m)$ is chosen to be the matrix of m eigenvectors corresponding to the m nonzero eigenvalues of S^+B subject to the condition that $A'SA = I_m$, B is the matrix of between mean sum of squares given by

$$B = \sum_{i=1}^{k} N_i (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})', \tag{2.3}$$

and

$$\bar{\boldsymbol{x}} = \sum_{i=1}^{k} N_i \bar{\boldsymbol{x}}_i \Big/ \sum_{i=1}^{k} N_i. \tag{2.4}$$

The advantage of such a method is that the canonical variables can be plotted two at a time for each group or population including the population to be classified. The relative position of the population to be classified in comparison to other populations will indicate to which population it is closest. Furthermore, often only a few canonical variables are important as the canonical variables corresponding to smaller eigenvalues do not have much discriminating power and so only one or two canonical variables may suffice for classification or discrimination of an individual into one of k populations. To show the equivalence of the two procedures, we start with the sample-squared distance function $\tilde{D}_i^{+2}$. Write

$$S = H'LH, \quad S^+ = H'L^{-1}H, \quad HH' = I_f, \tag{2.5}$$

$$L = diag(l_1, \dots, l_f), \quad l_1 > \cdots > l_f. \tag{2.6}$$

Consider the symmetric matrix

$$L^{-\frac{1}{2}} HBH' L^{-\frac{1}{2}}. \tag{2.7}$$

There exists an $f \times f$ orthogonal matrix Γ such that

$$\Gamma L^{-\frac{1}{2}} HBH' L^{-\frac{1}{2}} \Gamma' = \begin{pmatrix} D_m & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.8}$$

where $\Gamma'\Gamma = I_f$, and D_m is the diagonal matrix of the ordered eigenvalues of the matrix $L^{-\frac{1}{2}} HBH' L^{-\frac{1}{2}}$. Let $\Gamma' = (\Gamma_1', \Gamma_2')$, where $\Gamma_1 : m \times f$. Then $\Gamma_1 \Gamma_1' = I_m$. Let

$$\begin{aligned} P = \Gamma L^{-\frac{1}{2}} H &= \begin{pmatrix} \Gamma_1 L^{-\frac{1}{2}} H \\ \Gamma_2 L^{-\frac{1}{2}} H \end{pmatrix} \\ &= \begin{pmatrix} F \\ C \end{pmatrix}. \end{aligned} \tag{2.9}$$

Then

$$\begin{aligned} P'P &= H'L^{-\frac{1}{2}}\Gamma'\Gamma L^{-\frac{1}{2}}H = H'L^{-1}H \\ &= S^{+} = F'F + C'C, \end{aligned} \tag{2.10}$$

and

$$PBP' = \begin{pmatrix} D_m & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.11}$$

Hence,

$$FBF' = D_m, \quad CBC' = 0. \tag{2.12}$$

Also,

$$\begin{aligned} FSF' &= \Gamma_1 L^{-\frac{1}{2}} HSH' L^{-\frac{1}{2}} \Gamma_1' \\ &= \Gamma_1 \Gamma_1' = I_m. \end{aligned} \tag{2.13}$$

Thus, F' corresponds to A mentioned earlier. Furthermore we have,

$$\begin{aligned} 0 &= CBC' \\ &= \sum_{i=1}^{k} N_i C(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})'C'. \end{aligned} \tag{2.14}$$

Since each term in the summation is positive semidefinite, it follows that each term must be zero. Thus,

$$C\bar{\boldsymbol{x}}_i = C\bar{\boldsymbol{x}}, \quad i = 1, \ldots, k.$$

Hence, from (2.10) and (2.14), we get $\tilde{D}_i^{+2}$ given by

$$\begin{aligned} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'S^{+}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i) &= (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'[F'F + C'C](\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i) \\ &= (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'F'F(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i) + (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'C'C(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i) \\ &= (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'F'F(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i) + (\boldsymbol{x}_0 - \bar{\boldsymbol{x}})C'C(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}) \\ &= \sum_{j=1}^{m} [\boldsymbol{a}_j'(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)]^2 + (\boldsymbol{x}_0 - \bar{\boldsymbol{x}})C'C(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}). \end{aligned} \tag{2.15}$$

The second term in the last expression does not depend on the values of the *i*th population and hence has no discriminating power. Thus, the classification rule based on canonical variables is equivalent to the one based on the minimum distance rule. Although no expression for the errors of misclassification is available, estimates can be obtained by using methods given in Srivastava [12, pp. 250–251].

Remark 2.1. It may be noted that the $\tilde{D}_i^{+2}$ in (2.1) can also be defined by weighting it with $(1 + N_i^{-1})^{-1}$, namely,

$$D_i^{+2}(w_i) = (1 + N_i^{-1})^{-1}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)'S^{+}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i), \tag{2.16}$$

and the classification procedure (2.2) can be modified with this $D_i^{+2}(w_i)$.

2.2. An example

Wilbur et al. [16] analyzed the data in Nakatsu et al. [8] on soil DNA finger prints. Although, initially it had data on 10,000 finger prints on four groups consisting of 23, 22, 22, and 22 observations, they selected only 84 finger prints. Assuming that all the four groups have the same common covariance matrix Σ, the degrees of freedom available to estimate the unknown covariance matrix Σ is $89 - 4 = 85$, which is larger than 84, the number of finger prints. Thus, theoretically a positive definite estimate of Σ exists with probability one. However, the Fisher's linear discriminant rule or the minimum distance rule did not perform well since 24 eigenvalues of the sample covariance matrix are very close to zero. Thus, it would be desirable to drop some of the eigenvalues and eigenvectors from considerations. We define a quantity c_i called the ratio of the cummulative sum of the sample ordered eigenvalues (from highest to lowest) up to the ith eigenvalue divided by the sum of all the eigenvalues of the sample covariances matrix. For the example on finger prints data,

$$c_{35} = \left(\sum_{j=1}^{35} l_i \Big/ \sum_{j=1}^{84} l_i \right) = 90.23\%, \quad c_{60} = \left(\sum_{j=1}^{60} l_i \Big/ \sum_{j=1}^{84} l_i \right) = 99.99\%,$$

$$c_{80} = \left(\sum_{j=1}^{80} l_i \Big/ \sum_{j=1}^{84} l_i \right) = 100\%.$$

While Wilbur et al. [16] proposed two methods to reduce further from 84 finger prints to a smaller number to be used for analysis, we applied the method given above by considering 35, 60 and 80 eigenvalues, respectively, and compared the correct classification rates with their classification rules. The correct classification rates are obtained by leave-one-out cross validation method described, say, for example in Srivastava [12, pp. 322]. The results are shown in the following table (Table 1). It shows that a selection of 60 eigenvalues gives a total error rate of 5% as opposed to the best error rate of 12% obtained by Wilbur et al. [16]. It has also been found that by plotting two components at a time, namely $\boldsymbol{a}_1'(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})$, $\boldsymbol{a}_2'(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})$, $j = 1, \ldots, N_i$, $i = 1, \ldots, k$, see (2.15) the selection of 35 components do not provide a good separation between the four group while 60 components provide a good separation of the groups. These graphs can be obtained from the author.

2.3. Classification when the population covariances are unequal

When the covariances are unequal, we calculate the sample covariances S_i, $n_i S_i = V_i = \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'$, $n_i = (N_i - 1)$, $i = 1, \ldots, k$ and write

$$S_i = H_i' L_i H_i, \quad \text{and} \quad S_i^+ = H_i' L_i^{-1} H_i, \tag{2.17}$$

where $H_i H_i' = I_{n_i}$, and $L_i = diag(l_{i1}, \ldots, l_{in_i})$. We define the sample-squared distance between Π_0 and Π_i by

$$D_{ii}^{+2} = (1 + N_i^{-1})^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i)' S_i^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_i), \tag{2.18}$$

and use the minimum distance rule. That is, we classify $\boldsymbol{x}_0$ into Π_i if and only if

$$D_{ii}^{+2} = \min_{1 \leqslant j \leqslant k} D_{jj}^{+2}. \tag{2.19}$$

Table 1
Number of correctly classified samples in the cross-validations explained by PC and Wilbur et al. method

Proposed method						
Principal components	$c_i(\%)$	Treatment				Total
		1	2	3	4	
35	90.23	22	21	20	19	82
60	99.99	23	21	21	20	85
80	100	14	21	13	7	46
Wilbur et al. (2002) method						
Bernoulli	Multivariate	23	23	14	9	68
Logistic	Multivariate	20	22	14	9	65
Bernoulli	Univariate	22	18	22	17	79
Logistic	Univariate	22	18	18	14	72

3. Evaluation of misclassification errors: two groups case

To carry out the procedures described in Sections 1 and 2, the assumption of normality is not needed. However, to derive an expression for the errors of misclassification, the normality assumption is needed and we shall assume so. Under this assumption, we evaluate the probability of misclassifying an individual from group Π_1 (into Π_2), denoted by e_1 and called a misclassification error. That is, we evaluate

$$e_1 = P\{a(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1)' S^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1) > (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)' S^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2) \mid \boldsymbol{x}_0 \in \Pi_1\}, \tag{3.1}$$

where

$$a = (1 + N_1^{-1})^{-1}(1 + N_2^{-1}). \tag{3.2}$$

Similarly, the error e_2 of misclassifying an individual from Π_2 (into Π_1) is given by

$$e_2 = P\{a(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1)' S^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1) < (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)' S^+ (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2) \mid \boldsymbol{x}_0 \in \Pi_2\}. \tag{3.3}$$

And $\boldsymbol{x}_0, \ \bar{\boldsymbol{x}}_1, \ \bar{\boldsymbol{x}}_2$, and S are independently distributed as $\bar{\boldsymbol{x}}_i \sim N_p(\boldsymbol{\mu}_i, N_i^{-1}\Sigma), \ i = 1, 2, \ V = nS \sim W_p(\Sigma, n)$, and $\boldsymbol{x}_0 \sim N_p(\boldsymbol{\mu}_0, \Sigma), \ \boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, depending on which population it comes from. Since the classification rule defined in (1.9) is invariant under the orthogonal transformations: $\boldsymbol{x}_i \to G\boldsymbol{x}_i, \ S \to GSG'$, where $GG' = I_p$, we may assume without any loss of generality that the $p \times p$ positive definite matrix Σ is a diagonal matrix, given by

$$\Sigma = \Lambda = diag(\lambda_1, \ldots, \lambda_p). \tag{3.4}$$

Although, the errors of misclassification can be evaluated for a general classification rule, as in Srivastava and Khatri [14, pp. 246], in which a can take any positive real number, we shall confine to the case when $a = (1 + N_1^{-1})^{-1}(1 + N_2^{-1})$ as given in (3.2). We only evaluate e_1, as the calculation of e_2 is similar. The evaluation is, however, done for the case when the difference in the two mean vectors is of the order $n^{-\frac{1}{2}}$, that is,

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = n^{-\frac{1}{2}}\boldsymbol{\delta}, \tag{3.5}$$

where $\boldsymbol{\delta}$ is a nonnull vector of constants. Let

$$\begin{aligned} k_1^2 &= 2[(1 + N_2^{-1}) - a^{\frac{1}{2}}], \\ k_2^2 &= 2[(1 + N_2^{-1}) + a^{\frac{1}{2}}], \end{aligned} \tag{3.6}$$

and

$$\mathbf{u_1} = k_1^{-1}[a^{\frac{1}{2}}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1) - (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)],$$
$$\mathbf{u_2} = k_2^{-1}[a^{\frac{1}{2}}(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_1) + (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_2)]. \tag{3.7}$$

Then, when $\boldsymbol{x}_0 \in \Pi_1$

$$E(\mathbf{u_1}) = -k_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -k_1^{-1}\frac{\boldsymbol{\delta}}{\sqrt{n}},$$
$$E(\mathbf{u_2}) = k_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = k_2^{-1}\frac{\boldsymbol{\delta}}{\sqrt{n}},$$
$$\mathrm{Cov}(\mathbf{u}_1) = \Lambda, \quad \mathrm{Cov}(\mathbf{u}_2) = \Lambda, \quad \mathrm{Cov}(\mathbf{u}_1, \mathbf{u}_2) = 0. \tag{3.8}$$

Hence, when $\boldsymbol{x}_0 \in \Pi_1$

$$\begin{pmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{pmatrix} \sim N_{2p}\left[\begin{pmatrix} -k_1^{-1}\frac{\boldsymbol{\delta}}{\sqrt{n}} \\ k_2^{-1}\frac{\boldsymbol{\delta}}{\sqrt{n}} \end{pmatrix}, \begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda \end{pmatrix}\right] \tag{3.9}$$

and

$$e_1 = P\{\mathbf{u}_1' S^+ \mathbf{u}_2 > 0\}, \tag{3.10}$$

where $(\mathbf{u}_1', \mathbf{u}_2')'$ is distributed as normal, stated above, and is independently distributed of S. Thus, letting

$$A = (H\Lambda H')^{-\frac{1}{2}}, \tag{3.11}$$

we get

$$e_1 = P\{\mathbf{u_1'} H' L^{-1} H \mathbf{u_2'} > 0\}$$
$$= P\{\mathbf{u_1'} H' A (ALA)^{-1} A H \mathbf{u_2} > 0\}. \tag{3.12}$$

Note that given H,

$$AH\mathbf{u_i} \sim N_n((-1)^i n^{-\frac{1}{2}} k_i^{-1} AH\boldsymbol{\delta}_i, I_n), \quad i = 1, 2, \tag{3.13}$$

are independently distributed. Let Γ be an orthogonal matrix whose first row is $\frac{\boldsymbol{\delta}' H' A}{(\boldsymbol{\delta}' H' A^2 H \boldsymbol{\delta})^{\frac{1}{2}}}$. Then

$$e_1 = P\{\mathbf{u_1'} H' A \Gamma' \Gamma (ALA)^{-1} \Gamma' \Gamma AH \mathbf{u_2} > \mathbf{0}\}$$
$$= P\{\boldsymbol{w}_1' \Gamma (A\widetilde{L}A)^{-1} \Gamma' \boldsymbol{w}_2 > 0\}, \tag{3.14}$$

where $\boldsymbol{w}_i = \Gamma AH\mathbf{u}_i$, $\widetilde{L} = diag(\tilde{l}_1, \ldots, \tilde{l}_n)$, and $\tilde{l}_i$ are the eigenvalues of $V = nS$, $\tilde{l}_i = nl_i$. Given H, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are independently distributed as multivariate normal with covariances as the $n \times n$ identity matrix I_n and the mean vectors given by

$$E(\boldsymbol{w}_1 \mid H) = -\begin{pmatrix} k_1^{-1}\eta_n \\ \mathbf{0} \end{pmatrix},$$
$$E(\boldsymbol{w}_2 \mid H) = \begin{pmatrix} k_2^{-1}\eta_n \\ \mathbf{0} \end{pmatrix}, \tag{3.15}$$

where

$$\eta_n = (n^{-1}\boldsymbol{\delta}' H' A^2 H \boldsymbol{\delta})^{\frac{1}{2}}. \tag{3.16}$$

So far the results are exact. We now evaluate e_1 asymptotically as $p,\ n \to \infty$ and the difference in the mean vectors is assumed to be of the order $n^{-\frac{1}{2}}$, as given in (3.5). We also assume that

$$\begin{aligned}
&\text{(i) } 0 < a_{i0} = \lim_{p\to\infty} a_i < \infty,\\
&\text{(ii) } \theta_0^2 = \lim_{p\to\infty} \frac{\boldsymbol{\delta}'\Lambda\boldsymbol{\delta}}{pa_2} < \infty,
\end{aligned} \tag{3.17}$$

where $a_i = \frac{\mathrm{tr}\Sigma^i}{p}$, $i = 1, \ldots, 4$, and $\boldsymbol{\delta}$ is a nonnull vector of constants. Assumption (i) is needed to prove Lemma A.1 given in the Appendix.

From Lemma A.1, we get in probability

$$\lim_{n\to\infty}\lim_{p\to\infty}\left(\frac{\boldsymbol{\delta}' H' A^2 H\boldsymbol{\delta}}{n} - \frac{\boldsymbol{\delta}'\Lambda\boldsymbol{\delta}}{pa_2}\right) = 0, \tag{3.18}$$

and

$$\lim_{p\to\infty}\left(\frac{A\tilde{L}A}{p} - bI_n\right) = 0, \tag{3.19}$$

where $b = \frac{a_1^2}{a_2}$.

Thus,

$$\begin{aligned}
\lim_{n\to\infty}\lim_{p\to\infty} e_1 &= \lim_{n\to\infty}\lim_{p\to\infty} P\{4\boldsymbol{w}_1'\boldsymbol{w}_2 > 0\}\\
&= \lim_{n\to\infty}\lim_{p\to\infty} P\{(\boldsymbol{w}_1+\boldsymbol{w}_2)'(\boldsymbol{w}_1+\boldsymbol{w}_2) - (\boldsymbol{w}_1-\boldsymbol{w}_2)'(\boldsymbol{w}_1-\boldsymbol{w}_2) > 0\}\\
&= \lim_{n\to\infty} P\{\chi^2_{n,\gamma_1} - \chi^2_{n,\gamma_2} > 0\}\\
&= \lim_{n\to\infty} P\{\chi^2_{n,\gamma_2} - \chi^2_{n,\gamma_1} < 0\},
\end{aligned}$$

where

$$\begin{aligned}
\gamma_1 &= (k_2^{-1} - k_1^{-1})^2\theta_0^2,\\
\gamma_2 &= (k_2^{-1} + k_1^{-1})^2\theta_0^2,\\
\theta_0^2 &= \lim_{p\to\infty}\left(\frac{\boldsymbol{\delta}'\Lambda\boldsymbol{\delta}}{pa_2}\right) \equiv \lim_{p\to\infty}\theta^2,
\end{aligned}$$

and $\chi^2_{r,\gamma}$ denotes the noncentral chi-square with r degrees of freedom and noncentrality parameter γ; here χ^2_{n,γ_1} and χ^2_{n,γ_2} are independently distributed. Thus,

$$\begin{aligned}
\lim_{n\to\infty}\lim_{p\to\infty} e_1 &= \lim_{n\to\infty} P\left\{\sum_{j=1}^{n}(z_{2j}^2 - z_{1j}^2) < 0\right\}\\
&= \lim_{n\to\infty} P\left\{\sum_{j=1}^{n} w_j < 0\right\},
\end{aligned}$$

where z_{ij} are independently distributed as $N(\sqrt{\gamma_i/n}, 1)$, $j = 1, \dots, n$, $i = 1, 2$, and w_j are independently and identically distributed with

$$\begin{aligned} E(w_j) &= \frac{\gamma_2 - \gamma_1}{n} \\ &= \frac{4}{n} k_2^{-1} k_1^{-1} \theta_0^2, \\ \mathrm{Var}(w_j) &= 4 + 4\frac{\gamma_1 + \gamma_2}{n}, \\ \kappa_3(w_j) &= E[w_j - E(w_j)]^3 \\ &= 24\frac{\gamma_2 - \gamma_1}{n} \\ &= \frac{96}{n} k_2^{-1} k_1^{-1} \theta_0^2. \end{aligned}$$

Hence

$$\begin{aligned} &\lim_{n\to\infty} \lim_{p\to\infty} e_1 \\ &\quad = \lim_{n\to\infty} P \left\{ \frac{\sum_{j=1}^n w_j - 4k_2^{-1} k_1^{-1} \theta_0^2}{2(n + \gamma_1 + \gamma_2)^{\frac{1}{2}}} < -\frac{4k_2^{-1} k_1^{-1} \theta_0^2}{2(n + \gamma_1 + \gamma_2)^{\frac{1}{2}}} \right\}. \end{aligned}$$

Let

$$l = \frac{4k_2^{-1} k_1^{-1} \theta^2}{2[n + 2(k_1^{-2} + k_2^{-2})\theta^2]^{\frac{1}{2}}}. \tag{3.20}$$

Then from Edgeworth's expansion, see Cramer [1, pp. 229], we get the following theorem.

Theorem 3.1. *The error of misclassification e_1 is asymptotically given by*

$$e_1 = \Phi(-l) - \frac{l^2 - 1}{6} \phi(l) \frac{12k_2^{-1} k_1^{-1} \theta^2}{[n + 2(k_1^{-2} + k_2^{-2})\theta^2]^{\frac{3}{2}}} + o(n^{-\frac{1}{2}}). \tag{3.21}$$

Acknowledgments

I am grateful to the two anonymous referees for kindly reading this article and offering suggestions for improvements and to M. Shakhatreh for the computation in the example of Section 2.2. This research was supported by Natural Sciences and Engineering Research Council of Canada.

Appendix

Lemma A.1. *Let $V = YY' \sim W_p(\Lambda, n)$, $Y = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \overset{iid}{\sim} N_p(\mathbf{0}, \Lambda)$, $\Lambda > 0$, and $V = H'\widetilde{L}H$, where $HH' = I_n$, $\widetilde{L} = diag(\tilde{l}_1, \dots, \tilde{l}_n)$, a diagonal matrix consisting of the eigenvalues of V in the diagonal, and $\Lambda = diag(\lambda_1, \dots, \lambda_p)$ consisting of the eigenvalues of the covariance matrix $\Sigma > 0$. Then, in probability*

(a) $\lim_{p\to\infty} \frac{\widetilde{L}}{p} = a_{10} I_n$,

(b) $\lim_{p\to\infty} \frac{Y'Y}{p} = a_{10} I_n$,

(c) $\lim_{p\to\infty} H\Lambda H' = \frac{a_{20}}{a_{10}} I_n$,

(d) $\lim_{n\to\infty} \lim_{p\to\infty} \frac{\boldsymbol{a}'H'H\boldsymbol{a}}{n} = \lim_{p\to\infty} \frac{\boldsymbol{a}'\Lambda\boldsymbol{a}}{pa_1}$,
where $\boldsymbol{a} \neq \mathbf{0}$, is a vector of constants.

Proof. The n eigenvalues $\tilde{l}_1, \ldots, \tilde{l}_n$ of the diagonal matrix $\tilde{L}$ are the n nonzero eigenvalues of $V = YY'$, where the n columns of the $p \times n$ matrix Y are iid $N_p(\mathbf{0}, \Lambda)$. The n nonzero eigenvalues of YY' are also the n eigenvalues of $Y'Y$. Let U denote a $p \times n$ matrix where its n columns are iid $N_p(\mathbf{0}, I_p)$. Then, the eigenvalues of $Y'Y$ are in distribution the eigenvalues of

$$U'\Lambda U = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_n' \end{pmatrix} \Lambda(\mathbf{u}_1, \ldots, \mathbf{u}_n)$$
$$= \begin{pmatrix} \mathbf{u}_1'\Lambda\mathbf{u}_1, & \mathbf{u}_1'\Lambda\mathbf{u}_2, & \ldots, & \mathbf{u}_1'\Lambda\mathbf{u}_n \\ \mathbf{u}_2'\Lambda\mathbf{u}_1, & \mathbf{u}_2'\Lambda\mathbf{u}_2, & \ldots, & \mathbf{u}_2'\Lambda\mathbf{u}_n \\ \vdots, & \vdots, & \ddots, & \vdots \\ \mathbf{u}_n'\Lambda\mathbf{u}_1, & \mathbf{u}_n'\Lambda\mathbf{u}_2, & \ldots, & \mathbf{u}_n'\Lambda\mathbf{u}_n \end{pmatrix}.$$

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n) = (\mathrm{u}_{ij})$. Then u_{ij} are iid $N(0, 1)$ and $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are iid $N_p(\boldsymbol{0}, I_p)$. Hence,

$$E(\mathbf{u}_1'\Lambda\mathbf{u}_1) = \mathrm{tr}\,\Lambda, \quad E(\mathbf{u}_1'\Lambda\mathbf{u}_2) = 0,$$

and

$$E(Y'Y) = E(U'\Lambda U) = pa_1 I_n,$$

where

$$a_1 = (\mathrm{tr}\,\Lambda/p), \quad \text{and} \quad \lim_{p\to\infty} a_1 = a_{10}.$$

We also note that

$$E(\mathrm{u}_{ij}^2) = 1, \quad \mathrm{Var}(\mathrm{u}_{ij}^2) = 2.$$

Hence, from Chebyshev's inequality

$$\begin{aligned} P\left\{\left|\frac{\mathbf{u}_1'\Lambda\mathbf{u}_1}{p} - a_1\right| > \varepsilon\right\} &= P\left\{\left|\frac{\sum_{i=1}^p \lambda_i(\mathrm{u}_{1i}^2 - 1)}{p}\right| > \varepsilon\right\} \\ &\leqslant \frac{E[\sum_{i=1}^p \lambda_i(\mathrm{u}_{1i}^2 - 1)]^2}{p^2\varepsilon^2} \\ &= \frac{E[\sum_{i=1}^p \lambda_i^2(\mathrm{u}_{1i}^2 - 1)^2]}{p^2\varepsilon^2} \\ &= \frac{2\sum_{i=1}^p \lambda_i^2}{p^2\varepsilon^2}. \end{aligned}$$

Since $0 < \lim_{p \to 0} (\operatorname{tr} \Lambda^2 / p) < \infty$, it follows that

$$\lim_{p \to \infty} \frac{\sum_{i=1}^{p} \lambda_i^2}{p^2} = 0.$$

Hence,

$$\lim_{p \to \infty} \frac{\mathbf{u}_i' \Lambda \mathbf{u}_i}{p} \; \to \; a_{10}, \quad i = 1, \ldots, n$$

in probability. Similarly, it can be shown that in probability

$$\lim_{p \to \infty} \frac{\mathbf{u}_i' \Lambda \mathbf{u}_j}{p} = 0, \quad i \neq j,$$

and

$$\lim_{p \to \infty} \frac{Y'Y}{p} = a_{10} I_n \;\text{ in probability.}$$

This proves (a). Also, if $\tilde{l}_1, \ldots, \tilde{l}_n$ denote the nonzero eigenvalues of YY' then, from the above result, it follows that

$$\lim_{p \to \infty} \left(\frac{1}{p} \right) \tilde{L} = a_{10} I_n \;\text{ in probability.}$$

This proves (b). We note that

$$YY' = H' \tilde{L}^{\frac{1}{2}} G G' \tilde{L}^{\frac{1}{2}} H,$$

for an $n \times n$ orthogonal matrix G, $GG' = I_n$ depending on Y. Choosing $G = L^{\frac{1}{2}} H Y (Y'Y\, Y^{-1})$, we find that in distribution,

$$Y = H' \tilde{L}^{\frac{1}{2}} G \sim N_{p,n}(0, \Lambda, I_n).$$

Thus, in distribution

$$G Y' \Lambda Y G' = G U' \Lambda^2 U G' = \tilde{L}^{\frac{1}{2}} H \Lambda H' \tilde{L}^{\frac{1}{2}},$$

where $U = (\mathbf{u_1}, \ldots, \mathbf{u_n})$. We note that

$$E \left(\frac{\mathbf{u}_i' \Lambda^2 \mathbf{u}_j}{p} \right) = \frac{\operatorname{tr} \Lambda^2}{p} \, \delta_{ij},$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0, i \neq j, \; i, j = 1, \ldots, n$, the Kronecker symbol. Similarly,

$$\begin{aligned} \operatorname{Var} \left(\frac{\mathbf{u}_i' \Lambda^2 \mathbf{u}_j}{p} \right) &= \frac{2 \operatorname{tr} \Lambda^4}{p^2}, \quad i = j, \\ &= \frac{\operatorname{tr} \Lambda^4}{p^2}, \quad i \neq j. \end{aligned}$$

Since, $\lim\limits_{p\to\infty} \operatorname{tr}\Lambda^4/p = a_{40}$, and $0 < a_{40} < \infty$, it follows that

$$\lim_{p\to\infty} \frac{\operatorname{tr}\Lambda^4}{p^2} = 0,$$

Hence, in probability,

$$\lim_{p\to\infty} \left[G\left(\frac{Y'\Lambda Y}{p}\right)G'\right] = \left(\lim_{p\to\infty} \frac{\operatorname{tr}\Lambda^2}{p}\right) I_n = a_{20} I_n.$$

Thus, in probability

$$\lim_{p\to\infty} \left(\frac{\tilde{L}^{\frac{1}{2}} H\Lambda H' \tilde{L}^{\frac{1}{2}}}{p}\right) = a_{20} I_n.$$

Since, $\lim\limits_{p\to\infty} (\tilde{L}/p) = a_{10} I_n$ it follows that in probability

$$\lim_{p\to\infty} (H\Lambda H') = (a_{20}/a_{10}) I_n.$$

This proves (c).

To prove (d), consider a nonnull p-vector $\boldsymbol{a} = (a_1, \ldots, a_p)'$. Then, since $YY' = H'\tilde{L}H$, $HH' = I_n$, we get

$$\frac{\boldsymbol{a}'YY'\boldsymbol{a}}{pn} = \frac{\boldsymbol{a}'H'\tilde{L}H\boldsymbol{a}}{pn}.$$

With $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$, the left side

$$= \frac{1}{pn} \sum_{i=1}^{n} (\boldsymbol{a}'\boldsymbol{y}_i)^2$$

$$= \frac{1}{pn}\left[\sum_{i=1}^{n} \sum_{j=1}^{p} a_j^2 y_{ij}^2\right] + \frac{2}{pn} \sum_{i=1}^{n} \sum_{j<k}^{p} a_j a_k y_{ij} y_{ik},$$

of which the second term goes to zero in probability.

Hence, in probability

$$\lim_{n,p\to\infty} \frac{1}{pn} \sum_{i=1}^{n} \sum_{j=1}^{p} a_j^2 y_{ij}^2 = \lim_{n,p\to\infty} \frac{\boldsymbol{a}'H'\tilde{L}H\boldsymbol{a}}{pn}.$$

From the law of large numbers, the left side goes to $\lim_{n\to\infty} \lim_{p\to\infty} (\frac{\boldsymbol{a}'\Lambda\boldsymbol{a}}{p})$, and from the results in (a), we have in probability $\lim_{p\to\infty} p^{-1}\tilde{L} = a_{10} I_n$. Hence, in probability

$$\lim_{n,p\to\infty} \left(\frac{\boldsymbol{a}'H'H\boldsymbol{a}}{n}\right) = \lim_{p\to\infty} \left(\frac{\boldsymbol{a}'\Lambda\boldsymbol{a}}{p}\right) \Big/ a_{10} I_n.$$

This proves (d). □

This lemma along with other similar results appear in [13].

References

[1] H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, USA, 1946.

[2] S. DasGuspta, Probability inequalities and errors in classification, Ann. Statist. 2 (1974) 751–762.

[3] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tremors using gene expression data, J. Amer. Statist. Assoc. 97 (2002) 77–87.

[4] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[5] Y. Fujukoshi, Multivariate analysis for the case when the dimension is large compared to the sample size, J. Korean Statist. Soc. 33 (2004) 1–24.

[6] J. Kiefer, R. Schwartz, Admissible Bayes character of $T^2 - R^2$ and other fully invariant tests for classical multivariate normal problems, Ann. Math. Statist. 36 (1965) 747–770.

[7] E.L. Lehmann, Testing Statistical Hypotheses, Wiley, New York, 1959.

[8] C.H. Nakatsu, S.M. Brouder, J.D. Wilbur, F. Wanjau, R.W. Doerge, Impact of tillage and crop rotation on corn development and its associated microbial community, Proceedings of 15th Conference of the International Soil Tillage Research Organization, Fort Worth, Texas ISTRO, 2000.

[9] M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function, Ann. Math. Statist. 34 (1963) 1286–1301 Correction: 39, 1358–1359.

[10] H. Saranadasa, Asymptotic expansion of the misclassification probabilities of D- and A-criteria for discrimination from the two high dimensional populations using the theory of large dimensional metrices, J. Multivariate Anal. 46 (1993) 154–174.

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

[12] M.S. Srivastava, Methods of Multivariate Statistics, Wiley, New York, 2002.

[13] M.S. Srivastava, Multivariate theory for analyzing high dimensional data, J. Japan Statist. Soc. 37 (2007) to appear.

[14] M.S. Srivastava, C.G. Khatri, An Introduction to Multivariate Statistics, North Holland, New York, 1979.

[15] A. Wald, On the statistical problem arising in the classification of an individual into one of two groups, Ann. Math. Statist. 15 (1944) 145–162.

[16] J.D. Wilbur, J.K. Ghosh, C.H. Nakatsu, R.W. Doerge, Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA finger prints, Biometrics 58 (2002) 378–386.