Available online at www.sciencedirect.com

ScienceDirect

SoftwareX

SoftwareX 1–2 (2015) 26–31

www.elsevier.com/locate/softx

CAinterprTools: An R package to help interpreting Correspondence Analysis' results

Gianmarco Alberti

Department of Classics and Archaeology, Archaeology Farmhouse, University of Malta, Car Park 6, Msida, MSD 2080, Malta

Received 11 June 2015; received in revised form 16 July 2015; accepted 16 July 2015

Abstract

Correspondence Analysis (CA) is a statistical exploratory technique frequently used in many research fields to graphically visualize the structure of contingency tables. Many programs, both commercial and free, perform CA but none of them as yet provides a visual aid to the interpretation of the results. The 'CAinterprTools' package, designed to be used in the free R statistical environment, aims at filling that gap. A novel-to-medium R user has been considered as target. 15 commands enable to easily obtain charts that help (and are relevant to) the interpretation of the CA's results, freeing the user from the need to inspect and scrutinize tabular CA outputs, and to look up values and statistics on which further calculations are necessary. The package also implements tests to assess the significance of the input table's total inertia and individual dimensions.

Keywords: Correspondence analysis; Contingency tables; Interpretation; R; Package

Code Metadata Table

Current code version	*v 0.4*
Permanent link to code/repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00027
Legal Code License	*GPLv2*
Code versioning system used	*git*
Software code languages, tools, and services used	*R*(>=3.1.1)
Dependences	*ca, FactoMineR, InPosition, Hmisc*
If available Link to developer documentation/manual	https://github.com/gianmarcoalberti/CAinterprTools, http://cainarchaeology.weebly.com/cainterprtools-r-package.html
Support email for questions	gianmarcoalberti@tin.it gianmarco.alberti@um.edu.mt

1. Introduction

The use of contingency tables is widespread in many research fields. Archaeologists, political scientists, sociologists, biologists, linguistics (to cite a few) use contingency tables to summarize nominal data. They also need statistical tools to analyse cross-tabulations in order, for instance, to detect and measure the strength of the patterns of association between nominal variables. A number of statistical approaches are used for these purposes, encompassing hypothesis testing [1], logistic regression [2], and log-linear modelling [3]. Besides these approaches, Correspondence Analysis (hereafter CA) is an exploratory statistical technique frequently applied to contingency tables. Even though it has been slow in gaining popularity outside France before 1980s [4], CA is now widely used in fields as diverse as archaeology [5,6], marine biology [7], paleontology [8], marketing research [9], analysis of food preferences [10], textual analysis [11], crime studies [12], and other research [13,14].

E-mail addresses: gianmarco.alberti@um.edu.mt, gianmarcoalberti@tin.it.

http://dx.doi.org/10.1016/j.softx.2015.07.001

Referring the reader to existing literature for the mechanics, computation, and underlying logic [4,14–16], it suffices here to say that CA allows visually displaying the dependence between rows and columns of a contingency table in order to help the interpretation and to let patterns emerge. It reduces the number of dimensions needed to display the data points by decomposing the total inertia (i.e., the variability) of the table and isolating the smallest number of dimensions that can capture the data variability. CA returns a scatterplot where rows and/or columns are represented as points in a sequence of low-dimensional spaces. These spaces retain a decreasing amount of the total inertia, with the first dimension capturing the highest amount, while the second will capture the second largest proportion, and so on. On the scatterplot, the distance between data points of the same type (i.e., row-to-row) is related to the degree to which the rows have similar profiles (i.e., relative frequencies of column categories). The same applies for the column-to-column distance. The more the points are close to one another, the more similar their profiles will be. The origin of the axes represents the centroid (i.e., the average profile), and can be thought of as the place where there is no difference between profiles. The more different are the latter, the more the profile points will be spread on the plane away from the centroid. As for the relative distances between points of different type (i.e., row-to-column), it tells the analyst something about the "correspondence" between the categories that made up the table. In other words, the more a row point is close to a column point, the greater (i.e., the more distant from the average) is the proportion of that column category on the row profile.

2. Motivation and significance

Any statistical software, either commercial (e.g., Minitab, STATISTCA, JMP, XLSTAT, SYSTAT) or freeware (e.g., PAST) [17], perform CA. The same holds true for a number of packages that have been recently made available for the R statistical programming environment [18], such as 'ca' [19] and 'FactoMineR' [20] (for others, see [14]). The implementation of CA in R is also described in Greenacre's [15], and Beh and Lombardo's [14] books.

With the use of the available facilities it is easy to obtain one of the main output researchers are interested in, namely the scatterplot representing row and/or column points projected on a subspace chosen by the user. It must be noted, however, that in order to interpret the CA scatterplot and to have a sound comprehension of the data structure, the mere examination of that graph is not enough. The user has to consult a number of statistics reported on screen in tabular form [4,15]. Further, the user must perform from scratch some calculations on the basis of those raw statistics. Referring to the literature [14,15] for a guide to the use and interpretation of the CA outputs, I limit myself to cite few examples.

One of the most important step in understanding CA results is deciding how many dimensions can be considered important for interpretation. The analyst is faced with the need of a trade-off between the increasing explained data variability deriving by keeping many dimensions versus the increasing complexity that can make difficult the interpretation of more than two dimensions. One of the most used rule is the so-called 'average rule' [21]: analysts should retain all the dimensions that explain more than the average inertia (expressed in terms of percentages), the latter being equal to 100 divided by the dimensionality of the table (i.e., the number of rows or columns, whichever is smaller, minus 1). To apply this rule, the user has to calculate the dimensionality of the table, divide 100 by the latter, and then look up the table reporting the inertia explained by the CA dimensions and spotting which dimension is greater than that value. In another instance, users have to understand what row/column categories have a major contribution to the definition of given dimensions. If one is interested in spotting what row categories are actually contributing to the definition of the dimensions, say, 1 and 3, the user has to divide 100 by the number of rows, inspect the table listing the contribution of the categories to those specific dimensions, and keep trace of the row categories whose contribution to the inertia of those specific dimensions is greater than the devised figure.

These examples are meant to introduce the significance of the CAinterprTools package, whose aim is twofold. On the one hand, it provides charts that help (and are relevant to) the interpretation of the CA's results, freeing the user from the need to inspect and scrutinize the tabular CA output, and to look up values and statistics on which further calculations are necessary. This is not meant to suggest that the numerical output provided by other programs are not useful. I merely maintain that a visual aid to CA interpretation may prove easier and less time-consuming, while users can always go back to the numerical output if they need. On the other hand, the package also implement three functions that provide the facility to perform some hypothesis tests on the significance of the total inertia and of the inertia explained by individual dimensions. As for the latter, two different approaches have been used, one implementing the permutation test described by Greenacre [15], the other implementing a chi-square-based method called Malinvaud's test [22,23]. It is worth noting that the three functions, as well as the other ones implemented in the package, are not as yet available from any stats tool-pack, whether free or commercial, at the best of my knowledge. Last but not least, the package is freely available and can be easily downloaded and installed in the free R statistical programming environment, as described in the following paragraph.

3. Software description and illustrative example

The CAinterprTools package is available from a GitHub repository. It can be downloaded and installed into R by taking just few steps:

(1) installing the 'devtools' package:
install.packages ("devtools", dependencies =TRUE)

(2) loading that package: *library(devtools)*

(3) downloading the 'CAinterprTools' package from GitHub via the 'devtools''s command:
install_github("gianmarcoalberti/CAinterprTools")

Once installed, the package can be loaded by: *library(CAinterprTools)*.

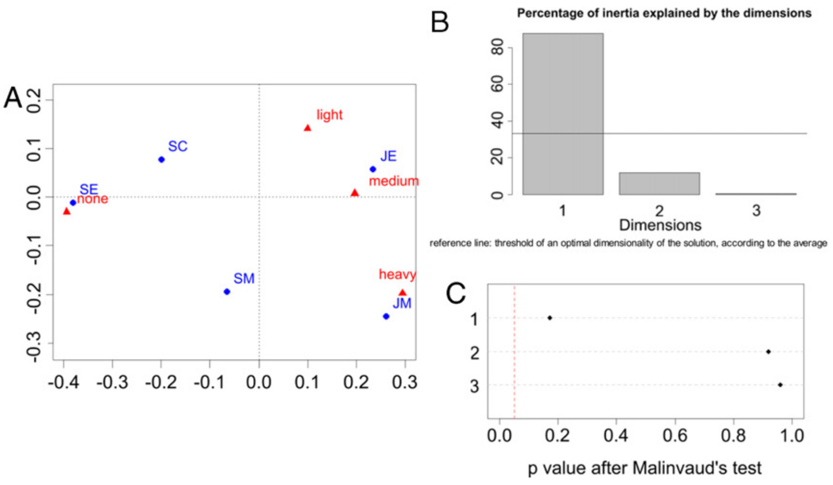

Fig. 1. (A) Correspondence Analysis scatterplot of the sub-space defined by dimension 1 and 2 (data from Table 1). (B) Bar-chart of the inertia explained by the three CA dimensions, and reference line indicating the threshold above which a dimension is important for the CA interpretation. (C) Chart reporting the statistical significance of the three CA dimensions according to Malinvaud's test.

The package depends on the 'ca', 'FactoMineR', 'InPosition', and 'Hmisc' (which are installed and loaded by default upon installing CAinterprTools) and features 15 commands, whose use is described in the help documentation and in an accompanying website (http://cainarchaeology.weebly.com/cainterprtools-r-package.html). The package also comes with a sample dataset after Greenacre's handbook [15]. It is worth stressing that, in designing and implementing the package's commands, a novel-to-medium R user has been considered as target. As consequence, it has been decided to keep commands as simple and short as possible, both in terms of commands' name length and number of arguments to be fed into each individual function.

A description of the package's features is provided by means of an illustrative example. It takes into account a (fictional) small dataset first devised by Greenacre [15], concerning a survey of 193 staff members of a company (classified into Senior Managers, Junior Managers, Senior Employees, Junior Employees, Secretaries) as to their smoking habits (None, Light, Medium, Heavy) (Table 1).

The choice of this dataset is meant to enhance the comparability of the described package's results with both literature and with commercial programs. In fact, as Greenacre notes, the dataset has been adopted as a test example in implementations of CA in many software. While the use of CA is desirable in relation to large contingency tables, using this rather small dataset helps keeping things simple for the sake of the illustrative purposes.

The scatterplot of the first two CA dimensions is in Fig. 1A.

In what follows, it is assumed that the dataset has been fed into R as an object named *smoke*. While the percentage of the inertia explained by each dimension is reported in the scatterplot, the package's *aver.rule(smoke)* command allows obtaining the bar-chart reproduced in Fig. 1B. Besides visually representing the decreasing amount of total inertia accounted for by all the CA dimensions, a reference line indicates the threshold above which a dimension can be considered important for the interpretation of the results, according the aforementioned so-called average rule. The chart shows that the first dimension accounts for most of the inertia and is well above the average-rule threshold.

Table 1
Sample dataset: staff members of a fictional company cross-tabulated against their smoking habits.

	None	Light	Medium	Heavy	
Senior managers	4	2	3	2	*11*
Junior managers	4	3	7	4	*18*
Senior employees	25	10	12	4	*51*
Junior employees	18	24	33	13	*88*
Secretaries	10	6	7	2	*25*
	61	*45*	*62*	*25*	**193**

Users could be interested in the significance of the total inertia as well as the significance of the dimensions. For these purposes, three functions have been implemented. The command *malinvaud(smoke)* returns a table in the R console and a chart (Fig. 1C) in which the significance of each dimension can be easily spotted. Referring the reader to the literature already provided, the Malinvaud test checks the significance of the remaining dimensions once the first k ones have been selected. In this example, none of the three dimensions turn out to be significant at alpha 0.05.

As for the permutation-based tests, the command *sig.tot.inertia.perm(smoke)* returns the frequency curve of the permuted total inertia (based on 999 simulated tables) (Fig. 2A). Two reference lines, one representing the observed total inertia (0.0852) and one the 95th percentile (0.10945, both reported on the R console) of the permuted total inertia, allow visually assessing the significance of the observed total inertia. In this case, the test yields a non-significant result, indicating that the hypothesis of independence between rows and columns cannot be rejected. This command can be used in place of the traditional chi-square test, which points in the same direction (chi-square: 16.441, df: 12, p: 0.171). A permutation-based test (using on 999 simulated tables) is implemented to test the significance of any pair of dimensions.

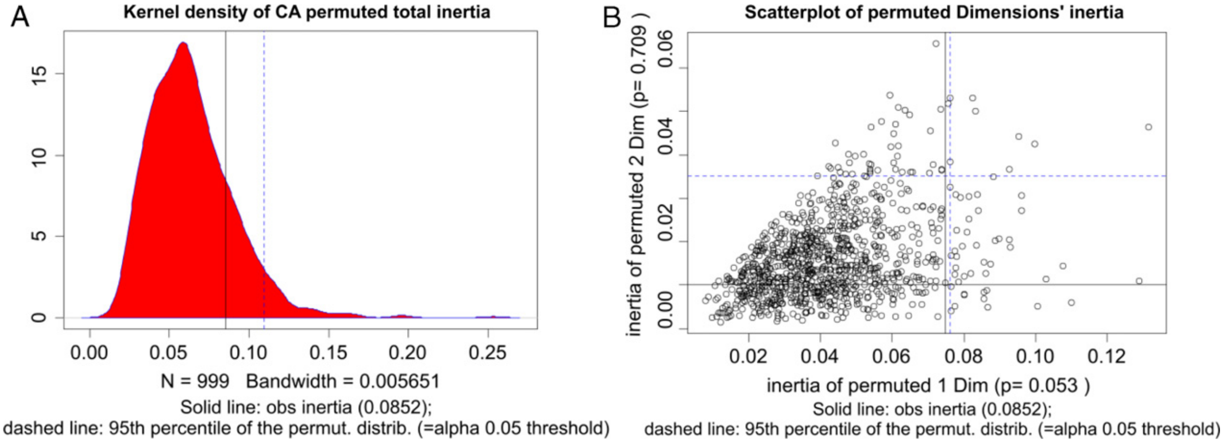

Fig. 2. (A) Density curve of the permuted total inertia, based on 999 simulated tables; observed total inertia and 95th percentile of the permuted total inertia are also reported. (B) Permuted inertia (based on 999 simulated tables) of dimension 1 and 2 plotted against one another; reference lines indicate the observed inertia and the 95th percentile of the permuted inertia; significance of the observed inertia is also reported.

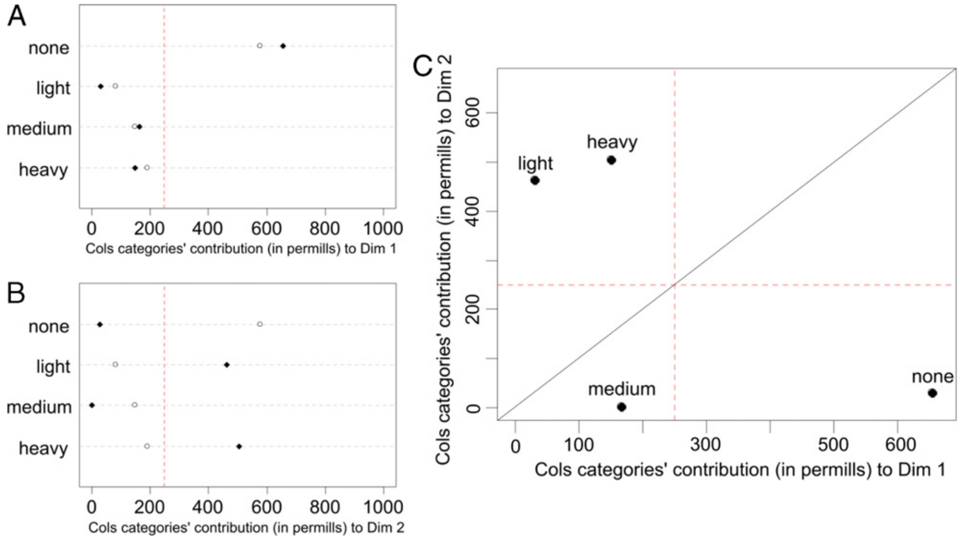

Fig. 3. Contribution of column categories to the definition of dimension 1 (A) and 2 (B); hollow points represent the contribution to the total inertia. (C) Scatterplot of the column categories' contribution to dimension 1 and 2.

The command *sig.dim.perm(smoke,1,2)* returns the scatterplot in Fig. 2B, where the permuted inertias of dimension 1 and 2 are plotted against one another. Again, reference lines representing the observed and permuted dimensions' inertia allow visually assessing the significance. The latter is also reported in the axes' labels. In the example, both dimensions are not significant at alpha 0.05, with dimension 1 and 2 having a p value of 0.053 and 0.709 respectively.

As for the interpretation of the CA scatterplot, since the user could be interested in understanding the similarities between the company's members as far as their smoking habits are concerned, it could be decided to interpret the position of the former (i.e., row categories) in the sub-space defined by the latter (i.e., column categories). As consequence, users may wish to know which smoking habit is actually defining the first two CA dimensions. The command *cols.cntr(smoke,1,T)* and *cols.cntr(smoke,2,T)* return the charts in Fig. 3A–B, showing the contribution (in permills) of the smoking habits to dimension 1 and 2 respectively (solid dots). A reference line helps in locating which habit has an important contribution to the determination of the dimension. Further, the parameter *T* enables to display in the same chart the contribution of the smoking habits to the total inertia (hollow dots). The 'none' smoking habit is contributing to the definition of the first dimension (and it is the major contributor to the total inertia as well), while the 'light' and 'heavy' categories have a major contribution to the definition of dimension 2. If one wants to couple the above information in the same chart, the command *cols.cntr.scatter(smoke,1,2)* returns the scatterplot in Fig. 3C, where the contributions to dimension 1 and 2 are plotted against one another. It can be easily eyeballed that different smoking habits are actually contributing to the determination of the two dimensions.

On the basis of these information, referring back to the CA scatterplot, it is quite easy to interpret the dimensions. The first, which as seen is determined by non-smokers, is actually accounting for the majority of the inertia of the data. Further, it is opposing non-smokers to the other smoking

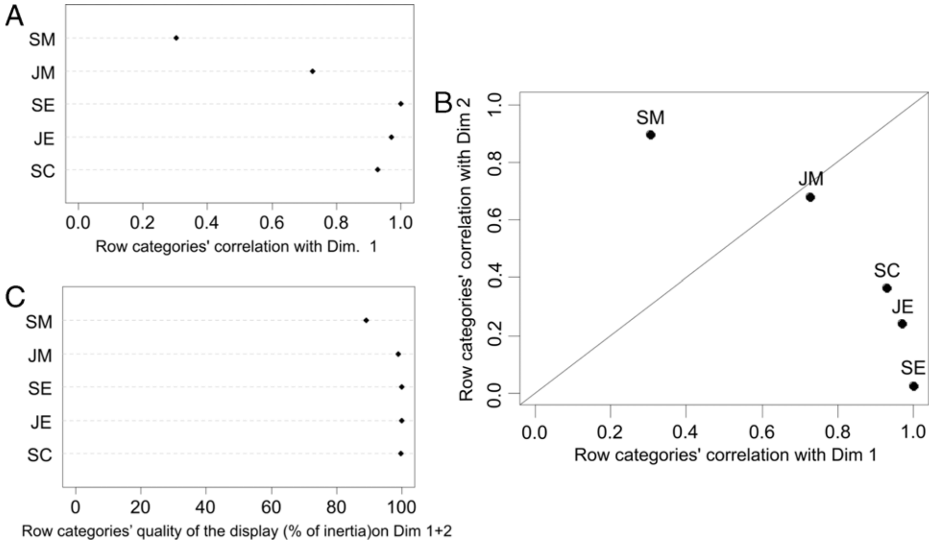

Fig. 4. (A) Row categories' correlation with dimension 1. (B) Scatterplot of the row categories' correlation with dimension 1 and 2. (C) Quality of the display (% of total inertia) of row categories on the sub-space defined by dimension 1 and 2.

categories, implying that most of the data variability is due to the difference between who smokes and who does not. The second dimension, which as seen is defined by light and heavy-smokers, is opposing the former to the latter category. After having interpreted the CA space defined by the smoking habit categories, the next step would be to understand which staff member category is associated with which dimension and, ultimately, with which smoking habits.

This step can be achieved with two commands: *rows.corr (smoke,1)*, which visually reports the correlation of the row categories with dimension 1, and *rows.corr.scatter(smoke,1,2)*, which displays a scatterplot of the correlation with dimensions 1 and 2 (Fig. 4A–B). It is very easy to see that Senior Employees and, to a slightly lesser extent, Secretaries have a very high correlation with dimension 1. This implies that there are relatively more non-smokers among them. Junior and Senior Managers have the highest (in relative terms) correlation with dimension 2 (specifically, with its negative pole), implying that there are relatively more heavy smokers among them. The Junior Managers category has also a high correlation with dimension 1 (specifically, with its positive pole) implying that there is a relatively higher proportion of smokers relative to non-smokers (represented, as seen, by the negative pole of the same dimension). The same applies to Junior Employees.

The user should be aware of the fact that not all the points could be well displayed in the chosen dimensions. To assess the quality of the display (in terms of the percentage of the total inertia that is captured by the selected dimensions) the command *rows.qlt(smoke,1,2)* can be used. It can be seen (Fig. 4C) that all the row categories in the illustrative example are well displayed on the plane defined by dimension 1 and 2; only the Senior Managers category has a relatively worse, yet still very high, quality of the display.

Finally, it is worth remembering that what described so far (in terms of dimensions interpretation, correlation to the dimensions, and quality of the display) can be also accomplished from the standpoint of the column categories. Each described command has its column counterpart.

4. Impact and conclusions

From the preceding description, it is apparent that the CAinterprTools package has not been conceived to pursue new research questions. Rather, its rationale and motivation lie in the very opportunity to provide users with a body of facilities that allow users to get the most of those CA statistics that are crucial to the understanding of the results. As a matter of fact, visually inspecting the described charts turns out to be easier and less time-consuming than going back and forth between columns of numerical values, which often need to be further processed. It is also important to stress that the commands for the calculation of some significance values may meet the users' need to add some inferential aspects to CA. In this respect, the permutation tests, as well as Malinvaud's test, provided by the package may prove useful and, notably, are not implemented elsewhere so far. Given the considerably wide use of CA in many research fields, and considering that R is admittedly not so user-friendly at the beginning and has a steep learning curve, it is likely that the package will meet the favour of many users in a wide arena of fields. The easiness of installation and its simplicity of use may assure a positive reception by both CA users and R enthusiasts. I already got few, yet encouraging, feedback about such a reception.

Acknowledgements

I thank the anonymous referees for their comments on the earlier version of the manuscript. My decision to build the described package came into being while co-supervising Miss Jessica Spiteri for her MA dissertation in archaeology at the University of Malta (Department of Classics and Archaeology; supervisor: Dr. Nicholas Vella). I realized that making the interpretation of Correspondence Analysis' results easier would had been useful for both the needs of the student and of a broader audience. I wish to thank the University of Malta for having been keen to cover the publication cost, long before the present author knew that the open access fee had been waived. I also take this opportunity to thank all the students and scholars who provided me with feedback about my R scripts, packages, and functions, directly and indirectly related to the use of CA. In alphabetical order: Professor Eric Beh, C Delgado, Josh Harlen, Joshua T Hogue, Sneha Koneru, Sebastian Kriesch, Professor Carl Lipo, Professor Rosaria Lombardo, Yonas Yohannes. Their kind gesture has been much appreciated.

References

[1] Reynolds HT. Analysis of nominal data. Iowa: SAGE University Press; 1977.

[2] Allison PD. Logistic regression using the sas system: Theory and application. Cary: Wiley-Blackwell; 2001.

[3] Agresti A. Categorical data analysis. 3rd ed. Hoboken: Wiley; 2012.

[4] Clausen SE. Applied correspondence analysis. In: An introduction. Sage university papers series in quantitative applications in the social science. Thousand Oaks: Sage; 1998.

[5] Bolviken E, Helskog E, Helskog K, Holm-Olsen IM, Solheim L, Bertelsen R. Correspondence analysis: An alternative to principal components. World Archaeol 1982;14:41–60.

[6] Alberti G. Making sense of contingency tables in archaeology: the aid of correspondence analysis to intra-site activity areas research. J Data Sci 2013;11:479–99.

[7] Ambroso S, Gori A, Dominguez-Carrió C, Gili JM, Berganzo E, Teixidó N, et al. Spatial distribution patterns of the soft corals alcyonium acaule and alcyonium palmatum in coastal bottoms (Cap de Creus, northwestern Mediterranean Sea). Mar Biol 2013;160:3059–70.

[8] Freudenthal M, Martín-Suárez E, Gallardo JA, Daroca AG, Minwer-Barakat R. The application of correspondence analysis in palaeontology. C R Palevol 2009;8:1–8.

[9] Bendixen M. Compositional perceptual mapping using chi-squared tree analysis and correspondence analysis. J Mark Manag 1995;11:571–81.

[10] Beh EJ, Lombardo R, Simonetti B. A european perception of food using two methods of correspondence analysis. Food Qual Prefer 2011;22:226–31.

[11] Blanco Abellan M. Textbooks on differential calculus in eighteenth century europe: a comparative stylistic analysis. J Data Sci 2007;5:597–612.

[12] Harcourt BE. Language of the gun: youth, crime, and public policy. Chicago-London: The University of Chicago Press; 2006.

[13] Blasius J, Greenacre M. Visualization of categorical data. San Diego-London: Academic Press; 1998.

[14] Beh EJ, Lombardo R. Correspondence analysis: Theory, practice and new strategies. Chichester: Wiley; 2014.

[15] Greenacre M. Correspondence analysis in practice. 2nd ed. Boca Raton-London-New York: Chapman & Hall/CRC-Taylor & Francis Group; 2007.

[16] Weller SC, Romney AK. Metric scaling. In: Correspondence analysis. Newbury Park-London-New Delhi: SAGE; 1990.

[17] Hammer Ø, Harper DAT, Ryan PD. Past: paleontological statistics software package for education and data analysis. Paleontol Electron 2001;4:1–9.

[18] Ihaka R, Gentelman RR. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.

[19] Nenadic O, Greenacre M. Correspondence analysis in R, with two and three-dimensional graphics: The ca package. J Stat Softw 2007;20:1–13.

[20] Lê S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. J Stat Softw 2008;25:1–18.

[21] Lorenzo-Seva U. Horn's parallel analysis for selecting the number of dimensions in correspondence analysis. Methodol Eur J Res Methods Behav Soc Sci 2011;7:96–102.

[22] Saporta G. Probabilités, analyse des données et statistique. Paris: Editions Technip; 2006.

[23] Camiz S, Gomes GC. Joint correspondence analysis versus multiple correspondence analysis: a solution to an undetected problem. In: Giusti A, editor. Classification and data mining studies in classification, data analysis, and knowledge organization. Berlin-Heidelberg: Springer; 2013. p. 11–8.