CLASSIFICATION BASED ON INCOMPLETE DATA RECORDS

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

Introduction

In this paper we discuss the situation where there are two distinct multivariate normal populations, Π_1 and Π_2, with common variance-covariance matrix. We observe a p-vector X and must assign X to Π_1 or Π_2 based on the components of X and our classification rule. If the parameters of the distributions of populations Π_1 and Π_2 are known, this information is utilized in the construction of a classification rule. If the parameters are not known, which is the usual situation, then random samples from Π_1 and Π_2 are used to estimate these parameters and to construct a classification rule. We shall use as our classification rule, Anderson's discriminant function.

One of the problems that arise in the practical applications of discriminant analysis is analytically measuring the goodness of the classification rule. This rule must be evaluated based on some criterion of goodness of classification. For our criterion in this study we shall use the total probability of misclassification. In general, since the parameters of the populations are usually unknown, the probability of misclassification must also be estimated from random samples.

The samples on which one bases the classification rule and estimates the probability of misclassification often contain incomplete observation vectors, that is, vectors in which one or more components are missing. In many such situations these incomplete vectors are not included in the construction of a classification rule or in the estimation of the probability of misclassification. The primary purpose of this paper is to investigate a method for incorporating these incomplete observation vectors in the construction of the classification rule and the estimation of the probability of miscalssification. This method and the commonly practiced method of ignoring these incomplete vectors will be compared by computer simulation.

The use of discriminant analysis techniques on incomplete data sets is an area where very little research has been done. Jackson (1968) investigated a classification problem which had missing values in a large data set. The missing values were estimated using means and regression techniques and for the problem under study, the estimation procedure using missing data gave better results than the procedure of ignoring the observations with missing values.

Chan and Dunn (1972) investigated the problem of constructing a discriminant function based on samples, which contained incomplete observation vectors. Several methods of estimating the missing components of these vectors were utilized and the resulting vectors were used to construct the discriminant function. They concluded that no method was best for every situation, and gave guidelines to use in choosing the best method for various situations.

Hocking-Smith Estimation Procedure

A generalization of the estimation procedure reported by Hocking-Smith (1968) will be applied to random samples from multivariate normal populations, which contain incomplete vectors. This procedure requires that optimal estimators of the mean vectors and dispersion matrices are available for each group of observations, the groups being collections of observation vectors with identical patterns of incompleteness. The procedure has been shown to be essentially equivalent to solving the maximum likelihood equations for the incomplete situation. The estimators have been shown to be consistent and asymptotically efficient.

Note that we are estimating the mean vectors and the variance-covariance matrices without estimating the missing components of the incomplete vectors. Those missing components could, however, be estimated by using the previously mentioned estimators and regression techniques.

To illustrate the form of the estimators, consider a set of observations which follow a p-variate normal distribution with unknown mean vector μ_1 and variance-covariance matrix Σ_1. Let there be n_1 independent complete observation vectors and n_2 independent incomplete observation vectors which follow the q-variate marginal distribution. We define the elementary matrix D such that $\mu_2=D\mu_1$ is the mean of the marginal and $D\Sigma_1D'=\Sigma_2$ the variance-covariance matrix for the marginal distribution.

The joint likelihood L for these two groups of observations is given by $L = L_1 \cdot L_2$, where L_1 and L_2 are the likelihood functions associated with the two groups of observation vectors. The Hocking-Smith estimates are given by

$$_2\hat{\mu}_1 = \hat{\mu}_1 - \frac{n_2}{N}\hat{R}'(D\hat{\mu}_1-\hat{\mu}_2)$$

$$_2\hat{\Sigma}_1 = \hat{\Sigma}_1 - \frac{n_2}{N}\hat{R}'(D\hat{\Sigma}_1D'-\hat{\Sigma}_2)\hat{R}$$

where $\hat{R}'=\hat{\Sigma}_1D'(D\hat{\Sigma}_1D')^{-1}$ and $N = n_1+n_2$. These estimates are in general maximum likelihood considering the combined likelihood function L if $\hat{\Sigma}_2$ is replaced by $\hat{\Sigma}_2 + \hat{H}_2$,

where $\hat{H}_2=n_2(\hat{\mu}_2-D_2\hat{\mu}_1)(\hat{\mu}_2-D_2\hat{\mu}_1)'$.

The extension to more than two groups follows sequentially and is easily adaptable to computer programming. For further information the interested reader is referred to Hocking et al. (1969).

Estimators of the Probability of Misclassification

The problem of estimating the probability of misclassification has received a considerable amount of attention in the statistical literature. A fairly complete review of the literature on this problem is given by Toussaint (1974). The estimators that are considered in this study are as follows:

1. The estimator $\phi(-D/2)$, where $D^2=(\hat{\mu}_1-\hat{\mu}_2)'\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$.
2. The estimator $\phi(-D'/2)$, where $D'^2=(n_1 + n_2 - p - 3)(n_1 + n_2 - 2)^{-1}D^2$.
3. The McLachlan estimator, which is defined in

McLachlan (1975).

For the discriminant functions based on only the complete observations these estimators will be denoted by P_1, P_2 and P_m, respectively. For the discrominant functions based on all of the observation vectors these estimates will be denoted by $\hat{\alpha}_1$, $\hat{\alpha}_2$ and α_m, respectively. The estimators $\hat{p}_1$ and $\hat{p}_2$ were studied by Lachenbruch and Mickey (1968) and Sorum (1972).

Simulation Procedure and Results

Random samples are generated from each of two populations and a specified percentage of vectors are randomly chosen and made incomplete. The groups for the Hocking-Smith estimator procedure are formed and the estimates of the parameters for each group calculated. The estimates from the group of complete vectors are utilized in the calculation of the discriminant function and the previously mentioned estimators for the probability of misclassification are calculated. Next, the Hocking-Smith estimates are used in calculating the discriminant function and the estimators for probability of misclassification are calculated. This procedure is repeated at least ten times for each specified set of simulation variables. The mean and standard deviation are calculated for each estimator of the probability of misclassification.

The simulation results obtained by varying the values of the correlation coefficient ρ from 0 to .9 by increments of .1 indicated that the relative performance of the estimators were not effected by this simulations variable. The same was found to be true for varying the form of the mean vector. Hence the simulation results presented in the tables and in the figures are for the pooled simulations of ρ and the form of the mean vector. The tables and figures presented in this paper are only for the number of variables equal to three and the Mahalanobis distance equal to four. The percentage of missing values were chosen to be 20, 40 and 80 with three groups of vectors for the Hocking-Smith estimation procedure. These simulation results are part of the simulations conducted by Bohannon (1976) and are representative of those results.

Frequencies and cumulative proportions of

$$e = |\alpha - \hat{\alpha}|$$

where α is the optimum probability of misclassification and $\hat{\alpha}$ as the estimator of this probability were calculated for the simulation combinations with end points .0125, .025, .0375, .05, .0625, .075, and the last interval greater than .075. Figures 1 to 3 present these results and Table 1 gives the means and standard deviations for these estimators.

In analyzing the previously mentioned tables and figures, there are several observations that are apparent. One being, that as the percentage of incomplete data increases, the variances of the estimators increase. However, the increase for the estimators based on the Hocking-Smith estimates is not as great as that for the estimators based only on complete vectors. Our simulations also indicate that the McLachlan estimator has a larger variance in general than the other estimators for our range of population parameters. The simulations indicate that $\hat{\alpha}_2$ is the best estimator of α based on the criterion of unbiasedness and minimum variance and in general the incomplete vectors do provide useful information for classifying the observation vectors.

References

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

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

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

Hocking, R. R., Smith, W. B., Waldron, B. R. and Oxspring, H. H. [1969]. "Estimation of parameters with incomplete data." Themis Report No. 12, Institute of Statistics, Texas A&M University, College Station, Texas.

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

Lachenbruch, P. A. and Mickey, M. R. [1968]. "Estimation of error rates in discriminant analysis." Technometrics 10, 1-11.

McLachlan, G. J. [1975]. "Confidence intervals for the conditional probability of misallocation in discriminant analysis." Biometrics 31, 161-167.

Rulon, Phillip J. [1951]. "Distinctions between discriminant and regression analysis and a geometric interpretation of the discriminant function." Harvard Educational Review 21, 80-90.

Sorum, M. [1972]. "Estimating the expected and the optimal probabilities of misclassification." Technometrics 14, 935-943.

Toussaint, G. T. [1974]. "Bibliography on estimation of misclassification." IEEE Transactions on Information Theory IT-20, 472-479.

TABLE 1

TABULATION FOR ESTIMATORS WITH $p = 3$, $\Delta = 2$, AND $\alpha = .1587$

			P_1	P_2	P_m	α_1	α_2	α_m
M	20%	mean	.1516	.1589	.1653	.1550	.1625	.1660
ND	126	std. dev.	.0337	.0337	.0357	.0337	.0318	.0332
M	40%	mean	.1462	.1591	.1642	.1513	.1588	.1621
ND	125	std. dev.	.0393	.0393	.0424	.0357	.0357	.0374
M	80%	mean	.1345	.1804	.1908	.1518	.1591	.1626
ND	115	std. dev.	.0580	.0588	.0743	.0409	.0411	.0480

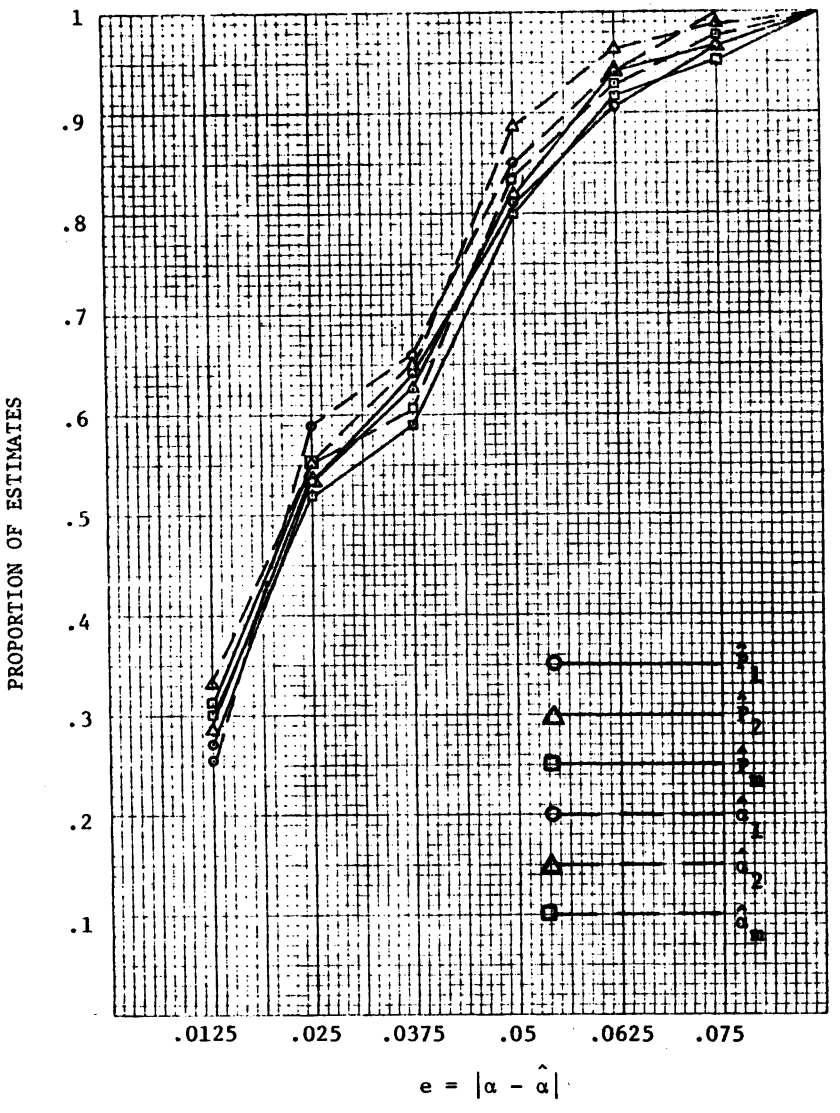

FIGURE 1

126 LINEAR DISCRIMINANT FUNCTIONS WITH p = 3, Δ = 2, AND m = 20%

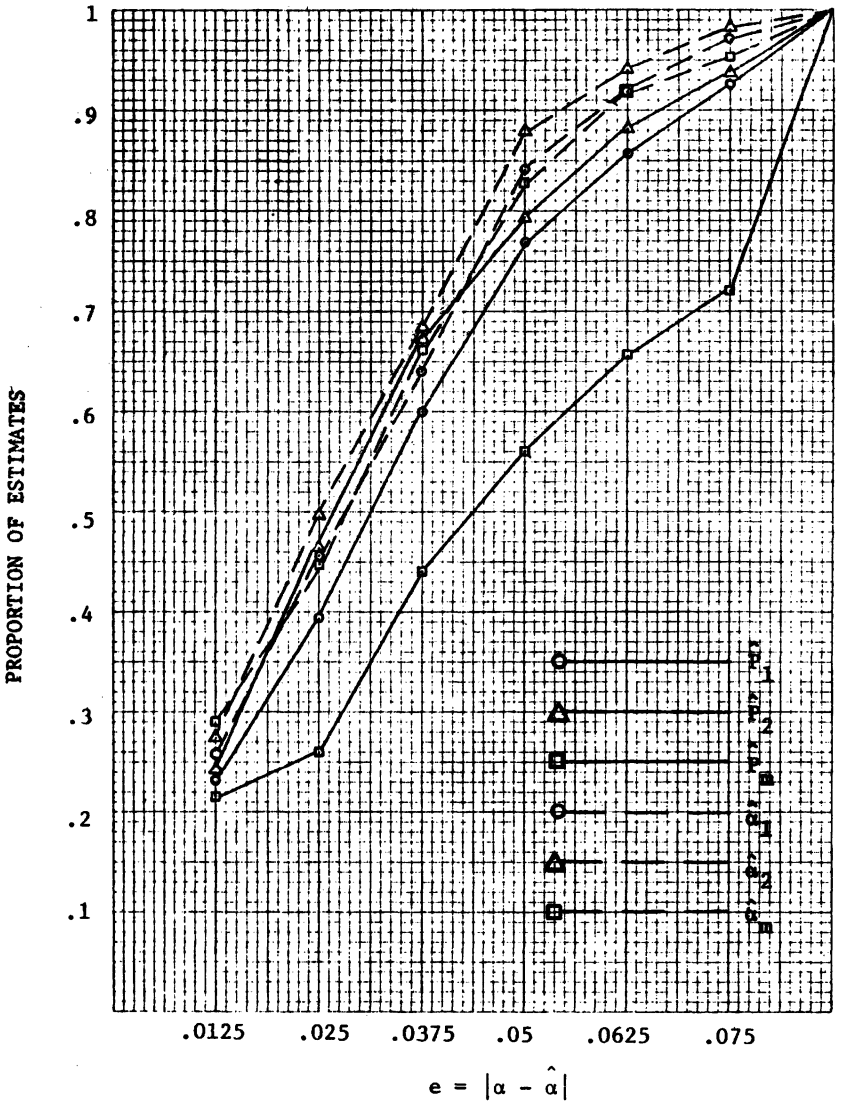

FIGURE 2

125 LINEAR DISCRIMINANT FUNCTIONS WITH p = 3, Δ = 2, AND m = 40%

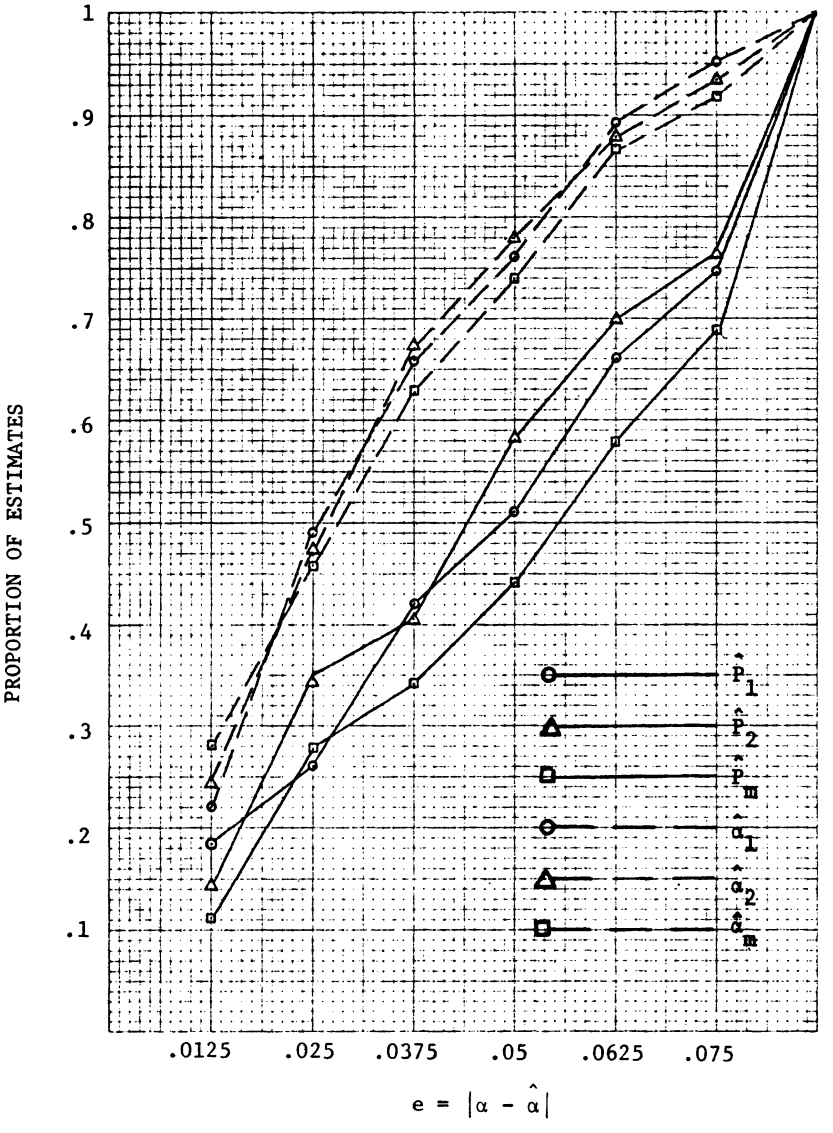

FIGURE 3

115 LINEAR DISCRIMINANT FUNCTIONS WITH p = 3, Δ = 2, AND m = 80%