Focus Article

Correspondence analysis

Michael J. Greenacre*

Correspondence analysis (CA) is a method of data visualization that is applicable to cross-tabular data such as counts, compositions, or any ratio-scale data where relative values are of interest. All the data should be on the same scale and the row and column margins of the table must make sense as weighting factors because the analysis gives varying importance to the respective rows and columns according to these margins. This method is one of a large class of methods based on the singular value decomposition and can be considered as the equivalent of principal component analysis for categorical and ratio-scale data or as a pair of classical scalings of the rows and columns based on their interpoint χ^2 distances, using the margins as weights. For categorical data, this method generalizes to multiple CA, a popular method for analyzing questionnaire data. A linearly constrained form of CA, canonical CA, is extensively used in ecological research where species abundance data at various sampling points are visualized subject to being linearly related to environmental variables measured at the same locations. When certain parameters are introduced into its definition, CA has been shown to have limiting cases of unweighted and weighted log-ratio analysis (the latter also known as the spectral map), as well as classical multidimensional scaling. *WIREs Comp Stat* 2010 2 613–619 DOI: 10.1002/wics.114

Keywords: multivariate analysis; visualization; categorical data; singular value decomposition/svd; principal component analysis

INTRODUCTION

Correspondence analysis (CA) has a long and interesting history of being defined, rediscovered, and redefined over many decades. Statisticians who have contributed to its origins in the first half of the 20th century have been H.O. Hartley (writing under his original name of Hirschfeld), R.A. Fisher, and Louis Guttman, among others. In the second half of the century, this method sprung up fairly independently in Japan, France, and Holland, respectively, guided by Chikio Hayashi (who saw it as a method of categorical data scaling), Jean-Paul Benzécri (who saw it as a method of data visualization), and Jan de Leeuw (who saw it as integrating categorical data into classical interval-level multivariate analysis). The method's use for multidimensional graphical display has proved to be very popular in research areas where large (and sometimes sparse) sets of categorical data are collected, in particular linguistics, the social sciences, ecology, archeology, marketing research, and genomics.

Technically, CA falls into the class of classical multivariate statistical methods of dimension reduction based on the singular value decomposition (SVD). To be suitable for CA, all data need to be on the same scale: examples are tables of counts, or ratio-scale data where all values are in dollars, say, or compositional data (sets of proportions or percentages with constant sums), or zero/one data. This method leads to a graphical display, which we call a *map* because of its spatial distance properties, where rows and columns are depicted as points. These points represent *profile vectors*, that is vectors of relative values in rows or in columns, expressed relative to their margins, while the margins themselves are used as weighting factors, called *masses*, giving varying importance to the respective row and column points. Distances between profile vectors are defined as χ^2 distances—these are weighted Euclidean distances based on the assumption that variance in each row or column is approximately proportional to the mean. Finally, the quality of display of the data matrix is measured as in principal component analysis (PCA), in terms of a percentage of explained variance.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: michael@upf.es

Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

DOI: 10.1002/wics.114

COMPUTATIONAL ASPECTS

Of the many equivalent ways to define CA, we choose the weighted PCA definition, which is close to Benzécri's approach—in the process the various concepts inherent to CA will be defined. In general, a set of m-dimensional points is denoted by $\mathbf{x}_i$, where $i = 1, 2, \ldots, n$, with weights w_i assigned to the ith point, and with metric between the points defined by the diagonal matrix $\mathbf{D}_q$ with positive values $q_1, \ldots, q_m$, on the diagonal. These points can be projected orthogonally (in the metric $\mathbf{D}_q$) onto a best-fitting low-dimensional subspace, where fit is measured by the weighted sum-of-squared distances between the points $\mathbf{x}_i$ and their projections $\hat{\mathbf{x}}_i$, $\sum_i w_i(\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}}\,\mathbf{D}_q\,(\mathbf{x}_i - \hat{\mathbf{x}}_i)$, as follows:

1. Collect the points as the rows of the $n \times m$ matrix $\mathbf{X}$, and the weights in the $n \times 1$ vector $\mathbf{w}$ and in the diagonal of the diagonal matrix $\mathbf{D}_{\mathrm{w}}$. The weights are positive and sum to 1: $\mathbf{1}^{\mathrm{T}}\mathbf{w} = 1$.
2. Center $\mathbf{X}$ with respect to its weighted column averages: $\mathbf{Y} = (\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathrm{T}})\mathbf{X}$ (it can be shown as a side result that the optimal subspace necessarily contains the weighted average, or *centroid*, of the points, so we center the points at the centroid from the start).
3. Perform a weighted SVD on $\mathbf{Y}$ by multiplying its rows by the square roots of the weights and its columns by the square roots of the elements of the metric, then calculate the usual (unweighted) SVD and transform back to the solution as follows:

$$\mathbf{D}_w^{1/2}\mathbf{Y}\mathbf{D}_q^{1/2} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathrm{T}} \qquad \mathbf{F} = \mathbf{D}_w^{1/2}\mathbf{U}\mathbf{D}_\alpha \quad (1)$$

The matrix $\mathbf{F}$ contains the *principal coordinates* of the points, i.e., their positions with respect to principal axes. For a map in two dimensions, for example, the first two columns of $\mathbf{F}$ provide the coordinates of the n points projected onto the planar subspace; the quality of display would be the sum of squares of the first two singular values relative to their total sum of squares: $(\alpha_1^2 + \alpha_2^2)/\sum_k \alpha_k^2$.

CA uses the above algorithm twice, on the rows and the columns, and both problems lead to the same SVD problem (this is unlike PCA, which gives different solutions for a matrix and for its transpose). Suppose the table of data, with all values on the same scale, is denoted by $\mathbf{N}$—the primary context is a two-way table of frequency counts in a contingency table, which we assume is the case henceforth, for reasons of terminology. Divide $\mathbf{N}$ by its grand total to get the correspondence matrix $\mathbf{P}$: $\mathbf{P} = (1/\sum_i \sum_j n_{ij})\mathbf{N}$, and define the row and column *masses* as the margins of $\mathbf{P}$: $\mathbf{r} = \mathbf{P}\mathbf{1}$, $\mathbf{c} = \mathbf{P}^{\mathrm{T}}\mathbf{1}$, also collected in the diagonals of the diagonal matrices $\mathbf{D}_r$ and $\mathbf{D}_c$. (Hence, we can think of $\mathbf{P}$ as an observed bivariate density and $\mathbf{r}$ and $\mathbf{c}$ as the marginal densities.) The *row profiles* of $\mathbf{N}$ (conditional densities) are its rows divided by their marginal totals, which are identical to the rows of $\mathbf{P}$ divided by their respective margins in $\mathbf{r}$: $\mathbf{D}_r^{-1}\mathbf{P}$; similarly, the *column profiles* (written as rows) are $\mathbf{D}_c^{-1}\mathbf{P}^{\mathrm{T}}$. The respective χ^2 *metrics* in the space of the row and column profiles are defined by the diagonal matrices $\mathbf{D}_c^{-1}$ and $\mathbf{D}_r^{-1}$.

Applying the above three-step algorithm, leading to solution (1), to the dimension reduction of the row profiles, on the one hand, and to that of the column profiles, on the other, leads to the same SVD problem, namely the SVD of the following matrix:

CA solution:

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}})\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathrm{T}} \quad (2)$$

$$\text{Principal row coordinates: } \mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha \quad (3)$$

$$\text{Principal column coordinates: } \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha. \quad (4)$$

Furthermore, if in either the row or the column problem we wish to know the positions of the unit vectors in the projected space, these are just the principal coordinates without the scaling of the dimensions by the singular values, called *standard coordinates*:

$$\text{Standard row coordinates: } \boldsymbol{\Phi} = \mathbf{D}_r^{-1/2}\mathbf{U} \quad (5)$$

$$\text{Standard column coordinates: } \boldsymbol{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}. \quad (6)$$

Once again, for an optimal low-dimensional display, for example, two-dimensional, use the first two columns of the selected coordinate matrices. The joint display of rows in columns, one in principal coordinates and the other in standard coordinates, called an *asymmetric map*, is a true *biplot*. For example, the combination of Eqs (3) and (6), called a *row principal map*, would depict the row profiles in their optimal positions, while the columns would be the projections of dummy row profiles each of which has a 1 in a particular column and 0s otherwise. The joint display of rows and columns in principal coordinates, called a *symmetric map*, is not strictly a biplot but optimizes the display of the interpoint distances between row profiles and those between column profiles. The total variance in the table is measured by the sum of squares of the matrix in Eq. (2), which is equal to χ^2/n, where χ^2 is the classical χ^2 statistic for measuring deviation from the independence hypothesis in a contingency table, and

n is the grand total of $\mathbf{N}$. This measure of variation in the table $\mathbf{N}$, called the *total inertia*, is decomposed along the principal dimensions of the solution as the sum-of-squared singular values $\sum_k \alpha_k^2$.

AN EXAMPLE

As a small illustrative example, consider Table 1, which was published in *El País* newspaper in Spain on May 28, 2009 [p. 35; source given as the Organization for Economic Co-operation and Development (OECD)]. Notice that this table is already in the form of a row profile matrix, although the published percentages do not always add up to exactly 100% due to rounding errors. The symmetric CA map in two dimensions is shown in Figure 1. The total inertia of the table is equal to 0.1285, of which 88.0% is explained in this two-dimensional map. The point OECD is treated as a *supplementary point*; that is, the solution is calculated on the 18 countries, and then the OECD profile is projected onto the display afterward. A three-dimensional view of the solution can be seen in the accompanying multimedia file (Multimedia 1), which shows a moving rotation of the points around the vertical second axis—Figure 2 shows a particular view of this three-dimensional map. The third axis accounts for a further 8.7% of the inertia, so that the three-dimensional display explains 96.7% and is an almost perfect representation of the data. The third dimension shows the distinction between sport and cultural activities, which was not apparent in the two-dimensional view, with Norway leaning more toward culture and Italy more toward sport.

TABLE 1 | Distribution of Leisure Time, as Percentages, in 18 Countries, Along With the Organization for Economic Co-operation and Development (OECD) Average (A Supplementary Point) and the 18-Country Average (These Last Average Values Are Used to Weight the Column Profiles As Well As—In Their Inverses—To Define Distances Between the Rows)

	TV/Radio	Sport	Friends	Cultural	Other
Australia	41	6	3	2	47
Belgium	36	5	8	8	42
Canada	34	8	21	2	34
Finland	37	8	7	8	40
Germany	28	7	4	15	46
Italy	28	8	6	10	48
Japan	47	6	4	0	42
South Korea	35	7	16	1	41
Mexico	48	5	10	4	33
New Zealand	25	5	24	2	45
Norway	31	8	14	15	33
Poland	41	6	6	8	38
Spain	31	12	4	12	41
Sweden	31	8	7	11	42
Turkey	40	2	34	0	25
UK	41	4	7	10	39
USA	44	5	16	2	32
OECD	36	7	11	4	40
Average	36	6	11	6	39

All figures have been rounded to the nearest percentage point.

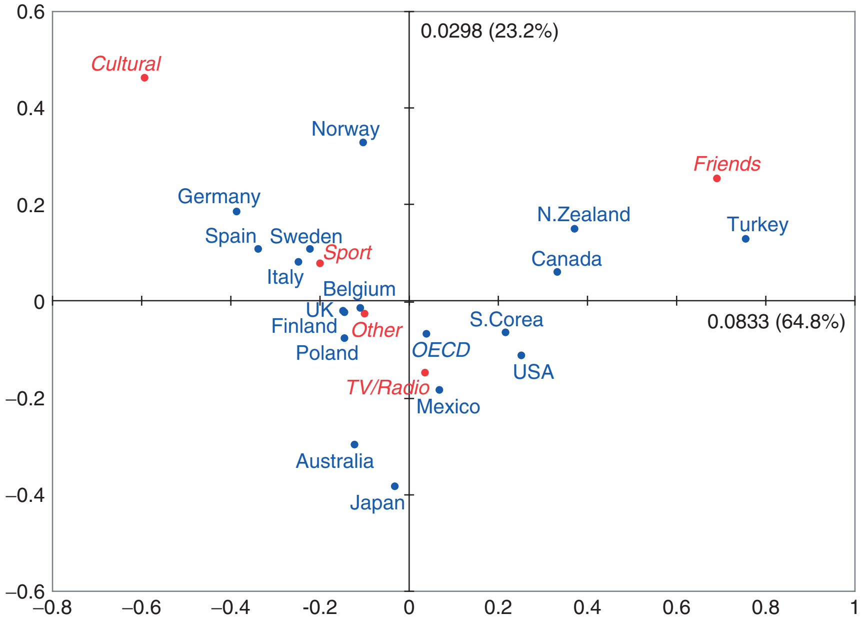

FIGURE 1 | Symmetric two-dimensional correspondence analysis (CA) map of Table 1. The percentage of inertia explained is 88.0%.

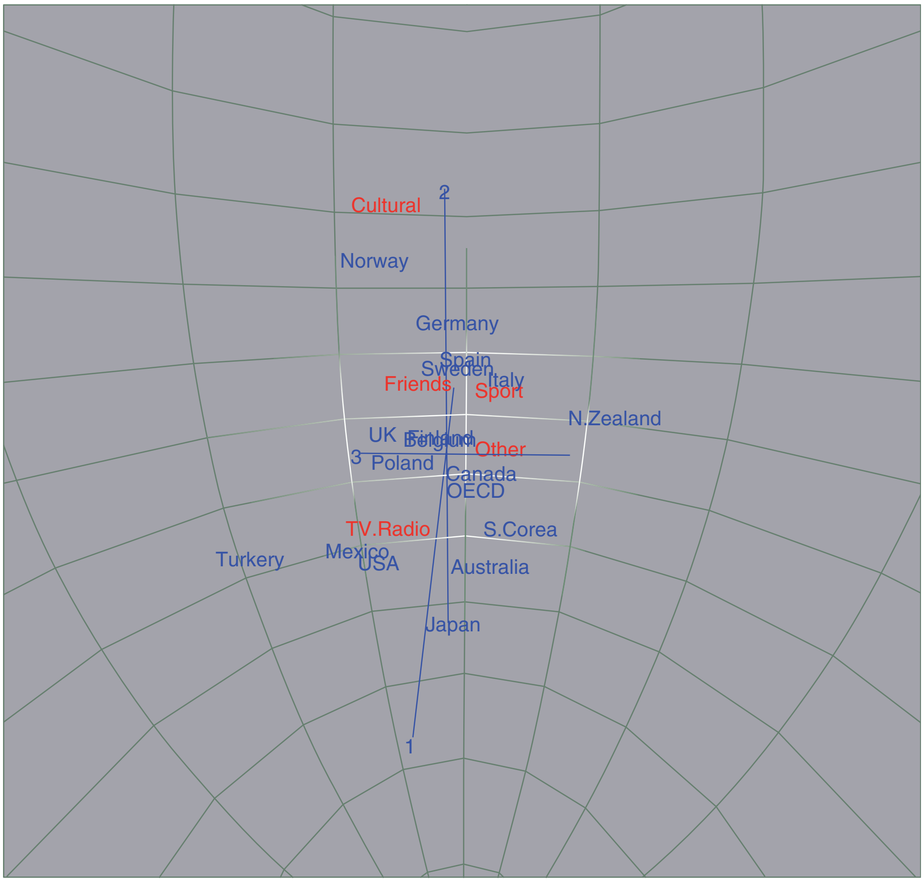

FIGURE 2 | Three-dimensional view of the correspondence analysis (CA) of the leisure data of Table 1, explaining 98.7% of the inertia (the accompanying multimedia file, Multimedia 1, shows the display rotating around the second axis).

MULTIPLE CORRESPONDENCE ANALYSIS

Multiple correspondence analysis (MCA) is the generalization of CA to several categorical variables, most commonly in the context of questionnaire data. Suppose there are Q questions (variables) and the qth question has J_q possible responses. The original data from N cases are sets of Q responses. This $N \times Q$ matrix is then converted into an $N \times J$ *indicator matrix*, where the columns are zero/one dummy variables for each of the J response categories: $J = \sum_q J_q$. MCA is simply the application of the CA algorithm to this large indicator matrix. As the treatment of the response categories is multiple nominal, it is a simple matter to include additional categories for nonsubstantive responses such as 'don't know' or missing values. The MCA solution has *optimal scaling* properties: the category coordinates on the first dimension maximize the variance of the case scores, subject to the inherent quadratic constraint, and also maximize Cronbach's alpha reliability measure between the questions (as quantified by the category coordinates) and the case scores (specifically, the average squared correlation between the questions and the case scores is maximized).

The display is usually an asymmetric map, for example, the row principal map using choices (3) and (6), where each case is represented at the average of his or her particular set of response categories. For example, we analyzed the responses of 10,112 people from six countries—Great Britain (GB), USA (US), Norway (NO), Japan (JA), Spain (SP), and France (FR)—to five statements about marriage (data from Ref 1; Figure 3). The MCA solution in Figure 3 shows a typical result for this type of data where the missing categories are strongly associated and define the strongest dimension in the data. The liberal-to-conservative attitude scale is seen on the left, and there are bands of case points depending on their level of missing responses. Each case has a position in this map, but we are more interested in groups of cases than individual cases—Figure 4 thus shows the average positions of male and female respondents in the six countries: positions stretch from the most liberal at bottom left (French and Norwegian females) to the

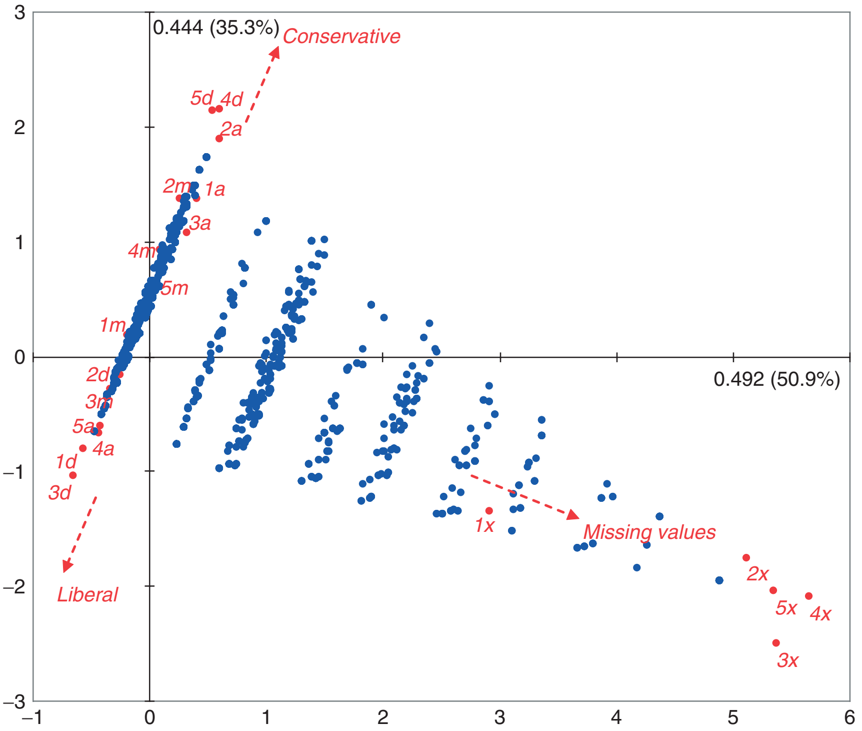

FIGURE 3 | Multiple correspondence analysis (MCA) of questionnaire data from six countries. Statements are (1) married people are generally happier, (2) bad marriage is better than no marriage, (3) marriage is better if people want kids, (4) couples can live together without marriage, and (5) couples can live together before getting married. Question responses can be (a) agree; (m) neither agree nor disagree; (d) disagree; (x) don't know/missing. Percentages of inertia are corrected according to Greenacre, 2007, p. 149 (see Further Reading).

most conservative at upper right (Japanese females and males). Generally, male and female attitudes from the same country are close together, the major exceptions being in Norway and France where males are noticeably more conservative. The country-gender points do not differ so much in the direction of the missing responses, which shows that the strong association of the missing categories is at an individual case level, not at an aggregate level (at least, not on the country-gender level; a similar analysis could be done using age and education groupings to see whether these showed any relationship with level of missing values).

OTHER VARIANTS OF CA

Two important variants of CA are subset correspondence analysis (SCA) and canonical correspondence analysis (CCA). Both these would be useful in the case of the second example presented above, to partial out the effect of the missing value categories. In SCA,[2] these categories are effectively deleted from consideration, while maintaining the original margins of the table (this is not the same as the supplementary point idea, where the margins would change if the missing values were declared supplementary). Geometrically, we maintain the same center and the same metric in the space but ignore the dimensions corresponding to the missing value categories.

In CCA, the display can be constrained to be linearly related (or linearly unrelated) to externally defined variables. This is used extensively in ecological research where biological variables are analyzed with the dimensions of the solution space constrained to be linear functions of environmental variables. In the questionnaire example above, we could define an external variable as the sum of the dummy variables for the five missing value categories, which is just the count of missing values for each respondent. Constraining by this variable forces the first dimension to align with the missing categories (so Figure 3 would be rotated approximately 30° anti-clockwise). Partial CCA would constrain the solution to be linearly unrelated to this external variable, and in this way would partial out this single dimension associated with the missing categories.

CA AND DATA CODING

An important feature of the CA approach is the wealth of coding schemes, which allow different data types to be transformed so that CA is suitable as a visualization method. Here are two examples of these.

Doubling

For ratings, rankings, and paired comparisons, each variable engenders two recoded variables that can be thought of as positive and negative poles. For example, a value of 2 on a 5-point rating scale ('negative' 1 to 'positive' 5, e.g. disagree to agree) is coded as two

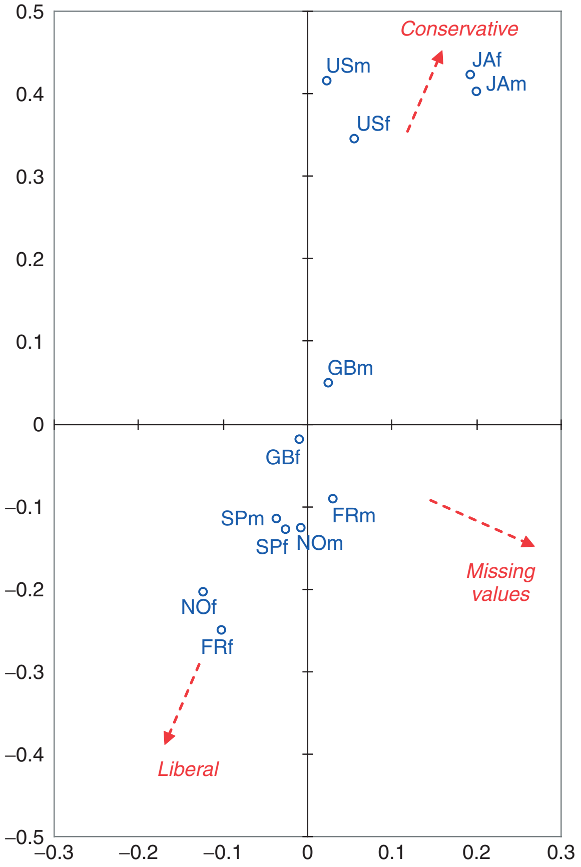

FIGURE 4 | Average positions of male (m) and female (f) respondents in Figure 3, for the six countries Great Britain (GB), USA (US), Norway (NO), Japan (JA), Spain (SP), and France (FR) (notice the considerably expanded scale compared to Figure 3).

values 1 and 3, respectively, because 2 has one scale point on the rating scale below it and three scale points above it. In a ranking of 10 objects (where 1 is the most preferred, say), a ranking of 3 would be coded $10 - 3 = 7$ and $3 - 1 = 2$, respectively, because there are seven objects less preferred and two more preferred. Note that the idea is to obtain measures of association between the respondent and the doubled variables, hence the positive pole in this last example has the higher value of 7 because the object is highly ranked.

Fuzzy coding

Continuous variables can be cut up into intervals and so converted into categorical data: for example, a variable such as temperature can be divided into three intervals using two cutpoints, so that a value of 2 represents a ‘medium’ temperature, eventually coded in dummy variable form as [0 1 0]. This ‘crisp’ coding clearly loses much information, and an alternative is ‘fuzzy’ coding, where the value is coded using so-called membership functions: for example, a recoded value of [0 0.692 0.308] would indicate a temperature somewhat higher up in the medium category toward the high category. The fuzzy-coded values also add up to 1 and so may be used in conjunction with other dummy-coded categorical data in a CA.

CONCLUSION

CA is a versatile method of dimension reduction and owes its good properties to the flexibility afforded by the weighting of the rows and columns proportional to their margins, and the dual concept of using the inverses of these margins to define interpoint χ^2 distances. It has some interesting theoretical links to other SVD-based methods, such as multidimensional scaling and the analysis of log ratios of positive data. It has been shown[3] that the CA of the matrix $k - d_{ij}^2$, where d_{ij} are interpoint distances, converges to the classical multidimensional scaling solution of the distances when k tends to infinity. Furthermore,[4] CA of power-transformed data n_{ij}^{λ} (or p_{ij}^{λ}) tends to the analysis of row or column log ratios—$\log(n_{ij}/n_{ij'})$ or $\log(n_{ij}/n_{i'j})$[5]—as the power parameter λ tends to zero. An alternative definition of CA applied to power transformation of the *contingency ratios* $[p_{ij}/(r_i c_j)]^{\lambda}$ tends in the limit to the weighted form of log-ratio analysis known in the biomedical literature as *spectral mapping* (see Ref 6 for details and references). These results mean that we can come arbitrarily close to classical MDS and to different forms of log-ratio analysis using the CA algorithm on appropriately transformed data.

An R package for performing CA and MCA with the supplementary point and subset options is described in Ref 7, while an R package[8] for ecologists includes CA and CCA.

REFERENCES

1. International Social Survey Programme: Survey on Family and Changing Gender Roles III. 2002. Available at: http://www.issp.org (Accessed July 18, 2010).
2. Greenacre M, Pardo R. Subset correspondence analysis: visualization of selected response categories in a questionnaire survey. *Sociol Methods Res* 2006, 35:193–218.

3. Carroll JD, Kumbasar E, Romney AK. An equivalence relation between correspondence analysis and classical metric multidimensional scaling for the recovery of Euclidean distances. *Br J Math Stat Psychol* 1997, 50:81–92.
4. Greenacre M. Power transformations in correspondence analysis. *Comput Stat Data Anal* 2009, 53:3107–3116.
5. Aitchison J, Greenacre M. Biplots of compositional data. *Appl Stat* 2002, 51:375–392.
6. Greenacre M, Lewi PJ. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio scale measurements. *J Classif* 2009, 26:29–54.
7. Nenadić O, Greenacre M. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *J Stat Softw* 20. Available at: http://www.jstatsoft.org/v20/i03/, package at http://cran.r-project.org/ (Accessed July 18, 2010).
8. Oksanen J, Kindt R, Legendre P, O'Hara RB. Vegan: Community Ecology Package version 1.8-3. Available at: http://cran.r-project.org/.

FURTHER READING

The book by Greenacre (2007) gives a practical introduction to all aspects of CA, MCA, SCA and CCA. The book by Murtagh (2005) pays special attention to different data coding systems to apply CA to different types of data. The books by Greenacre and Blasius (1994, 1998, 2006), were collectively written by over 100 statisticians and social scientists, with strict refereeing and editing, and reflect the development of the theoretical and practical aspects of CA and related methods.

Blasius J, Greenacre M, eds. *Visualization of Categorical Data*. San Diego: Academic Press; 1998.

Greenacre M. *Correspondence Analysis in Practice*. 2nd edn. 2007. Chapman & Hall/CRC, London. Published in Spanish as *La Práctica del Análisis de Correspondencias*, Fundación BBVA, Madrid; 2008.

Greenacre M, Blasius J, eds. *Correspondence Analysis in the Social Sciences*. London: Academic Press; 1994.

Greenacre M, Blasius J, eds. *Multiple Correspondence Analysis and Related Methods*. London: Chapman & Hall/CRC; 2006.

Murtagh F. *Correspondence Analysis and Data Coding with R and Java*. London: Chapman & Hall/CRC; 2005.