A Weighted Score Derived from a Multiple Correspondence Analysis Solution

de Souza, Márcio L. M.
Universidade Federal de Juiz de Fora, Departamento de Estatística
Rua José Lourenço Kelmer, s/n - Campus Universitário, Bairro São Pedro
36036-900 - Juiz de Fora - MG, Brazil
E-mail: souza.mlm@gmail.com

Bastos, Ronaldo R.
Universidade Federal de Juiz de Fora, Departamento de Estatística
Rua José Lourenço Kelmer, s/n - Campus Universitário, Bairro São Pedro
36036-900 - Juiz de Fora - MG, Brazil
E-mail: ronaldo.bastos@ufjf.edu.br

Vieira, Marcel de T.
Universidade Federal de Juiz de Fora, Departamento de Estatística
Rua José Lourenço Kelmer, s/n - Campus Universitário, Bairro São Pedro
36036-900 - Juiz de Fora - MG, Brazil
E-mail: marcel.vieira@ufjf.edu.br

Multivariate data analysis is performed in the presence of three or more variables and considers each variable as a dimension of a space R^J, J being the variables considered. This statistical technique has had its greater momentum in recent decades with the development and improvement of computer technology and one of its main purposes is the reduction of dimensionality. We propose a methodology for the calculation of weighted scores from a set of categorical data based on the mathematical properties of the Multiple Correspondence Analysis (MCA) paradigm.

Survey response data to address attitudes, satisfaction and other latent variables of interest to social scientists often rely on a set of Likert-type statements for which respondents choose one category among all possible ordinal categorical answers. For each respondent, scores are usually calculated as the summation of individual values obtained from each response (eg. Berrington, 2002). However, such scores are represented by integer values only and assume equal distances between each ordered category. Furthermore, summation scores may be less accurate in representing latent traits, as different profiles may result in identical score values among all possible response patterns, some summations over different patterns will accrue the same overall value. The scoring method proposed in this paper also tries to minimize this problem , as only identical profiles, that is, individuals with the same responses to all variables, will have repeated weighted scores as proposed.

The methodology proposed for the calculation of the weighted scores is included in the Section 'Weighted Score'. An illustrative example is presented in the Section 'Illustrative Example', while in the 'Simulation Study' some basic information of a simulation study that is currently being undertaken are presented. We make brief concluding remarks in the final Section.

Weighted Score

The main purpose of this work is to propose scores from the MCA solution considering the weighted average of the numerical values of the categories to which the respondent belongs to, given all the variables.

MCA allows the consideration of a finite number of factors (or dimensions), which has a maximum of $d = L - J$, where $L = \sum_{j=1}^{J} C_j$, C_j is the number of categories of the j^{th} variable, and J is the number of variables that are being considered in the analysis.

Let X be a $n \times d$ matrix that includes the object initial scores for each of the n individuals observed in the data set. Let Y_j, $j = 1, \ldots, J$, be also object score matrices that include the score of the j^{th} variable categories to which the objects (individuals) belong to. Let W_j, $j = 1, \ldots, J$, be $n \times n$ matrices given by

$$W_j = (X - Y_j)(X - Y_j)'.$$

Let P_1 be a $n \times J$ matrix whose j^{th} column is given by the square root of the main diagonal elements of W_j. Moreover, let M_j, $j = 1, \ldots, J$, be $n \times n$ matrices given by

$$M_j = (Y_j)(Y_j)'.$$

Let P_2 be a $n \times J$ matrix whose j^{th} column is given by the square root of the main diagonal elements of M_j. Let P be also a $n \times J$ matrix whose elements are defined such that $P_{ij} = P_{1ij} \bullet P_{2ij}$. Furthermore, let D_1 be a $n \times 1$ matrix given by the product $P \bullet U$, where U is a $J \times 1$ vector of ones. Let D be a $n \times n$ diagonal matrix whose diagonal elements are given by $1/D_1$.

Let K be the $n \times J$ profile matrix which includes numerical values of the categories to which each object (individual) belongs to for all the J variables. Finally, we propose the weighted scores to be calculated as the main diagonal of the matrix S_1 which is given by $S_1 = P \times K' \times D$. Therefore, let S be a score vector given by the main diagonal of S_1.

Equivalently, the weighted scores may be calculated using the following expression:

$$S_i = \sum_{j=1}^{J} K_{ij}.p1_{ij}.p2_{ij} \Big/ \sum_{j=1}^{J} p1_{ij}.p2_{ij}$$

Where:

K_{ij} represents the numeric value assigned to the category of variable j which the subject i belongs to;

$p1_{ij}$ and $p2_{ij}$ are the weights applied to the weighted average, and the elements of matrices P_1 and P_2 respectively, given by:

$$p1_{ij} = [(X_i \quad Y_{ij})^t (X_i \quad Y_{ij})]^{1/2}$$
$$p2_{ij} = [Y_{ij}{}^t Y_{ij}]^{1/2}$$

Note that both weights shown above are representations of the inverse matrix of the following Euclidean distances: (i) between the i^{th} individual and the category of the variable which such individual belongs to ($p1_{ij}$), and (ii) between the category of the variable which the i^{th} individual belongs to and the origin of the n-dimensional graph ($p2_{ij}$). In the above expression, X_i represents the position in n-dimensional graphical solution provided by MCA for the i^{th} individual and Y_{ij} represents the position of the j^{th} variable category to which the i^{th} individual belongs to.

Illustrative Example

We illustrate the methodology introduced in the previous section through an example: satisfaction with respect to Salary (SL) and satisfaction with respect to Number of Hours Worked (CH) for $n = 5$ employees of a fictitious company, whose profiles are presented in Table 1 below. Such variables have three categories represented by the numbers 1, 2 and 3, which mean, respectively, DISSATISFACTION, INDIFFERENCE and SATISFACTION.

Table 1. Profiles for 5 employees of a fictitious company

Employee	SL	CH	Summation Score
1	3	2	5
2	1	2	3
3	1	3	4
4	3	1	4
5	2	3	5

Note that, for our example, $d = 4$. MCA was applied to those fictitious data from Table 1 and a correspondence analysis map was generated. Figure 1 below presents results for the first two dimensions of the MCA solution.

Figure 1. Multiple Correspondence Analysis Solution for 5 employees of a fictitious company data set

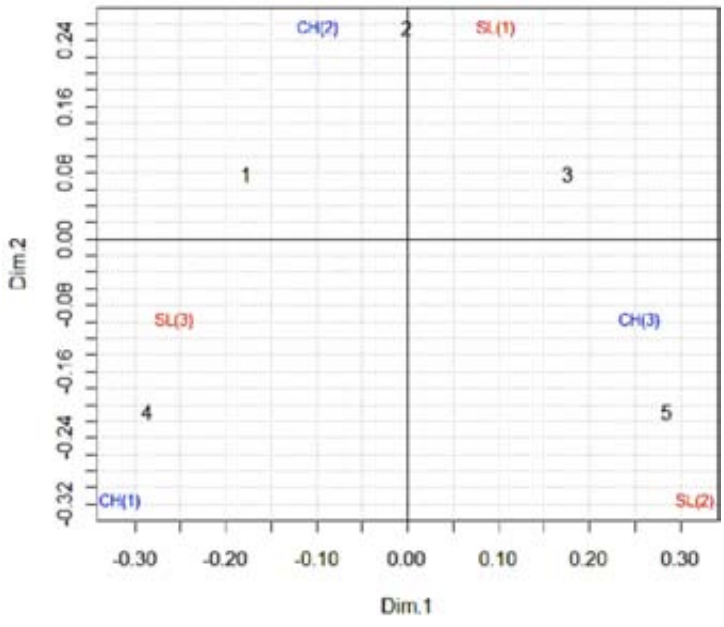

Table 2 below presents the object scores (or object/individual coordinates) for each of the four dimensi ons of the MCA solution.

Table 2. Object scores for each of the four dimensions of the MCA solution

Employee	Dimension 1	Dimension 2	Dimension 3	Dimension 4
1	-0.1768	0.0791	-0.1768	-0.0791
2	0.0000	0.2558	0.0000	0.0977
3	0.1768	0.0791	0.1768	-0.0791
4	-0.2860	-0.2070	0.1093	0.0302
5	0.2860	-0.2070	-0.1093	0.0302

Table 3 below presents the category scores (or category coordinates) for each of the variables considered in our analysis and for each of the four dimensions of the MCA solution.

Table 3. Category scores for each of the four dimensions of the MCA solution

Variable	Category	Dimension 1	Dimension 2	Dimension 3	Dimension 4
SL	**SL(1)**	0.0977	0.2558	0.2558	0.0977
	SL(2)	0.3162	-0.3162	-0.3162	0.3162
	SL(3)	-0.2558	-0.0977	-0.0977	-0.2558
CH	**CH(1)**	-0.3162	-0.3162	0.3162	0.3162
	CH(2)	-0.0977	0.2558	0.2558	0.0977
	CH(3)	0.2558	-0.0977	-0.0977	-0.2558

For our example, X includes the object scores that are presented in Table 2. Considering Table 1 and Table 3 from our example, Y_1 and Y_2 are given by:

$$Y_1 = \begin{bmatrix} 0.2558 & 0.0977 & 0.0977 & 0.2558 \\ 0.0977 & 0.2558 & 0.2558 & 0.0977 \\ 0.0977 & 0.2558 & 0.2558 & 0.0977 \\ 0.2558 & 0.0977 & 0.0977 & 0.2558 \\ 0.3162 & 0.3162 & 0.3162 & 0.3162 \end{bmatrix} \quad \text{and} \quad Y_2 = \begin{bmatrix} 0.0977 & 0.2558 & 0.2558 & 0.0977 \\ 0.0977 & 0.2558 & 0.2558 & 0.0977 \\ 0.2558 & 0.0977 & 0.0977 & 0.2558 \\ 0.3162 & 0.3162 & 0.3162 & 0.3162 \\ 0.2558 & 0.0977 & 0.0977 & 0.2558 \end{bmatrix}.$$

Following the methodology presented in the previous section, matrices P_1, P_2 and P were calculated and the results are as follows,

$$P_1 = \begin{bmatrix} 0.2738228 & 0.2738228 \\ 0.2738228 & 0.2738228 \\ 0.2738228 & 0.2738228 \\ 0.3708147 & 0.3707294 \\ 0.3707294 & 0.3708147 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.3872439 & 0.3872439 \\ 0.3872439 & 0.3872439 \\ 0.3872439 & 0.3872439 \\ 0.3872439 & 0.6324000 \\ 0.6324000 & 0.3872439 \end{bmatrix}, \text{ and } P = \begin{bmatrix} 0.1060362 & 0.1060362 \\ 0.1060362 & 0.1060362 \\ 0.1060362 & 0.1060362 \\ 0.1435957 & 0.2344493 \\ 0.2344493 & 0.1435957 \end{bmatrix}.$$

Moreover, matrices D_1, D, and S_1 were also calculated and the results are presented below:

$$D_1 = \begin{bmatrix} 0.2120725 \\ 0.2120724 \\ 0.2120725 \\ 0.3780450 \\ 0.3780450 \end{bmatrix}, \quad D = \begin{bmatrix} 4.71537 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 4.71537 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 4.71537 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 2.645188 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 2.645188 \end{bmatrix},$$

$$S_1 = \begin{bmatrix} 2.500000 & 1.500000 & 2.000000 & 1.121943 & 1.402429 \\ 2.500000 & 1.500000 & 2.000000 & 1.121943 & 1.402429 \\ 2.500000 & 1.500000 & 2.000000 & 1.121943 & 1.402429 \\ 4.242351 & 2.888137 & 3.993652 & 1.759675 & 2.620162 \\ 4.670759 & 2.459729 & 3.136836 & 2.240325 & 2.379838 \end{bmatrix}.$$

The weighted scores obtained for each of the five employees of our fictitious data are included in the following vector S,

$$S = \begin{bmatrix} 2.500000 & 1.500000 & 2.000000 & 1.759675 & 2.379838 \end{bmatrix}.$$

As can be seen, the proposed methodology for score calculation outputs unique values for each individual, something that the mere summing of original scores would not, wrongly considering the pair of individuals 1 and 5 and 3 and 4 as having the same overall score (see Table 1 summation score results).

Simulation Study

In order to evaluate the stability of the results we are currently undertaking simulation-based analyses with real attitudinal data from the British Household Panel Survey (see Taylor *et al.*, 2001) and also with data generated from different population scenarios. In our simulation study, the methodology is being implemented in the R program (R Development Core Team, 2010) considering he MCA solution obtained through the *ca* package, with the *mjca* function for a Burt matrix (Nenadic and Greenacre, 2007). For simulation purposes, we have considered nine variables, each of those with five categories. We have considered a sample size of 2000 individuals and data was generated from asymmetric product multinomial distributions. We have generated 1000 replicate samples.

Concluding Remarks

The work presented here proposes a new methodology for the construction of weighted scores, based on joint analysis of categorical variables, obtained by applying MCA. This work includes the first results for the proposed score, which, to our view has the potential of better representing the latent variable than the simple summation of values of categorical variables over all responses. As future work, we shall use the proposed methodology in order to define dependent variables of models that explain attitudes, satisfaction, and other multidimensional variables for both cross-sectional and longitudinal data, since the average profile of scores can change over time.

Acknowledgement: The authors thank FAPEMIG for grant number CEX-APQ-00467-2008 and for supporting the attendance at the ISI2011 Session.

REFERENCES

Berrington, A. (2002) Exploring Relationships Between Entry Into Parenthood and Gender Role Attitudes: Evidence from the British Household Panel Study. In Lesthaeghe, R. ed. Meaning and Choice: Value Orientations and Life Course Decisions. Brussels, NIDI.

Nenadic, O. and Greenacre, M. (2007). Correspondence Analysis in R, with Two-and-Three-dimensional graphics: The ca Package. *Journal of Statistics Software*, vol. 20, issue 3. http://www.jstatsoft.org/

R DEVELOPMENT CORE TEAM, 2010, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL.

Taylor, M. F. (ed), Brice, J., Buck, N. and Prentice-Lane, E. (2001) *British Household Panel Survey - User Manual - Vol. A: Introduction, Technical Report and Appendices*. Colchester, U. of Essex.

ABSTRACT

Multivariate data analysis considers each variable as a dimension of a space R^J of J variables and has had its greater momentum in recent decades with the development and improvement of computer technology. Such techniques have as one of their main purposes the reduction of dimensionality. In this paper, we propose a methodology for the calculation of weighted scores from a set of ordinal categorical variables based on the mathematical properties of the Multiple Correspondence Analysis (MCA) technique. Our results suggest that the proposed weighted score has the potential of better representing latent variables than the simple summation of values of categorical variables over all responses.