CLASSIFICATION INTO TWO MULTIVARIATE NORMAL DISTRIBUTIONS WITH EQUAL COVARIANCE MATRICES

by

Wie-Liem Loh
Purdue University

Technical Report #90-21

Department of Statistics
Purdue University

May 1990

◇

CLASSIFICATION INTO TWO MULTIVARIATE NORMAL DISTRIBUTIONS WITH EQUAL COVARIANCE MATRICES

BY WEI-LIEM LOH

Purdue University

In this paper, we consider the classical classification problem in which we have a training sample from each of two multivariate normal populations with equal covariance matrices and we wish to classify another observation $\mathbf{x}$ as coming from one of the two populations. Furthermore we assume that the parameters of these two distributions are unknown and that the priori probabilities that $\mathbf{x}$ comes from each of these populations are equal. It is well known that the most widely used discriminant procedure for this purpose was introduced by Fisher (1936). By extending the methods of Stein (1975), (1977), an alternative linear discriminant procedure is proposed. The idea here is to try to improve upon the usual procedure by using a better estimate of the smallest eigenvalue of the population covariance matrix. Although we have obtained some first-order asymptotic results concerning this alternative procedure, we have not been able to obtain an analytical treatment of the misclassification rate of this procedure. A Monte Carlo study is used instead to evaluate its performance relative to the usual discriminant rule. The results indicate that the misclassification rate of the alternative discriminant rule compares very favorably with the usual rule.

KEY WORDS: Linear discriminant analysis; Fisher's linear discriminant function; Monte Carlo; Multivariate normal distribution; Eigenvalue decomposition.

1 Introduction

In this paper we consider the problem in which we have a sample from each of two multivariate normal populations and we wish to classify another observation as coming from one of the two populations. More precisely, suppose we have a training sample $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}$ from $N_p(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and a training sample $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}$ from $N_p(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ where $n_1 + n_2 - p - 3 > 0$

and Σ nonsingular. We wish to classify another observation $\mathbf{x}$ as coming from one of these two distributions.

In the case where the two distributions are completely known, Wald (1944) proved that the classification procedure which minimizes the misclassification rate is given by the Fisher's discriminant function, namely: Classify $\mathbf{x}$ into $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ if

$$OPT(\mathbf{x}) = [\mathbf{x} - (\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})/2]'\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \geq \log(q_2/q_1), \qquad (1)$$

or into $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ otherwise, where q_1, q_2 are the priori probabilities that $\mathbf{x}$ comes from $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$, $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ respectively. For simplicity and definiteness, we shall assume throughout this paper that the priori probabilities q_1 and q_2 are equal.

However very often, the parameters of the two distributions $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ and $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ are unknown and need to be estimated from the training samples. If the maximum likelihood estimators are used to estimate these parameters in (1), we obtain the usual linear discriminant rule, that is, classify an observation $\mathbf{x}$ into $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ if

$$LDF(\mathbf{x}) = (n_1 + n_2 - 2)[\mathbf{x} - (\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})/2]'\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \geq 0,$$

or into $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ otherwise. Here for $i = 1, 2$,

$$\begin{aligned} \bar{\mathbf{x}}^{(i)} &= \sum_{j=1}^{n_i} \mathbf{x}_j^{(i)}/n_i, \\ \mathbf{S} &= \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})' \end{aligned}$$

and the MLE of Σ is scaled by a factor to remove the bias. As noted by Gnanadesikan, et. al. (1989), the above procedure is presently the most widely used rule for classifying an observation $\mathbf{x}$ into one of two populations. Its main advantages are simplicity, the availability of package programs and reasonable robustness against model violations especially for moderate sample sizes. Excellent accounts of this procedure can be found, for example, in Anderson (1984) and Gnanadesikan, et. al. (1989).

In the next section, we propose an alternative linear discriminant rule. Section 3 summarizes the results of a Monte Carlo study which compares the performance of this rule with that of the usual linear discriminant rule. Some asymptotic properties of the alternative procedure are proved in Section 4

and the derivation of this procedure is given in Section 5. We conclude with remarks on some issues that have not been addressed to in this paper. Technical details are deferred to the Appendix.

We end this section with the following note on notation. If a matrix $\mathbf{A}$ has entries a_{ij}, we shall indicate it by (a_{ij}). Given an $r \times s$ matrix $\mathbf{A}$, its $s \times r$ transpose is denoted by $\mathbf{A}'$. $\mathbf{A}^{-1}$, tr$\mathbf{A}$ denote the inverse of the square matrix $\mathbf{A}$, the trace of $\mathbf{A}$ respectively. If a $p \times p$ matrix $\mathbf{A}$ is diagonal and has entries a_{ij}, we shall write it as $\mathbf{A} = \text{diag}(a_{11}, \ldots, a_{pp})$. Finally the expected value of a random vector $\mathbf{X}$ is denoted by $E\mathbf{X}$.

2 Alternative Linear Discriminant Rule

The usual linear discriminant rule is obtained by 'plugging in' the MLE's into (1). It is conceivable that by using a better estimate of $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$, a better linear discriminant rule will emerge. This idea of replacing the MLE of $\boldsymbol{\eta}$ by a better estimate is definitely not new. This has been an area of intensive research over the years. The literature includes Das Gupta (1965), DiPillo (1976), (1979), Dey and Srinivasan (1986), Haff (1986), Greene and Rayens (1989), Rodriguez (1988) and Friedman (1989). Unfortunately, for the problem at hand, it appears that none of the procedures found in the literature dominates or nearly dominates the usual linear discriminant rule in terms of misclassification rate.

In this paper, we propose the following alternative linear discriminant rule: Classify $\mathbf{x}$ into $N_p(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ if

$$ALT(\mathbf{x}) = [\mathbf{x} - (\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})/2]'\hat{\boldsymbol{\eta}}_{ALT} \geq 0,$$

or into $N_p(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ otherwise, where $\hat{\boldsymbol{\eta}}_{ALT}$ is determined below.

First we need some additional notation. Let $\mathbf{H} = (h_{ij})$ be an $p \times p$ orthogonal matrix such that $\mathbf{HSH}' = \mathbf{L} = \text{diag}(l_1, \ldots, l_p)$ with $l_1 \geq \ldots \geq l_p$ and $h_{1i} \geq 0$ for $i = 1, \ldots, p$. Furthermore we let $\mathbf{y} = \sqrt{n_1 n_2/(n_1 + n_2)}\mathbf{H}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$.

Theorem 1 *$\hat{\boldsymbol{\eta}}(\mathbf{y}, \mathbf{S})$ is an orthogonally equivariant estimator of $\sqrt{n_1 n_2/(n_1 + n_2)}\boldsymbol{\eta}$ if and only if*

$$\hat{\boldsymbol{\eta}}(\mathbf{y}, \mathbf{S}) = \mathbf{H}'\boldsymbol{\Phi}(l_1, \ldots, l_p, y_1^2, \ldots, y_p^2)\mathbf{y}, \tag{2}$$

where $\boldsymbol{\Phi} = diag(\phi_1, \ldots, \phi_p)$.

PROOF. The proof is straightforward and is omitted. □

REMARK. If we take $\mathbf{\Phi} = (n_1 + n_2 - 2)\sqrt{(n_1 + n_2)/n_1 n_2}\mathbf{L}^{-1}$ in (2), we get the maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{MLE} = (n_1 + n_2 - 2)\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$. Again the MLE of $\mathbf{\Sigma}$ is scaled by a factor to remove the bias.

Next define

$$
\begin{aligned}
C &= \sum_{i=1}^{[p/2]} \{l_p + y_i^2(1 \wedge l_p y_p^{-2})\}/(l_i - l_p), \qquad (3)\\
\phi_1 &= (n_1 + n_2 - p - 3 + C)\sqrt{n_1 + n_2}/(l_1\sqrt{n_1 n_2}),\\
\phi_i &= (n_1 + n_2 - p - 3)\sqrt{n_1 + n_2}/(l_i\sqrt{n_1 n_2}), \qquad i = 2, \ldots, p-1,\\
\phi_p &= (n_1 + n_2 - p - 3 - C)\sqrt{n_1 + n_2}/(l_p\sqrt{n_1 n_2}),
\end{aligned}
$$

where $[p/2]$ denotes the greatest integer less than or equal to $p/2$. $\phi_1, \ldots, \phi_p$ are approximations to the diagonal elements of $\mathbf{\Phi}$. However the natural ordering of the ϕ_i's may be altered. The natural ordering of the ϕ_i's is given by $0 \le \phi_1 \le \cdots \le \phi_p$. To correct for this, Stein's (1975) isotonic regression is applied to the ϕ_i's. This results in ϕ_i^{ST}, $i = 1, \ldots, p$ where $0 \le \phi_1^{ST} \le \cdots \le \phi_p^{ST}$. For a detailed description of Stein's isotonic regression, we refer the reader to Lin and Perlman (1985). Now we define

$$\hat{\boldsymbol{\eta}}_{ALT}(\mathbf{y}, \mathbf{S}) = \mathbf{H}'\mathbf{\Phi}^{ST}\mathbf{y},$$

where $\mathbf{\Phi}^{ST} = \operatorname{diag}(\phi_1^{ST}, \ldots, \phi_p^{ST})$.

REMARK. As an alternative to Stein's isotonic regression, Haff's (1988) algorithm can be applied instead. The essential difference between these two algorithms is that the latter appears to give smoother estimates.

Here is a heuristic explanation for the form of $\hat{\boldsymbol{\eta}}_{ALT}$; its derivation is given in Section 5. Let $\lambda_1 \ge \cdots \ge \lambda_p$ denote the eigenvalues of $\mathbf{\Sigma}$. It is well known that the eigenvalues of $\mathbf{S}/(n_1 + n_2 - 2)$ are more spread out than the eigenvalues of $\mathbf{\Sigma}$. In particular, $l_p/(n_1 + n_2 - 2)$ underestimates λ_p. From the functional form of Fisher's linear discriminant function $OPT(\mathbf{x})$, it is clear that $OPT(\mathbf{x})$ is heavily influenced by the estimate of λ_p. $\hat{\boldsymbol{\eta}}_{ALT}$ tries to get a better estimate of λ_p by correcting for the bias of $l_p/(n_1 + n_2 - 2)$. We also wish to add that the correction to $l_1/(n_1 + n_2 - 2)$ is not significant here since $OPT(\mathbf{x})$ is only slightly influenced by the estimate of λ_1. The reason for doing so is to ensure that Stein's isotonic regression is formally applicable.

Another interesting point to note is that the correction factor C does not depend exclusively on $\mathbf{S}$ but also on the orientation of the eigenvectors

of $\mathbf{S}$ to that of $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$. More precisely, if $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$ lies in the subspace spanned by the eigenvectors corresponding to the smallest sample eigenvalues of $\mathbf{S}$, the correction factor should be small and if $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$ lies in the subspace spanned by the eigenvectors corresponding to the largest sample eigenvalues, C should be substantial. The reason is that in the former case, correcting the smallest sample eigenvalue reduces bias at the expense of introducing additional variance. If the Mahalanobis distance between the two populations is not too great, this may lead to increased variability of $ALT(\mathbf{x})$ and hence the misclassification rate. However in the latter case, correcting for the smallest eigenvalue reduces both bias and variance and hence would lead to a decrease in misclassification rate. This point was also raised by Friedman (1989). This is especially relevant in applications since it is usually the case that the 'signal' $(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$ lies in the subspace spanned by the eigenvectors corresponding to the largest eigenvalues of $\boldsymbol{\Sigma}$. This is also one of the main assumptions justifying the use of the first few principle components in classification, see for example Chang (1983).

3 Some Asymptotics

In this section, we assume that p is fixed and we let n_1, n_2 tend to infinity. It is well known that the usual linear discriminant rule is asymptotically optimal in that the law of $LDF(\mathbf{x})$ tends in distribution to the law of $OPT(\mathbf{x})$. The following is also true.

Theorem 2 *The law of $ALT(\mathbf{x})$ tends in distribution to that of $OPT(\mathbf{x})$.*

PROOF. This follows immediately from Slutsky's theorem and the fact that $\hat{\boldsymbol{\eta}}_{ALT} \to \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$ in probability. □

It follows from the above result that at least up to first-order asymptotics, $LDF(\mathbf{x})$ and $ALT(\mathbf{x})$ are equivalent.

4 Monte Carlo Study

Due to its rather complicated construction, we have not been able to obtain an analytical treatment of the misclassification rate of the alternative linear discriminant rule $ALT(\mathbf{x})$. We shall instead compare the behaviour of this procedure to that of the usual linear discriminant rule via a Monte Carlo study.

For the simulations, independent standard normal variates are generated by the IMSL subroutine DRNNOA and the eigenvalue decomposition uses the IMSL subroutine DEVCSF. We take $p = 5,\ n_1 = 5,\ n_2 = 5$ and $p = 10,\ n_1 = 7,\ n_2 = 7$. For simplicity, we write $\boldsymbol{\xi} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$. Due to the location and orthogonal equivariance of both procedures, we shall, without loss of generality, take $\boldsymbol{\mu}^{(2)} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ to be diagonal.

Each experiment consists of 2000 independent replications of the following procedure. First, training samples of sizes n_1 and n_2 are generated from $N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$ and $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ respectively. Next another test data set of size 100 is randomly generated such that each data point has prior probability $1/2$ of coming from $N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$ and prior probability $1/2$ as coming from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. This test set is then classified using the rules $ALT(\mathbf{x})$ and $LDF(\mathbf{x})$ derived from the training sets. This gives us estimates of the misclassification rates of both procedures.

Tables 1 and 2 give the average misclassification rates and their standard deviations of the procedures $LDF(\mathbf{x})$ and $ALT(\mathbf{x})$ over the 2000 independent replications. The numbers in italics in these tables correspond to that of $ALT(\mathbf{x})$. We observe that there is a high positive correlation between the average misclassification rates of these two procedures since the same training samples and test set are used for both. Thus the estimated standard deviation (as given in Tables 1 and 2) is probably a conservative indicator of the variability of the relative magnitude of the average misclassification rates.

We shall now summarize the results of this numerical study:

1. The misclassification rate of the alternative linear discriminant rule $ALT(\mathbf{x})$ compares very favorably with that of the usual discriminant rule $LDF(\mathbf{x})$ for the values of p, n_1, n_2 used. Savings in misclassification rate is achieved over most parts of the parameter space. In particular, this is most dramatic when the eigenvalues of $\boldsymbol{\Sigma}$ are spread far apart, $\boldsymbol{\xi}$ is in the subspace spanned by the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$ and the Mahalanobis squared distance, $\boldsymbol{\xi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}$, is moderately large. For example, Table 2 shows that when $p = 10,\ n_1 = 7,\ n_2 = 7,\ \boldsymbol{\Sigma} = \text{diag}(10^9, 10^8, \ldots, 1)$ and $\boldsymbol{\xi} = (158113.883, 0, \ldots, 0)$, the misclassification rates of the alternative and the usual linear discrimination procedures are 6.16% and 12.85% respectively. This gives an approximately 50% decrease in misclassification rate with the use of $ALT(\mathbf{x})$ over $LDF(\mathbf{x})$.

2. On the other hand, the usual discriminant rule is most favorable

against the alternative rule when the following situation occurs. The eigenvalues of Σ are far apart, ξ lies in the subspace spanned by the eigenvector corresponding to the smallest eigenvalue of Σ. However, here the difference between the misclassification rates of these procedures is small: at most an increase of 4% in misclassification rate with the use of $ALT(\mathbf{x})$ over $LDF(\mathbf{x})$. For example, Table 1 shows that when $p = 5$, $n_1 = 5$, $n_2 = 5$, $\Sigma = \text{diag}(10^8, 10^6, 10^4, 10^2, 1)$ and $\xi = (0, 0, 0, 0, 3)$, the misclassification rates of the alternative and the usual linear discriminant procedures are 18.63% and 18.09% respectively. This gives roughly an increase of 3% in misclassification rate with the use of $ALT(\mathbf{x})$ over $LDF(\mathbf{x})$.

3. For a fixed set of parameters (ξ, Σ), this study shows that the difference in the misclassification rates of these two procedures decreases with n_1 and n_2. Theorem 2 also points to this. Furthermore, the difference in misclassification rates also decreases with increasing p. This is intuitively evident since $ALT(\mathbf{x})$ essentially differs from $LDF(\mathbf{x})$ only in the smallest eigenvalue estimate of Σ.

TABLE 1

$p = 5 \quad n_1 = 5 \quad n_2 = 5$

Average misclassification rates of usual discriminant rule *and the alternative discriminant rule*

(Estimated standard errors are in parenthesis)

Eigenvalues of Σ	$\xi = (x, 0, \ldots, 0)'$	$\xi = (0, \ldots, 0, x)'$
$\xi'\Sigma^{-1}\xi = 1.00$		
(1,1,1,1,1)	42.29 (0.18)	42.17 (0.19)
	41.82 (0.18)	*41.77 (0.19)*
(25,1,1,1,1)	42.29 (0.18)	42.17 (0.19)
	40.91 (0.19)	*42.08 (0.18)*
(25,25,25,25,1)	42.29 (0.18)	42.17 (0.19)
	42.00 (0.18)	*42.35 (0.19)*
(25,5,1,0.5,0.04)	42.29 (0.18)	42.17 (0.19)
	41.24 (0.19)	*42.50 (0.18)*
(50,40,30,20,10)	42.29 (0.18)	42.17 (0.19)
	41.57 (0.19)	*42.18 (0.18)*
(16,8,4,2,1)	42.29 (0.18)	42.17 (0.19)
	41.34 (0.19)	*42.31 (0.18)*
$(10^8, 10^6, 10^4, 10^2, 1)$	42.29 (0.18)	42.17 (0.19)
	41.19 (0.19)	*42.56 (0.19)*
$\xi'\Sigma^{-1}\xi = 9.00$		
(1,1,1,1,1)	18.08 (0.20)	18.09 (0.21)
	15.97 (0.19)	*16.13 (0.19)*
(25,1,1,1,1)	18.08 (0.20)	18.09 (0.21)
	13.22 (0.19)	*17.39 (0.20)*
(25,25,25,25,1)	18.08 (0.20)	18.09 (0.21)
	16.07 (0.19)	*18.37 (0.21)*
(25,5,1,0.5,0.04)	18.08 (0.20)	18.09 (0.21)
	13.21 (0.19)	*18.54 (0.22)*
(50,40,30,20,10)	18.08 (0.20)	18.09 (0.21)
	14.68 (0.19)	*17.84 (0.20)*
(16,8,4,2,1)	18.08 (0.20)	18.09 (0.21)
	13.73 (0.19)	*18.09 (0.20)*
$(10^8, 10^6, 10^4, 10^2, 1)$	18.08 (0.20)	18.09 (0.21)
	13.14 (0.18)	*18.63 (0.22)*

TABLE 1 CONTD.

Eigenvalues of Σ	$\xi = (x, 0, \ldots, 0)'$	$\xi = (0, \ldots, 0, x)'$
$\xi'\Sigma^{-1}\xi = 25.00$		
(1,1,1,1,1)	6.29 (0.15)	6.30 (0.15)
	4.64 (0.13)	*4.63 (0.13)*
(25,1,1,1,1)	6.29 (0.15)	6.30 (0.15)
	3.22 (0.11)	*5.72 (0.14)*
(25,25,25,25,1)	6.29 (0.15)	6.30 (0.15)
	4.35 (0.12)	*6.35 (0.15)*
(25,5,1,0.5,0.04)	6.29 (0.15)	6.30 (0.15)
	3.01 (0.11)	*6.42 (0.15)*
(50,40,30,20,10)	6.29 (0.15)	6.30 (0.15)
	3.88 (0.12)	*5.92 (0.14)*
(16,8,4,2,1)	6.29 (0.15)	6.30 (0.15)
	3.41 (0.11)	*6.15 (0.14)*
$(10^8, 10^6, 10^4, 10^2, 1)$	6.29 (0.15)	6.30 (0.15)
	2.90 (0.11)	*6.46 (0.16)*

TABLE 2

$p = 10 \quad n_1 = 7 \quad n_2 = 7$

Average misclassification rates of usual discriminant rule *and the alternative discriminant rule*

(Estimated standard errors are in parenthesis)

Eigenvalues of Σ	$\xi = (x, 0, \ldots, 0)'$	$\xi = (0, \ldots, 0, x)'$
$\xi'\Sigma^{-1}\xi = 1.00$		
(1,1,1,1,1,	44.67 (0.16)	44.66 (0.16)
1,1,1,1,1)	*44.10 (0.16)*	*44.08 (0.16)*
(10,10,10,10,10,	44.67 (0.16)	44.66 (0.16)
1,1,1,1,1)	*43.65 (0.17)*	*44.39 (0.16)*
(50,1,1,1,1,	44.67 (0.16)	44.66 (0.16)
1,1,1,1,1)	*42.85 (0.17)*	*44.25 (0.16)*
(25,25,25,25,25,	44.67 (0.16)	44.66 (0.16)
25,25,25,25,1)	*44.31 (0.16)*	*44.94 (0.15)*
(20,20,20,5,5,	44.67 (0.16)	44.66 (0.16)
5,5,1,1,1)	*43.61 (0.17)*	*44.52 (0.15)*
(100,90,80,70,60,	44.67 (0.16)	44.66 (0.16)
50,40,30,20,10)	*43.87 (0.16)*	*44.58 (0.15)*
(512,256,128,64,32,	44.67 (0.16)	44.66 (0.16)
16,8,4,2,1)	*43.09 (0.17)*	*44.89 (0.15)*
$(10^9, 10^8, 10^7, 10^6, 10^5$	44.67 (0.16)	44.66 (0.16)
$10^4, 10^3, 10^2, 10, 1)$	*42.83 (0.18)*	*45.32 (0.15)*
$\xi'\Sigma^{-1}\xi = 9.00$		
(1,1,1,1,1,	25.48 (0.22)	25.90 (0.22)
1,1,1,1,1)	*22.42 (0.22)*	*22.82 (0.23)*
(10,10,10,10,10,	25.48 (0.22)	25.90 (0.22)
1,1,1,1,1)	*19.51 (0.23)*	*25.02 (0.22)*
(50,1,1,1,1,	25.48 (0.22)	25.90 (0.22)
1,1,1,1,1)	*17.58 (0.24)*	*23.63 (0.22)*
(25,25,25,25,25,	25.48 (0.22)	25.90 (0.22)
25,25,25,25,1)	*22.53 (0.22)*	*26.28 (0.22)*
(20,20,20,5,5,	25.48 (0.22)	25.90 (0.22)
5,5,1,1,1)	*19.40 (0.23)*	*25.28 (0.22)*
(100,90,80,70,60,	25.48 (0.22)	25.90 (0.22)
50,40,30,20,10)	*20.25 (0.23)*	*25.38 (0.22)*
(512,256,128,64,32,	25.48 (0.22)	25.90 (0.22)
16,8,4,2,1)	*17.79 (0.24)*	*26.01 (0.22)*
$(10^9, 10^8, 10^7, 10^6, 10^5,$	25.48 (0.22)	25.90 (0.22)
$10^4, 10^3, 10^2, 10, 1)$	*16.71 (0.24)*	*26.79 (0.24)*

TABLE 2 CONTD.

Eigenvalues of Σ	$\xi = (x, 0, \ldots, 0)'$	$\xi = (0, \ldots, 0, x)'$
$\xi'\Sigma^{-1}\xi = 25.00$		
(1,1,1,1,1,	12.85 (0.21)	13.30 (0.22)
1,1,1,1,1)	*9.56 (0.20)*	*9.91 (0.21)*
(10,10,10,10,10,	12.85 (0.21)	13.30 (0.22)
1,1,1,1,1)	*7.61 (0.20)*	*12.49 (0.21)*
(50,1,1,1,1,	12.85 (0.21)	13.30 (0.22)
1,1,1,1,1)	*6.73 (0.19)*	*10.87 (0.21)*
(25,25,25,25,25,	12.85 (0.21)	13.30 (0.22)
25,25,25,25,1)	*9.20 (0.20)*	*13.34 (0.22)*
(20,20,20,5,5,	12.85 (0.21)	13.30 (0.22)
5,5,1,1,1)	*7.55 (0.20)*	*12.67 (0.21)*
(100,90,80,70,60,	12.85 (0.21)	13.30 (0.22)
50,40,30,20,10)	*8.12 (0.20)*	*12.61 (0.21)*
(512,256,128,64,32,	12.85 (0.21)	13.30 (0.22)
16,8,4,2,1)	*6.74 (0.19)*	*13.21 (0.22)*
$(10^9, 10^8, 10^7, 10^6, 10^5,$	12.85 (0.21)	13.30 (0.22)
$10^4, 10^3, 10^2, 10, 1)$	*6.16 (0.19)*	*13.56 (0.22)*
$\xi'\Sigma^{-1}\xi = 49.00$		
(1,1,1,1,1,	6.41 (0.17)	6.78 (0.18)
1,1,1,1,1)	*4.53 (0.16)*	*4.71 (0.17)*
(10,10,10,10,10,	6.41 (0.17)	6.78 (0.18)
1,1,1,1,1)	*3.54 (0.15)*	*6.20 (0.18)*
(50,1,1,1,1,	6.41 (0.17)	6.78 (0.18)
1,1,1,1,1)	*3.12 (0.14)*	*5.56 (0.18)*
(25,25,25,25,25,	6.41 (0.17)	6.78 (0.18)
25,25,25,25,1)	*4.37 (0.16)*	*6.71 (0.18)*
(20,20,20,5,5,	6.41 (0.17)	6.78 (0.18)
5,5,1,1,1)	*3.55 (0.15)*	*6.36 (0.18)*
(100,90,80,70,60,	6.41 (0.17)	6.78 (0.18)
50,40,30,20,10)	*3.86 (0.15)*	*6.27 (0.18)*
(512,256,128,64,32,	6.41 (0.17)	6.78 (0.18)
16,8,4,2,1)	*3.11 (0.14)*	*6.67 (0.18)*
$(10^9, 10^8, 10^7, 10^6, 10^5,$	6.41 (0.17)	6.78 (0.18)
$10^4, 10^3, 10^2, 10, 1)$	*2.85 (0.14)*	*6.84 (0.18)*

5 Derivation of Alternative Discriminant Rule

From Section 2, we observe that it suffices to derive (3). First we observe that Fisher (1936) used the following criterion to estimate $\boldsymbol{\eta}$: Choose $\hat{\boldsymbol{\eta}}$ to maximize the ratio $\hat{\boldsymbol{\eta}}'(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})/\sqrt{\hat{\boldsymbol{\eta}}'\mathbf{S}\hat{\boldsymbol{\eta}}}$. This implies, as a first approximation, that the relevant loss functions to consider may be

$$L_1(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}) = -\hat{\boldsymbol{\eta}}'\boldsymbol{\xi}/(\sqrt{\hat{\boldsymbol{\eta}}'\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}}\sqrt{\boldsymbol{\xi}'\boldsymbol{\Sigma}\ \boldsymbol{\xi}})$$

and

$$L_2(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}) = -\hat{\boldsymbol{\eta}}'\mathbf{S}\boldsymbol{\eta}/(\sqrt{\hat{\boldsymbol{\eta}}'\mathbf{S}\hat{\boldsymbol{\eta}}}\sqrt{\boldsymbol{\xi}'\mathbf{S}\boldsymbol{\xi}}).$$

From the method of Lagrange multipliers, this is roughly equivalent to choosing $(\hat{\boldsymbol{\eta}}, \lambda)$ so as to 'maximize'

$$L_3 = \sqrt{n_1 n_2/(n_1+n_2)}\hat{\boldsymbol{\eta}}'\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi} - \lambda(\hat{\boldsymbol{\eta}}'\mathbf{S}\hat{\boldsymbol{\eta}} - 1).$$

The following theorem gives an unbiased estimate of L_3.

Theorem 3 *Let $\hat{\boldsymbol{\eta}}$ be defined as in (2). Then*

$$E(L_3) = \lambda + E(R),$$

where

$$\begin{aligned} R \;=\; & \sum_{i=1}^{p}\{(n_1+n_2-4)\phi_i y_i^2 - l_i\phi_i - 2(l_i\phi_i)^{(p+i)}y_i^2 \\ & + \phi_i y_i^2\Big(\sum_{j\neq i}\frac{l_j}{l_i-l_j}\Big) + 2(l_i\phi_i)^{(i)}y_i^2 \\ & - 2l_i y_i^2\Big(\sum_{j\neq i}\frac{y_j^2}{l_i-l_j}[\phi_i^{(p+j)} - \phi_i^{(p+i)}]\Big) + l_i\phi_i\Big(\sum_{j\neq i}\frac{y_j^2}{l_i-l_j}\Big) - \lambda l_i\phi_i^2 y_i^2\}, \end{aligned}$$

with $(l_i\phi_i)^{(p+j)} = \partial(l_i\phi_i)/\partial y_j^2$ and $(l_i\phi_i)^{(i)} = \partial(l_i\phi_i)/\partial l_i$.

The proof of this theorem is deferred to the Appendix.

Now we proceed with a heuristic maximization of R relative to the ϕ_i's. We observe that in order to compete well with the usual discriminant rule, our estimate of $\boldsymbol{\eta}$ must tend to $\hat{\boldsymbol{\eta}}_{MLE}$ when the l_i's are spread far apart and

that y_p^2 is large relative to the rest of the y_i^2's. This implies that if these conditions hold, we should have for $1 \leq i, j \leq p$

$$\begin{aligned} (l_i\phi_i)^{(p+j)} &\approx 0, \\ (l_i\phi_i)^{(i)} &\approx 0. \end{aligned} \tag{4}$$

Thus ignoring the derivative terms in R gives us

$$\begin{aligned} R &\approx \hat{R} \\ &= \sum_{i=1}^{p}\{(n_1+n_2-4)\phi_i y_i^2 - l_i\phi_i + \phi_i y_i^2(\sum_{j\neq i}\frac{l_j}{l_i-l_j}) \\ &\quad + l_i\phi_i(\sum_{j\neq i}\frac{y_j^2}{l_i-l_j}) - \lambda l_i\phi_i^2 y_i^2\}. \end{aligned}$$

$\hat{R}$ is maximized if for $1 \leq i \leq p$,

$$\phi_i = [n_1+n_2-p-3+(\sum_{j\neq i}\frac{l_i}{l_i-l_j})+(\sum_{j\neq i}\frac{l_i y_j^2}{y_i^2(l_i-l_j)})-\frac{l_i}{y_i^2}]/(2\lambda l_i). \tag{5}$$

With ϕ_i defined as such, we observe that condition (4) is approximately valid only when $i = p$. Furthermore, we observe that the absolute magnitude of $\hat{\boldsymbol{\eta}}$ does not matter here, only its direction. Hence we define for $i = 2, \ldots, p-1$,

$$\phi_i = \sqrt{(n_1+n_2)/n_1n_2}(n_1+n_2-p-3)/l_i.$$

A slight perturbation of the functional forms of ϕ_1 and ϕ_p results in

$$\begin{aligned} \phi_1 &= (n_1+n_2-p-3+C_{a,b})\sqrt{n_1+n_2}/(l_1\sqrt{n_1n_2}), \\ \phi_p &= (n_1+n_2-p-3-C_{a,b})\sqrt{n_1+n_2}/(l_p\sqrt{n_1n_2}), \end{aligned}$$

where for suitable constants a, b, we define

$$C_{a,b} = \sum_{i=1}^{a}\{l_p + y_i^2(b \wedge l_p y_p^2)\}/(l_i - l_p).$$

From Monte Carlo simulations, we find that taking $a = [p/2]$ and $b = 1$ gives a favorable alternative linear discriminant rule. This results in (3).

REMARK. We observe that (3) can also be obtained by using the following naive quadratic loss function get a better estimate of $\boldsymbol{\eta}$:

$$L_4(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}) = (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})'\mathbf{S}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}).$$

6 Final Remarks

There are a number of important issues that have not been addressed to in this paper. The first is the possible improvements in the proposed linear discriminant rule. It is evident that getting a better estimate of the smallest eigenvalue of Σ is of prime importance. However the alternative rule can be 'fine tuned' by getting better estimates of all eigenvalues of Σ, not just the smallest eigenvalue. A promising idea would be to iterate the methods introduced here.

We have assumed that the priori probabilities q_1 and q_2 are equal. If q_1 and q_2 are unequal, it is clear from the functional form of $ALT(\mathbf{x})$ that not only the direction of $\hat{\boldsymbol{\eta}}$, but also its magnitude should be taken into account.

Another issue is the performance of the alternative discriminant rule under model violations. In particular we are concerned with the performance of the alternative rule when (1) the two populations are multivariate normal but have unequal covariance matrices and (2) the populations may not be multivariate normal. More research is needed in this direction.

7 Acknowledgments

I would like to thank Professor Charles M. Stein for his advice and for introducing me to this problem.

8 Appendix

PROOF OF THEOREM 3.

First we need some additional definitions. A function $g : R^{p\times n} \to R$ is almost differentiable if, for every direction, the restrictions to almost all lines in that direction are absolutely continuous. If g on $R^{p\times n}$ is vector-valued instead of being real-valued, then g is almost differentiable if each of its coordinate functions are.

Theorem 4 (Normal Identity) *Let* $\mathbf{y} = (y_1, \ldots, y_p)' \sim N_p(\boldsymbol{\mu}, \Sigma)$ *and* $\mathbf{g} : R^p \to R^p$ *be an almost differentiable function such that* $E[\sum_{i,j} \mid \partial g_i(\boldsymbol{y}) / \partial y_j \mid]$ *is finite. Then*

$$E[\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\mathbf{g}'(\mathbf{y})] = E[\nabla \mathbf{g}'(\mathbf{y})],$$

where $\nabla = (\partial/\partial y_1, \ldots, \partial/\partial y_p)'$.

The Normal identity was first proved by Stein (1973). Let S_p denote the set of $p \times p$ positive definite matrices. Also we write for $1 \leq i, j \leq p$,

$$\tilde{\nabla} = (\tilde{\nabla}_{ij})_{p\times p}, \text{ where } \tilde{\nabla}_{ij} = (1/2)(1+\delta_{ij})\partial/\partial s_{ij},$$

where δ_{ij} denotes the Kronecker delta.

Theorem 5 (Wishart Identity) *Let* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *be a* $p \times n$ *random matrix, with the* $\mathbf{X}_k$ *independently normally distributed* p*-dimensional random vectors with mean* $\mathbf{0}$ *and unknown covariance matrix* $\mathbf{\Sigma}$*. We suppose* $n \geq p$*. Let* $\mathbf{g} : S_p \to R^{p\times p}$ *be such that* $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x}\mathbf{x}') : R^{p\times n} \to R^{p\times p}$ *is almost differentiable. Then, with* $\mathbf{S} = (s_{ij}) = \mathbf{X}\mathbf{X}'$*, we have*

$$E tr \mathbf{\Sigma}^{-1}\mathbf{g}(\mathbf{S}) = E tr[(n-p-1)\mathbf{S}^{-1}\mathbf{g}(\mathbf{S}) + 2\tilde{\nabla}\mathbf{g}(\mathbf{S})],$$

provided the expectations of the two terms on the r.h.s. exist.

The Wishart identity was proved by Stein (1975) and Haff (1977) independently. Next we need a few lemmas.

Lemma 1 *With the notation of Theorem 3,*

$$E[\mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'(\mathbf{y} - c\mathbf{H}\boldsymbol{\xi})] = E\sum_{i=1}^{p}\{l_i\phi_i + 2l_i\phi_i^{(p+i)}y_i^2\},$$

where $c = \sqrt{n_1 n_2/(n_1+n_2)}$.

PROOF. We observe from the Normal identity that

$$\begin{aligned} E[\mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'(\mathbf{y} - c\mathbf{H}\boldsymbol{\xi})] &= E(\nabla'\mathbf{L}\mathbf{\Phi}\mathbf{y}) \\ &= E\sum_{i=1}^{p}\partial(l_i\phi_i y_i)/\partial y_i \\ &= E\sum_{i=1}^{p}\{l_i\phi_i + 2l_i\phi_i^{(p+i)}y_i^2\}. \end{aligned}$$

This completes the proof. □

Lemma 2 *With the notation of Theorem 3, we have*

$$\begin{aligned} \tilde{\nabla}_{jk}l_i &= h_{ij}h_{ik}, \\ \tilde{\nabla}_{jk}h_{rs} &= \frac{1}{2}\sum_{i\neq r}\frac{h_{is}}{l_r - l_i}(h_{ij}h_{rk} + h_{ik}h_{rj}). \end{aligned}$$

Lemma 2 was first proved by Stein (1975).

Lemma 3 *Let* $n = n_1 + n_2 - 2$. *Then*

$$\begin{aligned} & E[\mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'\mathbf{y}] \\ = \; & E\sum_{i=1}^{p}\{n\phi_i y_i^2 + 2l_i\phi_i^{(i)}y_i^2 + (\sum_{j\neq i}\frac{l_j}{l_i - l_j})\phi_i y_i^2 \\ & -2l_i y_i^2(\sum_{j\neq i}\frac{y_j^2}{l_i - l_j}[\phi_i^{(p+j)} - \phi_i^{(p+i)}]) + l_i\phi_i(\sum_{j\neq i}\frac{y_j^2}{l_i - l_j})\}. \end{aligned}$$

PROOF. For simplicity, we write $\mathbf{y} = \mathbf{H}\mathbf{x}$. We observe from the Wishart identity that

$$\begin{aligned} & E[\mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'\mathbf{y}] \\ = \; & Etr(\mathbf{\Sigma}^{-1}\mathbf{H}'\mathbf{L}\mathbf{\Phi}\mathbf{H}\mathbf{x}\mathbf{x}') \\ = \; & E\{(n-p-1)\mathbf{y}'\mathbf{\Phi}\mathbf{y} + 2tr\tilde{\nabla}(\mathbf{H}'\mathbf{L}\mathbf{\Phi}\mathbf{H}\mathbf{x}\mathbf{x}')\}, \end{aligned}$$

and the result follows, after some tedious computation, from Lemma 2. □

Finally we observe that with $\hat{\boldsymbol{\eta}}$ defined as in (2) and $c = \sqrt{n_1 n_2/(n_1+n_2)}$,

$$\begin{aligned} & E\{c\hat{\boldsymbol{\eta}}'\mathbf{S}\mathbf{\Sigma}^{-1}\boldsymbol{\xi} - \lambda(\hat{\boldsymbol{\eta}}'\mathbf{S}\hat{\boldsymbol{\eta}} - 1)\} \\ = \; & E\{\mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'\mathbf{y} - \mathbf{y}'\mathbf{\Phi}\mathbf{L}\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}'(\mathbf{y} - c\mathbf{H}\boldsymbol{\xi}) - \lambda(\hat{\boldsymbol{\eta}}'\mathbf{S}\hat{\boldsymbol{\eta}} - 1)\}. \end{aligned}$$

Theorem 3 follows immediately from Lemmas 1 and 3.

References

[1] ANDERSON, T. W. (1984). *An Introduction to Multivariate Analysis*, 2nd ed. Wiley, New York.

[2] CHANG, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.* **32**, 267-275.

[3] DAS GUPTA, S. (1965). Optimum classification rules for classification into two multivariate normal populations. *Ann. Math. Statist.* **36**, 1174-1184.

[4] DEY, D. K. and SRINIVASAN, C. (1986). Estimation of normal covariance matrices with applications. *IMS Bulletin* **4**, 221.

[5] DIPILLO, P. J. (1976). The application of bias to discriminant analysis. *Commun. Statist.* **5**, 843-854.

[6] DIPILLO, P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. *Commun. Statist.* **8**, 1447-1457.

[7] FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7** (part 2), 179-188.

[8] FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84**, 165-175.

[9] GREENE, T. and RAYENS, W. S. (1989). Partially pooled covariance matrix estimation in discriminant analysis. *Commun. Statist.* **18**, 3679-3702.

[10] HAFF, L. R. (1977). Minimax estimators for a multinormal precision matrix. *J. Multivariate Anal.* **7**, 374-385.

[11] HAFF, L. R. (1986). On linear log-odds and estimation of discriminant coefficients. *Commun. Statist.* **15**, 2131-2144.

[12] HAFF, L. R. (1988). The variational form of certain Bayes estimators. Unpublished manuscript.

[13] LIN, S. P. and PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators for a covariance matrix. *Multivariate Analy.* **6**, 411-429, ed. P.K. Krishnaiah, North Holland, Amsterdam.

[14] RODRIGUEZ, A. F. (1988). Admissibility and unbiasedness of the ridge classification rules for two normal populations with equal covariance matrices. *Statistics* **19**, 383-388.

[15] STEIN, C. M. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymptotic Statist.*, 345-381.

[16] STEIN, C. M. (1975). Rietz Lecture. 39th annual meeting IMS. Atlanta, Georgia.

[17] STEIN, C. M. (1977). Lectures on the theory of estimation of many parameters. (In Russian.) *Studies in the Statistical Theory of Estimation, Part I* (Ibragimov, I. A. and Nikulin, M. S., eds.), *Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division* **74**, 4-65.

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