Discriminant multiple correspondence analysis

Abdesselam

Abstract The proposed approach leads to analyzing the associations between a set of quantitative variables and several qualitative variables measured on a same set of individuals. In a decision-making context, the proposed method can be considered as a generalization of discriminant analysis to the multiple groups' variables case. It's described as a principal component analysis of the centres of gravity tables. The decomposition of its duality diagram on a double diagram illustrates the link and the passing properties between multiple correspondence analysis and discriminant analysis. An application resulting from real data illustrates the utility of the discrimination model thus defined.

1 Introduction

In this work, we describe an approach which leads to analyzing the correspondences between a set of qualitative variables and several quantitative variables. In a context of prediction and classification, the proposed Discriminant Multiple Correspondence Analysis (DMCA) can be considered as a discriminant analysis on several groups' variables simultaneously. In a discrimination and classification aim, this method of decision-making explains simultaneously more than one qualitative group variable according to a set of quantitative variables. It's a multivariate statistical method derived from classical Discriminant Analysis (DA) used on the results of Multiple Correspondence Analysis (MCA). An example resulting from real data illustrates the results obtained with this method.
This approach leads to extending discrimination analysis to more than one group variable to be discriminated as generalized discriminant analysis proposed in [3]. It can be considered as a particular case of an extention of canonical analysis to more than two groups of variables.

CREM UMR CNRS 6211, Department of Economics and Management, University of Caen,
e-mail: rafik.abdesselam@unicaen.fr

2 Principle of the method

We use the following notations to explain the methodology of the discriminant model approach. Let us denote:
$X_{(n,p)}$ the quantitative data matrix associated to the set of p discriminant centered 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 groups' variables with $q=\sum_{l=1}^{m} q_l$ dummy variables $\{y_l^k;k=1,q_l\}_{\{l=1,m\}}$, that we wish to discriminate,
$Y_{l\,(n,q_l)}$ the dummy variables matrix associated to the q_l modalities of the variable y_l,
$Y_{(n,q)}=[Y_1,\ldots,Y_l,\ldots,Y_m]$ the global matrix, juxtaposition of the matrix $Y_{l\,(n,q_l)}$,
$E_x=R^p$ and $E_y=\oplus\{E_{y_l}\}_{\{l=1,m\}}=R^q$ are the individual subspaces associated by duality respectively to the data matrix $X_{(n,p)}$ and $Y_{(n,q)}$,
$M_x=V_x^+$ is Mahalanobis distance (Moore-Penrose generalized inverse of the covariance matrix V_x) in the explanatory subspace E_x,
$D_n=\frac{1}{n}I_n$ diagonal weights matrix of the n individuals, where I_n is the unit matrix with n order,
$D_q=\mathrm{Diag}(\mathrm{D}_{y_1},\cdots,D_{y_l},\cdots,D_{y_m})$ diagonal matrix of weights' matrix of the q centres of gravity,
$\chi_y^2=\mathrm{D}_q^{-1}=Diag(\chi_{y_1}^2,\cdots,\chi_{y_l}^2,\cdots,\chi_{y_m}^2)$ diagonal matrix of Chi-square distance,
$N_{y_l}=\{y_{l\,i}\,;\,i=1,n\}$ the configuration of individual-points associated to the rows of matrix Y_l.

Figure 1 shows the duality diagram corresponding to DMCA and its decomposition according to that of PCA triplet $(\widetilde{Y}=Y\chi_y^2\,,\,\frac{1}{m}D_q\,,\,D_n)$ corresponding to MCA of the q modalities associated to the m group variables.

$$
\begin{array}{ccccccccc}
E_x=R^p & & \overset{{}^tG_{(p,q)}}{\longleftarrow} & & F_x^\star=R^q & \qquad & E_x=R^p & \overset{{}^tX_{(p,n)}}{\longleftarrow} & F_y^\star=R^n \\
V_x^+\downarrow & & \boldsymbol{DMCA} & & \uparrow \frac{1}{m}D_q & & V_x^+\downarrow & & \uparrow D_n \\
E_x^\star=R^p & & \longrightarrow & & F_x=R^q & & E_x^\star=R^p & \longrightarrow & F_y=R^n \\
 & & G_{(q,p)}={}^t\widetilde{Y}\,D_n\,X & & & & & & X_{(n,q)}
\end{array}
\qquad
\begin{array}{ccc}
 & \overset{{}^t\widetilde{Y}_{(q,n)}}{\longrightarrow} & \\
 & \boldsymbol{MCA} & \\
 & \longleftarrow & \\
 & \widetilde{Y}_{(n,q)}=Y\,\chi_y^2 &
\end{array}
\qquad
\begin{array}{c}
E_y=R^q=F_x \\
\downarrow \frac{1}{m}D_q \\
E_y^\star=R^q=F_x^\star
\end{array}
$$

Fig. 1 DMCA duality diagrams.

The principal moments and principal vectors of MCA are eigenanalysis (normed) of the operator: ${}^t\widetilde{Y}\,D_n\,\widetilde{Y}\,\frac{1}{m}D_q={}^t(Y\chi_y^2)\,D_n\,Y\chi_y^2\,\frac{1}{m}D_q=\chi_y^2\,V_y\,\chi_y^2\,\frac{1}{m}D_q=\frac{1}{m}\chi_y^2\,V_y$, with inertia equal to: $I(N_{\widetilde{y}})=\frac{1}{m}trace(\chi_y^2\,V_y)=\frac{1}{m}\sum_{l=1}^{m}\chi_{y_l}^2\,V_{y_l}=\frac{1}{m}\sum_{l=1}^{m}q_l-1$.
Note that the diagonalized matrix $\chi_y^2\,V_y$ is the centered row profiles of Burt table.

Definition 1. DMCA consists to make the following principal component analysis:

$$PCA(G=\chi_y^2\,V_{yx}\;,\;V_x^+\;,\;\tfrac{1}{m}D_q).$$

Likewise, the principal moments and principal vectors of DMCA are eigenanalysis (normed) of the operator: ${}^tG\frac{1}{m}D_q G V_x^+ = \frac{1}{m}V_{xy}\chi_y^2 V_{yx}V_x^+$,
with explained inertia equal to:
$I_x(N_{\tilde{y}}) = I(N_g) = \frac{1}{m}trace(V_{xy}\chi_y^2 V_{yx}V_x^+) = \frac{1}{m}\sum_{l=1}^{m} V_{xy_l}\chi_{y_l}^2 V_{y_l x}V_x^+ = \frac{1}{m}\sum_{l=1}^{m} I_x(N_{y_l})$,
where, ${}^tG_{(p,q)} = [{}^tG_1,\ldots,{}^tG_l,\ldots,{}^tG_m]$ is the superposition of the m centres of gravity tables associated to $P_x(N_{\tilde{y}})$, with $G_{l(q_l,p)} = \chi_{y_l}^2 V_{y_l x}$, $V_{y_l x} = {}^tY_l D_n X$ the covariance matrix and P_x the orthogonal projection operator in subspace E_x.

Finally, in a practical point of view, DMCA appears as a classical DA, i.e., a PCA of the centres of gravity tables in the explanatory subspace E_x with Mahalanobis distance. The ratio of explained inertia is equal to $\frac{I_x(N_{\tilde{y}})}{I(N_{\tilde{y}})} = \frac{\sum_{l=1}^{m} I_x(N_{y_l})}{\sum_{k=1}^{m} q_k - 1}$.

3 Application example

To illustrate this approach, we use real data published in [8] which concerns characteristics of twenty-seven small cars of Belgium market. This data set contains seven quantitative variables: Price, Consumption, Cubic capacity, Speed, Volume of the boot, the Weight/Power ratio and the Length of the cars, and three qualitative groups' variables: the horsepower (4CV, 5CV, 6CV), the trademark (French, Foreign) and the range (Economic, Traditional, Polyvalent, Turbo) of the cars, with $q = 9$ modalities in total.

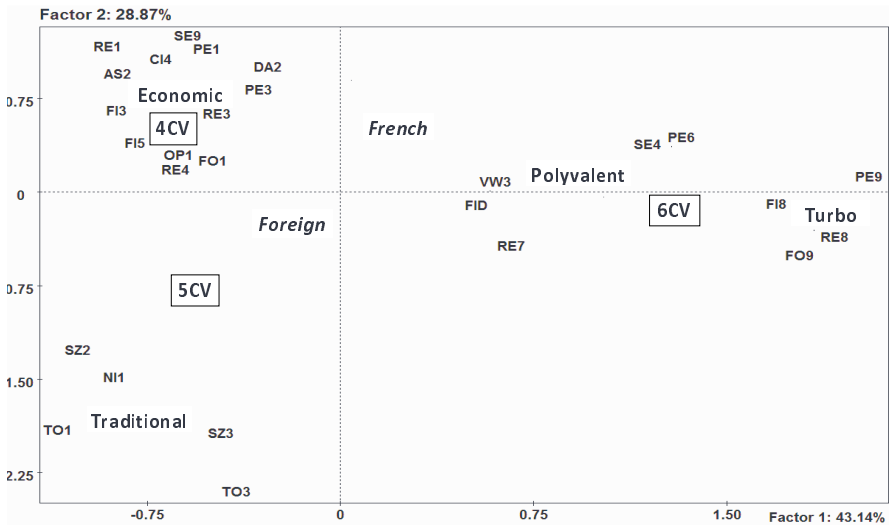

Fig. 2 Simultaneous representation of individuals-cars and centres of gravity.

The aim is to bring to the fore the mixed characteristics which differentiate well and separate the nine car groups constituted by the three qualitative target variables, according to the seven quantitative characteristics of the cars.
Figure 2 summarizes the main graphical result of DMCA, i.e. the good separation between the different groups on the first discriminant plane which explain 72.01% of the total variability, i.e. of the total explained inertia. Concerning the performance of the discrimination rule of DMCA, the ratio of explained inertia and the percentage of well classified, according to the three groups' variables, are respectively equal to 70.78% and 85.18%.

4 Discussion and conclusion

In this work, we present a methodology which extends discriminant analysis to several groups' variables simultaneously as a particular principal component analysis of results of multiple correspondence analysis. The main advantage of this method is its simplicity and facility, it finds interest in the context of the classification and scoring techniques; especially in sensometry, chimiometry, economics and insurance fields. With one group variable, DMCA is a classical DA. With two groups' variables, DMCA is a discriminant analysis on contingency table as correspondence discriminant analysis proposed in [9] and [10]. DMCA can also be used with mixed explanatory variables [1]. Finally, it will be interesting to compare the performance of this approach with that of canonical analysis with more than two groups of variables.

References

1. Abdesselam, R.: Mixed Discriminant Analysis. Sixth Scientific Meeting of the CLAssification and Data Analysis Group, the Italian Statistical Society, Macerata, 551-554 (2007)
2. Chessel, D., Lebreton, J.D., Yoccoz, N.: Propri*étés* de l'analyse canonique des correspondences. Revue de Statistique Appliqu*ée*, **Vol. 35 (4)**, 57-72 (1987)
3. Faraj, A.: Analyse de contiguit*é* : Une analyse discriminante g*én*ralis*ée à* plusieurs variables qualitatives. Revue de Statistique Appliqu*ée*, **XLI(3)**, 73-84 (1993)
4. Geoffrey, J. McLachlan: Discriminant Analysis and Data Statistical Pattern Recognition. Wiley-Interscience (2005)
5. Hand, D.: Discrimination and Classification. Wiley and Sons, New York (1981)
6. Hubert, M. and Van-Driessen, K.: Fast and robust discriminant analysis. Computational Statistics and Data Analysis, **45**, 301-320 (2004)
7. Lachenbruch, P.: Discriminant Analysis. Hafner Press, New York (1975)
8. Lambin, J.: La recherche marketing, Analyser - Mesurer - Pr*é*voir. McGraw-Hill (1990)
9. Perri*è*re, G., Lobry, J.R., Thioulouse, J.: Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. Comput. Appl. Biosci., **12**, 519-524 (1996)
10. Perri*è*re, G. and Thioulouse, J.: Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins. Computer Methods and Programs in Biomedicine. Elsevier, **70**, 99-105 (2003)