Available online at www.sciencedirect.com
ScienceDirect

SoftwareX 1–2 (2015) 26–31

SoftwareX
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,<br>http://cainarchaeology.weebly.com/cainterprtools-r-package.html |
| Support email for questions | gianmarcoalberti@tin.it<br>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)*.