CLASSIFYING AN INCOMPLETE VECTOR

Tom Bohannon, Appalachian State University
W. B. Smith, Texas A.& M. University

Introduction.

Discriminant analysis is concerned with the problem of assigning an observation vector, Z, of unknown origin to one of several distinct populations on the basis of some classification rule. Hodges [1950], Cacoullos [1973], and Lachenbruch [1975] give comprehensive lists of some case studies of various applications of discriminant analysis.

In our study, we shall only consider the situation where there are two p-variate normal populations Π_1 and Π_2 with distributions denoted by $N_p(\mu,\Sigma)$ and $N_p(\omega,\Sigma)$, respectively. We shall also assume that the probabilities of misclassification and the costs of misclassification are equal for the populations, thus the optimum classification rule is given by

$$D(X) = [X - 1/2(\mu + \omega)]'\Sigma^{-1}(\mu - \omega) .$$

In practice these parameters are usually not known and are estimated. These estimates are then substituted into the discriminant function, $D(X)$, to yield what is often referred to as Anderson's discriminant function, $W(X)$

$$W(X) = [X - (\frac{\hat{\mu} + \hat{\omega}}{2})]\hat{\Sigma}^{-1}(\hat{\mu} - \hat{\omega}) ,$$

where

$$\hat{\Sigma} = \frac{(n_1-1)\hat{\Sigma}_1 + (n - 1)\hat{\Sigma}_2}{n_1+ n_2- 1}$$

$\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ being the unbiased estimates of Σ_1 and Σ_2 based on the data collected from their respective populations. An additional unclassified observation Z will be classified into Π_1 if $W(Z)$ is non-negative, otherwise into Π_2.

We consider the problem of classifying an observation vector,

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

where Z_1 is a vector of observations and Z_2 is a vector with the components missing. The following notation will be used to denote the partitioned mean vectors and variance-covariance matrix:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Similar notation will be used for the estimates of these vectors and matrices.

Review of Discriminant Analysis With Missing Data

The use of discriminant analysis techniques on incomplete data sets is an area where very little research has been done. Jackson [1968] investigated a problem which had missing values in a discriminant problem where both the number of variables and the number of observations were large. Estimation of missing values and the number of observations were large. Estimation of missing values using mean and regression techniques were tried for the problem under study. The estimation procedure utilizing missing data gave more realistic results than the often used procedure of ignoring observation vectors with missing values.

Chan and Dunn [1972] investigated the problem of constructing a discriminant function based on samples which contain incomplete observation vectors. Several methods of estimating the components of the incomplete vectors were used to contruct the discriminant function. The effect on the performance of the discriminant function for each method was studied and compared. They concluded that no method is best for every situation, and gave guidelines to use in choosing the best method. Chan and Dunn [1974] studied the asymptotic behavior of these methods when the variables are equally correlated. They found that the differences of the asymptotic probability of correct classification from maximum were found to be small for all methods. Chan, Gilman and Dunn [1976] studied two additional methods and recommended their modified regression method.

Srivastava and Zaatar [1972] derived the maximum likelihood rule for incomplete data when a common covariance matrix is assumed known. Smith and Zeis [1973], using a generalization to the maximum likelihood estimation technique of Hocking and Smith [1968] and an application of the likelihood ratio criterion generalized the results of Srivastava and Zaatar to unknown and unequal covariance matrices. Bohannon [1976] compared this procedure to the standard procedure of ignoring the incomplete observations in the construction of the classification rule. The comparisons were made on the basis of the probability of misclassification and the proposed method performed best in the simulation study.

Classification Rules

Marginal Rule

If one ignores the variables in the vector Z_2 then the optimum rule is

$$V_1 = \left(Z_1 - \frac{\mu_1 + \omega_1}{2}\right)' \Sigma_{11}^{-1}\left(\mu_1 - \omega_1\right).$$

This function shall be referred to as the marginal discriminant function. This function also results if one uses the regression approach and estimates Z_2 by the regression equation

$$Z_2 = \frac{\mu_2+ \omega_2}{2} + \Sigma_{21}\Sigma_{11}^{-1}\left(Z_1 - \frac{\mu_1 + \omega_1}{2}\right)$$

and substitutes this value into the discriminant function of Z. That is, using $\frac{\mu_1 + \omega_1}{2}$ as the mean for Z_1 and $\frac{\mu_2+ \omega_2}{2}$ as the mean for Z_2 yields V_1 in the regression approach.

Two-Stage Rule

For the first stage of the classification rule use the marginal discriminant rule to classify Z_1 into population Π_1 or Π_2. Now we have the vector Z in either Π_1 or Π_2 and we shall utilize the Smith-Hocking estimation procedure to estimate the mean vector and covariance matrices. These new estimates shall be denoted

$_2\hat{\mu}$, $_2\hat{\Sigma}$ and $_2\hat{\omega}$. If Z was classified into Π_1 then Z_2 is estimated by $_2\hat{\mu}_2$ where

$$_2\hat{\mu} = \begin{pmatrix} _2\hat{\mu}_1 \\ _2\hat{\mu}_2 \end{pmatrix}$$

and in a similar manner we estimate Z_2 by $_2\hat{\omega}_2$ if Z was classified into Π_2 by V_1. The preliminary simulation studies indicate this procedure performs better than the rule utilizing only V_1.

Application

A student at Tarleton State University whose curriculum requires college algebra may take the College Algebra course for credit if the student is prepared or may take the Fundamentals of College Algebra for credit and then follow with the College Algebra for credit. It is most beneficial for the student who is prepared for the College Algebra course to be counselled into that course, rather than to lose interest in both mathematics and a semester in college by taking the remedial Fundamentals of College Algebra course. Even more serious, is the incorrect placement of a student in College Algebra who is deficient in his mathematical training and should be placed in the remedial course in algebra.

In an attempt to aid in the proper placement of the students, Bohannon [1976] utilized discriminant analysis to place the students. The variables used in the analysis are defined in Table 1.

Table I

Variables Utilized for Classification

Variable	Description of Variables
X_1	Student's High School Algebra I Score
X_2	Student's SAT Math Score
X_3	Student's SAT Verbal Score
X_4	Student's High School Geometry Score
X_5	Student's High School Algebra II Score

The samples utilized to construct a discriminant function were the students who enrolled at Tarleton State University during the Fall Semester of 1973 without having taken the Fundamentals of College Algebra course. Population one is defined to be the set of students who receive or will receive a grade of C or better in the College Algebra course, and the students who receive below a C, or drop the course, constitutes population two.

The study found estimates for the error of misclassification and for the purpose of cross-validation of the discriminant function, similar samples were drawn from the students who enrolled in the College Algebra in the Fall of 1974. The samples contained partial data records for some of the students and hence the two-stage discriminant rule was applied to those vectors and the results are shown in Table 2.

Table 2

Prediction Results

Marginal Rule

	Predicted	Group
Group	1	2
1	12	9
2	7	19

Two-Stage Rule

	Predicted	Group
Group	1	2
1	13	8
2	5	21

Thus we observe that the two-stage rule yields slightly better results than the marginal rule.

REFERENCES

1. Bohannon, Tom [1976]. "Discriminant Analysis with Missing Data." Ph.D. dissertation, Texas A.& M. University, College Station, Texas.

2. Cacoullos, T. [1972]. Discriminant Analysis and Applications. Academic Press, New York.

3. Chan, L. S. and Dunn, L. J. [1972]. "The treatment of missing values in discriminant analysis-I. The sampling experiment." J.A.S.A. 67, 473-477.

4. Chan, L. S. and Dunn, L. J. [1974]. "A note on the asymptotic aspect of the treatment of missing values in discriminant analysis." J.A.S.A. 69, 672-673.

5. Chan, Gilman and Dunn [1976]. "Alternative approaches to missing values in discriminant analysis." J.A.S.A. 71, 842-844.

6. Hocking, R. R. and Smith, W. B. [1968]. "Estimation of parameters in the multivariate normal distribution with missing observations." J.A.S.A. 63, 159-173.

7. Jackson, E. L. [1968]. "Missing values in linear multiple discriminant analysis." Biometrics 24. 835-844.

8. Lachenbruch, P. A. [1975]. Discriminant Analysis. Hafner Press, New York.

9. Smith, W. B. and Zeis, L. D. [1973]. "On classification for incomplete multinormal data." Communications in Statistics 2, 85-92.

10. Srvastava, J. N. and Zaatar, M. K. [1972]. "On the maximum likelihood classification rule for incomplete multivariate samples and its admissibility." Journal of Multivariate Analysis 22, 115-126.