

# 1 Topic

## Correspondence Analysis (CA) – Software comparison.

The correspondence analysis (or factorial correspondence analysis) is an exploratory technique which enables to detect the salient associations in a two-way contingency table. It proposes an attractive graphical display where the rows and the columns of the table are depicted as points. Thus, we can visually identify the similarities and the differences between the rows profiles (between the columns profiles). We can also detect the associations between rows and columns.

The correspondence analysis can be viewed as an approach to decompose the chi-squared statistic associated with a two-way contingency table into orthogonal factors. In fact, because CA is a descriptive technique, it can be applied to tables even if the chi-square test of independence is not appropriate. The only restriction is that the table must contain positive or zero values, the calculating the sum of the rows and the columns is possible, the rows and columns profiles can be interpreted<sup>1</sup>.

The correspondence analysis can be viewed as a factorial technique<sup>2</sup>. Factors are latent variables defined from linear combinations of the rows profiles (or columns profiles). We can use the factors scores coefficients to calculate the coordinate of supplementary rows or columns.

In this tutorial, we show how to implement the CA on a realistic dataset with various tools: Tanagra 1.4.48, which incorporates new features for a better reading of the results; R software, using the "ca" and "ade4" packages; OpenStat; and SAS (PROC CORRESP). We will see - as always - that all these software produce exactly the same numerical results (fortunately!). The differences are found mainly in terms of the organization of the outputs.

This paper completes a previous tutorial where we describe shortly the use of the correspondence analysis into Tanagra<sup>3</sup>.

## 2 Dataset

We use the dataset described into the Bendixen's paper (1996<sup>4</sup>). We compare the outputs of various tools with those of the author. The reader can refer also to this paper for the interpretation of the results, our main goal being to compare the behavior of different software in this tutorial.

The table comes from a survey where 100 housewives were asked which of the **L = 14** statements (rows) listed below they associated with any of **C = 8** breakfast foods (columns). We note that multiple responses were allowed. So, the total number of responses is **n = 1760**.

The goal is to highlight the main associations between the statements and the food. We will be able to detect also the foods which are characterized with the same statements.

---

<sup>1</sup> [http://en.wikipedia.org/wiki/Correspondence\\_analysis](http://en.wikipedia.org/wiki/Correspondence_analysis)

<sup>2</sup> <http://www.micheloud.com/FXM/COR/e/index.htm>

<sup>3</sup> <http://data-mining-tutorials.blogspot.fr/2008/11/correspondance-analysis.html>

<sup>4</sup> M. Bendixen, « A practical guide to the use of the correspondence analysis in marketing research », Marketing Research On-Line, 1 (1), pp. 16-38, 1996 ; [http://marketing-bulletin.massey.ac.nz/V14/MB\\_V14\\_T2\\_Bendixen.pdf](http://marketing-bulletin.massey.ac.nz/V14/MB_V14_T2_Bendixen.pdf)

		Foods							
		Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
Statement	Healthy	14	38	25	18	8	31	28	34
	Nutritious	14	28	25	25	7	32	26	31
	GoodSummer	42	22	11	13	7	37	16	35
	GoodWinter	10	10	32	26	6	11	19	8
	Expensive	6	33	5	27	3	9	18	10
	QuickEasy	54	33	8	2	15	26	8	20
	Tasty	24	21	16	34	11	33	26	26
	Economical	24	3	20	3	16	7	3	7
	ForATreat	5	3	3	31	4	4	16	17
	ForWeekdays	47	24	15	9	13	11	6	10
	ForWeekends	12	5	8	56	16	10	23	18
	Tasteless	8	6	2	2	0	0	2	1
	TooLongToPrepare	0	0	9	35	1	0	10	0
	FamilyFavourite	14	4	10	31	5	7	2	5

Figure 1 – Foods and statements (Bendixen, 1996)

### 3 Correspondence Analysis with Tanagra (14.4.48)

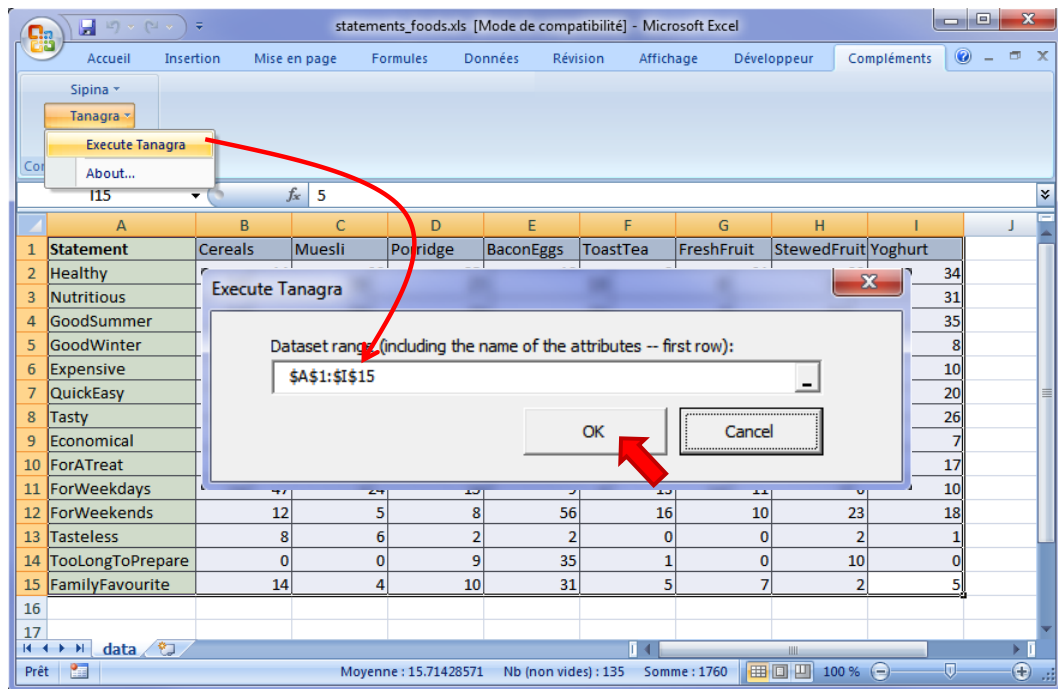
#### 3.1 Correspondence analysis with SAS

We use the results of SAS to check the outputs of the various tools. So, in a first time, we perform the analysis using the PROC CORRESP as follows<sup>5</sup>:

```
proc corresp data = mesdata.foods dimens=2;
var Cereals -- Yoghurt;
id Statement;
run;
```

We will describe the results below, by comparing them with those of Tanagra.

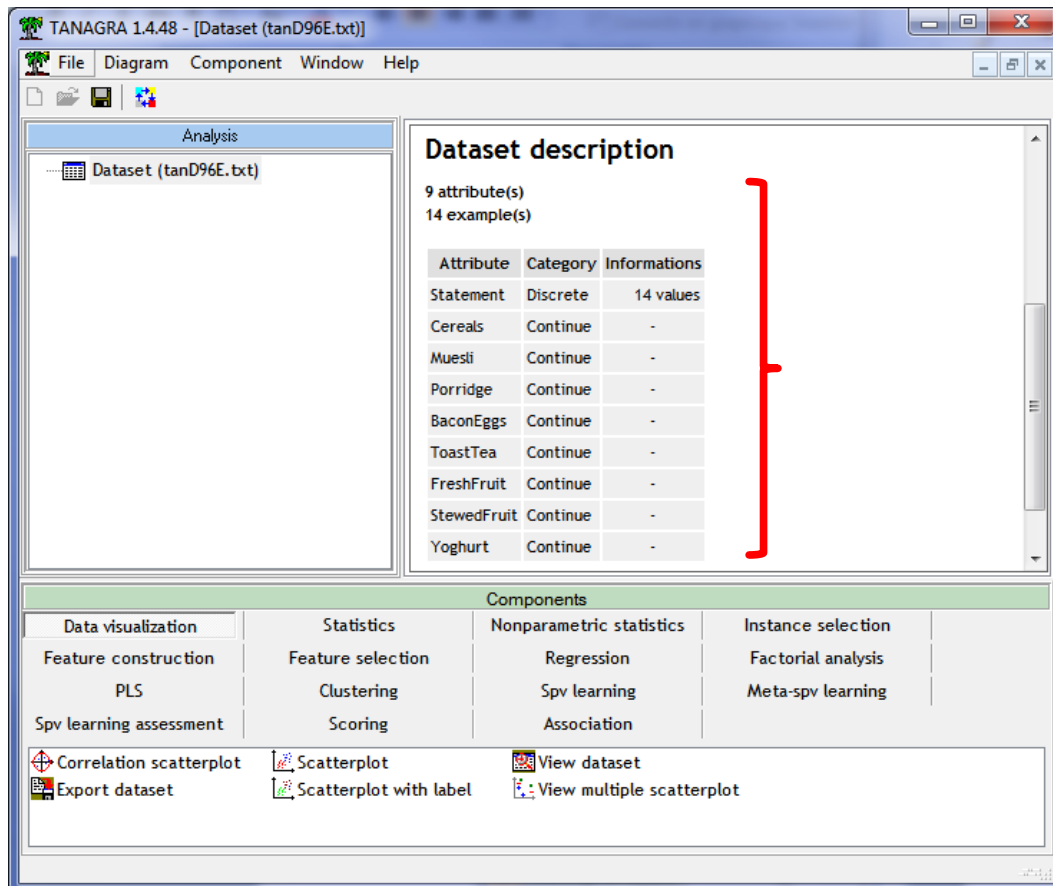
#### 3.2 Importing the data file into Tanagra



<sup>5</sup> See <http://data-mining-tutorials.blogspot.fr/2012/07/introduction-to-sas-proc-logistic.html> for the data importation (pages 2 and 3).

To import “statement\_foods.xls”, we use the add-in “**tanagra.xla**” which sends the dataset from the Excel spreadsheet to Tanagra<sup>6</sup>. A dialog box enables to check the data range (**\$A\$1:\$I\$15**). We confirm by clicking on the OK button.

Tanagra is automatically launched. The dataset is loaded (8 columns, 9 with the row labels; 14 rows, 15 with the column labels).

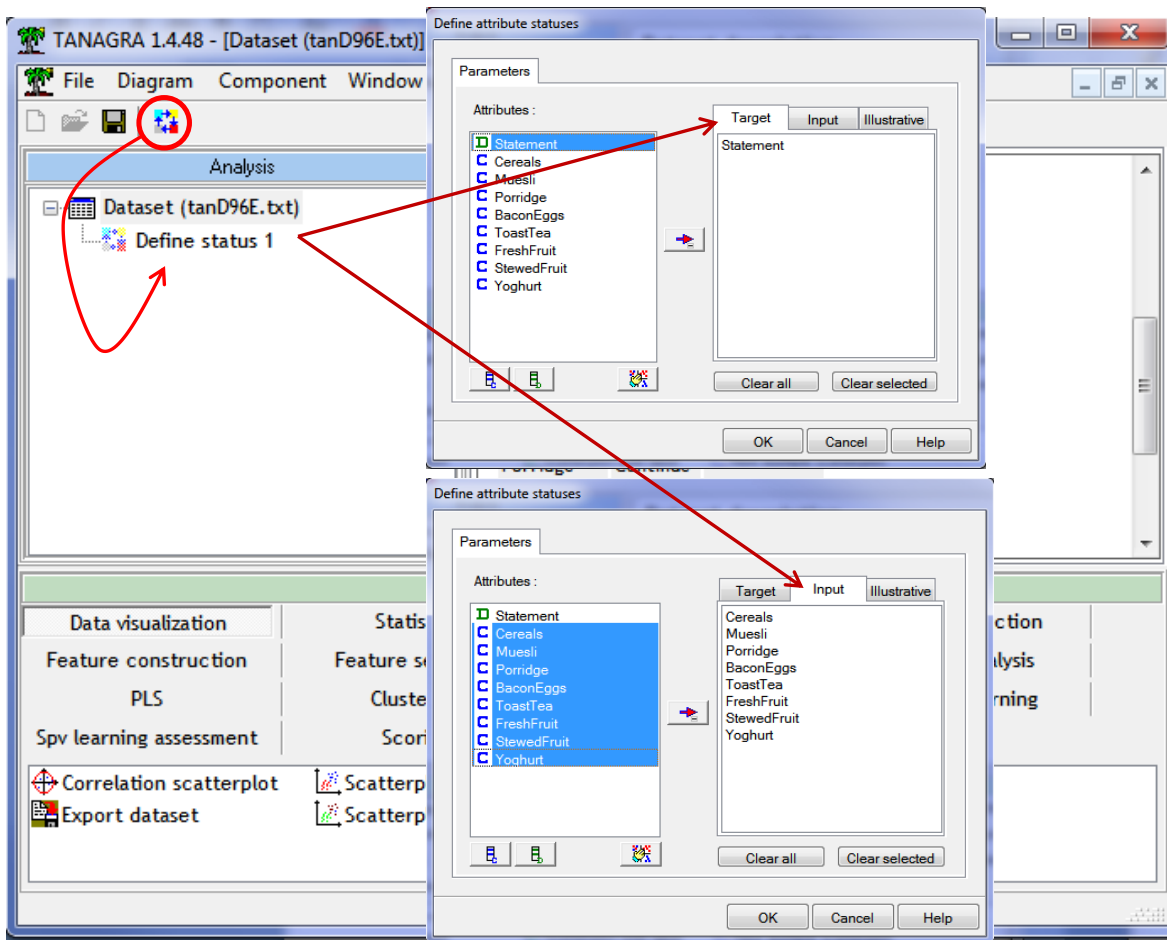


### 3.3 Specifying the parameters for the correspondence analysis

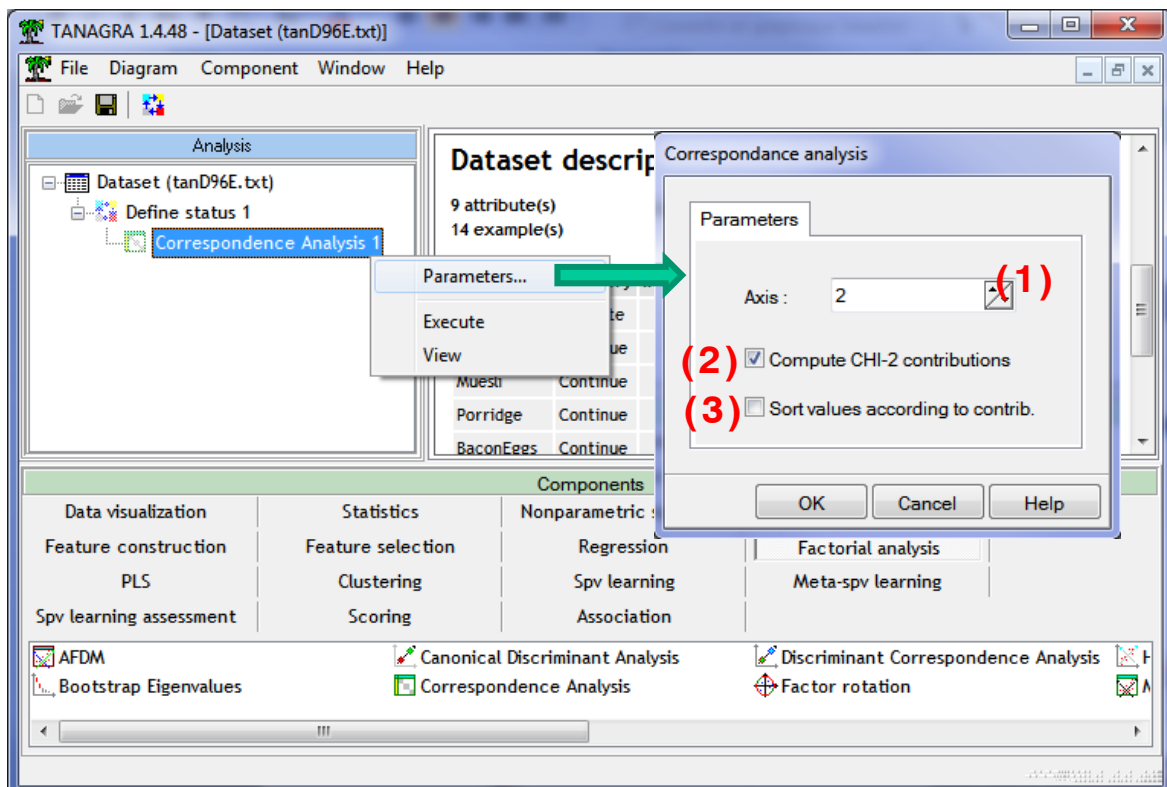
We must define the data used for the analysis using the DEFINE STATUS component. We do not really specify the role of the variables in the analysis, but rather to indicate to Tanagra the layout of the data in the contingency table.

We set STATEMENT (row labels) as TARGET, the other variables (columns) as INPUT.

<sup>6</sup> <http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html>; we can use a specific add-in for Open Office and Libre Office.



Then, we insert the CORRESPONDANCE ANALYSIS tool (FACTORIAL ANALYSIS tab) into the diagram. We click on the PARAMETERS menu to define the settings of the algorithm.



We ask 2 factors (1) (we explain why below); we calculate the contribution of each cells to the total inertia (2); we do not need to sort the output tables at this time (3).

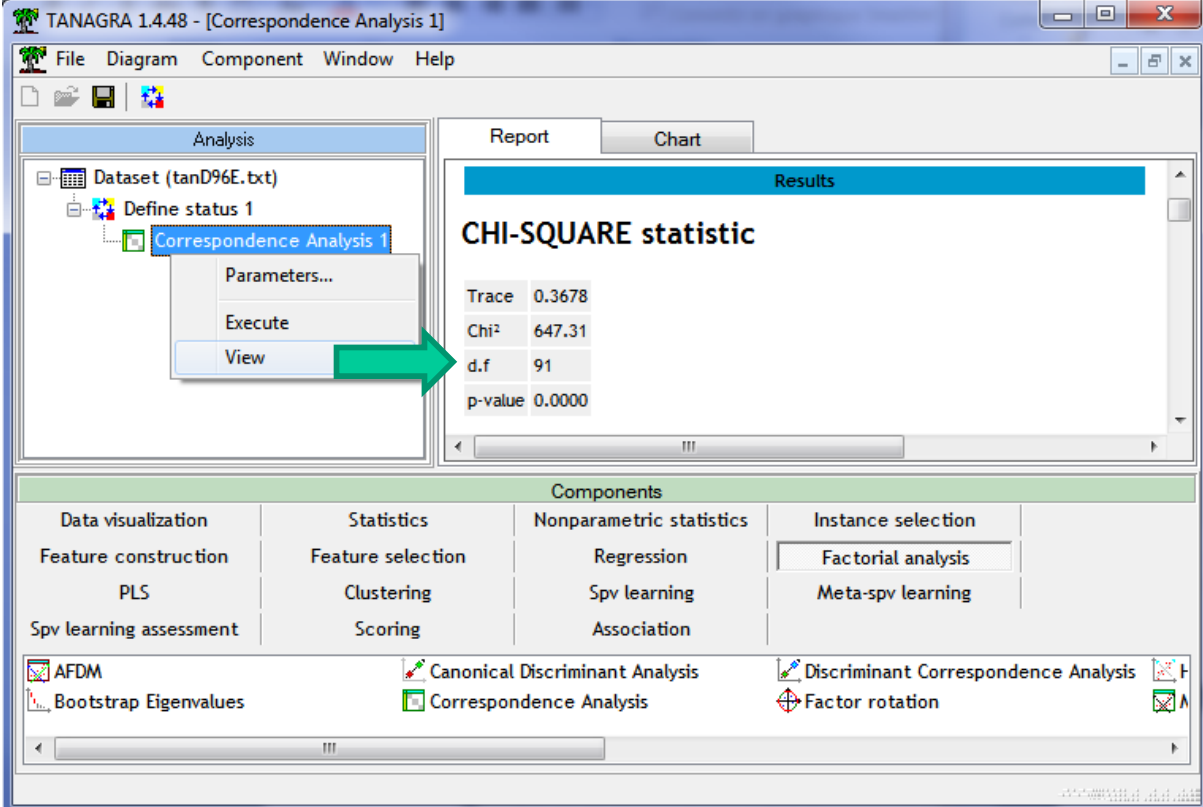
We confirm these settings and we click on the VIEW menu to obtain the results.

### 3.4 Reading of the outputs

The report is subdivided in several areas. We provide a detailed description in the next sections.

#### 3.4.1 CHI-squared statistic

The first table describes the results of the chi-square test of independence. This result is essential. Indeed, if the global association between the rows and the columns is too weak, the analysis of the associations between some rows and columns is not really useful. We must be sure that there is usable information into the table.



The screenshot shows the TANAGRA 1.4.48 software interface. The 'Analysis' tree on the left shows a project structure with 'Dataset (tanD96E.txt)', 'Define status 1', and 'Correspondence Analysis 1'. A context menu is open over 'Correspondence Analysis 1', with the 'View' option highlighted by a green arrow. The main window displays the 'Results' tab, which contains the following table:

CHI-SQUARE statistic	
Trace	0.3678
Chi <sup>2</sup>	647.31
d.f	91
p-value	0.0000

Below the results table, there is a 'Components' section with a grid of analysis options:

Components			
Data visualization	Statistics	Nonparametric statistics	Instance selection
Feature construction	Feature selection	Regression	Factorial analysis
PLS	Clustering	Spv learning	Meta-spv learning
Spv learning assessment	Scoring	Association	

At the bottom, there is a list of analysis methods: AFDM, Bootstrap Eigenvalues, Canonical Discriminant Analysis, Correspondence Analysis, Discriminant Correspondence Analysis, and Factor rotation.

Here, we obtain  $\chi^2_{\text{global}} = 647.31$ , with a degree of freedom equal to 91 [= (14 - 1) x (8 - 1)]. The association is statistically significant (p-value < 0.0001).

In addition, Tanagra provides the  $\phi^2$  statistic (**Trace**), with  $\phi^2 = \chi^2/n = 647.31 / 1760 = 0.3678$ . It is the total inertia of the cloud of points. The square root of the inertia corresponds to a kind of the correlation coefficient between the rows and the columns<sup>7</sup>.

The aim of the correspondence analysis is to decompose the  $\phi^2$  on a sequence of orthogonal factors.

<sup>7</sup> "As a rule of thumb, any value of  $\phi$  in excess of 0.2 indicates significant dependency" (Bendixen, page 7).

### 3.4.2 Eigenvalues table – Detecting the right number of factors

**Eigenvalues table.** Tanagra provides the eigenvalues table ( $\lambda_k$ ). They correspond to the part of the total inertia explained by the factors. Because we have orthogonal factors, the sum of all eigenvalues ( $0.193095 + 0.077731 + \dots + 0.002363$ ) is equal to the total inertia ( $\phi^2 = 0.3678$ ).

We can write the same information in the form of the proportion of the total inertia (e.g. % factor 1 =  $0.193095 / 0.3678 = 52.50\%$ ; % factor 2 =  $0.077731 / 0.3678 = 21.13\%$ ).

#### Eigen values

Matrix trace = 0.3678  
 SQR(Matrix trace) = 0.6065

Axis	Eigen value	% explained	Histogram	% cumulated
1	0.193095	52.50%		52.50%
2	0.077731	21.13%		73.64%
3	0.043854	11.92%		85.56%
4	0.032804	8.92%		94.48%
5	0.012257	3.33%		97.81%
6	0.005687	1.55%		99.36%
7	0.002363	0.64%		100.00%
Tot.	0.367791	-	-	-

(Tanagra)

The CORRESP Procedure

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	
0.43943	0.19309	339.846	52.50	52.50	*****
0.27880	0.07773	136.806	21.13	73.64	*****
0.20941	0.04385	77.183	11.92	85.56	*****
0.18112	0.03280	57.735	8.92	94.48	****
0.11071	0.01226	21.572	3.33	97.81	**
0.07541	0.00569	10.010	1.55	99.36	*
0.04861	0.00236	4.159	0.64	100.00	
Total	0.36779	647.312	100.00		

Degrees of Freedom = 91

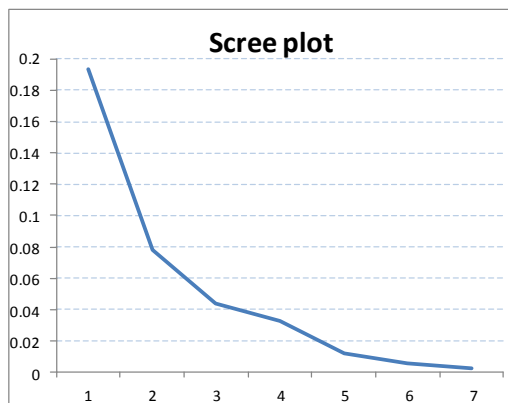
(SAS)

We observe a distinctive feature of SAS. It displays the decomposition of the chi-squared statistic  $\chi^2$ . The values correspond to the correspondence analysis eigenvalues multiplied by  $n = 1760$  (e.g.  $\chi^2_1 = 0.19309 \times 1760 = 339.846$ ; etc.). Of course, the sum of the chi-squared ( $\chi^2_k$ ) associated to the factors is equal to the global chi-squared ( $\chi^2_{\text{global}} = 647.312$ ).

**Selecting the right number of factors – The Kaiser-Gutman rule.** The determination of the right number of dimensions to retain is essential for an appropriate interpretation of the results. We know that the maximum number of factors that we can obtain is  $K_{\text{max}} = \text{MIN}(L - 1, C - 1) = \text{MIN}(13, 7) = 7$ . Thus, a really simple rule consists of selecting factors of which the proportion of explained variance is higher than  $(1 / K_{\text{max}}) = 14.3\%$ . For our dataset, these are factor 1 (52.50 %) and factor 2 (21.13 %).

This rule is very similar to the Kaiser-Guttman rule for principal component analysis (PCA). We retain the factors of which the eigenvalue is higher than the average of all eigenvalues ( $0.3678 / 7 = 0.0525$ ). We retain indeed the two first factors for our dataset ( $\lambda_1=0.19309$ ;  $\lambda_2= 0.07773$ ).

**Selecting the right number of factors – Scree plot.**



Similarly to the PCA, we can use also the scree plot for the detection of the right number of factors. We select the factors before (or before and including) the elbow in the plot of the eigenvalues according to the number of factors. Here, the solution with two factors seems to be suitable. This solution confirms the Kaiser-Guttman suggestion.

**Selecting the right number of factors – Malinvaud’s test.** In its book, Saporta (2009) describes a rule which is less empirical, based on a statistical approach. This is a sequential test which enables to check the significance of remaining factors if we select the K first ones.

It is based on the following test statistic:

$$Q_K = n \times (\lambda_{K+1} + \dots + \lambda_{K_{max}})$$

Under the null hypothesis, the K factors are sufficient, it follows a chi-squared distribution with  $(L - K - 1) \times (C - K - 1)$  degree of freedom.

The approach is statistically sound. But like all the process based on the chi-squared statistic, it tends to be always significant when the sample size ‘n’ increases. We observe this phenomenon here.

We describe into the table below the selection process:

K	Factor	Eigen value	CHI-2	ddl	p-value
0	1	0.193095	647.31	91	0.0000
1	2	0.077731	307.46	72	0.0000
2	3	0.043854	170.66	55	0.0000
3	4	0.032804	93.48	40	0.0000
4	5	0.012257	35.74	27	0.1211
5	6	0.005687	14.17	16	0.5862
6	7	0.002363	4.16	7	0.7613

For  $K = 0$ , we test the existence of at least one factor. It corresponds to the chi-square test for independence between the rows and the columns, indeed  $Q_0 = \chi^2_{global}$ .

For  $K = 3$  (i.e. 3 factors are enough or we need more?),

$$Q_3 = 1760 \times (0.032804 + 0.012257 + 0.005687 + 0.002363) = 93.48$$

With  $(L - K - 1) \times (C - K - 1) = (14 - 3 - 1) \times (8 - 3 - 1) = 40$  degree of freedom, we have a “p-value” < 0.0001. It seems that we need at least one additional factor.

For  $K = 4$  (i.e. 4 factors are enough or we need more?),

$$Q_4 = 1760 \times (0.012257 + 0.005687 + 0.002363) = 35.74$$

With  $(14 - 4 - 1) \times (8 - 4 - 1) = 27$  degree of freedom, we have p-value = 0.1211. It seems that  $K = 4$  factors is the right solution.

Obviously, this solution ( $K = 4$  axes) is not appropriate. It does not correspond to the result suggested by the Kaiser-Guttman rule and the scree test. It seems, according the reference used, that this test is only interesting on a moderate sample size. On a large dataset, it overestimates the number of factors to retain.

### 3.4.3 Row Points

This table describes information concerning the rows of the contingency table (Figure 2): some statistical information about each row (weight, squared distance from the origin, inertia) [A]; the row coordinates [B]; the row’s relative inertia contribution to the factor [C]; and the quality of the representation of each row ( $\text{COS}^2$ , individually and cumulatively on the factors) [D].

#### Rows analysis

Row Characterization				Coord.		Contributions (%)		COS <sup>2</sup>	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos <sup>2</sup> 1	cos <sup>2</sup> 2
Healthy	0.11136	0.16313	0.01817	0.08662	-0.34591	0.43	17.14	0.05 (0.05)	0.73 (0.78)
Nutritious	0.10682	0.10686	0.01141	-0.00861	-0.26886	0.00	9.93	0.00 (0.00)	0.68 (0.68)
GoodSummer	0.10398	0.22035	0.02291	0.34879	-0.13275	6.55	2.36	0.55 (0.55)	0.08 (0.63)
GoodWinter	0.06932	0.33945	0.02353	-0.27741	0.12713	2.76	1.44	0.23 (0.23)	0.05 (0.27)
Expensive	0.06307	0.40466	0.02552	-0.20067	-0.36125	1.32	10.59	0.10 (0.10)	0.32 (0.42)
QuickEasy	0.09432	0.46617	0.04397	0.64365	0.08476	20.24	0.87	0.89 (0.89)	0.02 (0.90)
Tasty	0.10852	0.03870	0.00420	-0.03963	-0.11273	0.09	1.77	0.04 (0.04)	0.33 (0.37)
Economical	0.04716	0.80423	0.03793	0.41692	0.65687	4.25	26.18	0.22 (0.22)	0.54 (0.75)
ForATreat	0.04716	0.54240	0.02558	-0.61946	-0.06492	9.37	0.26	0.71 (0.71)	0.01 (0.72)
ForWeekdays	0.07670	0.42038	0.03225	0.50675	0.33799	10.20	11.27	0.61 (0.61)	0.27 (0.88)
ForWeekends	0.08409	0.43297	0.03641	-0.56005	0.17123	13.66	3.17	0.72 (0.72)	0.07 (0.79)
Tasteless	0.01193	0.78910	0.00942	0.45096	0.15778	1.26	0.38	0.26 (0.26)	0.03 (0.29)
TooLongToPrepare	0.03125	1.82554	0.05705	-1.27778	0.33401	26.42	4.49	0.89 (0.89)	0.06 (0.96)
FamilyFavourite	0.04432	0.43902	0.01946	-0.38784	0.42183	3.45	10.15	0.34 (0.34)	0.41 (0.75)

(A)
(B)
(C)
(D)

Figure 2 – Row Points - Tanagra

SAS provides the same values, but they are spread across several tables (Figure 3). SAS normalizes in another way the inertia of rows.

Summary Statistics for the Row Points				Row Coordinates			Partial Contributions to Inertia for the Row Points			Squared Cosines for the Row Points		
	Quality	Mass	Inertia		Dim1	Dim2		Dim1	Dim2		Dim1	Dim2
Healthy	0.7795	0.1114	0.0494	Healthy	0.0866	-0.3459	Healthy	0.0043	0.1714	Healthy	0.0460	0.7335
Nutritious	0.6772	0.1068	0.0310	Nutritious	-0.0086	-0.2689	Nutritious	0.0000	0.0993	Nutritious	0.0007	0.6765
GoodSummer	0.6321	0.1040	0.0623	GoodSummer	0.3488	-0.1328	GoodSummer	0.0655	0.0236	GoodSummer	0.5521	0.0800
GoodWinter	0.2743	0.0693	0.0640	GoodWinter	-0.2774	0.1271	GoodWinter	0.0276	0.0144	GoodWinter	0.2267	0.0476
Expensive	0.4220	0.0631	0.0694	Expensive	-0.2007	-0.3612	Expensive	0.0132	0.1059	Expensive	0.0995	0.3225
QuickEasy	0.9041	0.0943	0.1195	QuickEasy	0.6437	0.0848	QuickEasy	0.2024	0.0087	QuickEasy	0.8887	0.0154
Tasty	0.3689	0.1085	0.0114	Tasty	-0.0396	-0.1127	Tasty	0.0009	0.0177	Tasty	0.0406	0.3284
Economical	0.7527	0.0472	0.1031	Economical	0.4169	0.6569	Economical	0.0425	0.2618	Economical	0.2161	0.5365
ForATreat	0.7152	0.0472	0.0695	ForATreat	-0.6195	-0.0649	ForATreat	0.0937	0.0026	ForATreat	0.7075	0.0078
ForWeekdays	0.8826	0.0767	0.0877	ForWeekdays	0.5068	0.3380	ForWeekdays	0.1020	0.1127	ForWeekdays	0.6109	0.2717
ForWeekends	0.7921	0.0841	0.0990	ForWeekends	-0.5600	0.1712	ForWeekends	0.1366	0.0317	ForWeekends	0.7244	0.0677
Tasteless	0.2893	0.0119	0.0256	Tasteless	0.4510	0.1578	Tasteless	0.0126	0.0038	Tasteless	0.2577	0.0315
TooLongToPrepare	0.9555	0.0313	0.1551	TooLongToPrepare	-1.2778	0.3340	TooLongToPrepare	0.2642	0.0449	TooLongToPrepare	0.8944	0.0611
FamilyFavourite	0.7480	0.0443	0.0529	FamilyFavourite	-0.3878	0.4218	FamilyFavourite	0.0345	0.1015	FamilyFavourite	0.3426	0.4053

(A)
(B)
(C)
(D)

Figure 3 – Row Points - SAS 9.3



The rows with high inertia have often a strong influence on the first factors. This is not really a problem. But, we must take into account this fact when we interpret the factors. For instance, 'TooLongToPrepare', 'QuickEasy' and 'Economical' will determine the first two factors if we consider the contributions.

The  $COS^2$  indicates the quality of the representation of the row on the factor. We have also the cumulative  $COS^2$  for the first K factors. For our dataset, we observe that only 'GoodWinter' and 'TasteLess' are not well represented on the selected factors.

For each factor, Tanagra highlights the coordinates of the rows for which:  $(CTR > 1/L)$  and  $(COS^2 > 1/K_{max})$ . The idea is to draw attention of the users on the most important results into the table. For our dataset, we observe that the first factor is determined by the opposition between (ForAtreat, ForWeekEnds, TooLongToPrepare) on the one hand, and (QuickEasy, ForWeekDays) on the other hand. About the second factor, it shows the opposition between (Healthy, Nutritious, Expensive) and (Economical, FamilyFavourite, ForWeekADays).

### 3.4.4 Column points

The representation of the columns follows the same process.

#### Columns analysis

Row Characterization				Coord.		Contributions (%)		COS <sup>2</sup>	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos <sup>2</sup> 1	cos <sup>2</sup> 2
Cereals	0.15568	0.48968	0.07623	0.56272	0.35791	25.53	25.66	0.65 (0.65)	0.26 (0.91)
Muesli	0.13068	0.33926	0.04433	0.31310	-0.31869	6.63	17.08	0.29 (0.29)	0.30 (0.59)
Porridge	0.10739	0.35450	0.03807	-0.05363	0.21310	0.16	6.27	0.01 (0.01)	0.13 (0.14)
BaconEggs	0.17727	0.66963	0.11871	-0.78344	0.14814	56.35	5.00	0.92 (0.92)	0.03 (0.95)
ToastTea	0.06364	0.37744	0.02402	0.17534	0.44702	1.01	16.36	0.08 (0.08)	0.53 (0.61)
FreshFruit	0.12386	0.19655	0.02434	0.24619	-0.24493	3.89	9.56	0.31 (0.31)	0.31 (0.61)
StewedFruit	0.11534	0.19120	0.02205	-0.31540	-0.24281	5.94	8.75	0.52 (0.52)	0.31 (0.83)
Yoghurt	0.12614	0.15878	0.02003	0.08599	-0.26416	0.48	11.32	0.05 (0.05)	0.44 (0.49)

Figure 4 – Column points - Tanagra

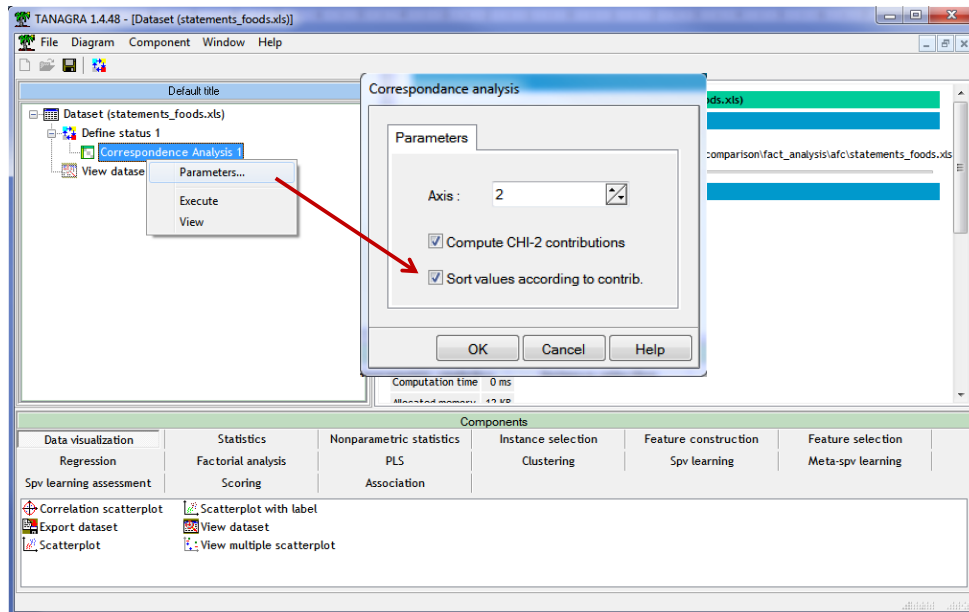
Because they have a high inertia, we know that “BaconEggs” and “Cereals” will have a high influence on the results. We observe that they determined largely the first two factors. For the column points, the coordinates are highlighted when  $(CTR > 1 / C)$  and  $(COS^2 > 1/K_{max})$ .

Of course, the SAS outputs are consistent with those of Tanagra.

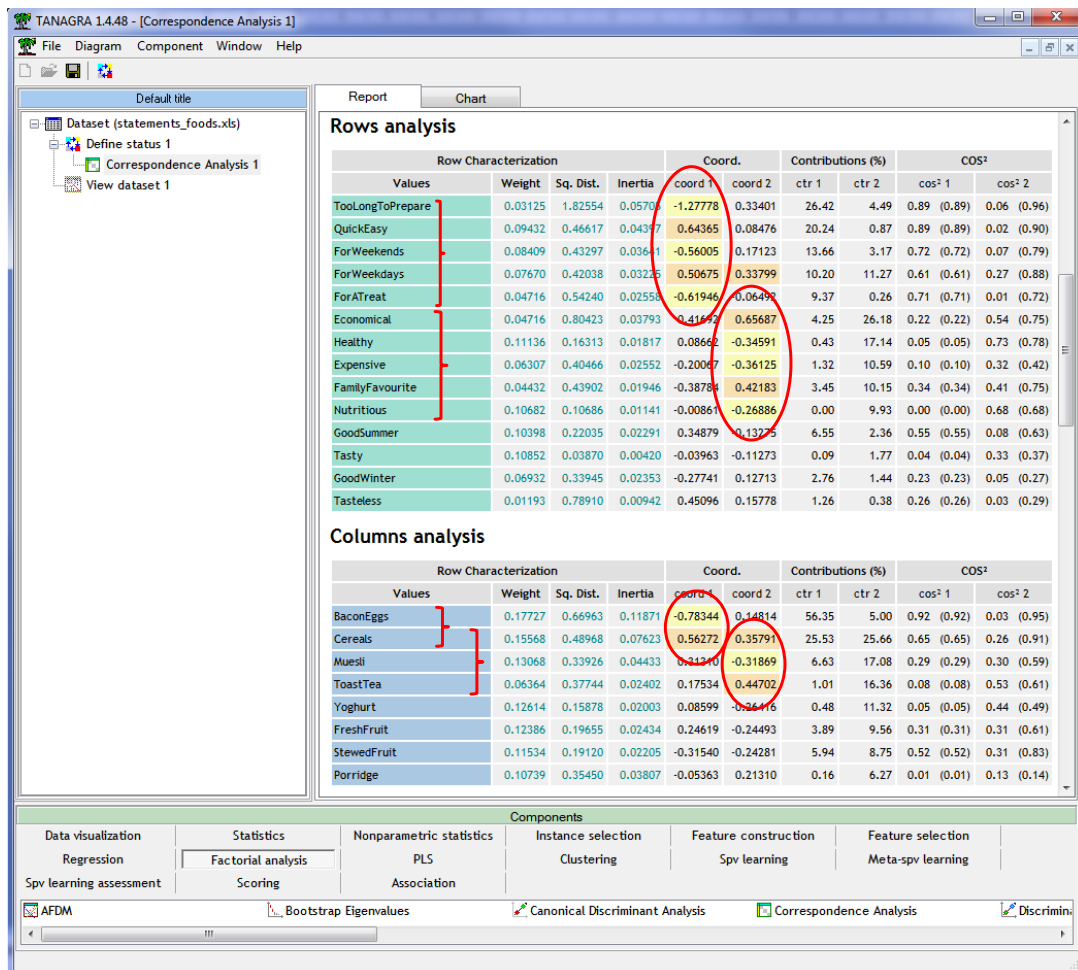
### 3.4.5 Sorting the column/row points according to their contribution

When we have a large number of rows (or columns), the reading of the tables above become difficult. We can improve the visualization by sorting the rows (columns) according to their contribution. Because the information related to the rows (columns) are spread on several factors, we cannot sort them according the first factor only. Tanagra uses the following strategy: for the first factor, it identifies the rows (columns) which have a contribution higher than  $(1/L)$   $(1/C)$  for columns), it sorts them according to the contribution; then, for the remaining rows (columns), it identifies

those which have a contribution larger than  $(1/L)$  ( $1/C$  for columns) on the second factor, it sorts them; etc. Thus, this process enables us to visualize easily the rows (columns) related to the first factor, to the second one, etc. We click on the PARAMETERS menu of the CORRESPONDENCE ANALYSIS component. We activate the "Sort values according to contrib." option.



We confirm and we click on the VIEW menu.



We identify better the group of values (rows or columns points) associated to each factor.

### 3.4.6 Simultaneous representation

The ability to produce graphical representation is most probably one of the reasons of the popularity of the correspondence analysis (and more generally speaking of the factor analysis methods). Tanagra proposes the symmetric representation of the rows and column points.

Into the CHART tab, we select the two first factors for the graphical representation.

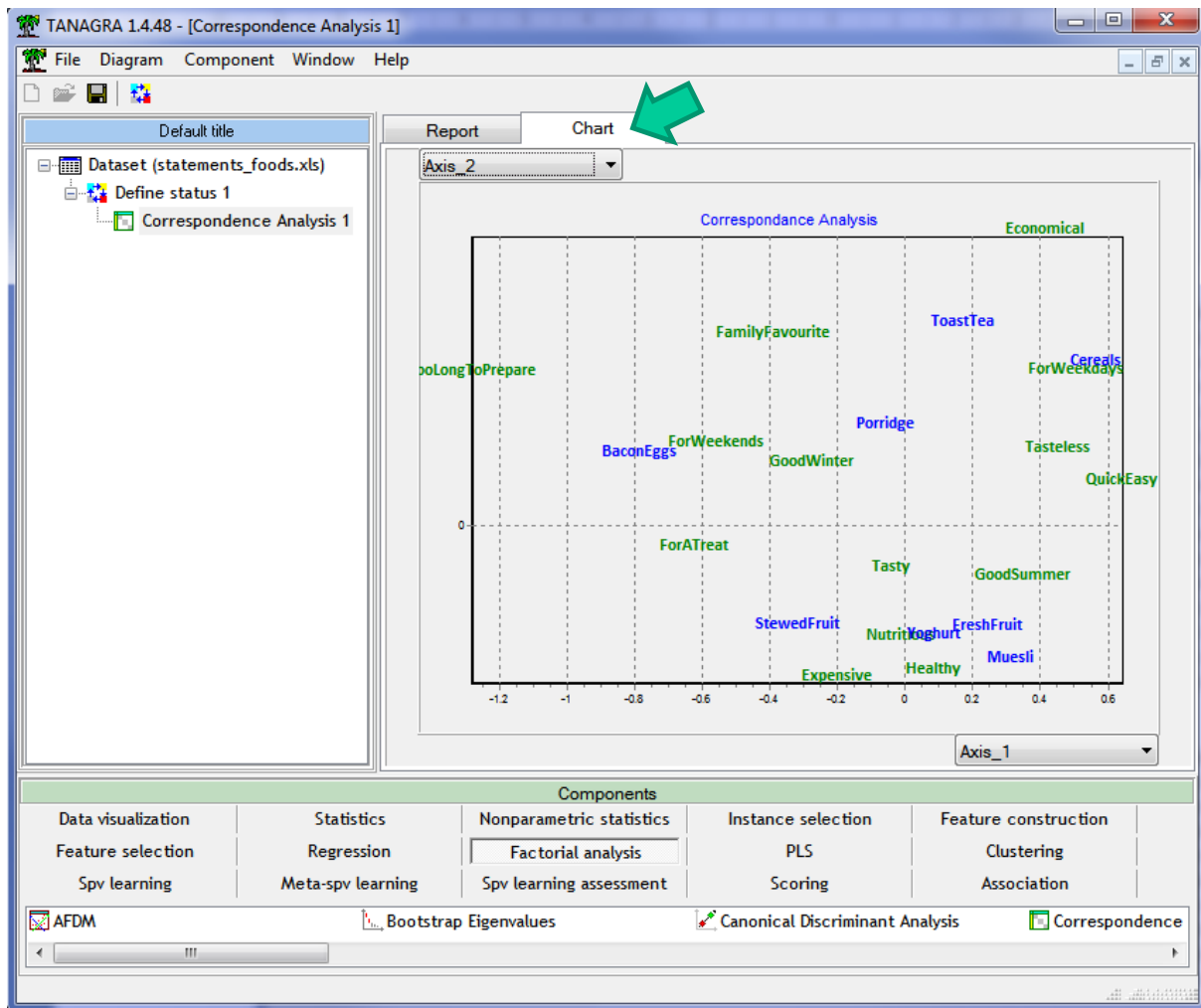


Figure 5 – Symmetric plot - Tanagra

This representation is the most popular but it is the most controversial also because the proximity between the rows and the columns must be considered cautiously, especially when they are located near to the origin. It does not correspond always to an association between the rows and the column of the contingency table. We must analysis the contribution to the chi-squared to confirm (or reject) the relation. For instance, the visual proximity between (nutritious, healthy) and (yogurt) does not correspond to an association when we analyze deeply the contingency table (see section 3.4.7).

On the other hand, when we consider the points at the extremities of the graph, for the rows and columns with high contribution or/and high quality (CTR and  $COS^2$ ), the proximities between rows and columns may correspond to an association (attraction or repulsion) into the contingency table.

However, each row point (column point) must be analyzed globally according to all the column points (row points). For instance, we report here the coordinates of the column points on the first factor.

Coord.Colonnes.Axe.1							
0.563	0.313	-0.054	-0.783	0.175	0.246	-0.315	0.086
Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt

We must calculate the coordinate of the row "TooLongTo Prepare" in relation to these column points. For this, we need to its row profile.

Profil ligne - TooLongToPrepare								
	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
TooLongToPrepare	0.00	0.00	0.16	0.64	0.02	0.00	0.18	0.00

We can guess that "BaconEggs" has the highest influence here. But this is not the only one. To calculate the coordinate of the row point, we need to the eigenvalue ( $\lambda_1 = 0.1931$ ) associated to the factor. We obtain the following coordinate (we exclude the null values):

$$\frac{1}{\sqrt{0.1931}} [0.16 \times (-0.054) + 0.64 \times (-0.783) + 0.02 \times 0.175 + 0.18 \times (-0.315)] = -1.278$$

This is the value reported in the rows point table (Figure 2).

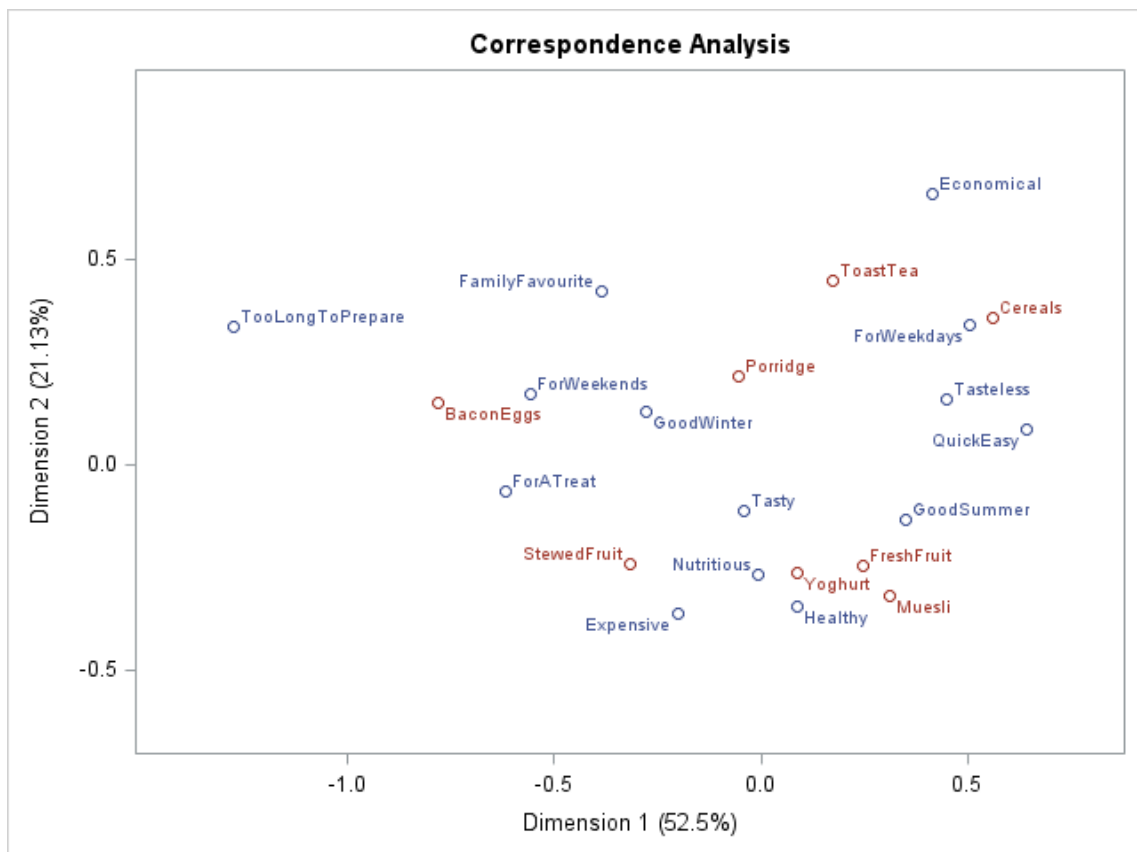


Figure 6 – Symmetric graph - SAS

We obtain the same symmetric graph with SAS (Figure 6).

### 3.4.7 Contribution to the chi-square

Tanagra provides optionally the table of the contributions to the total chi-square statistic  $\chi^2_{\text{global}}$ . For each combination of row and column, we have: the observed number; the theoretical number under the null hypothesis (the row and the columns are independent); the standardized residual; the absolute contribution to the total chi-square statistic (+: attraction, -: repulsion); the relative contribution (in percentage). Tanagra sorts the table according to the relative contribution. Only the cells with a relative contribution higher than  $[1 / (L \times C)]$  are showed.

#### CHI-2 contributions

Id	Row	Column	Value	Expected	Std.Resid.	Contrib.	%
1	TooLongToPrepare	BaconEggs	35	9.8	8.09	(+) 65.39	10.10
2	ForWeekends	BaconEggs	56	26.2	5.81	(+) 33.77	5.22
3	ForWeekdays	Cereals	47	21.0	5.67	(+) 32.12	4.96
4	QuickEasy	Cereals	54	25.8	5.54	(+) 30.68	4.74
5	GoodWinter	Porridge	32	13.1	5.22	(+) 27.26	4.21
6	QuickEasy	BaconEggs	2	29.4	-5.06	(-) 25.56	3.95
7	Expensive	Muesli	33	14.5	4.86	(+) 23.58	3.64
8	Economical	ToastTea	16	5.3	4.66	(+) 21.75	3.36
9	FamilyFavourite	BaconEggs	31	13.8	4.62	(+) 21.33	3.29
10	ForATreat	BaconEggs	31	14.7	4.25	(+) 18.03	2.78
11	Economical	Porridge	20	8.9	3.71	(+) 13.79	2.13
12	GoodSummer	BaconEggs	13	32.4	-3.41	(-) 11.65	1.80
13	ForWeekends	Muesli	5	19.3	-3.26	(-) 10.63	1.64
14	Economical	Cereals	24	12.9	3.08	(+) 9.50	1.47
15	Economical	BaconEggs	3	14.7	-3.05	(-) 9.33	1.44
16	ForWeekdays	BaconEggs	9	23.9	-3.05	(-) 9.32	1.44
17	GoodSummer	FreshFruit	37	22.7	3.01	(+) 9.06	1.40
18	Healthy	Cereals	14	30.5	-2.99	(-) 8.94	1.38
19	TooLongToPrepare	Cereals	0	8.6	-2.93	(-) 8.56	1.32
20	Healthy	BaconEggs	18	34.7	-2.84	(-) 8.07	1.25
21	Nutritious	Cereals	14	29.3	-2.82	(-) 7.96	1.23
22	Expensive	Cereals	6	17.3	-2.71	(-) 7.36	1.14
23	TooLongToPrepare	Muesli	0	7.2	-2.68	(-) 7.19	1.11
24	TooLongToPrepare	Yoghurt	0	6.9	-2.63	(-) 6.94	1.07
25	Tasteless	Cereals	8	3.3	2.62	(+) 6.85	1.06
26	TooLongToPrepare	FreshFruit	0	6.8	-2.61	(-) 6.81	1.05
27	QuickEasy	StewedFruit	8	19.1	-2.55	(-) 6.49	1.00
28	GoodSummer	Cereals	42	28.5	2.53	(+) 6.41	0.99
29	GoodSummer	Yoghurt	35	23.1	2.48	(+) 6.15	0.95
30	Healthy	Muesli	38	25.6	2.45	(+) 5.99	0.93
31	QuickEasy	Muesli	33	21.7	2.43	(+) 5.89	0.91
32	ForWeekdays	StewedFruit	6	15.6	-2.43	(-) 5.88	0.91

Figure 7 – Contributions to the  $\chi^2$  - Tanagra

It is clear that the information provided by the contingency table relies heavily on the association between "TooLongToPrepare" and "BaconEggs" (10.10%). We appraise better the situation by plotting the curve of decrease of contributions (Figure 8, this graph is not provided by Tanagra). *By considering this table, it is perhaps more judicious to set "BaconEggs" as supplementary point.*

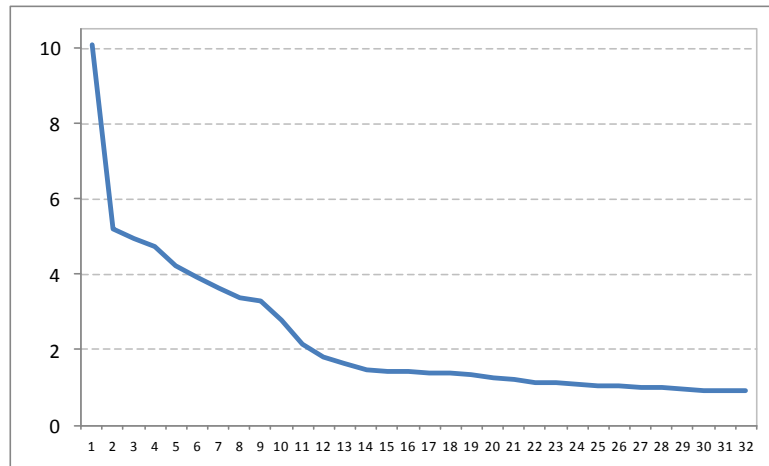


Figure 8 – Evolution of the contribution to the CHI-2

This table is really important because it allows to confirm (or reject) the associations highlighted in the symmetric graph. For instance, it seems that (yoghurt) is related to (nutritious and healthy). When we analyze the table of the contributions, we note that this association is not really true. The only characteristics which are really related to yoghurt are "TooLongToPrepare" (negative association) and "GoodToSummer" (positive association).

Id	Row	Column	Value	Expected	Std.Resid	Contrib.	%
24	TooLongToPrepare	Yoghurt	0	6.9	-2.63 (-)	6.94	1.07
29	GoodSummer	Yoghurt	35	23.1	2.48 (+)	6.15	0.95

### 3.4.8 Coordinates for supplementary rows

**Calculation of the coordinates.** We can compute the coordinates of supplementary rows from the correspondence analysis results. In its tutorial, Bendixen ([page 11](#)) provides the following table. It describes the frequency usage of the foods.

	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
I	24	3	4	8	18	2	9	11
II	58	15	8	13	16	10	10	29
III	6	10	12	46	8	14	15	8
IV	2	4	28	9	4	47	4	2
V	10	68	48	24	54	27	62	50

I : Daily ; II : Several times per week ; III : Several times per month ; IV : Every few months ; V : Never.

How to calculate the coordinates of these new rows (I, II, III, IV, V) in relation to the existing ones?

We can obtain the right values from the coordinates of the columns (e.g. <http://data-mining-tutorials.blogspot.fr/2008/11/correspondance-analysis.html>). But it is more convenient to have a set of coefficients that we can directly apply to the row profile. In addition, these coefficients can be easily deployed in an external tool (e.g. in a spreadsheet). In doing so, Tanagra provides the factor score coefficients table.

Factor score coefficients for supplementary row

From column values (relative frequency)

Column value	Factor 1	Factor 2
Cereals	1.280579	1.283725
Muesli	0.712525	-1.143072
Porridge	-0.122041	0.764323
BaconEggs	-1.782883	0.531349
ToastTea	0.399026	1.603371
FreshFruit	0.560261	-0.878505
StewedFruit	-0.717755	-0.870916
Yoghurt	0.195685	-0.947476

Figure 9 – Factor score coefficients for supplementary rows

Let us consider the row profiles of the supplementary rows above.

	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt
I	0.304	0.038	0.051	0.101	0.228	0.025	0.114	0.139
II	0.365	0.094	0.050	0.082	0.101	0.063	0.063	0.182
III	0.050	0.084	0.101	0.387	0.067	0.118	0.126	0.067
IV	0.020	0.040	0.280	0.090	0.040	0.470	0.040	0.020
V	0.029	0.198	0.140	0.070	0.157	0.079	0.181	0.146

Figure 10 – Row profiles of the supplementary rows

We can calculate their coordinates on the first two factors using the factor score coefficients.

	Scores	
	Factor.1	Factor.2
I	0.280	0.551
II	0.448	0.321
III	-0.562	0.082
IV	0.114	-0.161
V	0.042	-0.157

Figure 11 – Coordinates of the supplementary rows

We detail the calculations for « III: Several times per month » on the 1<sup>st</sup> factor.

$$C(\text{III}, \text{Factor 1}) = 1.280579 \times 0.050 + 0.712525 \times 0.084 + (-0.122041) \times 0.101 + (-1.782883) \times 0.387 + 0.399026 \times 0.067 + 0.560261 \times 0.118 + (-0.717755) \times 0.126 + 0.195685 \times 0.067 = -0.562$$

It is rather near to “TooLongToPrepare, ForATreat, ForWeekends”. It is not surprising knowing that the Bacon Eggs is the preferred food in this situation (38.7 %, see Figure 10).

**Quality of the representation.** To evaluate the quality of the representation of the supplementary point on a factor, we calculate the COS<sup>2</sup>. For that, we must calculate first the distance to the origin which corresponds to the marginal row profile.

Marge colonne	Cereals	Muesli	Porridge	BaconEggs	ToastTea	FreshFruit	StewedFruit	Yoghurt	Total
Effectif	274	230	189	312	112	218	203	222	1760
Profil	0.156	0.131	0.107	0.177	0.064	0.124	0.115	0.126	-

Figure 12 – Row profile of the marginal distribution

We calculate the squared distance to this marginal distribution for each supplementary row.

	Dist <sup>2</sup>
I	0.773
II	0.473
III	0.364
IV	1.616
V	0.407

Figure 13 – Squared distance to the origin for each supplementary point

Here is the detail of the calculations for « III: Several times per month » (from the values available in Figure 10 and Figure 12) :

$$D^2(III, Marge) = \frac{(0.050 - 0.156)^2}{0.156} + \frac{(0.084 - 0.131)^2}{0.131} + \dots + \frac{(0.050 - 0.126)^2}{0.126} = 0.364$$

Thus, the COS<sup>2</sup> for « III » on the first factor is:

$$COS^2(III, Facteur 1) = \frac{(-0.562)^2}{0.364} = 0.866$$

In the same way, we obtain the COS<sup>2</sup> of each supplementary point on the first two factors.

Cos <sup>2</sup>	Factor.1	Factor.2
I	0.101	0.393
II	0.425	0.218
III	0.866	0.019
IV	0.008	0.016
V	0.004	0.061

Figure 14 – Quality of the representation of the supplementary points

We observe that “III: Several times per month” is well described on the 1<sup>st</sup> factor (cos<sup>2</sup> = 0.866). It means that the proximity with (“Too Long to Prepare”, “For a Treat” and “For Weekends”) is meaningful. In addition, the association with “Bacon Eggs” is credible.

### 3.4.9 Supplementary columns

Factor score coefficients for supplementary column  
From row values (relative frequency)

Row value	Factor 1	Factor 2
Healthy	0.197122	-1.240706
Nutritious	-0.019587	-0.964348
GoodSummer	0.793731	-0.476147
GoodWinter	-0.631303	0.455973
Expensive	-0.456668	-1.295717
QuickEasy	1.464761	0.304003
Tasty	-0.090190	-0.404344
Economical	0.948796	2.356032
ForATreat	-1.409708	-0.232841
ForWeekdays	1.153220	1.212284
ForWeekends	-1.274503	0.614179
Tasteless	1.026245	0.565924
TooLongToPrepare	-2.907840	1.198004
FamilyFavourite	-0.882613	1.513020

Figure 15 - Factor score coefficients for supplementary columns - Tanagra



The same approach can be applied to supplementary columns. For that, Tanagra provides the factor score coefficients table. We can locate the coordinate of a new kind of food from its column profile.

## 4 Correspondence analysis with R

The multiplicity of packages is both an advantage and a weakness of R. This is an advantage because we have several points of view for the same problem. This can improve the analysis. A weakness because the (apparently) difference in results for the same problem is disturbing, especially when we do not know well the underlying statistical method. In addition, sometimes, the outputs of some packages do not correspond to the presentations found in reference books. This can trouble the user.

In this section, we will use two popular packages for the correspondence analysis. We will compare the results between them on the one hand, with SAS and Tanagra on the other hand. We will see that, through some intermediate transformations in some cases, we find exactly the same numerical results. That is what is important ultimately.

### 4.1 Chi-square test

In a first time, we perform the chi-square test of independence.

```
foods <- read.table(file="statements_foods.txt",header=T,sep="\t",row.names=1)
#chi-squared test
print(chisq.test(foods))
```

R shows a warning message because the number of instances in some cells under the null hypothesis is lower than 5 (the "tasteless" category is rare, it corresponds to the 1.2% of the responses).

```
> print(chisq.test(foods))

      Pearson's Chi-squared test

data:  foods
X-squared = 647.3121, df = 91, p-value < 2.2e-16

Message d'avis :
In chisq.test(foods) : Chi-squared approximation may be incorrect
```

### 4.2 The 'ca' package

The "ca" package is due to Michael Greenacre and Oleg Nenadic. The first one has published several books about the factorial analysis methods.

#### 4.2.1 Correspondence analysis – 2 factors

After we load the library, we perform the correspondence analysis by asking 2 factors.

```
#perform the correspondence analysis
library(ca)
foods.ca <- ca(foods,nd=2)
#eigen values and cumulative proportion of variance explained in percentage
print(cbind(foods.ca$sv^2, (100.0*cumsum(foods.ca$sv^2)/sum(foods.ca$sv^2))))
```

We obtain the eigenvalues and the cumulative proportion of variance explained by the factors.

```

      [,1]      [,2]
[1,] 0.193094526 52.50116
[2,] 0.077730798 73.63567
[3,] 0.043854131 85.55932
[4,] 0.032804216 94.47858
[5,] 0.012256794 97.81112
[6,] 0.005687400 99.35749
[7,] 0.002363091 100.00000

```

#### 4.2.2 Row points

We use the `'foods.ca'` object ("ca" class) to obtain the various results.

```

#row analysis
attach(foods.ca)
row.ca <- round(cbind(rowmass,rowdist^2,rowinertia,rowcoord[,1]*sv[1],rowcoord[,2]*sv[2]),5)
colnames(row.ca) <- c("weight","sq.dist","inertia","coord.1","coord.2")
rownames(row.ca) <- rownames
print(row.ca)

```

We obtain the weight, the squared distance to the origin, the inertia, and the coordinates of rows:

	weight	sq.dist	inertia	coord.1	coord.2
Healthy	0.11136	0.16313	0.01817	0.08662	0.34591
Nutritious	0.10682	0.10686	0.01141	-0.00861	0.26886
GoodSummer	0.10398	0.22035	0.02291	0.34879	0.13275
GoodWinter	0.06932	0.33945	0.02353	-0.27741	-0.12713
Expensive	0.06307	0.40466	0.02552	-0.20067	0.36125
QuickEasy	0.09432	0.46617	0.04397	0.64365	-0.08476
Tasty	0.10852	0.03870	0.00420	-0.03963	0.11273
Economical	0.04716	0.80423	0.03793	0.41693	-0.65687
ForATreat	0.04716	0.54240	0.02558	-0.61946	0.06492
ForWeekdays	0.07670	0.42038	0.03225	0.50675	-0.33799
ForWeekends	0.08409	0.43297	0.03641	-0.56005	-0.17123
Tasteless	0.01193	0.78910	0.00942	0.45096	-0.15778
TooLongToPrepare	0.03125	1.82554	0.05705	-1.27778	-0.33401
FamilyFavourite	0.04432	0.43902	0.01946	-0.38784	-0.42183

Figure 16 – Row points - Package 'ca'

We can compare these values to those of Tanagra (Figure 2).

#### 4.2.3 Column points

```

#column analysis
col.ca <- round(cbind(colmass,colldist^2,colinertia,colcoord[,1]*sv[1],colcoord[,2]*sv[2]),5)
colnames(col.ca) <- c("weight","sq.dist","inertia","coord.1","coord.2")
rownames(col.ca) <- colnames
print(col.ca)

```

Here also the results (Figure 17) are the same as those of Tanagra (Figure 4).

	weight	sq.dist	inertia	coord.1	coord.2
Cereals	0.15568	0.48968	0.07623	0.56272	-0.35791
Muesli	0.13068	0.33926	0.04433	0.31310	0.31869
Porridge	0.10739	0.35450	0.03807	-0.05363	-0.21310
BaconEggs	0.17727	0.66963	0.11871	-0.78344	-0.14814
ToastTea	0.06364	0.37744	0.02402	0.17534	-0.44702
FreshFruit	0.12386	0.19655	0.02434	0.24619	0.24493
StewedFruit	0.11534	0.19120	0.02205	-0.31540	0.24281
Yoghurt	0.12614	0.15878	0.02003	0.08599	0.26416

Figure 17 – Column points - Package 'ca'

#### 4.2.4 Symmetric graph

The plot() procedure is overwritten to handle the object "ca".

```
#plot rows and columns
plot(foods.ca)
```

We obtain the symmetric graph.

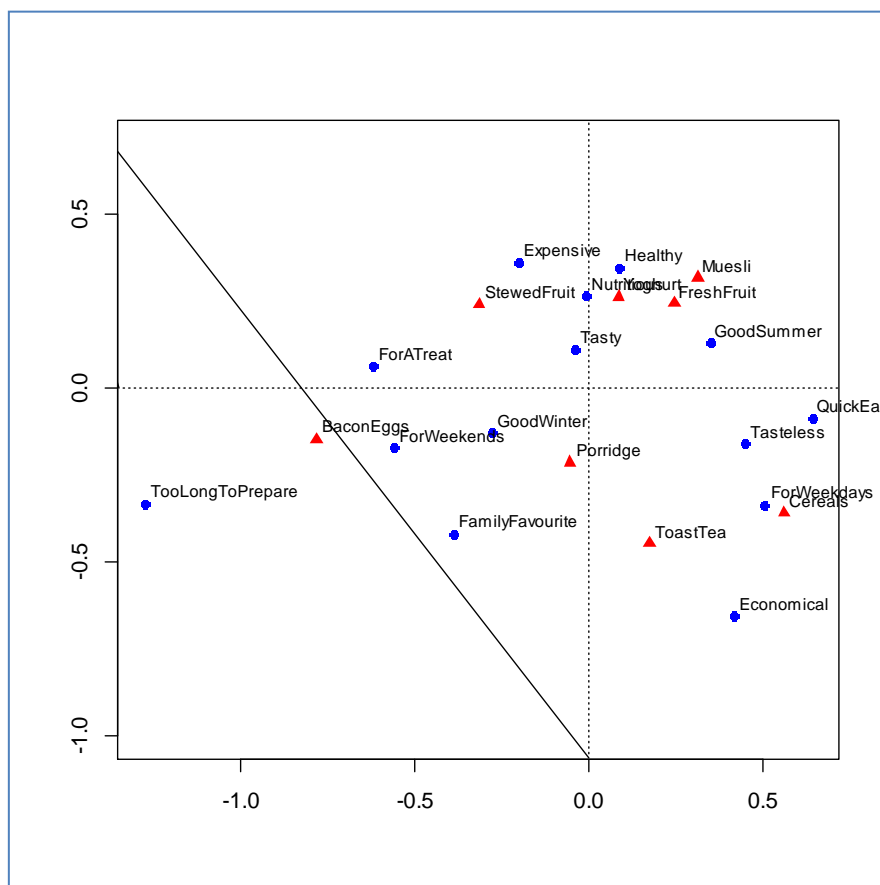


Figure 18 – Rows and columns points - 'ca' package

### 4.3 The 'ade4' package

The ade4 package is very popular with French searchers. Many tutorials and course materials are available online<sup>8</sup>.

<sup>8</sup> <http://pbil.univ-lyon1.fr/R/>

### 4.3.1 Row and column coordinates

We use the `dudi.coa()` procedure to perform the analysis.

```
library(ade4)
foods.coa <- dudi.coa(foods,scannf=F,nf=2)
#eigen values and cumulative proportion of variance explained in percentage
print(round(cbind(foods.coa$eig,100.0*cumsum(foods.coa$eig)/sum(foods.coa$eig)),4))
#row analysis - coordinates and contributions
print(cbind(foods.coa$li,inertia.dudi(foods.coa,row.inertia=T)$row.abs))
#column analysis - coordinates and contributions
print(cbind(foods.coa$co,inertia.dudi(foods.coa,col.inertia=T)$col.abs))
```

We obtain the same results as the other tools.

```
> library(ade4)
> foods.coa <- dudi.coa(foods,scannf=F,nf=2)
>
> #eigen values and cumulative proportion of variance explained in percentage
> print(round(cbind(foods.coa$eig,100.0*cumsum(foods.coa$eig)/sum(foods.coa$eig)),4))
      [,1]      [,2]
[1,] 0.1931  52.5012
[2,] 0.0777  73.6357
[3,] 0.0439  85.5593
[4,] 0.0328  94.4786
[5,] 0.0123  97.8111
[6,] 0.0057  99.3575
[7,] 0.0024 100.0000
>
> #row analysis - coordinates and contributions
> print(cbind(foods.coa$li,inertia.dudi(foods.coa,row.inertia=T)$row.abs))
      Axis1      Axis2 Axis1 Axis2
Healthy   -0.086620542 -0.34591187   43  1714
Nutritious  0.008607025 -0.26886266    0   993
GoodSummer -0.348785262 -0.13275108  655  236
GoodWinter  0.277410291  0.12712635  276  144
Expensive   0.200671515 -0.36124906  132 1059
QuickEasy  -0.643652755  0.08475682  2024  87
Tasty       0.039631673 -0.11273214    9   177
Economical -0.416925032  0.65686751   425 2618
ForATreat   0.619461485 -0.06491653  937   26
ForWeekdays -0.506754024  0.33798769 1020 1127
ForWeekends  0.560048658  0.17123472 1366  317
Tasteless  -0.450957714  0.15778118  126   38
TooLongToPrepare 1.277778139  0.33400636 2642  449
FamilyFavourite 0.387842351  0.42183368  345 1015
>
> #column analysis - coordinates and contributions
> print(cbind(foods.coa$co,inertia.dudi(foods.coa,col.inertia=T)$col.abs))
      Comp1      Comp2 Comp1 Comp2
Cereals   -0.56271870  0.3579056 2553 2566
Muesli    -0.31310163 -0.3186912  663 1708
Porridge   0.05362789  0.2130951   16  627
BaconEggs  0.78344380  0.1481415 5635  500
ToastTea  -0.17534202  0.4470236  101 1636
FreshFruit -0.24619292 -0.2449293  389  956
StewedFruit 0.31539949 -0.2428134  594  875
Yoghurt   -0.08598906 -0.2641586   48 1132
```

Figure 19 – Results provided by the 'ade-4' package

### 4.3.2 Symmetric graph

“Ade4” provides several graphical tools. We use the following command to obtain the symmetric graph (the option “method = 1” is the right option here).

```
#plotting rows and columns
scatter.coa(foods.coa,method=1)
```

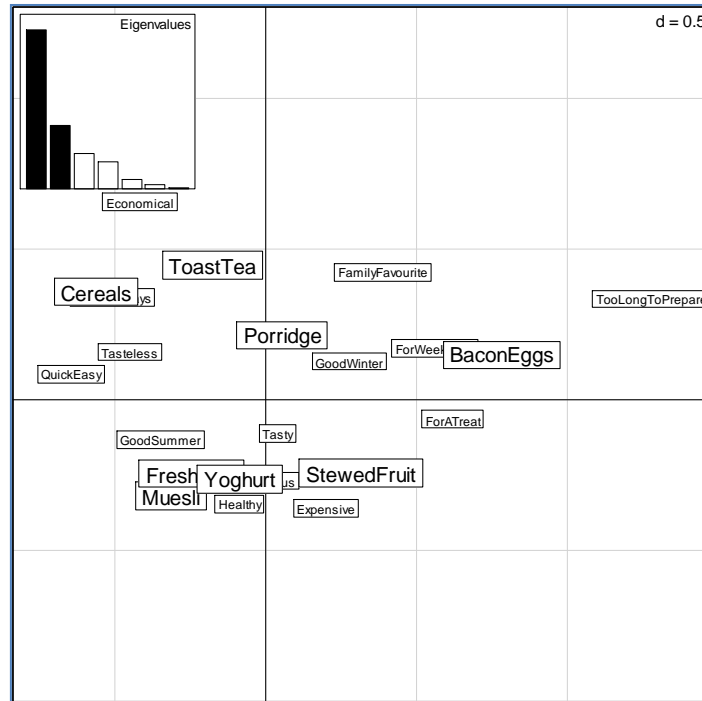


Figure 20 – Symmetric graph - Package 'ade4'

### 4.3.3 Visualizing the influence of the rows (columns)

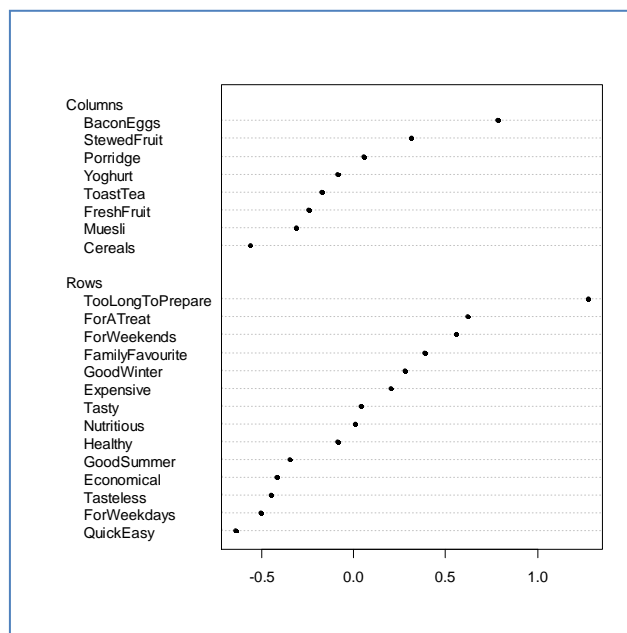


Figure 21 – Visualizing the influence of the rows (columns) - Package 'ade4'

Tanagra sorts the rows (columns) according to their contribution in order to provide a better visualization of their influence on the determination of the factors. Ade4 proposes a similar tool. The rows (columns) are sorted according their coordinate on each factor. Here we show the results for the first factor (Figure 21).

```
#canonical graph
```

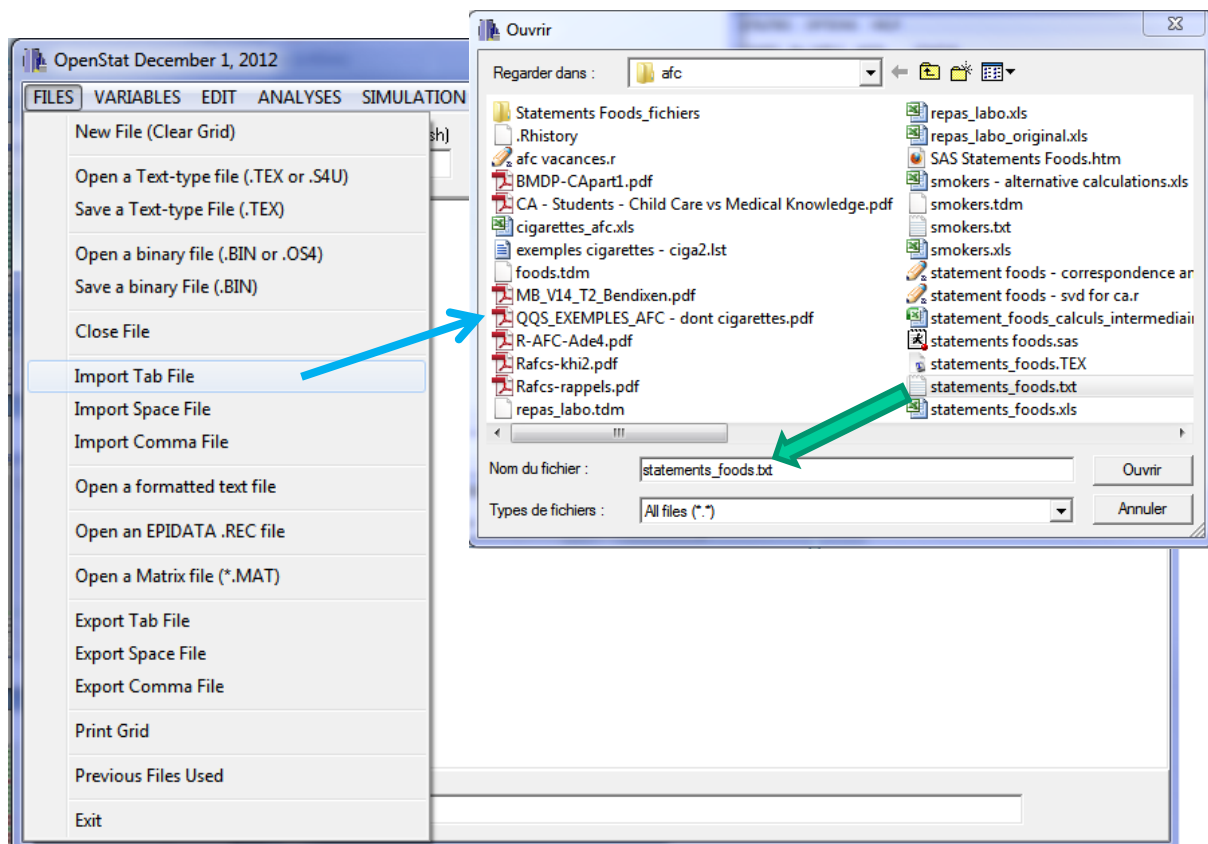
```
score.coa(foods.coa,xax=1,dotchart=T)
```

## 5 Correspondence analysis with OpenStat

OpenStat is a tool developed by William Miller, available on the web for many years. Its author has put online a large number of documentations (texts, videos) with the datasets used to illustrate the statistical methods. OpenStat is driven by menu. It is easy to handle. In this tutorial, we show only the main steps which enables to obtain the results highlighted in the previous sections.

### 5.1 Importing the data file

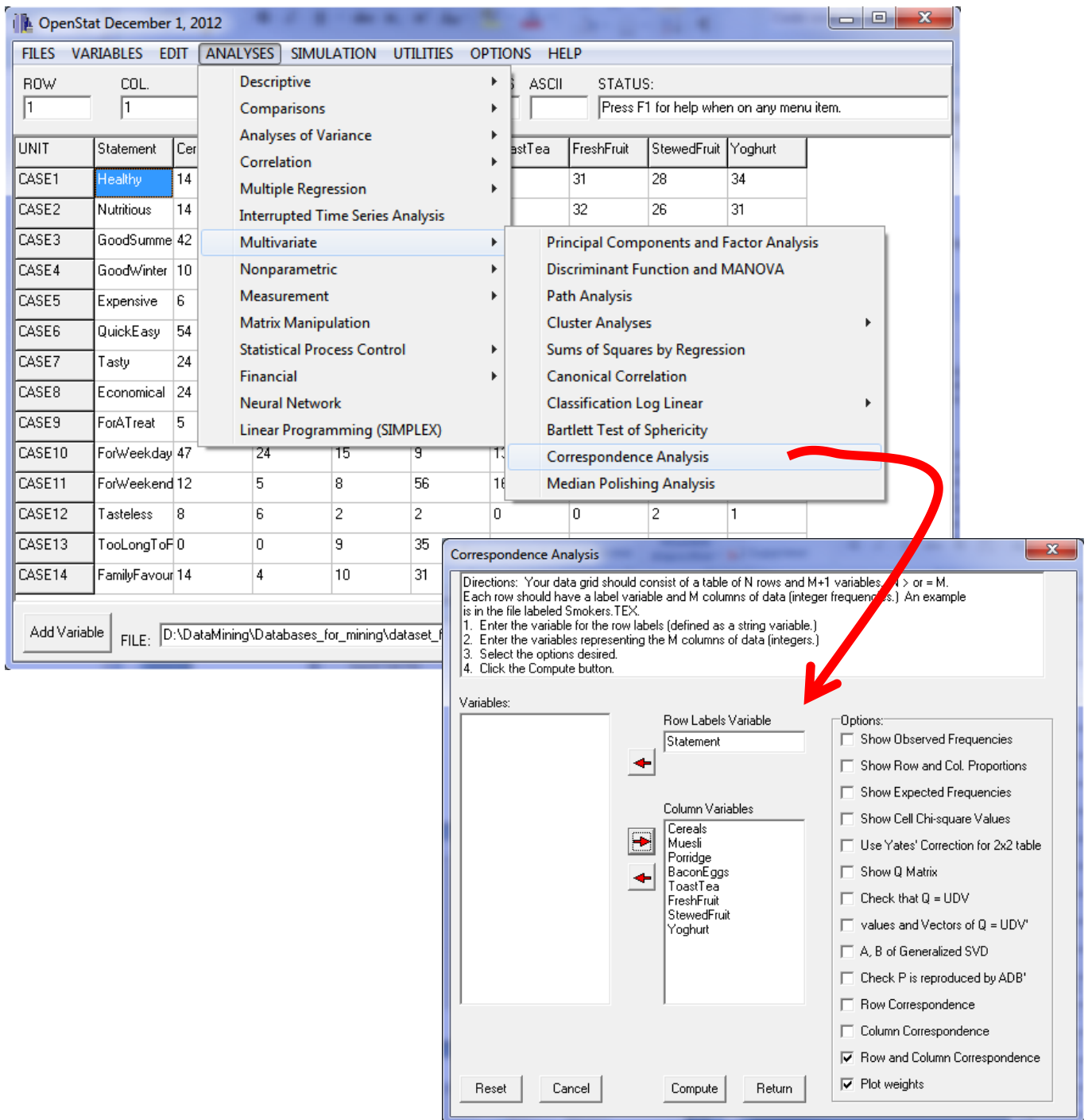
We launch OpenStat ([December 01, 2012](#)). We click on the FILES / IMPORT TAB FILE menu to load the “statements\_foods.txt” data file.



The raw data appears in the data grid.

### 5.2 Settings of the correspondence analysis

We click on the ANALYSES / MULTIVARIATE / CORRESPONDENCE ANALYSIS menu. The column STATEMENT corresponds to the label of the rows of the contingency table. The other variables (CEREALS ... YOGHOURT) correspond to its columns.



### 5.3 Reading the results

We click on the COMPUTE button to launch the calculations. Several Windows will then appear sequentially, describing various aspects of the results. To switch from one window to the other, we must click on the RETURN button at the top right.

#### 5.3.1 Chi-square test

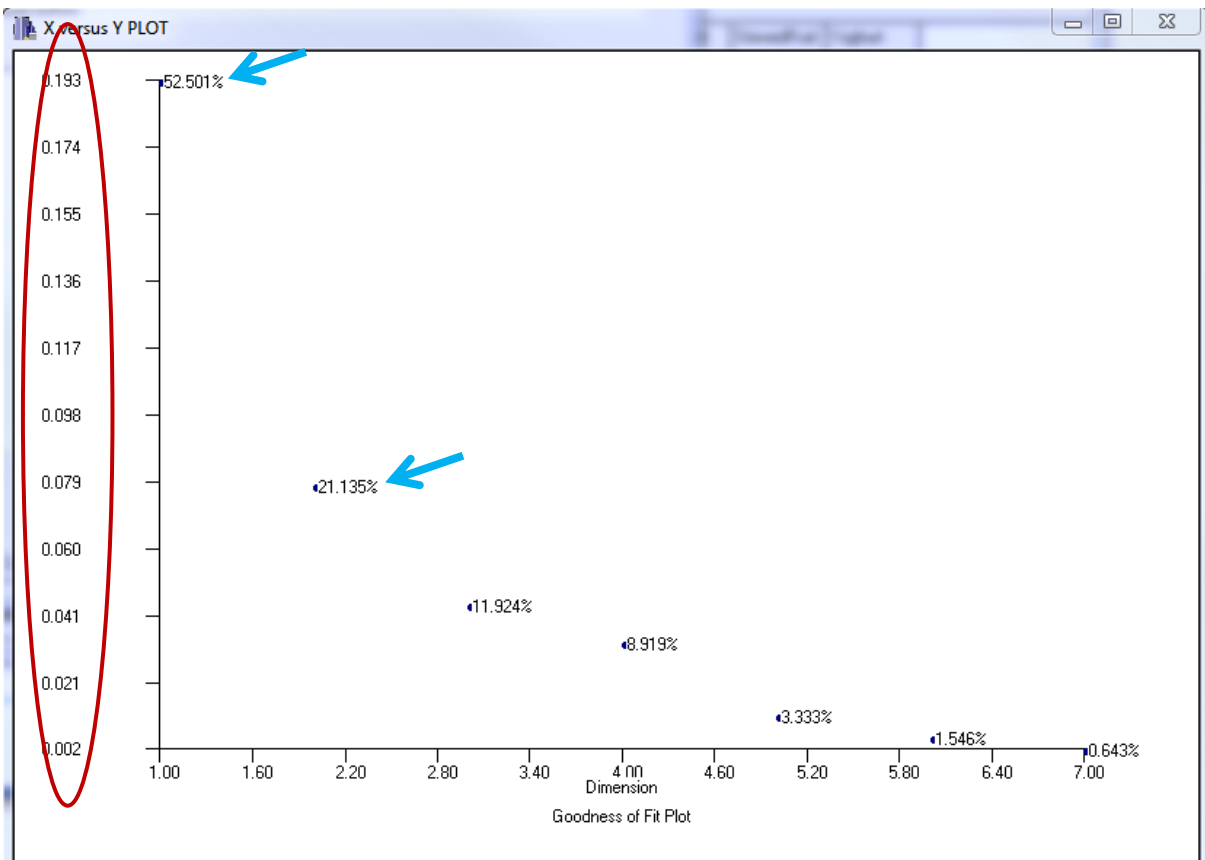
OpenStat provides several test statistics. Among other things, we have the standard  $\chi^2 = 647.312$ ; the squared of the phi coefficient  $\phi^2 = (0.6065)^2 = 0.3678$  which corresponds to the total inertia.

```

Chi-square = 647.312 with D.F. = 91. Prob. > value = 0.000
Likelihood Ratio = 652.128 with prob. > value = 0.0000
phi correlation = 0.6065
Pearson Correlation r = -0.0888
Mantel-Haenszel Test of Linear Association = 13.879 with probability > value = 0.0002
The coefficient of contingency = 0.519
Cramer's V = 0.229
    
```

### 5.3.2 Scree plot

OpenStat provides the scree plot (« Goodness of fit plot »). The points are labeled with the proportion of explained variance associated each factor.



### 5.3.3 Row points

OpenStat incorporates the trivial solution in the representation of the rows, specifying however that we must ignore it. It displays the coordinates of the rows for all factors. We find the results of the other tools. **Note:** the headers of each column in the table correspond to the factor number and not to the columns of the initial contingency table.



Results Window

Row Dimensions

	(Ignore Column 1)	Factor.1	Factor.2	BaconEggs	ToastTea	FreshFruit
Healthy	1.000	0.087	-0.346	0.173	-0.036	0.051
Nutritious	1.000	-0.009	-0.269	0.177	0.039	-0.027
GoodSummer	1.000	0.349	-0.133	-0.085	0.226	-0.146
GoodWinter	1.000	-0.277	0.127	0.457	-0.163	-0.066
Expensive	1.000	-0.201	-0.361	-0.219	-0.407	0.122
QuickEasy	1.000	0.644	0.085	-0.205	-0.019	0.011
Tasty	1.000	-0.040	-0.113	-0.002	0.121	-0.028
Economical	1.000	0.417	0.657	0.382	0.100	0.202
ForATreat	1.000	-0.619	-0.065	-0.217	0.229	0.097
ForWeekdays	1.000	0.507	0.338	-0.121	-0.179	0.008
ForWeekends	1.000	-0.560	0.171	-0.177	0.170	0.167
Tasteless	1.000	0.451	0.158	-0.306	-0.588	-0.127
TooLongToPrepare	1.000	-1.278	0.334	-0.046	-0.204	-0.186
FamilyFavourite	1.000	-0.388	0.422	-0.139	-0.017	-0.230

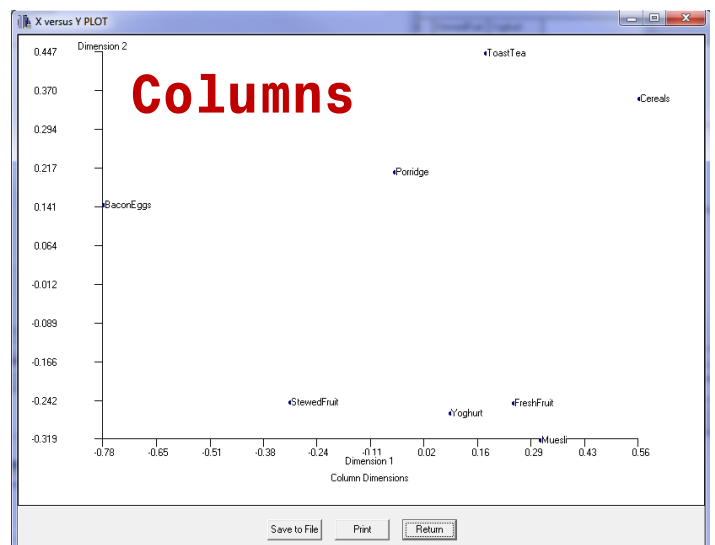
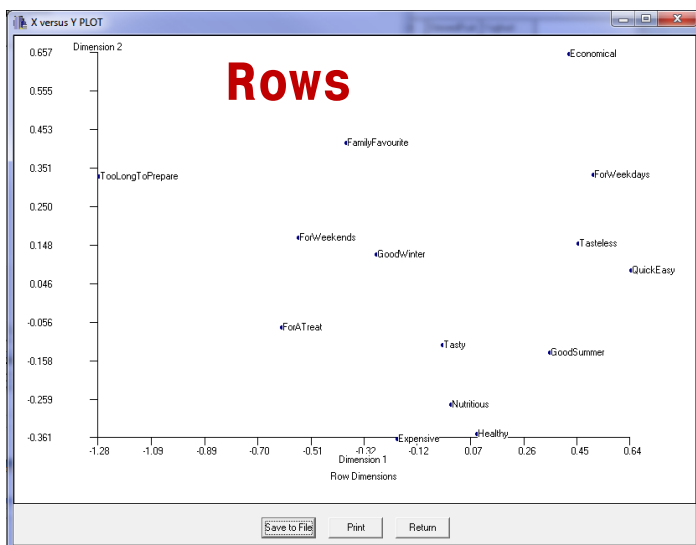
(Ignore Column 1)

	StewedFruit	Yoghurt
Healthy	0.004	-0.049
Nutritious	0.019	-0.023
GoodSummer	-0.038	0.001
GoodWinter	-0.063	0.047
Expensive	0.069	-0.007
QuickEasy	0.035	0.028
Tasty	0.051	0.080
Economical	0.027	-0.034
ForATreat	-0.204	-0.065
ForWeekdays	-0.051	-0.003
ForWeekends	0.030	0.026
Tasteless	-0.326	0.009
TooLongToPrepare	-0.009	0.056
FamilyFavourite	0.148	-0.127

We have the same presentation for the column points.

### 5.3.4 Graphical representations

OpenStat uses two separate charts for the representation of the rows and the columns. But we can very easily put them in parallel since we have the appropriate coordinates for a simultaneous representation.



## 6 Conclusion

The correspondence analysis is a very popular technique. It enables to quickly inspect large contingency tables by highlighting the most salient relationships.

We note it again in this tutorial. The various tools provide the same numerical results. From this point of view - I will never repeat it enough - there are no good or bad software. But they highlight with more or less intensity different aspects of the analysis. To fully appreciate the content of the results, we must understand the ins and outs of the underlying statistical method.

This comment is valid for R. Several packages enable to handle the same problem. The difference lies in the available documentation which allows us to easy use the tool. Ade4, offering many tutorials, is very interesting in this aspect.

## 7 References (French references)

L. Lebart, M. Piron, A. Morineau, « Statistique Exploratoire Multidimensionnelle », Dunod, 2000.

G. Saporta, « Probabilités, Analyse des Données et Statistique », Dunod, 2006.

M. Tenenhaus, « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.