A SAS MACRO FOR CALCULATING THE QUADRATIC DISCRIMINANT FUNCTION

Charles S. Davis, Mayo Clinic

ABSTRACT

An important assumption underlying the use of the linear discriminant function is homogeneity of the within-group covariance matrices. In practice, this equal covariance assumption is often not satisfied. The facility for quadratic discrimination, which does not require the assumption of equal covariances, is a major advantage of the SAS DISCRIM procedure. When the pooled covariance matrix is used in the discriminant model, the linear discriminant function is always printed. Unfortunately, when the within-group covariance matrices are used, the quadratic discriminant function (QDF) is not similarly available. A macro, utilizing PROC MATRIX and the output data set created by PROC DISCRIM, for computing and printing the QDF is described.

1. INTRODUCTION

The basic problem in discriminant analysis is to assign, based on a multivariate observation, an unknown subject to one of two or more distinct groups. Although methods of discrimination based on other assumptions have been proposed, in this paper the underlying probability distributions are assumed to be multivariate normal. Fisher's linear discriminant function is optimal in the sense of minimizing the overall probability of misclassification when the sampling is from multivariate normal populations with known means and known, equal covariance matrices (Lachenbruch, 1975).

In many applications, the assumption of equal covariance matrices is not justified. When population parameters are known and covariance matrices are unequal, the quadratic discriminant function (QDF) is optimal in the sense of minimizing the overall probability of misclassification. In spite of the theoretical evidence supporting the use of the QDF when covariance matrices are not equal, the use of quadratic discrimination is infrequent, due, in part, to the complicated form of the QDF (see e.g. Kendall and Stuart, 1976).

As was noted by Koonce and Icaza (1978) and Lachenbruch (1979), quadratic discrimination is a simple matter in SAS. The DISCRIM procedure incorporates a test of the equality of the within-group covariance matrices. The result of this test (or optionally, user specification) determines whether linear or quadratic discrimination is utilized. When DISCRIM is used for linear discrimination, the LDF is always printed. Unfortunately, the QDF is not similarly available. Thus, classification of future observations is possible only when DISCRIM's output data set, which contains the calibration information, is saved.

The purpose of this paper is to describe a macro, QUAD DIS, which was developed to compute and print the QDF. The motivation for this research was a consulting application of quadratic discrimination. In this application, an investigator, with no access to SAS, was required to personally classify future observations. As mentioned above, the necessary discriminating function is not provided by PROC DISCRIM.

First, the necessary notation and definitions are presented, followed by the derivation of an algebraic expression for the QDF. The structure of DISCRIM's output data set in the case of quadratic discrimination is then described. After a description of the macro, two examples of its usage are displayed. We conclude with a discussion of the strengths and weaknesses of the proposed technique, with special reference to some of the more recently published results concerning quadratic discrimination.

2. NOTATION AND DEFINITIONS

For observations from multivariate populations \( \mu_i \) (i=1, 2, ..., k), the n-dimensional random vector \( X \) is assumed to be normally distributed with mean \( \mu_i \) and covariance matrix \( \Sigma_i \). Based on samples from each of the \( k \) populations, the unknown mean vectors and covariance matrices are estimated by \( \hat{\mu}_i \) and \( \hat{\Sigma}_i \), respectively. Also, let \( p_i \) denote the a priori probability that an observation \( X \) comes from population \( \mu_i \).

The generalized squared distance, \( D_i^2(x) \), from an observation \( X = x \) to population \( \mu_i \) is defined as:

\[
D_i^2(x) = (x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (x - \hat{\mu}_i) + \ln |\hat{\Sigma}_i| - 2 \ln p_i.
\]

where \( \ln |\hat{\Sigma}_i| \) is the natural logarithm of the determinant of \( \hat{\Sigma}_i \). The quadratic discriminant function classifying an observation \( X = x \) into population \( \mu_j \) if \( D_j^2(x) \) is the minimum of the \( k \) distances \( D_i^2(x) \).

In the case of only two populations, the QDF reduces to:

\[
Q(x) = D_1^2(x) - D_2^2(x) = (x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1} (x - \hat{\mu}_1) - (x - \hat{\mu}_2)^T \hat{\Sigma}_2^{-1} (x - \hat{\mu}_2) + \ln |\hat{\Sigma}_1|/|\hat{\Sigma}_2| + 2 \ln \frac{p_1}{p_2}.
\]

An observation \( x \) is classified into \( \mu_j \) if \( Q(x)<0 \). When \( E_i \equiv \Sigma_i \) is estimated by the pooled covariance matrix \( \hat{\Sigma} \), Q(x) further reduces to:

\[
L(x) = (x - \hat{\mu}(x))^T \hat{\Sigma}^{-1} (x - \hat{\mu}(x)) + \ln \frac{p_1}{p_2}.
\]

In this case, \( x \) is classified into \( \mu_1 \) if \( L(x) \), the LDF, is less than zero.
3. An Algebraic Expression for the QDF

Given a new observation \( x=(x_1, x_2, \ldots, x_n)^T \), \( D_i(x) \) can readily be evaluated. However, we require an algebraic expression of the form:

\[
D_i(x) = \sum_{j=1}^{n} a_{ij} x_j^2 + \sum_{j<k} b_{ijk} x_j x_k + \sum_{j=1}^{n} c_{ij} x_j + d_i,
\]

where the \( a_{ij} \), \( b_{ijk} \), and \( d_i \) values are constants to be determined. With reference to formula (1),

\[
D_i(x) = (x-\bar{x}_i)^T S_i^{-1} (x-\bar{x}_i) + \ln |S_i| - 2n \ln p_i,
\]

where \( \bar{x}_i \) is the mean of the \( i \)-th population. The coefficients \( a_i, b_{ij}, c_i \), and \( d_i \) are determined by the method of maximum likelihood.

Thus \( c_i = (c_{i1}, c_{i2}, \ldots, c_{in}) \) = \(-2\bar{x}_i^T S_i^{-1} \) and \( d_i = \bar{x}_i^T S_i^{-1} \bar{x}_i + \ln |S_i| - 2n \ln p_i \). Expanding \( x^T S_i^{-1} x \) yields \( x_j \) = \( S_{jj}^{-1} \) and \( b_{ijk} = 2 \bar{x}_i \bar{x}_j \bar{x}_k \) where \( S_{jj}^{-1} \) denotes the element in the \( j \)-th row and \( k \)-th column of \( S_i^{-1} \).

4. The Structure of DISCRIM's Output Data Set

The output data set created by DISCRIM, although not described in the User's Guide, contains the information required for the evaluation of the coefficients of the QDF. The structure of this data set, in the case of a quadratic discrimination, is shown in Display 1.

In general, the output data set contains \( n+4 \) variables. Here CLASSVAR denotes the classification variable and VAR1, VAR2, \ldots, VARN are the variables used for discrimination (in the same order as listed in the VAR statement). The variables TYPE, LNDDET and PRIOR are special SAS variables.

The output data set contains \( k(n+2)+1 \) observations. In observation 1, the variable TYPE assumes one of two values ('NOEQ' if prior probabilities are equal and 'NOPR' otherwise). Also in observation 1, the variable LNDDET indicates the number of groups or populations.

The remaining observations occur in \( k \) groups (one for each value of the classification variable) of \( n+2 \) observations, as shown in Display 1. The value \( p_i \) of PRIOR is defined as:

\[
p_i = \begin{cases} 
1 & \text{if prior probabilities are equal}, \\
-2n \ln p_i & \text{if prior probabilities are not all equal}.
\end{cases}
\]

The values \( \bar{x}_i \) and \( S_i \) are the mean and standard deviation of the \( j \)-th variable in the \( i \)-th population. The \( n \times n \) matrix \( R_i \) is defined such that:

\[
S_i^{-1} = S_{ij}^{-1} R_i^{-1},
\]

where \( S_{jj}^{-1} = 1/s_{jj}, 1/s_{jj}, \ldots, 1/s_{jn} \).

5. Description of the Macro

The macro QUAD DIS (listed in Display 2) accesses DISCRIM's output data set through PROC MATRIX. The number of variables in the new matrix created by the FETCH command is dependent upon the number of numeric variables in DISCRIM's output data set. Thus, it is first necessary to determine whether the classification variable is numeric or character. Then, the number of populations, \( K \), and the number of variables, \( N \), are determined. The preliminary part of the program concludes with a check to insure that the macro is not inadvertently applied to a linear discrimination problem.

The output of the macro is a \( K \times (N+2)(N+1)/2 \) matrix of coefficients. The \( i \)-th row contains the coefficients of the generalized squared distance function for the \( i \)-th population. The coefficients \( a_i, b \), \( c \), and \( d \), corresponding to the columns of the printed output matrix, are listed in the same order as in equation (2). In the special case of only two populations, the two-group QDF is also printed.

6. Examples

Fisher (1936) provided measurements, in centimeters, of four variables on 50 flowers from each of three varieties of Iris. These measurements are reproduced on page 331 of Kendall and Stuart (1976). The variables are sepal length (SEPAL L), sepal width (SEPAL W), petal length (PETAL L), and petal width (PETAL W), while the classification variable (IRIS) assumes one of the three values 'SETOSA', 'VERSICOLOR' or 'VIRGINICA'.

In our first example, we wish to discriminate between the two groups versicolor and virginica using the variables petal width and petal length. The SAS program statements are:

\[
\text{PROC DISCRIM POOL=TEST OUT=QUADRAT;}
\text{CLASS IRIS;}
\text{VAR PETAL W PETAL L;}
\text{QUAD DIS *}
\]

Since the hypothesis of homogeneity of the within-group covariance matrices is rejected (\( \chi^2=31.2, p=0.0001 \)), quadratic discrimination is appropriate. The output of QUAD DIS is shown in Display 3. The two-group discriminant function is given by:

\[
52.3 \text{ PETAL W}_2 + 8.2 \text{ PETAL L}_2 - 39.7 \text{ PETAL W} \times \text{ PETAL L} + 44.9 \text{ PETAL W} - 11.3 \text{ PETAL L} = 39.5.
\]

Figure 1 displays the original data, along with the above discriminating function. For comparison, the LDF is also plotted on Figure 1.
As another example of the use of QUAD DIS, we discriminate among all three Iris populations using all four variables. The SAS program statements are:

```
PROC DISCRIM POOL=TEST OUT=QUADRAT;
CLASS IRIS;
VAR SEPAL_L SEPAL_W PETAL_L PETAL_W;
QUAD_DIS
```

Homogeneity of the within-group covariance matrices is again rejected ($\chi^2 = 1.4$, p<0.0001). The output from QUAD DIS, giving the three generalized squared distance functions, is shown in Display 4. For example, the function for Iris setosa is given by:

$$
18.9 \text{SEPAL}_L^2 + 15.6 \text{SEPAL}_W^2 + 38.8 \text{PETAL}_L^2 + 106.0 \text{PETAL}_W^2 - 24.8 \text{SEPAL}_L \times \text{SEPAL}_W - 9.0 \text{SEPAL}_L \times \text{PETAL}_L - 9.6 \text{SEPAL}_L \times \text{PETAL}_W + 2.2 \text{SEPAL}_W \times \text{PETAL}_L - 4.2 \text{SEPAL}_W \times \text{PETAL}_W - 35.9 \text{PETAL}_L \times \text{PETAL}_W - 89.1 \text{SEPAL}_L + 15.2 \text{SEPAL}_W - 67.1 \text{PETAL}_L + 62.5 \text{PETAL}_W + 224.9.
$$

7. DISCUSSION

The ability to obtain the coefficients of the QDF, for graphical display or classification of future observations, greatly extends the quadratic discrimination capabilities of PROC DISCRIM. In addition, the macro is very easy to use, enabling this additional information to be obtained with minimal effort. Although quadratic discrimination is nearly automatic using SAS, some cautions noted in published studies merit emphasis.

Based on the results of a Monte Carlo study, Marks and Dunn (1974) indicate that for small samples (21 observations from each population) from normal populations, the QDF performs poorly in comparison with the LDF when the number of variables is large (>6). The poor showing of the QDF is emphasized even more as covariance differences decrease. Wahl and Kronmal (1977) extend the results of Marks and Dunn (1974) to larger sample sizes and provide a simple rule-of-thumb relating sample size from each population to the dimension of the multivariate normal distributions. With 4 variables, 25 observations are sufficient; for every additional two variables, another 25 observations is desirable. A related rule (Lachenbruch and Goldstein, 1979) is that, from each population, there should be three times as many observations as there are parameters to be estimated.

In sharp contrast with earlier published results for the LDF, Lachenbruch (1979) showed that initial misclassification of some of the observations used to compute the QDF can have serious effects on the error rate. Clarke, Lachenbruch and Broffitt (1979) indicate that the QDF is robust to non-normality only when the marginal distributions are not highly skewed. Other studies which may be of interest to potential users of quadratic discrimination include those of Broffitt, Clarke and Lachenbruch (1980), Van Ness (1979), and Van Ness and Simpson (1976).

REFERENCES


Display 1
Structure of PROC DISCRIM's Output Data Set in the Case of Quadratic Discrimination

<table>
<thead>
<tr>
<th>OBS</th>
<th>TYPE</th>
<th>CLASSVAR</th>
<th>_LDET</th>
<th>_PRIOR</th>
<th>VAR1</th>
<th>VAR2</th>
<th>VARn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NOEQ, NOPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MEAN</td>
<td>class 1</td>
<td>ln</td>
<td>S1</td>
<td></td>
<td>p^1</td>
<td>s11</td>
</tr>
<tr>
<td>3</td>
<td>STD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>RINV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n+3</td>
<td>RINV</td>
<td>class 1</td>
<td>ln</td>
<td>S1</td>
<td></td>
<td>p^1</td>
<td>s11</td>
</tr>
<tr>
<td>n+4</td>
<td>MEAN</td>
<td>class 2</td>
<td>ln</td>
<td>S2</td>
<td></td>
<td>p^2</td>
<td>s21</td>
</tr>
<tr>
<td>n+5</td>
<td>STD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n+6</td>
<td>RINV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2(n+2)+1</td>
<td>RINV</td>
<td>class 2</td>
<td>ln</td>
<td>S2</td>
<td></td>
<td>p^2</td>
<td></td>
</tr>
<tr>
<td>(k-1)(n+2)+2</td>
<td>MEAN</td>
<td>class k</td>
<td>ln</td>
<td>S_k</td>
<td></td>
<td>p^k</td>
<td>s_k1</td>
</tr>
<tr>
<td>(k-1)(n+2)+3</td>
<td>STD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(k-1)(n+2)+4</td>
<td>RINV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k(n+2)+1</td>
<td>RINV</td>
<td>class k</td>
<td>ln</td>
<td>S_k</td>
<td></td>
<td>p^k</td>
<td></td>
</tr>
</tbody>
</table>

Display 2
Listing of the Macro QUAD_DIS

MACRO QUAD_DIS

* USING THE OUTPUT DATA SET CREATED BY PROC DISCRIM, THE GENERALIZED SQUARED DISTANCE FUNCTIONS FOR QUADRATIC DISCRIMINATION ARE CALCULATED AND PRINTED

PROC MATRIX:
* DETERMINE DIMENSIONS OF OUTPUT DATA SET:
FETCH QUAD1 NROWS=NRQW (QUAD1) NCOLS=NCOLS (QUAD1)
* DETERMINE WHETHER CLASSIFICATION VARIABLE IS NUMERIC OR CHARACTER:
IF QUAD1 (1,1) = . THEN C=21 * NUMERIC CLASSIFICATION VARIABLE;
ELSE C=11 * CHARACTER CLASSIFICATION VARIABLE;
* K=NUMBER OF GROUPS OR POPULATIONS, N=NUMBER OF VARIABLES:
K=QUAD1(1,1) N=NCOLS-C-11 NPLUS2=N+21
* QUAD_DIS CAN ONLY BE USED FOR QUADRATIC DISCRIMINATION:
IF NROWS<=(NPLUS2*K+1) THEN DO;
NOTE INVALID USE OF MACRO QUAD_DIS STOP;
END;
* INITIALIZE MATRIX OF GENERALIZED SQUARED DISTANCES:
DSQUARED=J(K,(NPLUS2*(N+1))#2,011)

534
Display 2 (continued)

Listing of the Macro QUAD_DIS

* CALCULATE DISTANCE FUNCTION FOR EACH OF THE K GROUPS:
  DO I=1 TO K
    J=(I-1)*NPLUS2+2
    QUAD2=QUAD1(J:(J+N+1),:NCOLS)
  * XBAR=VECTOR OF SAMPLE MEANS FOR THE I-TH GROUP:
    XBAR=QUAD2(1:3:NPLUS2)
  * STD=VECTOR OF SAMPLE STANDARD DEVIATIONS FOR THE I-TH GROUP:
    STD=QUAD2(2:3:NPLUS2)
  * SINV=INVERSE OF VARIANCE-COVARIANCE MATRIX FOR THE I-TH GROUP:
    SINV=RECIP(STD*STD)*RINV
  * COMPUTE COEFFICIENTS OF SQUARED TERMS:
    SQUARED=VECDIAG(SINV)
  * COMPUTE COEFFICIENTS OF CROSS-PRODUCT TERMS:
    CROSSPRO=J((N*(N-1))/2)+1)
    M=0
    DO J=1 TO (N-1)
    DO L=(J+1) TO N
      M=M+1
      CROSSPRO(I,M)=2*SINV(J,L)
    ENDDO
    ENDDO
  * COMPUTE COEFFICIENTS OF LINEAR TERMS:
    XSINV=XBAR*SINV
    LINEAR=(-2.0)*XSINV
  * COMPUTE CONSTANT TERM:
    CONSTANT=XSINV*XBAR+QUAD2(1:1)-QUAD2(1:2)
    DSQUARED(I):=SQUARED+LINEAR+CONSTANT
  ENDDO
  * PRINT GENERALIZED SQUARED DISTANCE FUNCTIONS:
    NOTE GENERALIZED SQUARE DISTANCE FUNCTIONS PRINT DSQUARED
  * PRINT TWO-GROUP DISCRIMINANT FUNCTION:
    IF K=2 THEN DO
      FUNCTION=DSQUARED(1)-DSQUARED(2)
      NOTE TWO-GROUP DISCRIMINANT FUNCTION PRINT FUNCTION
    ENDDO
  %

Display 3

Output from QUAD_DIS for Example 1

GENERALIZED SQUARED DISTANCE FUNCTIONS

<table>
<thead>
<tr>
<th>DSQUARED</th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
<th>COL5</th>
<th>COL6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW1</td>
<td>67.0897</td>
<td>11.8814</td>
<td>-44.205</td>
<td>11.3098</td>
<td>-42.3282</td>
<td>16.4443</td>
</tr>
<tr>
<td>ROW2</td>
<td>14.7915</td>
<td>3.6632</td>
<td>-4.3020</td>
<td>11.3098</td>
<td>-42.3282</td>
<td>16.4443</td>
</tr>
</tbody>
</table>

TWO-GROUP DISCRIMINANT FUNCTION

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
<th>COL5</th>
<th>COL6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW1</td>
<td>52.2981</td>
<td>8.2182</td>
<td>-39.6785</td>
<td>44.917</td>
<td>-11.2596</td>
<td>-39.4637</td>
</tr>
</tbody>
</table>

535
Figure 1
Plot of Data, LDF and QDF for Example 1

Display 4
Output from QUAD.DIS for Example 2

<table>
<thead>
<tr>
<th>DSQUARED</th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
<th>COL5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW1</td>
<td>18.9434</td>
<td>15.5705</td>
<td>38.7762</td>
<td>106.046</td>
<td>-24.8097</td>
</tr>
<tr>
<td></td>
<td>-89.1058</td>
<td>15.2319</td>
<td>-67.1191</td>
<td>62.5117</td>
<td>224.899</td>
</tr>
<tr>
<td></td>
<td>-17.2634</td>
<td>12.4009</td>
<td>4.23204</td>
<td>-38.9606</td>
<td>-53.8745</td>
</tr>
<tr>
<td></td>
<td>-36.0257</td>
<td>-31.9214</td>
<td>-6.53757</td>
<td>29.4251</td>
<td>134.344</td>
</tr>
<tr>
<td>ROW3</td>
<td>10.5339</td>
<td>15.8754</td>
<td>13.4058</td>
<td>19.3141</td>
<td>-6.95945</td>
</tr>
<tr>
<td></td>
<td>-19.9203</td>
<td>3.5763</td>
<td>2.20538</td>
<td>-16.9457</td>
<td>-5.78184</td>
</tr>
</tbody>
</table>