MULTIPLE CORRESPONDENCE ANALYSIS AND RELATED METHODS

Edited by Michael Greenacre and Jörg Blasius

Chapman & Hall/CRC, 2006.

This book gathers selected papers of the CARME 2003 conference held at Pompeu Fabra University on the topic of multiple correspondence analysis (MCA), as well as several chapters written specially to make the book self-contained. At present several books exist on correspondence analysis (CA) but none is devoted specifically to MCA, in this sense this book fills a gap in the scientific series of books. Professors Greenacre and Blasius should be thanked for their efforts to publish it in a very accurate and coherent way.

MCA is a more difficult topic to grasp than simple CA, hence it is often applied in a naïve manner, taking the default value of its parameters, without taking advantage of all its possibilities. MCA inherits from its ancestor, simple CA, its roots and its multiple facets. Different fields and hence difference schools have led to CA, ecology, psychometrics, linguistics, ... each one having its own contribution and particular point of view, one focusing on the quantification of categorical variables, another in finding the correlation between categorical variables, other in finding a latent variable among categorical variables, for others it is just a model for categorical data. In my opinion, however, the main strength of MCA is the visual representation of the information as popularized by the French "*Analyse des Données*" school and this has also been the main perspective of this book.

The book is very well written, it is composed of 23 chapters assembled in five parts, the first two being devoted to MCA, whereas the last three are about methods that are related in a general way. This is unavoidable in a book composed of contributions from a conference. In spite of that, the large majority of chapters present the state of the art of the methods in a very clear manner, using good and motivating examples that make enjoyable reading.

From a mathematical point of view MCA is a generalisation to categorical variables of the canonical correlation analysis, and this is the starting point of the book, but it goes too short in this path and moves to the geometric presentation as a particular case of PCA, which is the classic way of presenting CA. In this approach MCA consists of

the PCA of the triplet (**Y**,**M**,**N**), where **Y** represents a convenient matrix of profiles, **M** a chi-square metric and **N** is the diagonal matrix of relative weights. Then, it is easy to produce the desired visualisations and by modifying the metric to enlarge the possibilities of MCA. However, under this approach the introduction of the chi-square metric is rather arbitrary and so are the different possible visualisations.

In addition, it shows some drawbacks concerning the unexpected high importance of rare modalities and the problematic interpretation of the explained variance. All of this makes MCA a little messy and this gives to the book all its value.

CA is not a particular case of PCA indeed. CA deals with the analysis of two way tables and not with the analysis of data matrices, as PCA does, as John Gower points out very cleverly in Chapter 3. This is why CA needs to be presented as a double PCA, one for the row profiles and the other for the column profiles. However, the criticism of the chi-square metric has no justification; the more correct way of presenting CA is like a canonical correlation analysis of two categorical variables. Under this formalism, the chi-square metric emerges naturally as the categorical counterpart of the Mahalonobis metric for continuous data, well accepted to measure distances with multinormal data. Of course, it does not imply that other metrics can be used as in the case of continuous data, leading to Tucker analysis, taking the identity as metric for the row and column spaces, redundancy analysis, SIMPLS, (Tenenhaus, 1998). Nevertheless, the central role of the chi-square metric is clear, like the Mahalanobis one for continuous data.

Concerning the different possibilities for the representation of points, we know the infinity of different biplots available but only two have meaningful value, the row preserving or the column preserving biplot (the asymmetric visualisations), nevertheless the most used visualisation for CA and MCA is the symmetric one, which is not a biplot. In fact all these visualisations stem again from the CCA. In CCA, two representations are equally possible, one in the space of the first variable and the other in the space of the second variable, which correspond to the two asymmetric biplots. These two representations are not unique. It is possible to perform a representation in the direct sum space of the two variables, this representation holds the property (interesting) of maximizing the variance represented, yielding up to the so-called symmetric representation (Saporta, 1990).

All these representations have clear interpretation rules that give the user deep insight to pass from data to knowledge. This is what happens when reading the introduction to MCA with the ISSP (International Social Survey Program) data as an example in Chapters 1 and 2, starting from the simple case of a simple table, then going on to the stacking of tables and finally the analysis of all two-way tables, which is the situation of MCA, showing very clearly the interest and the differences between MCA and the CA of stacked tables.

Another important point for interpretative purposes is the adjustment of the inertia explained for each dimension, proposed by Greenacre in Chapter 2 – its simplicity merits it being routinely computed in the standard packages of MCA. Alternatively JCA (joint correspondence analysis) can be used, which by means of an iterative process annihilates the inertia of the block diagonal tables, like the method of principal factor analysis.

An endless debate is the link between modelling and visualisation. Chapter 3 makes bridges between both approaches, by bilinear models modelling the residuals with respect to some baseline model or by the visual approach using the SVD of the same residuals. Also here it is important to notice the difference between dealing with data matrices (two dimensions and two modes) and two-way tables (two dimensions but one mode) and the formal similarities between continuous and categorical variables, which leads John Gower to say that MCA is more related to PCA than to CA, and this is true. Saporta (1990) already presented MCA as a non-linear generalisation of PCA (PCA of an expansion of the variables space by splines of order 0). Also it is clever to point out that PCA and MCA yield up an approximation of matrix of residuals $\mathbf{Y}$, and not the correlation matrix $\mathbf{R}$ (which is the case for factor analysis). But the difference is very slightly, since both PCA and MCA are used to define latent variables from the observed ones (this is the explicit goal of homogeneity analysis and it is one of the foundations of the PLS path modelling community). The topics raised in Chapter 3 have attracted the attention of many researchers; Escofier (1984) presented the MCA respect to any model with the same margins; within the SPAD community biplots were usual from 1980 even though they were not called biplots. In Chapter 21 models are enlarged by Kroonenberg and Anderson to cover three-way interactions with additive terms and more sophisticated models with multiplicative terms but with lack of interpretability in this case. In Chapter 22 Groenen and Koning present a biplot for the visualisation of the interaction of an ANOVA model and in Chapter 23 Vicente-Villardon, Galindo and Blazquez visualize a logistic response via biplots.

Accepting that MCA is a non-linear analysis of data, it makes sense to constrain the ordering of the categories (for instance the Likert scale) to present the results in a meaningful way for the user. Chapter 4 presents it in a very general way, leading to non-linear PCA (NLPCA), also known as categorical PCA (CatPCA). In fact it is arguable whether a MCA would be preferable to CatPCA, as Nishisato says in Chapter 6, it depends on the application but in general it would be better to reveal all the non-linearities present in the data rather than to filter them. In fact CatPCA is a good alternative to PCA when analyzing ordinal data. Very often ordinal data is analyzed by PCA as if the variables were continuous, assuming equally sequenced scales for them. Then CatPCA and its comparison with standard PCA allow assessing whether the respondents have understood the questions and what subjective scales they have applied, as is shown in Chapter 20 by Blasius and Thiessen.

In fact, MCA is one of the most successful statistical techniques to analyze survey data, because MCA fits well the requisites for survey data: it can be applied to large data sets, very often collected with mild probabilistic assumptions, it reveals the salient parts of the information by looking at the multivariate distribution defined from a homogeneous group of variables (respect to the concept they measure, for instance opinions about a group of questions) which is called active, extracting their common denominators (the significant axes) and relating them to the external (supplementary) information, which very often forms a set of structuring factors (sex, age, level of studies, region, ...) leading to what Alain Morineau has called the "themascope" approach (Aluja and Morineau, 1999). This is the heart of the French "*Analyse des Données*" which is explained by Henry Rouanet in Chapter 5. This approach can be enhanced by incorporating inferential aspects like the confidence ellipses around the centroids of the categories of the structuring factors to assess its significance.

Ludovic Lebart goes deeply into this argument in Chapter 7, introducing validation tools of the revealed patterns. MCA gains all its value when applied to large data sets, where it is possible to rely on the central limit theorem and to compute t-test values for all centroids in every significant dimension. This allows linking the revealed pattern with the background information of individuals in a substantive way. In addition we can visualize the uncertainty inherent in the position of these centroids to compare pairs of them, by bootstrapping, to obtain confidence ellipses or convex hulls for supplementary category points. Another possibility for stabilizing the results is by regularized MCA (presented by Takane in Chapter 11), by means of a change of the metric of the column space in a similar way as in ridge regression; this leads to more stable estimates of the factorial axes and hence more reliable confidence ellipses, at a price of the greater complexity of the method.

Another problem when analyzing survey data is that of missing values, Figure 8.1 presents a typical display of its effect. Although it is arguable whether or not to eliminate these "don't know" points in the displays, since they do reveal some information about the individuals who had chosen this option, especially if we have socio-demographic supplementary points in the display, it is nevertheless interesting to focus the attention only on certain categories, for instance the expressed opinions or the most extreme ones, to get insight into the understanding of data. In Chapter 8 Greenacre and Pardo propose subset MCA as a clever and useful solution to this problem; it consists of performing MCA in a selection of categories with the same global margins to not lose the additive properties of each subset MCA. This topic was also addressed by Bénali and Escofier (1987), also it has been incorporated in the procedure COREM of SPAD. This problem of missing data is also treated in Chapter 12 by Matschinger and Angermeyer as a particular case of canonical correlation analysis but the rationale is more complicated than that of subset MCA.

We have stated that MCA is a correlational technique, however, what is the correlation between categorical variables is not an easy question to answer. The practice of just taking the correlation with the first solution of the MCA is obviously a partial solution. Nishishato in Chapter 6 proposes to measure the correlation between two variables in the space spanned by one of them, leading once again to the analysis of stacked tables. This measure turns out to be related to Cramer's V coefficient.

Another usage of MCA is to measure a latent variable, with maximal (squared) correlation with every question. Item response theory (IRT) serves to define scales measuring a latent unidimensional trait from a set of items. There are two main models for IRT, the dominance model, which implies a monotonic utility of the latent variable and the proximity model, implying a unimodal pattern, where the utility depends on the distance of one individual respect to the mode point. The Guttman model and the Rasch model are cases of the former type, whereas the unfolding model is an example of the latter. Chapters 9 and 10 (by Warrens and Heiser and van Schuur and Blasius respectively) deal with how MCA can be helpful to ascertain which type of data, dominance or proximity, we have at hand. The arch effect and the horseshoe are basic displays for these kinds of data. Here it is worth pointing out what van Schuur and Blasius state about the contradiction regarding the need of having simple questions in questionnaires and the difficulty of asking questions necessary to obtain unfolding data.

Sometimes we want to analyze a concatenation of tables coming from different studies, (like different countries, years, ...). This analysis will include an inertia effect between tables due to the different centroids of each table. Also it is required to give the same importance to all tables. Bécue and Escofier in Chapter 13 and Amaya and Goitisolo in Chapter 14 present similar methodologies to solve this problem (MFACT and SA respectively). First, each table is centred with respect to its own centroid to eliminate the between inertia (intra-analysis) and second the influence of each table is equalized respect its corresponding first dimension, that is, we operate a change of the metric in the space of columns (but other methods for balancing are possible). The two approaches differ in how they define the weight for the rows, which coming from different surveys, will be different in general and hence there exists no natural weight for each row. These kinds of approaches lean on the geometric facilities of the PCA. In Chapter 15 Abascal, Lautre and Landaluce propose a variant of the previous MFA method to treat mixed tables formed by categorical variables and continuous variables together. The obtained results are more difficult to interpret but this can be an alternative to CatPCA when we know *a priori* which variables have a non-linear effect.

The discrimination ability is another facet of CA. Very often we have to predict a categorical variable from a set of categorical predictors. CA provides a simple solution consisting of stacking the tables crossing the response variable with all the predictors (or interactions between the predictors). It stems from the method's very origins (Fisher, 1940) and other variants have been proposed such as PLS discrimination (Tenenhaus,

1998) or nonsymmetric CA (Lauro and d'Ambra, 1984). In Chapter 16 Saporta and Niang present the Disqual methodology, consisting of a MCA of the predictors followed by a linear discriminant analysis using the significant factors – in this sense Disqual acts like a principal component regression on categorical data. An important result is the reduction of the Vapnik-Cervonenkis dimension due to the reduction of dimensionality entailed by the use of MCA. Conjoint analysis can enter within the same framework of CA of the response variable (expressing a preference) crossed by the set of attributes of products, because very often conjoint analysis uses least-squares regression instead of monotonic regression, as shown by Torres-Lacomba in Chapter 19.

In Chapter 17, Bougeard, Hanafi, Noçairi and Qannari propose a unified method to obtain factors balancing the discrimination power of CA of the stacked table crossing the response variable and the predictors and the self explanation of factors issued from the MCA of the predictors. Then, successive solutions are obtained by deflation, in a similar way to PLS discriminant analysis.

Changing the metric of rows it is possible to highlight other features rather than the dispersion of individuals, in particular if we consider a metric downweighting the high distances to the global centroid, it is possible then to perform a robust MCA, to detect outliers, or choosing a metric to download the high inter-pair distances among individuals, it is possible to obtain a very flexible technique to reveal the natural clusters of the individuals in the display (in Chapter 18 by Caussinus and Ruiz-Gazen). A similar technique just retaining the distances lower to a given threshold has been proposed by Lebart (1995).

All these chapters emphasize the great versatility of MCA and all its possibilities; it can serve to analyze real problems in a large variety of different fields, social surveys, psychometry, marketing, ... and also to further develop the methodology to produce research papers, as shown in Figure 1.1. However, I would stress the need for having clear interpretation rules for a method to be useful. There is no a intelligent method, what is intelligent is the use of the method. This book serves the purpose of making the method useful.

Tomàs Aluja-Banet

Universitat Politècnica de Catalunya