CLASSIFICATION AS A TOOL OF RESEARCH
W. Gaul and M. Schader (Editors)
© Elsevier Science Publishers B.V. (North-Holland), 1986

MULTIPLE DISCRIMINANT ANALYSIS FOR PREDICTOR VARIABLES MEASURED AT VARIOUS SCALE LEVELS

Yoshio TAKANE
Department of Psychology, McGill University
1205 Dr. Penfield Avenue, Montreal, PQ H3A 1B1, Canada

Various methods have been developed for discriminant analysis (Lachenbruch, 1975). For continuous multivariate normal predictors a method based on canonical variates has been widely used (Fisher, 1936). The analytically same method is also used, though in a descriptive manner, for discrete data (Fisher, 1948; Hayashi, 1952), for which no definitive methods exist (Goldstein & Dillon, 1978). One promising method for discrete data is to use the log-linear model for discriminant analysis (e.g., Anderson, 1980). This method is appealing, since it is equipped with a built-in mechanism for choosing the best combination of predictor variables (including interactions among them.) A disadvantage is that it cannot be directly applied to continuous data. In this paper we discuss a method of multiple discriminant analysis which allows a mixture of continuous and discrete predictor variables, and which allows various statistical inferences.

In the proposed method the subjects (or any other sample units) are mapped into a multidimensional euclidern space. Coordinates of subject points are simple linear functions of the predictor variables. It is further assumed that there are ideal points corresponding to criterion groups, and that they are defined to be centroids of the subject points in the respective groups. The probability of a subject belonging to a particular criterion group is specified as a decreasing function of the (squared) distance between the subject point and the ideal point of the criterion group.

1. METHOD

Let $f_{k\alpha}$ ($k=1, \ldots, N$; $\alpha = 1, \ldots, n_g$) denote the observed frequency of observation (or response pattern) k on predictor variables in criterion group α . Observations on the predictor variables are denoted by G, where discrete predictors are coded into dummy variables and continuous variables are normalized appropriately. Let Y be the matrix of coordinates of subject points. We assume that

$$(1) \quad Y_A = GX_A ,$$

where X_A is the matrix of weights (analogous to the regression coefficients in the usual regression analysis), and A is the prescribed dimensionality of the space (the maximum value of A is $n_g - 1$). Let M denote the matrix of coordinates of group centroids. Then

$$(2) \quad M_A = (H'H)^{-1}H'Y_A = (H'H)^{-1}H'GX_A ,$$

where H is the matrix of dummy variables indicating the criterion groups. The squared euclidean distance between a subject point (k) and the centroid of a criterion group (α) is then given by

$$(3) \quad d^2_{k\alpha}(X_A) = \sum_a (y_{ka} - \mu_{\alpha a})^2$$

(which is the function of X_A), where y_{ka} and $\mu_{\alpha a}$ are appropriate elements of Y and M, respectively.

Let $p_{k\alpha}$ denote the conditional probability of group α given observation k on predictor variables. We postulate that

$$(4) \quad p_{k\alpha} = \frac{p_\alpha \exp(-d^2_{k\alpha})}{\sum_\beta p_\beta \exp(-d^2_{k\beta})} ,$$

where p_α is the prior probability of group α. Model (4) simply states that $p_{k\alpha}$ is proportional to $p_\alpha \exp(-d^2_{k\alpha})$; that is, $p_{k\alpha} = c\, p_\alpha \exp(-d^2_{k\alpha})$ for some constant $c(\neq 0)$. But since $\sum_\beta p_{k\beta} = 1$, $c = (\sum_\beta p_\beta \exp(-d^2_{k\beta}))^{-1}$; that is, the denominator of (4) is just a normalization factor. The model posits that $p_{k\alpha}$ increases proportionally to p_α and that $p_{k\alpha}$ decreases proportionally to $\exp(-d^2_{k\alpha})$ as $d_{k\alpha}$ increases. The likelihood of the total set of observations is now stated as

$$(5) \quad L = \prod_k \prod_\alpha (p_{k\alpha})^{f_{k\alpha}} .$$

Once model parameters are estimated so as to maximize (5), classifications may be made according to $\max_\alpha (p_{k\alpha})$.

An important feature of the above model is that the $\exp(-d^2_{k\alpha})$ part of the model remains intact, even if marginal frequencies of criterion groups are fixed apriori (separate sampling). The model can be fitted as if only the total sample size were fixed (joint sampling). The only difference is that an estimate of p_α other than f_α/f should be used in calculating $p_{k\alpha}$ for classification (Anderson, 1972), since in the separate sampling situation f_α/f does not reflect a population group size, where $f_\alpha = \sum_k f_{k\alpha}$ and $f = \sum_\alpha f_\alpha$.

2. CONSTRAINTS ON X_A

Different types of predictor variables are distinguished by different types of constraints on X.

(a) Unordered categorical variable: Multidimensional quantifications of categories will be obtained subject to

(6) $\Sigma\, f_{i(j)} x_{i(j)a} = 0$ for all a,

where $f_{i(j)}$ is the marginal frequency of category j in item i. The summation is over all categories in item i. This type of restriction is necessary in order to eliminate nonuniqueness of coefficients due to the linear dependency among categorical variables.

(b) Ordered categorical variable: Unidimensional quantifications that conform to apriori specified orders are obtained under a restriction similar to (6). Those unidimensionally quantified categories are weighted multidimensionally to yield multidimensional quantifications.

(c) Continuous variable: It is assumed that quantifications are apriori given, and only dimensional weights are estimated.

With the above constraints Fisher's scoring algorithm used to maximize (5) has been found remarkably efficient, even when the sample size is moderate to small.

3. EXAMPLE

We give just one example of applications of the above method. Data are from Maxwell (1961) and consist of three criterion groups (Schizophrenics, Manic-Depressive and Anxiety States) and four binary predictor variables, each indicating presence (1) or absence (0) of a certain symptom. The four symptoms are: 1 anxiety, 2 suspicion, 3 schizophrenic type of thought disorders, 4 delusions of guilt. The data are given in Table 1. There are sixteen possible response patterns taken on the four binary variables, and observed frequencies of the sixteen patterns in the three criterion groups are given.

The simplest way to perform a discriminant analysis on the data is to use $f_{k\alpha}/f_k$ (where $f_k = \sum_\alpha f_{k\alpha}$) as an estimate of $p_{k\alpha}$. This is often called the full multinomial model. It involves a minimal set of assumptions but uses a large number of parameters. Thus we might ask if our model performs better than this benchmark model, and if it does, what the best dimensionality is. For assessment of goodness of fit (GOF) we use the AIC statistic defined by AIC = -2 ln L + 2 x (number of model parameters). A model with a smaller AIC value is considered a better model. The AIC values are 854.7(32), 947.8(6) and 841.3(9) for full multinomial, one-dimensional and two-dimensional solutions, respectively, with numbers of model parameters in parentheses. This implies that the two-dimensional solution is the best fitting model. This solution also achieves the minimum attainable apparent error rate (.313) under the present circumstance.

Figure 1 displays the derived configuration of the sixteen response patterns and centroids of the three criterion groups (in parentheses) in the best fitting model. Dotted lines indicate boundary hyperplanes according to the maximum probability rule. Solid lines indicate boundary hyperplanes assuming equal group size. They are merely a translation of the original hyperplanes. This is in line with the point made above that prior probabilities do not affect $\exp(-d^2)$ part of the model.

Figure 2 shows plots of estimated weights for categories in the predictor variables. Paired numbers indicate item numbers followed by category numbers. We see close relationships between group 1 and 3-1, 2-1, group 2 and 3-0, 2-0, and group 3 and 1-1. (We should be cautious in this interpretation because of possible multicolinearity.)

We can determine if a particular predictor variable significantly contributes to discrimination. Table 2 shows estimated weights for categories along with their standard error estimates in parentheses. Under the asymptotic normality assumption, we may compute confidence intervals, and if they cover zero, we may conclude that the weights are not significantly different from zero. This can be done for each dimension, but we should remember that the configuration is rotatable. Those weights not significantly different from zero are underlined in the table. All four variables are useful in at least one of the dimensions. We may also assess the total contribution of each variable by fitting the model with the variable of concern deleted from the predictor set and comparing its GOF with that of the full model. The AIC values for variables 1, 2, 3 and 4, each deleted in turn are, respectively, 945.5, 870.9, 1012,5 and 915.2, none of which are as small as the AIC value for the full model. We thus conclude the four-predictor case is the best fitting model.

4. DISCUSSION

The proposed method can handle any mixture of different types of predictor variables. It can also handle both joint and separate sampling situations. It has a model evaluation feature, which enables the researcher to choose the best dimensionality and/or the optimal set of predictor variables thorugh the use of the AIC statistic (Akaike, 1974). It is also possible to determine whether a continuous predictor variable is better treated as it is, or better categorized and treated as such.

The proposed method is flexible enough to accommodate various changes (e.g., a different distance function) in the model to widen its applicability.

The proposed method is similar to the logistic discrimination (regression) model proposed by several authors at about the same time (Cox, 1966; Day & Kerridge, 1967; Walker & Duncan, 1967) for two-group situations, and later extended by Anderson (1972) to multiple-group situations. It can be shown that the present method gives similar results to the logistic discrimination model, if a maximum possible dimensionality (i.e., the number of criterion groups minus one) is taken. However, full dimensionality is often unnecessary or even harmful when the sample size is small. The logistic discrimination model has no provision for possible dimension reduction. With the proposed method, on the other hand, the best dimensionality can be chosen entirely on an empirical basis.

Table 1. Data from Maxwell (1961)

Pattern Number	Predictors				Observed Frequency in Groups		
					I	II	III
1	0	0	0	0	38	69	6
2	0	0	0	1	4	36	0
3	0	0	1	0	29	0	0
4	0	0	1	1	9	0	0
5	0	1	0	0	22	8	1
6	0	1	0	1	5	9	0
7	0	1	1	0	35	0	0
8	0	1	1	1	8	2	0
9	1	0	0	0	14	80	92
10	1	0	0	1	3	45	3
11	1	0	1	0	11	1	0
12	1	0	1	1	2	2	0
13	1	1	0	0	9	10	14
14	1	1	0	1	6	16	1
15	1	1	1	0	19	0	0
16	1	1	1	1	10	1	0
		Totals			224	279	117

Table 2. Estimated weights and their standard errors

Predictor variable	Cat.	dim 1	dim 2
1	1	.47 (.04)	-.48 (.06)
	2	- .39 (.03)	.40 (.05)
2	1	.16 (.03)	-.10 (.05)
	2	.40 (.07)	.25 (.12)
3	1	- .29 (.02)	-.10 (.04)
	2	1.11 (.08)	.37 (.17)
4	1	.00 (.03)	.40 (.04)
	2	- .01 (.08)	-1.13 (.11)

Coefficients (Standard Error)

Figure 1. Derived stimulus configuration of the sixteen response patterns and centroids of the three criterion groups

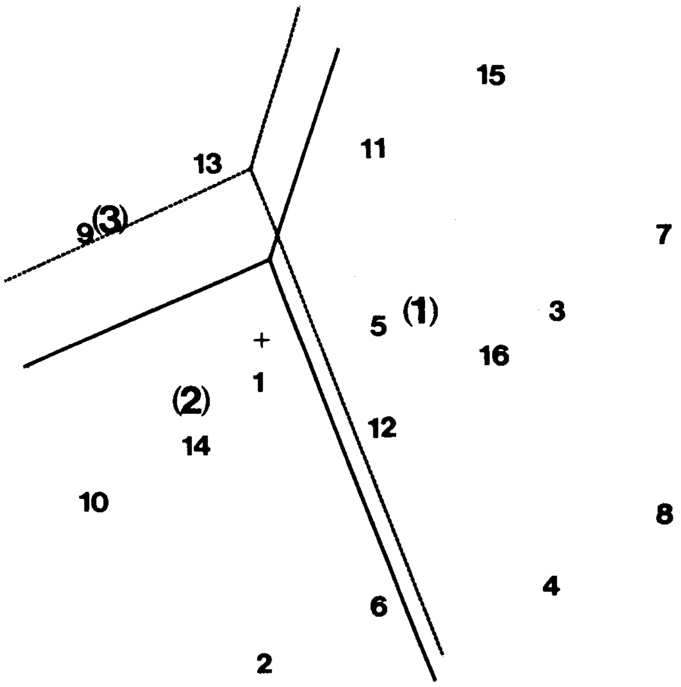
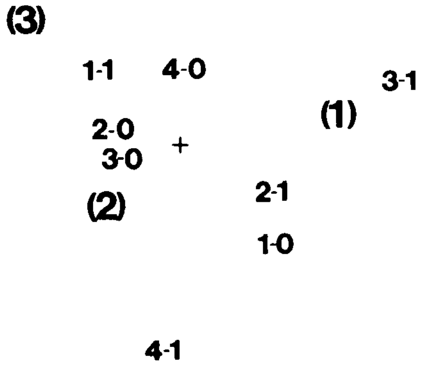


Figure 2. Plots of estimated weights for categories



Akaike, H. A new look at the statistical model identification. IEEE Transactions on automatic control, 1974, 19, 716-723.

Andersen, E.B. Discrete statistical models with social science applications. Amsterdam: North-Holland, 1980.

Anderson, J.A. Separate sample logistic discrimination. Biometrika, 1972, 59, 19-35.

Cox, D.R. Some procedure connected with the logistic qualitative response curve. In David, F.N. (Ed.), Research papers in statistics Festschrift for J. Neyman. New York: Wiley, 1966, 55-71.

Day, N.E ., aned Kerridge, D.F. A general maximum likelihood discriminant. Biometrics, 1967, 23, 313-323.

Fisher, R.A. Statistical methods for research workers. London: Oliver & Boyd, 1948, (7th printing).

Fisher, R.A. The use of multiple measurement in taxonomic problems. Annals of Eugenics, 1936, 7, 179-188.

Goldstein, M. and Dillon, W.R. Discrete discriminant analysis. New York: Wiley, 1978.

Hayashi, C. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. Analysis of the Institute of Statistical Mathematics, 1952, 2, 69-98.

Lachenbruch, P.A. Discriminant analysis. New York: Hafner, 1975.

Maxwell, A.E. Canonical variate analysis when the variables are dichotomous. Educational and Psychological Measurement, 1961, 21, 259-271.

Walker, S.H. & Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. Biometrika, 1967, 54, 167-179.