Comparison of different panels sorting tasks with hierarchical multiple factor analysis

Mónica Bécue[1], Berenice Colmenares[1], Sébastien Lê[2]

[1] *Universitat Politècnica de Catalunya*
Departament Estadística i Investigació Operativa
c/ Jordi Girona 1-3. 08034 Barcelona (Spain)
E-mail:monica.becue@upc.edu, berenice.colmenares@estudiant.upc.edu
[2] *AGROCAMPUS OUEST - Centre de Rennes.*
Laboratoire de Mathématiques Appliquées
65, rue de St-Brieuc; CS 84215; 35042 Rennes Cedex (France)
E-mail : sebastien.le@agrocampus-ouest.fr

Abstract: Hierarchical multiple factor analysis is a suitable tool for giving account of the evaluation of a same set of items by hierarchically structured sets of individuals. This method is applied to compare trained and non-trained panels in wine hall tests. Every panellist has to categorize the wines in clusters, describe them with free descriptive words and also give a hedonic score. Data coding leads to a wine × individual evaluation table in which the columns present a hierarchical structure. Hierarchical multiple factor analysis allows for exploring and visualizing the observed variability among both wines and panellists. Visualizing tools are also offered to evaluate the similarity between panels and sets of panels.

Keywords: Multiple table; Multiple factor analysis; Hierarchical multiple factor analysis; Wine hall tests; Sensometrics; Free-text description.

1 Introduction

The judgement of a set of items by a set of individuals is a very common situation in different fields: in sensory studies, panels of experts value food products, perfumes or industrial textiles; in socioeconomic studies, individuals give their opinion about society problems, government actions and/or value political leaders. In some cases, the individuals are gathered in nested partitions. We expose here how hierarchical multiple factor analysis (HMFA) (Le Dien & Pagès 2003a, 2003b) allows for a very flexible approach to this kind of data and eases the exploration, visualization and interpretation of the observed variability among both items and individuals.

In Section 2, we present the data collection. In Section 3, we set out their coding and notation. In Section 4, we recall HMFA principles and show how this method allows for a profitable approach to complex data. Section 5 offers some results and we conclude in Section 6.

2 Data collection and objectives

In the framework of a project meant to study the differences between assessments of a same set of wines by panels issued from different countries, three wine-tasting hall tests have been planned. Successively, a Catalan expert panel (C, 9 panellists), a French expert panel (F; 15 panellists) and, as to

contrast, a Catalan amateur panel (A; 10 panellists) have tasted eight Catalan wines. These eight wines correspond to the combinations of 3 factors: variety of wine (Grenache or Samsó), region (Priorat or Ampurdan, both in Catalonia) and production year (2005 or 2006). For the wine presentation, suitable designs have been used to balance the order and first-order carry-over effects. By performing a free sorting task, the panellists have categorised the wines in as many clusters as perceived, from two to a maximum of seven, and then described either the wines or the clusters by some words. The trained panellists have also qualified the wines with a hedonic score (from 0 to 10).

In this work, the main objective is to present the tools provided by HMFA to compare the panels, in particular the trained and non-trained panels, by computing both global similarities between panels and individual similarities between panellists.

3 Data coding and notation

Data coding is relevant as it strongly impacts on the results (Murtagh, 2005) and conditions the methodology. As usual in sensory analysis, the item-products (here, wines) are considered as the statistical units (rows of the table). The categorisation of the wines performed by each panellist is stored as a categorical variable, presenting as many categories as clusters, adopting thus the point of view presented in Lê, Cadoret & Pagès (2008). Equivalently, this categorical variable can correspond to one column (condensed coding, adopted in this work) or to as many columns as categories (complete disjunctive coding) as usual in multiple correspondence analysis (MCA). The categorical table so obtained is completed by the mean of the scores given by the trained panels (2 quantitative columns, one per panel) and the free description of the wines coded through a products × words frequency table. The words relative to a same characteristic, such as *amer* (bitter) and *amertume* (bitterness) are gathered and only those words whose frequency is over 3 are kept (31 Catalan and 58 French words). The resulting multiple table is presented in Figure 1.

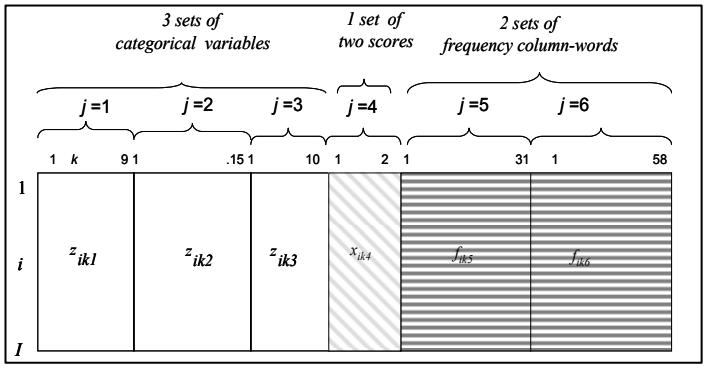

Fig. 1: Multiple data table

The objectives lead to consider a hierarchy on the set of categorical variables/partitions induced by the panellists. We want to compare a) the trained versus the non-trained panels b) the Catalan and French assessments c) the panellist behaviours. Thus, we have to balance trained and non-trained panels, Catalan and French panels within the trained panels set and, finally, the panellists' contributions within each of the three panels. Figure 2 summarises the nested partitions corresponding to this strategy.

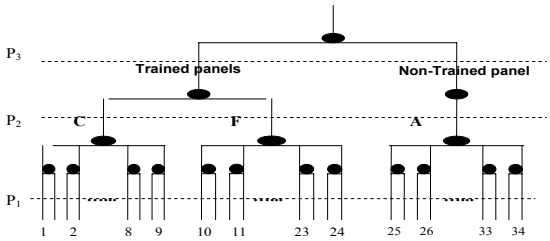

Fig. 2: The three nested partitions on the categorical columns

4 Methodology

We have to deal with a multiple table made of category columns structured according to a hierarchy. If it were not for this column structure, multiple correspondence analysis (MCA) would be a suitable method. Hierarchical multiple factor analysis (HMFA; Le Dien & Pagès, 2003a, 2003b) allows for both keeping a MCA-like approach and balancing the roles of the column sets at every level of the hierarchy. HMFA is an extension of multiple factor analysis (MFA). Thus, we first recall the principles of the latter method.

MFA analyses a multiple table in which a set of individuals is described by J sets of variables, quantitative or categorical. The method can be seen as a specific non-standardized principal component analysis (PCA) applied to the juxtaposed table, but overweighting the variables to balance the influence of the different sets in the determination of the first axis. For that purpose, the weight of the column variables of the set j is divided by λ_1^j, first eigenvalue obtained in the separate analysis–PCA or MCA depending on the type of variables– of the subtable j. Results offered by this method are:

- analogous to those of PCA or MCA, mainly a global representation of the rows (individuals) and columns (variables or categories);
- specific to multiple tables such as a synthetic visualisation of the sets of columns..

HMFA sticks to the general principles of MFA while taking into account a hierarchical structure on the columns. From the bottom up, HMFA scales the node weights, the way MFA does, at every level of the hierarchy in order to balance their influence. In our case, that means to place on the same footing all the one-column panellists sets (first level), then the three panels (second level) and, finally, both trained panels, on the one hand, and the non-trained panel, on the other hand (third level). HMFA can be seen as a weighted non-standardised PCA applied to the juxtaposed table. HFMA offers a representation of the sets of columns involved in the hierarchy by giv-

ing to every set a coordinate equal to L_g index that measures the relationship between the set and the axis (Escofier & Pagès, 1983-2008). This index ranges between 0 – no relationship between the axis and any category belonging to the set – and 1–the axis is equal to the axis computed in the separate MCA of the set. As in PCA, supplementary information can be used such as the hedonic scores and the free description of the wines.

5 Data analysis

HMFA is applied to the three nested partitions (Figure 2). A global representation of the wines, that is, from the whole of the weighted columns, is presented in Figure 3.a.

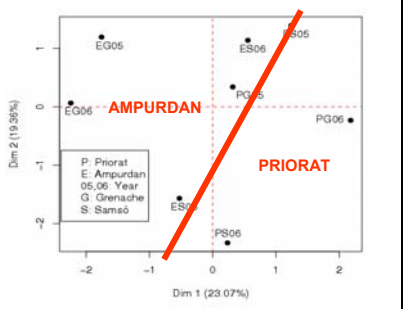

Fig.3.a: Representation of the row-wines

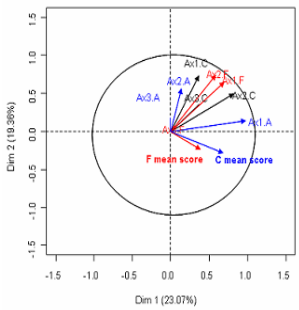

Fig.3.b: Correlations between global and separate axes/mean scores

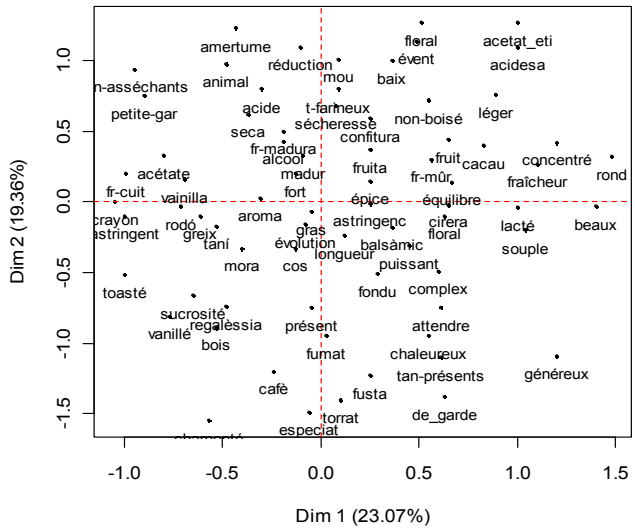

Fig.3.c: Excerpt of the representation of the column-words

Fig. 3: First principal plan issued from HMFA

The first two dimensions express 23.07% and 19.36% of the variation. As the categories do not reflect any quality (they are only the cluster number), the interpretation is mainly supported by the supplementary variables, that is, the scores given by the trained panellists and the words associated either to the wines or to the clusters but also the wine characteristics as known by the labels of the wines.

As in PCA, the scores are positioned on the axes by their correlation coefficients. Every word *(k,j)* is positioned on every axis at the weighted average of the wines that they describe, giving to every wine *i* the weight $f_{ikj}/f_{.kj}$ (Figure 3.c). Figure 3.b shows that the first axis is closely related to the mean scores (*corr* (axis, Catalan mean score)=0.82; *corr* (axis, French mean score)=0.75). It opposes wines that are perceived as *astringent* and *mouth drying* with a *vanilla* taste to those considered as *flexible*, *generous*, *concentrated* or *fresh* wines. The second axis opposes the wooden to the non-wooden wines. This debated and somewhat polemic characteristic is almost orthogonal to the scores: the wooden character of the wine can be successful, and the wine is *generous*, *warm* and *complex*, or not; in the latter case, the wine is perceived as *fatty*, *astringent* and *too tannic*.

If it were not for wine PG05 –very particular according to its chemical characteristics – Priorat and Ampurdan wines lie in differentiated zones on the first principal plane. This feature appears more clearly when only the trained panels are taken into account (study not reproduced here). Wine variety influence is not very strong, although close wines generally belong to the same variety –except for PG05. The production year is not relevant.

Figures 4.a and 4.b show that the non-trained and the trained panels do not behave similarly. Figure 4.a shows that the first dimension is slightly more related to the non-trained panel –in accordance with the correlations between the first global axis and the first separate axes (Figure 3.b). The trained panels seem to take into account the characteristics of the wines that are opposed on the second axis. The non-trained panel adopts a strategy more linked to a global appreciation of the quality level. Figure 4.b shows the great diversity existing among the individuals integrating a same panel. Nevertheless, the non-trained panelists present a high homogeneity in their relationship with the first axis. They seem divided into two subgroups, one of them more linked to the second axis and then closest to the trained panels.

6 Conclusions

HMFA deals with tables presenting a hierarchy structure on the columns. This method balances the influence of the sets, eventually of different types, at every level of the hierarchy. It provides representations of the rows but also of the columns and sets of columns. In the case of the application to wine hall tests, the latter type of representation has been used to represent the whole of the panellists. Thus, the variability of the wines as well as the variability of the panellists are visualised. The free description of the wines, coded through frequency tables, has resulted to be indispensable to give account of the criterions used in the wine categorisation. In particular, this de-

scription allows for interpreting the differences between the trained and non-trained panels, being the latter more influenced by the global quality of the wines. Nevertheless a subgroup of the non-trained panellists seems to have criterions more similar to the trained panellists.

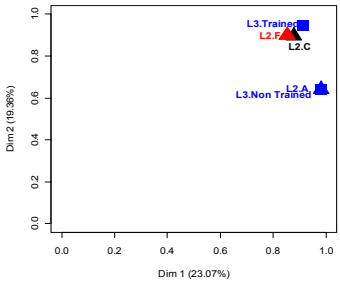

Fig. 4.a: Trained vs non-trained panels. Catalan vs French panels

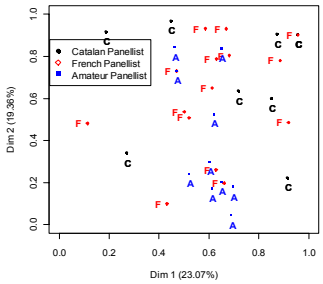

Fig. 4.b: Panellists

Fig.4: Synthetic representation of the sets at the different levels of the hierarchy

Acknowledgement

This work has received support from the Spanish Ministry of Science and Innovation and FEDER (grant ECO2008-01223/ECON). We acknowledge the *Associació Catalana d'Enòlegs (*Catalonia*,* Spain), the *Conseil Interprofessionnel des Vins du Roussillon (*Perpignan, France*)*, the *Parc Científic de les Industries Enològiques* (Falset, Catalonia), the *Institut Català de la Vinya i el Vi* (INCAVI, Catalonia) and the *Facultat de Matemàtiques i Estadística* (Barcelona, Catalonia*)* for their contribution to the hall tests.

Software

FactoMineR package, available in R (Lê, Josse & Husson, 2008).

References

Escofier, B., Pagès, J. 1983-2008. *Analyses factorielles simples et multiples*. Dunod, Paris.

Le Dien, S., and Pagès, J. 2003a. Analyse Factorielle Multiple Hiérarchique, *Revue de Statistique Appliquée* LI (2): 47-73.

Le Dien, S., and Pagès, J. 2003b. Hierarchical Multiple Factor Analysis: application to the comparison of sensory profiles, *Food Quality and Preference* 14: 397-403.

Lê, S., Cadoret M., and Pagès, J. 2008. A novel Factorial Approach for Sorting Task data. *First joint meeting of the SFC and the CLADAG*. Caserta (Italy), 11-13 juin 2008.

Lê, S., Josse, J., and Husson, F. 2008. FactoMineR: an R package for multivariate analysis, *Journal of Statistical Software* 25 (1): 1-18.

Murtagh, F. 2005. *Correspondence analysis and Data Coding with Java and R*. Chapman & Hall, Boca Raton.