UNIVERSITAT POLITÈCTICA DE CATALUNYA

PHD PROGRAM IN STATISTICS AND OPERATIONS RESEARCH

A Principal Component Method to Analyse Disconnected Frequency Tables by Means of Contextual Information

Author:
Belchin Adriyanov KOSTOV

Advisor:
Mónica Bécue BERTAUT

Co-advisor:
François HUSSON

Thesis submitted for the degree of Doctor in Statistics and Operations Research, Universitat Politèctica de Catalunya

July 2015

A Principal Component Method to Analyse Disconnected Frequency Tables by Means of Contextual Information

by

Belchin Adriyanov KOSTOV

Examination panel:

Montserrat Guillén Estany,
Departament d'Econometria, Estadística i Economia Espanyola,
Universitat de Barcelona, Spain

Fionn Murtagh,
Department of Computing and Mathematics, University of Derby, UK
Department of Computing, Goldsmiths University of London, UK

Ignacio García Lautre,
Departamento de Estadística e Investigación Operativa,
Universidad Pública de Navarra, Spain

External referees:

Fionn Murtagh,
Department of Computing and Mathematics, University of Derby, UK
Department of Computing, Goldsmiths University of London, UK

Julie Josse,
Applied Mathematics Department, Agrocampus Rennes, France
Department of Statistics, Stanford University, USA

Degree: PhD in Statistics and Operations Research

Thesis defence date: July 2015

To my loving angels, mom, grandma, my sister and my fiancée

Acknowledgements

The accomplishment of a thesis is a long journey to which many people contribute somehow. I would like to thank all these people who helped me to reach to the end of this thesis.

First and foremost, I would like to express my gratitude to my thesis advisors, Mónica Bécue and François Husson. Professor Mónica Bécue encouraged me to realise this thesis. Her tremendous dedication during my PhD, her supervision and support deserve many words of thanks. I can say the same for Professor François Husson. His guiding and his help related to programming aspects have been essential for this thesis. During the time that I spent in Agrocampus (France), I learned a lot except French. It was a pleasure to have you as my advisors.

It is not easy to work and to do a PhD at the same time. But, of course, it would be even harder without the support and the understanding of Dr. Antoni Sisó Almirall. More than a boss, he was a colleague from whom I learned a lot about research. My very special thanks go to Dr. Manel Ramos Casals. I learned from him that a statistician and a physician can work together and do a great job. Thanks to them, I found my vocation that is research.

I appreciate the support of my friends, Deniz, Tamer, Raúl, Adri and Dani. I would like to take this opportunity to thank as well my collegues from Department of Statistics and Operations Research, especially to Daría Hernández for her endless support and to Jordi Cortés for his unconditional help.

To realise a thesis requires an important knowledge gained at the university. I would like to thanks all my professors from School of Mathematics and Statistics of the Universitat Politècnica de Catalunya (UPC). Among many, especially I would like to thank to Klaus Langohr for his great courses about R statistical software and to Lidia Montero for her positive energy. I would also like to thank to all professors who evaluated my thesis proposal and gave me great advices that helped to improve the thesis.

It has been a real pleasure to be part of the Department of Statistics and Operations Research of UPC. I would like to thank to Jordi Castro, head of the PhD program, for his

management of all the things related to the doctorate program, to Carme Macias for her help on academic administration issues and to Toni Font for computer technical support.

There are other people who deserve words of appreciation. Thanks to Jordi Torrens and Pilar Urpi for collaborating on my thesis by means of the agreement between Freixenet and UPC, to Danièle Castin for English correction of papers, to Annie Morin for her support to the birth of this thesis, Vincent and Tam for their friendliness and finally to Fionn Murtagh and Julie Josse for the time spent on reading of my PhD dissertation and for the valuable comments which helped to improve the final document.

And the last, I would like to thank to my parents for their moral and economic support during many years. None of this would have been possible without their decision to leave everything behind and start from the beginning in a new place to be able to provide better opportunities to their children. Also I would like to thank to my grandfather, my grandmother, my uncle and my aunt for their moral support. I would like to thank especially to my sister, Melek. Thank you for standing by my side, particularly at the hardest moments, for supporting me through every decision and for advising me. I'm happy to have a sister like you. The journey was long but I would not have arrived to the end without you. And Nihan, my fiancée, you have been my biggest motivation during the last year. In the most difficult moments, you made me believe that I could do and I did. Thanks for making me believe that I could overcome the difficulties, sharing long hours of study in the university with me and giving a meaning to everything.

Abstract

This thesis arises from the need to deal with open-ended questions answered in different languages in international surveys. For every language, the free answers are encoded in the form of a individuals $\times$ words lexical table. An important feature is that the lexical tables, from one language to the other, have neither the row-individuals nor the column-words in common. However, the global analysis and the comparison of the different samples require to place all the words, in any language, in the same space.

As a solution, we propose to integrate the answers to the closed questions into the analysis, where the contextual variables the same for all the samples. This integration plays an essential role by permitting a global analysis. Thus, for every language, we have one lexical table and one categorical/quantitative table, a structure that we call "coupled tables". The global complex data structure is a sequence of "coupled tables".

To analyse these data, we adopt a Correspondence Analysis-like approach. We propose a method which combines: Multiple Factor Analysis for Contingency Tables, in order to balance the influence of the sets of words in the global analysis and Correspondence Analysis on a Generalised Aggregated Lexical Table, which places all the words in the same space.

The new method is called Multiple Factor Analysis on Generalised Aggregated Lexical Table. The results in an application show that the method provides outputs that are easy to interpret. They allow for studying the similarities/dissimilarities between the words including when they belong to different languages as far as they are associated in a similar/different way to the contextual variables. The methodology can be applied in other fields provided that the data are coded in a sequence of coupled tables.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

In social sciences data often come from different sources. For example, when a survey is performed in different countries, very specific problems arise, from obtaining equivalent questionnaires in the different languages. These surveys often include open-ended questions as a complement to classical closed questions to address complex and little known topics. However, the global analysis of this data requires to deal with a complex data structure that we will detail later.

This thesis aims at dealing with open-ended questions answered in different languages in international surveys. The general objective is to compare the free answers in different languages relating the samples of individuals and vocabularies, which can be summarised by the following questions: which are the similarities between words in the same language? Are the homologous words in different languages similarly used? Which are the similarities between languages from the point of view of the main topics that appear in respondents' free answers?

International survey by questionnaires is not the only field where such problems occur. Some sensory hall tests, including a verbalization task, have been proposed, with the intention of characterizing several regions' products (foods or beverages). In each region, the sensory panel integrates different panellists who use different vocabularies. However, in this case, another important problem exists. As discussed in Delarue and Sieffermann (2004), the ability of the panellists to communicate their sensory perceptions using a common base of vocabulary is sometimes doubtful. Moreover, the panellists understand and perceive some

words in a different way and thus associate them with different products (Cadoret et al., 2009; Veinand et al., 2011).

We turn back to our initial problem. The first issue to deal with, in the analysis of an open-ended question answered in different languages, is the data encoding. The free answers captured from open-ended question are encoded in the form of an individuals $\times$ words frequency table, called a lexical table (LT; Lebart et al., 1998) in textual analysis. Thus, we consider that we have L lexical tables, one lexical table per language. An important feature is that the L lexical tables have neither the row-individuals nor the column-words in common. However, the global analysis and the comparison of the different samples require to place all the words, in any language, in the same space. We propose to integrate the information of some closed questions into the analysis. The closed questions are encoded in an individuals $\times$ contextual variables table. As opposed to lexical tables, the contextual variables are common to all the samples. That plays an essential role by permitting a global analysis of all the lexical tables.

If we consider $\mathbf{Y_l}$ the individuals $\times$ words lexical table for language l $(l = 1, ..., L)$ and $\mathbf{X_l}$ the individuals $\times$ contextual variables table, the global resulting complex data structure is shown in Figure 1.1. We called this data structure a sequence of "coupled tables", each coupled table made of one frequency table and one quantitative/qualitative table. The "coupled table" $(\mathbf{Y_l}, \mathbf{X_l})$ corresponds to language l. Note that all the $\mathbf{X_l}$ $(l = 1, ..., L)$ have common columns.

Integration of the information of some closed questions into the analysis allows carrying out more detailed analyses and answering new questions such as: which are the variables that are globally similar (through all the samples)? Which words are associated with which variables? How similar or different are the variables structures from one sample to another, from the point of view of their association with the words? Which are the similarities between the samples of individuals, with two samples being much closer if the attractions (respectively, the repulsions) between variables and words induced by each sample are much similar?

Data sets with such a structure can be found in other fields provided that the data are coded in a similar structure such as social networks, machine learning and ecology. In the ecological studies, the species present on the sites of different ecological systems (or of the same system in different years) are counted. The species vary from a system to another, although some can be common. For every system, that is every subsample of sites, a frequency table sites $\times$ species is built. The frequency tables have neither the row-sites nor the column-species in common. From one system/subsample to another, the sites are

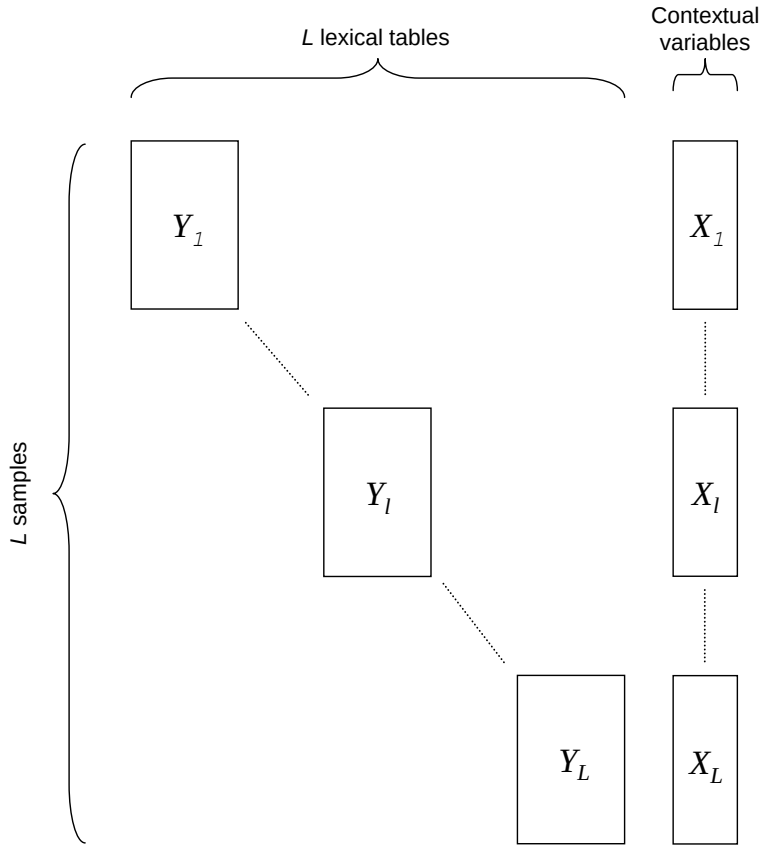


Fig. 1.1 Sequence of L coupled tables

described by the same environmental variables. The ecological systems have to be compared from the point of view of the relationships between species and environmental variables, giving to the latter an explicative role.

A similar data structure to the one we want to deal with, specially employed in ecology, is called a sequence of "paired tables". Each pair of ecological tables includes a first table containing environmental variables (usually quantitative) recorded in a set of sampling and a second table containing species frequency recorded at the same sampling sites. Several methods based on co-inertia analysis (COIA; Dolédec and Chessel, 1994; Dray et al., 2003) have been proposed to deal with sequence of paired ecological tables: Between-Group Co-Inertia Analysis (BGCOIA; Franquet et al., 1995), STATICO (Simier et al., 1999; Thioulouse et al., 2004) and COSTATIS (Thioulouse, 2011). Nevertheless, all of them share the same constraints on the species (frequency columns) and environmental variables, which should always be identical: same species and same environmental variables for all paired tables

(Thioulouse, 2011). In our case, rows and columns of the frequency tables are different from one sample to another. Thus, we prefer to call our data structure as "coupled tables".

Correspondence Analysis (CA) is the reference factorial method to analyze a frequency table (Benzécri, 1973, 1981; Lebart et al., 1998; Murtagh, 2005). CA applied to a lexical table (CA-LT), for example analysis of an open-ended question answered in one language, offers an optimal visualization of the similarities among the free answers and among the words as well as the associations between words and free answers.

The idea of jointly analysing open-ended and closed questions has been already studied. Lebart et al. (1998) emphasized difficulties for the interpretation of the similarities/oppositions among respondents without taking into account the respondents' characteristics. He proposed to build an aggregated lexical table (ALT; Lebart et al., 1998) crossing the categories of one categorical variable and the words. In this ALT, the row-respondents corresponding to a same category are collapsed into a single row while the column-words remain unchanged. CA applied to this ALT aims at establishing a typology of the categories from the words that they use and a typology of the words from the categories that use them (CA-ALT; Lebart et al., 1998).

CA considers only one frequency table. However, an extension of CA combined with Multiple Factor Analysis (MFA; Escofier and Pagès, 1988), allows for dealing with a multiple frequency table that juxtaposes row-wise several frequency tables. This method is called Multiple Factor Analysis for Contingency Tables (MFACT; Bécue-Bertaut and Pagès, 2004). For example, for ecological studies, the multiple frequency table could be the frequency of species on the same ecological sites along different years. In this case, each year corresponds to a different frequency table. MFACT combines Intra-tables Correspondence Analysis (ICA; Benzécri, 1983; Cazes and Moreau, 1991, 2000; Escofier and Drouet, 1983) and Multiple Factor Analysis. ICA centers each frequency table on its margin, allowing to consider the intra-tables independence model, while MFA balances the influence of each table.

Pagès and Bécue-Bertaut (2006) have proposed a first proposal to analyse an open-ended question answered in different languages by different samples. In this case, an aggregated lexical table $\mathbf{Y}_\mathbf{A}^\mathbf{l}$ crossing the categories of one categorical variable and the words is built for every sample l ($l = 1, ..., L$). In order to analyse and to compare the samples, the aggregated lexical tables are juxtaposed row-wise into a multiple frequency table $\mathbf{Y}_\mathbf{A}$ (Figure 1.2) and submitted to MFACT.

There are other approaches proposed to analyse textual information from different languages. Cross-Language Latent Semantic Indexing (CL-LSI; Littman et al., 1998), a method based on machine translation, aims at building a dual-language semantic space to represent

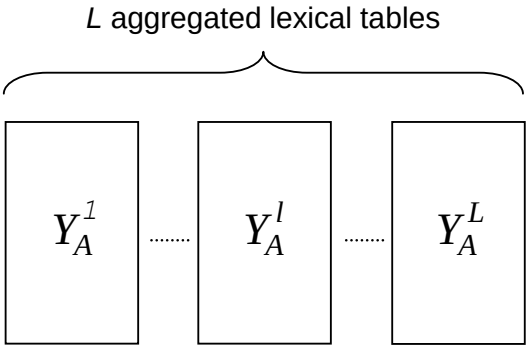


Fig. 1.2 The multiple frequency table $\mathbf{Y_A}$ that juxtaposes row-wise the L aggregated lexical tables

different languages in order to automate cross-language document retrieval without any translation query. For this, it requires a set of documents consisting of parallel text from each language translated by humans or by machine. However, this approach does not allow to study the similarities of the homologous words from different languages. They are translated automatically from one language to the other assuming that they have the same meaning for the individuals of different samples.

We aim at expanding the first proposal by Pagès and Bécue-Bertaut (2006) to the case of $\mathbf{X_l}$ including several quantitative or categorical variables. This requires to generalize the aggregated lexical table to several quantitative or categorical variables. This thesis contributes to the development of a new method called Correspondence Analysis on a Generalised Aggregated Lexical Table (CA-GALT; Bécue-Bertaut and Pagès, 2014; Bécue-Bertaut et al., 2014) that was proposed to generalize classical CA-ALT to the case of several quantitative, categorical and mixed variables. It aims to establish a typology of the contextual variables and a typology of the words from their mutual relationships.

The method that we propose to deal with a sequence of coupled tables combines MFACT to balance the influence of the sets of words and CA-GALT to place all the words in the same space and is called *Multiple Factor Analysis on Generalised Aggregated Lexical Table* (MFA-GALT).

1.2 Thesis outline

This thesis presents several methods that deal with data sets that are increasingly complex. Chapter 2 summarizes the general framework for principal components methods and then

presents the methods that support our approach: Principal Component Analysis (PCA), Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA) and Multiple Factor Analysis (MFA). In the next chapters, we present methods that bring new contributions to data processing. In chapter 3, we present MFACT and detail a new methodology to identify consensual words in sensory studies where free-text descriptions and equivalent methods are frequently used (Kostov et al., 2014). Chapter 4 presents CA-GALT and its extension to the case of several categorical variables (Bécue-Bertaut et al., 2014). Chapter 5 is devoted to the method that we propose to deal with a sequence of coupled tables: MFA-GALT. Chapter 6 sets out the implementation of the methods in R (Kostov et al., 2013, 2015). The main conclusions are set out in chapter 7.

Chapter 2

General framework for principal components methods

All the new methods and methodologies presented in this thesis are in the framework of exploratory data analysis in the "French way". Thus, they are based on principal components methods such as PCA, CA or MCA that deal with single data tables and they are also based on MFA that deals with several data tables. In this chapter, we present all these methods and their properties since they will be used in the next chapters.

2.1 Principal components methods

2.1.1 Data, clouds of rows and columns

Principal components methods (PCM), also known as General Factor Analysis (GFA; Lebart et al., 1997), offer a geometrical approach to extract and visualize the information contained in data tables such as individuals $\times$ variables or frequency/contingency tables.

PCM consider three matrices: the $(I \times K)$ data matrix $\mathbf{X}$, the weighting system for the rows, also used as the metric in the column space, stored in the $(I \times I)$ matrix $\mathbf{D}$, and the weighting system for the columns, and metric in the row space, stored in the $(K \times K)$ matrix $\mathbf{M}$.

The similarities between rows and between columns are studied from a geometrical viewpoint. The row cloud N_I (resp. the column cloud N_K) endowed with the weighting system $\mathbf{D}$ (resp. $\mathbf{M}$) are placed in $\mathbb{R}^K$ (resp. $\mathbb{R}^I$) endowed with metric $\mathbf{M}$ (resp. $\mathbf{D}$).

2.1.2 Inertia axes

PCM look for providing a visualization of the similarities/dissimilarities among the rows and among the columns as well as their associations (or repulsions) on a reduced dimension space spanned by the first main inertia axes.

The inertia axes in the row space (resp. column space) $\mathbb{R}^K$ (resp. $\mathbb{R}^I$) satisfy the equation 2.1 (resp. 2.2)

$$\mathbf{X^T D X M u_s} = \lambda_s \mathbf{u_s} \tag{2.1}$$

$$\text{with the restriction } ||\mathbf{u_s}||_{\mathbf{M}} = \mathbf{u_s^T M u_s} = 1.$$

$$\mathbf{X M X^T D v_s} = \lambda_s \mathbf{v_s} \tag{2.2}$$

$$\text{with the restriction } ||\mathbf{v_s}||_{\mathbf{D}} = \mathbf{v_s^T D v_s} = 1.$$

$\mathbf{u_s}$ and $\mathbf{v_s}$ correspond to the eigenvectors associated with the eigenvalue of rank s denoted λ_s. The eigenvalue λ_s is both the inertia of cloud N_I and cloud N_K projected on the axis of rank s in their own space.

2.1.3 Representation of the rows and the columns in a reduced dimension space

The vector of coordinates of the rows onto axis s, called principal component of rank s, is computed as

$$\mathbf{F_s} = \mathbf{X M u_s}. \tag{2.3}$$

In a similar manner, the vector of coordinates of the columns onto axis s is computed as

$$\mathbf{G_s} = \mathbf{X^T D v_s}. \tag{2.4}$$

2.1.4 Duality and the transition relations

The duality relationships are an essential property of PCM. Thanks to this property, the projection of rows (resp. columns) onto rank s axis in $\mathbb{R}^K$ (resp. $\mathbb{R}^I$) can be calculated from the coordinates of N_K (resp. N_I) onto rank s axis in $\mathbb{R}^I$ (resp. $\mathbb{R}^K$) by the way of the "transition relationships" expressed as

$$\mathbf{F_s} = \mathbf{X M G_s} \lambda_s^{-1/2} \tag{2.5}$$

$$\mathbf{G_s} = \mathbf{X^T D F_s} \lambda_s^{-1/2} \tag{2.6}$$

Table 2.1 General scheme shared by all principal components methods

	Row-points cloud N_I	Column-points cloud N_K
Space	$\mathbb{R}^K$	$\mathbb{R}^I$
Metric	$\mathbf{M}$	$\mathbf{D}$
Data matrix	$\mathbf{X}$	$\mathbf{X^T}$
Weights	$\mathbf{D}$	$\mathbf{M}$
Axes of inertia	$\mathbf{U}$	$\mathbf{V}$
Equation	$\mathbf{X^T DXMU} = \mathbf{U}\Lambda$ (2.7)	$\mathbf{XMX^T DV} = \mathbf{V}\Lambda$ (2.8)
Orthonormality	$\mathbf{U^T MU} = \mathbf{Id}$	$\mathbf{V^T DV} = \mathbf{Id}$
Principal components	$\mathbf{F} = \mathbf{XMU}$	$\mathbf{G} = \mathbf{X^T DV}$
Equation	$\mathbf{XMX^T DF} = \mathbf{F}\Lambda$ (2.9)	$\mathbf{X^T DXMG} = \mathbf{G}\Lambda$ (2.10)
Orthogonality	$\mathbf{F^T DF} = \Lambda$	$\mathbf{G^T MG} = \Lambda$
Equation (symmetrical form)	$\mathbf{M^{1/2} X^T DX M^{1/2} \tilde{U}} = \mathbf{\tilde{U}}\Lambda = \mathbf{M^{1/2}U}\Lambda$ (2.11)	$\mathbf{D^{1/2} XMX^T D^{1/2} \tilde{V}} = \mathbf{\tilde{V}}\Lambda = \mathbf{D^{1/2}V}\Lambda$ (2.12)
Transition relations	$\mathbf{F} = \mathbf{XMG}\Lambda^{-1/2}$	$\mathbf{G} = \mathbf{X^T DF}\Lambda^{-1/2}$

2.1.5 General scheme for the principal components methods

The general scheme for the principal components methods is summarised in Table 2.1. Eq.(2.7) and Eq.(2.8) give the expressions of the matrices to be diagonalized. The expression of the principal components and the relationships between Eq.(2.7) and Eq.(2.10), on the one hand, and between Eq.(2.8) and Eq.(2.9), on the other hand, show that either only Eq.(2.7) or only Eq.(2.8) have to be solved for computing both series of axes.

Principal Component Analysis (PCA), Multiple Correspondence Analysis (MCA) and Correspondence Analysis (CA) are introduced in the global scheme through specific definition for matrices **X**, **D** and **M** (Escofier and Pagès, 1988).

2.1.6 Tools and indicators for the interpretation

Several tools and indicators for principal components methods help in the interpretations of the results of principal components methods. The most important of these helps are:

- **Explained variance by component**: is equal to the ratio between the projected inertia and the total inertia. For component s is equal to

$$\frac{\lambda_s}{\sum_{s \in S} \lambda_s}. \tag{2.13}$$

  Multiplied by 100, this indicator gives the percentage of inertia expressed by the component of rank s.

- **Contribution of an element to the inertia of axis s**: the rows (resp. the columns) contribute to axis s inertia with

$$\frac{\mathbf{DF_s}^2}{\lambda_s} \times 100 \tag{2.14}$$

$$\frac{\mathbf{MG_s}^2}{\lambda_s} \times 100. \tag{2.15}$$

- **The quality of representation**: for a row i on the component s can be measured by the distance between the point within the space and the projection on the component

$$qlt_s(i) = \frac{Projected\ inertia\ of\ i\ on\ \mathbf{u_s}}{Total\ inertia\ of\ i} = cos^2\theta_i^s \tag{2.16}$$

  where θ_i^s is the angle between **Oi** (vector connecting the origin to the point i) and $\mathbf{u_s}$.

2.2 Specific methods

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is the principal components method that deals with individuals $\times$ quantitative variables table. The $(I \times K)$ matrix $\mathbf{X}$ contains K columns and I rows with generic term x_{ik} denoting the value of variable k for individual i.

In PCA, the data are column-centred to compute the inertia axes relative to the centroid of the cloud of individuals. Column-standardization must be considered if the variables are expressed in different units. Even when the units are the same, standardizing is preferable as it modifies the shape of the cloud by harmonising its variability in all the directions of the original variables giving the same importance to them. When the variables are centred and standardized PCA is called "standardized" and unstandardized PCA when the variables are only centred.

The principal component analysis (PCA) applied to a data matrix **X** (centred or standardized) with metric **M/D** and weighting system **D/M** in the row/column space is noted PCA(**X**,**M**,**D**).

PCA brings two individuals closer inasmuch as they assume similar values for the variables and two variables closer inasmuch as the same individuals assume high/low values for the variables. Variables are represented by means of the covariances/correlations between variable k and principal component s, visualizing whether or not a variable k is related to a dimension of variability.

The methods presented hereinafter will be shown as specific PCA.

2.2.2 Correspondence Analysis

CA applied to textual data

Correspondence analysis (CA) was proposed by Benzécri (1973, 1981) to deal with textual data. The corpus, collection of written or oral documents, is encoded into a $(I \times J)$ documents $\times$ words table **Y**, called a lexical table. The general term of this frequency table y_{ij} counts the frequency of word j in document i.

The term "correspondence analysis" stems from the fact that lexical table links two corresponding sets: one represented by the rows, and other represented by the columns, playing symmetric roles.

CA can be summarized through the following steps:

- The frequency table **Y** is transformed into a proportion table

$$\mathbf{P} = [p_{ij}] = [\frac{y_{ij}}{\sum_{i \in I} \sum_{j \in J} y_{ij}}] \tag{2.17}$$

thus: $\sum_{i\in I}\sum_{j\in J} p_{ij}$=1. The row and column margins of table **P** have for generic terms, respectively, $p_{i\bullet} = \sum_{j\in J} p_{ij}$ and $p_{\bullet j} = \sum_{i\in I} p_{ij}$ (filled in matrix **D** and **M**).

- The profiles of both sets of rows and columns are computed as $p_{ij}/p_{i\bullet}$ $(i = 1,...,I)$ and $p_{ij}/p_{\bullet j}$ $(j = 1,...,J)$.
- The distance between rows i and i' (resp. between columns j and j') is known as chi-square distance and is defined as

$$d^2(i,i') = \sum_{j\in J} \frac{1}{p_{\bullet j}}(\frac{p_{ij}}{p_{i\bullet}} - \frac{p_{i'j}}{p_{i'\bullet}})^2 \tag{2.18}$$

$$d^2(j,j') = \sum_{i\in I} \frac{1}{p_{i\bullet}}(\frac{p_{ij}}{p_{\bullet j}} - \frac{p_{ij'}}{p_{\bullet j'}})^2. \tag{2.19}$$

  This definition obeys the three principle imposed by the Benzécri (1973, 1981) viewpoint: the distance between two rows (respectively, two columns) has to be defined on the profiles, fulfil the distributional equivalence principle and be quadratic. This principle imposes that the distance between two rows i and i' does not change if we merge two columns j and j'. Similarly, the distance between two columns j and j' does not change if two rows i and i' are merged.

Classical CA results can be obtained through a PCA applied to

$$\mathbf{Q} = \mathbf{D^{-1}PM^{-1}} = [q_{ij}] = [\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}] \tag{2.20}$$

or equivalently, to its doubly centred form, the $(I \times J)$ matrix

$$\bar{\mathbf{Q}} = [\bar{q}_{ij}] = [\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{p_{i\bullet}p_{\bullet j}}] \tag{2.21}$$

with metrics/weights **M** and **D**, that is, PCA($\bar{\mathbf{Q}}$,**M**,**D**) (Escofier and Pagès, 1988; Pagès and Bécue-Bertaut, 2006). This computing places CA in the general scheme for the principal components methods evidencing that CA analyses the weighted deviation between **P** and the $(I \times J)$ independence model matrix $[p_{i\bullet}p_{\bullet j}]$.

Transition relationships in CA

The simultaneous representation of rows and columns relies on the transition relationships linking the coordinates $F_s(i)$ of row-points i $(i = 1,...,I)$ and the coordinates $G_s(j)$ of the column-points j $(j = 1,...,J)$ on the dispersion axes s $(s = 1,...,Min(I-1,J-1))$.

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J} \frac{p_{ij}}{p_{i\bullet}} G_s(j) \tag{2.22}$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{p_{ij}}{p_{\bullet j}} F_s(i) \tag{2.23}$$

For dimension s of this superimposed representation, up to the multiplicative factor $1/\sqrt{\lambda_s}$, a row i (resp. a column j) is at the barycentre of the columns (resp. of the rows), each column j (resp. each row i) having a weight $p_{ij}/p_{i\bullet}$ (resp. $p_{ij}/p_{\bullet j}$). This property is used to interpret the position of one row in relation to all of the columns and the position of one column in relation to all of the rows.

The similarity between rows and columns is expressed in a totally symmetrical way (characteristic of CA which differentiates it from other principal components methods). Two rows are all the closer as they are frequently associated with the same columns and two columns are all the closer that they are frequently associated with the same rows.

2.2.3 Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is a particular PCM used to tackle a table with I rows (individuals) and Q columns (qualitative variables). These data sets are coded into a complete disjunctive table $\mathbf{X}$ with K columns corresponding to the K categories of the Q qualitative variables with general term x_{ik} equals to 1 if individual i belongs to the category k and 0 otherwise.

We note I_k the number of individuals belonging to category k. From matrix $\mathbf{X}$, the proportion matrix

$$\mathbf{P} = [p_{ik}] = [\frac{x_{ik}}{IQ}] \tag{2.24}$$

is built. The marginal terms of this table are filled in the $(I \times I)$ diagonal matrix

$$\mathbf{D} = [d_{ii}] = [\frac{1}{I}] \tag{2.25}$$

and in the $(K \times K)$ diagonal matrix

$$\mathbf{M} = [m_{kk}] = [\frac{I_k}{IQ}]. \tag{2.26}$$

Classical MCA results can be obtained through a PCA applied to matrix

$$\mathbf{Z} = [z_{ik}] = [\frac{Ix_{ik}}{I_k} - 1] \tag{2.27}$$

with metrics/weights **M** and **D**, that is, PCA(**Z**,**M**,**D**) (Escofier and Pagès, 1988).

MCA considers three series of objects: individuals, variables and categories. Two individuals are similar if they share a great number of categories. Two categories are similar if they are frequently chosen by the same individuals.

2.3 Multiple Factor Analysis

Multiple Factor Analysis (MFA; Escofier and Pagès, 1988) deals with multiple tables in which individuals are described by several sets of variables. Within one set, variables must present the same type (quantitative or categorical) but sets of variables can belong to different types. MFA requires that the unit weights have to be identical across all the tables. To simplify this short synthesis of MFA, weights of individuals and weights of variables are supposed to be uniform.

The I individuals i $(i = 1,...,I)$ constitute the cloud N_I in the K-dimensional space $\mathbb{R}^K$; the K variables k $(k = 1,...,K)$ belonging to L sets of variables constitute the cloud N_K in the I-dimensional space $\mathbb{R}^I$.

If we consider the only (sub-)table l $(l = 1,...,L)$, individuals are denoted $i^l (i = 1,...,I)$ and constitute the cloud N_I^l in the K_l - dimensional space $\mathbb{R}^{K_l}$; the K_l variables constitute the cloud N_K^l in the I-dimensional space $\mathbb{R}^I$.

2.3.1 Balancing the sets of variables

The global analysis, where L sets of variables are simultaneously introduced as active, requires balancing the influences of the sets of variables. The influence of one set l derives from its structure in the different space directions. If a set presents a high inertia in one direction, this direction will strongly influence the first axis of the global analysis. That

suggests normalising the highest axial inertia of each set to 1. This is obtained by weighting each variable of set l by $1/\lambda_1^l$ where λ_1^l is the first eigenvalue issued from the principal component method applied to set l. This reweighting makes the highest axial inertia of each set equal to 1. However, it does not balance the total inertia of the different sets. Thus, a set with a high dimensionality will contribute to more axes than a set with low dimensionality.

The basic principle of MFA is a PCM applied to all sets of variables (global analysis) but balancing the influence of each set of variables on the computing of the first axis. MFA works with continuous variables as principal component analysis does, the variables being weighted; MFA works with categorical variables as multiple correspondence analysis does, the variables being weighted. MFA provides the classical outputs of general factor analysis:

- Coordinates, contributions and squared cosines of individuals
- Correlation coefficient between factors and continuous variables
- For each category, coordinate of the centroid of the individuals belonging to this category

MFA provides also specific outputs that are described below.

2.3.2 Superimposed representation of the *l* clouds of individuals

The partial cloud N_I^l of the individuals in the space $\mathbb{R}^{K_l}$ is associated with each set l. It contains "partial" individuals, noted i^l, that is individual i according to the set l.

To determine the resemblances, from one cloud to another, among distances between homologous points, the clouds N_I^l are projected upon the axes of the global analysis, as illustrative elements. The coordinate of i^l along axis s is denoted as $F_s(i^l)$ and is calculated from the coordinates of the variables $G_s(k)$, $k \in K_l$, by the way of the following relationship

$$F_s(i^l) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^l}} \sum_{k \in K_l} x_{ik} G_s(k) \tag{2.28}$$

that is, the usual transition formula but restricted to K_l variables of set l.

2.3.3 Synthetical representation of the sets

MFA also visualizes the proximities between the sets, each of them represented by a unique point. In this visualization, two sets are close to one another if they induce similar structures on the individuals.

For that, each set of variables K_l is represented by the $(I \times I)$ matrix $\mathbf{W_l}$ of scalar products between individuals ($\mathbf{W_l} = \mathbf{X_l X_l^T}$) from set l viewpoint. Each scalar product matrix $\mathbf{W_l}$ corresponds to one point in $\mathbb{R}^{I^2}$ which represents the set l. The L points constitute the set cloud N_L. In this cloud, the distance between two points $\mathbf{W_l}$ and $\mathbf{W_{l'}}$ decreases insofar the similarity between the structures (defined upon individuals) induced by sets K_l and $K_{l'}$ increases.

The representation provided by MFA is obtained in the following way. Each factor of rank s is considered as a set including a single variable; it is possible to associate with this set a scalar product matrix and thus a vector in $\mathbb{R}^{I^2}$. The normalised factor of rank s in $\mathbb{R}^K$, denoted $\mathbf{v_s}$ (Eq.2.2), induces the vector $\mathbf{w_s} = \mathbf{v_s v_s^T}$ in $\mathbb{R}^{I^2}$. Some properties of $\mathbf{v_s}$ induce corresponding properties for $\mathbf{w_s}$

$$\mathbf{v_s^T v_{s'}} = 0 \Longrightarrow \langle \mathbf{w_s}, \mathbf{w_{s'}} \rangle = 0 \tag{2.29}$$

$$||\mathbf{v_s}|| = 1 \Longrightarrow ||\mathbf{w_s}|| = 1 \tag{2.30}$$

The main interest of this projection space is that its reference axes (upon which N_L is projected) are interpretable in the same manner as the axes of the global analysis, due to factor analysis duality.

The coordinate of set l upon axis of rank s is computed as

$$Lg(l, \mathbf{v_s}) = \langle \mathbf{W_l D}, \mathbf{v_s D} \rangle = trace(\mathbf{W_l D v_s v_s^T D}) \tag{2.31}$$

that is the same as the sum of the contributions of K_l variables belonging to set l. Thus:

- Set coordinates are always between 0 and 1
- A small distance between two sets along axis s means that these two sets include the structure expressed by factor s each one with the same intensity. In other words, set representations shows which ones are similar (or different) from the point of view of global analysis factors.

Chapter 3

Multiple Factor Analysis for Contingency Tables

MFA deals with several data tables that are quantitative or qualitative. Since we are concerned by textual data sets, we need to use the extension of MFA to several frequency tables. This chapter presents the Multiple Factor Analysis for Contingency Tables (MFACT; Bécue-Bertaut and Pagès, 2004, 2008), an extension of the MFA to the case of frequency tables and/or mixture of quantitative, categorical and frequency tables.

Here, we present the first proposal by Pagès and Bécue-Bertaut (2006) to analyse an open-ended question answered in different languages using MFACT. Besides, a new methodology to identify consensual words in the context of the sensory studies is detailed (Kostov et al., 2014). We explain how this method can be applied in the case of the multi-language open-ended question to study the meaning of the homologous words.

The software implementation of MFACT (Kostov et al., 2013) is offered in chapter 6.

3.1 Multiple frequency table and notation

Several frequency tables $\mathbf{Y_1}, ..., \mathbf{Y_l}, ..., \mathbf{Y_L}$, of dimensions $(I \times J_l)$, are juxtaposed row-wise into the multiple frequency table $\mathbf{Y}$ of dimension $(I \times J)$.

$\mathbf{Y}$ is transformed into proportion table $\mathbf{P}$, globally computed on all the tables (Figure 3.1). Thus, p_{ijl} is the proportion of row i $(i = 1, ..., I)$ for column j $(j = 1, ..., J_l)$ of table l $(l = 1, ..., L)$; $\sum_{l \in L} \sum_{i \in I} \sum_{j \in J_l} p_{ijl} = 1$. The row and column margins of table $\mathbf{P}$ are respectively $p_{i\bullet\bullet} = \sum_{l \in L} \sum_{j \in J_l} p_{ijl}$ and $p_{\bullet jl} = \sum_{i \in I} p_{ijl}$.

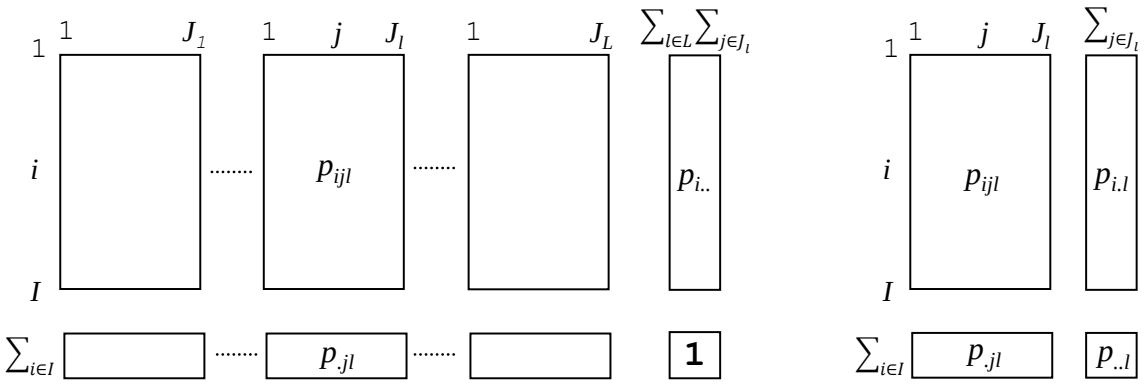


Fig. 3.1 Proportion table **P** and its global and separate margins

3.2 The algorithm

The introduction of the frequency tables as sets of variables induces a specific problem. Multiple Factor Analysis (MFA) requires that the unit weights be identical across all the tables. In the case of the classical MFA, analysing sets of either quantitative or categorical variables, uniform weights are adopted. In the case of a frequency table, CA imposes the coefficients of the row margin as row weights.

The CA of the multiple frequency table that juxtaposes several frequency tables allows a comparison, in a single framework, of the different sets of columns but not of the rows, described only by the columns as a whole. Furthermore, when the tables have no proportional column margins, CA does not actually compare the internal structures of the tables, because of the different centroids. Internal Correspondence Analysis (ICA; Benzécri, 1983; Cazes and Moreau, 1991, 2000; Escofier and Drouet, 1983) solves this problem by centring the subtables on their own margins. It can be seen as a CA that refers to the intra-tables independence model whose generic term is

$$m_{ijl} = \left(\frac{p_{i\bullet l}}{p_{\bullet\bullet l}}\right) p_{\bullet jl} \tag{3.1}$$

where $p_{i\bullet l} = \sum_{j\in J_l} p_{ijl}$ is the row margin of table l and $p_{\bullet\bullet l} = \sum_{i\in I}\sum_{j\in J_l} p_{ijl}$ is the sum of the terms of table l inside proportion table **P**.

Table **Z** is built as

$$\mathbf{Z} = \frac{p_{ijl} - m_{ijl}}{p_{i\bullet\bullet} \times p_{\bullet jl}} = \frac{p_{ijl} - \left(\frac{p_{i\bullet l}}{p_{\bullet\bullet l}}\right) p_{\bullet jl}}{p_{i\bullet\bullet} \times p_{\bullet jl}} = \frac{1}{p_{i\bullet\bullet}}\left(\frac{p_{ijl}}{p_{\bullet jl}} - \frac{p_{i\bullet l}}{p_{\bullet\bullet l}}\right). \tag{3.2}$$

Table **Z** generic term is the weighted residual with respect to the intra-tables independence model. This model neutralizes the differences between the separate average column profiles.

MFACT performs a non-standardized PCA on the global table **Z** giving the weight $p_{i\bullet\bullet}$ to the row i and the weight $p_{\bullet jl}/\lambda_1^l$ to the column j of table l with λ_1^l the first eigenvalue from the separate PCA of subtable $\mathbf{Z_l}$.

MFACT combines ICA which solves the problem of the different margins and MFA which balances the influence of the sets. The weighted rows and the weighted columns of each table l are centred. The influence of each subtable in the global analysis is balanced in a MFA-like way. MFA classical results are obtained and CA characteristics and interpretation rules are kept.

3.3 Analysis of an open-ended question answered in different languages

A first proposal to analyse an open-ended question answered in different languages by different samples of individuals using MFACT has been proposed by Pagès and Bécue-Bertaut (2006).

3.3.1 The dataset

In the example that they used, people from three cities (Tokyo, Paris, New-York) were asked the following open-ended question "Which dishes do you like and eat often?".

The data is encoded as follows. In each city, respondents were gathered into six groups by crossing gender (male, female) and age (in three age intervals: under 30, between 30 and 50, over 50). Then, for each city l ($l = 1, ..., L$), the $(K \times J)$ aggregated lexical table $\mathbf{Y_A^l}$ is constructed by crossing the six groups and the words. The general term y_{Akjl} is the frequency of word j by respondents belonging to group k in city l. The L aggregated lexical tables (in the example L=3) have the same K rows ($K = 6$): each row is a group of respondents. Columns are not homologous through the tables (they correspond to words in different languages).

3.3.2 MFACT applied to the multiple aggregated lexical table

The L aggregated lexical tables $\mathbf{Y_A^1}$,...,$\mathbf{Y_A^l}$,...,$\mathbf{Y_A^L}$ are juxtaposed row-wise into the $(K \times J)$ multiple aggregated lexical table $\mathbf{Y_A}$.

$\mathbf{Y_A}$ is transformed into the $(K \times J)$ proportion matrix

$$\mathbf{P_A} = \mathbf{Y_A}/N \tag{3.3}$$

with N the grand total over the table that gathers all the aggregated lexical tables; thus: $\sum_{l \in L} \sum_{k \in K} \sum_{j \in J_l} p_{Akjl}$=1

$p_{Ak\bullet l} = \sum_{j \in J_l} p_{Akjl}$ and $p_{A\bullet jl} = \sum_{k \in K} p_{Akjl}$ denote the row and column margins of the table l. $p_{Ak\bullet\bullet} = \sum_{l \in L} \sum_{j \in J_l} p_{Akjl}$ denote the row margin of $\mathbf{P_A}$. $p_{A\bullet\bullet l} = \sum_{k \in K} \sum_{j \in J_l} p_{Akjl}$ is the sum of the terms of table l inside table $\mathbf{P_A}$.

The table $\mathbf{Z}$ is built as

$$Z = [z_{kjl}] = \frac{p_{Akjl} - \left(\frac{p_{Ak\bullet l}}{p_{A\bullet\bullet l}}\right) p_{A\bullet jl}}{p_{Ak\bullet\bullet} \times p_{A\bullet jl}}. \tag{3.4}$$

MFACT performs a non-standardized PCA on the global table $\mathbf{Z}$ giving the weight $p_{Ak\bullet\bullet}$ to the row k and the weight $p_{A\bullet jl}/\lambda_1^l$ to the column j of table l with λ_1^l the first eigenvalue from the separate PCA of subtable $\mathbf{Z_l}$.

Authors proposed MFACT as a solution to the problem of separate analyses allowing to represent the words belonging to different languages in the same space. MFACT applied to the dataset allows for answering the following questions: Which are the gender $\times$ age categories that are globally similar (through all the samples)? Which are the similarities between dishes? Which dishes are associated with which gender $\times$ age categories? How similar or different are the category structures from one city to another (the partial representations)? Which are the similarities between cities?

3.4 Identification of consensual words

The analysis of open-ended question answered in different languages poses the problem of which is the meaning of the homologous words for the different samples. This problem has not been tackled by Pagès and Bécue-Bertaut (2006). They only mentioned different usages of *pizza* and *hamburger* made by the three samples.

Recently, we addressed a close problem but in another context, this of the sensory studies where free-text descriptions and equivalent methods are frequently used.

In this section, we expose this work and explain how it can be applied in the case of the multi-language open-ended question. We return to the text that we published in Food Quality and Preference with the title *An original methodology for the analysis and interpretation of word-count based methods: Multiple factor analysis for contingency tables complemented by consensual words* (Kostov et al., 2014).

3.4.1 Motivation

Free-text descriptions and frequency-of-citations based techniques (Campo et al., 2010; Varela and Ares, 2012) can be gathered under the name of word-count based methods that include open-ended questions (ten Kleij and Musters, 2003), check-all-that-apply (CATA; Lancaster and Foley, 2007), ultra-flash profiling (UFP; Perrin and Pagès, 2009) and labelled sorting task (Abdi et al., 2007; Cadoret et al., 2009). Although all these methods are executed in a different way, they share a common goal: understanding the consumers' sensory perceptions through collecting the product descriptions from the consumer's own vocabulary.

Word-count based methods are usually encoded into a products $\times$ words frequency table, called lexical table and analysed by means of a correspondence analysis. CA offers a visualization of: 1) the similarities between products: two products are all the closer as they are described by the same words; 2) the similarities between words: two words are all the closer as they are frequently associated with the same products and 3) the associations between products and words: a word is at the centroid of the products that it describes and a product is at the centroid of the words that describe it.

However, the large diversity of the vocabulary used by the panellists makes the products map difficult to interpret (Chollet et al., 2011). Moreover, the panellists understand and perceive some words in a different way and thus associate them with different products (Cadoret et al., 2009; Veinand et al., 2011). As discussed in Delarue and Sieffermann (2004), the ability of the panellists to communicate their sensory perceptions using a common base of vocabulary is sometimes doubtful. In fact, CA applied to a global products words table relies on the assumption that the same word mentioned by different panellists reports a similar perception. If this assumption is not verified, summing up the occurrences of the same word used by different panellists may not be meaningful because different perceptions are merged into the same word.

To deal with this problem, we propose to start from a different encoding of the results issued from word-count based methods which is detailed in section 3.4.2. This encoding preserves all the individual words provided by all the panellists that are included in a multiple frequency table. There are as many frequency tables as panellists. Then, MFACT is performed on the multiple frequency table, and an original technique is proposed to assess which words are consensual to guide the interpretation.

3.4.2 The dataset and encoding

Ninety eight consumers carried out a labelled sorting task on twelve luxury perfumes: Angel, Aromatics Elixir, Chanel nº5, Cinéma, Coco Mademoiselle, L'instant, Lolita Lempicka, Pleasures, Pure Poison, Shalimar, J'adore (eau de parfum), J'adore (eau de toilette). The consumers were mostly women (69.4%) and quite young (mean age: 25 years; range: 18-58).

Consumers were placed in individual booths, each perfume was sprayed on a small piece of cotton wool placed into a pill box, and all twelve pill boxes were presented at each consumer. The pill boxes were ordered according to William's design. Consumers had to evaluate the products in the presentation order but were allowed to go back to any sample; they were asked to make at least two and at most eleven groups of perfumes. After, they had to describe each group with a few words.

All the words are kept without applying any kind of spelling correction, lemmatization, stop-list or frequency threshold. One hundred and ninety-eight distinct-words and six hundred and eighty-four individual-words were used to characterize the perfumes.

The descriptions of the perfumes are encoded through perfumes $\times$ individual-words tables noted $\mathbf{Y_1}, ..., \mathbf{Y_l}, ..., \mathbf{Y_L}$, collecting and keeping all the data obtained from the consumers (Figure 3.2).

$\mathbf{Y_l}$, with dimensions $(I \times J_l)$, has as many columns as different individual-words as the consumer l $(l = 1, ..., L)$ used. Different individual-words, belonging to different consumers, can correspond to the same word. All the individual tables are juxtaposed row-wise into the $(I \times J)$ multiple frequency table $\mathbf{Y}$ (12 perfumes $\times$ 684 individual-words). Each cell y_{ijl} of $\mathbf{Y}$ is equal to "1" if perfume i is described with individual-word j used by consumer l and "0" if not.

The application of MFATC on multiple frequency table $\mathbf{Y}$ provides the representations of rows (perfumes), columns (individual-words) and groups (consumers). CA characteristics and interpretation rules are kept: 1) two perfumes are all the closer as they are described by

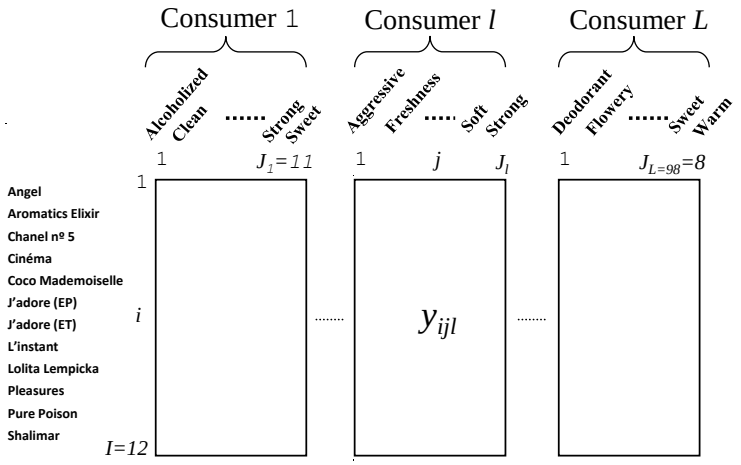


Fig. 3.2 Perfumes × individual-words tables

the same individual-words; 2) two individual-words are all the closer as they describe the same perfumes and 3) an individual-word is at the centroid of the perfumes that it describes and a perfume is at the centroid of the individual-words that describe it.

3.4.3 Assessing the consensual words

The consensual words are defined as those words that have the same meaning for most of the consumers as far as they describe the same products. The starting point to assess if a word is consensual or not, is to compute the within-inertia of the homologous individual-words as a measure of its level of consensus. First the centroid of each word w is computed as

$$C_d(w) = \frac{\sum_{l\in L}\sum_{j\in J_l}\frac{p_{\bullet jl}}{\lambda_1^l}G_s(j,l)\ \times\ 1_{word\ w=individual-word\ (j,l)}}{\sum_{l\in L}\sum_{j\in J_l}\frac{p_{\bullet jl}}{\lambda_1^l}\ \times\ 1_{word\ w=individual-word\ (j,l)}} \tag{3.5}$$

where $1_{word\ w=individual-word\ (j,l)}$ is equal to "1" if the individual-word j of table l corresponds to word w and "0" if not. $G_s(j,l)$ is the coordinate of individual-word j of table l on dimension s of MFACT and $\frac{p_{\bullet jl}}{\lambda_1^l}$ its weight as computed by MFACT.

Then, the within-inertia of word w is computed in a classical way:

$$WI_w = \sum_{s\in S}\sum_{l\in L}\sum_{j\in J_l}(G_s(j,l)-C_d(w))^2\frac{p_{\bullet jl}}{\lambda_1^l}\ \times\ 1_{word\ w=individual-word\ (j,l)}. \tag{3.6}$$

All the dimensions of MFACT are considered in equations 3.5 and 3.6. This choice is discussed in Section 3.4.5. An example is given with the words *heavy*, *peppery* and *young* (Figure 3.3). The word *young* was used by four consumers (four individual-words), *peppery* by six and *heavy* by seven. The coordinates of individual-words are computed through MFACT (Figure 3.3a). Then, centroids (Figure 3.3b) and within-inertias (Figure 3.3c) are computed.

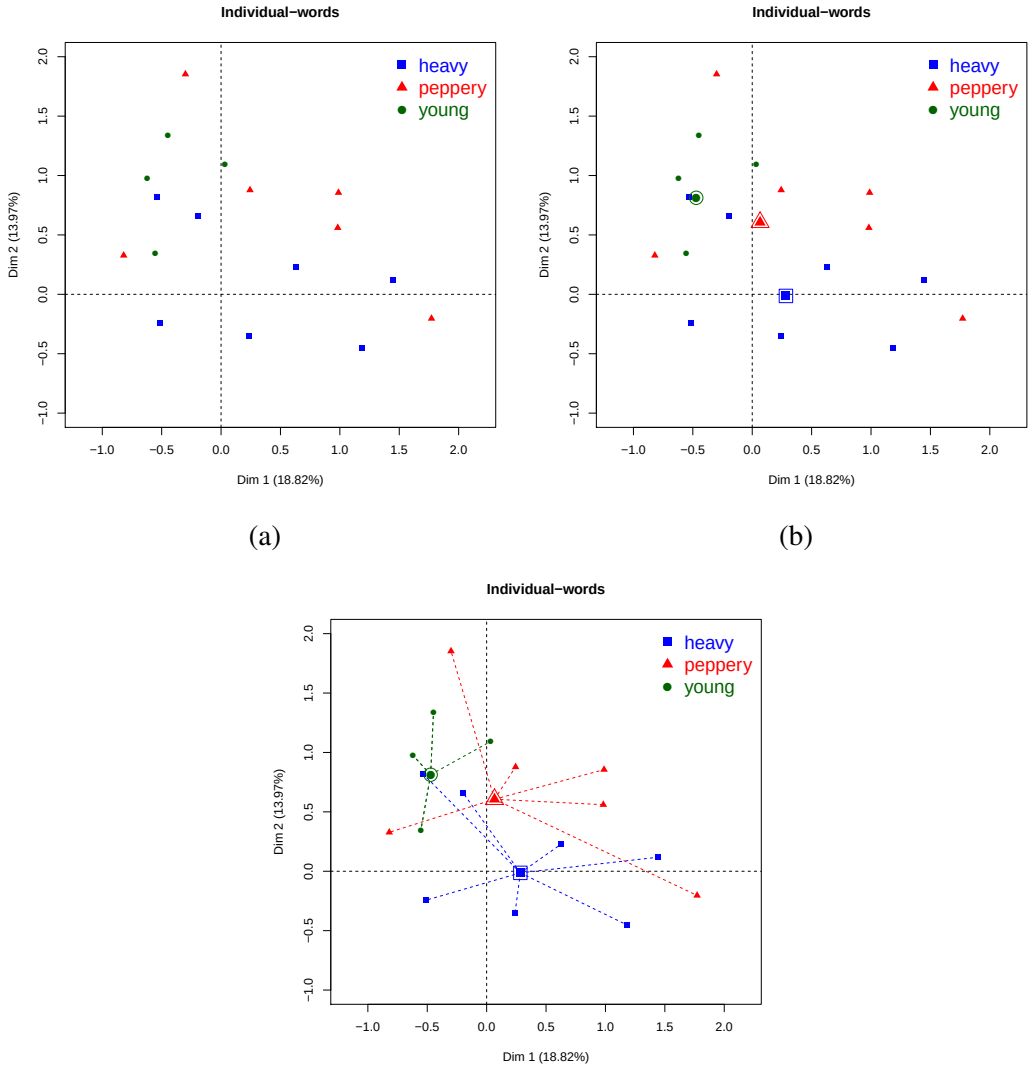
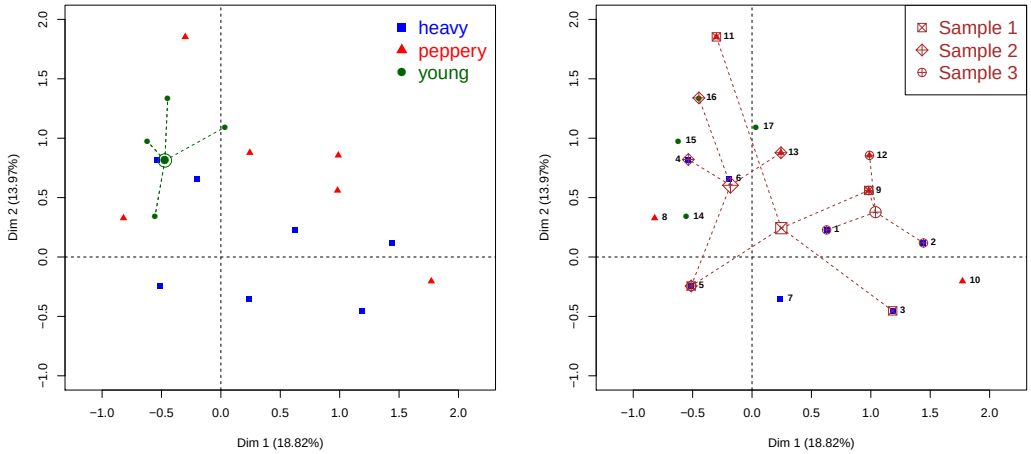


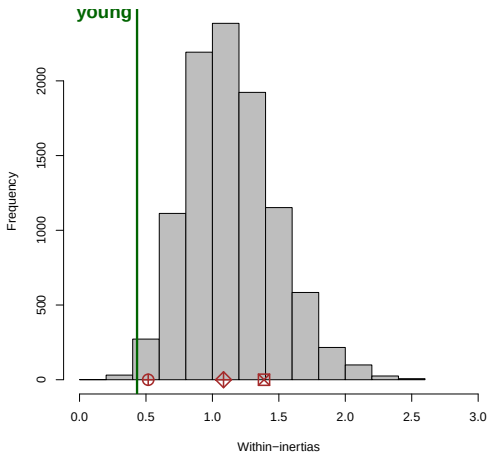
Fig. 3.3 Representation of individual-words, centroids and within-inertias of *heavy*, *peppery* and *young* on the first factorial plane issued from MFACT

A resampling technique is used to assess if a word is consensual. To achieve this purpose, for each distinct-word w used by m consumers, the null distribution is generated by choosing

m individual-words randomly from all the *N*=684 individual-words and then, their within-inertia is computed (in practice, 10000 random subsets are drawn). The within-inertia of word *w* is placed in the null hypothesis distribution (the association of the same word with different products by several consumers is interpreted as a lack of consensus for word *w*). Classically, the p-value is computed as the proportion of random subsets of size *m* having a within-inertia less than or equal to the within-inertia of the word *w*. If the p-value is less than 0.05, the word *w* is considered as consensual. This computing is performed for all the words cited at least by 3 consumers.



(a) (b)



(c)

Fig. 3.4 Representation of within-inertia of *young*, three samples with 4 randomly chosen individual-words and the null distribution of within-inertias for words used by 4 consumers

In the example, the word *young* has 4 occurrences (Figure 3.4a). So to assess if this word is consensual, samples with 4 individual-words are randomly chosen from all the individual-words and their within-inertias are computed (Figure 3.4b). The data set includes 684 individual-words but only 17 individual-words corresponding to *heavy*, *peppery* and *young* are represented in this example to ease the presentation of the methodology. The distribution of the within-inertias under the null hypothesis is generated by 10000 random samples with 4 individual-words and then, the observed within-inertia of the word *young* is situated on this distribution (Figure 3.4c). In this case, 43 samples have a smaller within-inertia than the within-inertia of *young* and the p-value of young is 43/10000=0.0043. Thanks to this, we can conclude that the observed within-inertia is statistically lower than the mean of the within-inertia under the null hypothesis. The algorithm is the same for the word *peppery* except that the random subsets consider 6 individual-words (and 7 for *heavy*).

Regarding the statistical software, the R function *WordCountAna* (Word-Count based methods Analysis) was developed and included into the package SensoMineR (Lê and Husson, 2008).

3.4.4 Results on the application

The total length of descriptions was 1984 occurrences composed of 684 individual-words and 198 distinct-words. *Flowery*, *soft*, *strong*, *sweet* and *fruity* were the five most frequent words, each one cited more than 100 times by more than 30 consumers.

MFACT was performed on the global table juxtaposing all the individual data sets. The first dimension opposed *Shalimar*, *Aromatics Elixir* and *Chanel n°5* to the other perfumes (Figure 3.5a). The second dimension opposed *Angel* and *Lolita Lempicka* to the rest. Consumers with high coordinates on the first dimension were those who detected differences between *Shalimar*, *Aromatics Elixir*, *Chanel n°5* and the rest of the perfumes (Figure 3.5b). *Angel* and *Lolita Lempicka* were individualized by those consumers with high coordinates on the second dimension. As it can be seen, the sensory map has shown that the differences between perfumes were identified by the consumers. However, the verbalization task was necessary to interpret the sensory dimensions and understand the reasons of the differences between the products from the consumer's point of view. Despite this, note that the representation of individual-words was difficult to interpret due to the high number of points and the diversity of the vocabulary (Figure 3.5c).

Among the 52 words pronounced by at least 3 consumers and corresponding to 511 individual-words, 12 words were assessed to be consensual (p-value less than 0.05). The

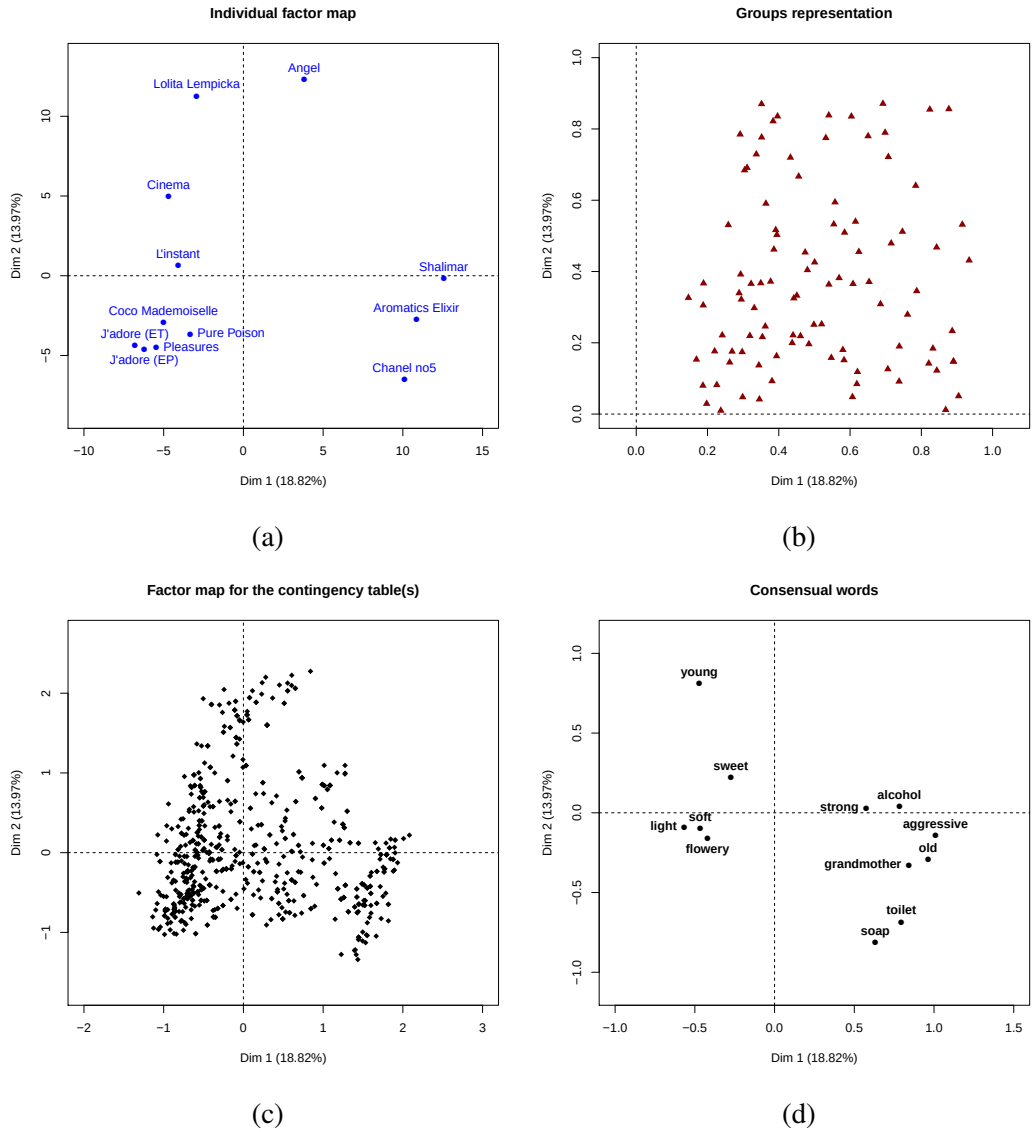


Fig. 3.5 Representation of perfumes, consumers, individual-words and consensual words in the plane defined by dimensions 1 and 2 of MFACT

four most frequent words (*flowery*, *soft*, *strong* and *sweet*) were consensual. But other widely pronounced words as *spicy*, *fresh*, *fruity* and *prickly* were not consensual. On the contrary, some words used by only few consumers as *young* and *alcohol* were consensual.

Finally, to represent the consensual words their centroids were used. Figure 3.5d shows that twelve consensual words were divided into two separate groups on the first dimension of MFACT. This dimension opposed the perfumes described by *flowery*, *soft*, *light*, *young* and *sweet*, to those described as rather *aggressive*, *strong*, *old*, *grandmother*, *alcohol*, *toilet* and *soap*. The second dimension opposed perfumes associated with *young* to the others.

3.4.5 Conclusions on the method to identify the consensual words

The methodology that we propose provides a map of the products taking into account all the panellists' points of view and easy to interpret through the consensual words.

In sensory analysis, assessing consensual properties of the words is important because the results from word-count based methods are customary analyzed through correspondence analysis assuming that the same word mentioned by different panelists corresponds to the similar perception. However, if the same word has different meanings for the panelists, it is not possible to use this word to interpret the products map since the reading can lead to different interpretations. In this sense, the methodology proposed to assess the consensual words eases the interpretation of the word-count based methods by solving problems arising from the large diversity of vocabulary and the different meanings possibly associated to a same word.

Regarding to the methodological aspects, a first remark concerns the computation of the within-inertia as consensus measure from all the dimensions of MFACT. As for any principal component method, the last dimensions could only be noise leading to unstable results and it could be advisable to remove them. However, when varying the number of dimensions considered to compute the within-inertia, we observe a high stability of the consensual words. Thus, the problem of choosing the number of dimensions can be avoided, keeping all of them. However, the number of consensual words seems to be related to the number of consumers. When the number of consumers is reduced to twenty, only two consensual words are found. If the threshold of the significance is increased up to 0.1, five consensual words are identified but the results should be analysed with caution in this case.

The synonymy of two words or two expressions with the similar meaning, e.g. *not very sweet* and *slightly sweet*, can also be studied applying the same methodology. To do so, the centroid and the within-inertia of all the individual-words corresponding to these two words are computed. Then, if the resampling test shows that the within-inertia is significantly small, the two words are considered as synonymous or have the same intensity meaning for most of the consumers.

Finally, if the free-text descriptions are obtained through a labelled sorting task, as in the example, multiple correspondence analysis (MCA) is an alternative to analyze the data (Cadoret et al., 2009). The resampling methodology proposed to detect the consensual words from the coordinates issued form MFACT can be easily adapted to the coordinates issued from MCA.

3.5 Study of the homologous words for different languages

There are other approaches proposed to analyse textual information from different languages. Cross-Language Latent Semantic Indexing (CL-LSI; Littman et al., 1998), a method based on machine translation, aims at building a dual-language semantic space to represent different languages in order to automate cross-language document retrieval without any translation query. For this, it requires a set of documents consisting of parallel text from each language translated by humans or by machine. However, this approach does not allow to study the similarities of the homologous words from different languages. They are translated automatically from one language to the other assuming that they have the same meaning for the individuals of different samples.

The methodology detailed in the previous section can be extended to study the meaning of the homologous words. We assume that the homologous words are perceived and understood in a similar way by different samples of individuals if they are used in a similar context. To assess so, the centroid and the within-inertia of all the individuals who pronounced the homologous words are computed. Then, if the resampling test shows that the within-inertia is significantly small, the homologous words are considered to have the same meaning for the individuals of different samples.

3.6 Conclusions

In textual analysis, MFACT can be used to deal with international surveys including open-ended questions answered in different languages. To do that, an aggregated lexical table is built crossing the categories of one categorical variable and the words for every sample and juxtaposed row-wise into a multiple frequency table submitted to MFACT.

The main drawback of this analysis is its restrictiveness. Only one categorical variable can be considered while often several categorical and quantitative variables are available and associated with the words.

We aim at expanding this analysis to the case of several quantitative/categorical variables. This requires first to generalize the aggregated lexical table to several quantitative/categorical variables as performed by *Correspondence Analysis on a Generalised Aggregated Lexical Table* (CA-GALT; Bécue-Bertaut and Pagès, 2014; Bécue-Bertaut et al., 2014).

Chapter 4

Correspondence Analysis on a Generalised Aggregated Lexical Table

In this chapter, we present the method called Correspondence Analysis on a Generalised Aggregated Lexical Table (CA-GALT; Bécue-Bertaut and Pagès, 2014; Bécue-Bertaut et al., 2014). This method was proposed to generalize correspondence analysis on an aggregated lexical table (CA-ALT; Lebart et al., 1998) to the case of several quantitative, categorical and mixed variables.

CA-GALT method was first proposed for quantitative contextual variables (Bécue-Bertaut and Pagès, 2014). Here, we present its extension to the case of several categorical variables. An application on a survey including an open-ended question is used to ease the presentation of the methodology. We return to the text of the article *Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach* published in SORT (Bécue-Bertaut et al., 2014).

The software implementation in the FactoMineR package through the *CaGalt* function (Kostov et al., 2015) is offered in chapter 6.

4.1 Motivation

Open-ended questioning is able to capture information in the form of free-text answers which could not be observed from closed questioning. The data coming from these open-ended questions usually are coded into a respondents $\times$ words frequency table. As mentioned in the previous chapters, correspondence analysis plays a central role in the statistical

analysis of open-ended questions. However, the direct analysis by CA of the lexical table, crossing respondents (rows) and words (columns), benefits from introducing the respondents' characteristics, such as age, education, etc. to obtain more robust results (Lebart et al., 1998).

A first manner consists in grouping the units depending on one categorical variable and building an aggregated lexical table (ALT) crossing the categories (rows) and the words (columns). In this ALT, the former respondent-rows corresponding to the same category are now collapsed into a single row while the word-columns remain unchanged. CA applied on this ALT is known as CA on the aggregated lexical table (CA-ALT; Lebart et al., 1998). CA-ALT offers a symmetric approach to the relationships between words and categories allowing for explaining the variability observed among the words by the variability observed among the categories and vice-versa. The attractions/rejections between certain words and certain categories are indicated and visualised on the principal planes. However, CA-ALT is too restrictive. Only one categorical variable can be considered while often several categorical and quantitative variables are available and required to better understand the variability observed in the lexical choices.

Two strategies were proposed when several categorical variables are considered. A first method proposes to perform a multiple correspondence analysis on categorical variables, clustering the respondents from their principal coordinates. Then, CA is performed on the aggregated lexical table crossing clusters and words. This strategy, called working demographic partition (WDP; Lebart et al., 1998), requires taking several decisions which are not obvious and any direct reference to the variables and categories is lost. A second method consists in applying the same idea as in CA-ALT. A multiple aggregated lexical table is built juxtaposing the aggregated lexical tables from each categorical variable. This approach has the drawback of not cancelling the associations among the variables and hence possible confusion effects remain.

We present here a methodology able to take into account several grouping variables while untangling their respective influence on the lexical choices and avoiding spurious relationships between certain categories and certain words. The new method, Correspondence Analysis on a Generalised Aggregated Lexical Table, has been proposed to generalize CA-ALT to the case of several quantitative, categorical and mixed variables.

4.2 Analysis of an aggregated lexical table

Before defining what we called a generalised aggregated lexical table, we recall the scheme of CA applied to an aggregated lexical table (CA-ALT) to facilitate the understanding of our proposal.

In this section, the columns of $\mathbf{X}$ are dummy variables corresponding to the categories of a single categorical variable. First, the $(J \times K)$ aggregated lexical table

$$\mathbf{Y_A} = \mathbf{Y^T X} \tag{4.1}$$

is built crossing the frequencies and the categories of the categorical variable (Figure 4.1). Its generic term y_{Ajk} is the frequency of word j in the free answers of the respondents belonging to category k.

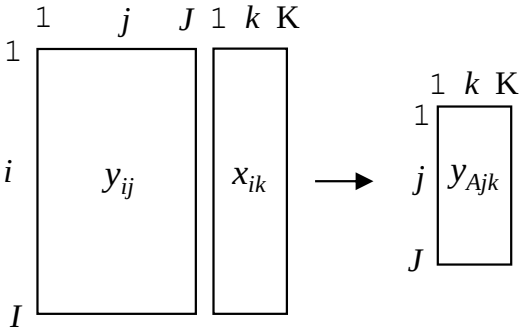


Fig. 4.1 Aggregated lexical table

Then, the $(J \times K)$ proportion matrix is computed as

$$\mathbf{P_A} = \mathbf{P^T X} \tag{4.2}$$

where the $(I \times J)$ matrix $\mathbf{P}$ corresponds to the proportion matrix issued from the lexical table. The $(J \times J)$ diagonal matrix

$$\mathbf{D_J} = [d_{Jjj}] = [p_{Aj\bullet}] \tag{4.3}$$

and the $(K \times K)$ diagonal matrix

$$\mathbf{D_K} = [d_{Kkk}] = [p_{A\bullet k}] \tag{4.4}$$

store, respectively, the row and column margins of $\mathbf{P_A}$. $\mathbf{D_J}$ (respectively, $\mathbf{D_K}$) corresponds to weighting system on the rows (respectively, on the columns). A double standardisation of

$\mathbf{P_A}$, on the rows and the columns, leads to the $(J \times K)$ data matrix analysed by CA-ALT

$$\mathbf{Q_A} = \mathbf{D_J^{-1}} \mathbf{P_A} \mathbf{D_K^{-1}}. \tag{4.5}$$

CA-ALT is performed through PCA($\mathbf{Q_A}, \mathbf{D_K}, \mathbf{D_J}$) that analyses both the dispersion of the cloud of category profiles insofar as explained by the dispersion of the cloud of frequency profiles and the dispersion of the frequency profiles insofar as explained by the dispersion of categories profiles. In other words, CA-ALT, as a double-projected analysis, allows for the variability of the rows to be explained in terms of the columns and the variability of the columns in terms of the rows.

Note that in this section, where only one categorical variable is considered, $\mathbf{X}$ includes non-centred dummy categories. The rows are endowed with the weighting system $\mathbf{D_I}$, the $(I \times I)$ diagonal matrix corresponding to the row margins of $\mathbf{P}$. $\mathbf{X}$ is a columnwise $\mathbf{D_I}$-orthogonal matrix making

$$\mathbf{D_K} = \mathbf{X^T D_I X} \tag{4.6}$$

diagonal. $d_{Kkk} = p_{A\bullet k}$ is the inertia of the dummy variable k relative to the origin (not the centroid). Thus, $\mathbf{D_K}$ acts as a covariance matrix when the analysis is performed on the non-centred matrix $\mathbf{X}$.

4.2.1 CA as a double projected analysis

The "inflated" $(N \times K)$ matrix $\mathbf{X_N}$ and the $(N \times J)$ matrix $\mathbf{Y_N}$ are considered. $\mathbf{X_N}$ and $\mathbf{Y_N}$ cross the N occurrences and, respectively, the K indicators corresponding to the column-categories of table $\mathbf{X}$ and the J column-words of table $\mathbf{Y}$. Each cell x_{Nnk} of table $\mathbf{X_N}$ is equal to "1" if occurrence n has been used by a respondent who presents category k and "0" otherwise. In like manner, each cell y_{Nnj} of table $\mathbf{Y_N}$ is equal to "1" if occurrence n corresponds to word j and "0" otherwise. A small example is given in Figure 4.2.

The "inflated" data matrices are defined by Legendre and Legendre (1998) as a solution to the fourth-corner problem that aims at studying the associations between the biological and behavioural characteristics of species and the habitat characteristics. Three matrices $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$ are defined. A matrix $\mathbf{A}$, with dimensions $J \times I$, contains data on the presence or absence of J species at I sites. A matrix $\mathbf{B}$, with dimensions $J \times T$, describes T biological or behavioural traits of the same J species. A matrix $\mathbf{C}$, with dimensions $K \times I$, contains information about K habitat characteristics at the I sites. Given the information in matrices $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$, the fourth-corner problem estimates the parameters in the fourth-corner matrix $\mathbf{D}$,

Inflated data matrices

	X_N			Y_N						
Occurrences	**< 21**	**21-50**	**>50**	**happy**	**in love**	**ill**	**tired**	**hungry**	**excited**	**sad**
Occ 1 @ Resp 1	1	0	0	1	0	0	0	0	0	0
Occ 2 @ Resp 1	1	0	0	0	1	0	0	0	0	0
Occ 3 @ Resp 2	0	0	1	0	0	1	0	0	0	0
Occ 4 @ Resp 3	0	1	0	0	0	0	1	0	0	0
Occ 5 @ Resp 3	0	1	0	0	0	0	0	1	0	0
Occ 6 @ Resp 4	1	0	0	0	0	0	0	0	1	0
Occ 7 @ Resp 4	1	0	0	0	1	0	0	0	0	0
Occ 8 @ Resp 5	0	0	1	0	0	1	0	0	0	0
Occ 9 @ Resp 5	0	0	1	0	0	0	1	0	0	0
Occ 10 @ Resp 5	0	0	1	0	0	0	0	0	0	1

Aggregated lexical table

	Y_A		
	< 21	21-50	>50
happy	1	0	0
in love	2	0	0
ill	0	0	2
tired	0	1	1
hungry	0	1	0
excited	1	0	0
sad	0	0	1

Fig. 4.2 Inflated data matrices (left): $\mathbf{X_N}$ and $\mathbf{Y_N}$. The aggregated lexical table (right): $\mathbf{Y_A}$

with dimensions $K \times T$, that crosses the biological and behavioural characteristics of species with the habitat characteristics.

A similar approach is considered to build the aggregated lexical table $\mathbf{Y_A}$ that crosses the words and the categories of the categorical variable (Figure 4.2)

$$\mathbf{Y_A} = \mathbf{Y_N^T X_N}. \tag{4.7}$$

The $(N \times N)$ diagonal matrix $\mathbf{D_N}$ corresponds to the uniform weighting system on the rows with a weight equal to $1/N$ for each row. Note that both the column-words and column-variables are in $\mathbb{R}^N$ space. The proportion matrix $\mathbf{P_A}$ can be rewritten as

$$\mathbf{P_A} = \mathbf{Y_N^T D_N X_N} \tag{4.8}$$

and matrix $\mathbf{Q_A}$ as

$$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A D_K^{-1}} = (\mathbf{Y_N^T D_N Y_N})^{-1}(\mathbf{Y_N^T D_N X_N})(\mathbf{X_N^T D_N X_N})^{-1}. \tag{4.9}$$

Eq.(4.9) shows that the columns of $\mathbf{Q_A}$ are the $\mathbf{D_N}$-orthogonal projection of the dummy-columns of $\mathbf{X_N D_K^{-1}} = \mathbf{X_N}(\mathbf{X_N^T D_N X_N})^{-1}$ on the subspace of $\mathbb{R}^N$ generated by the column-words of $\mathbf{Y_N}$. Similarly, the same Eq.(4.9) shows that the rows of $\mathbf{Q_A}$ are the $\mathbf{D_N}$-orthogonal projection of the column-words of $\mathbf{Y_N D_J^{-1}} = \mathbf{Y_N}(\mathbf{Y_N^T D_N Y_N})^{-1}$ on the subspace of $\mathbb{R}^N$ generated by the dummy-columns of $\mathbf{X_N}$. This viewpoint highlights that CA, as a double projected analysis, studies both the variability of the cloud of words, insofar as it is explained by the variability of the categories, and the variability of the cloud of categories, insofar it is explained by the variability of the words.

4.3 Analysis of a multiple aggregated lexical table

4.3.1 Classical correspondence analysis on a multiple aggregated lexical table

We may be interested in a broader context, such as a set of Q categorical variables (Q>1). As the starting point, the multiple aggregated lexical table is built juxtaposing row-wise the L aggregated lexical tables built from the L categorical variables (Figure 4.3).

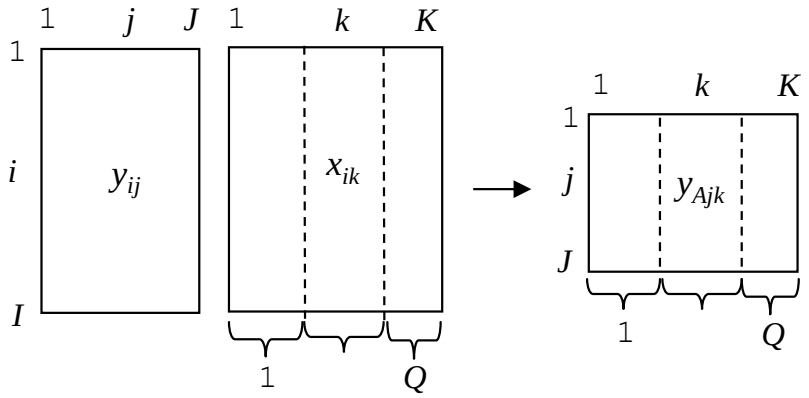


Fig. 4.3 Multiple aggregated lexical table

We follow a rationale akin to that of the former section. The aggregated lexical table

$$\mathbf{Y_A} = \mathbf{Y^T X} \tag{4.10}$$

is built and transformed into the proportion matrix

$$\mathbf{P_A} = \frac{\mathbf{Y_A}}{L \times N}. \tag{4.11}$$

Diagonal matrix $\mathbf{D_K}$ stores the column margins of $\mathbf{P_A}$ whose general term is the proportion of occurrences corresponding to category k. From $\mathbf{P_A}$, matrix

$$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A D_K^{-1}} \tag{4.12}$$

is computed. Then, PCA($\mathbf{Q_A}, \mathbf{D_K}, \mathbf{D_J}$) is performed.

The main difference with the former section is that $\mathbf{D_K}$ is no longer equal to $\mathbf{X^T D_I X}$. This latter matrix presents non-null off-diagonal terms because the column-categories of

$\mathbf{X}$ are generally not orthogonal when belonging to different variables. It is no longer a double-projected analysis and hence the influence of the associations among the categories of different variables is not filtered.

4.3.2 CA on a generalised aggregated lexical table

In this section, the dummy columns of $\mathbf{X}$ are centred. As in the former section, the aggregated lexical tables built from each categorical variable are juxtaposed row-wise into the $(J \times K)$ matrix $\mathbf{Y_A}$. This column-centred multiple aggregated lexical table is called *generalised aggregated lexical table* (GALT).

To maintain a double projected analysis as in CA, the starting point consists in substituting the matrix $\mathbf{D_K^{-1}}$ by the Moore-Penrose pseudoinverse $\mathbf{C}^-$ of

$$\mathbf{C} = (\mathbf{X^T D_I X}). \tag{4.13}$$

The $(K \times K)$ matrix $\mathbf{C}$ is the covariance matrix between the columns of $\mathbf{X}$ taking into account that the respondents are endowed with weighting system $\mathbf{D_I}$.

$\mathbf{C}^-$ substitutes $\mathbf{D_K^{-1}}$ in the expression of the $(J \times K)$ matrix

$$\bar{\mathbf{Q}}_\mathbf{A} = \mathbf{D_J^{-1} P_A C^-}. \tag{4.14}$$

The matrix $\mathbf{C}^-$ operates a multivariate standardisation. The Mahalanobis transformation, a linear transformation

$$\bar{\mathbf{X}} = \mathbf{X}(\mathbf{C}^-)^{1/2} \tag{4.15}$$

of data matrix $\mathbf{X}$, standardizes the variance of each variable and eliminates the correlation between the variables making them uncorrelated (Brandimarte, 2011; Härdle and Simar, 2012).

Remark: to compute $(\mathbf{C}^-)^{1/2}$, $\mathbf{C}$ is diagonalised and the whole of its S_C non-null eigenvalues, all positive, are ranked in descending order and stored in the $(S_C \times S_C)$ diagonal matrix $\Lambda_\mathbf{C}$. S_C is equal to the dimension of the space spanned by the columns of $\mathbf{X}$, that is, the number of independent dummy-columns of $\mathbf{X}$. The corresponding eigenvectors are stored in the columns of the $(K \times S_C)$ matrix $\mathbf{U_C}$. The S_C columns of $\mathbf{X}(\mathbf{C}^-)^{1/2}$, with $(\mathbf{C}^-)^{1/2} = \mathbf{U_C}\Lambda_\mathbf{C}^{-1/2}$, are standardised and uncorrelated.

CA-GALT is performed through PCA($\bar{\mathbf{Q}}_\mathbf{A}, \mathbf{C}, \mathbf{D_J}$) that is equivalent to analyse the column-centred multiple aggregated lexical table $\mathbf{P_A}$ through CA with a modified metric $\mathbf{C}$ in the row space.

As in any PCA, the eigenvalues are stored into the $(S \times S)$ diagonal matrix Λ, and the eigenvectors into the $(K \times S)$ matrix $\mathbf{U}$. The coordinates of the row-frequencies are computed as

$$\mathbf{F} = \bar{\mathbf{Q}}_\mathbf{A}\mathbf{C}\mathbf{U} \tag{4.16}$$

and the column-categories as

$$\mathbf{G} = \bar{\mathbf{Q}}_\mathbf{A}^\mathbf{T}\mathbf{D_J}\mathbf{F}\Lambda^{-1/2} \tag{4.17}$$

using the transition relationships. The interpretation rules of the results of this specific CA are the usual CA interpretation rules (Escofier and Pagès, 1988; Greenacre, 1984; Lebart et al., 1998)

The matrix

$$\mathbf{F} = \bar{\mathbf{Q}}_\mathbf{A}\mathbf{C}\mathbf{G}\Lambda^{-1/2} = \mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{C}^-\mathbf{C}\mathbf{G}\Lambda^{-1/2} = \mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{G}\Lambda^{-1/2} \tag{4.18}$$

shows that, as in classical CA, a word is placed on axis s, up to a coefficient varying from one axis to the other, at the centroid of the categories that use it, endowing the categories with the weighting system $p_{Ajk}/p_{Aj\bullet}$ $(k = 1, ..., K)$.

Considering the matrices $\mathbf{Y_N} = [\mathbf{y_{Nnj}}]$ and $\mathbf{X_N} = [\mathbf{x_{Nnk}}]$ defined similarly to those in section 4.2.1, but $\mathbf{X_N}$ now comprising the K centred dummy columns corresponding to all the categories of the selected categorical variables, the matrix

$$\mathbf{G} = \bar{\mathbf{Q}}_\mathbf{A}^\mathbf{T}\mathbf{D_J}\mathbf{F}\Lambda^{-1/2} = (\mathbf{D_J^{-1}}\mathbf{P_A}\mathbf{C}^-)^\mathbf{T}\mathbf{D_J}\mathbf{F}\Lambda^{-1/2} = \mathbf{C}^-\mathbf{P_A^T}\mathbf{D_J^{-1}}\mathbf{D_J}\mathbf{F}\Lambda^{-1/2} = \mathbf{C}^-\mathbf{P_A^T}\mathbf{F}\Lambda^{-1/2} \tag{4.19}$$

can be rewritten as

$$\mathbf{G} = (\mathbf{X^T}\mathbf{D_I}\mathbf{X})^- \frac{\mathbf{X^T}\mathbf{Y}}{N \times L}\mathbf{F}\Lambda^{-1/2} = ((\mathbf{X_N^T}\mathbf{D_N}\mathbf{X_N})^-\mathbf{X_N^T}\mathbf{D_N}\mathbf{Y_N}/L)\mathbf{F}\Lambda^{-1/2} = \mathbf{B}\mathbf{F}\Lambda^{-1/2}. \tag{4.20}$$

The $(K \times J)$ matrix

$$\mathbf{B} = (\mathbf{X_N^T}\mathbf{D_N}\mathbf{X_N})^-\mathbf{X_N^T}\mathbf{D_N}\mathbf{Y_N}/L = [b_{kj}] \tag{4.21}$$

is, except for the scaling coefficient $1/L$, the matrix of regression coefficients (strictly, analysis of variance coefficients given that the regressors are dummy variables) of all the column-words of $\mathbf{Y_N}$ on the regressor column-categories of $\mathbf{X_N}$. These coefficients are issued from the simultaneous, or multivariate, linear regression of all the column-words of $\mathbf{Y_N}$ on the column-categories of $\mathbf{X_N}$ (Finn, 1974).

A category k is placed on axis s, up to a coefficient varying from one axis to the other, at the centroid of the words, endowing them with the weighting system b_{kj} ($j = 1, ..., J$). The weight given to word j, equal to b_{kj}, is the coefficient of category k in the regression of column-word j on all the categories. Thus, a category is placed in the direction of the words that the respondents belonging to this category tend to use, all things being equal.

The respondents also can be reintroduced in the analysis by positioning the columns of $\mathbf{Q} = \mathbf{D_I^{-1}PD_J^{-1}}$ as supplementary columns in PCA($\bar{\mathbf{Q}}_\mathbf{A}, \mathbf{C}, \mathbf{D_J}$). So, the respondents are placed for each dimension at the weighted centroid, up to a constant, of the frequencies that they use. Thus, their coordinates are computed via the transition relationships as

$$\mathbf{G}^+ = \mathbf{D_I^{-1}PF}\Lambda^{-1/2}. \tag{4.22}$$

CA-GALT for quantitative contextual variables

The case of the quantitative variables is detailed in Bécue-Bertaut and Pagès (2014). The same rationale is followed, but defining:

- The GALT $\mathbf{Y_A}$ is built as $\mathbf{Y_A} = \mathbf{Y^T X}$ and transformed into the matrix $\mathbf{P_A} = \mathbf{P^T X}$ whose generic term $p_{Ajk} = \sum_i p_{ij} x_{ik}$ is equal to the weighted sum of the values assumed for variable k by the respondents who used word j.
- The matrix $\mathbf{D_K^{-1}}$ no longer exists because the column-margins of $\mathbf{P_A}$ do not correspond to a weighting system on the columns.

CA-GALT on principal components

A stability problem arises when addressing multicollinearity among the contextual variables and/or noisy data, which is very usual in survey data. Substituting the original variables by their principal components and eliminating those with low inertia relieve these problems. Thus, proposal involves performing a previous principal component method, PCA($\mathbf{X}, \mathbf{I}, \mathbf{D_I}$). The S (S$\leq$ K) first principal components corresponding to the contextual variables are stored

into the $(I \times S)$ matrix Ψ. The first eigenvalues are stored in the $(S \times S)$ diagonal matrix Δ. The final CA-GALT procedure substitutes $\mathbf{X}$ by Ψ and $\mathbf{C}$ by Δ in Eq.(4.14). If all of the components were used, the results would be identical to those obtained when operating from the original contextual variables.

Validation tools

Several validation tools were provided to validate and improve the results obtained with CA-GALT as tests of significance on the first eigenvalue and the sum of eigenvalues, confidence ellipses, association between vocabulary and contextual variables and association between words and contextual variables (Bécue-Bertaut and Pagès, 2014).

- **Tests of significance on the first eigenvalue and the sum of eigenvalues**: have proposed tests to contrast whether the uncovered structure is actually due to the relationship between the lexical table $\mathbf{Y}$ and the contextual table $\mathbf{X}$. These tests are suitable for CA-GALT. Two statistics are used, the sum of eigenvalues and the first eigenvalue. For each statistic, the null distribution is generated by randomly permuting the row-respondents of solely table $\mathbf{X}$ (equivalently, of solely table $\mathbf{Y}$) and computing CA-GALT and the values of the statistic for each permutation. By placing the observed value in this null distribution, the corresponding p-value is computed.

- **Confidence ellipses**: To validate and to better interpret the representation of the words and variables, confidence ellipses based on the bootstrap principles (Efron, 1979) are performed. To do that, replicated samples are randomly extracted, with replacement, from the original sample. From each replicated sample, the replicated tables words $\times$ principal-components and words $\times$ variables are built. The row-words and column-variables of these tables are projected as illustrative elements on the corresponding planes provided by the analysis of the "true" table. The cloud of replicated points corresponding to each word/variable are represented through an ellipse designed such that the $\alpha\%$ more extreme points are excluded when a confidence level equal to (1-α)% is selected.

- **Association between vocabulary and contextual variables**: To facilitate the interpretation, the convenient contextual variables are selected, studying the association between the vocabulary and contextual variables. The vocabulary is considered non-associated with a given contextual variable if the words do not significantly differ in terms of the values assumed by the respondents who use them. In the case of categorical contextual variables, the association between the vocabulary and a categorical

variable is assessed by building the table counting the frequency of every word in every category; then, the classical chi-square statistic is used and the significance computed through a permutation test. In the case of quantitative variables, word j assumes, for contextual variable k (k=1,...,K), the weighted average $\bar{x}_k^j = \sum_{i \in I}(p_{ij}/p_{i\bullet})x_{ik}$ of the values assumed by the respondents who used it. The problem is placed within the framework of the analysis of variance (ANOVA). The variability of the occurrences is decomposed being the total variability equal to (N-1)s_k^2, where s_k^2 is the variance of the contextual variable k. This total variability is decomposed into the variability between words (SS_B as explained variability) and the variability within-words (SS_W as non-explained variability). The usual indicator, referred to as the association-with-words ratio, that measures the intensity of the association between the vocabulary and a contextual variable k is defined as

$$R^2 = \frac{SS_B}{(N-1)s_k^2}. \tag{4.23}$$

In this case, the assumptions regarding the validity of the usual F test are not verified as the occurrences are not independent. Thus a permutation procedure is used. At each step, the occurrences of the respondents' answers are randomly extracted, without replacement, among the N occurrences, disregarding the words to which they belong.

- **Association between words and contextual variables**: When the association between the vocabulary and a contextual variable k is assessed, the words responsible for this association are identified defining the words characteristic of each category as detailed in Lebart et al. (1998). The statistic used (one-tail test) for this purpose is

$$\frac{\bar{x}_k^j - \bar{x}_k}{s_{\bar{x}_k^j}} = \frac{(\bar{x}_k^j - \bar{x}_k)\sqrt{y_{\bullet j}}}{s_k}\sqrt{\frac{N-1}{N-y_{\bullet j}}} \tag{4.24}$$

where $s_{\bar{x}_k^j}$ is the variance of the contextual variable k assumed by word j and $y_{\bullet j}$ the number of occurrences of word j. The low count of many words as well as the non-independence of the occurrences belonging to the same free answer favour resorting to permutations to compute the significance.

Comparison with other methodologies

Bécue-Bertaut and Pagès (2014) also details the comparison of CA-GALT with other methodologies dealing with the same data structure such as Canonical Correspondence Analysis

(CCA; ter Braak, 1986), method developed to analyse ecological data studying the relationships between sites and species insofar as they are explained by the environmental variables, and MFACT.

Comparison with CCA, applied to textual data, and CA-GALT result in the same factors on the row-words. However, the factors on the contextual variables and the respondents differ from CA-GALT to CCA. In the latter method, the contextual variables are reintroduced further as illustrative and placed on the axes through the correlations between the columns of $\mathbf{X}$ and the factors on the respondents (active elements). The correlations are not eliminated and, consequently, similarities among variables reflect both their correlations and their possible associations with the words.

These latter results are in accordance with the different objectives of both methods. CA-GALT studies the relationships between the words and the contextual variables, leaving the respondents in a second level. CCA favours the relationships between the respondents and the words insofar as they are explained by the contextual variables.

Regarding to the comparison with MFACT, this latter aims at establishing a global typology of the respondents from all of the sets of columns and, possibly, to compare this typology with the typologies that could be separately obtained from each set of columns. It places the word and contextual variables in the same space, and their representation offers a framework for comparison. However, the relationships between words and contextual variables are created through the respondents without eliminating the correlations among the contextual variables. MFACT does not offer a direct view that highlights the associations between words and contextual variables.

The main point to stand out is multivariate standardisation operated by CA-GALT. This standardisation cancels the associations among the variables avoiding "confusion effect". This is an essential property of CA-GALT that differentiates it from the other methods dealing with the same data structure. Application to a survey presented in the next section will help to ease understanding of this idea.

4.4 Application to a survey

The example is extracted from a survey intended to better know the definitions of health that the non-experts give. An open-ended question "*What does health mean to you?*" was asked to 392 respondents who answered through free-text comments. The respondents $\times$ words table is built keeping only the words used at least 10 times among all respondents. This

minimum threshold is used to obtain statistically interpretable results (Lebart et al., 1998; Murtagh, 2005). Thus, 126 different words and 7751 occurrences are kept.

The respondents' characteristics are also collected. In this example, we use age in groups (under 21, 21-35, 36-50 and over 50), gender (man and woman) and health condition (poor, fair, good and very good health) as they possibly condition the respondents' viewpoint.

CA-GALT is able to determine the main dispersion dimensions as much as they are related to the respondents' characteristics.

The data is coded into two matrices (Figure 4.4). The $(I \times J)$ matrix $\mathbf{Y}$, with generic term y_{ij}, contains J frequency columns (words) and I rows (respondents). The $(I \times K)$ matrix $\mathbf{X}$, with generic term x_{ik}, stores the K variables (categories): x_{ik}=1 if i belongs to the category k and x_{ik}=0 otherwise.

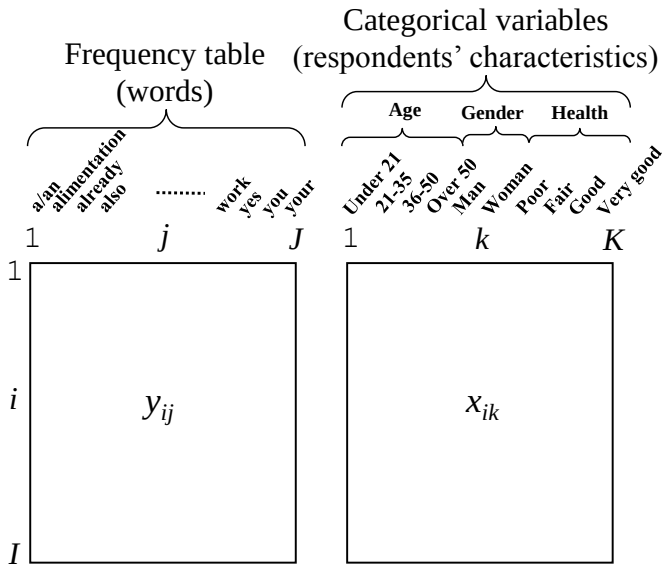


Fig. 4.4 The data set. On the left, the frequency table $\mathbf{Y}$; on the right the categorical table $\mathbf{X}$. In the example, $I = 392$ (respondents), $J = 126$ (words), $K = 10$ (categories)

First, a classical CA is applied on the multiple aggregated lexical table built by juxtaposing the three aggregated lexical tables issued from using age group (four categories), gender (two categories) and health condition (four categories). Then, CA-GALT is performed with the three previous variables as contextual variables. Confidence ellipses are performed to validate and better interpret the representation of the words and the variables. The comparison of the results obtained from these two methods allows for demonstrating the effectiveness of CA-GALT.

4.4.1 Classical CA on the multiple aggregated lexical table

CA is applied to the multiple aggregated lexical table. The total inertia is equal to 0.072. *Age group*, *health condition* and *gender* contribute to this total inertia bringing, respectively 49.4%, 33.0% and 17.6% of this total inertia. The first two axes, whose inertia are respectively 0.026 and 0.013, keep together 54.6% of the total inertia.

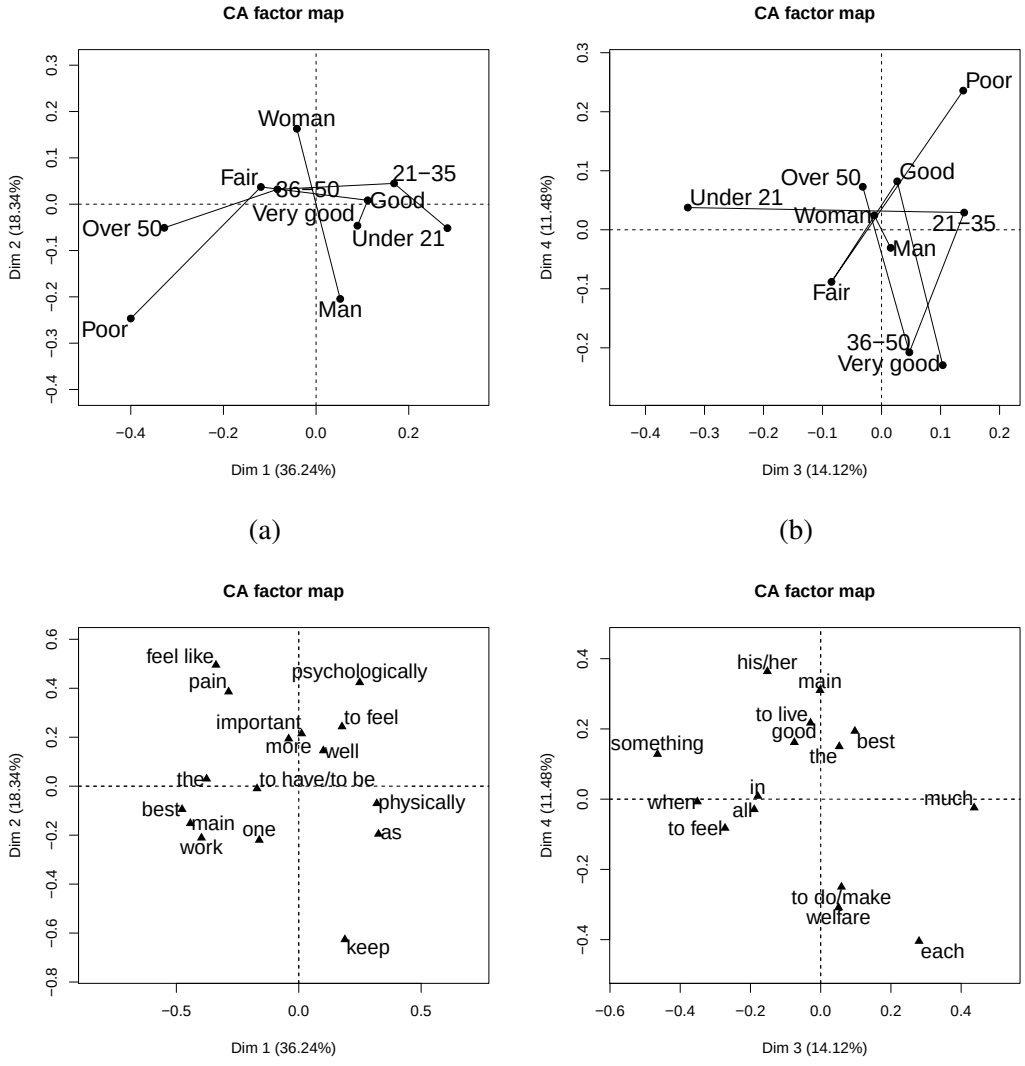


Fig. 4.5 Categories and contributory words on the CA planes (1,2) and (3,4)

Figure 4.5a offers the representation of the categories on the plane (1,2). The trajectory of *age group* categories notably follows the first axis, outlining a weak arch effect. This axis ranks, in their natural order, the *health condition* categories except for the inversion

between *very good health* and *good health* which lie very close. The extreme categories of this variable, very particularly *poor health*, are opposed to the intermediate categories on the second axis, indicating a more pronounced arch effect than *age group*. However, the main opposition on the second axis concerns the two *gender* categories so that *age group* and *gender* are practically orthogonal, this in terms of the vocabulary that they use. Regarding the plane (3,4) (Figure 4.5b), the third axis shows that young people (*under21*), besides using words close to those used by the following *age group* as revealed by axis one, also express themselves with their own words. No clear pattern stands out on the fourth axis.

The representation of the words with a high contribution (Figures 4.5c and 4.5d) brings information about the meaning of the oppositions and trajectories, showing for example that the words *the*, *best*, *main* and *work* are words both used by the oldest and/or less healthy categories and avoided by the youngest and/or more healthy categories. However, one might wonder if the choice/rejection of these words is related to *age* or to *health condition* or to both.

Two variables are strongly associated but still that the association is sufficiently loose as to allow for untangling the influence of both variables on the vocabulary, provided that an adequate method is applied. Precisely, CA-GALT offers a suitable approach because the associations between the variables are cancelled.

4.4.2 CA-GALT

The total inertia is equal to 0.21. The first two axes are moderately dominant with eigenvalues equal to 0.0636 (30.81% of the inertia) and 0.03882 (18.78%).

Figures 4.6a and 4.7a display, respectively, the contextual variables and the words with a high contribution on the CA-GALT first principal plane. These representations are completed by drawing confidence ellipses. Only the confidence ellipses around the words *the*, *best*, *main* and *work* are represented, because these words are favoured as examples to show the effectiveness of the approach that we propose (Figures 4.7c and 4.7d). If all the ellipses were drawn, only those around *he/she* and *to be able*, on plane (1,2) and around *to be able* and *from* on plane (1,4) would overlap the centroid.

As in the former analysis, the trajectory of *age* group notably follows the first axis (Figure 4.6a). The extreme categories of this variable, *over 50* (at the left); *under 21* (at the right) bring, respectively, 52.1% and 23.1% of this axis inertia. However, *health condition* representation differs after cancelling the associations between *age* and *health condition*. The categories of this variable now lie close to the centroid on the first plane and their confidence

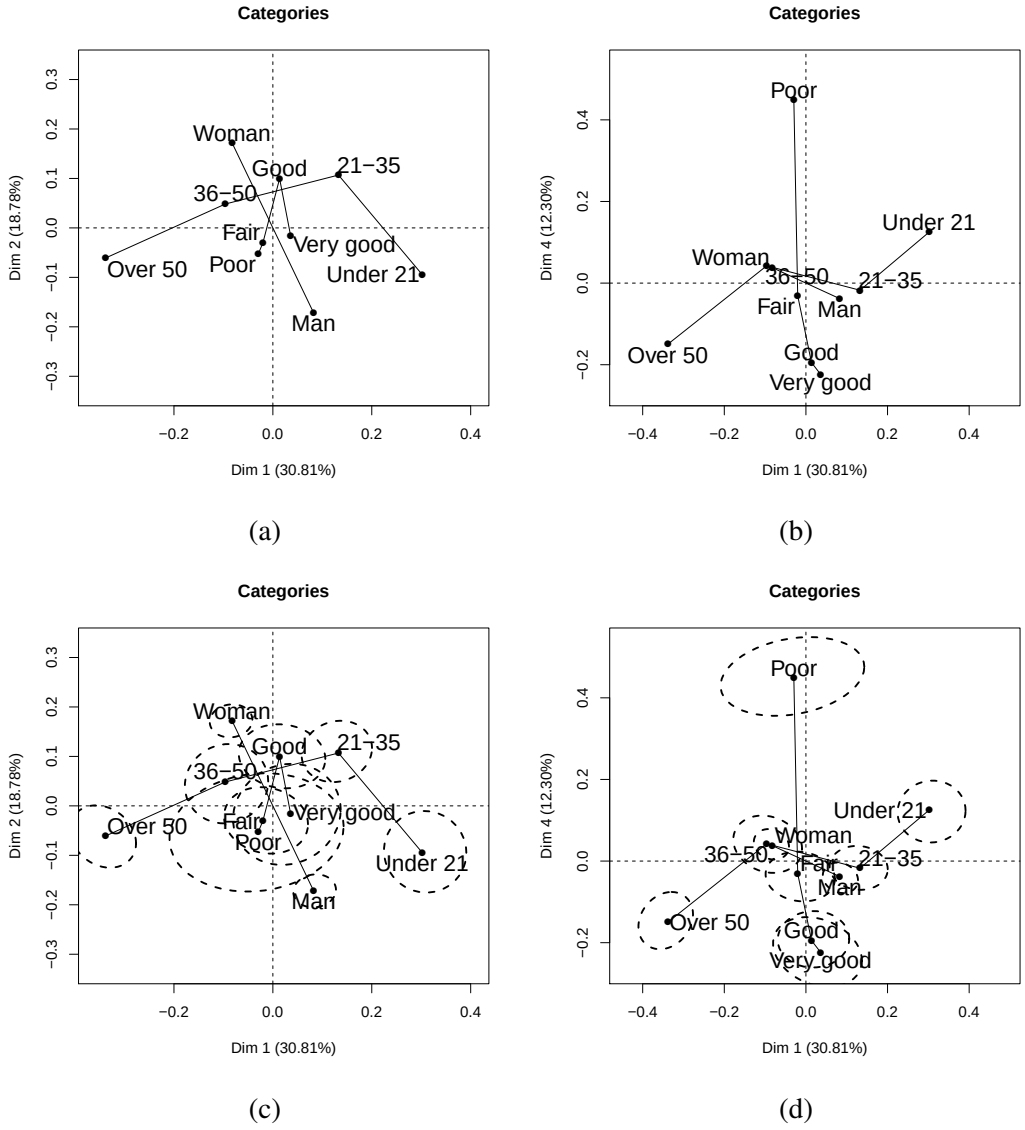


Fig. 4.6 Categories on the CA-GALT planes (1,2) and (1,4) completed by confidence ellipses

ellipses extensively overlap one another (Figure 4.6c). Regarding the words, we find again *the*, *best* and *main* with high coordinates on the left of the axis (Figure 4.7a), indicating that they are words both very used by the oldest categories and avoided by the youngest. The word *work* is no longer present on this graphic, since it is close to the centroid and thus not a key word for the oldest categories.

We detail neither the second axis, which opposes *Man* and *Woman*, nor the third (not reproduced in the graphic), which highlights the specific use of the vocabulary by the *under 21* respondents. Both axes are close to those computed in the former CA. However, the fourth

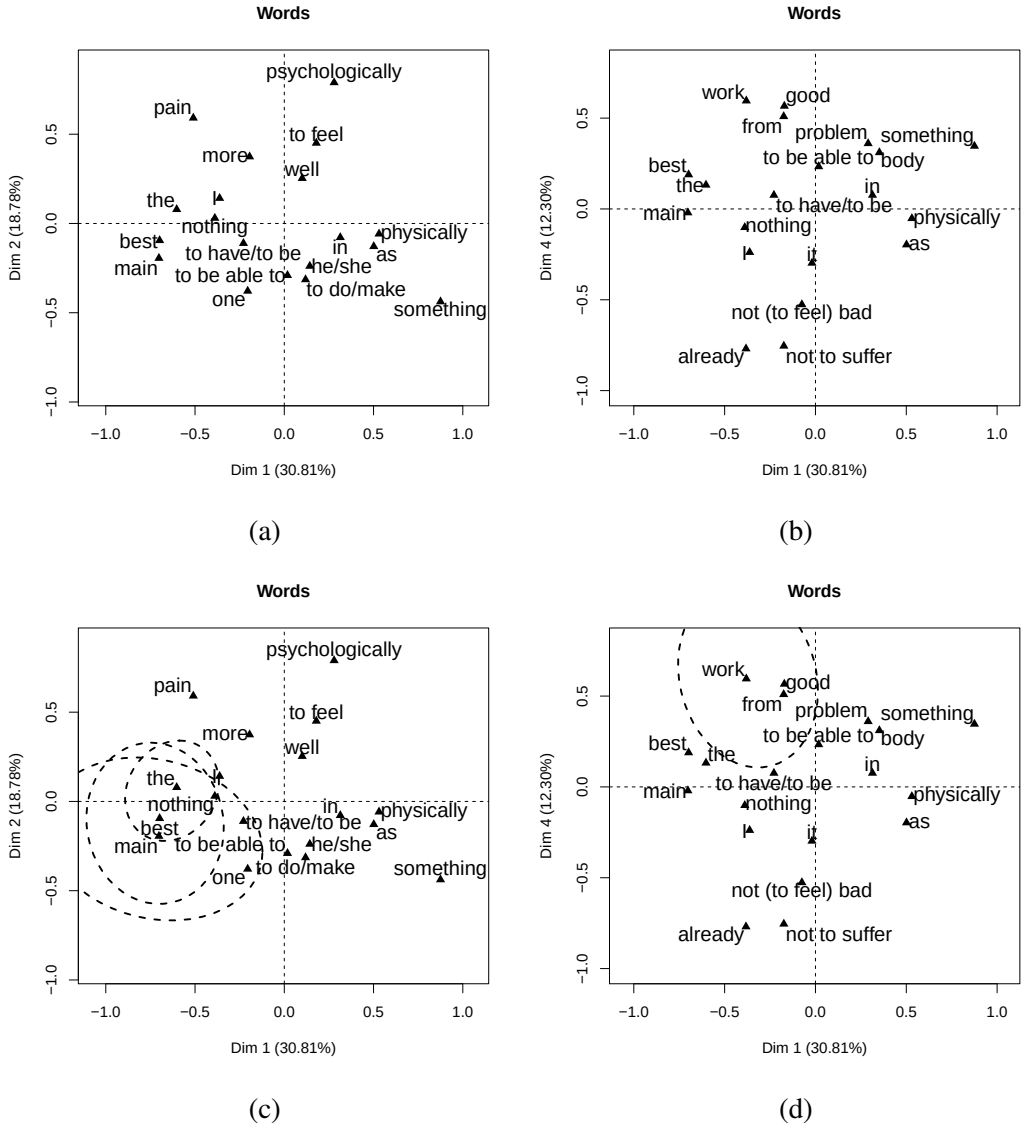


Fig. 4.7 Contributory words on the CA-GALT planes (1,2) and (1,4) completed by confidence ellipses

axis turns out to be of interest because of ranking *health condition* categories in their natural order (Figure 4.6b). These categories, which provide together 75% of the axis inertia, are well separated, except for the two better health categories whose confidence ellipses overlap (Figure 4.6d). The word *work* lies close to *poor health* category in the positive part of the fourth axis (Figures 4.6b and 4.7b), pointing out a strong association between this word and this category and also little use of this word by the most healthy categories. The word *work* contrasts on the fourth axis with *bad*, *suffer* and *already* which are associated with *good health*, *very good health* and *over 50*. These latter words are used in free answers where

health is defined through negative expressions such as *not to feel bad*, *not to be bad*, *not to suffer*, *not to suffer* from any disease or pain.

4.4.3 Conclusions on CA-GALT application

The application of the method to a real data set has demonstrated how free-text and closed answers combine to provide relevant information.

The comparison of the results obtained from two methods, classical CA on the multiple aggregated table and CA-GALT, allows for demonstrating the effectiveness of the second.

In the first method, the projection of the categories, at the centroid of the respondents who belong to them, allows for detecting which variables (and which categories) are strongly associated with the words. However, the effects of the different variables are merged.

Meanwhile, CA-GALT takes into account several variables untangling their respective influence on the lexical choices and avoiding spurious relationships between certain categories and certain words. The influence of each variable on the lexical choices is visualised, avoiding "confusion effect".

Although this methodology was developed mainly to give an answer to the problem of analyzing open-ended questions with several quantitative/categorical variables, it can be applied to any kind of frequency/contingency table connected by contextual variables such as ecological studies that aim to study the relationships between species distribution and environmental variables.

Chapter 5

Multiple Factor Analysis on Generalised Aggregated Lexical Table

This chapter is devoted to detail the method that we propose to deal with a sequence of coupled tables. This method combines MFACT and CA-GALT. The presentation will be illustrated by using the results obtained in two examples. The first example corresponds to an open-ended question answered in different languages. The second example deals with ecological data.

5.1 Methodology

We first remember the data structure and the notation. For every sample l $(l = 1, .., L)$, there is a frequency table $\mathbf{Y_l}$ of dimensions $(I_l \times J_l)$ and a contextual variables table $\mathbf{X_l}$ of dimensions $(I_l \times K)$. $(\mathbf{Y_l}, \mathbf{X_l})$ is the "coupled table" that corresponds to sample l. The global resulting data structure is a sequence of L coupled tables (Figure 5.1).

5.1.1 Starting point

To ease the reading, we describe the starting point in terms of a concrete example. This is related to the open-ended question answered. The data structure and notation are presented in Figure 5.2. In each sample, we have the coupled tables $(\mathbf{Y_l}, \mathbf{X_l})$. $\mathbf{Y_l}$, with dimensions $(I_l \times J_l)$, is a lexical table containing the frequency with which every respondent has used each of

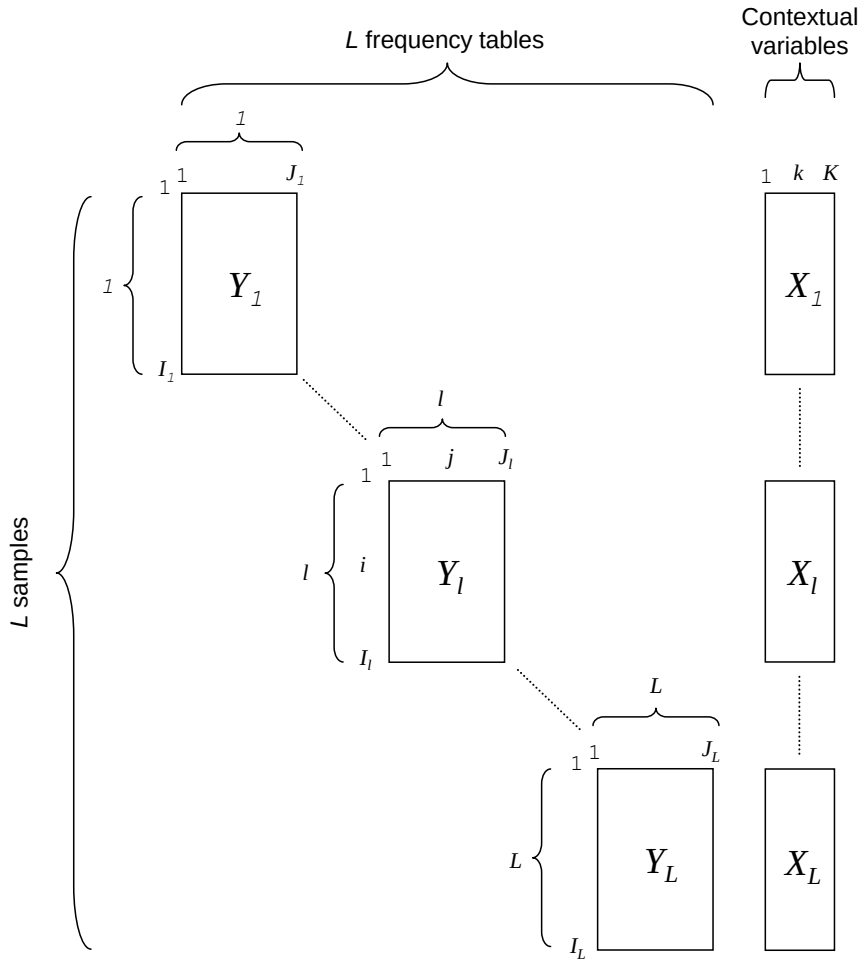


Fig. 5.1 Data structure

the J_l words used in language l. The $(I_l \times K)$ matrix $\mathbf{X_l}$, individuals×quantitative variables, contains the values presented by the respondents for every variable (here, satisfaction scores).

By applying CA-GALT rationale, separately in each sample, both generalised aggregated lexical tables (GALT) $\mathbf{Y_A^1}$ and $\mathbf{Y_A^2}$ are built as $\mathbf{Y_A^l} = \mathbf{Y_l^T X_l}$. In the GALT $\mathbf{Y_A^l}$, every word is placed on the variables at the weighted sum of the values of the respondents who use them. As a result, both subsets of words, rows of $\mathbf{Y_A^1}$ and $\mathbf{Y_A^2}$, respectively, are placed in $\mathbb{R}^K$.

These GALT, $\mathbf{Y_A^1}$ and $\mathbf{Y_A^2}$, play the role that the aggregated lexical tables play in MFACT (section 3.3), except that we now have to deal with quantitative variables.

$\mathbf{Y_A^1}$ and $\mathbf{Y_A^2}$ are juxtaposed into a global table $\mathbf{Y_A}$. How the juxtaposed table is constructed suggests the starting point of the methodology, that is, to combine CA-GALT-like viewpoint (to adopt metrics similar to those used in CA-GALT) and MFA-like viewpoint (to balance

the influence of the subsets in the global analysis and to conserve the separation of the sets in order to compare the structures of the tables through partial analyses).

Remark: in most cases, MFA considers a set of column-variables; however in this case, sets of row-words are considered.

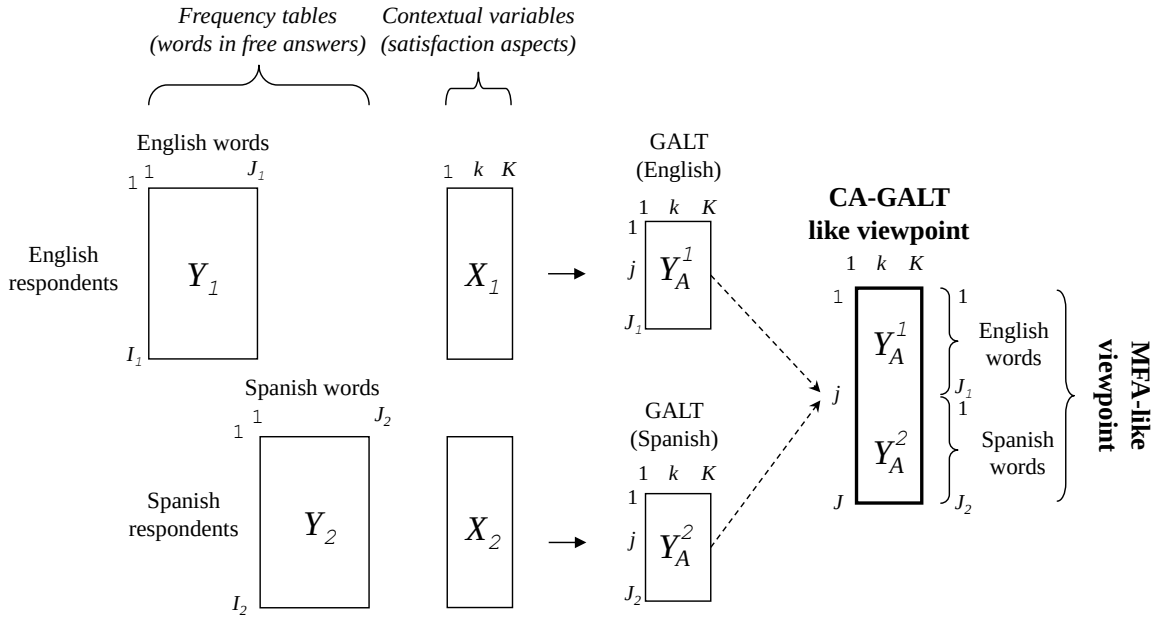


Fig. 5.2 Methodology

5.1.2 Rationale of MFACT

We turn back to the rationale of MFACT, briefly presented in section 3.3.2. We will detail in more depth the algorithm in a way that eases its extensions to several categorical or quantitative variables, extensions that are presented in the following sections.

The data structure is the one presented in Figure 5.1, but a single categorical variable is considered. The columns of $\mathbf{X_l}$ are non-centred dummy variables corresponding to its categories.

The rationale can be summed up by five points: separate proportion tables, multiple aggregated lexical table building, definition of the weights and metrics, model matrix and analysis through MFACT.

Proportion tables

For each set l $(l = 1,...,L)$, the $(I_l \times J_l)$ proportion matrix $\mathbf{P_l}$ is computed as

$$\mathbf{P_l} = \mathbf{Y_l}/N_l \tag{5.1}$$

with $N_l = \sum_{i \in I_l} \sum_{j \in J_l} y_{ijl}$, grand total of table $\mathbf{Y_l}$ (Figure 5.3). $N = \sum_{l \in L} N_l$ is the grand total through all the frequency tables $\mathbf{Y_l}$.

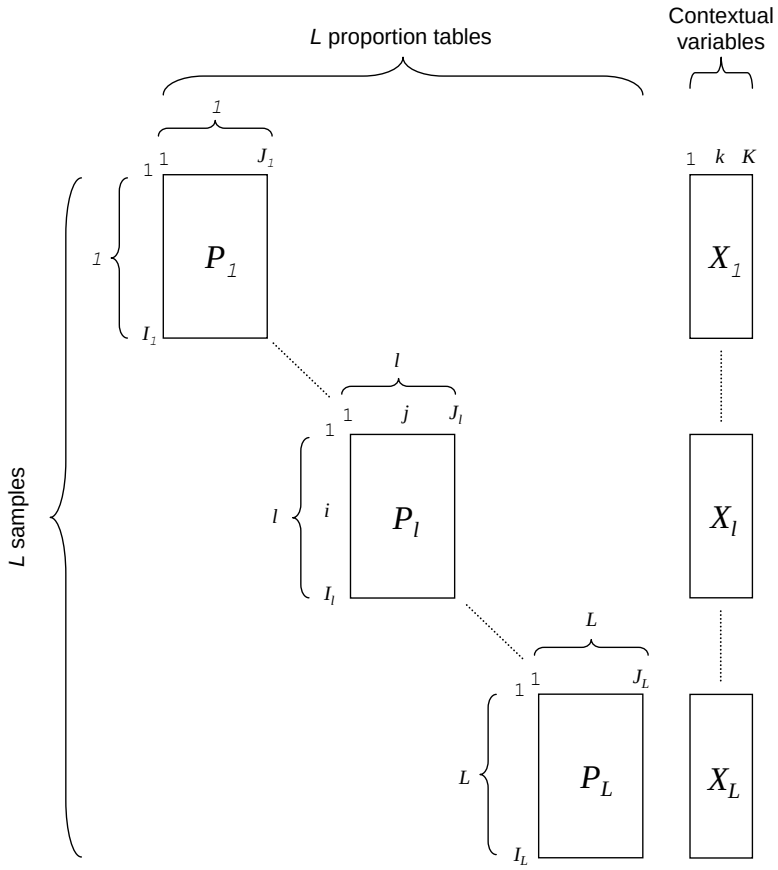


Fig. 5.3 Proportion tables

The generic term of the row and column margins of $\mathbf{P_l}$ are devoted by $p_{i\bullet l} = \sum_{j \in J_l} p_{ijl}$ and $p_{\bullet jl} = \sum_{i \in I_l} p_{ijl}$. The sum of terms of matrix $\mathbf{P_l}$ is equal to 1: $\sum_{i \in I_l} \sum_{j \in J_l} p_{ijl}$=1.

The relative importance of each frequency table is

$$\omega_l = N_l/N. \tag{5.2}$$

Multiple aggregated lexical table building

For each set l $(l = 1, ..., L)$, the $(J_l \times K)$ aggregated lexical table

$$\mathbf{Y_A^l} = \mathbf{Y_l^T X_l} \tag{5.3}$$

is built, crossing the words used in language l (in rows) and the categories of the categorical variable (in columns).

The L aggregated lexical tables $\mathbf{Y_A^1}$,...,$\mathbf{Y_A^l}$,...,$\mathbf{Y_A^L}$ are juxtaposed column-wise into the $(J \times K)$ multiple aggregated lexical table $\mathbf{Y_A}$. $\mathbf{Y_A}$ generic term y_{Ajkl} is the number of occurrences of the word j used by the respondents of category k in language l.

The $(J \times K)$ frequency table $\mathbf{Y_A}$ is transformed into the $(J \times K)$ proportion matrix

$$\mathbf{P_A} = \mathbf{Y_A}/N \tag{5.4}$$

whose generic term p_{Ajkl} is equal to the proportion of occurrences of word j used by the respondents of category k in language l; $\sum_{l \in L} \sum_{j \in J_l} \sum_{k \in K} p_{Ajkl}$=1.

The global proportion matrix $\mathbf{P_A}$ juxtaposes column-wise the L separate proportion matrices (Figure 5.4)

$$\mathbf{P_A^l} = \mathbf{Y_A^l}/N. \tag{5.5}$$

MFACT, as an extension of CA, analyses the weighted deviation between $\mathbf{P_A}$ and a model matrix $\mathbf{M}$. The row and column spaces are endowed with metrics and weighting systems chosen, as well as the model $\mathbf{M}$, such as to adopt a CA-like approach.

Metrics and weights

The $(J \times J)$ diagonal matrix

$$\mathbf{D_J} = [d_{Jjj}] = [p_{Aj\bullet l}] \tag{5.6}$$

and the $(K \times K)$ diagonal matrix

$$\mathbf{D_K} = [d_{Kkk}] = [p_{A\bullet k\bullet}] \tag{5.7}$$

issued from the margins of $\mathbf{P_A}$ correspond, respectively, to the weighting system on the row-words and the column-categories.

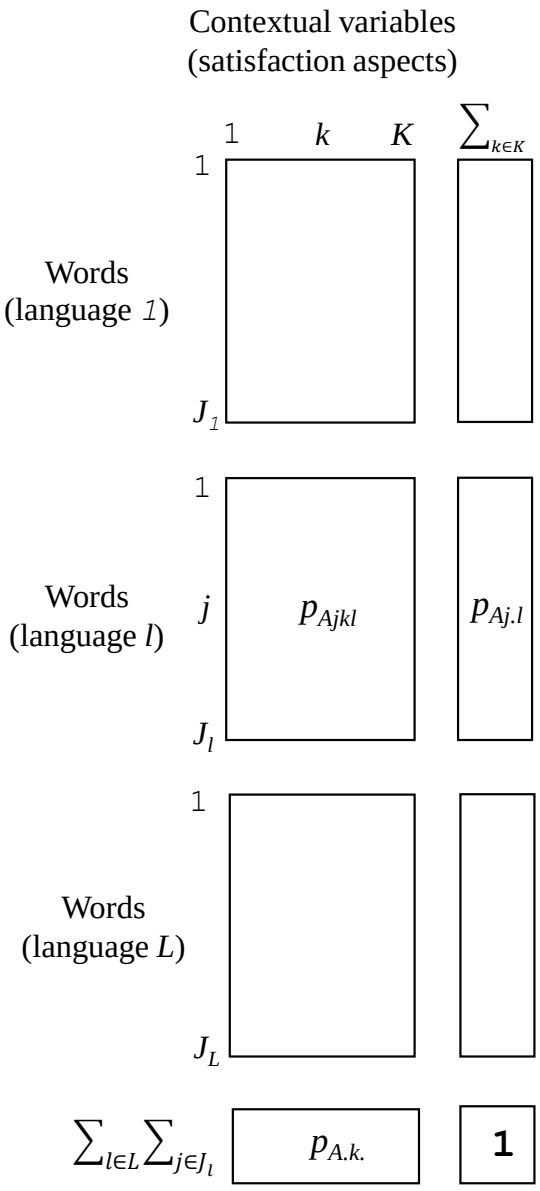


Fig. 5.4 Multiple aggregated proportion matrix $\mathbf{P_A}$ and its global and separate margins

Model Matrix

The global model $\mathbf{M}$ juxtaposes column-wise the L intra-tables independence models $\mathbf{M_l}$. The $(J_l \times K)$ intra-tables independence model $\mathbf{M_l}$ is expressed as

$$\mathbf{M_l} = \mathbf{D_J^l 1 D_K^l} \omega_l^{-1} \tag{5.8}$$

where the $(J_l \times J_l)$ diagonal matrix $\mathbf{D_J^l}$ and the $(K \times K)$ diagonal matrix $\mathbf{D_K^l}$ store, respectively, the row and column margins of $\mathbf{P_A^l}$. $\mathbf{1}$ denotes the $(J_l \times K)$ matrix whose generic term is equal to 1.

MFACT

For each set l, the table $\mathbf{Z_l}$, whose general term is the weighted residual with respect to the intra-table independence model, is built as

$$\mathbf{Z_l} = \mathbf{D_J^{l-1}}(\mathbf{P_A^l} - \mathbf{M_l})\mathbf{D_K^{-1}}. \tag{5.9}$$

Adopting MFA-like viewpoint to balance the influence of the subsets in the global analysis, first L separate analyses are performed via PCA($\mathbf{Z_l}$,$\mathbf{D_K}$,$\mathbf{D_J^l}$) to compute the L first eigenvalues λ_1^l $(l = 1, ..., L)$.

The $(J \times K)$ multiple table $\mathbf{Z}$ juxtaposes column-wise the $(J_l \times K)$ tables $\mathbf{Z_l}$. A non-standardized weighted PCA is performed on $\mathbf{Z}$ with d_{Jjj}/λ_1^l as row weights and as metric in the column space (stored into the matrix $\mathbf{D_{J\lambda}}$) and d_{Kkk} as column weights and as metric in the row space (stored into the matrix $\mathbf{D_K}$).

Note that, although a partition structure on the rows is considered, this is completely equivalent to the partition structure on the columns considered in classical MFACT (section 3.3.2). The results obtained in this analysis are the same as those issued from MFACT applied to the $(K \times J)$ global frequency table $\mathbf{Y_A^T}$.

Remarks: 1. The role of $\mathbf{D_J^{-1}}$ and $\mathbf{D_K^{-1}}$ are to perform a double standardisation on the rows and columns, as in CA. The rows and columns are seen from their profiles. In particular, this standardization eliminates the influence of the number of respondents in category k.

2. The intra-tables independence model $\mathbf{M_l}$ can be rewritten at the "respondent level". This rewriting will be useful for generalising the method to several categorical or quantitative variables.

The respondents weights corresponding to set l are stored in the $(I_l \times I_l)$ diagonal matrix $\mathbf{D_I^l}$. They are issued from the row margins of $\mathbf{P_l}$ but multiplied by ω_l in order to have the sum of respondents weights equal to the relative importance of table $\mathbf{P_l}$, and not to 1,

$$\mathbf{D_I^l} = [d_{Iii}] = [p_{i\bullet l}] \times \omega_l. \tag{5.10}$$

Then, the respondents weights across all the frequency tables are juxtaposed into the $(I \times I)$ diagonal matrix $\mathbf{D_I}$; $\sum_{i \in I} d_{Iii} = 1$.

In a similar way, the column margin of $\mathbf{P_l}$, but multiplied by ω_l, allows for rewriting the weighting system on the words

$$\mathbf{D_J^l} = [d_{Jjj}] = [p_{\bullet jl}] \times \omega_l. \tag{5.11}$$

Since a single categorical variable is considered, the columns of $\mathbf{X_l}$ are $\mathbf{D_I^l}$-orthogonal, making $\mathbf{D_K^l} = \mathbf{X_l^T D_I^l X_l}$ diagonal. Thus, the matrix $\mathbf{M_l}$ can be rewritten as

$$\mathbf{M_l} = \mathbf{D_J^l 1 X_l^T D_I^l X_l} \omega_l^{-1}. \tag{5.12}$$

The $(J_l \times I_l)$ matrix

$$\mathbf{1^*} = \mathbf{1 X_l^T} \tag{5.13}$$

with generic term the constant 1, substitutes $\mathbf{1X_l^T}$ in Eq.(5.12)

$$\mathbf{M_l} = \mathbf{D_J^l 1^* D_I^l X_l} \omega_l^{-1}. \tag{5.14}$$

In this way, the respondents are considered, instead of the categories, in the expression of the intra-tables independence model. This property allows for generalizing this methodology to the case of several variables as discussed in the next sections.

5.1.3 Extension to several contextual categorical variables

We may be interested in a broader context, such as a set of Q categorical variables. In the following, we use the same notation as the one used in the former section to highlight that the same rationale is followed. The formulas of the former section corresponding to the separate proportion tables are identical. However, the other formulas differ due to $\mathbf{X_l}$ being built from several variables.

Matrix $\mathbf{X_l}$ includes now the centred dummy columns corresponding to several categorical variables. The columns of $\mathbf{X_l}$ are centred by set and its generic term is $x_{ikl}^C = x_{ikl} - \bar{x}_k^l$, where $\bar{x}_k^l = \sum_{i \in I_l} x_{ikl} p_{i\bullet l}$ denotes the mean of the variable k^l restricted to the rows of set l. In this way, the global contextual table $\mathbf{X}$, juxtaposing column-wise the contextual tables $\mathbf{X_1}$,...,$\mathbf{X_l}$,...,$\mathbf{X_L}$, is also mean-centred by the weighting system $\mathbf{D_I}$.

The $(J_l \times K)$ generalised aggregated lexical table $\mathbf{Y_A^l}$ is built crossing the words in language l (in rows) and categories of the categorical variables (in columns) (Eq.5.3). The L generalised aggregated lexical tables $\mathbf{Y_A^1}$,...,$\mathbf{Y_A^l}$,...,$\mathbf{Y_A^L}$ are juxtaposed column-wise into the $(J \times K)$ multiple generalised aggregated lexical table $\mathbf{Y_A}$. Then, $\mathbf{Y_A}$ is transformed into the $(J \times K)$ matrix $\mathbf{P_A}$ (Eq.5.23).

To maintain a double standardisation on the rows and columns of matrix $\mathbf{P_A}$, as in CA-GALT, the Moore-Penrose pseudoinverse $\mathbf{C}^-$ of the $(K \times K)$ covariance matrix

$$\mathbf{C} = \mathbf{X^T D_I X} \tag{5.15}$$

substitutes $\mathbf{D_K^{-1}}$ in the expression of

$$\mathbf{Z_l} = \mathbf{D_J^{l-1}}(\mathbf{P_A^l} - \mathbf{M_l})\mathbf{C}^-. \tag{5.16}$$

$\mathbf{C}$ is computed taking into account that the respondents are endowed with weighting system $\mathbf{D_I}$.

The model matrix $\mathbf{M_l} = \mathbf{D_J^l 1^* D_I^l X_l} \omega_l^{-1}$ has for generic term

$$m_{jkl} = \sum_{i \in I_l} \frac{\omega_l^2 p_{\bullet jl} p_{i\bullet l} x_{ikl}^C}{\omega_l} = \omega_l p_{\bullet jl} \sum_{i \in I_l} p_{i\bullet l} x_{ikl}^C \tag{5.17}$$

where $\sum_{i \in I_l} p_{i\bullet l} x_{ikl}^C$ denotes the mean of k^l equal to 0. Therefore, m_{jkl} is also equal to 0. In other words, mean-centring the columns of $\mathbf{X_l}$ corresponds to use model $\mathbf{M_l}$. Eq.(5.16) is now expressed as

$$\mathbf{Z_l} = \mathbf{D_J^{l-1} P_A^l C}^-. \tag{5.18}$$

Then, the same steps as in the former section are followed. For each set l, a separate analysis is performed via PCA($\mathbf{Z_l}$,$\mathbf{C}$,$\mathbf{D_J^l}$) and the first eigenvalue λ_1^l is stored. MFA-GALT is performed through PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{D_{J\lambda}}$) as a non-standardized weighted PCA performed on the multiple table $\mathbf{Z}$ with $\mathbf{D_{J\lambda}}$ as row weights and as metric in the column space and $\mathbf{C}$ as column weights and as metric in the row space.

5.1.4 MFA-GALT for contextual quantitative variables

The method proposed in the case of several categorical variables can be easily generalized to quantitative variables. The same rationale is followed and can be summed up by five

points: separate proportion tables, mean-centring quantitative variables, multiple generalised aggregated lexical table building, definition of the weights and metrics and analysis through MFA-GALT.

I) Proportion tables

The formulas of the section 5.1.2 corresponding to the separate proportion tables are identical. For each set l $(l = 1,...,L)$, the $(I_l \times J_l)$ proportion matrix $\mathbf{P_l}$ is computed as

$$\mathbf{P_l} = \mathbf{Y_l}/N_l \tag{5.19}$$

with $N_l = \sum_{i \in I_l} \sum_{j \in J_l} y_{ijl}$, grand total of table $\mathbf{Y_l}$. $N = \sum_{l \in L} N_l$ is the grand total through all the frequency tables $\mathbf{Y_l}$.

The generic term of the row and column margins of $\mathbf{P_l}$ are devoted by $p_{i\bullet l} = \sum_{j \in J_l} p_{ijl}$ and $p_{\bullet jl} = \sum_{i \in I_l} p_{ijl}$. The relative importance of each frequency table is

$$\omega_l = N_l/N. \tag{5.20}$$

II) Mean-centring quantitative variables

The quantitative variables of $\mathbf{X_l}$ are mean-centred by set and its generic term is $x^C_{ikl} = x_{ikl} - \bar{x}^l_k$, where $\bar{x}^l_k = \sum_{i \in I_l} x_{ikl} p_{i\bullet l}$ denotes the mean of the variable k^l restricted to the rows of set l. In this way, the global contextual table $\mathbf{X}$, juxtaposing column-wise the contextual tables $\mathbf{X_1}$,...,$\mathbf{X_l}$,...,$\mathbf{X_L}$, is also mean-centred by the weighting system stored into the $(I \times I)$ diagonal matrix

$$\mathbf{D_I} = [d_{Iii}] = [p_{i\bullet l}] \times \omega_l. \tag{5.21}$$

III) Multiple generalised aggregated lexical table building

For each set l $(l = 1,...,L)$, the $(J_l \times K)$ generalised aggregated lexical table

$$\mathbf{Y^l_A} = \mathbf{Y^T_l X_l} \tag{5.22}$$

is built crossing the words in language l (in rows) and quantitative variables (in columns). The L generalised aggregated lexical tables $\mathbf{Y^1_A}$,...,$\mathbf{Y^l_A}$,...,$\mathbf{Y^L_A}$ are juxtaposed column-wise into the $(J \times K)$ multiple generalised aggregated lexical table $\mathbf{Y_A}$.

$\mathbf{Y_A}$ is transformed into the $(J \times K)$ matrix

$$\mathbf{P_A} = \mathbf{Y_A}/N \tag{5.23}$$

whose generic term $p_{Ajkl} = \sum_{i \in I_l} y_{ijl} x_{ikl}/N$ is equal to the weighted sum of the values assumed for variable k by the respondents who used word j in language l.

IV) Definition of the weights and metrics

The column margin of $\mathbf{P_l}$, but multiplied by ω_l,

$$\mathbf{D_J^l} = [d_{Jjj}] = [p_{\bullet jl}] \times \omega_l. \tag{5.24}$$

corresponds to the weighting system on the row-words and the metric on the column space. The row space metric, the $(K \times K)$ non-diagonal covariance matrix of the columns of $\mathbf{X}$, is computed as

$$\mathbf{C} = \mathbf{X^T D_I X}. \tag{5.25}$$

V) MFA-GALT

A double standardisation of $\mathbf{P_A}$, on the rows and the columns, leads to the $(J \times K)$ table $\mathbf{Z}$ analysed by MFA-GALT. The table $\mathbf{Z}$ juxtaposes column-wise the $(J_l \times K)$ tables

$$\mathbf{Z_l} = \mathbf{D_J^{l-1} P_A^l C^{-1}}. \tag{5.26}$$

If $\mathbf{C}$ is not invertible, $\mathbf{C^{-1}}$ is substituted by the general inverse $\mathbf{C}^-$. Then, the same steps as in the former section are followed.

For each set l, a separate analysis is performed via PCA($\mathbf{Z_l}$,$\mathbf{C}$,$\mathbf{D_J^l}$) and the first eigenvalue λ_1^l is stored. MFA-GALT is performed through PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{D_{J\lambda}}$) as a non-standardized weighted PCA performed on the multiple table $\mathbf{Z}$ with $\mathbf{D_{J\lambda}}$ as row weights and as metric in the column space and $\mathbf{C}$ as column weights and as metric in the row space.

Remark: In matrix $\mathbf{P_A}$, the respondents are weighted by y_{ijl}/N that is the relative frequency of word j in language l used by respondent i and the grand total over all the frequency tables N. Thus, the respondents giving the longest answers would be favoured. However, the standardisation on the rows makes that the words are placed on the variable from the weighted average of the values taken by the respondents who use them, through $\mathbf{D_J^{-1} P_A}$,

that balances the importance of the words. Moreover, following the same rationale as in CA-GALT (section 4.3.2), metric $\mathbf{C^{-1}}$ operates a multivariate standardisation. The columns of the matrix $\mathbf{X}(\mathbf{C}^{-})^{1/2}$ are those of $\mathbf{X}$ standardised by a Mahalanobis transformation, a multivariate standardisation that not only separately standardises the columns of $\mathbf{X}$ but, in addition, makes them uncorrelated.

5.1.5 Main properties of MFA-GALT

MFA-GALT provides the classical outputs of the principal components methods:

- coordinates, contributions and the quality of representation of row-words
- coordinates of categories at the centroid of the row-words used by individuals belonging to this category
- coordinates of quantitative variables as covariances and correlation coefficient between factors and quantitative variables

Furthermore, outputs related to MFACT as partial representation of the variables, representation of the sets and similarity measures for sets are also provided.

Representation of the row-words and the column-variables

The cloud of row-words N_J is placed in $\mathbb{R}^K$ (row space). In the row space, the inertia axis with rank s corresponds to the eigenvector $\mathbf{u_s}$ ($\|\mathbf{u_s}\|_{\mathbf{C}}$=1) of the $\mathbf{Z^T D_{J\lambda} ZC}$, associated with the eigenvalue λ_s

$$\mathbf{Z^T D_{J\lambda} ZCu_s} = \lambda_s \mathbf{u_s}. \tag{5.27}$$

The eigenvalues λ_s are stored into the $(K \times K)$ diagonal matrix Λ and the eigenvectors into the columns of the $(K \times K)$ matrix $\mathbf{U}$. The coordinates of the row-words are computed as

$$\mathbf{F} = \mathbf{ZCU}. \tag{5.28}$$

The cloud of column-variables N_K is placed in $\mathbb{R}^J$ (variable space). In the variable space, the inertia axis with rank s corresponds to the eigenvector $\mathbf{v_s}$ ($\|\mathbf{v_s}\|_{\mathbf{D_{J\lambda}}}$=1) of the matrix $\mathbf{ZCZ^T D_{J\lambda}}$, associated with the same eigenvalue λ_s

$$\mathbf{ZCZ^T D_{J\lambda} v_s} = \lambda_s \mathbf{v_s}. \tag{5.29}$$

The eigenvectors are stored into the $(J \times J)$ matrix $\mathbf{V}$. The coordinates of the column-variables are computed as

$$\mathbf{G} = \mathbf{Z^T D_{J\lambda} V}. \tag{5.30}$$

Transition formulae

The factors on the row-words and on the column-variables, columns of $\mathbf{F}$ and $\mathbf{G}$, are linked by the usual transition relationships

$$\mathbf{F} = \mathbf{ZCG}\Lambda^{-1/2} = \mathbf{D_J^{-1} P_A G}\Lambda^{-1/2} \tag{5.31}$$

$$\mathbf{G} = \mathbf{Z^T D_{J\lambda} F}\Lambda^{-1/2}. \tag{5.32}$$

General transition formulae allow to express the coordinates of a column-variable depending on the coordinates of all the row-words and vice versa. So, only Eq.(5.27) or Eq.(5.29) have to be solved for computing both the coordinates of rows and columns.

Superimposed representation of the *l* clouds of variables

We associate the "partial" cloud N_K^l of the variables in the space $\mathbb{R}^{J_l}$ with set l. As in MFA, it is possible to represent simultaneously the L scatter plots of the variables N_K^l in the space $\mathbb{R}^J$.

The coordinate of k^l along axis s is denoted as $G_s(k^l)$ and can be calculated from the coordinates of the row-words $F_s(j)$, $j \in J_l$, by the restriction of the transition relationships to set l:

$$G_s(k^l) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^l}} \sum_{j \in J_l} z_{jkl} d_{Jjj} F_s(j). \tag{5.33}$$

The representations of the sets and similarity measures

Two similarity measures are used for the synthetic interpretation and comparison of the sets as in MFA: Lg (dimensionality index) and RV (similarity between the configurations of partial points).

The Lg coefficient (Escofier and Pagès, 1988) is defined as the scalar product between the matrices associated with each set. These coefficients are displayed in a matrix with

an interpretation analogous to that of a covariance matrix. Lg coefficient measures the richness of the common structure between the two sets. It takes the value 0 when there is no relationship between sets, and it increases with the strength of the relationship. The larger is the Lg coefficient, the larger is the common structure and the dimensionality of the sets (Abascal et al., 2006; Pagès, 2004).

The RV coefficient (Escoufier, 1973) is a measure of relationship between two sets of variables and it is based on the principle that two sets of variables are perfectly correlated if there exists an orthogonal transformation that makes the two sets coincide. The RV coefficient is equal to 0 if each variable of the first table is uncorrelated to each variable of the second table and is equal to 1 if the configurations of the individuals induced by the two data sets are homothetic. Between the two extreme bounds 0 and 1, the value of an RV coefficient is in itself not informative because it depends on the number of individuals, on the number of variables and on the dimensionality. So, it is advisable to test the significance of the RV coefficient (Josse et al., 2008).

To calculate these measures, the $(K \times K)$ matrix of scalar products $\mathbf{W_l}$ between the K column-variables of set l is computed as

$$\mathbf{W_l} = \mathbf{Z_l^T D_{J\lambda}^l Z_l}. \tag{5.34}$$

The number of common dispersion dimensions between two sets l and l' is computed by

$$Lg(l,l') = \langle \mathbf{W_l C}, \mathbf{W_{l'} C} \rangle = trace(\mathbf{W_l C W_{l'} C}). \tag{5.35}$$

The similarity between two configurations of partial points is computed by

$$RV(l,l') = \frac{\langle \mathbf{W_l C}, \mathbf{W_{l'} C} \rangle}{||\mathbf{W_l C}||\ ||\mathbf{W_{l'} C}||}. \tag{5.36}$$

The coordinate of set l upon axis of rank s is computed as

$$Lg(l, \mathbf{u_s}) = \langle \mathbf{W_l C}, \mathbf{u_s C} \rangle = trace(\mathbf{W_l C u_s u_s^T C}). \tag{5.37}$$

5.2 Application to international survey

A railway company conducted a survey to know the opinion and satisfaction of their passengers concerning high-quality night rail service. Passengers were asked to rate their satisfaction

about 14 different aspects related to comfort (general, cabin, bed, seat), cleanliness (common areas, cabin, toilet), staff (welcome attention, trip attention, language skills) and others (cabin room, air conditioning, punctuality, general aspects). Each aspect was scored on a 11 point Likert scale from 0 (very bad) to 10 (excellent). Additionally, an open-ended question was added to the questionnaire asking for the aspects that should be improved. This question required free and spontaneous answers and could be answered in English or in Spanish.

The data is coded into the data structure presented in Figure 5.5. For each language, open-ended question is coded into the $(I_l \times J_l)$ matrix $\mathbf{Y_l}$ with generic term y_{ijl} corresponding to the frequency of word j used by respondent i in language l. The $(I_l \times K)$ matrix $\mathbf{X_l}$, with generic term x_{ikl}, stores the K satisfaction scores evaluated by I_l individuals responding in language l.

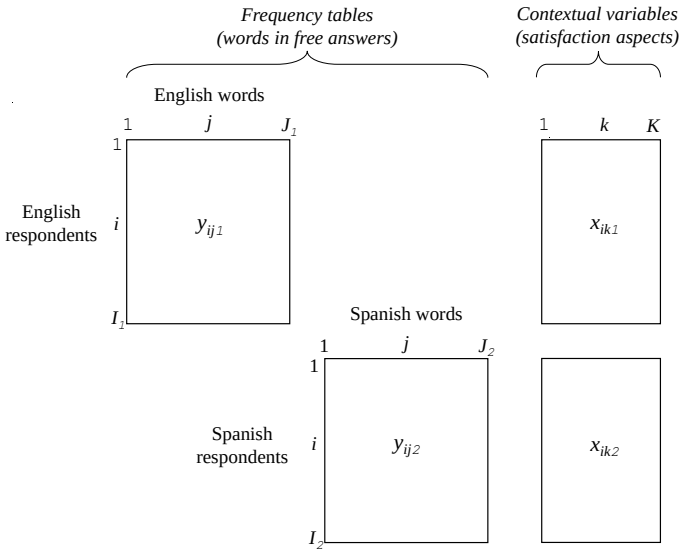


Fig. 5.5 The dataset. On the left, the frequency tables;on the right contextual variables. In the example, I_1 = 274 (English respondents), I_2 = 1003 (Spanish respondents), J_1 = 60 (English words), J_2 = 80 (Spanish words), K = 14 (satisfaction scores).

Respondents $\times$ words frequency tables $\mathbf{Y_l}$ were built following the classical preprocessing steps. Stop words were used and lemmatization from plural to singular form was performed. Only the words used at least 10 times among all the Spanish answers and at least 5 times among all the English answers were kept. Thus, 80 distinct words and 3092 occurrences were kept for 1003 respondents who answered in Spanish. Meanwhile, the total length of 274 answers in English was 829 occurrences composed of 60 distinct words. Regarding the satisfaction scores, the average of the number of missing values were 78 (range 49-213) for

the satisfaction scores answered in Spanish and 6 (range 1-23) for the satisfaction scores answered in English. A single imputation approach was used to impute missing data (Josse and Husson, 2012). This approach is based on the regularized iterative PCA algorithm that first replaces missing values by the mean of each variable. Then, a PCA on the completed data set is performed to estimate the parameters used by the regularized reconstruction formulae that imputes the missing values. These steps of estimation of the parameters via PCA and imputation of the missing values using the (regularized) fitted matrix are repeated until convergence.

Regarding to statistical analysis, first, the glossary of the most frequent words and the mean satisfaction scores provided a first overview of the complaints. Moreover, the associations between the vocabulary and contextual variables were studied. A first approximation to analyse the whole data based on CA-GALT was considered. A generalised aggregated lexical table was constructed for each sample. Two separated CA-GALT were performed on these tables that allowed for studying the similarities among the words, among the contextual variables and the relationships between them for each sample separately. However, it was not possible to represent the words pronounced by English and Spanish speakers in a common framework to study their similarities. MFA-GALT was proposed as a solution to this problem. A global analysis conserving the separation of the samples allowed for representing the whole data in a common framework. The similarities between words (in the same language), the similarities between the homologous words (in different languages), the associations between words and satisfaction scores, the similarities between satisfaction scores structures (partial representations) and the similarities between two samples were studied. These results are detailed below.

5.2.1 Univariate approach from open-ended and closed questions

The most frequent words used by the Spanish speakers were *espacios*/spaces (400), *cabinas*/cabins (290), *precios*/prices (183), *aseos*/bathrooms (115), *asientos*/seats (93) and *baños*/toilets (80). The most of these words were also very used by the English speakers: *cabins* (75), *prices* (43), *space* (41), *seats* (37), *size* (36) and *toilets* (32).

Table 5.1 shows the mean satisfaction scores for Spanish and English speakers. According to the results, *punctuality* obtained the highest satisfaction score from both Spanish and English speakers respondents (average scores 8.22 and 8.42, respectively) and *cabin room* the lowest mean scores (4.94 and 5.40, respectively). English speakers are less satisfied with

staff welcome, *trip attention* and *language skills* than the Spanish speakers. The correlation structure between the scores are similar from one group to the other.

The association between vocabulary and contextual variables, defined as the association-with-words ratio (section 4.3.2), shows that *Air conditioning* is the score with the highest association with vocabulary for both Spanish (0.097) and English (0.167) speakers. *Toilet cleanliness* has the second highest association ratio with words. The main difference between the two groups corresponds to *staff language skills*, highly associated with words used by English speakers and lowly associated with words used by Spanish speakers.

Table 5.1 Mean satisfaction scores and association with words ratios

	Spanish respondents		English respondents	
Satisfaction scores	mean (SD)	ass.ratio (p-value)	mean (SD)	ass.ratio (p-value)
General comfort	6.63 (1.82)	0.054 (0.000)	6.40 (2.06)	0.091 (0.148)
Cabin comfort	6.08 (2.06)	0.060 (0.000)	6.02 (2.16)	0.084 (0.188)
Cabin room	4.94 (2.39)	0.063 (0.000)	5.40 (2.45)	0.105 (0.013)
Bed comfort	6.55 (2.01)	0.038 (0.004)	6.53 (2.17)	0.051 (0.954)
Seat comfort	5.79 (2.22)	0.056 (0.000)	5.62 (2.51)	0.112 (0.007)
Air conditioning	6.19 (2.63)	0.097 (0.000)	6.13 (2.92)	0.167 (0.000)
Common areas cleanliness	7.23 (1.94)	0.039 (0.002)	7.46 (1.93)	0.076 (0.398)
Cabin cleanliness	7.42 (1.90)	0.045 (0.000)	7.48 (1.87)	0.093 (0.131)
Toilet cleanliness	5.84 (2.64)	0.067 (0.000)	5.94 (2.56)	0.121 (0.000)
Staff welcome attention	7.90 (1.94)	0.035 (0.018)	7.12 (2.60)	0.082 (0.213)
Staff trip attention	7.96 (1.85)	0.034 (0.054)	7.16 (2.39)	0.082 (0.264)
Punctuality	8.22 (1.57)	0.042 (0.001)	8.42 (1.42)	0.072 (0.491)
General aspects	7.58 (1.60)	0.034 (0.031)	7.35 (1.97)	0.076 (0.395)
Staff language skills	7.58 (2.12)	0.043 (0.002)	6.98 (2.67)	0.115 (0.004)

The methodology detailed in the previous section 4.2.2 was applied to identify the words associated with the satisfactions scores. A lack of control over the *temperature/temperatura* and dissatisfaction with *air conditioning/aire acondicionado* adjustment generates major complaints about this aspect. Passengers also insist on *toilet cleanliness/limpieza de los lavabos* considering *hygiene* as a main concern. Regarding the *cabin room*, one of the main problems is related to the *cabin size/tamaño de las cabinas*.

Table 5.1 shows that average scores and association-with-words ratios do not share the same reverse order (a low satisfaction indicating high association-with-words ratio). This means that, according to the opinion of the passengers, aspects which should be improved the most are not corresponding to the aspects of which they were less satisfied. This fact justifies the interest and need to collect information by means of open-ended questions as a complement to classical closed-ended questions.

5.2.2 Separate CA-GALT

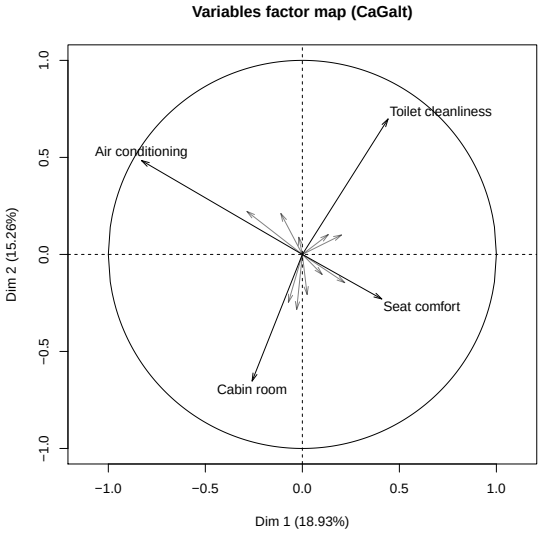
Two separate CA-GALT were performed on each set. The overall inertia of the English set was higher than the one of the Spanish set (1.53 vs. 0.73) as well as the first eigenvalue (0.237 vs. 0.138). The higher inertia of the English set suggests a higher structural relationship between vocabulary and contextual variables for English speakers. However, tests on CA-GALT total inertia and first eigenvalue (section 4.3.2) confirms that the relationships between free answers and satisfaction scores are strong for both sets.

The percentages of variance explained by the first two dimensions were 18.9% vs 15.4% for the first and 15.2% vs 14.9% for the second. The first two axes of the separate CA-GALT correspond to general trends. The third and fourth dimensions also explain an important part of the total variance (around 20% between two axes) and it may be interesting to study also these two dimensions.

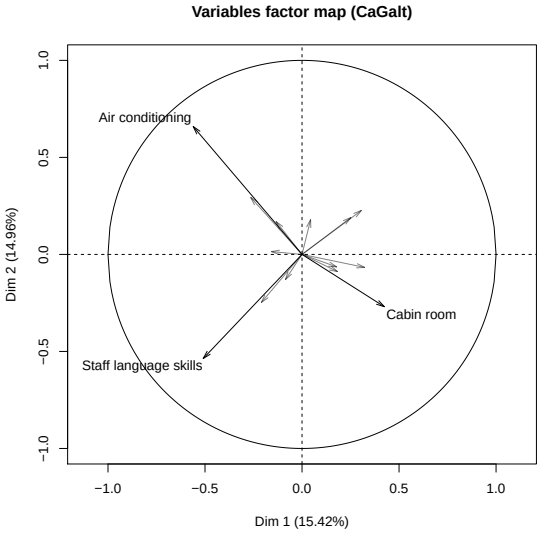
Figure 5.6 displays the satisfaction scores (as correlations) and the words on the first principal plane of the separated CA-GALT. To ease the interpretation, the sign of the scores were inverted (a high positive value indicates high dissatisfaction) and only the words with the highest contributions (fifty words for each language) were represented.

Air conditioning, *toilet cleanliness*, *cabin room* and *seat comfort* are the satisfaction scores highly correlated with the first dimension corresponding to the aspects that the Spanish speakers strongly emphasize in their free answers (Figure 5.6a). English speakers also accentuate their dissatisfaction related to *staff language skills* besides of *air conditioning* and *cabin room* (Figure 5.6b).

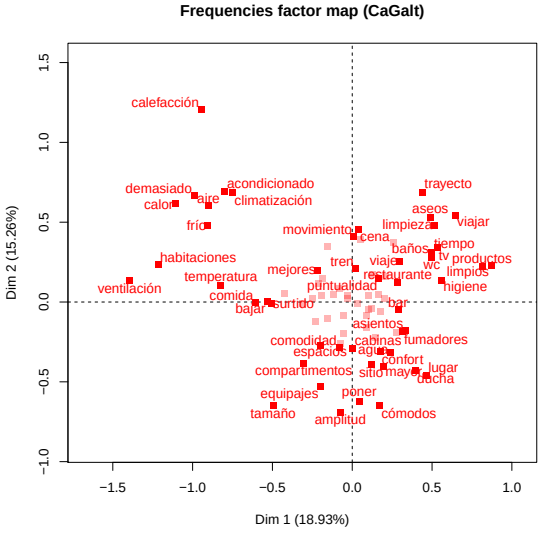
Regarding the words used by the respondents answering in Spanish (Figure 5.6c), *calefacción*/heating, *aire*/air, *acondicionado*/conditioning, *climatización*/air-conditioning, *fríio*/cold, *calor*/hot, *ventilación*/ventilation and *temperatura*/temperature are highly associated with dissatisfaction related to *air conditioning*. A lack of *toilet cleanliness* is characterised by *limpieza*/cleanliness, *aseos*/bathroom, *baños*/toilets, *limpios*/clean and *higiene*/hygiene. Words as *tamaño*/size, *espacios*/spaces and *comodidad*/convenience indicate the main prob-



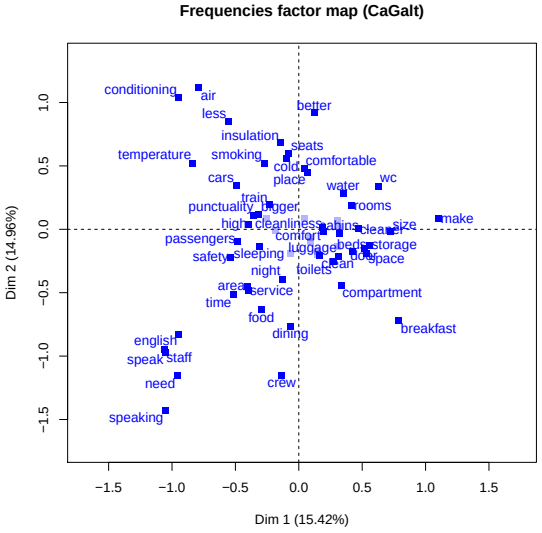
(a) Satisfaction scores (Spanish)



(b) Satisfation scores (English)



(c) Words (Spanish)



(d) Words (English)

Fig. 5.6 Scores and words on the first principal planes of separate CA-GALT

lems related to *cabin room*. In relation to *seat comfort*, more *cómodos*/comfortable *asientos*/seats would be appreciated.

English speakers expressed their dissatisfaction related to *air conditioning* and *cabin room* in a similar way using words as *air*, *conditioning*, *size/space*, *cabins/compartment* (Figure 5.6d). Whereas, the English speakers used several words as *english*, *staff*, *speak*, *need*, *speaking* and *crew* to express their dissatisfaction related to *staff language skills*.

Secondary patterns for satisfaction scores stand out on the third and fourth dimension. Spanish speakers insist on their dissatisfaction about aspects related to *staff*. Meanwhile, English speakers insist on lack of *toilet cleanliness* and dissatisfaction related to *general aspects*.

5.2.3 MFA-GALT applied to whole data

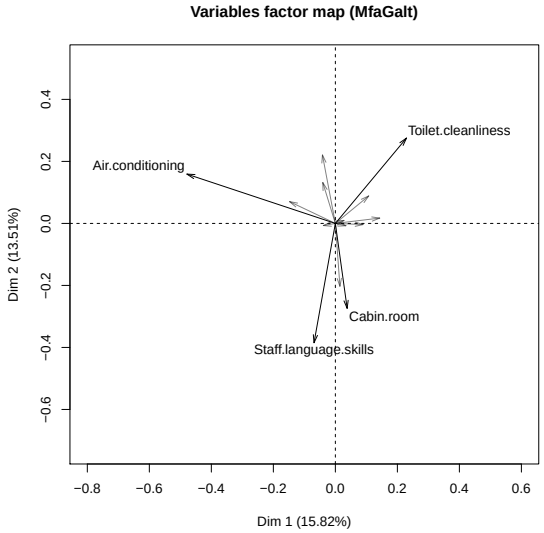
MFA-GALT is applied on the multiple generalised aggregated lexical table. The total inertia is equal to 10.45. The first eigenvalue 1.65, is close to the number of sets (English and Spanish samples) meaning that the two sets share the dispersion direction corresponding to the first global axis. The first two axes are moderately dominant with eigenvalues 1.65 (15.82% of the total inertia) and 1.41 (13.51% of the total inertia). The third and fourth axis also indicate important directions of variance (21.85% of the total variance together).

Global representation of the words and the satisfaction scores

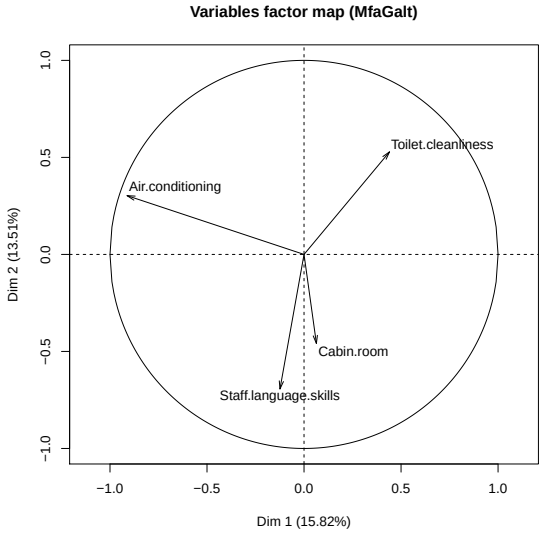
Figure 5.7a represents the satisfaction scores through their covariances with the axes on the first principal plane of MFA-GALT (performed through a non-standardized PCA). The scores highly related with the first axes are *air conditioning*, *toilet cleanliness*, *staff language skills* and *cabin room*. To ease the interpretation and comparison of the scores, the correlations between factors and scores are visualized (Figure 5.7b). Only the highest correlations were represented. The results show a tripolar structure. The three poles refer to inconvenience associated with *air conditioning*, lack of *toilet cleanliness* and problems related to *staff language skills* and *cabin room*.

Regarding the words, those with the highest contributions are represented (Figure 5.7c). *aire/air*, *acondicionado/conditioning*, *temperatura/temperature*, *fríO/cold*, *climatización*/air-conditioning, *ventilación*/ventilation and *calefacción*/heating are words highly associated with *air conditioning*. A lack of *toilet cleanliness* is characterised by *limpieza/cleanliness*, *limpios/clean(er)*, *aseos/bathrooms* and *baños/toilets*. On the second axis, English speaking respondents express their dissatisfaction related to *staff language skills* with words as *english*, *staff*, *speak*, *need*, *speaking* and *crew*. Meanwhile, on the same axis, Spanish speaking respondents express their dissatisfaction related to *cabin room* with words as *espacios*/spaces, *equipajes*/baggages, *tamaño*/size and *cabinas/cabins*.

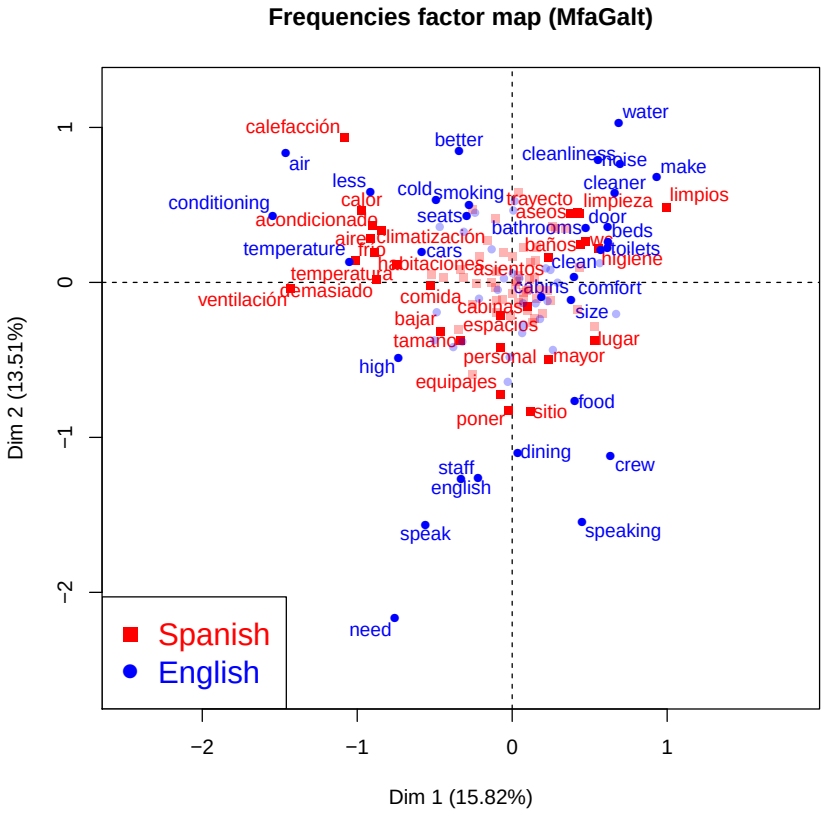
Secondary patterns for satisfaction scores stand out on the third and fourth dimension (Figure 5.8a). Third dimension insists more on problems related to *cabin room*, *toilet*



(a) Satisfaction scores (covariances)



(b) Satsifaction scores (correlations)



(c) Sixty words with the highest contributions

Fig. 5.7 Satisfaction scores and contributory words on the first principal plane of MFA-GALT

cleanliness and *staff language skills* (Figure 5.8b). Fourth dimension indicates dissatisfaction on *seat comfort*, specially expressed by Spanish speakers, with words as *asientos/seats*, *confort/comfort*, *cómodos*/comfortable and *comodidad*/convenience. Problems related to *general aspects* and *general comfort*, are expressed, mainly by English speakers, with words as *puntualidad/punctuality*, *agua/water*, *velocidad*/speed, *food*, *insulation*, *sleeping*, *class*, *ticket*, *train* and *service*.

Partial representation of the satisfaction scores

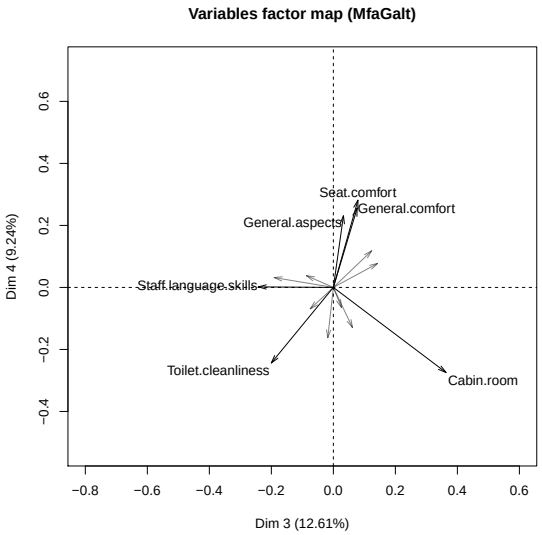
The partial representation of the satisfaction scores shows the similarities and differences between the two sets regarding the association between words and scores on the axes. *Air conditioning* and *toilet cleanliness* are highly related with the first axes for both sample (Figure 5.9a). However, on the second axis, *cabin room* is associated much more with Spanish words and *staff language skills* with words used by English speakers. On the forth dimension, English speakers give priority to express a lack of *general comfort* and problems related to *general aspects*. Meanwhile, Spanish speakers prefer to express their dissatisfaction related to a lack of *seat comfort* (Figure 5.9b)

Representation of the sets

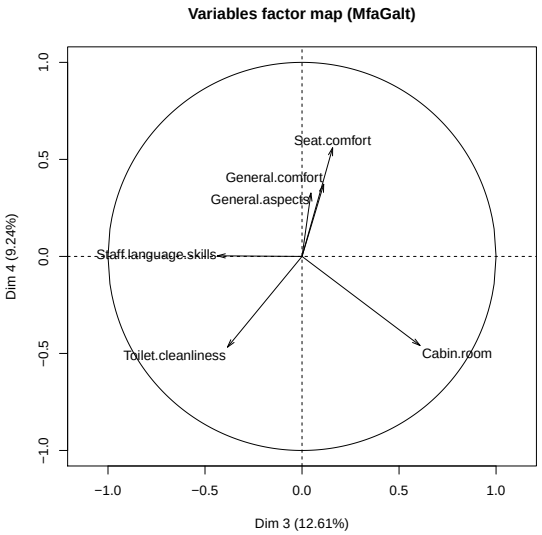
Similarity measures confirm that both sets share dispersion directions. The Lg coefficient suggests 2 common dispersion dimensions. The value of the RV coefficient (0.74, $p<0.001$) (Josse et al., 2008) confirms that the partial configurations are relatively close but not homothetic.

Regarding to the representation of the sets on the first dimension (Figure 5.10), the coordinate of the Spanish sample is 0.88 whereas the coordinate of English sample takes a slightly smaller value but also quite high (0.77) (Figure 5.10a). It means that the first axis provided by MFA-GALT is an axis of high importance for both sets that consequently they share some common structure. Actually, this common structure corresponds to the relation between the first axis and the satisfaction scores as *toilet cleanliness* and *air conditioning*.

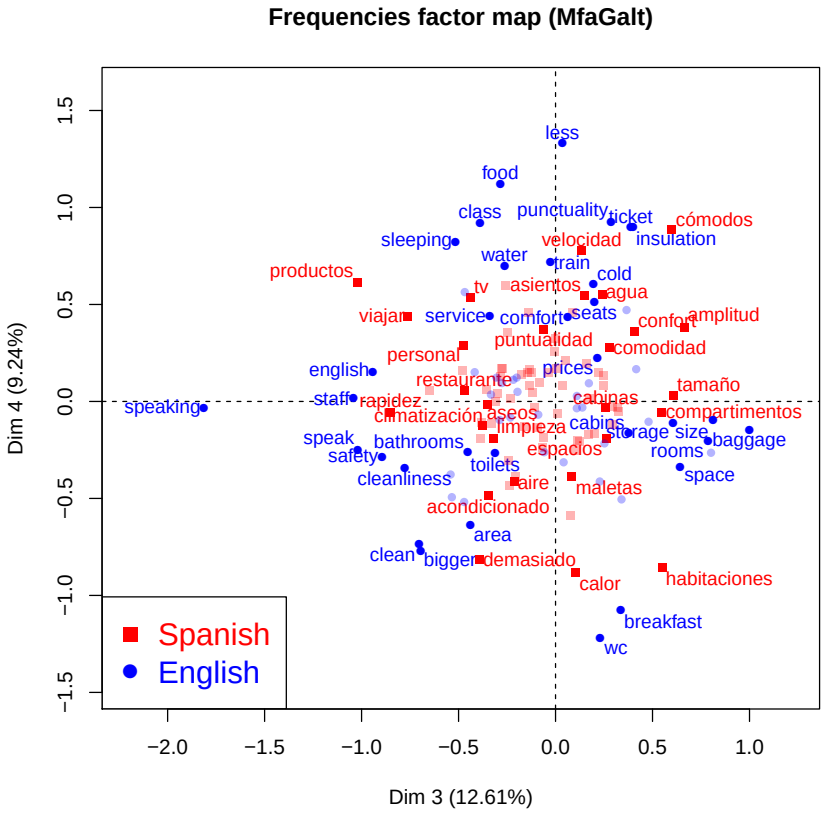
On the second dimension, the coordinate of the English sample is 0.84 whereas the coordinate of the Spanish sample is slightly smaller (0.57). This means that the second axis provided by MFA-GALT is an axis of high importance for English sample and of low importance for Spanish Sample. Thus differences are related to *staff language skills* highly correlated with the second axis. The coordinate of the Spanish sample on the third dimension



(a) Satisfaction scores (covariances)

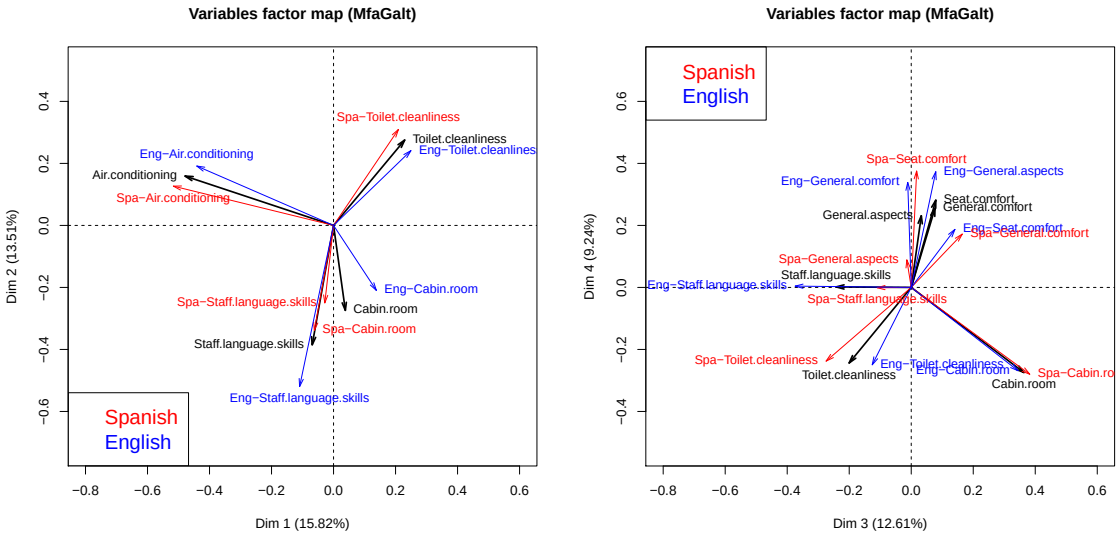


(b) Satisfaction scores (correlations)



(c) Sixty words with the highest contributions

Fig. 5.8 Satisfaction scores and contributory words on the MFA-GALT plane (3,4)



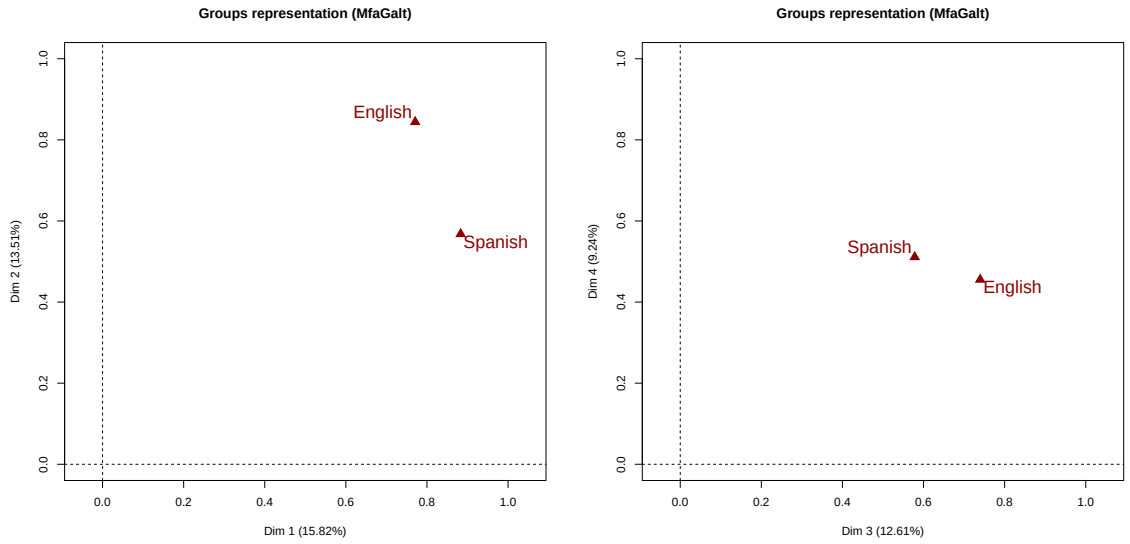
(a) MFA-GALT plane (1,2) (b) MFA-GALT plane (3,4)

Fig. 5.9 Partial representation of the satisfaction scores

is 0.58 and on the fourth dimension 0.51 (Figure 5.10b). Both third and fourth dimensions are not of high importance for Spanish sample. However, the third dimension is of high importance for English sample (0.74).

5.2.4 Concluding remarks

Results on the application show that this methodology allows for studying the similarities between words from different languages related to contextual variables in a common framework. In this sense, similarities between two structures of two samples are identified. These are mainly related to the association between words and scores as *toilet cleanliness* and *air conditioning*. Some differences were also observed. The most important correspond to *staff language skills* and *cabin room*. Secondary patterns were also identified. English speakers give priority to express lack of *general comfort* and problems related to *general aspects*. Meanwhile, Spanish speakers prefer to express their dissatisfaction related to lack of *seat comfort*.



(a) MFA-GALT plane (1,2)

(b) MFA-GALT plane (3,4)

Fig. 5.10 Representation of the sets of variables

5.3 Application to ecological data

In this example, we use the data set "meau" from the R package "ade4" (Dray et al., 2007). The data set includes species data (frequency table) and environmental data (contextual variables). The species data consist of 13 species (abundance of Ephemeroptera taxa) in 6 sites, measured 4 times in the same year (spring, summer, autumn and winter), and the environmental data consists of 10 physicochemical variables measured in the same sites and at the same dates. This data set can be considered as a sequence of 4 "paired tables" adopting the terminology used in "ade4". Each pair is made of one environmental variables table with 6 rows and 10 columns and one species composition table with 6 rows and 13 columns (Figure 5.11).

This example has been used by Thioulouse (2011) for the comparison of several methods for the simultaneous analysis of a sequence of paired tables: Between-Group Co-Inertia Analysis (BGCOIA; Franquet et al., 1995), STATICO (Simier et al., 1999; Thioulouse et al., 2004) and COSTATIS (Thioulouse, 2011). The BGCOIA is the simplest case where the mean of the variables in each paired table is computed and stored in two tables. A Co-Inertia Analysis is then performed on these new tables. In STATICO, that means STATIS (Escoufier, 1973; Lavit et al., 1994) and Co-Inertia, l cross-covariance tables are computed from l paired table and stored in l-tables. A Partial Triadic Analysis (PTA; Thioulouse and Chessel, 1987) is then performed on these l-tables. In COSTATIS, that means Co-Inertia and STATIS, two

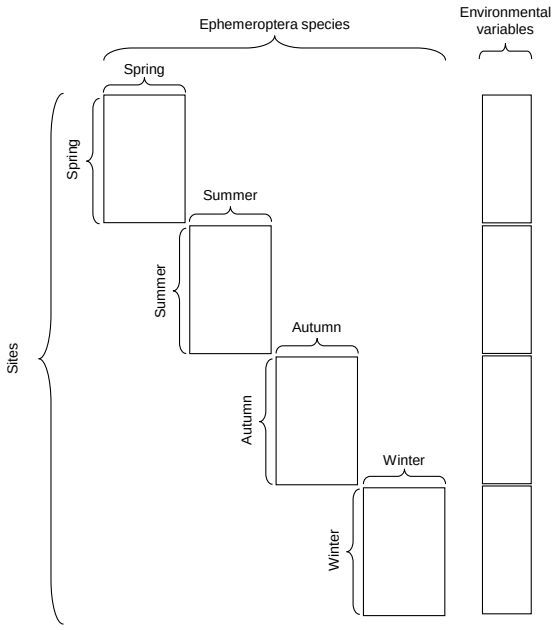


Fig. 5.11 Data structure corresponding to ecological data

PTA are used to compute the compromises of the two *l*-tables. A Co-Inertia Analysis is then used to analyze the relationships between these two compromises.

All three methods aims at studying the evolution of species-environment relationships. However, some differences exist. STATICO displays the stable component of species-environment relationship variations. COSTATIS looks for the relationchips between two stable structures. BGCOIA is only recommended when there are good reasons to give the priority to space or to time. More details on these methods and the results on the example are presented in Thioulouse (2011).

Regarding to the our methodology, we do not aim at carrying out a detailed analysis of these data but at showing that the MFA-GALT can be applied on this kind of data to provide interesting results complementary to those obtained by the already existing methods.

The species are as follows: Eda = Ephemera danica, Bsp = Baetis sp., Brh = Baetis rhodani, Bni = Baetis niger, Bpu = Baetis pumilus, Cen = Centroptilum, Ecd = Ecdyonurus, Rhi = Rhithrogena, Hla = Habrophlebia lauta, Hab = Habroleptoides modesta, Par = Paraleptophlebia, Cae = Caenis, Eig = Ephemerella ignita.

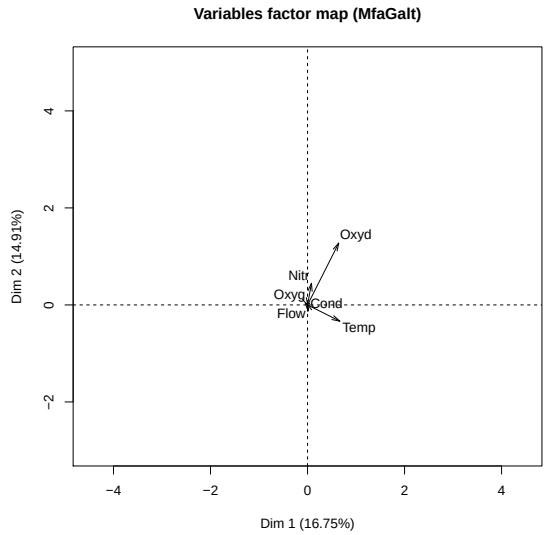
In this example, the frequency of the species varies. These frequencies are (ordered from the highest to the lowest): *Brh* (206), *Bsp* (140), *Hla* (102), *Rhi* (67), *Bpu* (54), *Eig* (51), *Cen* (44), *Ecd* (39), *Par* (38), *Eda* (22), *Bni* (18), *Cae* (10), *Hab* (5).

The physico-chemical measures are: water temperature, flow, pH, conductivity, oxygen, biological oxygen demand (BOD5), oxidability, ammonium, nitrates and phosphates.

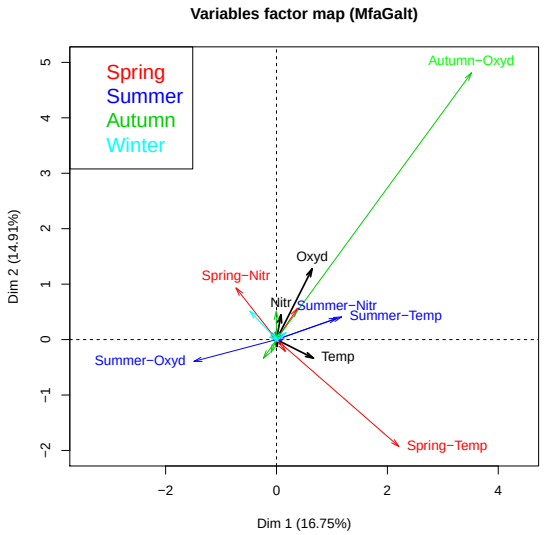
We show the main results obtained applying MFA-GALT to this data set. MFA-GALT on this data set allows us to answer the following questions: "Is there a season where habitats preferences of the species are highly associated with the environmental characteristics? Which is this season? And which are the species and environmental variables that are highly associated?"

Figure 5.12 shows the environmental variables, partial representations and seasons (=sets) on the first principal plane of MFA-GALT. The variables highly related with the first axes are *oxydability* and *temperature* (Figure 5.12a). The partial representations of the variables depending on the season show that the *autumn* (resp. *spring*) is the season when *oxydability-autumn* (resp. *temperature-spring*) is much more related with the first axes (Figure 5.12b). The representation of the seasons shows that the first axes are of high importance for *autumn* and *spring* and of low importance for *winter* and *summer* (Figure 5.12c). Figure 5.13 displays the representation of the species distribution for those two seasons. The species are ordered on the first bisector depending on *oxydability-autumn* and on the second bisector by *temperature-spring*. In autumn, species *Eda-autumn*, *Bpu-autumn* and *Cen-autumn* are more abundant in sites with lower *oxydability*. Meanhwile, in spring, species as *Bni-spring*, *Par-spring* and *Eda-spring* are more abundant in sites with lower *temperature*, while specie *Ecd-spring* is characteristic of sites with higher *temperature*. Regarding to species observed in almost all sites (*Brh*, *Bsp* and *Hla*), their coordinates are close from one season to the other. In this case, the habitat preferences related to different characteristics of the sites are not identified.

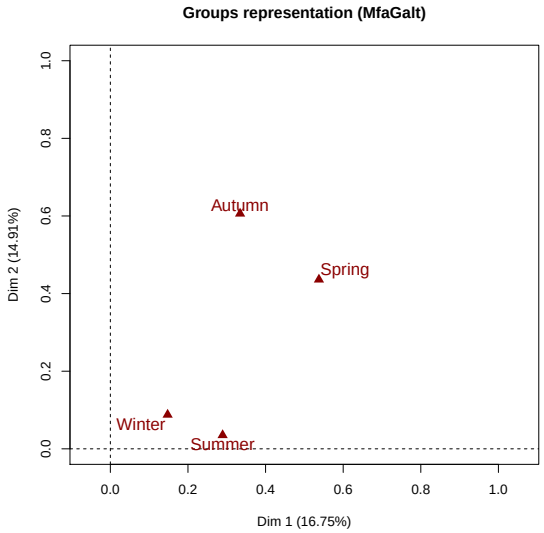
All the methods proposed to deal with sequence of paired tables as BGCOIA, STATICO and COSTATIS aim at highlighting the relationships between species and environmental variables. The main relationships concern the sites where the species with the highest frequencies are abundant. However, MFA-GALT looks for specific relationships between the habitats preferences of the species and the environmental characteristics in a given season. This analysis does not favour the species that are very abundant and observed in almost all sites. The reason for these differences in the results provided by both approaches can be explained by the double standardisation performed by MFA-GALT on rows and columns of matrix $\mathbf{P_A}$. This double standardisation balances the importance of the species in the global analysis and cancels the associations among the environmental variables. Thus, the influence of the environmental variables on the species distribution is visualised, avoiding a confusion effect.



(a) Environmental variables



(b) Partial representations



(c) Seasons

Fig. 5.12 Representations of the environmental variables, partial representations and seasons on the first principal plane of MFA-GALT

5.4 The other fields of application

Besides the text mining and the ecology, the methodology proposed to deal with a sequence of coupled tables could be applied in other fields such as social networks, graph or machine learning.

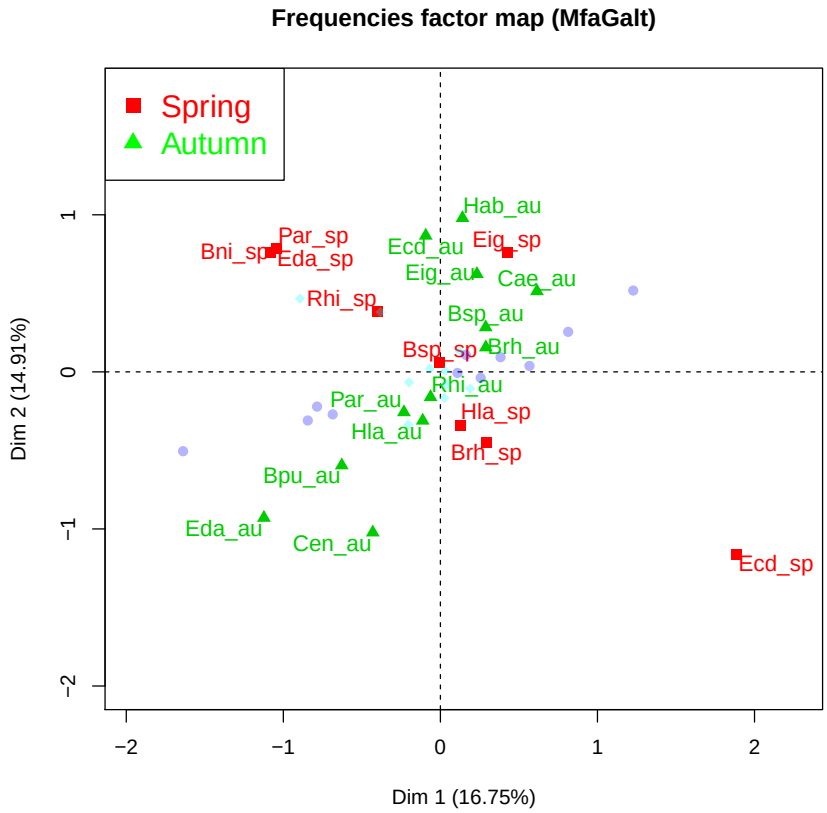


Fig. 5.13 Representations of the species on the first principal plane of MFA-GALT

A social network is defined as a social structure made up of individuals (or organizations) called "nodes", which are connecte by one or more specific types of interdependency, such as friendship, common interest, financial exchange, knowledge, etc. (Passmore, 2011). Analysis of such a network aims at studying social relationships in terms of network theory consisting of nodes and connections. For example, a simple social network could be a map of friendship connections between the nodes being studied. This network also can be represented by a socio-matrix containing binary data (link/no link).

Correspondence Analysis is considered as one of the suitable methods to analyse this kind of social and communication networks (Roberts, 2000; Scott, 2002; Wasserman, 1994). However, the resulting graph-based structures of these networks are often very complex. Recent works insist on collecting the social contextual information instead of considering only social network structures (Guo et al., 2014; Jiang et al., 2014). D'Esposito et al. (2014) proposed the use of Multiple Correspondence Analysis that makes possible to add covariates to the analysis of affiliation networks in order to improve results interpretation. MFA-GALT would be a useful tool to analyse social networks structures in different countries

complemented by contextual information. This analysis allows for comparison between the characteristics of social networks from one country to the other.

Regarding to machine learning, the computer-based analysis and organization of large document repositories is one of today's great challenges. Latent semantic indexing (LSI; Deerwester et al., 1990) is one of the reference machine learning methods to analyze textual data. This vector-space technique has shown great promise in addressing the problem of synonymy in text collections, offering improved recall over standard keyword searches.

A standard step in LSI is the creation of frequency matrix documents$\times$words. This matrix also can be represented by a contextual network graph. In this case, a bipartite graph of term and document nodes where each non-zero value in the documents$\times$words table corresponds to an edge connecting a term node to a document node. In this model, every term is connected to all of the documents in which the term appears, and every document has a link to each term contained in that document. The frequency values in the documents$\times$words table correspond to weights placed on the edges of the graph.

Compared with social networks graphs, kinds of contextual information obtained from contextual network graph are different: dependency between words, sentence co-occurrence, and proximity, that is, co-occurrence with other word (Hagiwara et al., 2006). However, as in social network, there is a growing interest in collecting effective contextual information to improve the automatic processes (Mallat et al., 2015). Thus, machine learning also would be another important field of application for MFA-GALT.

5.5 Summary

The main features of the new method called Multiple Factor Analysis on Generalised Aggregated Lexical Table (MFA-GALT) have been detailed. MFA-GALT was proposed to deal with a sequence of coupled tables. The method has been illustrated by using an example issued from an international survey including an open-ended question answered in different languages by different subsamples of individuals. A small application to ecological data shows that this method can also be applied in other fields provided that the data are coded into a similar structure.

Chapter 6

Implementation in R

6.1 MFACT

MFACT was implemented in the FactoMineR package through an option of the *MFA* function (Kostov et al., 2013). Argument *type* of variables should be defined as "f" to consider a frequency set. Then, results for the frequencies are provided as outputs. The rest of the arguments and values of *MFA* function are identically used as in previous versions of the function.

6.1.1 Application

We illustrate the function through a mortality data set in France from 1979 to 2006. The aim of this analysis is to study the evolution of the causes of mortality from 1979 to 2006. For each year, a crossed table contains 62 mortality causes (in rows) and age intervals (in columns). At the intersection of a row-mortality cause and a column-age interval, the counts of deaths corresponding to this cause and this age interval during this year is given.

The data are read from the FactoMineR package:

```
> library(FactoMineR)
> data(mortality)
```

MFACT is performed on the multiple table "causes of mortality" as follows:

```
> mfact<-MFA(mortality,group=c(9,9),type=c("f","f"),
name.group=c("1979","2006"))
```

The *group* argument specifies that each table has 9 columns, *type* specifies the type of the tables, here frequency tables, *name.group* gives the name of each table. The study of mortality causes in France in 1979 and 2006 will show the kind of results offered by MFA extended to deal with a multiple frequency/contingency table.

6.1.2 Numerical Outputs

The numerical outputs of MFA are a series of numerical indicators as results of the separate analysis, eigenvalues, coordinates, contributions, correlations, etc.

```
> mfact
**Results of the Multiple Factor Analysis (MFA)**
The analysis was performed on 62 individuals, described by 18 variables
*Results are available in the following objects :

  name                 description
1 "$eig"               "eigenvalues"
2 "$separate.analyses" "separate analyses for each group of variables"
3 "$group"             "results for all the groups"
4 "$partial.axes"      "results for the partial axes"
5 "$inertia.ratio"     "inertia ratio"
6 "$ind"               "results for the individuals"
7 "$freq"              "results for the frequencies"
8 "$global.pca"        "results for the global PCA"
```

The sequence of eigenvalues identifies two dominating dimensions that, together, account for 81.68% of the total inertia. That leads to focus on the first principal plane.

```
> round(mfact$eig,3)[1:4,]
       eigenvalue percentage of variance cumulative percentage of variance
comp 1      1.790                52.420                             52.420
comp 2      0.999                29.269                             81.689
comp 3      0.262                 7.659                             89.348
comp 4      0.149                 4.367                             93.715
```

We recall that in MFA the first eigenvalue varies between 1 and the number of groups (two in this case). A first eigenvalue close to the maximum means that the first dimension of

the separate analyses (here the CAs on each contingency table) are very similar. Thus, the value of the first eigenvalue (equal to 1.79) indicates that the first principal component is an important dispersion dimension in both tables, and is similar to the first axes of the two separate CAs.

6.1.3 Graphical outputs

The *MFA* function returns a series of graphics as individuals/rows, superimposed representation, variables/columns, partial axes, groups. However, the high number of elements drawn on the graphs (points and labels) can make the reading and the interpretation difficult. We can use *plot.MFA* function to improve graphical outputs.

Representation of rows

Figure 6.1 visualizes the age intervals columns on the first principal plane. The trajectories of the age intervals of both years are drawn in different colours to ease the comparison.

```
> plot.MFA(mfact,choix="freq",invisible="row")
> lines(mfact$freq$coord[1:9,1],mfact$freq$coord[1:9,2],col="red")
> lines(mfact$freq$coord[10:18,1],mfact$freq$coord[10:18,2],col="green")
> legend("topleft",c("1979","2006"),text.col=c("red","green"))
```

The first dimension perfectly ranks the age intervals, opposing the youngest to the oldest. The second dimension opposes the extreme to the medium-age intervals. The homologous age intervals corresponding to 1979 and 2006 are very close. In both years, the youngest intervals differ more from one another than the oldest. That indicates that the predominant causes of mortality in young people change rapidly with age.

Representation of columns

The visualization of the mortality causes on the principal plane shows that some causes are clearly separated from the others (Figure 6.2).

```
> sel <- c(2,8:10,15,38,58)
> plot.MFA(mfact,lab.ind=FALSE,habillage="group")
> text(mfact$ind$coord[sel,1],mfact$ind$coord[sel,2],
rownames(mortality)[sel],pos=c(2,2,2,2,4,2,4))
```

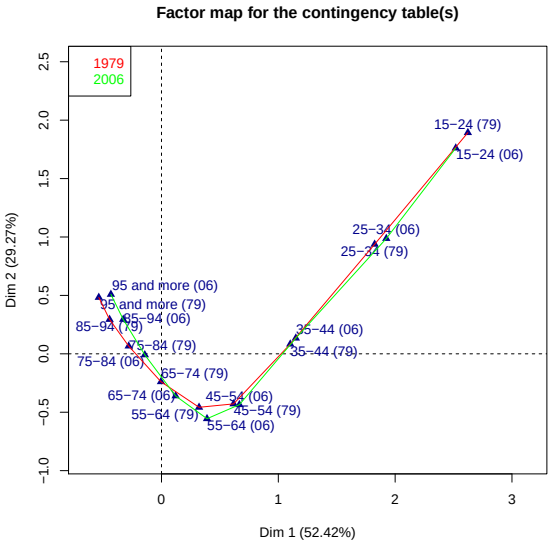


Fig. 6.1 Trajectories of the age intervals on the first principal plane.

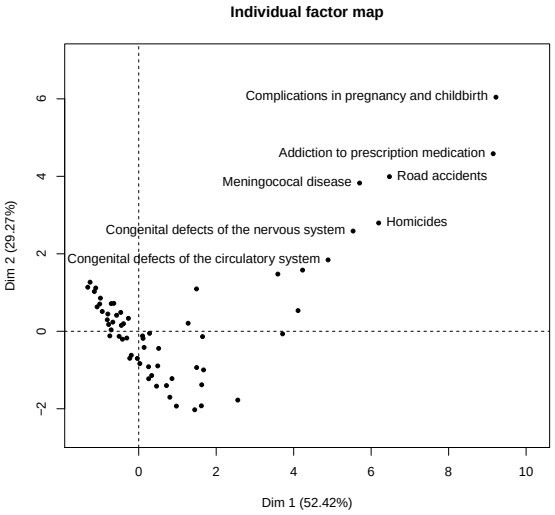


Fig. 6.2 Mortality causes representation on the first principal plane. Only the "young" causes are labelled.

The causes highly associated with young age are very specific (Figure 6.2): complications in pregnancy and childbirth, addiction to prescription medication, road accidents, meningococcal disease, homicides, congenital defects of the nervous system and congenital defects of the circulatory system.

Dimensions of the separate analyses

The dimensions of the separate analyses can be represented through their correlations with the MFA dimensions. The first and second dimensions of the global analysis are very correlated with the first and second dimensions of the separate analyses (Figure 6.3a).

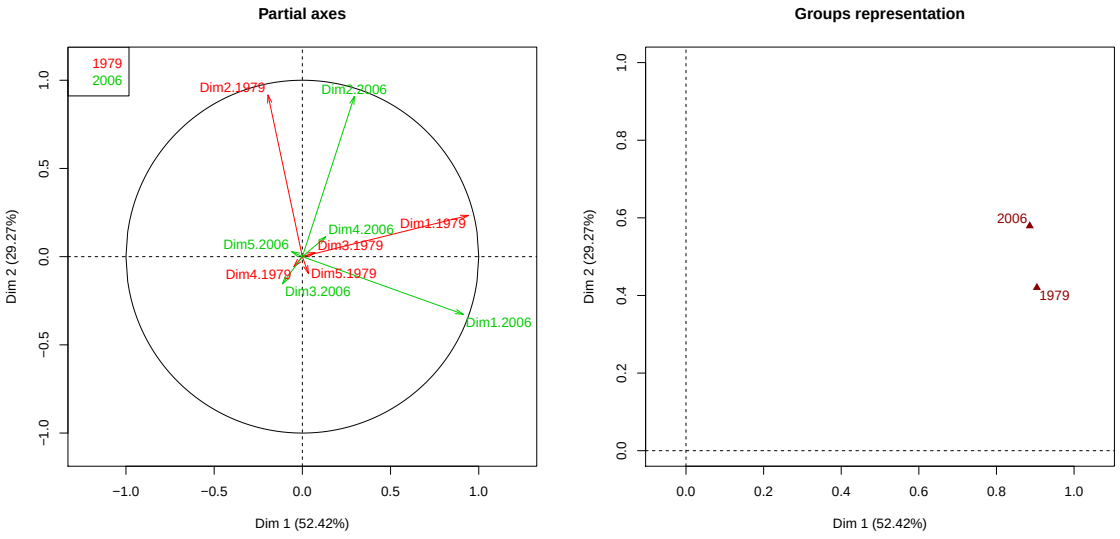
```
> plot.MFA(mfact,choix="axes",habillage="group")
```

Synthetic representation of the groups

As there are only two groups, their representation provides little information. However, in the case of a high number of groups, this kind of representation would be very useful for a global comparison.

```
> plot.MFA(mfact,choix="group")
```

Figure 6.3b shows that both groups have coordinates on dimension 1 that are very close to 1 showing that they are sensitive to the age ranking as reflected by this dimension. There are some differences between two groups on dimension 2: the oppositions between causes is highlighted on the second dimension are slightly more pronounced in 2006.



(a) Dimensions of the separate analyses

(b) Representation of the groups

Fig. 6.3 Other graphical outputs

Superimposed representation

The superimposed representation of the partial rows shows the more important changes between 1979 and 2006 (Figure 6.4).

```
> sel <- c(2,10,41,58)
> plot.MFA(mfact,lab.ind=FALSE,habillage="group",
partial=rownames(mortality)[sel])
> text(mfact$ind$coord[sel,1],mfact$ind$coord[sel,2],
rownames(mortality)[sel],pos=4)
```

Some important changes concerning the mortality causes distribution among the age intervals between 1979 and 2006 are identified. Addiction to prescription medication is the cause of death showing the highest difference. It moves from close to the centroid (1979) to a position with a high coordinate (2006) on the first dimension: deaths related with this cause concern younger people in 2006 than in 1979.

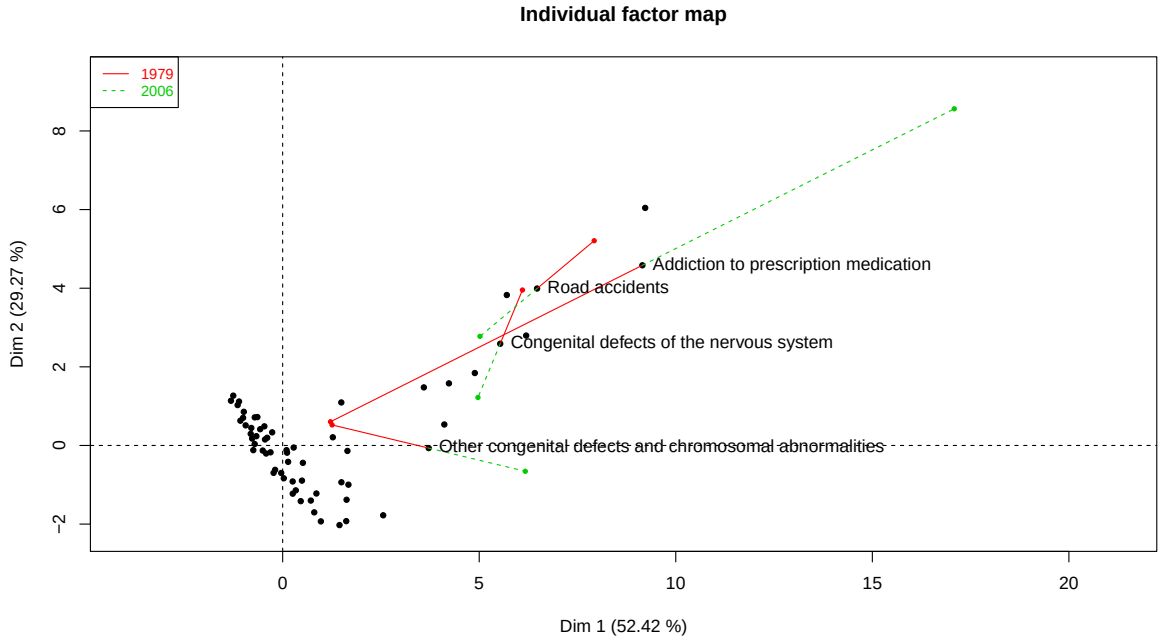


Fig. 6.4 Excerpt of the superimposed representation of the partial and global mortality causes.

6.2 WordCountAna

The R function *WordCountAna* is currently part of the SensoMineR package. This function performs a multiple factor analysis for contingency tables keeping all the information in the comparison of the products. The identification of the consensual words, assessed by a resampling technique, eases the interpretation of the word-count based methods and solves the problems arising from the large diversity of vocabulary as the different meanings possibly associated to a same word.

The default input for the *WordCountAna* function in R is

```
WordCountAna (base, sep.word = NULL, ncp = Inf, nb.panel = 3,
nb.simul = 500, proba = 0.05, graph = TRUE, axes = c(1,2))
```

6.2.1 Arguments and values

WordCountAna includes the following arguments

- *base*: a data frame with n rows (products) and p columns (panellists). Each cell corresponds to a free-text description used to describe a product by a panellist
- *sep.word*: a string with all the characters which correspond to separator of words (by default, NULL and is considered equal to "; (),?./:'!$=+;<>[]@-")
- *ncp*: number of dimensions kept in the results and to compute the within-inertia
- *nb.panel*: minimum number of panellists who used the same word in order to define consensual words (by default 3)
- *nb.simul*: number of bootstrap simulations (by default 500)
- *proba*: significance threshold considered to define consensual words (by default 0.05)
- *graph*: boolean, if TRUE a graph is displayed
- *axes*: a length 2 vector specifying the components to plot

The values returned by *WordCountAna* are

- *mfact*: a list of matrices containing all the results for multiple factor analysis for contingency tables

- *dist.words*: a matrix containing the results for distinct words (number of times that used and number of panellists that pronounced)
- *centroids*: a matrix containing the coordinates of the centroids of distinct-words
- *cons*: a matrix containing the results of bootstrap resampling for distinct-words pronounced by at least "nb.panel" panellists (number of times that used, number of panellists that pronounced and the significance of the consensus)
- *cons.words*: a vector of consensual words assessed by bootstrap resampling

6.2.2 Application

To illustrate the outputs and graphs of *WordCountAna*, we use the data set presented in chapter 3 but restricted to 30 panellists. Thus, the data frame includes 12 rows (the number of perfumes) and 30 columns (categorization of perfumes by consumers).

The data are read from the SensoMineR package:

```
> library(SensoMineR)
> data(perfume)
```

The code to perform *WordCountAna* is

```
> res<-WordCountAna(base=perfume,sep.word=";")
```

6.2.3 Numerical outputs

The numerical outputs of *WordCountAna* are results of MFACT, distinct words, centroids, permutation test and consensual words.

```
> names(res)
[1] "mfact"      "dist.words" "centroids"  "cons"       "cons.words"
```

The total length of descriptions was 565 occurrences composed of 90 distinct-words.

```
> sum(res$dist.words[,1])
[1] 565
> length(rownames(res$dist.words))
[1] 90
```

Strong, *flowery*, *soft*, *sweet* and *fruity* were the five most frequent words, each one cited more than 25 times by more than 7 consumers.

```
> res$dist.words[1:5,]
        Nb times Nb panellists
strong        40            16
flowery       49            10
soft          41            10
sweet         23            10
fruity        28             8
```

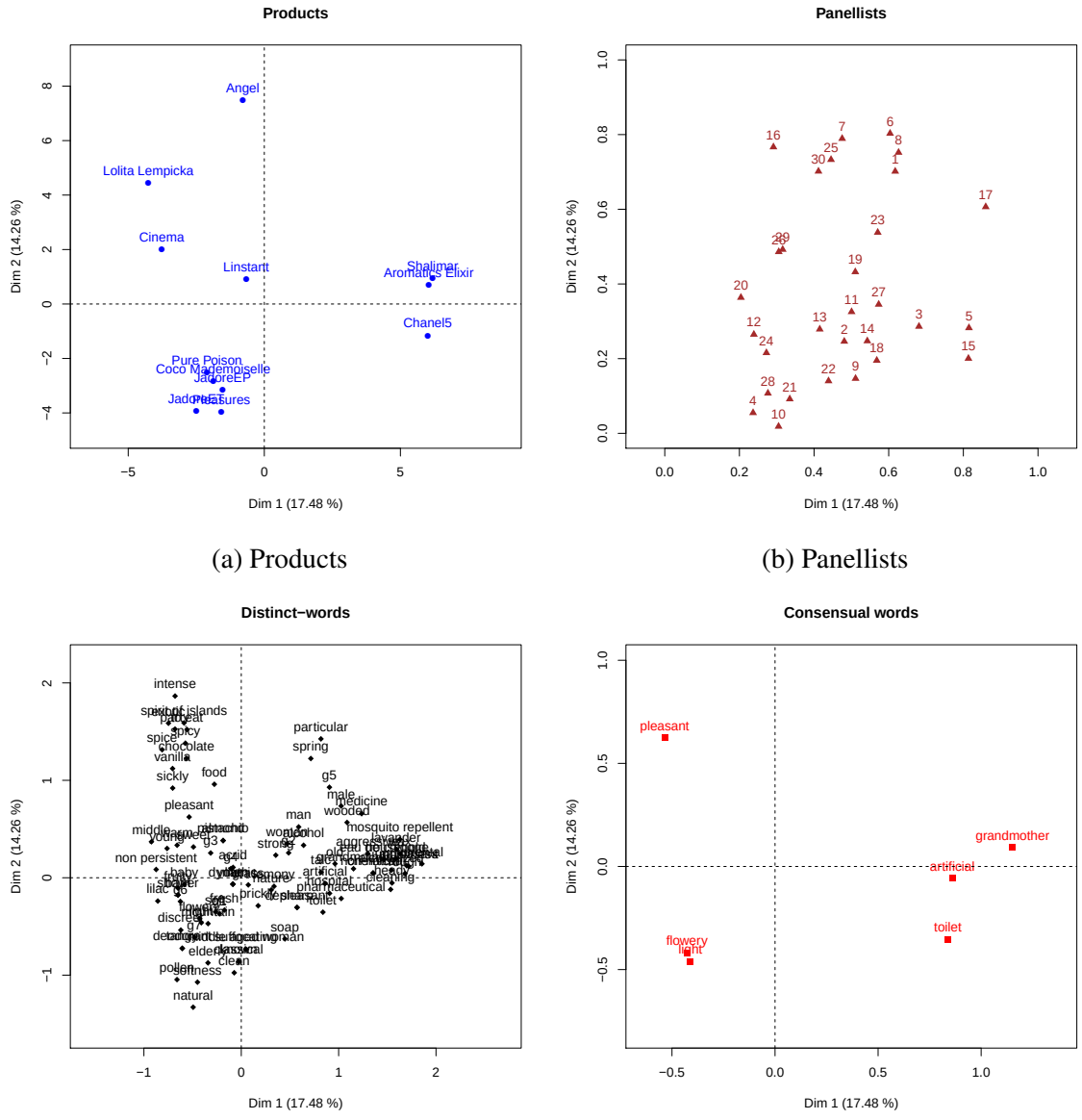
Among the 15 words pronounced by at least 3 consumers, 6 words were assessed to be consensual (p-value less than 0.05).

```
> res$cons
               Nb times Nb panellists Pvalue
grandmother          10             4  0.004
flowery              49            10  0.008
toilet                6             4  0.016
light                23             5  0.018
pleasant              7             3  0.038
artificial            8             3  0.042
soft                 41            10  0.062
fruity               28             8  0.156
eau de cologne        4             3  0.166
old                   7             3  0.192
fresh                15             5  0.238
soap                 10             4  0.250
sweet                23            10  0.312
strong               40            16  0.718
hospital              5             3  0.864
```

```
> res$cons.words
[1] "artificial"  "flowery"     "grandmother" "light"       "pleasant"
[6] "toilet"
```

6.2.4 Graphical Outputs

By default, the *WordCountAna* function gives 4 graphs: one for the products, one for the panellists, one for the distinct words and one for the consensual words (Figure 6.5). The interpretation of these graphs was done in chapter 3.



(a) Products (b) Panellists

(c) Distinct words (d) Consensual words

Fig. 6.5 Graphical outputs of *WordCountAna*

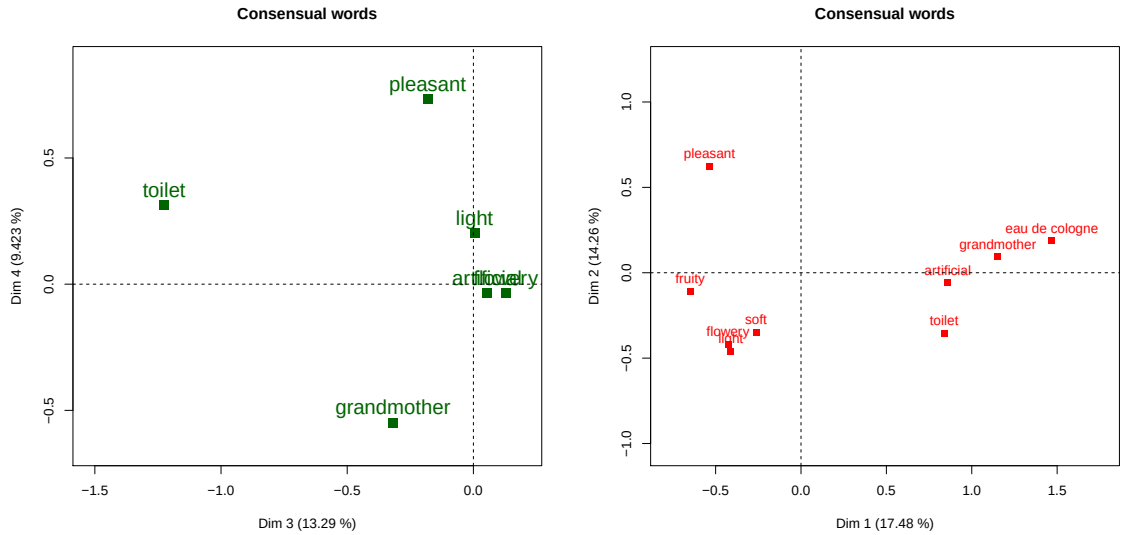
The function *plot.WordCountAna* can be used to improve graphical outputs. To give but two examples:

```
> plot.WordCountAna(res,choix="cons",axes=c(3,4),col="darkgreen",cex=1.5)
```

The consensual words (*choix="cons"*) were plotted for the dimensions 3 and 4 (*axes=c(3,4)*) in darkgreen (*col="darkgreen"*) using a magnification 1.5 (*cex=1.5*) (Figure 6.6a).

```
> plot.WordCountAna(res,choix="cons",proba=0.2)
```

The threshold of the significance is increased up to 0.2 (*proba=0.2*) and three new consensual words were identified (Figure 6.6b).



(a) Consensual words for the dimensions 3-4 (b) Consensual words with threshold 0.2

Fig. 6.6 Examples *plot.WordCountAna*

6.3 CaGalt

The R function *CaGalt* is currently part of the FactoMineR package (Kostov et al., 2015). The default input for the *CaGalt* function in R is

```
CaGalt(Y,X,type="s",conf.ellip=FALSE,nb.ellip=100,level.ventil=0,
sx=NULL,graph=TRUE,axes=c(1,2))
```

6.3.1 Arguments and values

CaGalt includes the following arguments

- *Y*: a data frame with n rows (individuals) and p columns (frequencies)
- *X*: a data frame with n rows (individuals) and k columns (quantitative or categorical variables)
- *type*: the type of variables: "c" or "s" for quantitative variables and "n" for categorical variables. The difference is that for "s" variables are scaled to unit variance (by default, variables are scaled to unit variance)
- *conf.ellip*: boolean (FALSE by default), if TRUE, draw confidence ellipses around the frequencies and the variables when "graph" is TRUE
- *nb.ellip*: number of bootstrap samples to compute the confidence ellipses (by default 100)
- *level.ventil*: proportion corresponding to the level under which the category is ventilated; by default, 0 and no ventilation is done. Available only when type is equal to "n"
- *sx*: number of principal components kept from the principal axes analysis of the contextual variables (by default this is NULL and all principal components are kept)
- *graph*: boolean, if TRUE a graph is displayed
- *axes*: a length 2 vector specifying the components to plot

The returned value of *CaGalt* is a list containing

- *eig*: a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
- *ind*: a list of matrices containing all the results for the individuals (coordinates, square cosine)
- *freq*: a list of matrices containing all the results for the frequencies (coordinates, square cosine, contributions)

- *quanti.var*: a list of matrices containing all the results for the quantitative variables (coordinates, correlation between variables and axes, square cosine)
- *quali.var*: a list of matrices containing all the results for the categorical variables (coordinates of each categories of each variables, square cosine)
- *ellip*: a list of matrices containing the coordinates of the frequencies and variables for replicated samples from which the confidence ellipses are constructed

To improve the output of the results and the graphs, two complementary functions, *summary.CaGalt* and *plot.CaGalt*, were also implemented in FactoMineR.

6.3.2 Application

To illustrate the outputs and graphs of *CaGalt*, we use the data set presented in chapter 4. The first 115 columns correspond to the frequencies of the words in respondents' answers and the last three columns correspond to the categorical variables corresponding to respondents' characteristics (age, gender, health condition).

The data are read from the FactoMineR package:

```
> library(FactoMineR)
> data(health)
```

The code to perform the *CaGalt* is

```
> res.cagalt<-CaGalt(Y=health[,1:115],X=health[,116:118],type="n")
```

6.3.3 Numerical Outputs

The results are given in a list for the individuals, the frequencies, the variables and the confidence ellipses.

```
> res.cagalt
**Results for the Correspondence Analysis on Generalised Aggregated
Lexical Table (CaGalt)**
*The results are available in the following objects:
   name               description
```

```
1  "$eig"             "eigenvalues"
2  "$ind"             "results for the individuals"
3  "$ind$coord"       "coordinates for the individuals"
4  "$ind$cos2"        "cos2 for the individuals"
5  "$freq"            "results for the frequencies"
6  "$freq$coord"      "coordinates for the frequencies"
7  "$freq$cos2"       "cos2 for the frequencies"
8  "$freq$contrib"    "contributions of the frequencies"
9  "$quali.var"       "results for the categorical variables"
10 "$quali.var$coord" "coordinates for the categories"
11 "$quali.var$cos2"  "cos2 for the categories"
12 "$ellip"           "coordinates to construct confidence ellipses"
13 "$ellip$freq"      "coordinates of the ellipses for the frequencies"
14 "$ellip$var"       "coordinates of the ellipses for the variables"
```

The interpretation of the numerical outputs can be facilitated using *summary.CaGalt* function which prints summaries of the *CaGalt* objects.

```
> summary.CaGalt(res.cagalt)
Eigenvalues
                       Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7
Variance               0.057  0.036  0.026  0.024  0.020  0.013  0.012
% of var.             30.21  19.02  13.78  12.95  10.85   6.82   6.37
Cumulative % of var.  30.21  49.23  63.01  75.96  86.81  93.63 100.00

Individuals (the 10 first individuals)
                  Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
6              |  0.120  0.037 | -0.551  0.781 | -0.065  0.011 |
7              | -0.134  0.019 | -0.788  0.649 | -0.166  0.029 |
9              |  0.056  0.002 |  0.272  0.047 | -0.211  0.028 |
10             |  0.015  0.001 | -0.262  0.342 | -0.084  0.035 |
11             | -1.131  0.293 |  0.775  0.138 | -0.613  0.086 |
13             | -0.909  0.231 | -0.340  0.032 |  0.464  0.060 |
14             |  0.097  0.026 |  0.070  0.014 |  0.236  0.154 |
15             | -0.718  0.117 | -1.524  0.526 | -0.717  0.116 |
17             | -0.924  0.372 |  0.074  0.002 |  0.954  0.397 |
18             | -0.202  0.050 |  0.563  0.389 |  0.404  0.200 |
```

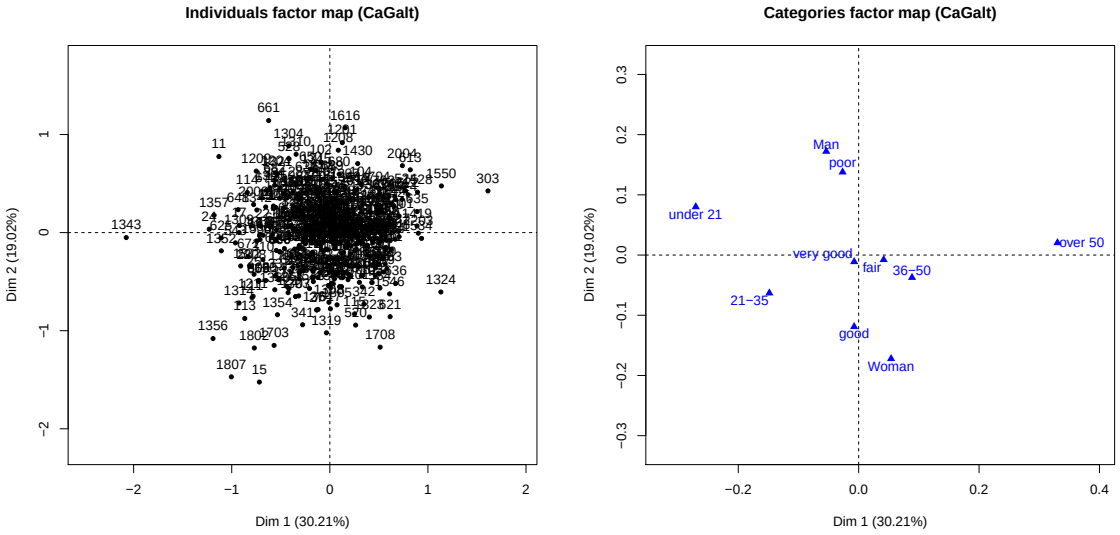
```
Frequencies (the 10 first most contributed frequencies on the first
principal plane)
                  Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr  cos2
physically      | -0.51  6.06  0.94 | -0.04  0.05  0.01 | -0.05  0.10  0.01 |
to have         |  0.24  5.65  0.73 |  0.05  0.45  0.04 | -0.03  0.16  0.01 |
well            | -0.12  0.79  0.17 | -0.26  5.25  0.70 | -0.07  0.59  0.06 |
to feel         | -0.22  1.17  0.22 | -0.32  4.06  0.48 | -0.16  1.35  0.12 |
hungry          |  0.55  1.50  0.25 | -0.63  3.16  0.34 | -0.37  1.48  0.11 |
I               |  0.36  2.93  0.54 | -0.21  1.62  0.19 | -0.11  0.64  0.05 |
one             |  0.25  0.96  0.24 |  0.37  3.45  0.54 |  0.12  0.50  0.06 |
something       | -0.83  2.96  0.43 |  0.44  1.35  0.12 | -0.73  5.12  0.34 |
best            |  0.67  4.18  0.60 |  0.09  0.12  0.01 |  0.23  1.04  0.07 |
psychologically | -0.37  0.56  0.16 | -0.73  3.44  0.60 |  0.19  0.32  0.04 |

Categorical variables
               Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
21-35        | -0.148  0.347 | -0.063  0.063 |  0.148  0.347 |
36-50        |  0.089  0.108 | -0.037  0.019 |  0.120  0.199 |
over 50      |  0.330  0.788 |  0.020  0.003 | -0.028  0.006 |
under 21     | -0.271  0.484 |  0.080  0.042 | -0.240  0.382 |
Man          | -0.054  0.081 |  0.172  0.826 |  0.018  0.009 |
Woman        |  0.054  0.081 | -0.172  0.826 | -0.018  0.009 |
fair         |  0.042  0.029 | -0.008  0.001 | -0.144  0.342 |
good         | -0.007  0.001 | -0.119  0.185 | -0.077  0.077 |
poor         | -0.027  0.002 |  0.138  0.061 |  0.193  0.120 |
very good    | -0.007  0.000 | -0.011  0.001 |  0.027  0.005 |
```

The numerical outputs corresponding to the individuals, frequencies and variables are useful especially as a help to interpret the graphical outputs.

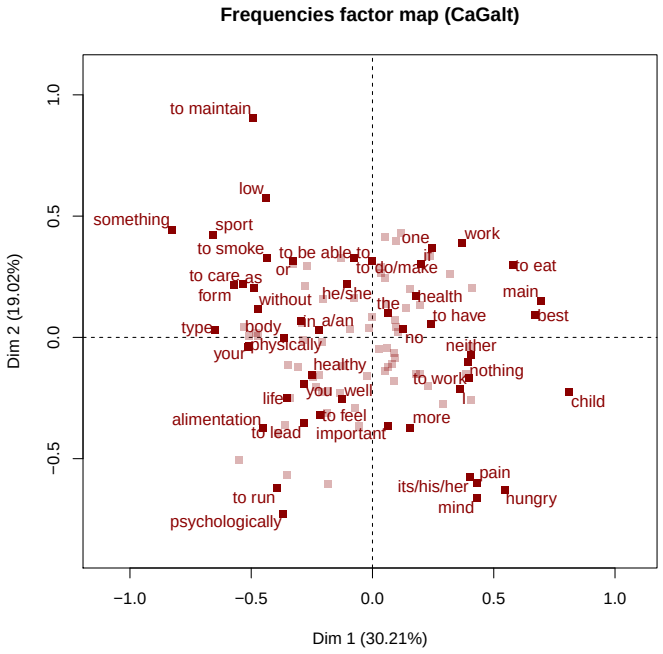
The arguments of the *summary.CaGalt* function *nbelements* (number of written elements), *nb.dec* (number of printed decimals) and *ncp* (number of printed dimensions) can also be modified to obtain more detailed numerical outputs.

6.3.4 Graphical Outputs

The *CaGalt* function returns a series of graphics as representation of the individuals, the variables and the frequencies (Figure 6.7).



(a) Individuals (b) Categories



(c) Frequencies

Fig. 6.7 CA-GALT results on the first principal plane

The high number of elements drawn on the graphs (points and labels) can make the reading and the interpretation difficult. The function *plot.CaGalt* can improve them through replacing the labels, modifying their size, changing the colours, adding confidence ellipses or only selecting a part of the elements. For example, for the categorical variables we will use the code

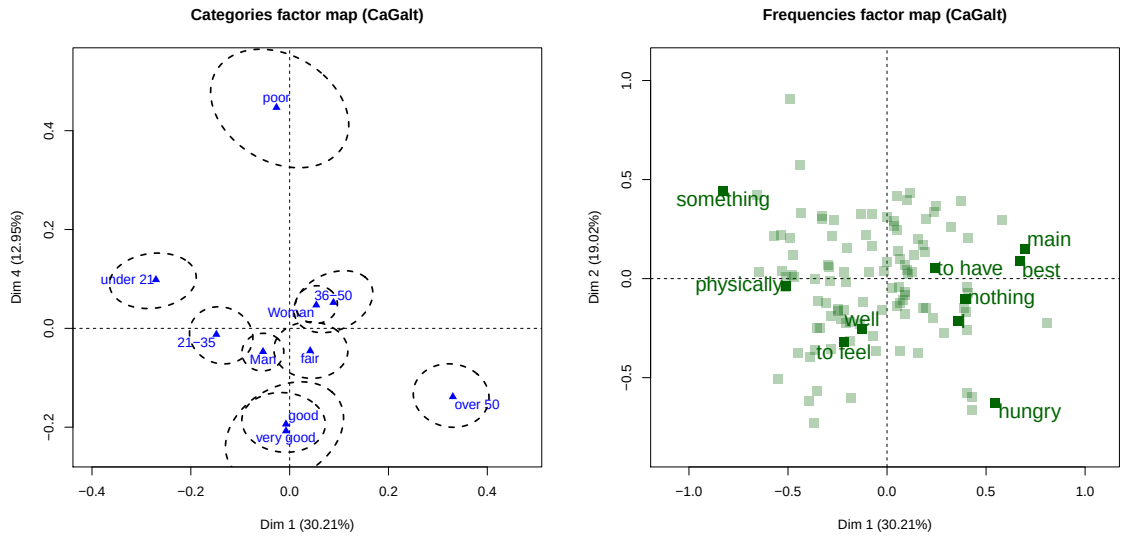
```
> plot.CaGalt(res.cagalt,choix="quali.var",conf.ellip=TRUE,axes=c(1,4))
```

The parameter *choix="quali.var"* indicates that we plot the graph of the categorical variables, the parameter *axes=c(1,4)* indicates that the graph is done for the dimensions 1 and 4, and the parameter *conf.ellip=TRUE* indicates that confidence ellipses are drawn around the categorical variables (Figure 6.8a).

Another example is provided through the following code

```
> plot.CaGalt(res.cagalt,choix="freq",cex=1.5,col.freq="darkgreen",
select="contrib 10")
```

The ten frequencies (*choix="freq"*) with highest contributions (*select="contrib 10"*) are selected and plotted in darkgreen (*col.freq="darkgreen"*) using a magnification 1.5 (*cex=1.5*) relative to the default (Figure 6.8b).



(a) Categories for plane (1,4) with conf. ellipses (b) Ten words with the highest contributions

Fig. 6.8 Examples *plot.CaGalt*

6.4 MfaGalt

The R function *MfaGalt* (Multiple Factor Analysis on Generalised Aggregated Lexical Table) has been developed. This function will be included in the next release of package FactoMineR. Meanwhile, it can be requested from the author. The default input for *MfaGalt* function in R is

```
MfaGalt(Y,X,type="c",name.group=NULL,graph=TRUE,axes=c(1,2))
```

6.4.1 Arguments and values

MfaGalt includes the following arguments

- *Y*: a list of data frames where each data frame "l" contains n(l) rows (individuals) and p(l) columns (frequencies)
- *X*: a list of data frames where each data frame "l" contains n(l) rows (individuals) and k columns (quantitative or categorical variables)
- *type*: the type of variables: "c" for quantitative variables and "n" for categorical variables (by default, variables are considered as quantitative)
- *name.group*: a vector containing the name of the groups (by default, NULL and the group are named group.1, group.2 and so on)
- *graph*: boolean, if TRUE a graph is displayed
- *axes*: a length 2 vector specifying the components to plot

The returned value of *MfaGalt* is a list containing

- *eig*: a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
- *freq*: a list of matrices containing all the results for the frequencies (coordinates, square cosine, contributions)
- *quanti.var*: a list of matrices containing all the results for the quantitative variables (coordinates, correlation between variables and axes, square cosine, partial coordinates)

- *quali.var*: a list of matrices containing all the results for the categorical variables (coordinates of each categories of each variables, square cosine, partial coordinates)
- *group*: a list of matrices containing all the results for the groups (Lg and RV coefficients, coordinates, square cosine, contributions)

To improve the output of the graphs a complementary function *plot.MfaGalt* is also implemented in R.

6.4.2 Application

To illustrate the outputs and graphs of *MfaGalt*, we use the ecological data set presented in chapter 5.

The data are read from the ade4 package.

```
> library(ade4)
> data(meau)
```

Then, a sequence of 4 coupled tables is built.

```
Y<-X<-list()
X[[1]]<-meau$env[1:6,]
X[[2]]<-meau$env[7:12,]
X[[3]]<-meau$env[13:18,]
X[[4]]<-meau$env[19:24,]
Y[[1]]<-meau$spe[1:6,]
Y[[2]]<-meau$spe[7:12,]
Y[[3]]<-meau$spe[13:18,]
Y[[4]]<-meau$spe[19:24,]
colnames(Y[[1]])<-paste(colnames(Y[[1]]),"_sp",sep="")
colnames(Y[[2]])<-paste(colnames(Y[[2]]),"_su",sep="")
colnames(Y[[3]])<-paste(colnames(Y[[3]]),"_au",sep="")
colnames(Y[[4]])<-paste(colnames(Y[[4]]),"_wi",sep="")
Y[[1]]<-Y[[1]][,apply(Y[[1]],2,sum)!=0]
Y[[2]]<-Y[[2]][,apply(Y[[2]],2,sum)!=0]
Y[[3]]<-Y[[3]][,apply(Y[[3]],2,sum)!=0]
Y[[4]]<-Y[[4]][,apply(Y[[4]],2,sum)!=0]
```

The code to perform the *CAMGALT* is

```
res.mfagalt<-MfaGalt(Y,X,type="c",
name.group=c("Spring","Summer","Autumn","Winter"))
```

The *type* argument specifies the type of the contextual tables, here quantitative tables, *name.group* gives the name of each table.

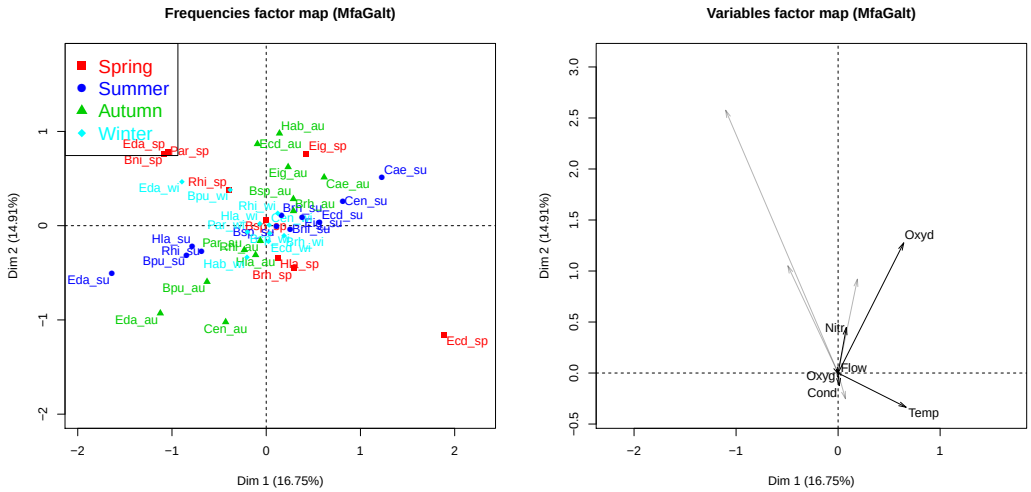
6.4.3 Numerical Outputs

The results are given in a list for the frequencies, the variables and the groups.

```
> print.MfaGalt(res.mfagalt)
**Results for the Multiple Factor Analysis on Generalised Aggregated
Lexical Table (MFA-GALT)**
*The results are available in the following objects:
   name                        description
1  "$eig"                      "eigenvalues"
2  "$freq"                     "results for the frequencies"
3  "$freq$coord"               "coordinates for the frequencies"
4  "$freq$cos2"                "cos2 for the frequencies"
5  "$freq$contrib"             "contributions of the frequencies"
6  "$quanti.var"               "results for the quantitative variables"
7  "$quanti.var$coord"         "coordinates for the quantitative variables"
8  "$quanti.var$cor"           "correlations between quantitative variables
and dimensions"
9  "$quanti.var$cos2"          "cos2 for the quantitative variables"
10 "$quanti.var$coord.partial" "partial coordinates for the quantitative
variables"
11 "$group"                    "results for all the groups"
```

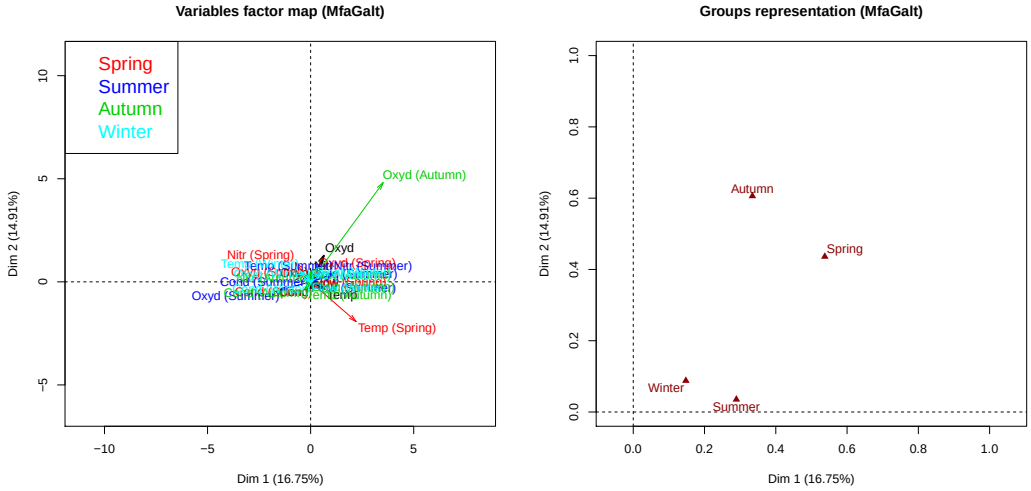
6.4.4 Graphical Outputs

By default, the *MfaGalt* function gives representation of the frequencies, variables, partial variables and groups (Figure 6.9).



(a) Frequencies

(b) Variables



(c) Partial variables

(d) Groups

Fig. 6.9 Graphical outputs of *MfaGalt*

The graphical outputs could be improved using *plot.MfaGalt* function. Although, the number of arguments of this function is high, the most important arguments to know are four:

- *choix*: a string corresponding to the graph that you want to do ("freq" for the frequencies, "var" for the variables, "group" for the groups representation, "partial" for the partial representations, "cor" for the representation of quantitative variables as correlations)
- *axes*: a length 2 vector specifying the components to plot

- *pos.legend*: a single keyword from the list "bottomright", "bottom", "bottomleft", "left", "topleft", "top", "topright", "right" and "center" to specify the location of the legend with group names)
- *select*: a selection of the elements that are drawn

For example, using the code

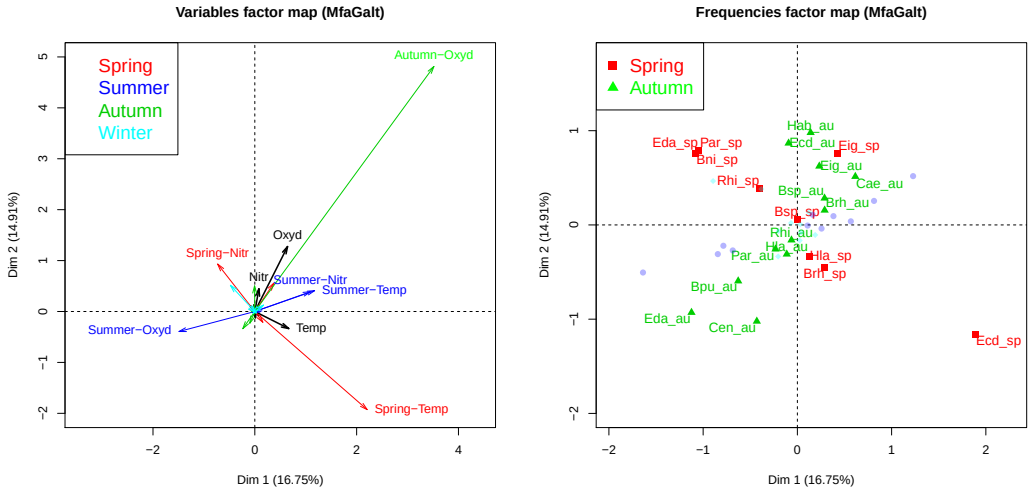
```
plot.MfaGalt(res.mfagalt,choix="partial",label=FALSE,lim.cos2=0.2,
xlim=c(-2,3),ylim=c(-2,5))
coord.part<-res.mfagalt$quanti.var$coord.partial
text(coord.part[[3]][c(7),1],coord.part[[3]][c(7),2],"Autumn-Oxyd",
pos=c(3),col="green")
text(coord.part[[1]][c(1,9),1],coord.part[[1]][c(1,9),2],
paste("Spring-",rownames(coord.part[[1]][c(1,9),]),sep=""),
pos=c(4,3),col="red")
text(coord.part[[2]][c(1,7,9),1],coord.part[[2]][c(1,7,9),2],
paste("Summer-",rownames(coord.part[[2]][c(1,7,9),]),sep=""),
pos=c(4,2,3),col="blue")
text(res.mfagalt$quanti.var$coord[c(1,7,9),1],
res.mfagalt$quanti.var$coord[c(1,7,9),2],
rownames(res.mfagalt$quanti.var$coord[c(1,7,9),]),pos=c(4,3,3))
```

we plot the partial variables (*choix="partial"*) with a quality of representation higher than 0.2 (*lim.cos2="0.2"*) without labels (*label=FALSE*) changing range for x values (*xlim=c(-2,3)*) and y value (*ylim=c(-2,5)*). Then, we add the labels of the highest partial variables (Figure 6.10a).

Another example is provided through the following code

```
spring<-c("Eda_sp","Bsp_sp","Brh_sp","Bni_sp","Ecd_sp","Rhi_sp","Hla_sp",
"Par_sp","Eig_sp")
autumn<-c("Eda_au","Bsp_au","Brh_au","Bpu_au","Cen_au","Ecd_au","Rhi_au",
"Hla_au","Hab_au","Par_au","Cae_au","Eig_au")
plot.MfaGalt(res.mfagalt,choix="freq",select=c(spring,autumn),cex=1.1,
pos.legend=NA)
legend("topleft",c("Spring","Autumn"),pch=c(15,17),cex=1.5,
text.col=c("red","green"),col=c("red","green"))
```

This time frequencies (*choix="freq"*) of the sets spring and autumn (*select=c(spring,autumn)*) are chosen (Figure 6.10b).



(a) (b)

Fig. 6.10 Examples *plot.MfaGalt*

Chapter 7

Conclusions

A new principal axis method has been developed to deal with data sets encoded into a sequence of coupled tables, each one made of one frequency table and one quantitative/qualitative table. The main difficulty to deal with this complex data structure was that the L frequency tables have neither the rows nor the columns in common, when the comparison of the different sets requires to place all the rows in the same space. The starting point considers a first proposal by Pagès and Bécue-Bertaut (2006) to analyse an open-ended question answered in different languages by different samples. We expand this first proposal to the case of several quantitative/qualitative variables by combining two approaches: CA-GALT-like viewpoint (to adopt metrics similar to those used in CA-GALT) and MFA-like viewpoint (to balance the influence of the subsets in the global analysis and to conserve the separation of the sets in order to compare the structures of the tables through partial analyses). The new method is called *Multiple Factor Analysis on Generalised Aggregated Lexical Table* (MFA-GALT).

The main features of MFA-GALT have been illustrated by using two examples. The first example corresponds to data issued from the initial problem, that is, the analysis of an open-ended question answered in different languages in international surveys. The outputs provided by MFA-GALT allow to study the similarities between words (in the same language), the similarities between the homologous words (in different languages), the associations between words and satisfaction scores, the similarities between satisfaction scores structures (partial representations) and the similarities between groups. The results on this application show that MFA-GALT provides a good synthesis of the separate analysis. The homologous words through the two samples show the similarities between the two structures regarding *air conditioning* and *toilet cleanliness*. The words used by the English speakers to express their

dissatisfaction related to *staff language skills* and the words used by the Spanish speakers to express their dissatisfaction related to *cabin room* point out the main differences.

Another example corresponds to an application to ecological data. The data set is a sequence of 4 paired tables with 6 rows for each paired table. So, in this second example, we deal with a very small data set. This data set also has been studied by other methods proposed to deal with a sequence of ecological paired tables (Thioulouse, 2011). The results on this application show that MFA-GALT can be applied with benefit to this kind of data, providing interesting results that are complementary to those obtained by the already existing methods.

The development of MFA-GALT required first to generalize what are the aggregated lexical table to several quantitative/categorical variables to expand the first proposal by Pagès and Bécue-Bertaut (2006) to the case of several quantitative/categorical variables. Precisely, CA-GALT defines the generalised aggregated lexical table (Bécue-Bertaut et al., 2014). The application of the method to a real data set has demonstrated how free-text and closed answers combine to provide relevant information.

Besides, this thesis also offers another original contribution. A new methodology is proposed to identify the consensual words in sensory studies where free-text descriptions and equivalent methods are frequently used (Kostov et al., 2014).

All the methods presented in this thesis are implemented in R. CA-GALT and MFACT, through an option of the *MFA* function in the package FactoMineR, software implementations favouring simple function syntax try to make easier dealing with these complex data sets. To ease the understanding of how to use these functions and interpret the outputs, a statistical computing paper is published in The R Journal (Kostov et al., 2013) and another one has recently been accepted (Kostov et al., 2015). The R function *WordCountAna*, the method proposed to identify the consensual words in sensory studies, has been integrated into the package SensoMineR. The function developed to perform MFA-GALT will be included in the next release of the package FactoMineR.

As future works, we can mention first the method proposed to deal with a sequence of coupled tables can be extended to mixed contextual variables. However, this is not straightforward because it is necessary to adopt a MFA-like approach balancing the influences of the sets of variables (quantitative or categorical). However, MFA-like approach is already considered by MFA-GALT to balance the influence of the sets of rows. Furthermore, the weighting system on the variables is not a diagonal matrix. Thus, the classical way to normalise the highest axial inertia of each set to 1 is not valid in this case. Secondly, the methodology proposed to identify the consensual words in sensory studies can be applied in

the case of the multilanguage open-ended question to study the meaning of the homologous words as has been detailed in chapter 3.

In summary, this thesis proposed a new method to deal with a sequence of coupled tables. Several methodological and software contributions were made. Results on the two applications show that the method provides outputs that are easy to interpret.

References

Abascal, E., García Lautre, I., and Landaluce, M. I. (2006). Multiple factor analysis of mixed tables of metric and categorical data. In Greenacre, M. and Blasius, J., editors, *Multiple correspondence analysis and related methods*, pages 351–367. Chapman & Hall / CRC PRESS.

Abdi, H., Valentin, D., Chollet, S., and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, 18(4):627–640.

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

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

Bécue-Bertaut, M. and Pagès, J. (2014). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. *Advances in Data Analysis and Classification*.

Bécue-Bertaut, M., Pagès, J., and Kostov, B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. a statistical approach. *SORT*, 38:285–302.

Benzécri, J. P. (1973). *Analyse des Données*. Dunod.

Benzécri, J. P. (1981). *Pratique de l'analyse des données: Linguistique & lexicologie*, volume 3. Dunod.

Benzécri, J. P. (1983). Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence. *Les Cahiers de l'Analyse des Données*, 8(3):351–358.

Brandimarte, P. (2011). *Quantitative methods: an introduction for business management*. John Wiley & Sons.

Cadoret, M., Lê, S., and Pagès, J. (2009). A factorial approach for sorting task data (fast). *Food Quality and Preference*, 20(6):410–417.

Campo, E., Ballester, J., Langlois, J., Dacremont, C., and Valentin, D. (2010). Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for

odor profiling: An application to burgundy pinot noir wines. *Food Quality and Preference*, 21(1):44–55.

Cazes, P. and Moreau, J. (1991). Analysis of a contingency table in which the rows and the columns have a graph structure. In Diday, E. and Lechevallier, Y., editors, *Symbolic-Numeric Data Analysis and Learning*, pages 271–280. Nova Science Publishers, NewYork,.

Cazes, P. and Moreau, J. (2000). Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique. In Moreau, J., Doudin, P., and Cazes, P., editors, *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*, page 87–103. Springer, Berlin-Heidelberg.

Chollet, S., Lelièvre, M., Abdi, H., and Valentin, D. (2011). Sort and beer: Everything you wanted to know about the sorting task but did not dare to ask. *Food quality and preference*, 22(6):507–520.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Delarue, J. and Sieffermann, J.-M. (2004). Sensory mapping using flash profile. comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food quality and preference*, 15(4):383–392.

D'Esposito, M. R., De Stefano, D., and Ragozini, G. (2014). On the use of multiple correspondence analysis to visually explore affiliation networks. *Social Networks*, 38:28–40.

Dolédec, S. and Chessel, D. (1994). Co-inertia analysis: An alternative method for studying species-environment relationships. *Freshwater Biology*, 31:277–294.

Dray, S., Chessel, D., and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84(11):3078–3089.

Dray, S., Dufour, A.-B., et al. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, 22(4):1–20.

Efron, B. (1979). Bootstrap methods: Another look at jackknife. *The Annals of Statistics*, 7:1–26.

Escofier, B. and Drouet, D. (1983). Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données*, 8(4):491–499.

Escofier, B. and Pagès, J. (1988). *Analyses factorielles simples et multiples*. Dunod.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29:751–760.

Finn, J. D. (1974). *A general model for multivariate analysis.* Holt, Rinehart & Winston.

Franquet, E., Dolédec, S., and Chessel, D. (1995). Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. *Hydrobiologia*, 300:425–431.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

Guo, L., Ma, J., Chen, Z., and Zhong, H. (2014). Learning to recommend with social contextual information from implicit feedback. *Soft Computing*, 19(5):1351–1362.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2006). Selection of effective contextual information for automatic synonym acquisition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 353–360. Association for Computational Linguistics.

Härdle, W. and Simar, L. (2012). *Applied multivariate statistical analysis*. Springer Verlag, Heidelberg.

Jiang, M., Cui, P., Wang, F., Zhu, W., and Yang, S. (2014). Scalable recommendation with social contextual information. *IEEE Transactions on Knowledge and Data Engineering*, 26(11):2789–2802.

Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99.

Josse, J., Pagès, J., and Husson, F. (2008). Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91.

Kostov, B., Bécue-Bertaut, M., and Husson, F. (2013). Multiple factor analysis for contingency tables in FactoMineR package. *The R Journal*, 5:29–38.

Kostov, B., Bécue-Bertaut, M., and Husson, F. (2014). An original methodology for the analysis and interpretation of word-count based methods: Multiple factor analysis for contingency tables complemented by consensual words. *Food Quality and Preference*, 32:35–40.

Kostov, B., Bécue-Bertaut, M., and Husson, F. (2015). Correspondence analysis on generalised aggregated lexical table (CA-GALT) in the FactoMineR package. *The R Journal*.

Lancaster, B. and Foley, M. (2007). Determining statistical significance for choose-allthat-apply question responses. In *Seventh Pangborn Sensory Science Symposium*, Minneapolis, USA.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The act (statis method). *Computational Statistics and Data Analysis*, 18:97–119.

Lê, S. and Husson, F. (2008). Sensominer: A package for sensory data analysis. *Journal of Sensory Studies*, 23:14–25.

Lebart, L., Morineau, A., and Piron, M. (1997). *Statistique exploratoire multidimensionnelle*. Dunod.

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

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*. Elsevier Science, Amsterdam.

Littman, M., Dumais, S., and Landauer, T. (1998). Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette, G., editor, *Cross-Language Information Retrieval*, volume 2 of *The Springer International Series on Information Retrieval*, pages 51–62. Springer US.

Mallat, S., Hkiri, E., Maraoui, M., and Zrigui, M. (2015). Lexical network enrichment using association rules model. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 59–72. Springer International Publishing.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall / CRC Press.

Pagès, J. (2004). Multiple factor analysis: main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1):1–26.

Pagès, J. and Bécue-Bertaut, M. (2006). Multiple factor analysis for contingency tables. In Greenacre, M. and Blasius, J., editors, *Multiple correspondence analysis and related methods*, pages 299–326. Chapman & Hall / CRC PRESS.

Passmore, D. L. (2011). Social network analysis: Theory and applications.

Perrin, L. and Pagès, J. (2009). Construction of a product space from the ultra-flash profiling method: Application to 10 red wines from the Loire Valley. *Journal of Sensory Studies*, 24(3):372–395.

Roberts, J. M. (2000). Correspondence analysis of two-mode network data. *Social Networks*, 22(1):65–72.

Scott, J. (2002). *Social networks: Critical concepts in sociology*, volume 4. Taylor & Francis.

Simier, M., Blanc, L., Pellegrin, F., and Nandris, D. (1999). Approche simultanee de k couples de tableaux: Application à l'étude des relations pathologie végétale-environnement. *Revue de Statistique Appliquée*, 47:31–46.

ten Kleij, F. and Musters, P. A. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food quality and preference*, 14(1):43–52.

ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179.

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics*, 5(4):2300–2325.

Thioulouse, J. and Chessel, D. (1987). Les analyses multitableaux en ecologie factorielle. i : de la typologie d'etat a la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica Oecologia Generalis*, 8(4):463–480.

Thioulouse, J., Simier, M., and Chessel, D. (2004). Simultaneous analysis of a sequence of paired ecological tables. *Ecology*, 85:272–283.

Varela, P. and Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. a review of novel methods for product characterization. *Food Research International*, 48(2):893–908.

Veinand, B., Godefroy, C., Adam, C., and Delarue, J. (2011). Highlight of important product characteristics for consumers. comparison of three sensory descriptive methods performed by consumers. *Food Quality and Preference*, 22(5):474–485.

Wasserman, S. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press.

Appendix A

Publications, conferences and software

A.1 Publications related with this thesis

1. Kostov B, Bécue-Bertaut M, and Husson F. (2015). Correspondence analysis on generalised aggregated lexical table (CA-GALT) in the FactoMineR package. *The R Journal*, (Accepted).

2. Bécue-Bertaut M, Pagès J, and Kostov B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. a statistical approach. *SORT*, 38:285-302.

3. Kostov B, Bécue-Bertaut M, and Husson F. (2014). An original methodology for the analysis and interpretation of word-count based methods: Multiple factor analysis for contingency tables complemented by consensual words. *Food Quality and Preference*, 32:35-40.

4. Kostov B, Bécue-Bertaut M, and Husson F. (2013). Multiple factor analysis for contingency tables in FactoMineR package. *The R Journal*, 5:29-38.

A.2 Conferences related with this thesis

1. Kostov B, Bécue-Bertaut M and Husson F. (2013). Análisis factorial múltiple dual de tablas de frecuencias agregadas generalizadas: AFMD-TFAG. *Congreso Nacional de Estadística e Investigación Operativa*, Castellón (Spain) 11-13 September.

2. Kostov B, Bécue-Bertaut M and Husson F. (2013). A new principal component method to analyse frequency tables with different row and column sets using common instrumental variables: application to a multilingual survey. *Stochastic Modeling Techniques and Data Analysis International Conference*, Mataró (Catalonia) 25-28 Juny.

3. Kostov B, Bécue-Bertaut M, Husson F, Pagès J, Cadoret M, Torrens J and Urpi P. (2012). A tool for detecting words with consensual meaning in labeled categorized napping. *Sensometrics*, Rennes (France) 11-13 July.

4. Kostov B, Bécue-Bertaut M, Husson F and Hernández D. (2012). Multiple Factor Analysis for Contingency Tables in FactoMineR Package. *Rencontres R*, Bordeaux (France) 3-4 July.

5. Kostov B, Bécue-Bertaut M, Pagès J, Cadoret M, Torrens J and Urpi P. (2011). Verbalisation tasks in Hall test sessions. *International Classification Conference*, St. Andrews (Scotland) 11-15 July.

A.3 Software contributions

1. Multiple Factor Analysis (MFA) {FactoMineR} http://factominer.free.fr

2. Word-Count based methods Analysis (WordCountAna) {SensoMineR} http://sensominer.free.fr

3. Correspondence Analysis on Generalised Aggregated Lexical Table (CaGalt) {FactoMineR} http://factominer.free.fr

4. Multiple Factor Analysis on Generalised Aggregated Lexical Table (MfaGalt)

Appendix B

MFA-GALT code

```
MfaGalt<-function(Y,X,type="c",name.group=NULL,graph=TRUE,axes=c(1,2)){
library(FactoMineR)
library(MASS)
num.group<-length(Y)
for(i in 1:num.group) Y[[i]]<-as.matrix(Y[[i]])
if (is.null(name.group)) name.group<-paste("GROUP",1:num.group,sep = ".")
if (length(Y)!=length(X))
stop("Number of frequency tables should be equal to the number of cont. tables")
if (length(unique(sapply(X,ncol)))!=1)
stop("Number of variables should be the same for all the samples")
if (!type%in%c("c","s","n")) stop("not convenient type definition")
mean.p<-function(V, poids)
res <- sum(V * poids, na.rm = TRUE)/sum(poids[!is.na(V)])
sd.p<-function(V, poids)
res <- sqrt(sum(V^2 * poids, na.rm = TRUE)/sum(poids[!is.na(V)]))
N<-sum(unlist(Y))
Nl<-sapply(Y,sum)
num.ind<-sapply(Y,nrow)
num.freq<-sapply(Y,ncol)
names(num.freq)<-names(num.ind)<-names(Nl)<-name.group
Pl<-mapply(function(y,nl) y/nl,Y,Nl,SIMPLIFY = FALSE)
pi.l<-lapply(Pl,function(p) apply(p,1,sum))
p.jl<-lapply(Pl,function(p) apply(p,2,sum))
wl<-Nl/N
```

```
DIl<-mapply(function(pi,wl) pi*wl,pi.l,wl,SIMPLIFY = FALSE)
DJl<-mapply(function(pj,wl) pj*wl,p.jl,wl,SIMPLIFY = FALSE)
DI<-unlist(DIl)
DJ<-unlist(DJl)
Xl<-vector(mode='list',length=num.group)
if(type=="n") for(i in 1:num.group) Xl[[i]]<-as.matrix(tab.disjonctif(X[[i]]))
else for(i in 1:num.group) Xl[[i]]<-as.matrix(X[[i]])
if(type!="n"|unique(sapply(X,ncol))!=1)
Xl<-mapply(function(x,pi) sweep(x,2,apply(x,2,mean.p,pi),"-"),
Xl,pi.l,SIMPLIFY = FALSE)
XG<-data.frame()
for (i in 1:num.group) XG<-as.matrix(rbind(XG,Xl[[i]]))
Yal<-mapply(function(y,x) crossprod(y,x),Y,Xl,SIMPLIFY = FALSE)
Pal<-mapply(function(ya) ya/N,Yal,SIMPLIFY = FALSE)
C<-crossprod(sweep(XG,1,DI^(1/2),"*"),sweep(XG,1,DI^(1/2),"*"))
Ml<-mapply(function(di,dj,w,x)
diag(dj)%*%matrix(nrow=length(dj),ncol=length(di),1/w)%*%diag(di)%*%
x,DIl,DJl,wl,Xl,SIMPLIFY = FALSE)
Zl<-mapply(function(pa,m,dj) sweep(pa-m,1,dj,"/")%*%ginv(C),Pal,Ml,DJl,
SIMPLIFY=FALSE)
ponde<-mapply(function(z,dj) Re(eigen(t(z)%*%diag(dj)%*%z%*%C)$values[1]),
Zl,DJl,SIMPLIFY = FALSE)
row.w<-unlist(mapply(function(dj,pond) dj/pond,DJl,ponde,SIMPLIFY = FALSE))
Z<-data.frame()
for (i in 1:num.group) Z<-as.matrix(rbind(Z,Zl[[i]]))
diag.MFAGALT<-eigen(t(Z)%*%diag(row.w)%*%Z%*%C)
eig<-Re(diag.MFAGALT$values)
if(type=="n") eig<-eig[1:(length(eig)-unique(sapply(X,ncol)))]
vp<-as.data.frame(matrix(NA,length(eig),3))
rownames(vp)<-paste("comp",1:length(eig))
colnames(vp)<- c("eigenvalue","percentage of variance",
"cumulative percentage of variance")
vp[,"eigenvalue"]<-eig
vp[,"percentage of variance"]<-(eig/sum(eig))*100
vp[,"cumulative percentage of variance"]<-cumsum(vp[,"percentage of variance"])
U<-sweep(Re(diag.MFAGALT$vectors[,1:length(eig)]),2,
```

```
sqrt(diag(t(Re(diag.MFAGALT$vectors[,1:length(eig)]))%*%C%*%
Re(diag.MFAGALT$vectors[,1:length(eig)]))),"/")
coord.freq<-Z%*%C%*%U
contrib.freq<-sweep(sweep(coord.freq^2,1,row.w,FUN="*")*100,2,eig,FUN="/")
dist2.freq<-diag(Z%*%C%*%t(Z))
cos2.freq<-sweep(as.matrix(coord.freq^2),1,dist2.freq,FUN="/")
coord.var<-sweep(U,2,sqrt(eig),"*")
dist2.var<-apply(sweep(Z,1,sqrt(row.w),FUN="*")^2,2,sum)
cos2.var<-sweep(as.matrix(coord.var^2),1,dist2.var,FUN="/")
coord.freq.partiel<-mapply(function(z) z%*%C%*%U,Zl,SIMPLIFY=FALSE)
coord.var.partiel<-mapply(function(z,f,dj,pond)
sweep(t(z)%*%sweep(f,1,dj/pond,"*")*num.group,2,sqrt(eig),"/"),
Zl,coord.freq.partiel,DJl,ponde,SIMPLIFY=FALSE)
if(type!="n"){
cor.var<-sweep(sweep(t(Z)%*%sweep(coord.freq,1,row.w,"*"),1,
apply(Z,2,function(V, poids)
res <- sqrt(sum(V^2 * poids, na.rm = TRUE)),row.w),"/"),2,sqrt(eig),"/")
}
tab<-(Z%*%C%*%t(Z))^2
tab<-sweep(sweep(tab,2,row.w,"*"),1,row.w,"*")
Lg<-matrix(0,num.group+1,num.group+1)
ind.gl<-0
for (gl in 1:num.group) {
ind.gc<-0
for (gc in 1:num.group) {
Lg[gl,gc]<-Lg[gc,gl]<-sum(tab[(ind.gl+1):(ind.gl+num.freq[gl]),
(ind.gc+1):(ind.gc+num.freq[gc])])
ind.gc<-ind.gc+num.freq[gc]
}
Lg[num.group+1,gl]<-Lg[gl,num.group+1]<-sum(tab[(ind.gl+1):
(ind.gl+num.freq[gl]),1:ncol(tab)])/eig[1]
ind.gl<-ind.gl+num.freq[gl]
}
Lg[num.group+1,num.group+1]<-sum(tab[1:ncol(tab),1:ncol(tab)])/(eig[1]^2)
RV<-sweep(Lg,2,sqrt(diag(Lg)),"/")
RV<-sweep(RV,1,sqrt(diag(Lg)),"/")
```

```
contrib.group<-matrix(NA,num.group,ncol(U))
dist2.group<-vector(length=num.group)
freq.gr<-0
for (g in 1:num.group) {
if (g==1) contrib.group[g,]<-apply(contrib.freq[1:num.freq[g],],2,sum)
else contrib.group[g,]<-apply(contrib.freq[(freq.gr+1):(freq.gr+num.freq[g]),],
2,sum)
ponde.tot<-eigen(t(Zl[[g]])%*%diag(DJl[[g]])%*%Zl[[g]]%*%C)$values
dist2.group[g]<-sum((ponde.tot/ponde[[g]])^2)
freq.gr<-freq.gr+num.freq[g]
}
coord.group<-sweep(contrib.group/100,2,eig,"*")
cos2.group<-sweep(coord.group^2,1,dist2.group,"/")
rownames(coord.var)<-rownames(cos2.var)<-colnames(XG)
colnames(coord.freq)<-colnames(contrib.freq)<-paste("Dim",c(1:ncol(U)),sep = ".")
colnames(cos2.freq)<-paste("Dim",c(1:ncol(U)),sep = ".")
colnames(coord.var)<-colnames(cos2.var)<-paste("Dim",c(1:ncol(U)),sep = ".")
if(type!="n"){
rownames(cor.var)<-colnames(XG)
colnames(cor.var)<-paste("Dim",c(1:ncol(U)),sep = ".")
}
names(coord.var.partiel)<-name.group
for(i in 1:num.group){
rownames(coord.var.partiel[[i]])<-colnames(XG)
colnames(coord.var.partiel[[i]])<-paste("Dim",c(1:ncol(U)),sep = ".")
}
rownames(coord.group)<-rownames(contrib.group)<-rownames(cos2.group)<-name.group
colnames(coord.group)<-colnames(contrib.group)<-paste("Dim",c(1:ncol(U)),sep = ".")
colnames(cos2.group)<-paste("Dim",c(1:ncol(U)),sep = ".")
rownames(Lg)<-colnames(Lg)<-rownames(RV)<-colnames(RV)<-c(name.group,"MFA-GALT")
res<-list()
res$eig<-vp
res$freq<-list(coord=coord.freq,cos2=cos2.freq,contrib=contrib.freq)
if (type=="n") res$quali.var<-list(coord=coord.var,cos2=cos2.var,
coord.partial=coord.var.partiel)
else res$quanti.var<-list(coord=coord.var,cor=cor.var,cos2=cos2.var,
```

```
coord.partial=coord.var.partiel)
res$group<-list(Lg=Lg,RV=RV,coord=coord.group,contrib=contrib.group,
cos2=cos2.group)
res$call<-list(num.groups=num.group,name.groups=name.group,
num.freq=num.freq,type=type)
class(res)<-c("MFAGALT","list")
if (graph) {
plot.MfaGalt(res,choix="freq",axes=axes,select="contrib 60")
warning("The first 60 frequencies that have the highest contribution
on the 2 dimensions of your plot are drawn")
plot.MfaGalt(res,choix="var",axes=axes,new.plot=TRUE,select="cos2 0.2")
plot.MfaGalt(res,choix="partial",axes=axes,new.plot=TRUE,lim.cos2.var=0.2)
plot.MfaGalt(res,choix="group",axes=axes,new.plot=TRUE,cex=1.5)
}
return(res)
}
```