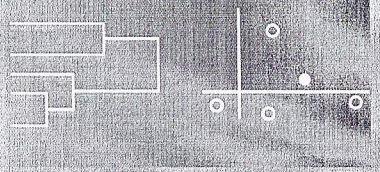
AND KNOWLEDGE ORGANIZATION

Data Analysis and Classification

Francesco Palumbo
Carlo Natale Lauro
Michael J. Greenacre
Editors



Springer

Data Analysis and Classification

The volume provides results from the latest methodological developments in data analysis and classification and highlights new emerging subjects within the field. It contains articles about statistical models, classification, cluster analysis, multidimensional scaling, multivariate analysis, latent variables, knowledge extraction from temporal data, financial and economic applications, and missing values. Papers cover both theoretical and empirical aspects.

ISSN 1431-8814

ISBN 978-3-642-03738-2

9 783642 037382

springer.com

Discriminant Analysis on Mixed Predictors

Rafik Abdesselam

Abstract The processing of mixed data – both quantitative and qualitative variables – cannot be carried out as explanatory variables through a discriminant analysis method. In this work, we describe a methodology of a discriminant analysis on mixed predictors. The proposed method uses simultaneously quantitative and qualitative explanatory data with a discrimination and classification aim. It's a classical discriminant analysis carried out on the principal factors of a Mixed Principal Component Analysis of explanatory mixed variables, i.e. both quantitative and transformed qualitative variables associate to the dummy variables. An example resulting from real data illustrates the results obtained with this method, which are also compared with those of a logistic regression model.

1 Introduction

The methodology of quantification qualitative variables evolved in the context of Mixed Principal Component Analysis (MPCA) (Abdesselam 2006) is used here in a discrimination and classification aim on explanatory mixed variables. Discriminant analysis in its usual version use only quantitative predictors (Fisher 1938). Since, a methodology called DISQUAL method (Saporta 1977) allows to extend the context of discriminant analysis to qualitative predictors. The proposed Mixed Discriminant Analysis (MDA) approach allows to implement a discriminant analysis with the two types of predictors, this is the main aim of this work; to extend the discriminant model context for using mixed predictors like, for example, logistic model or discriminant partial least squares (PLS) approach. The proposed approach is evaluated then compared to the logit model on the basis of real mixed data. These analyses are carry out by discrimination with two groups on principal factors procedure of

R. Abdesselam
ERIC EA 3038, University of Lyon 2, 69676 Bron, France
e-mail: rafik.abdesselam@univ-lyon2.fr

SPAD software for MDA and by logistic procedure of SAS software for the logistic model.

2 Mixed Discriminant Predictors

We use the following notations to explain the methodology which consists in transforming the qualitative explanatory variables on quantitative variables for the discriminant model. Let us denote:

- $Z_{(n,r)}$ the qualitative data matrix associated to $\{z^t; t = 1, r\}$, the dummy variables of the variable z with r modalities or groups that we wish to discriminate.
- $X_{(n,p)}$ the quantitative data matrix associated to the set of p discriminant variables $\{x^j; j = 1, p\}$, with n rows-individuals and p columns-variables.
- $(y_1, \ldots, y_l, \ldots, y_m)$ the set of m qualitative discriminant variables with $q = \sum_{l=1}^{m} q_l$ dummy variables $\{y_l^k; k = 1, q_l\}_{\{l=1,m\}}$.
- $Y_{l\,(n,q_l)}$ the dummy variables matrix associated to the q_l modalities of the variable y_l.
- $Y_{(n,q)} = [Y_1, \ldots, Y_l, \ldots, Y_m]$ global matrix, juxtaposition of the matrix $Y_{l\,(n,q_l)}$.
- $E_z = R^r$, $E_x = R^p$ and $E_y = \oplus\{E_{y_l}\}_{\{l=1,m\}} = R^q$ are the individual subspaces associated by duality respectively to the data matrix $Z_{(n,r)}$, $X_{(n,p)}$ and $Y_{(n,q)}$.
- $D = \frac{1}{n} I_n$ diagonal weights matrix of the n individuals and I_n the unit matrix with n order.
- $N_x = \{x_i \in E_x\,;\, i = 1, n\}$ and $N_{y_l} = \{y_i \in E_{y_l}\,;\, i = 1, n\}$ are the configurations of the individual-points associated to the rows of the matrix $X_{(n,p)}$ and $Y_{l\,(n,q_l)}$.
- $M_x = V_x^+$ and $M_{y_l} = \chi^2_{y_l}$ are the matrix of inner product, the Mahalanobis distance in E_x and the Chi-square distance in E_{y_l}.
- $V_{xy_l} = {}^t X D Y_l$ the matrix of covariances.
- $P_{E_{y_l}}$ the orthogonal projection operator in subspace E_{y_l}.

The quantification of qualitative data is made with the statistical and geometrical construction of m configurations of individual-points $\hat{N}_x^{y_l} = \{P_{E_{y_l}}(x_i); x_i \in N_x\} \subset E_{y_l}$. For all l=1 to m, we note $\hat{X}^{y_l} = X V_x^+ V_{xy_l}$ the data matrix of order (n, q_l) associated to the project configuration of individual points $\hat{N}_x^{y_l}$; the subspace E_{y_l} is considered as an explanatory subspace on which we project the configuration of individual points N_x of quantitative data in the explain subspace E_x.

It is shown in Abdesselam (2006) the following remark and property concerning the Mixed Principal Component Analysis (MPCA).

Remark. The PCA $(\hat{X}^{y_l}; \chi^2_{y_l}; D)$ is equivalent to Multivariate ANalysis Of Variance (MANOVA) between the p quantitative variables and the q_l dummy variables

associated to the levels of the explained factor y_l, which $I(\hat{N}_x^{y_l}) = trace(V_{y_l x} V_x^+ \times V_{xy_l} \chi^2_{y_l})$, the explained inertia is equal to Pillai's trace.

Property. The MPCA of the mixed data table $[X \mid Y]_{(n;p+q)}$ consists to carry out the standardized PCA of the data table $[X \mid \tilde{Y}]_{(n;p+q)}$.

Where, $\tilde{Y}_{(n,q)} = [\tilde{Y}_1, \ldots, \tilde{Y}_l, \ldots, \tilde{Y}_m]$ is a juxtaposition matrix of transformed qualitative data, with $\tilde{Y}_l = Y_l - {}^t G_l$ the quantitative data matrix of order (n, q_l) associated to the configuration of individual points $N_{\tilde{y}_l} \subset E_{y_l}$ that inertia $I(N_{\tilde{y}_l}) = q_l - 1$, where $G_l = {}^t \hat{X}^{y_l} D\, 1_n$, is the mean vector of the variables $\hat{X}^{y_l}$ and 1_n the unit vector with n order.

Note that MPCA is equivalent to Mixed Data Factorial Analysis (MDFA) (Pagés 2004). The main aim of these two methods is to research principal components, noted F^s, which maximize the following mixed criterion, proposed in square correlation terms in Saporta (1990) and geometrically in terms of square cosinus of angles in Escofier and Pagés, J. (1979):

$$\sum_{j=1}^{p} r^2(x^j, F^s) + \sum_{l=1}^{m} \eta^2(y_l, F^s) = \sum_{j=1}^{p} \cos^2 \theta_{js} + \sum_{l=1}^{m} \cos^2 \theta_{ls}$$

where, r^2 and η^2 are respectively the square of the linear correlation coefficient of quantitative variables and the correlation ratio of qualitative variables with the s^{th} factor, and θ the angle between the correspondent vectors. These two expressions are equal in view of fact that the variables are normalized.

In a methodological point of view, the MDA appears as a chain of two procedures: a projection procedure of configurations of points corresponding to the MANOVA coordinates to quantify the qualitative variables, we take into account the correlation ratios, then a standardized PCA procedure to synthesize the linear correlations between all variables, quantitative and transformed qualitative variables.

Definition 1. The MDA $[X \mid Y]_{(n;p+q)} \longrightarrow Z_{(n;r)}$ consists to carry out a discriminant analysis on the principal factors of the MPCA of mixed data table $[X \mid Y]_{(n;p+q)}$.

So, this extension methodology of discriminant analysis on mixed variables, that we can call DISMIX method (DIScrimination on MIXed variables), is like DISQUAL method (DIScrimination on QUALitative variables), which consists to make a discriminant analysis on factors of Multiple Correspondence Analysis (MCA) of explanatory variables (Saporta 1977).

We can note that the first principal factors of MPCA (respectively MCA) are not necessary the better discriminant factors of DISMIX (respectively DISQUAL) method, but we can select only the significant discriminant factors. We obtain satisfactory discrimination results with these methods.

3 Application Example

To illustrate this approach then to compare it with logistic model, we use data of an application example taken from the library SAS System. In this study of the analgesic effects of treatments on elderly patients with neuralgia, two test treatments and a placebo are compared. This data set contains the responses of $p = 2$ explanatory quantitative variables: Age of the patients and the Duration of complaint before the treatment began and $m = 2$ explanatory qualitative variables with $q = 5$ modalities in total: Treatment (A, B, Placebo) and Sex (Female, Male) of the patients according to the response explain variable Pain with two groups: is whether the patient reported pain or not (Yes_{25}, NO_{35}).

This sample of size $n = 60$ patients is subdivided into two samples: a basic-sample or "training set" composed of $n_1 = 55$ (90%), randomly drawn from the whole data set for the discriminant rule and a test-sample or "validation set" of size $n_2 = 5$ (10%) for next evaluated the performance of this rule.

Moreover the fact to compare the two test treatments and a placebo, the aim is to bring to the fore the mixed characteristics which well differentiate the two groups of patients.

3.1 Predictor Analysis

First, we analyze and describe only the predictors using Mixed Principal Component Analysis (MPCA). This analysis extracts in total five factors $(p + q - m)$ given in Table 1.

Table 2 gives the linear correlations between mixed predictors and MPCA factors. Figure 1 shows the graphical representations of the quantitative and transformed qualitative variables, on the MPCA factorial planes which explain 90.16% of the total variability. The first axis (30.18%) opposes men to women patients, the second one (22.69%) compares treatment B and placebo. While the third axis (21.21%) summarizes the transformed variable treatment A, the fourth axis (16.07%) synthesizes and opposes the age variable to duration variable.

Table 1 MPCA eigenvalues

Number	Eigenvalue	Proportion (%)	Cumulative (%)
1	2.1129	30.18	30.18
2	1.5886	22.69	52.88
3	1.4850	21.21	74.09
4	1.1249	16.07	90.16
5	0.6885	09.84	100.00

Table 2 Correlation mixed variables – factors

Iden.	Wording variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
AGE	Age of the patient	−0.37	−0.03	−0.13	+0.73	+0.55
DURA	Duration	+0.14	+0.49	+0.04	−0.65	+0.57
TREA	Treatment A	+0.31	−0.01	+0.94	+0.15	+0.05
TREB	Treatment B	−0.15	+0.82	−0.49	0.14	−0.18
TREP	Treatment placebo	−0.17	−0.82	−0.45	−0.29	+0.13
FEMA	Female-sex	+0.95	−0.05	−0.26	+0.15	+0.06
MALE	Male-sex	−0.95	+0.05	+0.26	−0.15	−0.06

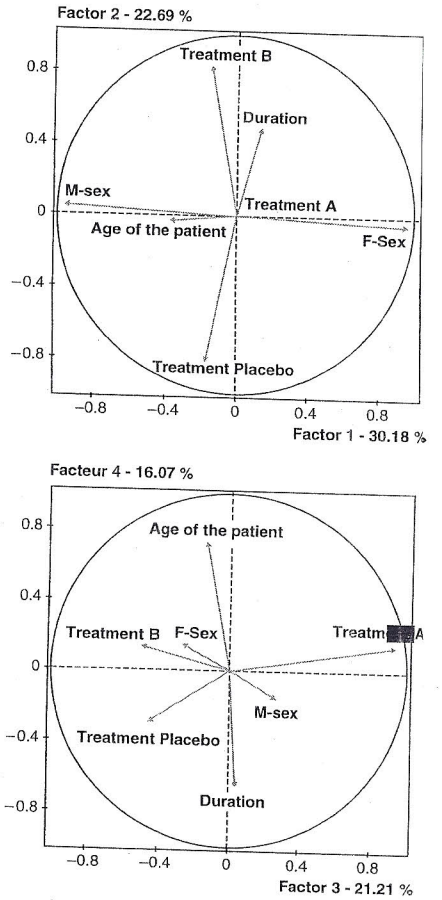


Fig. 1 Circles of correlations: mixed predictors on the first and second MPCA factorial planes

3.2 Discriminant Analysis

We use a discriminant analysis on the significant MPCA factors (corresponding to the four first components with an eigenvalue larger than unity) which explain 89.90% of the variance kept for the discrimination (see Table 1).

Table 3 presents the Fisher discriminant linear function of the MDA with two groups on MPCA factors of explanatory mixed variables. This discriminant rule is computed from the training set of 55 observations. The obtained results show that the discriminant model overall is very significant, the probability (PROBA = 0.0001) is less than the classical significance level of 5%.

So, among the four introduced mixed variables, we can note that, with a significance level less or equal to 5%, neither the duration nor the treatment A differentiate the two groups of patients (PROBA > 5%).

Indeed, the patients who did not report pain are women less elderly who had been given treatment B. However, the group patients reporting the most pain are more elderly men who had been given the placebo.

Table 4 presents some results of logistic model applied to the same training set, implement with the logistic procedure of SAS System. The estimation and the significance of the parameters estimated by the binary logistic model are presented. In this model, the reference modalities for explain variable "Pain" and explanatory variables "Treatment" and "Sex" are respectively "No pain", "Placebo" and "Male". The likelihood ratio, score and Wald tests lead all to reject the nullity hypothesize of the set of coefficients. So, with a classical error risk of 5%, only Duration and Treatment A don't have a significant marginal apport in this full model.

3.3 Comparison

In this part, we use the criterion of misclassification rates to evaluate and compare the performances of the discrimination rules of MDA and Logistic methods.

Table 3 Mixed discriminant analysis – SPAD results

FISHER'S LINEAR FUNCTION

VARIABLES NUM IDEN LABEL	PARAMETER ESTIMATE FUNCTION DISC.	REGRESSION (RES. TYPE REG.)	STANDARD DEVIATION	T STUDENT	PROBA
2 AGE Age of the patient	-0.2186	-0.0646	0.0218	2.97	0.005[a]
3 DURA Duration	0.0137	0.0041	0.0097	0.42	0.677
4 TREA Treatment A	0.8076	0.2387	0.1547	1.54	0.129
5 TREB Treatment B	1.1590	0.3426	0.1584	2.16	0.036[b]
6 TREP Placebo	-1.9666	-0.5814	0.1551	3.75	0.000[a]
7 FEMA Female patient	0.9656	0.2855	0.1111	2.57	0.013[b]
8 MALE Male patient	-0.9656	-0.2855	0.1111	2.57	0.013[b]
INTERCEPT	14.606855	4.248122			

R2 = 0.42246 F = 7.16850 PROBA = 0.0001
D2 = 2.89710 T2 = 38.76842 PROBA = 0.0001

[a]Significance less or equal than 1%
[b]Significance 1–5%

Table 4 Binary logistic model – SAS results

Model fit statistics

Criterion	Intercept only	Intercept and covariates
AIC	76.767	57.280
SC	78.74	69.324
−2 Log L	74.767	45.280

Testing global null hypothesis: BETA = 0

Test	Chi-square	DF	Pr > ChiSq
Likelihood ratio	29.4864	5	<0.0001[a]
Score	23.2353	5	0.0003[a]
Wald	13.2742	5	0.0209[b]

Analysis of maximum likelihood estimates

Parameter	DF	Estimate	Standard error	Wald chi-square	Pr > ChiSq
Intercept	1	17.4418	6.8320	6.5176	0.0107[b]
Treatment A	1	0.7498	0.5324	1.9836	0.1590
Treatment B	1	1.2554	0.6128	4.1970	0.0405[b]
Sex female	1	0.9682	0.4119	5.5247	0.0188[b]
Age	1	−0.2457	0.0953	6.6448	0.0099[a]
Duration	1	0.0183	0.0350	0.2737	0.6009

[a]Significance less or equal than 1%
[b]Significance 1–5%

Table 5 Comparison – number of observations (percent) well classified into group

	Reported pain groups	MDA	Logistic	Total
Basic sample (90%)	No pain	30 (93.75%)	28 (87.50%)	32
	Yes pain	15 (65.22%)	18 (78.26%)	23
	Total	45 (81.82%)	46 (83.64%)	55
Test sample (10%)	No pain	3 (100.00%)	3 (100.00%)	3
	Yes pain	1 (50.00%)	1 (50.00%)	2
	Total	4 (80.00%)	4 (80.00%)	5

Table 5 shows that the classification results obtained by these two methods on the basic and test samples, are very similar. Indeed, on the training set of 55 observations, the estimations of well classification probabilities are practically the same, namely 81.82% for MDA and 83.64% for logistic model. This corresponds with 45 and 46 observations, respectively.

When we estimate the misclassification probabilities based on the validation set that consists of the remaining five observations, we obtain the same results for MDA and Logistic model.

4 Conclusion

In this work, the methodology to extend discriminant analysis to mixed variables is presented as a methodological chain of known factorial methods. Simple in concept and easy to use, it finds interest in the context of the classification and prediction techniques, when user is confronted with analyzing objects characterized by mixed variables, as is often the case, especially in economics, financial and insurance fields.

The Mixed Discriminant Analysis proposed allows to implement a discriminant analysis on the two types of predictors. This method comes up to one of the disadvantages of discriminant analysis in relation to logistic regression. The latter being a rival if we look at it from discrimination and prediction method point of view.

Finally, it will be interesting to compare the performances of this approach with those of PLS Discriminant Analysis.

References

Abdesselam, R. (2006). Mixed principal component analysis. In M. Nadif & F. X. Jollois (Eds.), *Actes des XIIIémes Rencontres SFC-2006* (pp. 27–31). Metz, France.

Escofier, B., & Pagés, J. (1979). Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *Cahier de l'analyse des données, 4*(2), 137–146.

Fisher, R. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics, VIII*, 376–386.

Geoffrey, J., & McLachlan (2005). *Discriminant analysis and data statistical pattern recognition*. New York: Wiley.

Hand, D. (1981). *Discrimination and classification*. New York: Wiley.

Hubert, M., & Van-Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis, 45*, 301–320.

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

Pagés, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée, LII*(4), 93–111.

Saporta, G. (1977). *Une méthode et un programme d'analyse discriminante sur variables qualitatives*. Journées internationales, Analyse des données et informatique, INRIA.

Saporta, G. (1990). Simultaneous analysis of qualitative and quantitative data. In *Atti XXXV Riunione Scientifica della Societa Italiana di Statistica* (pp. 63–72).

Sjöström, M., Wold, S., & Söderström, B. (1986). PLS discrimination plots. In: E. S. Gelsema & L. N. Kanals (Eds.), *Pattern recognition in practice II*. Amsterdam: Elsevier.

Tenenhaus, M. (1998). *La régression PLS: Théorie et pratique*. Paris: Technip.

Tomassone, R., Danzart, M., Daudin, J. J., & Masson, J. P. (1988). *Discrimination et classement* (172 pp.). Paris: Masson.