A MACRO FOR TWO POPULATION NONPARAMETRIC UNIVARIATE DISCRIMINANT ANALYSIS WITH EXTENSIONS TO HIGHER DIMENSIONAL SPACES

Donald J. Henderson, ORI, Inc.
William A. Blattner, National Cancer Institute

I. Introduction

Discriminant analysis is a widely used technique, particularly in distinguishing between normal and diseased individuals. If the distributions of the populations are normal, procedure DISCRIM can be used. For data that is not normal, the SAS user is limited to procedure NEIGHBOR or one of several user-contributed logistic discrimination procedures. There are, however, many methodologies in addition to logistic or nearest neighbor methods for discriminant analysis. Unfortunately, none are readily available in SAS.

The MACRO presented here was developed in response to the need for greater nonparametric discriminant analysis capability. It determines the point at which the difference of the two cumulative distribution functions for the two populations is at its maximum. Overlayed plots of the two cumulative distribution functions, the plot of the difference of the cumulative distributions and a print of the range(s) at which the maximum difference occurs are also produced.

In section II, we discuss the statistical background of the algorithm used in MACRO _DISCRIM. Included is a discussion of methods for extending this standard univariate procedure to multivariate data. Section III gives instructions for using the MACRO and describes two examples. The first, a standard univariate application, and the second, an application utilizing a bivariate ranking algorithm to reduce the dimensionality of the data to one.

II. A Rank Based Discriminant Rule

Consider the two population discrimination problem in which the populations have continuous real-valued cumulative distribution functions F and G with corresponding densities f and g. If the distributions are completely known, then the optimal discriminant rule is based on the ratio f/g. Unfortunately, the distributions are rarely completely known, and so it is common practice to substitute sample-based estimators for f and g. For normally distributed data with only the means and variance(s) unknown, the formulation leads to the usual linear or quadratic discriminant rules which are available in procedure DISCRIM. While the linear and quadratic discriminant rules are widely applicable, they are sensitive to deviations from normality (Lachenbruch, Sneeringer and Revo 1973).

If one is not willing to assume a distributional form, then a nonparametric procedure is necessary. SAS procedure NEIGHBOR, which utilizes a k-Nearest Neighbor rule (Fix and Hodges 1951), provides some non-parametric discrimination capability; however it does have several major limitations. First, the choice of an appropriate k is not straightforward. Second, due to the discreteness of the computed posterior probabilities, it is difficult to incorporate prior probability or cost information into the rule in a meaningful way. Third, the classification regions that result for any given population are not necessarily contiguous. Last, in addition to having to select a value for k, it is also necessary to specify a distance metric. In small samples, the specification of an appropriate metric is almost as crucial for nonparametric procedures as the specification of the true distributional form is for parametric procedures. It has been shown in some limited Monte Carlo studies (Henderson 1979) that the selection of an inappropriate metric can cause serious performance problems for nonparametric discrimination rules.

There are many other nonparametric discriminant analysis procedures. Most notable are those based on kernel density estimators (Parzen 1962) or more recently, the rank-based methods of Randles, Broffitt, Ramberg and Hogg (1978). Most of them require the imposition of some type of metric. As discussed above, however, this can cause serious problems.

There are methods, however, which do not require the imposition of a distance metric. For the univariate case, both Stoller (1954) and Hudimoto (1956) proposed nonparametric rules based on the sample cumulative distribution functions (CDF's). They observed that the rule which has its classification regions defined by the inequality f/g > c, is mathematically equivalent to the rule which uses the point at which F - G is maximized in order to define the classification regions. Such rules are nonparametric or distribution-free since the sample CDF's are distribution-free and further do not require the specification of a distance metric. All that is required
is that some ordering or ranking be applied to the sample data. Since this is straightforward for univariate data, sample based CDF rules are very attractive for the univariate case.

Such sample based CDF rules can be extended to the multivariate case more easily than first impressions would indicate. It is first necessary to recognize that many discriminant rules are based on first reducing the dimensionality of the data to one. For example, the simplest form of the linear discriminant rule involves reducing the data to one dimension by first computing the difference of the Mahalanobis distances to the two populations and then classifying an unknown observation based on the sign of this difference. For the k-Nearest Neighbor rule and most density estimation based rules, the data is reduced to one dimension by computing the distance from an unknown to each known. The classification is then based on these distances. From this perspective, it is clear that the sample based CDF rules can be easily extended to the multivariate case. One approach would be to use a multivariate CDF as described by Das Gupta (1964). Alternatively, one could utilize any algorithm which assigns an ordering or ranking to the data. As discussed in Henderson (1979), the methods of constructing nonparametric multivariate tolerance regions (Tukey 1947; Fraser 1953) can be used to rank or order multivariate data.

Thus, through the use of data dependent or adaptive ranking schemes for multivariate data, the practitioner of discriminant analysis can avoid the pitfalls associated with having to choose a distance metric. Such ranking schemes are very intuitive and can be easily tailored to the application at hand. A simple example of such an adaptive ranking scheme is included in section III.

III. MACRO _DISCRIM - Description and Examples

MACRO _DISCRIM does univariate nonparametric discriminant analysis for the two population problem using the sample-based CDF rule discussed in the previous section. The MACRO does differ from that rule in that the constant c is assumed to be equal to 1. The MACRO can be easily modified to handle more general values for c.

The following statements are necessary to execute MACRO _DISCRIM:

MACRO _DATAIN Input datasetX
MACRO _GROUP classification variableX
MACRO _VAR dependent variableX
_DISCRIM

The dependent variable (_VAR) must be numeric. The classification variable (_GROUP) can be either character or numeric. The MACRO deletes all observations which have missing values for either the classification or dependent variables. If the classification variable is a character variable, blank is assumed to be missing. Several lines of error messages indicating illegal character to numeric conversion may result if the classification variable is character. These may be ignored. The presence of a MISSING statement for any of the characters _ or A-Z will cause observations to be deleted if the classification variable is character and its value is one of the single characters included in the MISSING statement. For example, the statement

MISSING A will cause observations with a value of "A" for a character classification variable to be deleted.

The printed output of _DISCRIM includes: the value of the maximum difference of the CDF's; a listing of all of the intervals of the dependent variable for which the difference of the sample CDF's is at its maximum; for each interval, the percent of the data from the stochastically smaller distribution which is less than all values in the interval and the percent of the data from the stochastically larger distribution which is greater than all values in the interval; an overlayed plot of the two CDF's; and a plot of the difference of the CDF's.

The first example is a standard univariate application of _DISCRIM. The variable MLCLNCPH has been measured on both cancer cases and normal control subjects. The cases are categorized into these two groups by the character variable GROUP. The data is stored in SAS dataset LTN.ALLTESTS. The following statements will execute _DISCRIM on this data:

MACRO _DATAIN LTN.ALLTESTSX
MACRO _GROUP GROUPX
MACRO _VAR MLCLNCPHX
_DISCRIM

The output will then indicate whether the variable MLCLNCPH can discriminate cancer cases from normals.

MACRO _DISCRIM can be applied to multivariate data. The user must first, however, create a variable whose values represent the ordering or ranking desired for the original multivariate data. Consider an extension of the above example in which we have two variables, MLCLNCPH and CONLNCMPH, which we want to use jointly. The desired ordering or ranking algorithm is illustrated in Figure 3.1.
STANDARDIZED MLCLNCMP

FIGURE 3.1 BIVARIATE RANKING ALGORITHM

We wish to assign low ranks if either of the two variables has a small (depressed) value. We also wish to assign low ranks to observations for which the individual values may not be depressed, but the combination of values is depressed. This is illustrated by the point in Figure 3.1 which is assigned a rank of 3. The diagonal ranking functions are used to determine whether an observation is depressed in a bivariate sense.

A PROC MATRIX program was written to perform the ranking illustrated in Figure 3.1; unfortunately, this program used excessive amounts of CPU time. For this reason, a PL/I program was written to do the ranking. We execute the PL/I program from within SAS through the use of the external procedure feature. The following statements were then used to apply _DISCRIM to the multivariate data by first computing bivariate ranks from MLCLNCMP and then executing _DISCRIM on the ranks:

```plaintext
PROC STANDARD MEAN=0 STD=1
DATA=LTN.ALLTESTS OUT=A;
%STANDARDIZE THE DATA;
VAR MLCLNCMP CONLNCPM;

DATA _NULL_
%WRITE OUT THE DATA TO BE RANKED;
SET A;
FILE INDATA;
PUT (MLCLNCMP CONLNCPM) (9.6 9.6);

PROC BIVRANK; %ASSIGN THE RANKS;
DATA A;
%READ IN THE ORIGINAL DATA;

SET A;
%MERGE THE RANKS IN WITH IT;
INFILE OUTDATA;
INPUT BIVRANK;

%EXECUTE _DISCRIM;
MACRO _DATAIN A;
MACRO _GROUP GROUPX;
MACRO _VAR BIVRANK;
_DISCRIM
```

The output produced is given in Figure 3.2.

References


NONPARAMETRIC UNIVARIATE DISCRIMINANT ANALYSIS MACRO

DEPENDENT VARIABLE: BIVRANK  CLASSIFICATION VARIABLE: GROUP

LISTING OF ALL RANGES FOR THE OPTIMAL CUTPOINT
A MAXIMUM DIFFERENCE OF THE CDF'S OF -28.7713 OCCURS IN THE
INTERVAL [156, 157].
68.87% OF POPULATION 2 (CANCER) IS BELOW CUTPOINT
59.90% OF POPULATION 1 (NORMAL) IS ABOVE CUTPOINT

PLOT OF THE CUMULATIVE DISTRIBUTION FUNCTIONS
PLOT OF CDF1*BIVRANK  SYMBOL USED IS 1
PLOT OF CDF2*BIVRANK  SYMBOL USED IS 2

CDF1

100  
75  
50  
25  
0  

CDF2

2222222222 111111  
2222222222 111111  
2222222222 111111  
2222222222 111111  

BIVRANK

PLOT OF THE DIFFERENCE OF THE CDF'S
PLOT OF DIFF*BIVRANK  LEGEND: A = 1 OBS, B = 2 OBS, ETC.

DIFF

0  + BC  
-10  + AB  
-20  + BA  
-30  + CC

CCDA  
AEDBA  
BD A  
ACD  

DIFF

0  + BC  
-10  + AB  
-20  + BA  
-30  + CC

BIVRANK

FIGURE 3.2 EXAMPLE OUTPUT FROM BIVARIATE EXTENSION OF MACRO _DISCRIM
BIVRANK IS BIVARIATE RANK COMPUTED FROM MLCLNCPM AND CONLNCPM
APPENDIX A. LISTING OF MACRO _DISCRIM

MACRO _DISCRIM
*THIS MACRO DOES UNIVARIATE NONPARAMETRIC DISCRIMINANT ANALYSIS FOR THE TWO POPULATION PROBLEM BY DETERMINING THE POINT AT WHICH THE DIFFERENCE OF THE TWO CDF'S IS AT ITS MAXIMUM.*
DATA DISCRIM(KEEP=_GROUP _VAR)
   COUNTS(KEEP=POPLAB1 POPLAB2 POPN1 POPN2);
*CHECK OUT DATA AND OBTAIN POPULATION COUNTS;
SET _DATAIN(KEEP=_GROUP _VAR) END=EOF;
*CHECK FOR MISSING VALUES;
OPTIONS MISSING=' ' ERRORS=1;
SPECMISS=(_GROUP=, OR A<=_GROUP<=Z);}SPECIAL NUMERIC OR NUMERIC MISSING;
ORDMISS=(VERIFY(_GROUP, ' ')=0));}ORDINARY CHARACTER OR NUMERIC MISSING;
IF _VAR LE 2 OR SPECMISS OR ORDMISS THEN /*MISSING VALUES*/;
ELSE DO /*DATA NOT MISSING*/;
OUTPUT DISCRIM;
*CHECK THAT THERE ARE ONLY TWO LEVELS OF THE GROUPING VARIABLE AND COUNT THE NUMBER OF OBSERVATIONS FOR EACH LEVEL;
RETAIN POPLAB1 POPLAB2 POPN1 POPN2; RETAIN GT3LEVS 'N';
IF POPN1= OR POPN2= THEN DO /*DEFINE LEVELS OF GROUPING VARIABLE*/;
   IF POPN1= THEN
   DO /*DEFINE FIRST LEVEL*/;
      POPLAB1=_GROUP;
   POPN1+1
   END /*DEFINE FIRST LEVEL*/;
   ELSE DO /*DEFINE SECOND LEVEL*/;
      IF _GROUP NE POPLAB1 THEN
      DO /*DEFINE SECOND LEVEL*/;
         POPLAB2=_GROUP;
      POPN2+1
      END /*DEFINE SECOND LEVEL*/;
   ELSE POPN1+1; /*FIRST LEVEL STILL. ADD 1 TO COUNTER*/
   END /*CHECK OR SECOND LEVEL*/;
ELSE DO /*VERIFY ONLY TWO LEVELS AND COMPUTE OBSERVATION COUNTS*/;
   IF _GROUP EQ POPLAB1 THEN POPN1+1;
   ELSE IF _GROUP EQ POPLAB2 THEN POPN2+1;
   ELSE DO /*MORE THAN TWO LEVELS*/;
      GT3LEVS='Y';
      PUT _GROUP= ' VALUE DISCOVERED. TWO DIFFERENT LEVELS,';
      POPLAB1= POPLAB2= ' ALREADY IDENTIFIED.';
   END /*MORE THAN TWO LEVELS*/;
   END /*VERIFY ONLY TWO LEVELS AND COMPUTE OBSERVATION COUNTS*/;
END /*DATA NOT MISSING*/;
IF EOF THEN
*FINAL CHECK OF DATA AND OUTPUT LABELS AND COUNTS*/;
IF POPN1 EQ . OR POPN2 EQ . OR GT3LEVS EQ 'Y' THEN DO /*ERRORS IN DATA*/;
   PUT 'EXECUTION ABORTING - 5';
   IF GT3LEVS EQ 'Y' THEN PUT 'MORE THAN 2 POPULATIONS';
   ELSE IF POPN1 EQ . THEN PUT 'NO NON-MISSING OBSERVATIONS';
   ELSE PUT 'ONLY 1 POPULATION: ' _GROUP=;
   ABORT;
END /*ERRORS IN DATA*/;
ELSE OUTPUT COUNTS /*DATA OK*/;
END /*FINAL CHECK OF DATA AND OUTPUT LABELS AND COUNTS*/;
PROC SORT DATA=DISCRIM by _VAR _GROUP; *ORDER THE DATA;
DATA CDFS(RENAME=(XVALUE=_VAR)
   KEEP=XVALUE CDF1 CDF2 DIFF POPLAB1 POPLAB2)
   MAXPTRS(KEEP=OBSNOS);
*READ INPUT DATA AND COMPUTE THE CDF'S. STORE THEM IN DATA SET CDFS. STORE POINTERS TO MAX DIFFERENCE IN DATA SET MAXPTRS;
PROC PLOT DATA='CDFS';
     PLOT CDF1=_VAR='1' CDF2=_VAR='2';/OVERLAY VAXIS=0 TO 100 BY 25;
     TITLE9 'PLOT OF THE CUMULATIVE DISTRIBUTION FUNCTIONS';
PROC PLOT DATA='CDFS';
     PLOT DIFF=_VAR;
     TITLE10 'PLOT OF THE DIFFERENCE OF THE CDF`S';
RUN; OPTIONS OBS=MAX;