Analysis of cross-language open-ended questions through MFACT

Mónica Bécue[1], Jérôme Pagès[2], and Campo-Elías Pardo[3]

[1] EIO. Universitat Politècnica de Catalunya - 08028 Barcelona - Spain
monica.becue@upc.es
[2] ENSA/INFSA - 65 rue de Saint-Brieuc, CS 84215- F-35042 Rennes Cedex - France jerome.pages@agrorennes.educagri.fr
[3] Departamento de Estadística. Universidad Nacional de Colombia - Bogotá - Colombia cepardot@unal.edu.co

1 Introduction

One of the reference methodologies to analyse open-ended questions is correspondence analysis (CA) (Lebart et al., 1998), applied to individual lexical tables (individual answers × words tables) or to aggregated lexical tables (category-documents × words tables), built up by gathering the answers of the individuals belonging to the same category (e.g. young women, etc.).

International surveys lead to deal with cross-language open-ended questions. A solution could be to translate the whole of the answers into one of the languages, but it presents two different natured drawbacks. This solution is costly, on the one hand, and important features and nuances can be destroyed by translation, on the other hand.

The methodology that we propose, multiple factor analysis for contingency tables (MFACT) (Bécue and Pagès, 2004), tackles the whole of the responses without any translation, operating from the various aggregated lexical tables, one by country, which can be juxtaposed row-wise when using the same categories. The goal consists in analyzing this multiple contingency table in order to describe the categories from a global (here international) point of view but also from each country point of view, looking for structures common to all the countries or specific to only some of them. MFACT, an extension of multiple factor analysis (MFA), can be seen as a particular multicanonical method.

Section 2 shows the close connection existing between MFA and Carroll's generalized canonical correlation analysis. In section 3, after introducing the notation (Sect. 3.1), CA is presented as a particular principal component analysis (Sect. 3.2) and the main properties of MFACT applied to lexical tables are exposed (Sects. 3.3 and 3.4). Section 4 offers some results obtained from an international survey.

2 MFA as a multicanonical analysis

Among multicanonical methods, Carroll's generalized canonical correlation analysis (Carroll, 1968) and multiple factor analysis MFA (Escofier and Pagès, 1998) adopt a similar strategy. Both methods deal with a multiple table individual×quantitative variables $\mathbf{X}$ of order $I \times J$, where the variables are divided into groups K_t containing the variables $v_{j,t}$ $(t = 1, \cdots, T;\ j = 1, \cdots, J_t;\ \sum\limits_{t=1}^{T} J_t = J)$, and search for a series of S not correlated latent general variables, z_s, related as much as possible to the T groups K_t of variables. Then, for each general variable z_s, they look for their representatives in every group, called canonical variables, linear combination of the variables of the group having the maximum relationship with the corresponding general variable. The difference between these methods comes from the measure of the relationship between a variable and a group of variables.

In Carroll's generalized canonical correlation analysis, the relationship between a quantitative variable z and the set of variables of group K_t is the cosines of the angle between z and the subspace generated by the variables of this group. In MFA (Pagès and Tenenhaus, 2001), the relationship between a quantitative variable z and the set of variables of group K_t is the weighted total inertia of the variables of this group in the direction of z as given by (1):

$$\mathcal{L}_g\left(z, K_t\right) = \sum_{j=1}^{J_t} m_t cor^2\left(z, v_{j,t}\right) \tag{1}$$

with $m_t = 1/\lambda_1^t$ being λ_1^t the first eigenvalue of the separate principal component analysis of group K_t. The weights m_t balance the influence of the groups of variables in the determination of the first general variable. Using this relationship measure, the general variables z_s is thus defined by $\sum\limits_t \mathcal{L}_g\left(z_s, K_t\right)$ maximum with the constraints $Var\left(z_s\right) = 1$ and $Corr\left(z_s, z_t\right) = 0\ \forall t \neq s$.

Both methods can also be seen as principal component analysis with specific metrics. The general variables are the standardized principal components.

Carroll's generalized canonical correlation analysis can also be performed as a PCA applied to the global table $\mathbf{X}$, using as metric the bloc diagonal matrix composed of the inverses of the variance-covariance matrices internal to every group K_t. In the case of MFA, from reexpressing $\mathcal{L}_g\left(z_s, K_t\right)$ and $\sum\limits_t \mathcal{L}_g\left(z_s, K_t\right)$ respectively as $\mathcal{L}_g\left(z_s, K_t\right) = \frac{1}{I^2}\mathbf{z}_s^{'}\mathbf{X}_t\mathbf{M}_t\mathbf{X}_t^{'}\mathbf{z}_s$ and $\sum\limits_t \mathcal{L}_g\left(z_s, K_t\right) = \frac{1}{I^2}\mathbf{z'}_s\mathbf{XMX'}\mathbf{z}_s$, it can be deduced that the set of variables z_s are the standardized principal components ($\mathbf{z}_s = \frac{\mathbf{F}_s}{\sqrt{\lambda_s}}$) of global table $\mathbf{X}$ using matrix $\mathbf{M}$ as weights for columns (and metric in individuals space). $\mathbf{M}$ is the diagonal matrix of order J×J, composed by T diagonal submatrices, of order J_t× J_t, with term m_t repeated J_t times (Pagès and Tenenhaus, 2001).

3 MFACT

3.1 Notation

When considering only one country, a lexical table $\mathbf{X}$, of order $I \times J$, is built up. Its general term f_{ij} is the relative number of occurrences of the word j ($j = 1, \cdots, J$) in the whole of the answers of category i ($i = 1, \cdots, I$) such that $\sum\limits_{ij} f_{ij} = 1$. $\mathbf{D}_I$ is the diagonal matrix with general term $f_{i.} = \sum\limits_{j} f_{ij}$, ($i = 1, \cdots, I$). $\mathbf{D}_J$ is the diagonal matrix with general term $f_{.j} = \sum\limits_{i} f_{ij}$, ($j = 1, \cdots, J$). When dealing with several countries, T lexical tables $\mathbf{X}_t$, of order $I \times J_t$, are built up and juxtaposed row-wise to form global table $\mathbf{X}_G$ of dimension $I \times J$. The general term f_{ijt} is the proportion, in table t ($t = 1, \cdots, T$), with which category i ($i = 1, \cdots, I$) is associated with word j ($j = 1, \cdots, J_t$; $\sum\limits_{t} J_t = J$). $\sum\limits_{ijt} f_{ijt} = 1$. We denote the row margin of table $\mathbf{X}_G$ as $f_{i..} = \sum\limits_{jt} f_{ijt}$ and the column margin of table $\mathbf{X}_G$ as $f_{.jt} = \sum\limits_{i} f_{ijt}$. The row margin of table t, as a subtable of table $\mathbf{X}_G$, is $f_{i.t} = \sum\limits_{j} f_{ijt}$, and the sum of the terms of table t inside table $\mathbf{X}_G$ is $f_{..t} = \sum\limits_{ij} f_{ijt}$. $\mathbf{D}_{I_T}$ is the diagonal matrix with general term $f_{i..}$ and $\mathbf{D}_{J_T}$ is the diagonal matrix with general term $f_{.jt}$.

3.2 CA as a Principal Component Analysis

To apply classical CA to table $\mathbf{X}$ is equivalent to perform a principal component analysis (Escofier and Pagès, 1998, pp. 95-97) on the table $\mathbf{W}$ whose general term is given by (2), using $\mathbf{D}_I$, as row weights and metric in the column space, and $\mathbf{D}_J$, as column weights and metric in the row space.

$$w_{ij} = \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}} \tag{2}$$

3.3 MFACT as a specific PCA

MFACT applied to table $\mathbf{X}_G$ consists in a PCA on the table $\mathbf{Y}$ whose general term is given by (3), using $\mathbf{D}_{I_T}$ as row weights and metric in the column space and $\mathbf{D}_{J_T}$ as column weights and metric in the row space.

$$y_{ijt} = \frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) f_{.jt}}{f_{i..} f_{.jt}} = \frac{1}{f_{i..}} \left[\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}} \right] \tag{3}$$

The global principal components are $\mathbf{F}_s = \frac{1}{\lambda_s} \sum\limits_{t} \mathbf{Y}_t \mathbf{D}_{J_t} \mathbf{Y'}_t \mathbf{D}_{I_T} \mathbf{F}_s$. The general variables are the standardized global principal components $\mathbf{z}_s = \frac{\mathbf{F}_s}{\sqrt{\lambda_s}}$. The canonical components $\mathbf{F}^t_s$, associated with the global principal components $\mathbf{F}_s$ in each group t, are defined by $\mathbf{F}_s = \sum\limits_{t} \mathbf{F}^t_s$ and then $\mathbf{F}^t_s = \frac{1}{\lambda_s} \mathbf{Y}_t \mathbf{D}_{J_t} \mathbf{Y'}_t \mathbf{D}_{I_T} \mathbf{F}_s$.

3.4 Application of MFACT to cross language surveys: main features

Global and partial documents. The whole of the answers corresponding to category i ($i = 1, \cdots, I$) through the different countries form the global category-document i, characterized through the coordinates given by (4); the weighting by $\frac{1}{\sqrt{\lambda_1^t}}$ balances the importance of each country in the global document.

$$\frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) f_{.jt}}{f_{i..} f_{.jt}\sqrt{\lambda_1^t}} = \frac{1}{\sqrt{\lambda_1^t} f_{i..}} \left[\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}}\right] (t = 1, \cdots, T; j = 1, \cdots, J_t) \tag{4}$$

The *partial* document i^t consists in the whole of the answers of category i but only in country t. Its coordinates are those given by (4) restricted to t.

MFACT analyzes the global-document×words table and defines distances between global documents and between words. In MFACT, each partial document i^t is considered as a supplementary row by completing the columns j, r ($r \neq t$) by zeroes and projected on the global axes. In this way, the representations of the partial documents and of the global documents (or average documents) are superimposed. So, the relative positions of the separate documents corresponding to a same global document can be studied.

Distances between global documents. The squared distance between documents i and l, calculated from coordinates given in (4) is:

$$d^2(i,l) = \left[\sum_t \frac{1}{\lambda_1^t} \sum_{j \in J_t} \left(\frac{f_{ijt}}{f_{i..}} - \frac{f_{ljt}}{f_{l..}}\right)^2 \frac{1}{f_{.jt}}\right] - \left[\sum_t \frac{1}{\lambda_1^t f_{..t}} \left(\frac{f_{i.t}}{f_{i..}} - \frac{f_{l.t}}{f_{l..}}\right)^2\right] \tag{5}$$

In expression (5), disregarding weighting by the reverse of the first eigenvalue, the first term corresponds to the distance (between profiles i and l) in the CA of the juxtaposed tables. The second term corresponds to the distance (between profiles i and l) in the CA of the table containing the sums by row and by subtable. The general term $i.t$ in this table is the sum of row i in table t. We see here how this last table is neutralized by recentering each subtable on its own margins.

Distances between words. The squared distance between word j (belonging to table t) and word k (belonging to table r) is given by 6.

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i\cdot\cdot}} \left[\left(\frac{f_{ijt}}{f_{\cdot jt}} - \frac{f_{ikr}}{f_{\cdot kr}}\right) - \left(\frac{f_{i\cdot t}}{f_{\cdot\cdot t}} - \frac{f_{i\cdot r}}{f_{\cdot\cdot r}}\right)\right]^2 \tag{6}$$

Case 1: the words belong to a same table ($t = r$). The proximity between two words is interpreted in term of resemblance between profiles, as in CA.

Case 2: the words belong to different tables ($t \neq r$). Equation (6) shows how the differences between word profiles are relativized by the differences between average profiles.

Distributional equivalence property. The distance between words and between documents induced by MFACT conserves the distributional equivalence property. It carries along that, as in CA applied to one lexical table, gathering synonyms (in terms of a synonyms dictionary) changes neither the distances between words nor the distances between documents when they have the same profiles; but such gathering does change these distances if the synonyms profiles are different. That suggests avoiding this data transformation and claims for operating without any translation which could collapse different words with different profiles in a unique column-word.

Transition formulae

Documents among words. The relation giving (along the s-axis) the coordinate $F_s(i)$ of global document i from the coordinates $\{G_s(j,t);\ j = 1,\ldots,\ J_t;\ t= 1,\ldots,\ T\}$ of words is:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_t \frac{1}{\lambda_1^t} \frac{f_{i.t}}{f_{i..}} \left[\sum_{j \in J_t} \frac{f_{ijt}}{f_{i.t}} G_s(j,t) \right] \tag{7}$$

Except for a constant, each category-document lies in the centroid of the words associated with this document. Globally, a document is attracted by the words with which it is associated.

Partial documents among words. The superimposed representation of the partial documents benefits from CA properties. In particular, these partial representations can be related to word representation by means of a “restricted” transition formula:

$$F_s^t\,(i) = \frac{1}{\sqrt{\lambda_s}} \frac{f_{i.t}}{f_{i..}} \left[\sum_{j \in J_t} \frac{f_{ijt}}{f_{i.t}} \frac{G_s\,(j,t)}{\lambda_1^t} \right] \tag{8}$$

In the graph superimposing partial representations (see Fig. 3 in Sect. 4), the coordinates of the partial documents are dilated by the coefficient T (number of tables). Thus, a global document point is located in the centroid of the corresponding partial document points.

Words among documents. The expression (along the s-axis) for the coordinate $G_s(j,t)$ of word j,t from the coordinates $\{F_s(i),\ i{=}1,\ldots,I\}$ of documents is:

$$G_s(j,t) = \frac{1}{\sqrt{\lambda_s}} \left[\sum_i \left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}} \right) F_s(i) \right] \tag{9}$$

As the coefficient of $F_s(i)$ can be negative, the words are not in the centroid of the documents, except when the document weights are the same in all the tables. This coefficient measures the discrepancy between the profile of words j,t and the column margin of table t. A word is attracted (or repelled) by the documents that are more (or less) associated with it than if there was independence between documents and words in table t.

4 Example

The data are extracted from a large international survey (Hayashi et al., 1992)[4]. People from four countries (Great Britain, France, Italy, Japan) are asked several closed questions and, moreover, the open-ended question: *"What is the most important thing to you in life?"* is considered. The Japanese answers are romanized.

In each country, the free answers are grouped into 18 category-documents by crossing gender (male, female), age (into three categories: 18-34, 35-44, 55 and over) and education level (into three categories: low, medium and high). Then, for each country, from the count of words in the whole answers, the lexical table arises by crossing the 18 documents and the most frequent words. Only the words used at least 20 times are kept.

4.1 Results obtained performing MFACT

Visualization of the global documents. The visualization of the documents obtained by MFACT is given by Fig. 1. On this figure, the 6 trajectories of age intervals are drawn; they show a rather regular structure, compromise between the representations that would have been offered by the separate CA. Age increases along the first axis, and the second axis opposes the genders. The categories with the high education degree have, on the first axis, coordinates which correspond to younger people with lower degrees.

Visualization of the words. Figure 2 shows an excerpt of the representation of the words. We can see, for example, that the word *husband* and its translation in French (*mari*) and Japanese (*shuzin* and *otto*) are rather quoted by the same categories.

Superimposed representation of the partial documents. In order to compare the structures induced on the documents by the four sets of words, we superimpose the global description of the documents and those induced by each country (partial documents). We can interpret the relative positions of partial documents of a same country: for example, Fig. 3 suggests that males and females in 35-54 interval, whatever the qualification, almost do not differ in

[4] We thank Profesor Lebart to put at our disposal this data set.

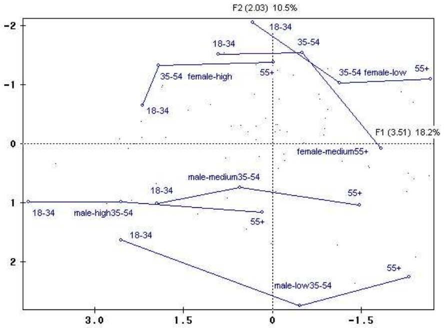

Fig. 1. Representation of the global documents

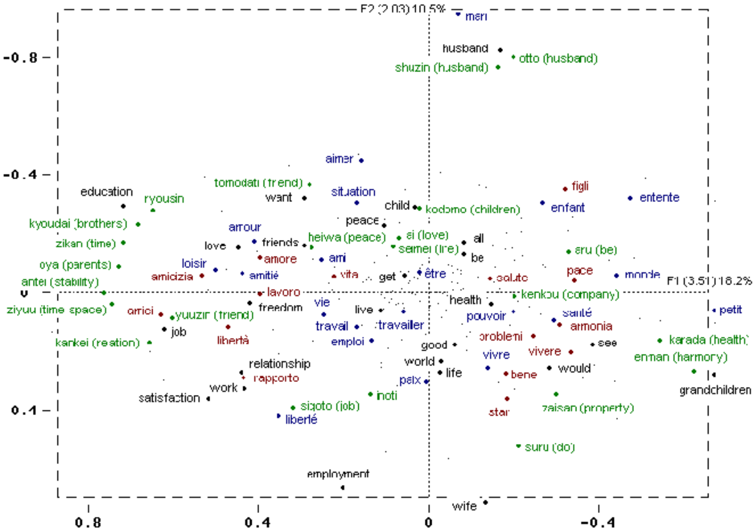

Fig. 2. Excerpt representation of the words

Italy. We can also interpret the relative positions of partial documents corresponding to a same global document: Fig. 3 suggests that males in 35-54 interval with high degree or with only medium degree almost use the same vocabulary in United Kingdom but, on the contrary, greatly differ in Japan.

5 Conclusion

MFACT allows for adopting a direct multicanonical approach to analyze multiple contingency table. Its application to cross language aggregated lexical tables presents interesting properties such as to locate the whole of the words in a same representation space, on the one hand, and all the documents in

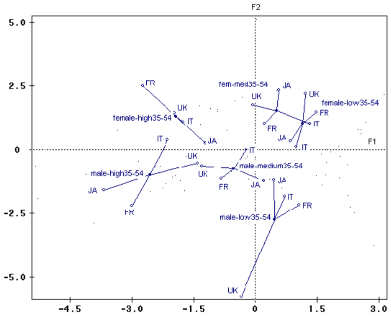

Fig. 3. Excerpt of the superimposed representation: Global and partial documents corresponding to all the categories between 35 and 54 years old

a same representation space, on the other hand, being both spaces linked through transition formulae. So, the similarities between documents can be interpreted in terms of semantics and the similarities between words in terms of users'profiles.

Acknowledgements

This work has been supported by the European project NEMIS (IST- 2001 - 37574), by the Spanish Science and Technology Ministery (SEC2001-2581-C02-02) and by a grant from National University of Colombia - Bogotá.

References

Bécue, M. and Pagès, J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Comp. Statistics & Data Analysis* (in press, available online in Science Direct).

Carroll, J. (1968). Generalization of canonical correlation analysis to three or more set of variables. *Proceedings of the 76th Convention of the American Psychology Association* **3** 227–228.

Escofier, B. and Pagès, J. (1998). *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*. Dunod.

Hayashi, C., Suzuki, T. and Sasaki, M. (1992). *Data Analysis for Social Comparative Research: International Perspective*. North Holland.

Lebart, L., Salem, A. and Berry, E. (1998). *Exploring Textual Data*. Kluwer.

Pagès, J. and Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemometric Intell.Lab.Syst.* **58** 261–273.

Index