

Journal of Statistical Software

September 2016, Volume 73, Issue 6. doi: 10.18637/jss.v073.i06

Jan de Leeuw and the French School of Data Analysis

François Husson
Agrocampus Ouest Rennes

Julie Josse
Agrocampus Ouest Rennes

Gilbert Saporta
CNAM Paris

Abstract

The Dutch and the French schools of data analysis differ in their approaches to the question: How does one understand and summarize the information contained in a data set? The commonalities and discrepancies between the schools are explored here with a focus on methods dedicated to the analysis of categorical data, which are known either as homogeneity analysis (HOMALS) or multiple correspondence analysis (MCA).

Keywords: HOMALS, multiple correspondence analysis, missing values, analyse des données, French and Dutch schools.

1. Introduction

In the 1960s two currents of research emerged in the spirit of Tukey's exploratory data analysis (Tukey 1962): the French school and the Dutch school. Researchers in these schools were outliers in the statistical landscape of the time, in which most research was performed in the framework of probability models. What can be highlighted is that many of the modern arguments about data sciences, machine learning, statistics (see Donoho 2015, for a nice overview), and inference (see for instance the ASA statement on p-values in Wasserstein and Lazar 2016) were already debated. We feel that the way both schools tackled problems and data was a bit ahead of its time.

The French school of "analyse des données" (data analysis) was led by Jean-Paul Benzécri, a mathematician and linguist, who encouraged the idea of "letting the data speak for themselves". One of his famous quotes (Benzécri 1973, p 6, Tome 2) starts with, "The model must follow the data, not the other way around," and ends with, "What we need is a rigorous method which extracts structures from the data."[1] He described "statistical analysis as a tool

[1]In French: "Le modèle doit suivre les données, non l'inverse ... ce dont nous avons besoin, c'est d'une méthode rigoureuse qui extraie des structures à partir des données."

to make patterns emerge from data (Benzécri 1986).” When presenting methods of “analyse des données”, we say that such methods allow description, exploration and visualization of the data. Furthermore, they involve reducing data dimensionality in order to provide a subspace that best represents the data in the sense of maximizing the variability of the projected points. A great importance is attached to graphical displays and often the representation of rows is as interesting as the representation of the columns. Methods based on principal component analysis have roughly similar aims, such as studying similarities between rows, similarities between columns, and associations between rows and columns, but they differ with respect to the nature of the data: principal components analysis for continuous data, correspondence analysis for contingency tables, and multiple correspondence analysis for categorical data. An intrinsic characteristic of the approaches is that they are presented using geometrical considerations without any references to a probabilistic model[2]. From a technical point of view, the core of all these methods is the singular value decomposition (SVD) of certain matrices with specific row and column weights and metrics (used to compute the distances). In the words of Benzécri, “all in all, doing a data analysis, in good mathematics, is simply searching eigenvectors; all the science (or the art) of it is in finding the right matrix to diagonalize.”[3] Although many contributions in the French school were never translated to English, many references are available and include Benzécri (1982); Le Roux and Rouanet (2004); Murtagh (2005); Holmes (2008); Lebart (2008); Lebart and Saporta (2014); Lebaron and Le Roux (2015). One feature that can be mentioned is the strong connection between J-P. Benzécri and P. Bourdieu (Lebaron and Le Roux 2015) which has a role in explaining the popularity of such methods in the social sciences. However, this popularity is mainly due to the prevalence of categorical data in this area.

It is of course more difficult for us to talk about the Dutch school and to reflect on Jan’s views of statistics, models, and inference without taking the risk of misrepresenting his thoughts. In addition, his views may have evolved through the years. So we would advise the reader to consider two articles we really enjoyed and that may reflect some of his current ideas: “Models of Data” (De Leeuw 2005) and “Statistics and the Sciences” (De Leeuw 2011b). We also refer the reader to the interview of De Leeuw (De Leeuw 2011a) given on the occasion of the International Conference on Correspondence Analysis and Related Methods (CARME) in 2011. We see there many similarities with our day to day practice of statistics, in which we think about the encoding of the data, the use of statistics “as tools for data analysis”, concerns about stability, etc. What we can say for sure is that the absence of models is also a strong characteristic of the Dutch school. In addition, methods of this school known as Gifi’s methods (Gifi 1990; Michailidis and De Leeuw 1998) also reduce the dimensionality and respect the nature of the data, whether categorical or ordinal, for instance. For both schools, coding categorical variables with the indicator matrix of dummy variables and considering them as Gaussian, for instance, is almost a crime. Another strong feature of both schools is that the approaches are completely unsupervised in the sense that very often there is no distinction between explanatory variables and a response variable, or in other words, there is no Y variable.

What is different between the schools is the manner in which problems are of presented and solved problems, mainly based on projections in the French school, and on the definition of a

[2]It does not mean that the concepts of stability, replicability, etc. are not covered.

[3]From our point of view, specific choices of weights and metrics can be seen as inducing specific models for the data under analysis.

loss function solved by an alternating least squares (ALS) algorithm and transformation of the variables in the Dutch school. The difference between these points of view implies different research focuses and developments. It is very interesting to see how the way a problem is written influences the stream of ideas. As we will see in what follows, the projections point of view facilitates the introduction of supplementary elements to enhance the interpretation of the graphical outputs whereas the ALS point of view easily enables the introduction of constraints in the optimization problem.

In this paper, to illustrate the commonalities and discrepancies between the French and Dutch schools we focus on the method dedicated to analyse categorical data, known either as multiple correspondence analysis (MCA) or homogeneity analysis (HOMALS). We start by reviewing both approaches and by presenting how these methods have been extended to deal with missing values. Then, we illustrate the approaches on a survey data set describing genetically modified organisms. Finally, we show how Jan's developments influenced the French school.

2. HOMALS and multiple correspondence analysis

HOMALS and MCA have been successfully applied to describe the relationship between categorical variables in many fields such as the social sciences, marketing, health, psychology, educational research, political science, genetics, etc. (Greenacre and Blasius 2006). They are often used to analyse survey data where participants answer many questions. Let us consider a dataset with n rows and J categorical variables, $\mathbf{v}_{j_{j=1,\ldots,J}}$ with K_j categories each. The data are coded using the indicator matrix of dummy variables denoted $\mathbf{G}_{n\times K}$, $K = \sum_j K_j$ with $g_{ijk} = 1$ if person i selects category k of variable j and $g_{ijk} = 0$ otherwise as illustrated below for three variables with respectively three, three and two categories.

$$\mathbf{G} = [\mathbf{G}_1|\mathbf{G}_2|\mathbf{G}_3] \quad = \quad \left[\begin{array}{ccc|ccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$$

2.1. Classical MCA presentation

Historically, Lebart (Lebart and Tabard 1973) had the idea to apply correspondence analysis (CA) to the indicator matrix $\mathbf{G}$ (Lebart and Saporta 2014). This strategy yields very interesting results with new properties: this is how MCA was born, and this remains its most common definition. Another nearly-equivalent way to perform MCA consists of applying CA on the Burt matrix $\mathbf{B} = \mathbf{G}^\top\mathbf{G}$, which is the matrix of all pairwise associations between the variables. Note that in this table, the information on rows is lost. The final common presentation of MCA consists of performing PCA onto the indicator matrix $\mathbf{G}$ with specific row and column weights. The choice of weights ensures the properties of the method such as the Chi-square interpretation of the distances between rows as well as the fact that the principal components are "new" variables that are maximally related to the set of variables, with the

relationship measured here by the squared correlation ratio (η^2) of analysis of variance. More precisely, let us denote a matrix $\mathbf{X}_{n\times S}$ which represents the principal components (scaled to 1) also known as the normalized scores, i.e., the normalized coordinates of the n observations on the S axes. They satisfy the following property (Saporta 1988a):

$$\mathbf{x}_s = \underset{\mathbf{x}_s\in\mathbb{R}^n}{\arg\max} \quad \frac{1}{J}\sum_{j=1}^{J}\eta^2(\mathbf{x}_s,\mathbf{v}_j) \tag{1}$$

with the constraint that $\mathbf{x}_s$ (the s^{th} column of $\mathbf{X}$) has norm equal to 1 and is orthogonal to $\mathbf{x}_s'$ for all $s' < s$. This expression strengthens the presentation of MCA as an extension of PCA. It also strengthens the practice of performing a clustering method, such as k-means or a hierarchical clustering algorithm onto the S first principal components of MCA. Indeed, it allows both working with continuous variables that summarize the categorical variables and also removes some noise (assuming that the last dimensions are restricted to noise), which stabilizes the clustering (Husson, Lê, and Pagès 2010). Note that the complementarity between clustering and principal components methods is usual in the French school of data analysis.

That three presentations of MCA can be seen as a strength of the method (Husson and Josse 2014). Whatever the point of view used, the MCA solution can be obtained by performing the generalized SVD (Greenacre 1984) of the triplet *data, column weights, row weights* $\left(\mathbf{G}-\mathbf{M}, J^{-1}{\mathbf{D}_\mathbf{\Sigma}}^{-1/2}, n^{-1}\mathbb{I}_n\right)$ with $\mathbf{D}_\mathbf{\Sigma}$, the diagonal matrix of the column margins of the matrix and $\mathbf{M}$ the matrix where each row is equal to the vector of the means of each column of $\mathbf{G}$. It boils down to performing the following SVD: $\mathbf{G}-\mathbf{M} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top$ with $\mathbf{U}^\top(n^{-1}\mathbb{I}_n)\mathbf{U} = \mathbb{I}$ and $\mathbf{V}^\top(J^{-1}\mathbf{D}_\mathbf{\Sigma}^{-1/2})\mathbf{V} = \mathbb{I}$.

MCA can also be defined as finding the best low rank approximation of $\mathbf{G}-\mathbf{M}$ with a matrix of rank S according to the Hilbert-Schmidt norm $\parallel \mathbf{T} \parallel^2_{J^{-1}\mathbf{D}_\mathbf{\Sigma}^{-1/2},\frac{1}{n}\mathbb{I}_n} = \text{tr}\left(\mathbf{T}J^{-1}\mathbf{D}_\mathbf{\Sigma}^{-1/2}\mathbf{T}^\top\frac{1}{n}\mathbb{I}_n\right)$:

$$L_{\text{MCA}}(\mathbf{X},\mathbf{A}) \quad = \quad \|(\mathbf{G}-\mathbf{M}) - \mathbf{X}\mathbf{A}^\top\|^2_{J^{-1}\mathbf{D}_\mathbf{\Sigma}^{-1/2},\frac{1}{n}\mathbb{I}_n}$$

with $\mathbf{A}^\top = [\mathbf{A}_1^\top|\ldots|\mathbf{A}_J^\top]$ and $\mathbf{A}_j$ the $K_j\times S$ matrix representing the K_j categories of variable j. The solution is given by $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ and $\mathbf{X} = \mathbf{U}$ truncated at order S.

MCA analysis mainly consists of interpreting the graphical outputs where rows are represented with $\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{D}_\mathbf{\Sigma}^{1/2}$ and categories are represented with $\mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{D}_\mathbf{\Sigma}^{1/2}$. There are different choices regarding the graphical representations; the previous system is known as the French coordinates. In addition, we usually emphasize the (pseudo) barycentric principle which helps in interpreting simultaneous graphical displays: a column category point is, apart from scaling factors, the centroid of observations belonging to that category and a row point is, also apart from scaling factors, at the barycenter of the categories it belongs to. This property is at the origin of an additional way to introduce MCA known as "dual scaling" and popularized by Nishisato (1980). Note also that examing both rows as well as columns of a data set is already a step away from the classical inferential framework, where the rows are often a sample from a larger population and useful only in that they provide information on the relationship between variables.

2.2. Classical HOMALS presentation

HOMALS (De Leeuw and Van Rijckevorsel 1980; Gifi 1990; Michailidis and De Leeuw 1998)

is defined using the concept of "quantification". The quantification of the rows is represented with a matrix $\mathbf{X}_{n\times S}$ and the quantification of the categories is represented with the matrix $\mathbf{Y}_{S\times K} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_J)$ and the quantification of each variable is $\mathbf{G}_j\mathbf{Y}_j$ of size $n \times S$.

Homogeneity analysis is defined using a loss function which represents a criterion of departure of homogeneity:

$$L(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{J} \|\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j\|^2 \tag{2}$$

The HOMALS solution minimizes criterion (2):

$$(\mathbf{X}, \mathbf{Y}) = \arg\min_{\mathbf{X},\mathbf{Y}} L(\mathbf{X}, \mathbf{Y})$$

with the constraint that $\mathbf{x}_s$ has norm equal to 1 and is orthogonal to $\mathbf{x}'_s$ for all $s' < s$.

Contrary to MCA which is solved by SVD, homogeneity analysis uses an alternating least-squares algorithm where at step ℓ there are three substeps (at iteration $\ell = 0$, arbitrary rows scores $\mathbf{X}_0$ is used):

1. $\mathbf{Y}_j^\ell = (\mathbf{G}^\top_\mathbf{j}\mathbf{G_j})^{-\mathbf{1}}\mathbf{G}_\mathbf{j}^\top\mathbf{X}$; it corresponds to the centroid of row quantifications.
2. $\mathbf{Z}^\ell = \frac{1}{J}\sum_j \mathbf{G}_j\mathbf{Y}_j^\ell$; it corresponds to the centroid of category quantifications.
3. $\mathbf{X}^\ell$ is defined as the orthonormalized version of $\mathbf{Z}^\ell$.

The homogeneity analysis framework makes it easy to add constraints. It is common, for instance, to impose a rank constraint on the $\mathbf{Y}_j$; often rank 1 is chosen. It can be done simply by adding after step (1) in the algorithm a step where $\mathbf{Y}_c^\ell$ is defined as the best rank 1 approximation to $\mathbf{Y}^\ell$. De Leeuw and Mair (2009) highlighted the fact that such a constraint may make the interpretation easier since it leads to a more parsimonious representation. In addition, such a constraint can be a way to avoid horseshoe effects if such effects are not desirable. In addition to a rank constraint, a level constraint can be imposed to reflect the data type, i.e., ordinal or numerical variables. The idea is to respect the nature of the variables by preserving the original order of the categories, for instance. Thus, the categories of an ordinal variable will be ordered as well on the low-dimensional graphical representation.

Note also that the HOMALS frameworks allows definition of variable transformations with other restrictions on the quantification matrix $\mathbf{Y}_J$, which gives new methods such as non linear version of PCA (De Leeuw 2014).

We should mention that Jan was aware of the work of Benzécri and was influenced by Van De Geer's books on multivariate analysis from a graphical perspective. So he gave an extra perspective by including the optimization framework and this point of view was favored by the Dutch school.

2.3. Connection between HOMALS and MCA

Both HOMALS and MCA are dedicated to the analysis of categorical data and represent the data in a lower dimensional space with row coordinates $\mathbf{X}$ and category coordinates $\mathbf{Y}$. The connection between both criteria (1) and (2) is straightforward by plugging-in the centroid

of the $\mathbf{X}$ points assigned to each level of variable j, $\mathbf{Y}_j = (\mathbf{G}_j^\top\mathbf{G}_j)^{-1}\mathbf{G}_j^\top\mathbf{X}$ back into the objective function (2) for a dimension s as follows:

$$\underset{\|\mathbf{x}_s\|=1}{\arg\min}\, L(\mathbf{X},\mathbf{Y}) = \underset{\|\mathbf{x}_s\|=1}{\arg\max}\;\; \frac{1}{J}\sum_{j=1}^{J}\mathbf{X}^\top\mathbf{G}_j(\mathbf{G}_j^\top\mathbf{G}_j)^{-1}\mathbf{G}_j^\top\mathbf{X} = \underset{\|\mathbf{x}_s\|=1}{\arg\max}\;\; \frac{1}{J}\sum_{j=1}^{J}\eta^2(\mathbf{X}_s,\mathbf{v}_j)$$

Thus, MCA and HOMALS (in its simplest form without constraints) lead to exactly the same graphical representations and analysis.

However, due to the difference of starting points, we feel that the algorithms as applied in practice are more different than they initially appear. The strongest point of the Gifi methods is their use of advanced optimization techniques and Jan is a pioneer in this domain; one can mention his works on majorization algorithms known as majorization by minorization (MM) algorithms (De Leeuw and Heiser 1977), for instance. On the other hand, the extensive use of SVD has led to the developments in the matrix completion framework as illustrated in Section 2.4. Once again, this highlights the very modern aspects of these schools since both optimization techniques and the SVD have gained huge popularity in the past decade due to their ability to address problems involving high dimensional data.

In the next section, we discuss missing values. HOMALS and MCA approach missing values differently, which can be explained by the differing formulation of the methods.

2.4. Handling missing values in HOMALS and MCA

A first possibility to manage missing values consists of adding an additional column to the indicator matrix for each variable with missing data. In this case, missing values for a variable are considered as a new category and not as one of the observed categories. Then, classical HOMALS or MCA can be applied on this new complete data set. Note that this strategy makes sense for missing not at random data (MNAR) (Little and Rubin 1987, 2002), for instance, or to inspect the missing data pattern (Josse, Chavent, Liquet, and Husson 2012).

Other ways are available and they differ in HOMALS and MCA with respect to their strategy and their results.

Missing values in HOMALS

In HOMALS, missing observations are simply coded as zero rows in the matrix $\mathbf{G}$; if object i is missing on variable j, then row sum i of $\mathbf{G}_j$ is 0, otherwise row sum becomes 1 since the category entries are disjunctive. Then, whatever the coding, all row sums of $\mathbf{G}_j$ are collected in a diagonal matrix $\mathbf{M}_j$ and the criterion $L(\mathbf{X},\mathbf{Y})$ is written by introducing these matrices $(\mathbf{M}_j)_{j=1,\ldots,J}$:

$$\underset{\mathbf{X},\mathbf{Y}}{\arg\min}\, L(\mathbf{X},\mathbf{Y}) = \frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j)^\top\mathbf{M}_j(X-\mathbf{G}_j\mathbf{Y}_j).$$

We note that this approach seems a very natural way to "skip" the missing values in the optimization problem. This strategy is also known as *missing passive* (Meulman 1982) and has been extended in the framework of MCA by Escofier (1987) with *missing passive modified margin*. Van der Heijden and Escofier (2003) and Josse *et al.* (2012) discuss the advantages and drawbacks of both approaches.

Missing values in MCA

Since MCA can be presented as a particular PCA with some metrics, the approach used to handle missing values in PCA has been extended to MCA by Josse *et al.* (2012). In PCA, for a data matrix $\mathbf{Z}$, it consists of ignoring the missing values by minimizing the reconstruction error over all non-missing elements. This is done by introducing a weighted matrix $\mathbf{W}$ (with $w_{ij} = 0$ if z_{ij} is missing and $w_{ij} = 1$ otherwise) in the PCA least squares criterion:

$$\mathcal{C} = \|\mathbf{W} * (\mathbf{Z} - \mathbf{X}\mathbf{A}^\top)\|^2,$$

with $*$ the Hadamard product. This criterion can be minimized either using the alternating weighted least squares algorithm or iterative PCA (Kiers 1997; Josse, Pagès, and Husson 2009). This latter consists of randomly imputing the missing entries, performing PCA on the completed matrix and then using the principal components and loadings to impute missing values. The steps of estimation and imputation are repeated until convergence. From this iterative PCA algorithm, an algorithm called "iterative MCA" has been derived in Josse *et al.* (2012) and it takes into account the features of MCA such as updates for the column margins.

Comparison between both strategies

Both approaches aim at skipping missing values by introducing a weighted matrix in the criterion. However, both approaches lead to very different results as discussed in Josse *et al.* (2012). As mentioned in Section 2.3, the criterion and the choice of an algorithm have an impact on the properties highlighted or sometimes worse, overlooked. For instance iterative MCA can be seen as a matrix completion method which can be interesting in itself (Audigier, Husson, and Josse 2016). Of course, imputation is also possible with HOMALS, although it is less natural since it does not show up in the algorithm.

What can be noted is that the strategy *missing passive* used in HOMALS was proposed in Benzécri (1973, p.327) but it has been criticized by the French (Van der Heijden and Escofier 2003; Josse *et al.* 2012) due to the fact that many MCA properties are lost. On the contrary, in iterative MCA, the strategy to handle missing values is based on a criterion that is minimized with an iterative algorithm, which is more in the spirit of the Dutch school.

2.5. Example: Survey on the perception of genetically modified organisms

To illustrate the methods, we use an example of a survey describing genetically modified organisms (GMOs). These data are described in Husson, Josse, Le, and Mazet (2011) and are available at Husson, Josse, Lê, and Mazet (2009). The questionnaire contains 16 questions directly linked to the participants' opinion of GMOs. For instance "Do you feel implicated in the debate about GMOs (a lot, to a certain extent, a little, not at all)?"; "What is your view of GMO cultivation in France (very favourable, favourable, somewhat against, totally opposed)?" and so on. The questionnaire also contains five socio-demographic variables: sex, professional status (farmer, student, manual labourer, senior management, civil servant, accredited professional, technician, retailer, other profession, unemployed, retired), age (–25 years, 25–40 years, 40–60 years, +60 years), "Is your profession or education in any way linked to agriculture, the food industry or the pharmaceutical industry (Yes/No)?", "Which political movement do you most adhere to (extreme left, green, left, liberal, right, extreme right)?". The aim of the data analysis is first to characterise the respondents in terms of their

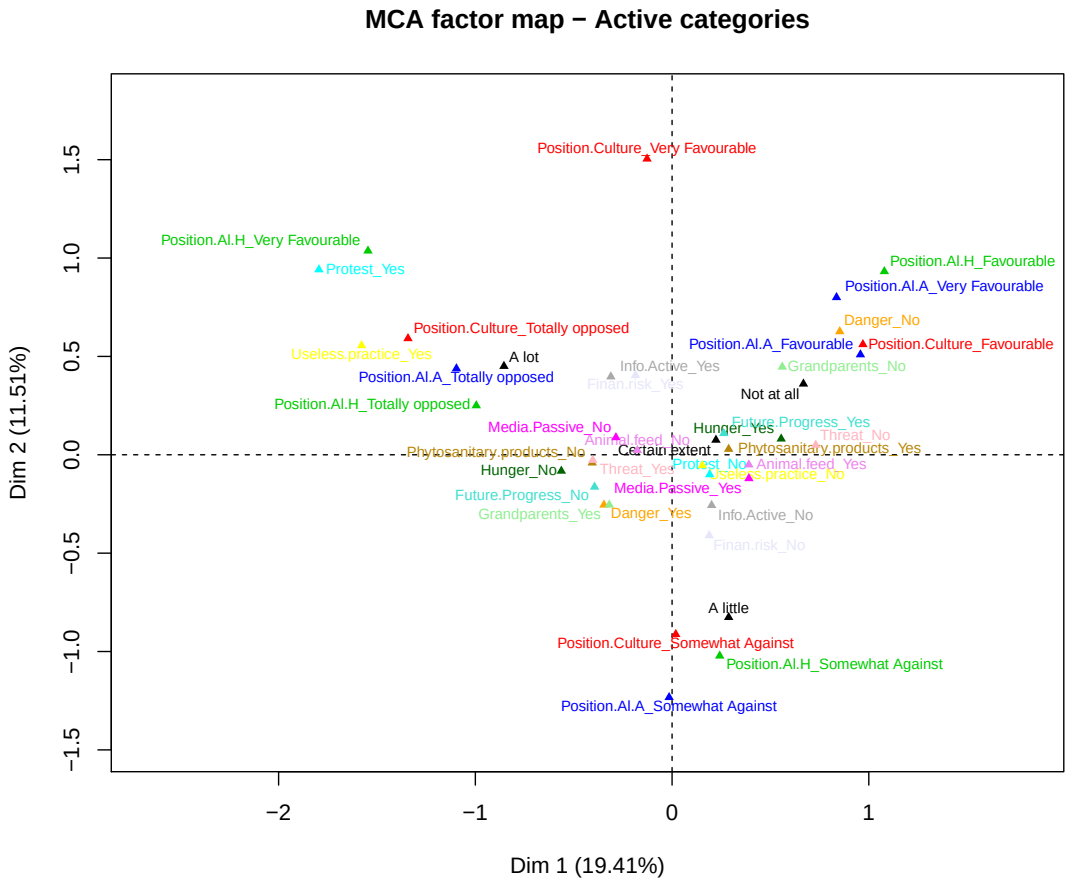


Figure 1: HOMALS and MCA representation for the active categories

relationship with GMOs; and on the other hand, to see if this characterisation has any relation with the sociodemographic variables. Consequently, these latter variables are considered as supplementary variables. It means that they are not used to build the scores, but only projected afterward onto the graphical representations that have been obtained. They are used to enhance the interpretation. Note that introducing supplementary variables is possible in both approaches, even though, historically, such variables are most used in the French school due to the prominence of the graphical outputs. In the Dutch school, supplementary points are refered to as passive variables. HOMALS and MCA are implemented in the R packages **homals** (De Leeuw and Mair 2009) and **FactoMineR** (Lê, Josse, and Husson 2008).

Figure 1 gives the graph obtained by HOMALS and MCA for the active categories (the two graphs are the same). On the negative side, represented by the first principal component, we can observe those people who feel implicated by the debate surrounding GMOs and who are somewhat against their use (through the categories they chose). On the positive side, we can see those people who do not feel implicated by the debate surrounding GMOs and who are in favour of their use. Along the second principal component, we can also observe those people with less distinct opinions who feel somewhat implicated by the debate surrounding GMOs and who are somewhat against their use. More interpretation is given in Husson *et al.* (2010).

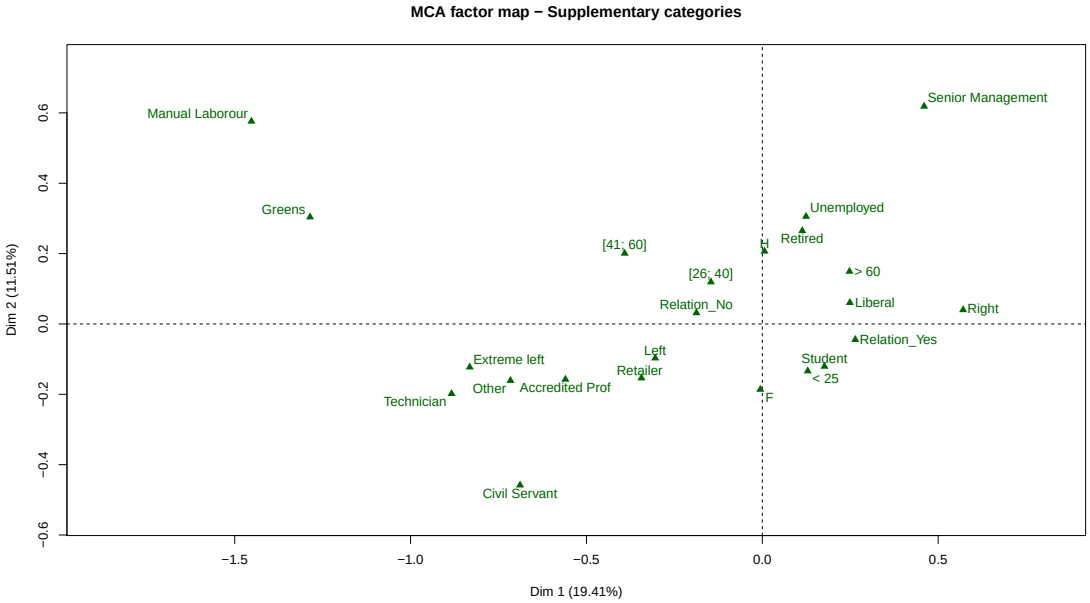


Figure 2: Supplementary categories representation.

The representation of supplementary variables in Figure 2 reveals a strong structure for both of the variables *profession* and *identification with a political movement*, and second, it fails to identify any particular structure with the variables of age, sex, or profession in relation to agriculture, the food industry, and the pharmaceutical industry. The categories *senior management*, *unemployed*, and *retired* are in opposition to the categories *technician* and *manual labourer* to *civil servant* between the two groups. Similarly, the category *right* is opposed to the categories *green* and *extreme left*, to, in the middle, *left*.

Since in the questionnaire, some variables are naturally ordered, we can use HOMALS with the rank 1 constraint (Figure 3) and with the constraint that the categories are ordered (Figure 4). As expected, the categories of the ordinal variables are on a straight line and the order is preserved.

Note also that some attempts have been made in MCA to add constraints; see Benzécri (1973, p. 261–287) and Beh and Lombardo (2014, Chapter 6). However, the inclusion of constraints is less straightforward than in HOMALS and they are consequently not as used by the French school. The lack of use of constraints can also be explained by the fact that they have never been implemented. Software is an incredibly powerful tool to popularize methods, and the implementation and availability of methods in software may explain why some practices (even when flawed) are still in use.

3. Influence of Jan de Leeuw's work on French works

3.1. Influence on the early works

Jan de Leeuw was well aware of French works in data analysis and especially in correspondence analysis. In his 1973 thesis (De Leeuw 1973), there are several references to Jean-Paul

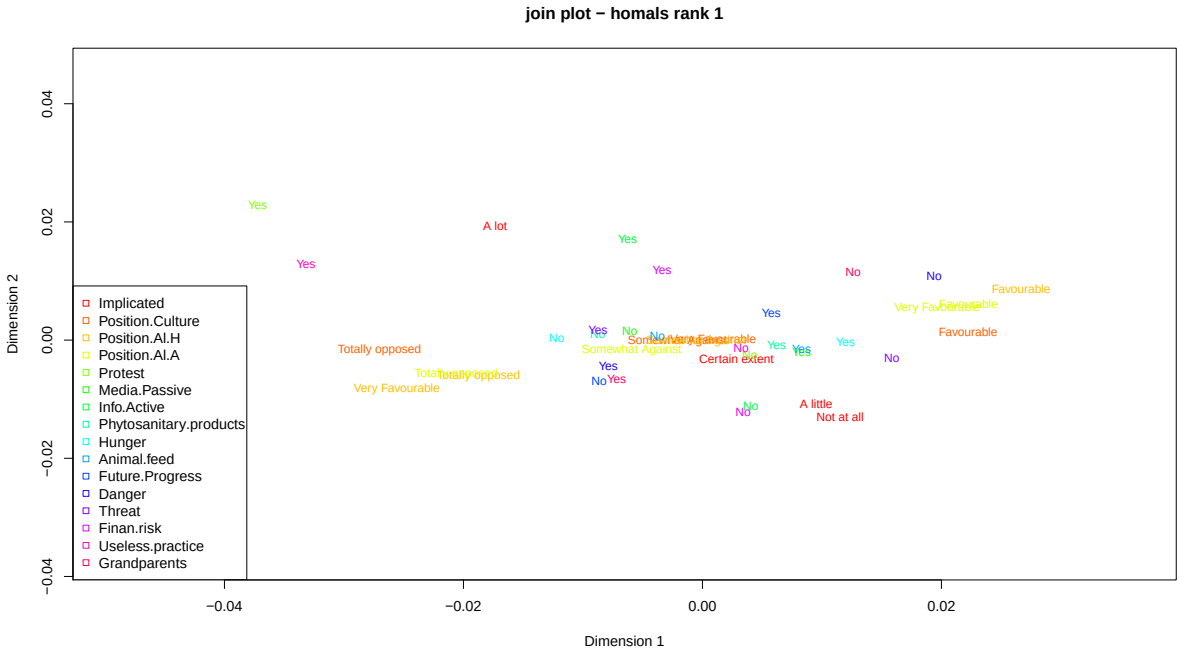


Figure 3: GMO data: HOMALS with the constraint of rank 1.

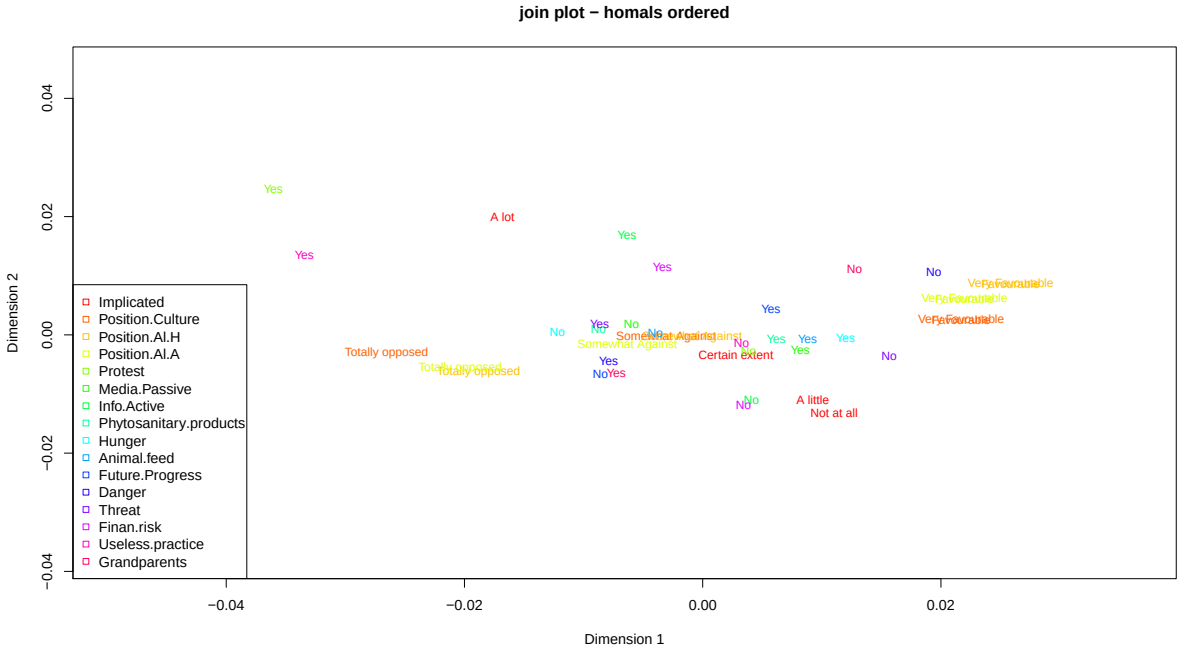


Figure 4: GMO data: HOMALS with the constraint of rank 1 and ordered categories.

Benzécri, Henri Caussinus, and above all to the dissertation of Brigitte Escofier-Cordier entitled “l’analyse des correspondances” (Cordier 1965). In turn the influence of Jan de Leeuw was felt as early as the late 70s in two areas. The main media were the Revue de Statistique Appliquée (Revue de Statistique Appliquée) and Statistique et Analyse des Données (Statistique et Analyse des Données).

The two areas in which the influence of De Leeuw was felt were:

- Optimal scaling where categorical variables, either ordinal or nominal are optimally transformed into discrete numerical variables enabling the use of methods like regression, PCA and discriminant analysis. De Leeuw's dissertation was rapidly known in France and was referred to in Bouroche, Saporta, and Tenenhaus (1975), Saporta (1975), Tenenhaus (1977). The series of papers by Young, Takane, De Leeuw in Psychometrika were very influential: see e.g., Dupont-Gatelmand (1979). Nonlinear PCA, which is intimately connected to optimal scaling (De Leeuw 1988; Gifi 1990), inspired many French works more or less directly till the end of the 90s; see for instance Ferraty (1997).
- Analysis of ordinal data. This area includes multidimensional scaling, preferences analysis and multicriteria choice. The preprint of Takane, Young, and De Leeuw (1977) inspired Drouet d'Aubigny (1976) and independently Lemaire (1977) in multivariate data analysis. Jacquet-Lagreze (1977), Siskos (1980), Jacquet-Lagreze and Siskos (1982) developed the multicriteria decision method. Recently one may notice Bennani-Dosse (1995) who gave a generalization of SMACOF, using the majorization algorithm, and Husson and Pagès (2006) with contributions to multidimensional scaling.

HOMALS, often presented in France as a Dutch version of multiple correspondence analysis (Desbois 2008), has not been really used in the social sciences, despite having some interesting features (see the previous sections). This is certainly due to the wide diffusion of software packages made in France like SPAD (Coheris 2016), XLSTAT (Addinsoft 2016), **FactoMineR** and to the leading position of SAS (with `proc corresp`, SAS Institute Inc. 2011) over IBM SPSS (IBM Corporation 2016). Jan de Leeuw was invited several times to France, especially by statisticians from Toulouse in the 80s, and to the conferences "Data Analysis and Informatics" organised by Edwin Diday. Despite the undeniable influence of Jan de Leeuw on the work of French statisticians, one can find among the numerous collaborations of Jan de Leeuw only a single publication with a French co-author: Antoine de Falguerolles (Van Der Heijden, De Falguerolles, and De Leeuw 1989).

3.2. Contribution of optimization methods in recent work

Jan de Leeuw has been a forerunner in developing and using optimisation techniques (see for instance De Leeuw 2016). The block-relaxation algorithms (De Leeuw 1994) or, in more modern words, block coordinate descent, together with majorization by minorization (De Leeuw and Heiser 1977), are used for instance in the regularized generalized canonical correlation analysis (RGCCA) method (Tenenhaus and Tenenhaus 2011) for multi-block data analysis, which concerns the analysis of several sets of variables (blocks) observed on the same group of individuals. The main aims of RGCCA are: (i) to study the relationships between blocks and (ii) to identify subsets of variables of each block which are active in their relationships with the other blocks. RGCCA is based on a monotonically convergent iterative algorithm and has the distinct advantage of being formulated as an explicit optimization problem.

References

Addinsoft (2016). *XLSTAT Software, Version 2016.5*. URL https://www.xlstat.com/.

Audigier V, Husson F, Josse J (2016). "MIMCA: Multiple Imputation for Categorical Variables with Multiple Correspondence Analysis." *Statistics and Computing*, pp. 1–18. doi:10.1007/s11222-016-9635-4.

Beh EJ, Lombardo R (2014). *Correspondence Analysis. Theory, Practice and New Strategies*. John Wiley & Sons.

Bennani-Dosse M (1995). "Positionnement Multidimensionnel D'un Tableau à 3 Voies." *Revue de Statistique Appliquée*, **43**, 63–75.

Benzécri J (1973). *L'Analyse des Données. Tome II: L'Analyse des Correspondances*. Dunod.

Benzécri JP (1982). *Histoire et Préhistoire de L'Analyse des Données*. Dunod.

Benzécri JP (1986). "Statistical Analysis as a Tool to Emerge Patterns from the Data." In S Watanabe (ed.), *Methodologies of Pattern Recognition*, pp. 34–74. Academic Press, New York.

Bouroche JM, Saporta G, Tenenhaus M (1975). "Generalized Canonical Analysis of Qualitative Data." In *US Japan Seminar on Theory, Methods and Applications of Multidimensional Scaling and Related Techniques*. MIT Press, San Diego. URL http://cedric.cnam.fr/%7Esaporta/SanDiego1975.pdf.

Bouroche JM, Saporta G, Tenenhaus M (1977). "Some Methods of Qualitative Data Analysis." In JR Barra, F Brodeau, G Romier, B Van Cutsem (eds.), *Recent Developments in Statistics*, pp. 749–755. North Holland Publishing Company, Amsterdam.

Coheris (2016). *Coheris Analytics SPAD Software, Version 8.2*. URL http://www.coheris.com/.

Cordier B (1965). *L'Analyse des Correspondances*. Ph.D. thesis, University of Rennes.

De Leeuw J (1973). *Canonical Analysis of Categorical Data*. Ph.D. thesis, University of Leiden, The Netherlands.

De Leeuw J (1977). "Applications of Convex Analysis to Multidimensional Scaling." In JR Barra, F Brodeau, G Romier, Van Cutsem (eds.), *Recent Developments in Statistics*, pp. 133–145. North Holland Publishing Company, Amsterdam.

De Leeuw J (1988). "Multivariate Analysis with Optimal Scaling." In *Proceedings of the International Conference on Advances in Multivariate Statistical Analysis*, pp. 127–160. Indian Statistical Institute, Calcutta.

De Leeuw J (1994). "Block-Relaxation Algorithms in Statistics." In *Information Systems and Data Analysis*, pp. 308–324. Springer-Verlag.

De Leeuw J (2005). "Models of Data." *Technical report*, University of California, Los Angeles. URL http://escholarship.org/uc/item/74t5g61m.

De Leeuw J (2011a). "Interview Given on the Occasion of the International Conference on Correspondence Analysis and Related Methods (CARME)." Conference website: http://carme2011.agrocampus-ouest.fr/, URL http://gifi.stat.ucla.edu/CARME/.

De Leeuw J (2011b). "Statistics and the Sciences." *Technical report*, University of California, Los Angeles. URL http://escholarship.org/uc/item/46b4s8m3.

De Leeuw J (2014). "Nonlinear Principal Component Analysis and Related Techniques." In J Blasius, M Greenacre (eds.), *Visualization and Verbalization of Data*, pp. 107–134. CRC Press, Taylor & Francis.

De Leeuw J (2016). "Block Relaxation Algorithms in Statistics." URL https://www.gitbook.com/@jandeleeuw.

De Leeuw J, Heiser WJ (1977). "Convergence of Correction Matrix Algorithms for Multidimensional Scaling." In JC Lingoes, E Roskam, I Borg (eds.), *Geometric Representations of Relational Data*, pp. 735–752. Mathesis Press.

De Leeuw J, Mair P (2009). "Gifi Methods for Optimal Scaling in R: The Package **homals**." *Journal of Statistical Software*, **31**(4), 1–20. doi:10.18637/jss.v031.i04.

De Leeuw J, Van Rijckevorsel J (1980). "HOMALS and PRINCALS – Some Generalizations of Principal Components Analysis." In E Diday (ed.), *Data Analysis and Informatics*, pp. 231–242. Springer-Verlag.

Desbois D (2008). "L'Analyse des Correspondances Multiples "a la Hollandaise": Introduction à l'Analyse d'Homogénéité." *Revue Modulad*, **38**, 194–244.

Donoho D (2015). "50 Years of Data Science." URL http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

Drouet d'Aubigny G (1976). "L'Utilisation des Méthodes Ordinales en Analyse des Données." *Statistique et Analyse des Données*, **1**, 63–84.

Dupont-Gatelmand C (1979). "Une Méthode de Classification Automatique sur Variables Hétérogènes." *Revue de Statistique Appliquée*, **27**, 23–37.

Escofier B (1987). "Traitement des Questionnaires avec Non Réponse, Analyse des Correspondances avec Marges Modifiée et Analyse Multicanonique avec Contraintes." *Publications de l'institut de statistique de l'université de Paris*, **32**, 33–70.

Ferraty F (1997). "Estimations de Transformations Optimales en ACP Curvilinéaire." *Revue de Statistique Appliquée*, **45**, 5–39.

Gifi A (1990). *Nonlinear Multivariate Analysis*. John Wiley & Sons, Chichester.

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

Greenacre M, Blasius J (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC. doi:10.1201/9781420011319.

Holmes S (2008). "Multivariate Data Analysis: The French Way." *Institute of Mathematical Statistics Collections*, pp. 219–233. doi:10.1214/193940307000000455.

Husson F, Josse J (2014). "Multiple Correspondence Analysis." In J Blasius, M Greenacre (eds.), *Visualization and Verbalization of Data*, pp. 165–184. Chapman & Hall/CRC.

Husson F, Josse J, Lê S, Mazet J (2009). "GMO Survey Data." URL http://factominer.free.fr/book/gmo.csv.

Husson F, Josse J, Le S, Mazet J (2011). ***FactoMineR**: Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.16, URL https://CRAN.R-project.org/package=FactoMineR.

Husson F, Lê S, Pagès J (2010). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC. doi:10.1201/b10345.

Husson F, Pagès J (2006). "Aspects Méthodologiques du Modèle INDSCAL." *Revue de Statistique Appliquée*, **54**, 83–100.

IBM Corporation (2016). *IBM SPSS Statistics 24*. Armonk. URL http://www.ibm.com/analytics/us/en/technology/spss/.

Jacquet-Lagreze E (1977). "Méthodes Explicatives en Analyses de Préférences Ordinales." *Statistique et Analyse des Données*, **2**, 45–58.

Jacquet-Lagreze E, Siskos J (1982). "Assessing a Set of Additive Utility Functions for Multicriteria Decision-Making, the UTA Method." *European Journal of Operational Research*, **10**, 151–164. doi:10.1016/0377-2217(82)90155-2.

Josse J, Chavent M, Liquet B, Husson F (2012). "Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis." *Journal of Classification*, **29**(1), 91–116. doi:10.1007/s00357-012-9097-0.

Josse J, Pagès J, Husson F (2009). "Gestion des Données Manquantes en Analyse en Composantes Principales." *Journal de la Société Française de Statistique*, **150**(2), 28–51.

Kiers HAL (1997). "Weighted Least Squares Fitting Using Ordinary Least Squares Algorithms." *Psychometrika*, **62**(2), 251–266. doi:10.1007/bf02295279.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18. doi:10.18637/jss.v025.i01.

Le Roux B, Rouanet H (2004). *Geometric Data Analysis: From Correspondence Analysis to Stuctured Data Analysis*. Kluwer Academic Publishers.

Lebaron F, Le Roux B (2015). *La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données*. Dunod.

Lebart L (2008). "Exploratory Multivariate Data Analysis from Its Origins to 1980: Nine Contributions." *Electronic Journal for History of Probability and Statistics*, **4**(2). URL http://www.jehps.net/decembre2008.html.

Lebart L, Morineau A, Warwick KM (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, New York.

Lebart L, Saporta G (2014). "Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis." In J Blasius, M Greenacre (eds.), *Visualization and Verbalization of Data*, pp. 31–44. Chapman & Hall/CRC.

Lebart L, Tabard N (1973). "Recherches sur la Description Automatique des Données Socio-Economiques." In *CORDES-CREDOC, Research Convention N○13/1971*.

Lemaire J (1977). "Agrégation Typologique de Données de Préférences." *Mathématiques et Sciences Humaines*, **58**, 31–50.

Little RJA, Rubin DB (1987, 2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Meulman J (1982). *Homogeneity Analysis of Incomplete Data*. DSWO Press, Leiden.

Michailidis G, De Leeuw J (1998). "The Gifi System of Descriptive Multivariate Analysis." *Statistical Science*, **13**, 307–336. doi:10.1214/ss/1028905828.

Murtagh F (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall/CRC. doi:10.1201/9781420034943.

Nishisato S (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press, Toronto.

Psychometrika (1976). "Program of the 1976 Annual Spring Meeting." *Psychometrika*, **41**(3), 421–424. doi:10.1007/bf02293567.

Revue de Statistique Appliquée (2006). Published 1953-2006, URL http://www.numdam.org/numdam-bin/browse?j=RSA&sl=0.

Saporta G (1975). *Liaisons entre Plusieurs Ensembles de Variables et Codage de Données Qualitatives*. Ph.D. thesis, Université Pierre et Marie Curie. URL https://tel.archives-ouvertes.fr/tel-00649127.

Saporta G (1988a). "About Maximal Association Criteria in Linear Analysis and in Cluster Analysis." In HH Bock (ed.), *Classification and Related Methods of Data Analysis*, pp. 541–550. Springer-Verlag.

Saporta G (1988b). "Correspondence Analysis for Categorical Stochastic Processes." In *Proceedings of the International Conference on Advances in Multivariate Statistical Analysis*, pp. 365–376. Indian Statistical Institute, Calcutta.

SAS Institute Inc (2011). *SAS/STAT Software, Version 9.3*. Cary. URL http://www.sas.com/.

Siskos J (1980). "Comment Modéliser Les Préférences Au Moyen de Fonctions d'Utilité Additives." *Revue Française d'Automatique, d'Informatique et de Recherche Opérationnelle*, **14**, 53–82.

Statistique et Analyse des Données (1991). Published 1976-1991, URL http://www.numdam.org/numdam-bin/browse?j=SAD&sl=0.

Takane Y, Young FW, De Leeuw J (1977). "Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika*, **42**, 7–67. doi:10.1007/bf02293745.

Tenenhaus A, Tenenhaus M (2011). "Regularized Generalized Canonical Correlation Analysis." *Psychometrika*, **76**, 257–284. doi:10.1007/s11336-011-9206-8.

Tenenhaus M (1977). "Analyse en Composantes Principales d'un Ensemble de Variables Nominales ou Numériques." *Revue de Statistique Appliquée*, **25**, 39–56.

Tukey JW (1962). "The Future of Data Analysis." *Annals of Mathematical Statistics*, **33**, 1–67. doi:10.1214/aoms/1177704711.

Van Der Heijden PGM, De Falguerolles A, De Leeuw J (1989). "A Combined Approach to Contingency Table Analysis with Correspondence Analysis and Log-Linear Analysis." *Applied Statistics*, **38**, 249–292. doi:10.2307/2348058.

Van der Heijden PGM, Escofier B (2003). "Multiple Correspondence Analysis with Missing Data." In *Analyse des Correspondances*, pp. 153–170. Presses Universitaires de Rennes.

Wasserstein RL, Lazar NA (2016). "The ASA's Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician*, **70**, 129–133. doi:10.1080/00031305.2016.1154108.

Appendix[4]

My first, half-missed, encounter with Jan was in April 1976 on the occasion of a symposium on Optimal Scaling during the spring meeting of the Psychometric Society. It was my first trip to the US and, suffering from jet-lag, I collapsed early in my bed. That's when I got a phone call from Jan offering to get acquainted. I stammered a few words and then I fell asleep again. We finally met the next day.

There is an unfortunate typo in the title of my talk: it was about "nominal", and not about "normal", variables.

The European Meeting of Statisticians, organised from 6 to 11 September 1976 in Grenoble (France) under the auspices of the European Regional Committee of the Bernoulli Society, gave us the opportunity to form a better relationship. I especially remember a lunch organised by Gérard Drouet d'Aubigny in Sassenage, a village nearby Grenoble: Jan van Rijckevorsel, Jean-Marie Bouroche, Michel Tenenhaus, and a few others were there. On the menu there was a very French, and not very vegetarian special sausage: the "andouillette", but I do not remember if Jan de Leeuw tasted it! In the proceedings of the Grenoble meeting, one can find a paper by Jan (De Leeuw 1977) as well as the one by the French trio (Bouroche, Saporta, and Tenenhaus 1977), which referred to Jan's communication at the Spring Meeting of the Psychometric Society a few months before.

We had many opportunities to see each other afterwards in various meetings as well as at Ph.D. defenses in the Netherlands. I have a particular remembrance of the International Conference on Advances in Multivariate Statistical Analysis held in Calcutta in December 1985 at the Indian Statistical Institute in P.C. Mahalanobis' domain: Jan told me one day

[4]Written by Gilbert Saporta.

PROGRAM OF THE 1976 ANNUAL SPRING MEETING

The annual spring meeting of the Psychometric Society and the Mathematical Psychology Group was held on April 1–3, 1976 at Bell Laboratories in Murray Hill, New Jersey. R. Duncan Luce was program chairman, while J. Douglas Carroll, Joseph B. Kruskal, and Myron Wish were in charge of local arrangements. The program follows. * signifies presenter.

SYMPOSIUM: OPTIMAL SCALING, chaired by Forrest W. Young, University of North Carolina

Forrest W. Young; University of North Carolina *Introduction and history*

Shizuhiko Nishisato; Ontario Institute for Studies in Education *Optimal scaling as applied to different types of models*

Forrest W. Young; University of North Carolina *Optimal scaling with different types of models*

Jan de Leeuw; Datatheorie Instituut, Leiden *Optimal scaling and canonical analysis*

G. Saporta; University of Paris *Discriminant analysis when all the variables are normal: A stepwise method*

Discussant: Joseph B. Kruskal; Bell Laboratories

Figure 5: Program of the annual spring meeting of the Psychometric Society (Psychometrika 1976).



Figure 6: Lunch time, Indian Statistical Institute, Calcutta, December 1985. C. R. Rao sitting at left.

that it was his 40th birthday. Jan's paper De Leeuw (1988) cites J. P. Benzécri three times, and the book by Lebart, Morineau, and Warwick (1984), while my paper Saporta (1988b) cites Jan's works three times.

Affiliation:

François Husson
Department of Statistics
Agrocampus Ouest Rennes
35042 Rennes, France
E-mail: husson@agrocampus-ouest.fr
URL: http://www.agrocampus-ouest.fr/math/husson/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
September 2016, Volume 73, Issue 6 *Submitted:* 2016-03-24
doi:10.18637/jss.v073.i06 *Accepted:* 2016-06-01