Correspondence Analysis on Generalised Aggregated Lexical Tables (CA-GALT) in the FactoMineR Package

by Belchin Kostov, Mónica Bécue-Bertaut and François Husson

Abstract Correspondence Analysis on Generalised Aggregated Lexical Tables (CA-GALT) is a method that generalizes classical CA-ALT to the case of several quantitative, categorical and mixed variables. It aims to establish a typology of the external variables and a typology of the events from their mutual relationships. In order to do so, the influence of external variables on the lexical choices is untangled cancelling the associations among them and to avoid the instability issued from multicollinearity, they are substituted by their principal components.The `CaGalt` function, implemented in the **FactoMineR** package, provides numerous numerical and graphical outputs. Confidence ellipses are also provided to validate and improve the representation of words and variables. Although this methodology was developed mainly to give an answer to the problem of analyzing open-ended questions, it can be applied to any kind of frequency/contingency table with external variables.

1. Introduction

Frequency tables are a common data structure in very different domains such as ecology (specie abundance table), textual analysis (documents $\times$ words table) and public information systems (administrative register such as mortality data). This type of table counts the occurrences of a series of events (species, words, death causes) observed on different units (ecological sites, documents, administrative area). Correspondence analysis (CA) is a reference method to analyse this type of tables offering the visualization of the similarities between events, the similarities between units and the associations between events and units (Benzécri, 1973; Lebart et al., 1998; Murtagh, 2005; Greenacre, 2007; Beh and Lombardo, 2014). However, this method presents two main drawbacks, when the frequency table is very sparse:

1. The first axes frequently show the relationships between small sets of units and small sets of events and do not reveal global trends.
2. The interpretation of the similarities/oppositions among units cannot be understood without taking into account the unit characteristics (such as, for example, climatic conditions, socio-economic description of the respondents or economic characteristics of the area).

In order to solve these drawbacks, contextual variables are also observed on the units and introduced in the analysis. A first manner consists in grouping the units depending on one categorical variable and building an aggregated frequency table (AFT) crossing the categories (rows) and the events (columns). In this AFT, the former row-units corresponding to a same category are now collapsed into a single row while the event-columns remain unchanged. Then, CA is applied on this AFT, often called, in textual analysis, aggregated lexical table (ALT; Lebart et al., 1998).

CA on the aggregated lexical table (CA-ALT) usually leads to robust and interpretable results. CA-ALT visualizes the similarities among categories, the similarities among words and the associations between categories and words. A same approach can be applied in other domains. The main drawback of CA-ALT is its restrictiveness. Only one categorical variable can be considered while often several categorical and quantitative contextual variables are available and associated to the events.

Recently, Correspondence Analysis on Generalised Aggregated Lexical Tables (CA-GALT; Bécue-Bertaut and Pagès, 2015; Bécue-Bertaut et al., 2014) has been proposed to generalize CA-ALT to the case of several quantitative, categorical and mixed variables. CA-GALT brings out the relationships between the vocabulary and the several selected contextual variables.

This article presents an R function implementing CA-GALT into **FactoMineR** package (Lê et al., 2008; Husson et al., 2010) with the following outline. Section 2 describes the example used to illustrate the method and Section 3 introduces the notation. Section 4 recalls the principles of CA-GALT methodology. Section 5 details the function and the algorithm. The results obtained on the example are provided in Section 6. We conclude in Section 7 with some remarks.

2. Example

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 documents × 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, 115 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.

3. Notation

The data is coded into two matrices (see Figure 1). The $(I \times J)$ matrix $\mathbf{Y}$, with generic term y_{ij}, contains the frequency of the J words in the I respondents' answers. The $(I \times K)$ matrix $\mathbf{X}$, with generic term x_{ik}, stores the K respondents' characteristics, codified as dummy variables from the L categorical variables.

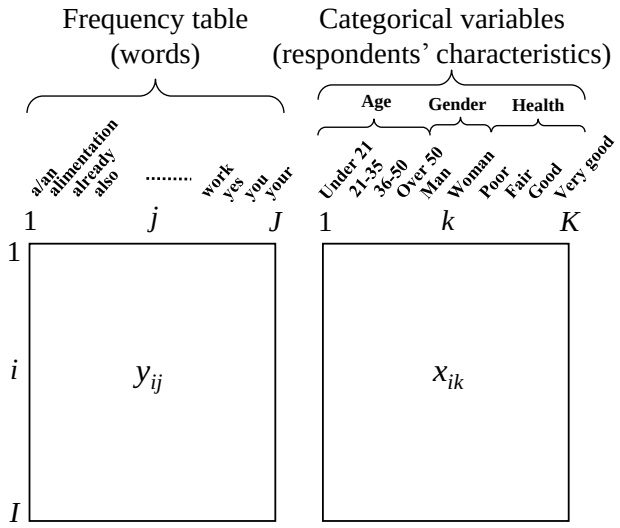

Figure 1: The data set. On the left, the frequency table $\mathbf{Y}$; on the right the categorical table $\mathbf{X}$. In the example, $I = 392$ (respondents), $J = 115$ (words), $K = 10$ (categories).

The proportion matrix $\mathbf{P}$ is computed as $\mathbf{P}=\mathbf{Y}/N$ with generic term $p_{ij} = y_{ij}/N$. The row margin (respectively, column margin) with generic term $p_{i\bullet} = \sum_j p_{ij}$ (respectively, $p_{\bullet j} = \sum_i p_{ij}$) is stored in the $(I \times I)$ diagonal matrix $\mathbf{D_I}$ (respectively, the $(J \times J)$ diagonal matrix $\mathbf{D_J}$). From $\mathbf{P}$, the $(I \times J)$ matrix $\mathbf{Q}$ is defined as $\mathbf{Q}=\mathbf{D_I^{-1}PD_J^{-1}}$ with generic term $q_{ij} = p_{ij}/(p_{i\bullet}p_{\bullet j})$. This matrix, or equivalently, its doubly centred form, the $(I \times J)$ matrix $\bar{\mathbf{Q}}$ with generic term $\bar{q}_{ij} = (p_{ij} - p_{i\bullet}p_{\bullet j})/(p_{i\bullet}p_{\bullet j})$ is the matrix analysed by CA. $\bar{\mathbf{Q}}$ evidences that CA analyses the weighted deviation between $\mathbf{P}$ and the $(I \times J)$ independence model matrix $[p_{i\bullet}p_{\bullet j}]$.

The principal component analysis (PCA) applied to a data matrix $\mathbf{Z}$ with metric **M/D** and weighting system **D/M** in the row/column space is noted PCA($\mathbf{Z}$,**M**,**D**).

4. Methodology

CA-GALT method is detailed in Bécue-Bertaut and Pagès (2015) for quantitative contextual variables and Bécue-Bertaut et al. (2014) for categorical variables. We recall here the principles of this method. To ease the presentation, we recall first that classical CA is a double projected analysis. Then, we show that CA-GALT maintains a similar approach when several quantitative or categorical variables are taken into account.

4.1. Classical CA as a particular PCA

It is established that classical CA($\mathbf{Y}$) is equivalent to PCA($\mathbf{Q}$,$\mathbf{D_J}$,$\mathbf{D_I}$) or to PCA($\bar{\mathbf{Q}}$,$\mathbf{D_J}$,$\mathbf{D_I}$) (Escofier and Pagès, 1988; Böckenholt and Takane, 1994; Bécue-Bertaut and Pagès, 2004). This point of view presents the advantage of placing CA rationale in the general scheme for the principal components methods.

4.2. CA of an aggregated lexical table

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

$$\mathbf{Y_A} = \mathbf{Y^T X} \tag{1}$$

is built crossing the words and the categories of the categorical variable. Then the $(J \times K)$ proportion matrix is computed as

$$\mathbf{P_A} = \mathbf{P^T X}. \tag{2}$$

The $(J \times J)$ diagonal matrix

$$\mathbf{D_J} = [d_{Jjj}] = [p_{Aj\bullet}] = [p_{\bullet j}] \tag{3}$$

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

$$\mathbf{D_K} = [d_{Kkk}] = [p_{A\bullet k}] \tag{4}$$

store, respectively, the row and colum margins of $\mathbf{P_A}$. $\mathbf{D_J}$ (respectively, $\mathbf{D_K}$) corresponds to weighting system on the rows (respectively, on the columns). As a single categorical variable is considered

$$\mathbf{D_K} = \mathbf{X^T D_I X}. \tag{5}$$

CA-ALT, that is CA($\mathbf{Y_A}$), is performed through PCA($\mathbf{Q_A}$, $\mathbf{D_K}$, $\mathbf{D_J}$) where the $(J \times K)$ matrix

$$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A D_K^{-1}} \tag{6}$$

is the double standardized form of $\mathbf{P_A}$.

PCA($\mathbf{Q_A}$, $\mathbf{D_K}$, $\mathbf{D_J}$) analyses both the dispersion of the cloud of category profiles insofar as explained by the dispersion of the cloud of word profiles and the dispersion of the word 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 (Bécue-Bertaut et al., 2014).

4.3. Correspondence analysis on generalised aggregated lexical tables (CA-GALT)

In this section, $\mathbf{X}$ includes the centred dummy columns corresponding to several categorical variables. The aggregated lexical tables built from each categorical variable are juxtaposed row-wise into the generalised aggregated lexical table (GALT) $\mathbf{Y_A}$ of dimensions $J \times K$ (see Figure 2).

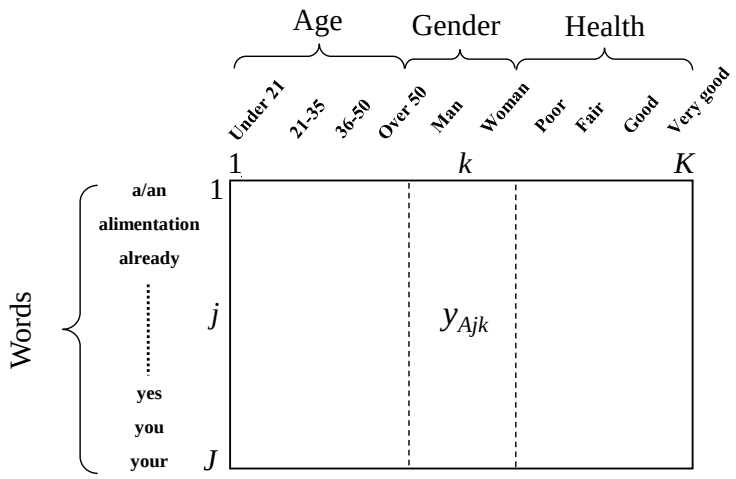

Figure 2: Generalised aggregated lexical table with the words in rows and the categories of age, gender and health condition in columns

In the following, we use the same notation that this used in the former section to highlight that the same rationale is followed regardless the differences existing between $\mathbf{X}$ structure in both cases. As in the former section

$$\mathbf{Y_A} = \mathbf{Y^T X}. \tag{7}$$

$\mathbf{Y_A}$ is transformed into the matrix

$$\mathbf{P_A} = \mathbf{Y_A}/N. \tag{8}$$

To maintain a double projected analysis, $\mathbf{D_K}$ is substituted by the $(K \times K)$ covariance matrix $\mathbf{C} = (\mathbf{X^T D_I X})$. As $\mathbf{C}$ is not invertible, the Moore-Penrose pseudoinverse $\mathbf{C}^-$ is used and the former Eq.(6) becomes

$$\mathbf{Q_A} = \mathbf{D_J^{-1} P_A C^-} \tag{9}$$

of dimensions $J \times K$. CA-GALT is performed through PCA($\mathbf{Q_A, C, D_J}$). As in any PCA, the eigenvalues are stored into the $(S \times S)$ diagonal matrix $\mathbf{\Lambda}$, and the eigenvectors into the $(K \times S)$ matrix $\mathbf{U}$. The coordinates of the row-words are computed as

$$\mathbf{F} = \mathbf{Q_A C U} \tag{10}$$

and the column-categories as

$$\mathbf{G} = \mathbf{Q_A^T D_J F \Lambda^{-1/2}}. \tag{11}$$

The respondents can be reintroduced in the analysis by positioning the columns of $\mathbf{Q}$ as supplementary columns in PCA($\mathbf{Q_A, C, D_J}$). So the respondents are placed on the axes at the weighted centroid, up to a constant, of the words that they use. Thus, their coordinates are computed via the transition relationships as

$$\mathbf{G^+} = \mathbf{D_I^{-1} P F \Lambda^{-1/2}}. \tag{12}$$

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

4.4. Quantitative contextual variables

The case of the quantitative variables is detailed in Bécue-Bertaut and Pagès (2015). 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.

4.5. Other aspects

- **CA-GALT on principal components**: To obtain less time consuming computation, the original variables are substituted by their principal components, computed either by PCA if $\mathbf{X}$ stores quantitative variables or multiple correspondence analysis if $\mathbf{X}$ stores categorical variables. The components associated to the low eigenvalues can be discarded to solve the instability problem issued from multicolinearity.
- **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.

5. Using the CaGalt function in R

The R function `CaGalt` is currently part of the **FactoMineR** package. 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))
```

with 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 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, corresponding print, plot and summary methods were also implemented in **FactoMineR**.

6. Application of the CaGalt function

To illustrate the outputs and graphs of `CaGalt`, we use the data set presented in section 2. 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 (whose type is defined as "n"). The code to perform the `CaGalt` is

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

6.1. 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 Tables (CaGalt)**
*The results are available in the following entries:
   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` entries.

```
> 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.             30.207  19.024  13.776  12.953  10.847   6.819   6.374
Cumulative % of var.  30.207  49.230  63.007  75.960  86.807  93.626 100.000

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      | -0.508  6.062  0.941 | -0.036  0.050  0.005 | -0.045  0.102  0.007 |
to have         |  0.241  5.654  0.727 |  0.054  0.451  0.037 | -0.027  0.157  0.009 |
well            | -0.124  0.790  0.168 | -0.255  5.254  0.703 | -0.073  0.593  0.057 |
to feel         | -0.217  1.174  0.222 | -0.321  4.059  0.484 | -0.158  1.353  0.117 |
hungry          |  0.548  1.504  0.254 | -0.630  3.158  0.336 | -0.367  1.478  0.114 |
I               |  0.360  2.927  0.537 | -0.213  1.623  0.188 | -0.114  0.639  0.053 |
one             |  0.246  0.964  0.241 |  0.369  3.449  0.542 |  0.120  0.500  0.057 |
something       | -0.826  2.959  0.431 |  0.444  1.353  0.124 | -0.734  5.121  0.340 |
best            |  0.669  4.182  0.602 |  0.091  0.123  0.011 |  0.225  1.041  0.068 |
psychologically | -0.369  0.560  0.155 | -0.727  3.444  0.601 |  0.190  0.323  0.041 |

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 |
```

Regarding the interpretation of these results, the percentage of variance explained by each dimension is given: 30.21% for the first axis and 19.02% for the second one. The third and fourth dimensions also explain an important part of the total variability (13.78% and 12.95%, respectively) and it may be interesting to plot the graph for these two dimensions. 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. For more information please check the help file of this function.

6.2. Graphical outputs

By default, the `CaGalt` function gives three graphs: one for the individuals, one for the variables and one for the frequencies. The variables graph shows the map of the categories (see Figure 3). The trajectory of *age group* categories notably follows the first axis. The second axis opposes *Man* category and *Woman* category. The categories of *health condition* are close to the centroid on the first dimension indicating a low association with the words on this dimension. The categories *poor* and *good* of *health condition* are opposed on the second axis; they lie close to *Man* and *Woman* respectively.

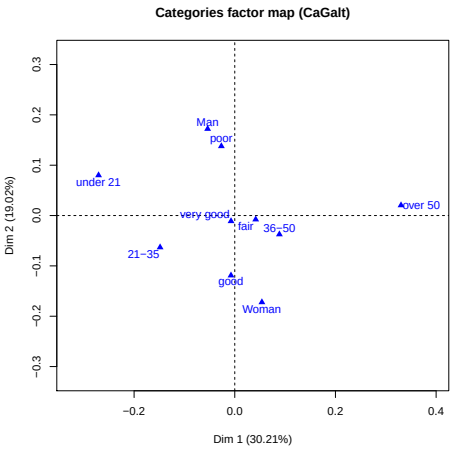

Figure 3: Categories for dimensions 1 and 2

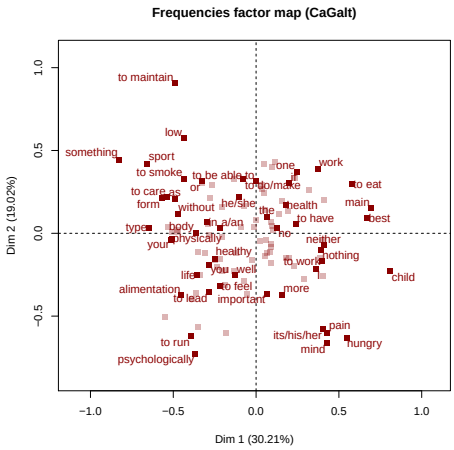

Figure 4: Frequencies for dimensions 1 and 2

Regarding the word graph (see Figure 4), *best*, *main* and *child* are the words with the highest coordinates at the right of the first axis. From the transition relationships between words and categories, we can say that both words are very used by the oldest categories and avoided by the youngest. These words contrast with *something* and *sport*, at the left of the first axis, being these latter words used by the youngest categories and avoided by the oldest. Comparing vocabulary choices depending on the gender (second axis), we see that men more frequently consider health from a physical point of view (*sport, to maintain, and so on*) and women from a psychological point of view (*psychologically, mind, and so on*).

The high number of elements drawn on the graphs (points and labels) makes 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. Analyzing the graph which was drawn with this code (see Figure 5), it shows that the fourth axis turns out to be of interest because of ranking health condition categories in their natural order. These categories are well separated, except for the two better health categories whose confidence ellipses overlap.

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 (see Figure 6).

7. Conclusion

The main features of the R function `CaGalt` have been detailed and illustrated by using an example issued from a survey by questionnaire including an open-ended question. This application to a

real data set demonstrates how both open-ended and closed questions combine to provide relevant information. Although this methodology was developed mainly to give an answer to the problem of analyzing open-ended questions with several quantitative, categorical and mixed contextual variables, it can be applied to any kind of frequency/contingency table with external variables. The function `CaGalt` can also be used to perform a CA-ALT because there is no other function in R for the same purpose.

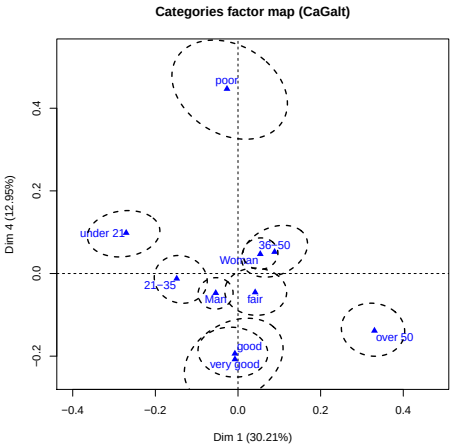

Figure 5: Categories for dimensions 1 and 4 completed by confidence ellipses

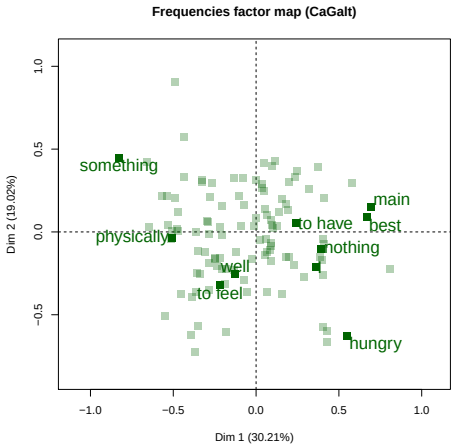

Figure 6: Ten words with the highest contributions on the first principal plane

Bibliography

M. Bécue-Bertaut and J. Pagès. A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, 45:481–503, 2004. [p3]

M. Bécue-Bertaut and J. Pagès. Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. *Advances in Data Analysis and Classification*, 9(2):125–142, 2015. [p1, 2, 4]

M. Bécue-Bertaut, J. Pages, and B. Kostov. Untangling the influence of several contextual variables on the respondents' lexical choices. a statistical approach. *SORT – Statistics and Operations Research Transactions*, 38(2):285–302, 2014. [p1, 2, 3]

E. J. Beh and R. Lombardo. *Correspondence Analysis: Theory, Practice and New Strategies*. John Wiley & Sons, 2014. [p1]

J. Benzécri. *Analyse des Données*. Dunod, 1973. [p1]

U. Böckenholt and Y. Takane. Linear constraints in correspondence analysis. In M. Greenacre and J. Blasius, editors, *Correspondence Analysis in the Social Sciences*. Academic Press, 1994. [p3]

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. [p4]

B. Escofier and J. Pagès. *Analyses factorielles simples et multiples*. Dunod, 1988. [p3, 4]

M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984. [p4]

M. Greenacre. *Correspondence analysis in practice*. CRC Press, 2007. [p1]

F. Husson, S. Lê, and J. Pagès. *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall / CRC Press, 2010. [p1]

S. Lê, J. Josse, and F. Husson. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008. [p1]

L. Lebart, A. Salem, and L. Berry. *Exploring Textual Data*. Kluwer, 1998. [p1, 2, 4]

F. Murtagh. *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall / CRC Press, 2005. [p1, 2]

Belchin Kostov
Transverse group for research in primary care, IDIBAPS
Mejia Lequerica, s / n.
08028 Barcelona
Spain
badriyan@clinic.ub.es

Mónica Bécue-Bertaut
Department of Statistics and Operational Research
Universitat Politècnica de Catalunya
North Campus - C5
Jordi Girona 1-3
08034 Barcelona
Spain
monica.becue@upc.edu

François Husson
Agrocampus Rennes
65 rue de Saint-Brieuc
35042 Rennes France
husson@agrocampus-ouest.fr