ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
20,2 (1990), pp. 219–226

E. KRUSIŃSKA and J. LIEBHART (Wrocław)

CLEAR GRAPHICAL REPRESENTATION OF THE LOCATION MODEL IN MIXED VARIABLE DISCRIMINANT ANALYSIS

1. Introduction. Practical problems with mixed continuous and discrete predictor variables quite often appear in discriminant analysis. In spite of the wide range of applications till now only a few procedures have been elaborated to handle such mixtures.

Lachenbruch [13] in his book on discriminant analysis mentioned the logistic approach and the foundations of nonparametric methods as applied to discrimination. 9 years later Seber [14] in the monograph on multivariate statistics gave a comprehensive review of discriminant analysis methods. Besides logistic discriminant functions and more extended studies on nonparametric procedures he summarized the results in the location model approach. The idea of the latter was given by Krzanowski [6] in 1975, and then extended in his next papers ([9], [12]).

The method consists in creating different linear classification rules on the basis of continuous variables for each cell of the contingency table defined by discrete variable values. It is simple enough and computationally feasible in comparison with logistic discrimination and nonparametric density estimation which need a lot of computer time.

From the medical point of view, discriminant analysis methods can be a tool of assistance in medical diagnosis for differentiating between considered diseases and a control group.

As the predictor variables to perform discrimination, results of many examinations are used. Except the laboratory findings, which are mostly of continuous character, anamnesis, physical examination and often other additional examinations, such as for instance ecg, rtg, are coded as discrete. For this reason the location model approach seems to be a suitable strategy to perform assistance of medical diagnosis. However, some difficulties occur for a great number of discrete variables because in such a case the method is untractable, i.e., the parameters of the model are unestimable. So the selection of the most discriminative variables is necessary. Two procedures – the first

one of Krzanowski [10] which enables to reduce the number of discrete variables and the second one of Daudin [3] for the simultaneous choice of discrete and continuous variables – have been elaborated to solve the problem of selection.

The new procedure (Krusińska [5]) based on the multivariate discriminatory measure T^2 (Ahrens and Läuter [1]) was inspiration for further studies. They resulted in obtaining the canonical representation of the location model, via a suitable linear transformation of the data, other than that given by Krzanowski [8] and called *between-cell analyses*. The analysis presented in the paper is a *between-group analysis* and has a particular reference to the discrimination problem, because it gives the best separation between considered groups. It is a simple generalization of the canonical analysis for continuous data. The presented method enables us to represent cells, groups or individuals on the plane in the space of the two first most significant canonical variates.

From the medical point of view it should be added that graphical assistance of diagnosis has for the physicians one considerable advantage. It is much more clear and intuitively acceptable than numerical results of the complicated multivariate statistical procedures.

2. Preliminaries. Following Krzanowski [6] let us introduce the location model technique. Suppose that each individual is described by a vector

$$y' = (y_1, y_2, \ldots, y_p)$$

of p continuous variables and a vector

$$x' = (x_1, x_2, \ldots, x_q)$$

of q binary variables. Discrete variables with the number of states greater than two are recoded to binary ones [9].

The problem is in classifying an individual $w' = (x', y')$ to one of two (or generally more) populations Π_1 and Π_2 on the basis of the observed values of x and y. It is assumed that the continuous variables follow different multivariate normal distributions for each possible combination of values of binary variables. This means that

$$y \sim N(\mu_i^{(m)}, \Sigma) \quad (i = 1, 2;\ m = 1, 2, \ldots, 2^q).$$

The covariance matrix is assumed to be equal for all 2^q "locations" (cells of the contingency table defined by binary variables).

The optimal classification rule in that model is to allocate w falling into the m-th cell to Π_1 if

$$(\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1} \{y - \tfrac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})\} \geqslant \log(p_{2m}/p_{1m})$$

and otherwise to Π_2. Thus, it is equivalent to the classical linear discrimination but performed separately for each cell of the contingency table.

The parameters of the model (unknown in practice) are estimated from the following data: the a priori probabilities p_{im} ($i = 1, 2$; $m = 1, 2, \ldots, 2^q$) by the iterative scaling procedure of Haberman [4] which allows, for empty cells in the contingency table, the parameters related to continuous variables, i.e., $\mu_i^{(m)}$, Σ, by the use of the linear additive model imposed on the mean vectors. This enables to obtain smoothed estimates of $\mu_i^{(m)}$ and Σ (for more details see [6]). The classification of an individual $w' = (x', y')$ is performed by the leaving-one-out method, i.e., the unknown parameters are calculated after throwing away the actually classified individual and then the classification is performed on their basis. As has been indicated by Krzanowski [6], the location model gives better results of discrimination than the linear discriminant function with both continuous and binary features as predictor variables when there are interactions between binary variables and populations.

A generalization of the described technique to mixtures of both continuous and discrete variables and to the multiple discrimination problem is possible (see [9] and [12]).

3. Model selection and canonical representation. As indicated by Lachenbruch [13], testing that the considered set of variables has a sufficient discriminatory power is equivalent to performing a test on between-group differences. Thus the problem of discriminant analysis can be reformulated in the terminology of analysis of variance.

First let us consider a multivariate general linear model in the form

$$\underset{n\times p}{Y'} = \underset{n\times k}{X} \cdot \underset{k\times p}{B} + \underset{n\times p}{E}, \quad n > k, \tag{1}$$

where Y is the observation matrix.

In model (1) we test the hypothesis about the parameters B:

$$H_0\colon \underset{s\times k}{K}\ \underset{k\times p}{B} = \underset{s\times p}{O}, \quad 1 \leqslant s \leqslant k,\ \mathrm{rk}(K) = s. \tag{2}$$

Let $M = KB$, $X = [X_1 X_2]$, $\mathrm{rk}(X_1) = r > 0$, and let X_1 be nonsingular.

Let $\hat{M} = X_1(X_1'X_1)^{-1}X_1'Y$ be the least square estimate of M. Now we may define the matrix of residual sums of squares and products as

$$G = (Y - \hat{M})(Y - \hat{M})'$$

and the matrix of sums of squares and products due to H_0 as

$$H = \hat{\Delta}'[K_1(X_1'X_1)K_1']^{-1}\hat{\Delta},$$

where

$$\underset{s\times p}{\hat{\Delta}} = K_1(X_1'X_1)^{-1}X_1'Y.$$

The partition $K = [K_1, K_2]$ corresponds to the partition of X.

The hypothesis H_0 is tested by test statistics which are functions of the matrices H and G (see [1]). One of them is the Lawley–Hotelling trace statistic

$$T^2 = \mathrm{tr}(HG^{-1}). \tag{3}$$

Its distribution is approximated by Snedecor F. This enables us to use the distributional approach when T^2 is applied as a criterion for a subset choice. Alternatively, the generalized discriminatory measure $T_1^2 = rT^2$ (where r is defined by (2)) may be used [1].

Further, the new discriminant variables, the so-called canonical variates, can be defined as a solution of the eigenvalue problem:

$$He = \lambda Ge. \tag{4}$$

If $\lambda_1 > \lambda_2 > \ldots > \lambda_t$ are the eigenvalues of the matrix HG^{-1} and $e_1, e_2, \ldots, e_t$ are the corresponding eigenvectors, the canonical variables w_h $(h = 1, 2, \ldots, t)$ are given as

$$w_h = e_h' y = \sum_{i=1}^{p} e_{ih} y_i,$$

where $y = (y_1, y_2, \ldots, y_p)$ is the observational vector of p continuous variables.

Ahrens and Läuter in their book [1] have used the multivariate discriminatory measure T^2 (formula (3)) in the simple problem of differentiating between several groups described only by continuous variables. Such a problem is equivalent to one-way-classification multivariate analysis of variance (MANOVA).

Now let us consider the location model case. In the terminology of analysis of variance we have here a nested model of two-way-classification MANOVA with the first factor, A, corresponding to cells and the second one, B, corresponding to groups. The factor A has at most $l = 2^q$ levels (some cells may be empty in all groups), and the factor B has g_i levels for $i = 1, 2, \ldots, l$.

Now we test the hypothesis of the influence of the factor B (groups), so we test between group differences, i.e.,

$$H_B: \begin{cases} \beta_{1(1)} = \beta_{2(1)} = \ldots = \beta_{g1(1)}, \\ \beta_{1(2)} = \beta_{2(2)} = \ldots = \beta_{g2(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \beta_{1(l)} = \beta_{2(l)} = \ldots = \beta_{gl(l)}, \end{cases} \tag{5}$$

where $\beta_{j(i)}$ corresponds to the influence of the j-th level of the factor B (observed for the i-th level of A) on the values of the variables y.

The hypothesis H_B (formula (5)) is equivalent to the hypothesis on the equality of means in all groups for each cell of the contingency table. The problem of discriminatory power of variables in the location model was defined in such a way by Krzanowski [11], however he considered also the equality of cell probabilities.

The hypothesis H_B can be written in the matrix form (3), and then the formulae for the matrices H and G may be obtained. This way, however, is a complicated one. Much simpler is to obtain them basing on the form of the

test statistic in the univariate analysis of variance (ANOVA) (see [1]). For the nested model in two-way classification ANOVA the test statistic is given as (see, e.g., [2])

$$F_B = \frac{\sum_{i=1}^{l}\sum_{j=1}^{g_i} n_{ij}(\bar{y}_{ij.}-\bar{y}_{i..})^2}{g_.-l} : \frac{\sum_{i=1}^{l}\sum_{j=1}^{g_i}\sum_{h=1}^{n_{ij}} (y_{ijh}-\bar{y}_{ij.})^2}{N-g_.}, \tag{6}$$

where

$$N = \sum_{i=1}^{l}\sum_{j=1}^{g_i} n_{ij}, \qquad g_. = \sum_{i=1}^{l} g_i,$$

y_{ijh} is the value of the variable y for the h-th observation, the j-th level of B and the i-th level of A, $\bar{y}_{ij.}$ is the mean value for the j-th level of B and the i-th level of A, $\bar{y}_{i..}$ is the mean value for the i-th level of A.

Now the matrices H and G are obtainable by analogy to the sums of squares in the numerator and the denominator of statistic (6). Thus

$$\begin{aligned} H &= \sum_{i=1}^{l}\sum_{j=1}^{g_i} n_{ij}(\bar{y}_{ij.}-\bar{y}_{i..})(\bar{y}_{ij.}-y_{i..})', \\ G &= \sum_{i=1}^{l}\sum_{j=1}^{g_i}\sum_{h=1}^{n_{ij}} (y_{ijh}-\bar{y}_{ij.})(y_{ijh}-\bar{y}_{ij.})', \end{aligned} \tag{7}$$

where $\bar{y}_{ij.}$, $\bar{y}_{i..}$, y_{ijh} are vectors of p components.

H is called here the *matrix of between group adjusted squares and products*, and G is the *matrix of within group adjusted squares and products.*

The linear transformation of original data into the space of canonical variates is now possible after solving the eigenvalue problem (4) with the matrices H, G defined as in (7). This enables a clear graphical representation of single individuals (cells or groups) on the plane in the space of the most significant canonical variates which correspond to the largest eigenvalues of the matrix HG^{-1}.

The linear transformation of the data obtained basing on the matrix of between group adjusted squares and products enables the best separation of the considered groups and is a new one and different from those given by Krzanowski [8] as the result of between-cell analyses. In addition, it has a straightforward connection with the selection procedure by T^2 statistic, which gives the possibility of the simultaneous choice of continuous and discrete variables to the location model.

Further, it should be noted that the significance of canonical variates in that particular case of the location model may be tested by the χ^2 statistic as in the one-way classification problem. For the general model the χ^2 statistic is given as (see [1])

$$\chi^2 = (n-r-p+t_1+1)(\lambda_{t_1}+\lambda_{t_1+1}+\ \dots\ +\lambda_t) \sim \chi^2_{s(p-t_1)}, \tag{8}$$

where n, r, s, p are defined in (1) and (2).

The test χ^2 (formula (8)) enables us to verify that the last $t-t_1$ canonical variates are redundant in reference to the first t_1 variates. The test may be performed for $t_1 = 0, 1, \ldots, t-1$.

Now, if for all $t_1 < v$ the $t-t_1$ canonical variables are nonredundant and redundant for $t_1 \geqslant v$, then the dimensionality of the discriminant space equals v.

In the location model case, the statistic (8) takes the form (for specifying r and s compare (6))

$$\chi^2 = (N-g_{.}-p+t_1+1)(\lambda_{t_1}+\lambda_{t_1+1}+\ \ldots\ +\lambda_t) \sim \chi^2_{(g_{.}-1)(p-t_1)}.$$

Another canonical representation of the location model studied by Krzanowski [7] in the dichotomous problem was based on the matrix of intercell distances. The latter was obtained for all $2l$ ($l = 2^q$) cells.

4. Example. The presented example of application is a part of the more extended studies on the chronic obturative lung disease. The sample of patients consists of 164 persons suffering from uncomplicated bronchial asthma ($n_1 = 112$) and bronchial asthma complicated by lung emphysema ($n_2 = 54$). 14 predictor variables, i.e., 6 continuous ones, called here $C_1, C_2, \ldots, C_6$ (5 spirometric examinations and smoking index), and 8 binary ones denoted by $B_1, B_2, \ldots, B_8$ (disease symptoms such as cough, dyspnea, findings of the X-ray examination of the chest) are considered.

As the first, a selection of variables basing on the multivariate discriminatory measure T^2 was performed. The distribution of T^2 is approximated by Snedecor F with the numbers of degrees of freedom depending on the number of nonempty cells and the number of continuous variables. Thus, comparing the values of T^2 for different models, the distributional approach should be used or the generalized measure $T_1^2 = g_{.}T^2$ (see [5]) ought to be alternatively applied.

Let us consider here a distributional approach. For the whole set of variables we have the probability

$$\Pr(F > F_{\text{calc}}) = 1.2702_{10}-1.$$

For the subsets of 9 ($C_1, C_2, \ldots, C_6, B_3, B_6, B_7$) and 6 variables ($C_1, C_3, C_4, C_5, C_6, B_3$) obtained by the backward elimination procedure it equals $5.8884_{10}-7$ and $3.0072_{10}-10$, respectively. The best result was obtained for the subset of only 3 continuous variables which is not considered here, but it was at a similar level as for 6 variables. The canonical representations in the space of the two variates corresponding to the largest eigenvalues of the problem (4) are presented in Fig. 1 for both sets considered. The first representation (a) concerns the subset of 9 variables with three binary ones (B_3, B_6, B_7) among them. It is possible to have here $l = 2^3$ cells, but only 6 of them, denoted by A, B, C, D, E, F, are nonempty. This notation is enriched by the number of the group (1 or 2). Three cells — $A2$, $B1$ and

$F2$ – are placed in the same position and denoted by $*$. The graphical representation is prepared in the way to obtain the best separation between both groups.

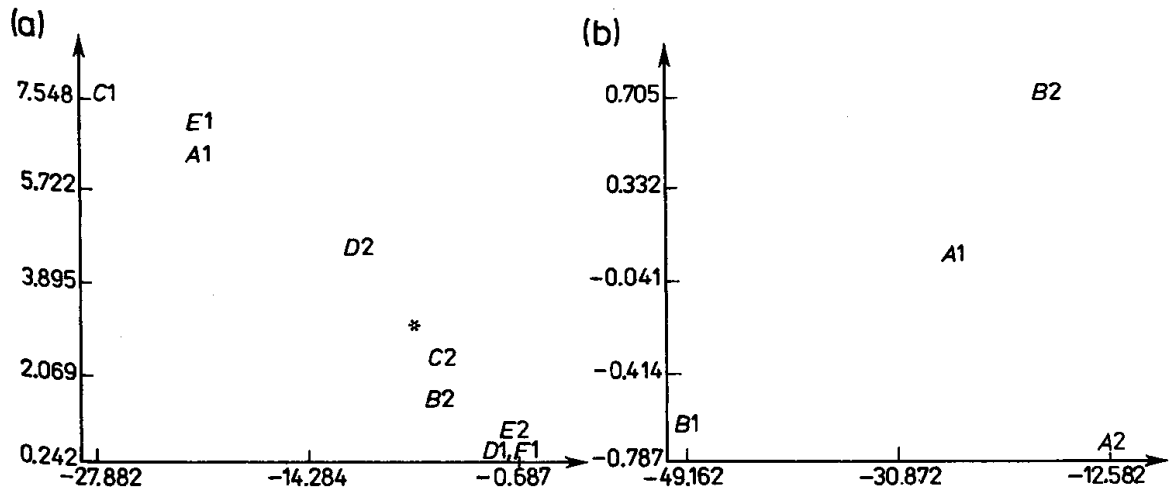

Fig. 1. Canonical representations of the location model

(a) subset of 9 variables ($*$ indicates cells $A2$, $B1$, $F2$), (b) subset of 6 variables

The cells $A1$ and $A2$, $C1$ and $C2$, $D1$ and $D2$, $E1$ and $E2$ are fine separated, the cells $F1$ and $F2$ a little worse, but the cells $B1$ and $B2$ are placed quite not far away. The reclassification of the sample with the whole location model procedure by the leaving-one-out method and basing on the smoothed estimates has given 49 incorrectly classified individuals. The second representation (b) is prepared for the subset of 6 variables which is better than the former according to the discriminatory power (compare: the results of selection). There are only $l = 2$ cells, both nonempty. The cells $A1$ and $A2$ as well as $B1$ and $B2$ are fine separated. This is confirmed by the better results of reclassification (only 44 incorrectly classified individuals).

References

[1] H. Ahrens and J. Läuter, *Mehrdimensionale Varianzanalyse*, Akademie Verlag, Berlin 1974.

[2] A. Bartkowiak, *Opis merytoryczny programów statystycznych*, Wrocław University Press, Wrocław 1982.

[3] J. J. Daudin, *Selection of variables in mixed-variable discriminant analysis*, Biometrics 42 (1986), pp. 473–482.

[4] S. J. Haberman, *Log linear fit for contingency tables. Algorithm AS51*, Applied Statistics 21 (1972), pp. 218–225.

[5] E. Krusińska, *New procedure for selection of variables in location model for mixed variable discrimination*, Biometrical J. (to appear).

[6] W. J. Krzanowski, *Discrimination and classification using both binary and continuous variables*, J. Amer. Statist. Assoc. 70 (1975), pp. 782–790.

[7] — *Canonical representation of the location model for discrimination or classification*, ibidem 71 (1976), pp. 845–848.

[8] — *Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis*, Biometrika 66 (1979), pp. 33–39.

[9] — *Mixtures of continuous and categorical variables in discriminant analysis*, Biometrics 36 (1980), pp. 493–499.

[10] — *Stepwise location model choice in mixed-variable discrimination*, Applied Statistics 32 (1983), pp. 260–266.

[11] — *On the null distribution of distance between two groups, using mixed continuous and categorical variables*, Journal of Classification 1 (1984), pp. 243–253.

[12] — *Multiple discriminant analysis in the presence of mixed continuous and categorical data*, Comput. Math. Appl. 12a (1986), pp. 179–185.

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

[14] G. A. F. Seber, *Multivariate Observations*, J. Wiley, New York 1984.

INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
51-151 WROCŁAW

DEPARTMENT OF INTERNAL DISEASES
MEDICAL ACADEMY OF WROCŁAW
50-477 WROCŁAW

Received on 1987.05.14