The CHIC Analysis Software v1.0

Angelos Markos[1], George Menexes[2], and Iannis Papadimitriou[1]

[1] Department of Applied Informatics, University of Macedonia, Greece
amarkos@uom.gr, iannis@uom.gr
[2] Lab of Agronomy, School of Agriculture, Aristotle University of Thessaloniki, Greece gmenexes@uom.gr

Summary. In this paper we describe CHIC (Correspondence & HIerarchical Cluster) Analysis, a specialized software package for Correspondence Analysis-CA (Simple and Multiple) and Hierarchical Cluster Analysis (Benzécri's chi-square distance, Ward's linkage criterion). The implementation of CA is in line with both the French approach and the Gifi System of data analysis. CHIC Analysis combines the graphical interface features of CodeGear Delphi with the computational power of MatLab. The software was implemented as an attempt to contribute to the effectiveness and reliability of CA. For this purpose, it offers a variety of aids to the results' interpretation and tools for the construction of special data tables. A modified version of the CA algorithm is implemented in the multivariate case. Special emphasis has been put on the graphical options for biplots, maps and dendrograms.

Key words: correspondence analysis, hierarchical cluster analysis, singular value decomposition.

1 Introduction

Correspondence Analysis (CA) is a multidimensional data analytic method, suitable for graphically exploring the association between two or more, nonmetric variables without a priori hypotheses or assumptions. Similar to Principal Component Analysis, CA results in elegant but simple lower-dimensional displays, so that the principal dimensions (usually two or three) capture the most variance (or inertia) possible. There are two popular ways to treat CA; the geometrical point of view of the French school of data analysis [2] and the optimal scaling framework of the Gifi System [4].

A common practice among researchers and practitioners is the complementary use of CA and a hierarchical cluster analysis (HCA) procedure, based on Ward's minimum-distance criterion and Benzécri's chi-square distance [2, 9]. This specific Ward clustering provides a decomposition of inertia with respect to the nodes of a dendrogram, analogous to the decomposition in the CA con-

text [6]. More details on the theoretical background of HCA, CA and various extensions can be found in [2, 4, 5, 6, 7, 17].

CA has become increasingly popular over the last decades and simple and multiple CA were introduced into most of the mainstream statistical software packages. General purpose software such as SAS [21], SPSS [16] and XL-STAT [1], implement CA offering a variety of options. However, apart from the XL-STAT software, none of the major programs offers recent developments [18]. Additionally, the widespread adoption of **R** [20] within the statistics community led to some important open source CA implementations. The **ca** package [18] provides functions for Simple, Multiple and Joint CA. Simple and Canonical CA are provided by **anacor** [11], a package which offers alternative plotting options and scaling methods. Multiple CA also known as Homogeneity Analysis (HOMALS) along with various Gifi extensions can be computed by means of the **homals** package [12]. **FactoMineR** performs CA (simple and multiple) offering a variety of interpretation options [8]. For most **R** packages a strong level of familiarity with the command line is kind of assumed.

In this paper we present CHIC (Correspondence & HIerarchical Cluster) Analysis, a specialized software which implements CA (Simple and Multiple) and HCA as a complementary method. The software combines two different development tools; Codegear Delphi 7, a visual programming language and MATLAB [14], a high-level scripting language. This scheme offers a high degree of flexibility since MATLAB is useful for implementing matrix computations, while Delphi offers a variety of tools for the design of graphical user interfaces. The implementation of CA is accompanied by a variety of options for empirical interpretation, statistical inference and visualization, inherent either in the Gifi System or the French approach. Moreover, it offers a modification of the main CA algorithm, so that the analysis of "tall" data sets (objects $\gg$ variables) becomes both feasible and effective. Finally, it is important to note that the development of CHIC Analysis was motivated by the need to teach CA and related methods to students with little or no statistical background and familiarity with the command line.

The paper is organized as follows: Section 2 describes in brief the data entry and data management options. The various interpretation options and relative criteria for simple and multiple CA, available in CHIC Analysis, are described in Sections 3, 4 and 5. A hierarchical clustering procedure as a complementary method is discussed in Sect. 6. A numerical example is given to demonstrate the use of new or interesting features. The paper concludes in Sect. 7.

2 Data Entry and Data Management

CHIC Analysis offers a customized data spreadsheet for direct data entry in the form of either a raw data table (*observations* $\times$ *variables*) or a contin-

gency table of variable categories. Additionally, data can be imported into the spreadsheet from an MS Excel or a text delimited file. There are, also, tools for the discretization of quantitative variables into ordinal and the recoding of categorical variables.

Additionally, the software offers a graphical tool for the direct construction of a Burt subtable or a two-way pivot table, in an abstract form (see Fig. 1). The user can select variables of interest from a list of all available variables in the data set and then drag and drop the desired variables into the available row and column lists. The corresponding variable categories correspond to the rows and columns of a Burt subtable, a contingency table which will be subsequently analyzed by CA.

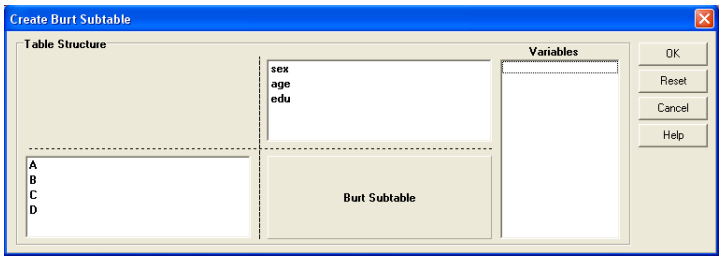

Fig. 1. A tool for Burt Subtable construction

3 Simple Correspondence Analysis

The implementation of the CA algorithm follows that of [3, 6] and its crucial step is the Singular Value Decomposition (SVD) of the standardized residuals matrix. The standard output of CA contains the eigenvalues and the relative and cumulative percentages of explained inertia for all available dimensions. The selection of significant axes can be based on the scree plot and on three different statistical significance tests of the principal inertias proposed by Nishisato [19], Van de Geer [22] and Greenacre [5, 6], respectively. Table 1 shows the output of CA on the *smoke* dataset, which contains frequencies of smoking habits for staff groups in a fictional company [18, 6]. According to the first two criteria, only the first principal component is statistically significant at the 5% level ($p(VdG) < 0.05$, $p(Nish.) < 0.05$). The options for row and column points include principal coordinates with respect to the dimensionality of the solution, total quality (QLT), inertias (INR), masses (MASS), squared correlations (COR) and contributions (CTR) (see [6] for more details about these concepts). Additional significance criteria of individual points include correlations (SQCOR), which is the equivalent of the factor loadings in PCA [3] and the Best index, which, similar to CTR, is an indicator of which points

Table 1. Eigenvalues, percentages of inertia and statistical significance

Axis	χ^2	df (VdG)	p (VdG)	χ^2 (Nish.)	df (Nish.)	p (Nish.)	Inertia	%	Cum.%
1	14.429	6	0.025	14.608	6	0.024	0.075	87.756	87.756
2	1.933	4	0.748	1.893	4	0.755	0.010	11.759	99.515
3	0.080	2	0.961	0.078	2	0.962	0.000	0.485	100.000

Critical χ^2 value = 15.24 ($\alpha = 0.05$)

best explain the inertia of each dimension [21]. In the case of supplementary variables, an (S) is appended to the supplementary variable names in the output, which includes only the QLT, INR and COR indices. Table 2 exhibits the row output for the first significant axis of the *smoke* dataset. Optionally, the

Table 2. Principal row projections, contributions and correlations

	QLT	MASS	INR	Best	F1	INR1	COR1	SQCOR1	CTR1	Best1
SM	0.092	0.057	0.003	3	-0.066	0.000	0.092	-0.304	0.003	0
JM	0.526	0.093	0.012	2	0.259	0.006	0.526	0.726	0.084	0
SE	1.000	0.264	0.038	1	-0.381	0.038	1.000	-1.000	0.512	1
JE	0.942	0.456	0.026	1	0.233	0.025	0.942	0.971	0.331	1
SC	0.865	0.130	0.006	2	-0.201	0.005	0.865	-0.930	0.070	0

user can ask for the expected frequencies, three kinds of residuals (plain, standardized and adjusted) of the original contingency table, variable chi-square contributions and the reconstructed input data for a given dimensionality of the solution, as described in [3].

4 Multiple Correspondence Analysis

Multiple CA is in fact a simple CA that can be carried out in terms of the SVD on either the indicator matrix or the Burt matrix, a choice which depends on the purpose of the analysis. The indicator matrix is a binary representation of the different categorical values of each variable, while the Burt matrix is equal to the cross-product of the indicator matrix and concatenates all two-way cross-tabulations between pairs of variables [5, 6]. In cases where a CA on the indicator matrix $\mathbf{Z}$ is of interest, we perform alternatively the SVD on the standardized residuals matrix, calculating on the Burt matrix $\mathbf{B}$. Then, we use the well-known transition formulae of CA [5, 6] and the relation between $\mathbf{Z}$ and $\mathbf{B}$, to obtain the results of the CA on $\mathbf{Z}$. This scheme bypasses the decomposition of $\mathbf{Z}$ and can be efficient in the case of "tall" data sets, where the number of objects is much greater than the number of variables. It is important to note that the same CA solution can be also efficiently

obtained by means of an Alternating Least Squares algorithm (ALS), which iteratively minimizes a least-squares loss function [4]. A thorough description of the modified CA version and its efficiency can be found in [13].

The standard output of MCA remains the same as in the simple case (see Sect. 3). Additional options include the summary of contributions (CTR) for each variable and the discrimination measures [4, 16], an important interpretation option inherent in the Gifi System. Furthermore, test values can be calculated for supplementary variables, as a measure of the significance between a variable and an axis [10]. The selection of significant axes can be based on a statistical significance test of the principal inertias, proposed by Nishisato [19] and on Cronbach's alpha, as a measure of reliability of each principal inertia [6]. Finally, in case the analysis is based on the Burt matrix, two inertia adjustment options are offered for solving the low percentage of inertia problem, proposed by Grenacre [6] and Menexes [15], respectively. Both options are based on the average inertia in Burt's off-diagonal blocks.

5 Visualization Options

The basic plot in CA and MCA is the symmetric map where both rows and columns are plotted in principal coordinates. Depending on the situation, other types of display are appropriate. This can be set with the normalization options for CA and MCA. Following [18], in Table 3 we give a brief overview over the available options and their meanings.

Table 3. Normalization Options in CA and MCA maps

option	*description*
RPN - Row Principal	Rows in principal and columns in standard coordinates
CPN - Column Principal	Rows in standard and columns in principal coordinates
SN - Symmetrical	Row and column coordinates are scaled to have variances equal to the singular values
PN - Principal	Rows and columns in principal coordinates (default)

The first three scaling options lead to a biplot, while the last one results in a symmetric map [6]. The intepretation of biplots is enhanced with the option to draw the biplot axes passing through each row (or column) point, as shown in Fig.2. The dots represent the intersections of the orthogonal projections of points on these biplot axes. In the case of RPN or CPN, the corresponding biplot is likely to be crowded; in that case the interpretation can be based on a table of distances and correlations (Table 4). The distances are in ascending order and indicate a ranking or ordering of the projected points, while Cos2 indicates the square cosine of the angle between a biplot axis and a position

vector of a point. For example, for the biplot axis passing through the point "SM", the distance of the projection of the point "Heavy", on the biplot axis, from the point "SM" is 1.791 and the squared correlation between "Heavy" and "SM" is 0.14.

Table 4. Distances and correlations on biplot axes

SM	Heavy	None	Medium	Light
Distance	1.791	1.873	1.982	2.103
Cos2	0.140	0.039	0.998	0.307
JM	**Heavy**	**None**	**Medium**	**Light**
Distance	2.469	2.540	2.636	2.744
Cos2	0.390	0.780	0.615	0.109
SE	**Heavy**	**None**	**Medium**	**Light**
Distance	1.373	1.383	1.403	1.434
Cos2	0.043	0.025	0.754	0.675
JE	**Light**	**Medium**	**None**	**Heavy**
Distance	0.936	1.009	1.082	1.141
Cos2	1.000	0.874	0.204	0.430
SC	**Light**	**Medium**	**None**	**Heavy**
Distance	0.987	1.058	1.129	1.187
Cos2	0.404	0.718	0.085	0.949

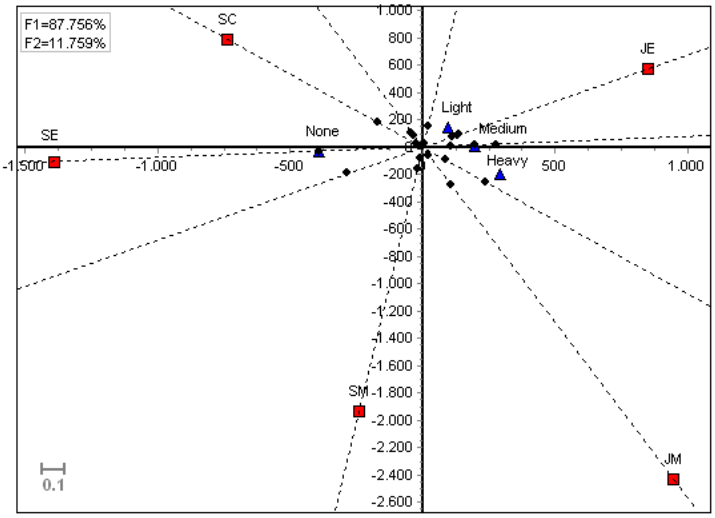

Fig. 2. Asymmetric map of the smoking data with CPN and biplot axes

6 Ward Clustering as a Complementary Method

Hierarchical Cluster Analysis (HCA) can be used as a complementary method to CA, in order to identify relatively homogeneous clusters either in the original data or in the low-dimensional space. A Ward clustering procedure takes into account the chi-square distances between the profiles and the associated masses. This way it provides a decomposition of inertia with respect to the nodes of a dendrogram, analogous to the decomposition of inertia with respect to principal axes in CA [6]. The total inertia (or equivalently the chi-square statistic) of the table is reduced by a minimum at each successive level of merging of the rows (or columns). More details on the HCA algorithm implementation can be found in [6, 9, 17].

Furthermore, CHIC Analysis offers a group of interpretation options traditionally called VACOR, which allow the user to explore the representation of clusters derived from the hierarchical trees in factor space and describe the cluster dipoles which take account of the cluster components. More details on the VACOR implementation can be found in [2, 17].

7 Summary

We have presented CHIC Analysis, a specialized software for simple, multiple correspondence analysis and hierarchical clustering. The software contains most of the features of present available software packages as well as various new features that are not available elsewhere. Amongst the main advantages of this program is that it is menu driven, available for free and offers a large variety of complementary options to facilitate data interpretation. The included data entry and data management utilities make it possible to handle directly almost any data table, and this gives the user a great deal of flexibility. The software and its user's manual can be downloaded from http://www.amarkos.gr/en/research/chic.

In future releases, we plan to take advantage of the common mathematical foundation of many multivariate data analysis methods, as a basis for incorporating CA variations and related methods.

References

1. Addinsoft. *XLSTAT Statistical software for MS Excel.* URL http://www.xlstat.com/, 2007.
2. J.-P. Benzécri. *Correspondence Analysis Handbook.* Marcel Dekker, New York, 1992.
3. J. Blasius and M.J. Greenacre. Computation of Correspondence Analysis. In M.J. Greenacre, J. Blasius, editors, *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, pages 53-75. Academic Press, London, 1994.

4. A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, 1990.
5. M.J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.
6. M.J. Greenacre. *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton, 2007.
7. M.J. Greenacre and J. Blasius, editors. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, London, 2006.
8. S. Le, J. Josse, and F. Husson. FactoMineR: an R package for multivariate analysis, *Journal of Statistical Software*, 25(1):1–18, 2008.
9. L. Lebart. Complementary use of correspondence analysis and cluster analysis. In M.J. Greenacre and J. Blasius, editors, *Correspondence Analysis in the Social sciences. Recent Developments and Applications*, pages 162–178. Academic Press, London, 1994.
10. L. Lebart. Validation techniques in multiple correspondence analysis. In M.J. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*, pages 179–194. Chapman & Hall, London, 2006.
11. P. Mair and J. de Leeuw, Simple and canonical correspondence analysis using the R package anacor, *Journal of Statistical Software*, 31(5):1–18, 2009.
12. P. Mair and J. de Leeuw, Gifi methods for optimal scaling in R: The package homals, *Journal of Statistical Software*, 31(4):1–21, 2009.
13. A. Markos, G. Menexes, and T. Papadimitriou. Multiple correspondence analysis for "tall" data sets, *Intelligent Data Analysis*, forthcoming, 2009.
14. The MathWorks, Inc. MATLAB - The Language of Technical Computing, Version 7.5. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/, 2007.
15. G. Menexes and I. Papadimitriou. Relations of inertia, In *Abstracts of International Conference on Correspondence Analysis and Related Methods* (CARME 2003).
16. J.J. Meulman and W.J. Heiser. *SPSS Categories 14.0*. Chicago, IL: SPSS, Inc., 2005.
17. F. Murtagh. *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall, 2005.
18. O. Nenadic and M.J. Greenacre. Correspondence analysis in R, with two- and three dimensional graphics: The ca Package. *Journal of Statistical Software*, 20(3):1–13, 2006.
19. S. Nishisato. *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press, Toronto, 1980.
20. R Development Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2007.
21. SAS Institute Inc. SAS/STAT Software, Version 9.1. Cary, NC. URL http://www.sas.com/, 2003.
22. J.P. Van de Geer. *Multivariate Analysis of Categorical Data: Applications*. Advanced Quantitative Techniques in the Social Sciences, Vol. 3. Sage, Newbury Park, 1993.