

A VARIABLE SELECTION ALGORITHM FOR QUADRATIC DISCRIMINANT ANALYSIS

Samuel L. Seaman, Baylor University, P.O. Box 98005, Waco, Texas 76798
Dean M. Young, Baylor University, P.O. Box 98005, Waco, Texas 76798

INTRODUCTION

One difficulty in the application of statistical discriminant analysis obtains when the researcher must select the most important variables to be included in a statistical analysis. Typically, the relevant variables have been identified using some objective criterion from the substantive discipline in which the application is made. However, the resulting set of variables can contain redundancies, making a further reduction in variables desirable. Indeed, there are other reasons that a researcher might wish to perform variable selection for a statistical discrimination algorithm.

Firstly, variable selection can effectively reduce the conditional error rate by increasing the ratio of the training sample size to the dimension (see Van Ness (1979)). Secondly, the cost of gathering and processing a large number of variables may be prohibitive, and variable selection may be necessary to minimize costs. Thirdly, the goal of the analysis may be to identify only those variables most relevant for use in allocating future observations. Finally, variable selection may be used to obtain parsimony of the data set.

The problem of variable selection in discriminant analysis has received much attention in the literature. Reviews by McKay and Campbell (1982) and Huberty (1984) provide excellent summaries of this work. For the allocatory variable selection problem, McKay and Campbell conclude that "The basic problem (of variable selection for allocation purposes) is to determine the subset (or subsets) for which the (conditional) error rate is a minimum. Statistically it may only be possible to determine a subset or subsets for which the error rate is essentially no greater than that for the original set."

Numerous methods for selecting variables for a linear discriminant analysis have been suggested in the literature, but little has been written on the topic of variable selection for the quadratic discriminant function. Fatti and Hawkins (1986) derive a variable selection algorithm for quadratic discrimination using a likelihood ratio test statistic. Also, Young and Odell (1986) have formulated a variable selection method based on a linear sufficiency criterion. Pynnönen (1987) has proposed a variable selection algorithm for the heteroscedastic discrimination model based on the Akaike information criterion.

In this paper we are interested in variable selection for allocatory discrimination and, in particular, the development of a computationally efficient classificatory variable selection algorithm for use with the quadratic discriminant model.

THE VARIABLE SELECTION ALGORITHM

The new variable selection algorithm for quadratic discrimination which we propose has three main ingredients: a method for calculating the inverse of a sample covariance matrix for a subset of the original variables using the inverse of the original p -dimensional sample covariance matrix, an estimator of the conditional error rate, and a stopping rule.

A mathematical relationship described by Narendra and Fukunaga (1977) and Kittler (1978) allows one to apply the computations for the inverse of the covariance matrix of a full set of variables to obtain the inverse(s) of the covariance matrices for any subset(s) of the original variables, making for more efficient use of computer time. This relationship is described as follows. Let A be a symmetric $p \times p$ nonsingular matrix. Furthermore, let A_1 be

the symmetric nonsingular matrix obtained by deleting the i th row and i th column from A . The inverse of A_i is

$$A_i^{-1} = B - rr'/v_{ii}$$

where B is the $(p-1) \times (p-1)$ matrix formed by deleting the i th row and i th column of A^{-1} and the p -dimensional vector r is the i th column of A with the i th element, v_{ii} , omitted.

The proposed variable selection algorithm is designed to select a subset of predicate variables which has minimal estimated conditional error rate. The estimated conditional error rate used in the proposed algorithm, is obtained with the leave-one-out (LOO) method suggested by Lachenbruch (1967). The method is characterized by a two-step process. In the first step, an observation is deleted and the linear classification functions are computed for the remaining $(N-1)$ observations. In the second step, the QDF computed in step one is used to allocate the deleted observation into one of the labeled groups. This process is repeated n times and the proportions of observations misclassified may be employed to determine the estimated error rate.

A third characteristic of our variable selection algorithm is the stopping rule. Since we adopt a backward elimination strategy, we shall need a rule to decide when to discontinue the variable selection process. Our stopping rule is similar to one proposed by Habbema and Hermans (1977), and may be described as follows:

Stop the variable selection algorithm if

1. The LOO conditional PMC estimated for the selected subset exceeds the LOO conditional PMC estimated for the original set of p variables by more than some tolerance value, say $\epsilon \geq 0$.
2. The LOO conditional PMC estimated for a subset at the j th iteration exceeds the LOO conditional PMC estimated for the most recently selected subset. Stop at subset S_q if $P(S_p) - P(S_{j-1}) \geq \epsilon$, $j = p, p-1, \dots, q+1$, while $P(S_{q-1}) - P(S_q) > 0$.

Note that estimates of the conditional error rates could be calculated for all

possible subsets of the original variables, but this would be very costly, especially when the dimension (p) is large. Instead, we adopt a backward elimination strategy and search for the subset of variables having an estimated PMC which cannot exceed the estimated PMC of the full set of variables by more than a specified threshold value, ϵ .

We shall now describe the essential steps which constitute the new variable selection algorithm for quadratic discrimination.

- 1.) Compute the leave-one-out estimated probability of misclassification (PMC) for the original set of variables, and store the inverse of the pooled covariance matrix, S .
- 2.) Delete one of the p variables and compute the leave-one-out PMC for the remaining subset of $(p-1)$ variables. Repeat until having swept through all p variables. The variable associated with the maximum PMC is deleted and a comparison of PMC's according to the suggested stopping rule is conducted.
- 3.) Repeat steps one and two until a stopping rule has been satisfied.

We note that the variable selection algorithm described above will not necessarily yield the subset which has the minimal conditional overall error rate. However, the algorithm is designed to yield a feature subset that has a conditional overall error rate that is relatively close to the minimal conditional overall error rate.

AN EXAMPLE

Sample data for our comparison derive from a bankruptcy prediction study. In this example, balance sheet and income statement data have been collected for a sample of 39 publicly listed firms entering bankruptcy and 39 nonbankrupt firms for the period 1977 to 1983. Bankrupt firms have been identified using The Wall Street Journal Index and a list of bankrupt firms compiled by the Securities and Exchange

Commission (SEC). Financial statement data for each bankrupt firm have been obtained from the last annual (10-K) financial report filed with the SEC prior to bankruptcy. Nonbankrupt firms have been selected based on the same criteria, which include fiscal year-end, industry, and asset size. Financial statement data for the nonbankrupt firms also have been obtained from the 10-K reports. A set of twenty two financial ratios, constructed from this financial statement data, make up the full set of independent variables used to classify firms into the solvent or insolvent group. We note here that both the test for equal covariance matrices and the test for multivariate normality have been rejected at the .0001 level. The proposed comparison may appear to be inappropriate here, since the Stepdisc procedure assumes equal covariance matrices for the two competing populations. However, the SAS stepwise procedure is often misapplied in the heteroscedastic case simply because it is readily available.

In order to evaluate the efficacy of the proposed variable selection algorithm for quadratic discrimination we shall compare the performance of the SAS Stepdisc variable selection procedure with the performance of the newly proposed variable selection algorithm. Our algorithm selected a subset of 5 predictor variables, which had an estimated conditional error rate of .1473. The SAS stepdisc variable selection procedure yielded a subset of 3 variables, with an estimated conditional error rate of .2876.

COMMENTS

A reasonable criterion for the selection of variables in allocatory discriminant analysis is the minimization of the estimated conditional overall error rate. The algorithm we propose employs this criterion and, yet, is computationally efficient. The SAS Stepdisc procedure, uses the F statistic criterion for variable selection. Eisenbeis, Gilbert, and Avery (1973) have shown the F statistic criterion to be inferior to most others in allocatory

discrimination. This may be due, in part, to the sensitivity of the F-statistic to nonnormality. McLachlan (1986) notes, for example, that "caution should be exercised with the use of parametric estimators of the error rates as they may not be reliable under departures from the parametric model adopted, even though the sample rule itself may be robust." This caution certainly applies to the data analyzed in this study as there is strong evidence of nonnormality. In contrast, the proposed algorithm adopts a robust criterion and appears to be less sensitive to the departures from normality realized in this example. It is for this reason, perhaps, that the proposed algorithm provides greater classification accuracy and can therefore be recommended over the SAS Stepdisc procedure. The proposed algorithm is also preferable to the SAS procedure in that it yields a robust estimate of the overall conditional error rate for the selected subset. This estimate of the conditional error rate should provide a useful index of the accuracy in classification that one might expect when using the selected set of variables. This added information is a highly desirable outcome when the purpose of the discriminant analysis is allocation.

REFERENCES

- Anderson, T.W. (1951). Classification by multivariate analysis. *Psychometrika* 16, 31-50.
- Anderson, T.W. (1984). *Introduction to Multivariate Statistical Analysis*, John Wiley and Sons: New York.
- Fatti, L.P. and Hawkins, D.M. (1986). Variable Selection in Heteroscedastic discriminant analysis. *Journal of the American Statistical Association* 81, 494-500.
- Habbema, J.D.F. & Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics* 19, 487-493.
- Huberty, C.J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin* 95, 156-171.

Huberty, C.J. & Smith, D.U. (1976). Variable contribution in discriminant analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kittler, J. (1978). Classification of incomplete pattern vectors using modified discriminant functions. *IEEE Transactions on Computers* C-27, 367-374.

Lachenbruch, P.A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics* 23, 639- 645.

Lachenbruch, P.A. (1975). *Discriminant Analysis*. Hafner Press: New York.

McKay, R.J. and Campbell, N.A. (1982a). Variable selection techniques in discriminant analysis I. Description. *British Journal of Mathematical and Statistical Psychology* 35, 1-29.

McKay, R.J. and Campbell, N.A. (1982b). Variable selection techniques in discriminant analysis II. Allocation. *British Journal of Mathematical and Statistical Psychology*, 35, 30-41.

Narendra and Fukunaga (1977). A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput.* C-26, 917-922.

Pynnonen, S. (1987). Selection of variables in nonlinear discriminant analysis by information criteria. *Proc. Second International Ampere Conference in Statistics*, 627-636.

Wyman, F., Young, D.M., and Turner, D. (1990). A comparison of asymptotic error rates for the sample linear discriminant function. To appear in *Pattern Recognition*

Young, D.M. and Odell, P.L. (1986). Feature-subset selection for statistical classification problems involving unequal covariance matrices. *Communications in Statistics--Theory and Methods*

.....