Multiple Factor Analysis for Contingency Tables in FactoMineR Package

Belchin Adriyanov Kostov, Mónica Bécue Bertaut, François Husson, Daría Hernández

Multiple Factor Analysis for Contingency Tables in FactoMineR Package

Belchin Adriyanov-Kostov [a], **Mónica Bécue-Bertaut** [b], **François Husson** [c], **Daría Hernández** [d]

[a] Primary Health Care Center Les Corts, CAPSE
Mejia Lequerica, s / n, 08028 Barcelona (Spain)
badriyan@clinic.ub.es

[b] Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona (Spain)
monica.becue@upc.edu

[c] Agrocampus Rennes
65 rue de Saint-Brieuc, 35042 Rennes (France)
husson@agrocampus-ouest.fr

[d] Centro Mexicano de Estudios Económicos y Sociales
Napoleón 54, Col. Moderna 3500, México D.F. (México)
dari_hdez@yahoo.com.mx

Keywords: Multiple Contingency Tables, Multiple Factor Analysis, Multiple Factor Analysis for Contingency Tables, Scientometrics, FactoMineR

FactoMineR package [1] offers the most commonly used principal component methods: principal component analysis (PCA), correspondence analysis (CA), multiple correspondence analysis (MCA) and multiple factor analysis (MFA) [2]. MFA function has been recently extended to consider contingency/frequency tables as proposed by Bécue-Bertaut and Pagès (multiple factor analysis for contingency tables, MFACT) [3-4]. MFACT is used in different domains such as sensometrics, ecology and text mining.

Multiple factor analysis [2] deals with a multiple table, composed of sets of either quantitative or categorical variables balancing the influence of the different sets on the first principal axis by dividing the weight of their variables/columns by the first eigenvalue of the separate analysis of this set (PCA or MCA depending on the type of the variables). Thus, the highest axial inertia of each group is standardized to 1. MFA offers the usual results in any PCA (global representation of rows and columns) and also tools for comparing the different sets such as the superimposed representation of the rows as induced separately by every set of columns (partial rows). Initially multiple factor analysis for contingency tables [3] was proposed to simultaneously analyze several frequency/contingency tables. Afterward, it has been extended to multiple tables with a mixture of quantitative, categorical and frequency sets [4].

This method is presented through its application to a scientometric study in medicine. 457 abstracts relative to randomized clinical trials on *Systemic Lupus Erythematosus* (SLE) from 1994 to 2011 were downloaded from PubMED, the most important scientific data base in medical research. The abstracts×words matrix was constructed, keeping only the words repeated at least 10 times. The publication year was also considered. Thus, the data base juxtaposes a quantitative set (publication year) and a frequency set (abstracts×words a 457×1046 matrix). The aim was to study the evolution in the research concerning this rare disease, through the vocabulary changes. The superimposed representation provides a graph where the abstracts are seen globally (one global

point) and partially (two partial points) from the two different points of view that are the vocabulary of the abstract and its publication year. This representation is considered to look for those abstracts evidencing an important difference between both points of view to detect either pioneering works or works returning to topics treated in the past.

Figure 1 displays the first principal global map provided by MFACT. The interpretation rules are those of CA for the words and those of PCA for the quantitative columns, here reduced to the publication year. The different years are projected as supplementary categorical columns. Year is highly correlated (0.94) with the first dimension, opposing words related to symptoms and drugs (at the left) to etiology and genetics (at the right). The part of the evolution of the vocabulary linked to chronology is reflected on this axis. The early research in SLE was dedicated to detect its symptoms and to test the effectiveness of drugs already known and previously used for other diseases. The most recent research works are concerned by etiology, in particular through genetics. The second dimension opposes topics present in the research at a same moment such as symptoms and drugs at the beginning and genetics and innovative drugs such as “Belimumab” in the recent years.
Figure 2 offers the superimposed representation of the global and partial points on the first principal plane. The partial points with the highest differences on the first dimension (indicating a chronological gap) are underlined. Two recent works (noted by 4 and 5) are related to former topics (drugs and symptoms). Three works (noted by 1, 2 and 3) can be considered as pioneering works, as they are dedicated to a genetic approach at a date as early as 1996.

Figure 1. Global representation

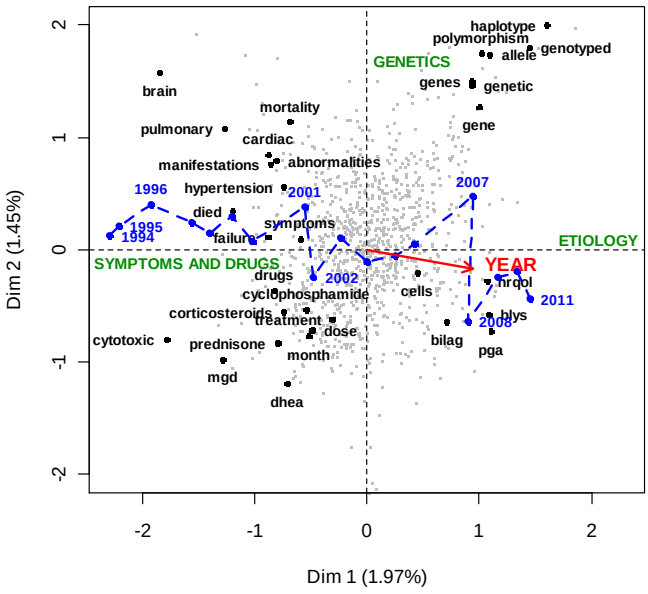
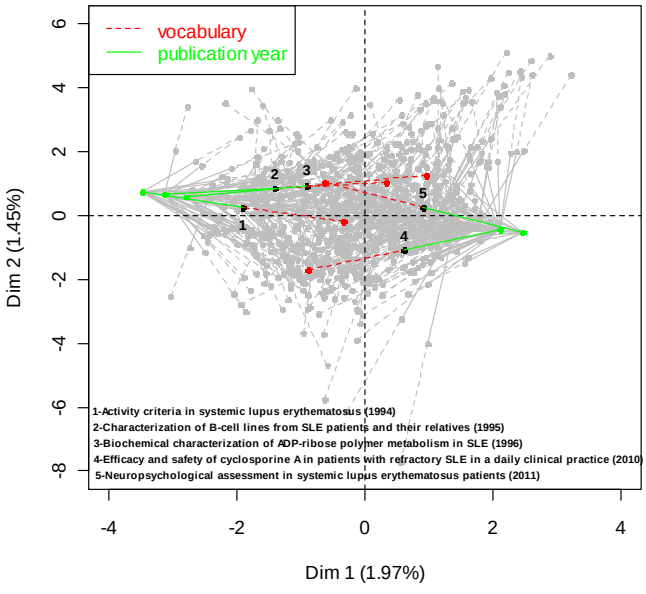


Figura 2. Superimposed representation



Références

[1] Lê, S., Josse, J., Husson, F. (2008). Factominer: An r package for multivariate analysis. *Journal of Statistical Software*, **25**(1), 1–18

[2] Escofier, B., Pagès, J. (1990). *Analyses factorielles simples et multiples: objectifs, méthodes, interprétation.* Dunod, Paris

[3] Bécue-Bertaut, M., Pagès, J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, **45**, 481–503

[4] Bécue-Bertaut, M., Pagès, J. (2008). Multiple factor análisis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis,* **52**, 3255–3268