Small excerpts adapted from “Exploring textual data”, Ludovic Lebart, André Salem, Lisette Berry, Kluwer Academic Publishers, 1998.

Correspondence analysis - a simple numerical example

Table 1 is a frequency table. In this table the *14* rows are words used in responses to an open-ended question given by *2000* respondents[1]. The five columns are the educational levels[2] of the respondents. Recall that the statistical unit here is not the respondent but the occurrence of a word. The columns constitute a partitioning of the set of respondents but the rows do not: a single respondent can use several words in the list within his or her response. The rows constitute a partitioning of the set of occurrences of the words.

Table 1 Cross-tabulation of words with Educational level. Raw frequencies

	No	Elem.	Trade	High	Coll	Total
Words	degree	Sch.	Sch.	Sch.	ege	
Money	*51*	*64*	*32*	*29*	*17*	*193*
Future	*53*	*90*	*78*	*75*	*22*	*318*
Unemployment	*71*	*111*	*50*	*40*	*11*	*283*
Decision	*1*	*7*	*5*	*5*	*4*	*22*
Difficult	*7*	*11*	*4*	*3*	*2*	*27*
Economic	*7*	*13*	*12*	*11*	*11*	*54*
Selfishness	*21*	*37*	*14*	*26*	*9*	*107*
Occupation	*12*	*35*	*19*	*6*	*7*	*79*
Finances	*10*	*7*	*7*	*3*	*1*	*28*
War	*4*	*7*	*7*	*6*	*2*	*26*
Housing	*8*	*22*	*7*	*10*	*5*	*52*
Fear	*25*	*45*	*38*	*38*	*13*	*159*
Health	*18*	*27*	*20*	*19*	*9*	*93*
Work	*35*	*61*	*29*	*14*	*12*	*151*
Total	*323*	*537*	*322*	*285*	*125*	*1592*

The table is read as follows: the word *Money* for example, was used *51* times by persons belonging to the category "no degree". The row totals represent the number of occurrences of each word whereas the column totals represent the total number of words (within the list) used by the various categories of respondents.

Table 2 shows the row-profiles expressed as percentages; they are obtained by dividing each element of the table by the corresponding row-sum: the row-profile of the word *Fear*, for example, is obtained by dividing each number in that row by *159*.

By comparing two row-profiles, we learn how the words represented by these two profiles are associated with the categories (columns)[3].

This comparison is more difficult if we use table 1 alone, because the frequencies of the words vary a lot. Thus it is not immediately obvious in table 1 that *Decision* is used relatively often by college graduates. But this is easy to see in table 2.

[1] "*What are the reasons that might cause a couple or a woman to hesitate having children ?*" Survey about aspirations and life styles of the French.

[2] No degree, Elementary School degree, Trade School degree, High School degree, College degree.

[3] To improve the readability of this table, the numbers have been multiplied by *100*.

Table 2
Row-profiles of table 1

	No	Elem.	Trade	High	Coll	Total
Words	degree	Sch.	Sch.	Sch.	ege	
Money	*26.4*	*33.2*	*16.6*	*15.0*	*8.8*	*100.0*
Future	*16.7*	*28.3*	*24.5*	*23.6*	*6.9*	*100.0*
Unemployment	*25.1*	*39.2*	*17.7*	*14.1*	*3.9*	*100.0*
Decision	*4.5*	*31.8*	*22.7*	*22.7*	*18.2*	*100.0*
Difficult	*25.9*	*40.7*	*14.8*	*11.1*	*7.4*	*100.0*
Economic	*13.0*	*24.1*	*22.2*	*20.4*	*20.4*	*100.0*
Selfishness	*19.6*	*34.6*	*13.1*	*24.3*	*8.4*	*100.0*
Occupation	*15.2*	*44.3*	*24.1*	*7.6*	*8.9*	*100.0*
Finances	*35.7*	*25.0*	*25.0*	*10.7*	*3.6*	*100.0*
War	*15.4*	*26.9*	*26.9*	*23.1*	*7.7*	*100.0*
Housing	*15.4*	*42.3*	*13.5*	*19.2*	*9.6*	*100.0*
Fear	*15.7*	*28.3*	*23.9*	*23.9*	*8.2*	*100.0*
Health	*19.4*	*29.0*	*21.5*	*20.4*	*9.7*	*100.0*
Work	*23.2*	*40.4*	*19.2*	*9.3*	*7.9*	*100.0*
Total	*20.3*	*33.7*	*20.2*	*17.9*	*7.9*	*100.0*

Table 3 shows column-profiles: these are obtained in analogous fashion by dividing the elements of each column by their sum and multiplying the result by *100*.

Table 3
Column-profiles of table 1

	No	Elem.	Trade	High	Coll-	Total
Words	degree	Sch.	Sch.	Sch.	ege	
Money	*15.8*	*11.9*	*9.9*	*10.2*	*13.6*	*12.1*
Future	*16.4*	*16.8*	*24.2*	*26.3*	*17.6*	*20.0*
Unemployment	*22.0*	*20.7*	*15.5*	*14.0*	*8.8*	*17.8*
Decision	*.3*	*1.3*	*1.6*	*1.8*	*3.2*	*1.4*
Difficult	*2.2*	*2.0*	*1.2*	*1.1*	*1.6*	*1.7*
Economic	*2.2*	*2.4*	*3.7*	*3.9*	*8.8*	*3.4*
Selfishness	*6.5*	*6.9*	*4.3*	*9.1*	*7.2*	*6.7*
Occupation	*3.7*	*6.5*	*5.9*	*2.1*	*5.6*	*5.0*
Finances	*3.1*	*1.3*	*2.2*	*1.1*	*.8*	*1.8*
War	*1.2*	*1.3*	*2.2*	*2.1*	*1.6*	*1.6*
Housing	*2.5*	*4.1*	*2.2*	*3.5*	*4.0*	*3.3*
Fear	*7.7*	*8.4*	*11.8*	*13.3*	*10.4*	*10.0*
Health	*5.6*	*5.0*	*6.2*	*6.7*	*7.2*	*5.8*
Work	*10.8*	*11.4*	*9.0*	*4.9*	*9.6*	*9.5*
Total	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*

By comparing two column-profiles we learn about similarities that exist among the various educational categories with respect to vocabulary used.

. . . .

Reduction of dimensionality

Correspondence analysis and principal components analysis are used under different circumstances: principal components analysis (that can be viewed as a purely descriptive variant of factor analysis) is used for tables consisting of continuous measurements. Correspondence analysis is best adapted to contingency tables (cross-tabulations). By extension, correspondence analysis also provides a satisfactory description of tables with binary coding.

Most of these methods provide the user with a sequence of nested subspaces. That means that the best one dimensional subspace (a straight line) is included in the best two-dimensional subspace (a plane) which in turn is included in the best three-dimensional subspace, etc.. In such a series of nested subspaces, the two-dimensional one is a very special case, since it is compatible with our most usual communication devices such as sheets of paper or video screens.

As a consequence, the following sections mainly deal with two-dimensional displays, which rarely give an exact representation of the distances between profiles, but are easier to inspect.

An example of a two-dimensional map

Figure 1 is a two-dimensional display generated by a correspondence analysis of table 1. It is a visual representation, or map, which simultaneously describes the similarities among row-profiles and similarities among column-profiles.

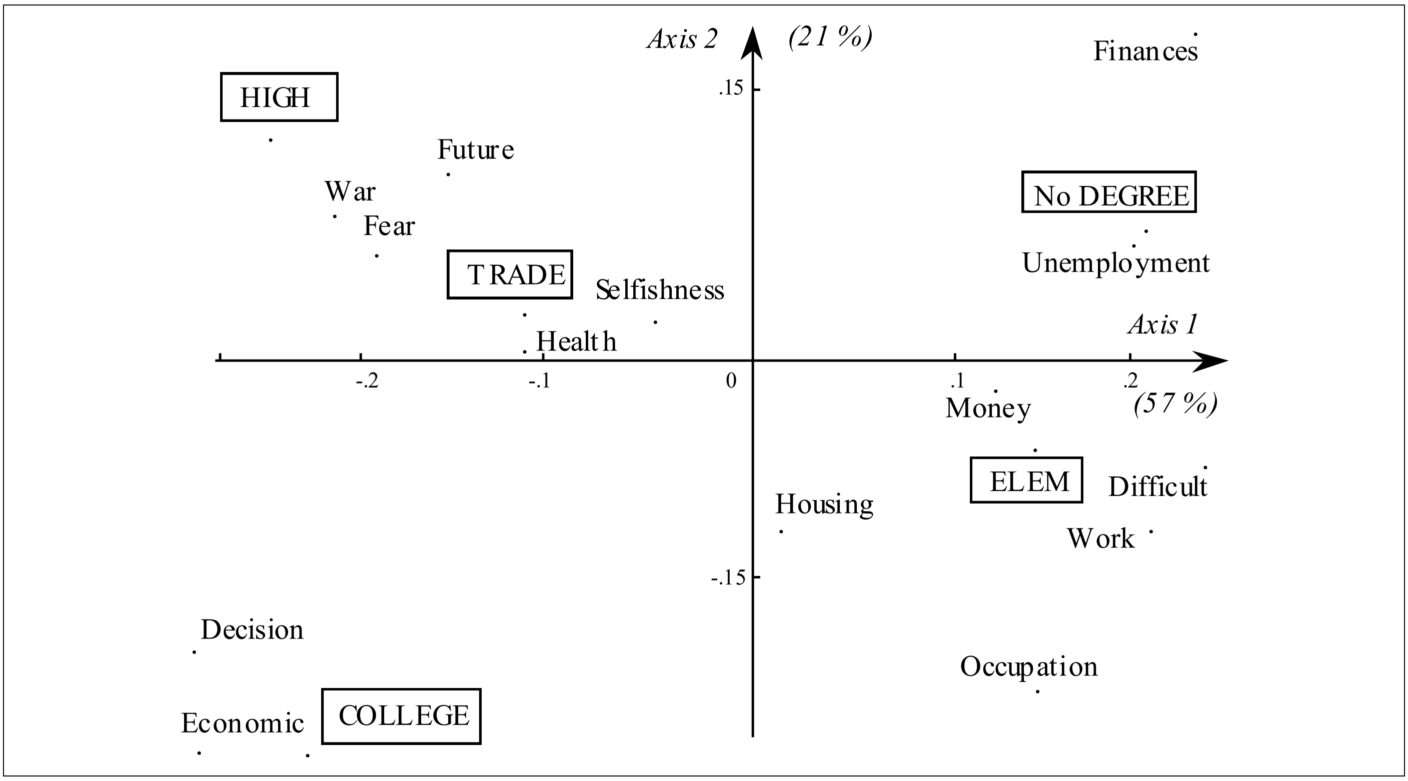

Figure 1 Proximities among words and among educational levels
Correspondence analysis of table 1 . *(the boxed elements correspond to the columns of table 1)*

For someone who has a good grasp of the rules for interpreting such a map, this is a quick process for assimilating information. However one should not expect any major surprises or revelations from a graphical representation of a table of such modest dimensions. In this case correspondence analysis has a purely descriptive role: it simply makes results easier to look at.

Let us show, with the help of this example, how simple the principles and rules for interpretation of the method are.

How does one read figure 1 ?

If two row-points *i* and *i'* have identical or similar profiles, they appear in exactly, or almost the same, position on each of the principal axes. In analogous fashion if two column-points *j* and *j'* have identical or similar profiles, they are in the same position or very close.

The origin of the axes represents the mean profiles (i.e., the marginals of the table of frequencies).

The components of the mean row-profile are: $f_{.j}$, $j=1,\ldots p$

The components of the mean column-profile are: $f_{i.}$, $i=1,\ldots n$

Thus, a column-point (boxed elements) such as ***Trade School***, which is rather close to the origin, has a similar profile to that of the *Total* column (the vertical marginal) of table 3. Similarly, a row-point such as *Health* has a profile that is similar to the *Total* row (horizontal marginal) which is the last row of table 2.

Why a simultaneous representation?

We now know how to interpret the proximity between two row-points or two column-points, as well as their respective positions relative to the origin of the axes.

But figure 1 shows us row-points and column-points simultaneously, and thereby displays additional proximities that we are tempted to interpret: it is not surprising that the row-point *Unemployment* should be close to the column-point ***NoDegree***. But the proximity between *Health* and ***Trade School*** is less obvious.

As a matter of fact, it is not possible to interpret these cross-proximities between a row-point and a column-point, because these two points do not come from the same initial space. Nevertheless, it is possible to interpret the position of a single row-point with respect to the set of column-points or of a single column-point with respect to the set of row-points.

Active and supplementary variables

Correspondence analysis is used for finding subspaces to represent proximities among profiles. But it can also be used for positioning *supplementary* rows and columns of the data matrix in this subspace.

The elements or variables used to calculate the two dimensional display are called *active elements* or *active variables*. These elements or variables must be a homogeneous set in order for the distances among individuals or observations to make sense, and therefore the observed graphical proximities to be interpretable. Elements or variables that are projected *a posteriori* on the two dimensional display are the *supplementary* or *illustrative* elements. It is not necessary for these illustrative elements (rows or columns) to constitute a homogeneous set. The computation is executed separately for each one. This dichotomy between active variables and illustrative variables is fundamental from a methodological viewpoint.

Example

The distributions of four words having small overall counts are shown on table 4. These words were not part of the preceding analysis.

Table 4

Four supplementary (illustrative) rows

	No	Elem.	Trade	High	Coll	Total
Words	degree	Sch.	Sch.	Sch.	ege	
Comfort	*2*	*4*	*3*	*1*	*4*	*14*
Disagreement	*2*	*8*	*2*	*5*	*2*	*19*
World	*1*	*5*	*4*	*6*	*3*	*19*
Live	*3*	*3*	*1*	*3*	*4*	*14*

We would like to see where they are situated with respect to the other words represented on the two dimensional display of figure 2. Their row-profiles can be positioned in the same 5 dimensional space and can therefore be projected onto the plane of figure 1.

Table 5

Three supplementary (illustrative) columns

Word	Age-30	Age-50	Age+50
Money	*59*	*66*	*70*
Future	*115*	*117*	*86*
Unemployment	*79*	*88*	*177*
Decision	*9*	*8*	*5*
Difficult	*2*	*17*	*18*
Economic	*18*	*19*	*17*
Selfishness	*14*	*34*	*61*
Occupation	*21*	*30*	*28*
Finances	*8*	*12*	*8*
War	*7*	*6*	*13*
Housing	*10*	*27*	*17*
Fear	*48*	*59*	*52*
Health	*13*	*29*	*53*
Work	*30*	*63*	*58*
Total	*433*	*575*	*663*

In analogous fashion, table 5 contains three supplementary columns (age categories) that were not included in the set of active columns due to the heterogeneous nature of the themes: interpreting the proximities among rows, and thus among words, would have been more difficult.

Are two words close because of their distribution with respect to educational level or with respect to age categories? This type of decision is not easy to make if the distances among words are calculated on the basis of both variables simultaneously.

Figure 2 below shows us that three supplementary words are relatively closely associated with the responses of persons with a higher degree, whereas the fourth, *disagreement*, is less characteristic, being closer to the center that represents the mean profile.

It is good practice to start by using, for active tables, homogeneous data sets that describe proximities from a single point of view.

The representation can then be enriched by illustrating it with supplementary information.

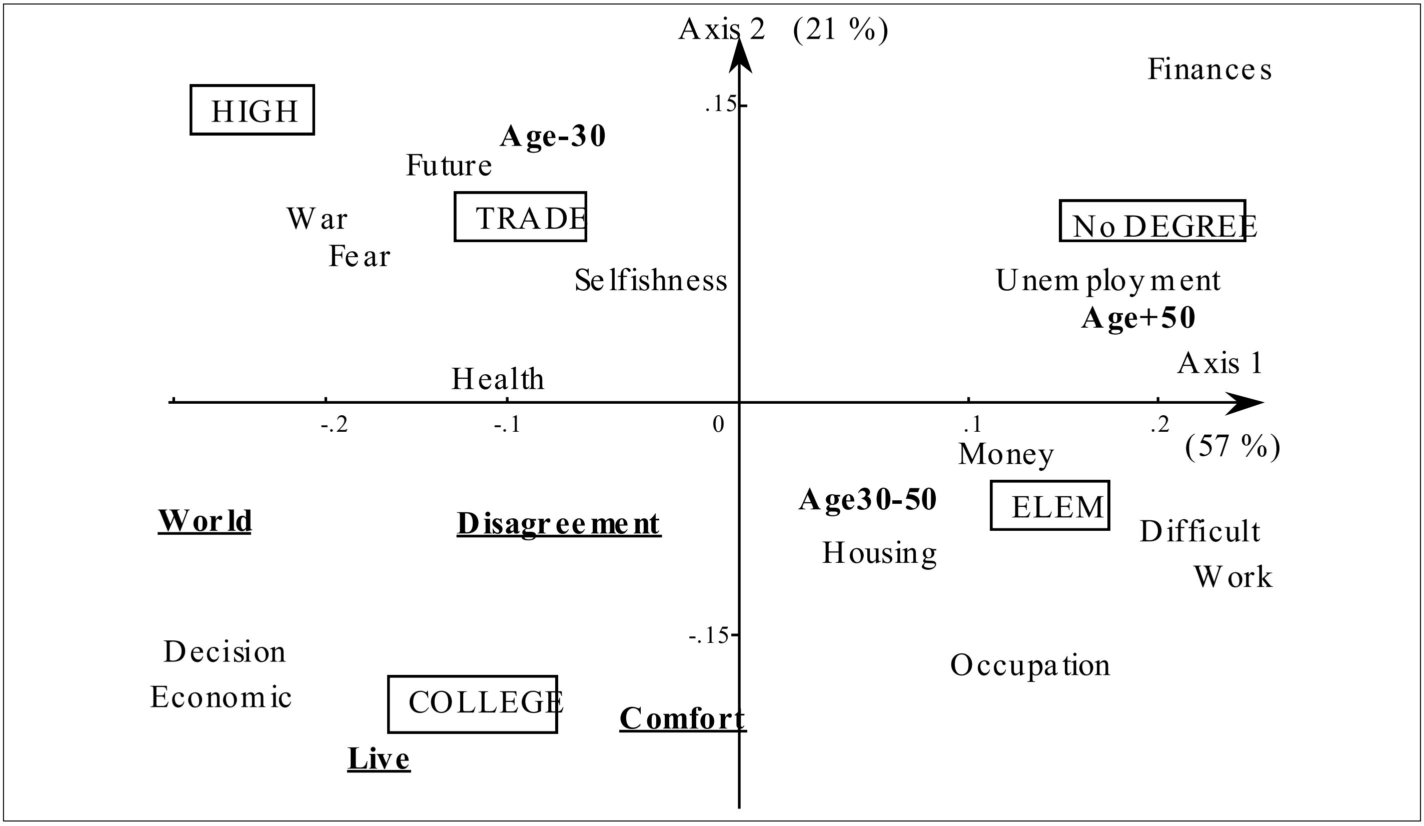

Figure 2

Associations among words and Educational Level (*continuation*)
Positioning of illustrative elements (in bold) in plane of figure 1

On figure 2 the three age categories are ordered along the horizontal axis just as the educational levels are: increasing age groups correspond to decreasing educational levels. This is a structural trait of the population under analysis: the younger respondents have more schooling, and this complicates interpretation in terms of causality. One is led to wonder whether the effect of Educational Level on open-ended responses can be separated from the effect of age.