Robustness of Fisher's Linear Discriminant Function under Two-Component Mixed Normal Models

T. Ashikaga and P. C. Chang

College of Engineering, Mathematics and Business Administration, University of Vermont, Burlington, VT, 05405

College of Medicine, University of Vermont, Burlington, VT, 05405

School of Public Health, University of California, Los Angeles, CA, 90024, USA

Published online: 12 Mar 2012.


To link to this article: http://dx.doi.org/10.1080/01621459.1981.10477703

Please scroll down for article.
Robustness of Fisher's Linear Discriminant Function Under Two-Component Mixed Normal Models

T. ASHIKAGA and P. C. CHANG*

Robustness of Fisher's linear discriminant function is evaluated when the distributions of the two populations are characterized by two-component mixed normal distributions with known parameters. The results suggest that the linear discriminant function is rather robust when the two distributions do not markedly deviate from normality and are moderately distant, particularly if they are similar in shape.

KEY WORDS: Robustness; Linear discriminant function; Mixed normal model.

1. INTRODUCTION

Fisher's linear discriminant function (LDF) is frequently used for the two-group discrimination problem. Its application is justified, in part, by the fact that when the distributions of the two groups are multivariate normal with identical covariance matrices and the parameters are known, its use minimizes the expected loss in the case of known a priori probabilities and it is an admissible procedure when a priori probabilities are not known (Anderson 1958). However, in most practical situations, only sample data (one set from each population) are available and the parameters as well as the shapes of the two distributions are not known. In this case, use of the LDF is defended by the argument that it maximizes the sample Mahalanobis distance between the two data sets (Fisher 1936). This argument, nevertheless, does not imply that the LDF is the best procedure for these situations.

Although Fisher's LDF has been used in many practical applications, its statistical properties under nonoptimal conditions have not received much attention until recently. Performance of the LDF relative to other procedures was examined by Gilbert (1968) and Moore (1973) for the case of discrete variables; by Krzanowski (1977) and Sugano (1976) for the case of mixtures of discrete and continuous variables; and by Zhezhel (1968), Lachenbruch, Sneeringer, and Revo (1973), Lachenbruch, Sneeringer, and Revo (1973), Marks and Dunn (1974) for the case of continuous variables.

This paper examines the robustness of the linear discriminant function by comparing the error rate for the discriminant procedure based on the LDF \( P_L \) and the error rate for the Bayes procedure \( P_B \) when the distribution of the variables follows various two-component mixed normal distributions with known parameters.

The model is different from that used by Lachenbruch and Kupper (1973) in that both distributions are allowed to be nonnormal, from that used by Zhezhel (1968) in that the two populations do not necessarily have identical covariance matrices, from that used by Lachenbruch, Sneeringer, and Revo (1973) in that the covariances are not necessarily zero, and from that used by Marks and Dunn (1974) in that the joint distributions are not normal.

Section 2 describes the two-component mixed normal model. Section 3 specifies the linear and the Bayes procedures and their error rates. The parameter configurations included in this study are given in Section 4. The results are summarized in Section 5.

2. MODEL

2.1 The Two-Component Mixed Normal Model

The two-component mixed normal model assumes that the density for the random vector \( X = (X_1, \ldots, X_p)' \) from the population \( i \), \( (i = 1, 2) \), has the form

\[
 f(x) = (1 - \alpha^i)n(\mu_1^i, \Sigma_1) + \alpha^i n(\mu_2^i, \Sigma_2),
\]

where \( 0 \leq \alpha^i \leq 1 \) is the mixture level, \( \mu_1^i, \mu_2^i \) are the component mean vectors, \( \Sigma_1, \Sigma_2 \) are the component covariance matrices, and \( n(\mu, \Sigma) \) denotes the density function of a multivariate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \).

One advantage of this model is that it allows various types and degrees of deviations from normality. It also has been considered to be a reasonable model in many areas of research, such as fisheries (Smiles and Pearcy 1970), genetics (Harris and Smith 1948; Odell, Jackson, and Friday 1970), and psychology (Isaac 1969).

For this study, the parameters for each population are restricted such that \( \mu_2^i - \mu_1^i = \delta (i = 1, 2) \), \( \Sigma_2 = \sigma^2 \Sigma_1 \), where \( \Sigma_1 \) is positive definite and \( \sigma^2 > 0 \). For given \( \alpha^1 \) and \( \alpha^2 \), \( \delta \) specifies the degree of the location and \( \sigma^2 \) the degree of the scale mixture.

The mean vector and the covariance matrix of \( X \) from

* Takamaru Ashikaga is Associate Professor of Mathematics, College of Engineering, Mathematics and Business Administration, and Director of the Biometry Facility, College of Medicine, University of Vermont, Burlington, VT 05405. Potter C. Chang is Associate Professor of Biostatistics, and Head of the Division of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90024.

© Journal of the American Statistical Association
September 1981, Volume 76, Number 375
Theory and Methods Section
II, are
\[ \mu^{(i)} = \mu_{1}^{(i)} + \alpha^{(i)} \delta \]  
and
\[ \Sigma^{(i)} = (1 - \alpha^{(i)} + \sigma^2 \alpha^{(i)}) \Sigma_{1} + \alpha^{(i)} (1 - \alpha^{(i)}) \delta \delta^t. \]  
Thus when \( \alpha^{(1)} = \alpha^{(2)} \), the distributions of \( \Pi_{1} \) and \( \Pi_{2} \) are identical except for a location shift.

For convenience, the distance between \( \Pi_{1} \) and \( \Pi_{2} \) is measured by
\[ \Delta^2 = (\mu^{(1)})^t \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) \]  
where
\[ \Sigma = \frac{1}{2}(\Sigma^{(1)} + \Sigma^{(2)}). \]  
This distance measure is used to organize the computation of error rates as well as the presentation of results.

Since in most applications of Fisher's linear discriminant function the a priori probabilities and the costs of misclassification are assumed equal, these assumptions are made in this study.

### 2.2 A Canonical Form for the Mixed Normal Model

Through appropriate transformations the densities for \( \Pi_{1} \) and \( \Pi_{2} \) can be written as
\[ f_{1}(x) = (1 - \alpha^{(1)}) n(0, I) + \alpha^{(1)} n(\delta, \sigma^2 I) \]  
and
\[ f_{2}(x) = (1 - \alpha^{(2)}) n(\mu, I) + \alpha^{(2)} n(\mu + \delta, \sigma^2 I), \]  
where \( \theta = (0, \ldots, 0)' \), \( \delta = (\delta_1, \delta_2, 0, \ldots, 0)' \), and \( \mu = (\mu, 0, \ldots, 0)' \) all have nonnegative vector components. The population means are then
\[ \mu^{(1)} = \alpha^{(1)} \delta \]  
and
\[ \mu^{(2)} = \mu + \alpha^{(2)} \delta \]  
with population covariance matrices \( i = 1, 2 \)
\[ \Sigma^{(i)} = (1 - \alpha^{(i)} + \sigma^2 \alpha^{(i)}) I + \alpha^{(i)} (1 - \alpha^{(i)}) \delta \delta^t. \]

It follows that the distance measure \( \Delta^2 \) becomes
\[ \Delta^2 = \mu' \Sigma^{-1} \mu + 2 \mu' \Sigma^{-1} ((\alpha^{(2)} - \alpha^{(1)}) \delta) \]  
\[ + ((\alpha^{(2)} - \alpha^{(1)}) \delta)' \Sigma^{-1} \delta, \]  
where \( \Sigma \) was given in (2.5).

A major advantage of this canonical form is that the number of unspecified parameters is reduced. Thus the task of parameter selection for the numerical comparison of error rates (Sec. 4) is simplified.

### 3. THE PROCEDURES AND THEIR RATES

Error rate (the average probability of misclassification) was used as the criterion of comparison. Expressions for error rates are stated in this section.

### 3.1 \( P_L \): The Error Rate Using the Linear Rule

For the two-component mixed normal models specified in Section 2, the linear procedure classifies \( x \), an observation of \( X \), into population \( \Pi_{1} \) if the linear function
\[ z = x' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) \]  
\[ - \frac{1}{2} (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} + \mu^{(2)}) \geq 0, \]
and into population \( \Pi_{2} \) otherwise. Here, (3.1) is the linear discriminant function based on \( \mu^{(1)} \) and \( \mu^{(2)} \) (2.2) and \( \Sigma \) (2.5). It is the population counterpart of the linear discriminant function determined from sample data when the distributions of the two populations are specified by (2.1). Equation (3.1) defines a plane in a \( p \)-dimensional space. The average misclassification probability for this procedure under a two-component mixed normal model is
\[ P_L = \frac{1}{2} \left\{ \int_{R_{1}} f_{1}(t) \, dt + \int_{R_{2}} f_{2}(t) \, dt \right\}. \]  
where \( R \) represents the \( p \)-dimensional region associated with classification into \( \Pi_{1} \), \( i = 1, 2 \). \( P_L \) is actually a weighted sum of four univariate normal probabilities. These four probabilities are determined by the orthogonal distances of the mixed normal component means from the separating plane defined by inequality (3.1) and by the magnitude of \( \sigma^2 \). The distances \( d_{(i)} (i = 1, 2; j = 1, 2) \) are
\[ d_{(i)} = |\mu_{(i)}' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})| \]  
\[ - c / || \Sigma^{-1} (\mu^{(1)} + \mu^{(2)}) ||, \]
where \( c = \frac{1}{2} (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} + \mu^{(2)}) \) and \( || \cdot \|| \) represents the usual Euclidian norm of a vector.

Under the canonical form of the mixed normal model, \( P_L \) is completely determined by the parameters \( \alpha^{(i)}, \mu, \delta_{1}, \delta_{2}, \) and \( \sigma^2 \) for all values of \( \rho \).

### 3.2 \( P_B \): The Error Rate Using the Bayes Rule

Since the costs and the a priori probabilities are assumed to be equal, the Bayes classification procedure is to classify an observation \( x \) into population \( \Pi_{1} \) if \( f_{1}(x) / f_{2}(x) \geq 1 \), and into \( \Pi_{2} \) otherwise.

The average probability of misclassification for this procedure can be expressed as
\[ P_B = \frac{1}{2} \left( \int_{f_{1} > f_{2}} f_{2}(t) \, dt + \int_{f_{1} < f_{2}} f_{1}(t) \, dt \right) \]  
or
\[ P_B = \frac{1}{2} - \frac{1}{2} \int | f_{1}(t) - f_{2}(t) | \, dt. \]

### 3.3 Special Cases

Several cases are of special interest. One occurs when \( \alpha^{(1)} = \alpha^{(2)} = \alpha \) and \( \delta = (0, \delta_{2}, \ldots, 0)' \). It can be shown that in this case the linear procedure is equivalent to the Bayes procedure. Since under the mixed model of this study the distributions of the two populations deviate
markedly from normality when $\alpha = .5$ and $\beta_2$ and $\sigma^2$ are large, this special case suggests that the linear procedure can be optimal even if the populations deviate from normal.

Another special case occurs when $\sigma^2 = 1$. In this case it can be shown that the ratio $f_1(x)/f_2(x)$ is completely determined by the first two elements of $x = (x_1, x_2, \ldots, x_p)$.

4. PARAMETER SELECTION

To evaluate the robustness of the linear procedure in more detail, error rates for the Bayes and the linear procedure were computed. For convenience, the canonical form of the mixed normal model was used to specify the distributions for which error rates were determined. The parameters for the canonical form of the two-component mixed normal model were obtained by using the IBM System/360 Model 91 of the Health Sciences Computing Facility, University of California at Los Angeles. The cumulative distribution function of a normal distribution was computed by using the approximation of Hastings (1955). The error rates for the Bayes procedure were computed by using the method of importance sampling (Hammersley and Handscomb 1958). This technique took advantage of the fact that the integral $\int |f_1(t) - f_2(t)| \, dt$ in (3.4) could be expressed as

$$\int \frac{|f_1(t) - f_2(t)|}{g(t)} g(t) \, dt,$$

where $g(t)$ was the density of any convenient continuous distribution function $G(t)$. Numerical integration of (4.1) was then obtained through summation of the ratio $|f_1(t) - f_2(t)|/g(t)$ evaluated at values of $t$ generated from the distribution $G(t)$. For this study $g(t)$ was taken as $\frac{1}{2}f_1(t) + \frac{1}{2}f_2(t)$. Each value of $P_n$ was obtained by generating 3,000 observation vectors from $G(t)$. The accuracy of this procedure was checked by setting $f_1(x) = 1 - f_2(x)$. It was noticed that $P_n$ equaled $\Phi(-\Delta/2)$ where $\Phi(\cdot)$ is the cumulative distribution of the standard normal. For various values of $\Delta^2 (1.0, 4.0, 9.0)$, the average of the deviations between $\Phi(\Delta)/2$ and $P_n$ determined by the important sampling procedure were less than .001.

5. RESULTS

Several measures of robustness were considered. These included $(P_L - P_B)$, $(P_L - P_B)/P_B$, and $(P_L - P_B)/(1 - P_B)$. Since similar results were obtained, it was decided to present $(P_L - P_B)$ and to use $\Phi(-\Delta/2)$ as a reference value for each given $\Delta^2$. Here $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and $\Phi(-\Delta/2)$ is the error rate of LDF if the distribution of the two populations were $N(\mu_i^*, \Sigma), i = 1, 2$.

Several measures of nonnormality were used to organize the results. These included the measures of skewness and kurtosis used by Malkovich and Affifi (1973) and $N^*$, a general measure of nonnormality based on the principle of the likelihood ratio test. The values of $(P_L - P_B)$ were plotted against each of these measures of nonnormality. It was noticed that $(P_L - P_B)$ did not seem to be related to the measures of skewness and kurtosis in any specific fashion. They were, however, roughly related to $N^*$. For this reason, this measure of normality was used to present the data and is briefly described in the following subsection. Further details on the relationship of $(P_L / P_B)$ to the measures of skewness and kurtosis can be found in Ashikaga (1973).

5.1 $N^*$: A Measure of Nonnormality

This measure of nonnormality was constructed by using the likelihood ratio test procedure. Suppose that

<table>
<thead>
<tr>
<th>Table 1. Parameter Configurations Examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension $p$</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>
the interest is to test whether a distribution is either some known multivariate mixed normal distribution of form (2.1) or a multivariate normal with known mean vector (2.2) and known covariance matrix (2.3). It can be shown that

\[ N^* = 4(\sigma_0 + \sigma_A)^2((\mu_0 - \mu_A)^2 \]

was the sample size needed to detect the above distributional difference with significance level .05 and power .95. Here \( \mu_0, \sigma_0^2, \mu_A, \sigma_A^2 \) represent the means and variances of the log-likelihood ratio when the null and the alternative hypotheses are respectively a normal and a mixed normal distribution with a specific parameter configuration.

Determined in this manner, \( N^* \) is a general measure of normality. Larger values of \( N^* \) represent minute deviations from normality, and smaller values indicate more marked deviations. Since the degree of nonnormality of a mixed normal distribution is affected by each of the parameters \( \alpha^{(0)}, \delta, \sigma^2 \), similar values of \( N^* \) were obtained for various parameter configurations. For example, for \( p = 2 \), \( N^* \) is greater than 10,000 when \( \| \delta \| = 1, \sigma^2 = 1, \) and \( .1 \leq \alpha^{(2)} \leq .9; \) \( N^* \) is less than 100 when \( \| \delta \| = 3, \sigma^2 = 9, \) and \( .1 \leq \alpha^{(1)} \leq .7 \).

5.2 Maximum of \( (P_L - P_H) \)

Values of \( (P_L - P_H) \) corresponding to each value of \( N^* \) were examined. It was noticed that for each class of \( N^* \), values of \( (P_L - P_H) \) varied substantially with zero (when \( \alpha^{(1)} = \alpha^{(2)} = 0, 0, 0, \ldots, 0) \) as the lower bound. Since the linear discriminant function is robust if the maximum of \( (P_L - P_H) \) is small and is not necessarily robust if only the mean of \( (P_L - P_H) \) is small, max\( (P_L - P_H) \) for each value of \( \Delta^2 \) and each class of \( N^* \) is presented in Tables 2 and 3. It is, however, important to note that for each \( \Delta^2 \) and each class of \( N^* \), max\( (P_L - P_H) \) is only the maximum for the parameter configurations included in this study and is not the maximum of all possible parameter configurations.

In Tables 2 and 3 the values of \( d(-\Delta/2) \) are given as a reference in the parentheses following each value of \( \Delta^2 \).

5.3 Identical Distributions

When \( \alpha^{(1)} = \alpha^{(2)} \), the distributions of the two populations are identical in shape and differ only by a location shift. Table 2 gives values of max\( (P_L - P_H) \) for this case. It is observed that

1. Generally max\( (P_L - P_H) \) is larger when \( p = 2 \) than when \( p = 1; \)
2. \( \max(P_L - P_H) \) tended to decrease when the distance measure between the two populations \( \Delta^2 \) was increased;
3. When \( \Delta^2 \leq 2 \) or \( N^* > 100 \), max\( (P_L - P_H) \) decreased in a monotone fashion as the measure of nonnormality \( N^* \) was increased;
4. Values of max\( (P_L - P_H) \) were small \((<.007) \) when both \( N^* \geq 600 \) and \( \Delta^2 \geq 4 \), and not large \((<.029) \) when either \( N^* \geq 600 \) or \( \Delta^2 \geq 4 \).

5.4 Nonidentical Distributions

Table 3 gives the values of max\( (P_L - P_H) \) for the case of nonidentical population distributions \((i.e., \alpha^{(1)} \neq \alpha^{(2)}). \) In general, values of max\( (P_L - P_H) \) are larger for this case than for the corresponding cases when the two distributions have identical form. The same effect of dimensionality was observed as was in Table 2. When \( p \)

<table>
<thead>
<tr>
<th>Table 2. ( \max(P_L - P_H): ) Identical Distributions With Location Mix ( [\alpha^{(1)} = \alpha^{(2)} = 0.1, 1.9] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta^2 ) &amp; ( N^* ) &amp; ( p = 1 ) &amp; ( p = 2 ) &amp; ( p = 1 ) &amp; ( p = 2 ) &amp; ( p = 1 ) &amp; ( p = 2 ) &amp; ( p = 1 ) &amp; ( p = 2 )</td>
</tr>
<tr>
<td>( \Delta^2 = 1(31) ) &amp; ( \leq 100 ) &amp; .075 &amp; .090 &amp; .050 &amp; .060 &amp; .024 &amp; .041 &amp; .012 &amp; .024</td>
</tr>
<tr>
<td>100-300 &amp; .080 &amp; .098 &amp; .044 &amp; .052 &amp; .030 &amp; .039 &amp; .020 &amp; .026 &amp; .010 &amp; .021</td>
</tr>
<tr>
<td>300-600 &amp; .041 &amp; .056 &amp; .043 &amp; .040 &amp; .024 &amp; .028 &amp; .012 &amp; .017 &amp; .008 &amp; .010</td>
</tr>
<tr>
<td>600-1000 &amp; .010 &amp; .029 &amp; .025 &amp; .007 &amp; .014 &amp; .008 &amp; .006 &amp; .007 &amp; .007 &amp; .004</td>
</tr>
<tr>
<td>1000-10000 &amp; .006 &amp; .008 &amp; .006 &amp; .006 &amp; .003 &amp; .004 &amp; .003 &amp; .005 &amp; .003 &amp; .003</td>
</tr>
<tr>
<td>&gt;10000 &amp; .001 &amp; .003 &amp; .005 &amp; .005 &amp; .002 &amp; .004 &amp; .004 &amp; .005 &amp; .003 &amp; .003</td>
</tr>
</tbody>
</table>

\(* \) The value given in the parenthesis following each \( \Delta^2 \) is the error rate of LDF if the distributions of the two populations were \( N(\mu^1, \Sigma), i = 1, 2. \)
of max(P, \(P_a\)) tended to decrease as \(\Delta^2\) was increased. But this was not the case when \(p = 1\). Max(P, \(P_a\)) was not a monotone function of \(N^*\) and was small only when both \(N^* \geq 600\) and \(\Delta^2 \geq 4\).

6. DISCUSSION

Since the values of (\(P_L - P_a\)) were computed for the case of known parameters, the results of this study are useful only when LDF's are determined from large data sets. However, they also are valuable references when sampling experiments are designed to investigate the robustness of LDF for small samples.

The contrast between the values of max(P, \(P_a\)) in Table 2 and those in Table 3 suggest that a more important issue than nonnormality is whether the distributions of the two populations are similar in shape. When the shapes are identical, LDF is robust (max(P, \(P_a\)) \(\leq .007\)) as long as the two populations are not too close (\(\Delta^2 \geq 4\)) and the distributions do not markedly deviate from normality (\(N^* \geq 600\)); and is relatively robust (max(P, \(P_a\)) \(\leq .025\)) as long as \(N^* \geq 600\) or \(\Delta^2 \geq 4\). When the two distributions differ in shape, LDF is not as robust. It is only relatively robust when both \(N^* \geq 600\) and \(\Delta^2 \geq 4\).

This result is rather encouraging. For the use of discrimination procedures is of more interest when the two populations are not too close (say, \(\Delta^2 \geq 4\)) and the potential error rate is not too large (say, \(.16\) or smaller). This is when LDF is relatively robust, particularly if the distributions of the two populations are not substantially different in shape and do not markedly deviate from normality.

Also note that max(P, \(P_a\)) is not a monotone function of \(N^*\). This is partly because values of (P_L - P_a) are small for some mixed normal distributions that are quite nonnormal. A case in point is the special case of identical \(N^*\). This is partly because values of (PL - PA) are small for some mixed normal distributions that are quite nonnormal. A case in point is the special case of identical \(N^*\).

REFERENCES


