Stopping rule for variable selection using stepwise discriminant analysis

C. S. Munita,[1]* L. P. Barroso,[2] P. M. S. Oliveira[2]

[1] *Laboratório de Análise por Ativação Neutrônica Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, C.P. 11049, 05422-970 São Paulo, Brazil*

[2] *Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil*

(Received April 6, 2006)

In general, when characterizing samples, such as ceramic samples or other types of samples, for first time by means of chemical elements, the analyst measures a large number of variables, many of which may not be very informative. In fact, some may even be unrelated to the issue at hand and blur the picture instead of making it clearer. In subsequent studies the analyst may wish to measure fewer variables for several reasons, such as being very time consuming; in cases where measurement time is important, such as on-line monitoring; in order to reduce cost or effort; etc. Therefore, the hope is to determine those variables that are most relevant without losing essential information and to remove the less productive information. The problem is how to perform this in an objective way and to capture crucial information using a multivariate analysis. This paper aims to describe and illustrate a stopping rule for the identification of redundant variables, and the selection of variable subsets, preserving multivariate data structure using stepwise discriminant analysis, selecting those variables that are in some senses adequate for discrimination purposes. One illustrative example using data sets obtained via INAA of ceramic samples from two archaeological sites is provided.

Introduction

Very often chemists face analytical problems whose properties are unknown or not clearly defined, thus a goal in analytical chemistry could be to unravel information hidden in experimental data. Moreover, problems can be difficult to handle not only because of the multivariate nature of the data but also because investigators do not know what are the characteristic variables or if some of them are more important than others. Even when the main goal was clearly defined, there could be serious doubts about particular parameters to measure. Thus, various times, as many variables as possible are measured in order to capture crucial information. This is naturally very time consuming and expensive. Such multidimensional data sets must be closely examined to draw useful information. Moreover, investigators would have to decide if it is really necessary to measure all the variables on a particular set of samples to describe the problem. In such cases the problem is choosing a subset of the available variables, which hopefully will in some sense be almost as informative as the entire set of variables. One approach is to consider every subset of the variables, and to choose an optimal subset by some suitable criterion. For more than about 14 or 20 variables, the number of subsets involved becomes unmanageable.

The problem is how to perform this in an objective way and to capture the crucial information using a multivariate analysis. There are a number of statistical methods, which have been suggested and used as rules for selection of a subset of variables.[1–4] COSTANZA and AFIFI[5] studied seven methods of the selection of variables. Similar techniques were discussed by SCHAAFSMA and VARK[6] to decide how to use different stopping rules. Possibly the best known is the stepwise *F*-procedure.[7]

The selection of the most useful variables in discriminant analysis is an important contribution in analytical chemistry. The analyst may wish to measure fewer variables in subsequent studies, to reduce cost or effort of measurement or to try to reduce the complexity of the problem.

In this work, a stopping rule for the identification of redundant variables, and the selection of variable subsets, preserving multivariate data structure for stepwise discriminant analysis is presented, i.e., selecting those variables which are in some sense adequate for discrimination purposes, without losing essential information. Although many criteria are available, the stopping rule most commonly employed uses a sequence of standard *F*-tests to determine the significance of the additional distance contributed by each forward stepwise entry. Stopping occurs just before the first insignificant entry. The rule is based on the maximum estimated unconditional probability that often performs better than the strict use of all variables. The procedure was illustrated using a data set of 114 ceramics samples analyzed by INAA from two archaeological sites, named A, and B.

Variable selection procedure

Suppose that p variables have been measured on each of n samples, and that the essential dimensionality of the data to be used in any comparison is k. A criterion for assessing a particular variable x_{p+1} increases the separation provided by variables x_1, …, x_p which is obtained by means of an analysis of covariance, treating

* E-mail: camunita@ipen.br

x_{p+1} as the response, and $x_1, \ldots, x_p$ as covariants. Then x_{p+1} provides significant additional information at level α if the partial F-statistic:[5]

$$F_{(q)} = \frac{(n_1 + n_2 - p - 1)}{(p - q)} \cdot \frac{n_1 n_2 (\Delta^2_{(p)} - \Delta^2_{(q)})}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 \Delta^2_{(q)}} \qquad (1)$$

where $\Delta^2_{(q)} = (\bar{x}_{1(q)} - \bar{x}_{2(q)})' S^{-1}_{(qq)} (\bar{x}_{1(q)} - \bar{x}_{2(q)})$ (2)

and $q = 0, 1, \ldots, (p-1)$

exceeds the critical value:

$$F_{\text{crit}} \leq F_{1-\alpha}(p-q, n_1+n_2-p-1) \qquad (3)$$

A sequence of partial F-statistics results:

$$F_1, F_{2.1}, F_{3.12}, \ldots, F_{p+1.12\ldots p}, \ldots, F_{k,12\ldots(k-1)}$$

F_1 is the usual analysis of variance F-statistics for testing whether x_1 separates the population.

The procedure starts by using $F_{\text{k},12....(k-1)}$ to test whether x_k can be deleted; if it can, x_{k-1} is examined in the same way, and so on until a deletion is not justified by the corresponding test. The variables remaining after the sequence of deletions are then considered adequate. If all k tests are carried out without a significant result, the conclusion is that $x_1, \ldots, x_k$ provide no separation.

In other words, for each variable, the F statistic is computed. The variable corresponding to the largest of these statistics is the first selected, provided the statistic exceeds a specific value. Variables are then added one at a time based on an examination of partial F-statistics. Suppose that variables $x_1, \ldots, x_p$ have been selected. The partial F-statistic reflecting additional information supplied by each of the remaining variables independently of $x_1, \ldots, x_p$ is computed. The variable corresponding to the largest of these statistics is selected, provided that statistics exceed the specified critical point calculated by Eq. (3). The procedure terminates when none of the selected variables can be excluded, and no further variables can be included.

Results and discussion

The procedure was applied on a real database, for this the data set obtained by MUNITA et al.[8] was considered. These data comprise the determination of 13 elements (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U) in 114 ceramic fragments from two archaeological sites named A and B. The full data matrix and details of the analyses are given in MUNITA et al.[8]

Initially, the results were transformed to log base 10 to compensate for the large differences of magnitudes between the measured elements for the trace level and the larger ones.[9] The log base 10 transformation of data before a multivariate statistical method is common. One reason for this is the belief that, within the manufacturing of raw material elements there is a natural log-normal distribution, and this data normality is desirable. Another reason is that a logarithmic transformation tends to stabilize the variance of the variables and would thus give them approximately equal weight in a non-standardized multivariate statistical analysis. All individual determinations in each data set were tested for discordant results. The Mahalanobis distance, D_i, is suggested by many authors as a method for detecting outliers in multivariate data.[10,11] For each of the n observations (samples) in a p variable data set, a distance value D_i was calculated. Let $\bar{x}$ be the sample mean vector and let S be the sample covariance matrix,

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T / (n-1) \qquad (4)$$

and $$D_i = \sqrt{\{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})\}}$$

for $i = 1, \ldots, n$, where $(x_i - \bar{x})$ is the difference vector between the measured values in one group and the mean values of the other group. WILKS[12] suggested the use of:

$$p\,(n-1)^2\, F_{p,n-p-1} / n(n-p-1+pF_{p,n-p-1}) \qquad (5)$$

to calculate the critical values for D_i when searching for a single outlier. WILKS[12] used so-called scatter ratios to search for outliers in multivariate normal data. To search for a single outlier, the author calculates a scatter ratio R_i:

$$R_i = |A_i|/|A|$$

where $$A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

and $|A|$ = determinant (A), A_i is calculated as for A with observation i eliminated from the sample. The most outlying observation is that which has the smallest scatter ratio R_i, where $R_1 = \min\{R_i\}$, i.e., the observation whose removal leads to the greatest reduction in $|A|$. This procedure at confidence level of 95% was applied at data set and the D values were calculated for all samples. When the D calculated in the sample was higher than D critical value the sample was eliminated and calculated a new D. The procedure ended when the $D_\text{crit} > D_\text{cal}$. In all two samples were eliminated (one in each site).

The variable with the smallest partial F-statistic was eliminated with the purpose to study a subset of variables, of the partial F-statistics and to examine whether that variable supplied additional information independently of the remaining $k-1$ variables or not. All this is done provided that the statistic does not exceed a specified critical value. If the variable can be eliminated, the process is repeated on the remaining $k-1$ variables, and so on.

Table 1. Partial F-statistic for variable selection, for sites A and B

Variable	Step			
	1	2	3	4
As	0.020	0.021	0.022	0.022
Ce	0.00009*			
Cr	0.001	0.0026	0.0016*	
Eu	0.015	0.015	0.015	0.019
Fe	0.095	0.098	0.097	0.096
Hf	0.018	0.018	0.018	0.017
La	0.562	0.628	0.632	0.674
Na	0.001	0.0018*		
Nd	0.014	0.014	0.015	0.014
Sc	0.083	0.085	0.135	0.135
Sm	0.234	0.241	0.242	0.248
Th	0.168	0.173	0.182	0.189
U	0.018	0.019	0.019	0.019
Critical value:**	0.0039	0.0039	0.0039	0.0039

* Variable deleted at each step.
** Nominal 5% test.

Table 1 contains the partial F-statistic involved in the database. The smallest value in the first column, 0.00009, corresponds to the variable Ce. For the sake of discussion, this is compared with the critical value at 95% of confidence level (0.0039). Hence, Ce can be eliminated. The smallest partial F-statistic in the second column, 0.0018, corresponds to the variable Na. The comparison of this with the critical value at 95% of confidence level (0.0039) leads to the deletion of Na. Continuing in this way, Cr is eliminated at the third step. The procedure terminates at the fourth step when the smallest partial F-statistic, 0.014, exceeds the critical value at 95% of confidence level (0.0039). Thus the variables selected are As, Eu, Fe, Hf, La, Nd, Sc, Sm, Th, and U.

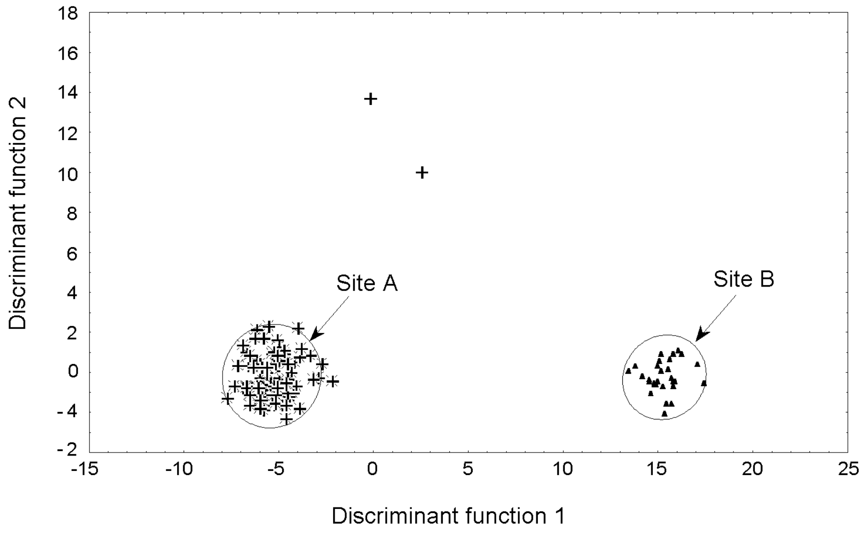

Fig. 1. Discriminant functions for all variables. Ellipses represent 95% confidence level

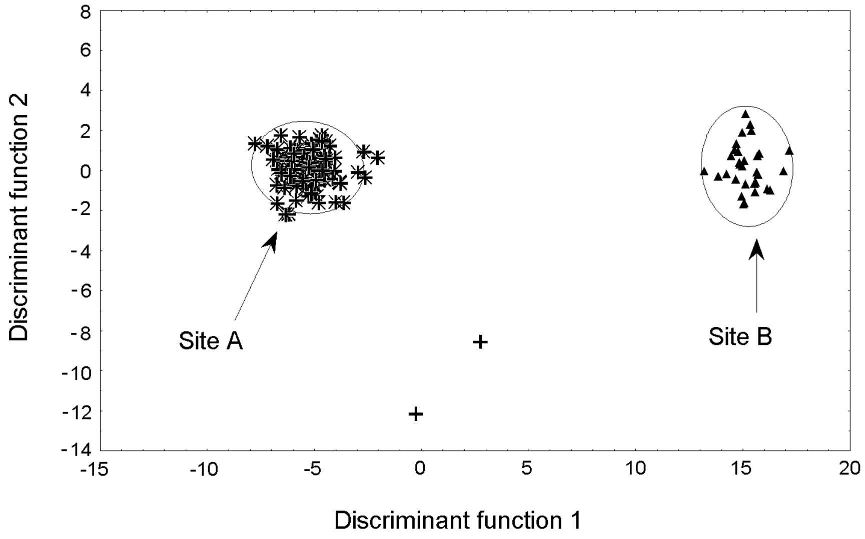

Fig. 2. Discriminant functions for selected variables. Ellipses represent 95% confidence level

To determine how well these subsets capture the structure of the complete data, Fig. 1 shows the plot for discriminant function 2 versus discriminant function 1 for all the variables (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th, and U), and Fig. 2 shows the plot of discriminant function 2 versus discriminant function 1 using the variables selected (As, Eu, Fe, Hf, La, Nd, Sc, Sm, Th and U). The comparison of Figs 1 and 2 confirms that discriminant analysis based on ten variables produce similar results to a discriminant analysis using all variables.

Conclusions

The procedure presented in this paper provides a useful descriptive tool in discriminant analysis to eliminate variables maintaining low probabilities of misclassification. The performance of the method was validated using concentration data of the ceramic sample. The plots of the discriminant functions using all the variables or with the selected variables are similar. The procedure aims to isolate subsets of variables, which separate the groups involved essentially in the same extent as the set of all available variables. The method can be useful in archaeometric studies to consider only the variables that might contribute to discrimination.

References

1. G. DIJKSTERHUIS, M. B. FROS, D. V. BYRNE, Food Qual. Prefer., 13 (2002) 89.
2. Q. GUO, W. WU, D. L. MASSART, C. BOUCON, S. DE JONG, Chemom. Intell. Lab. Syst., 61 (2002) 123.
3. N. T. TRENDSFILOV, L. A. LIPPERT, Linear Algebra Appl., 349 (2002) 245.
4. R. J. MCKAY, N. A. CAMPBELL, Brist. J. Math. Stat. Psych., 35 (1982) 1.
5. M. C. COSTANZA, A. A. AFIFI, J. Amer. Stat. Assoc., 74 (1979) 777.
6. W. SCHAAFSMA, G. N. VAN VARK, Stat. Nederlandica, 33 (1979) 91.
7. J. ROY, Ann. Math. Stat., 29 (1958) 177.
8. C. S. MUNITA, R. P. PAIVA, M. A. ALVES, E. F. MOMOSE, P. M. S. OLIVEIRA, J. Trace Microprobe Techn., 21 (2003) 697.
9. G. HARBOTLE, in: Radiochemistry: A Specialist Periodical Report, Vol. 3, G. W. A. NEWTON (Ed.), The Chemical Society, London, 1976, p. 33.
10. K. I. PENNY, Appl. Stat., 45 (1996) 73.
11. C. S. MUNITA, M. A. SILVA, F. A. SILVA, P. M. S. OLIVEIRA, Instrum. Sci. Technol., 33 (2005) 161.
12. S. S. WILKS, Sankhya, 25 (1963) 407.