

Benko, Michal; Lejeune, Michel

Working Paper

Correspondence analysis

SFB 373 Discussion Paper, No. 2000,64

Provided in Cooperation with:

Collaborative Research Center 373: Quantification and Simulation of Economic Processes,
Humboldt University Berlin

Suggested Citation: Benko, Michal; Lejeune, Michel (2000) : Correspondence analysis, SFB 373 Discussion Paper, No. 2000,64, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin,
<https://nbn-resolving.de/urn:nbn:de:kobv:11-10047907>

This Version is available at:

<https://hdl.handle.net/10419/62249>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Correspondence Analysis

Michal Benko and Michel Lejeune

Correspondence analysis (CA) is a descriptive method which allows us to analyze and to XploRe the structure of contingency tables (or, by extension, non-negative tables where rows and columns are the entities of interest). It is similar to **principal component analysis** (PCA) in the sense that it attempts to obtain a representation of either the I row items or the J column items in a low dimensional space, while preserving at best total variation in the table.

1 Introduction

In **contingency table**, the data are classified according to each of two characteristics. The attributes on each characteristic are represented by the row and the column categories. We will denote by n_{ij} the number of individuals with the i -th row and j -th column attributes. The contingency table itself is the $(I \times J)$ matrix containing the elements n_{ij} .

1.1 Singular Value Decomposition

Total variation in the contingency table is measured by departure from independence, i.e., more precisely, by the χ^2 statistic

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J (n_{ij} - E_{ij})^2 / E_{ij},$$

where n_{ij} , $i = 1, \dots, I$, $j = 1, \dots, J$ are the observed frequencies and E_{ij} is the estimated expected value in the cell (i, j) under the assumption of independence

$$E_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}.$$

We define

$$M = (n_{ij} - n_{i\bullet}n_{\bullet j}/n).$$

The matrix M contains the differences between the observed frequency and the frequency estimated under assumption of independence.

The χ^2 statistic which measures the departure of independence can be rewritten as

$$ntr(M^T R M C),$$

where $R = diag(1/n_{i\bullet})$ and $C = diag(1/n_{\bullet j})$.

The CA itself consists of finding the **singular value decomposition** (SVD) of the matrix $R^{1/2} M C^{1/2}$. In this way, we obtain approximations of the matrix $R^{1/2} M C^{1/2}$ by matrices of lower rank:

$$M = (g_1)^{-1/2} r_1 c_1^T + (g_2)^{-1/2} r_2 c_2^T + \dots + (g_u)^{-1/2} r_u c_u^T,$$

where $(g_1)^{-1/2} r_1 c_1^T$ is the matrix of rank one closest to M in the chi-square norm, $(g_1)^{-1/2} r_1 c_1^T + (g_2)^{-1/2} r_2 c_2^T$ is the matrix of rank two closest to M in the chi-square norm and so on. The g_k 's are the eigenvalues of $M^T R M C$ in decreasing order and $c_k^T c_k = r_k^T r_k = g_k$.

1.2 Coordinates of Factors

The $I \times 1$ vector r_k , defines the coordinates of the rows corresponding to the k -th factor. Similarly, the $J \times 1$ vector c_k defines the coordinates of columns corresponding to the k -th factor.

A set of u coordinates for row (resp. column) items, where $u = \min(I, J) - 1$ is hierarchically constructed via singular value decomposition. Thus the construction is similar to the PCA, however with a different matrix norm in order to take into account the specific frequency nature of the data.

For the sake of simplicity, the vector of first row coordinates is called the first factor (as well as the vector of the first coordinates for columns), and so on up to the u -th factor.

2 XploRe Implementation

```
corresp(fadata{, fsldata, fscdata, titl, fal, fac, fsl,
                                             fsc, outdoc})
runs correspondence analysis on fadata
```

Notice that the only obligatory parameter is the name of the active data file `fadata`. The explanation of parameters is the following.

`fadata`

active data file, it is a $I \times J$ matrix containing analyzed cross table

`fsldata`

K rows supplementary data file, it is a $K \times J$ matrix (see the explanation of supplementary items in Subsection 3.6)

`fscdata`

Q columns supplementary data file, it is a $I \times Q$ matrix

`fal`

active row labels file used in biplot, it is a $I \times 1$ string vector

`fac`

active column labels file used in biplot, it is a $J \times 1$ string vector

`fsl`

K row supplementary labels file used in biplot, it is a $K \times 1$ string vector

`fsc`

Q column supplementary labels file used in biplot, it is a $Q \times 1$ string vector

3 Example: Eye-Hair

3.1 Description of Data


The data set given in Table 1 is a contingency table of **hair colors** (4 categories) and **eye colors** (4 categories) for 592 women (Lebart, L., Morineau, A., and Piron, M. 1995).

EYE \ HAIR COLOR	black	brown	red	blond	total
dark brown	68	119	26	7	220
light brown	15	54	14	10	93
green	5	29	14	16	64
blue	20	84	17	94	215
total	108	286	71	127	592

Table 1: Contingency table for eye-hair color data.

3.2 Calling the Quantlet

The following XploRe code explains how to run correspondence analysis using quantlet `corresp` in XploRe.

```
library("stats")
corresp("e.dat", "null", "null", "EYE-HAIR", "eltxt.dat",
        "ectxt.dat", "null", "null", "null")
 corre01.xpl
```

In this example, we use the active data file `e`. The file `e` contains the Hair-eye contingency table given in Table 1.

```
68 119 26 7
15 54 14 10
5 29 14 16
20 84 17 94
```


Row labels are given in the file `eltxt`:

```
dark-brown
light-brown
green
blue
```

Column labels are in the file `ectxt`:

BLACK
BROWN
RED
BLOND

3.3 Documentation of Results

The output of CA from  `corre01.xpl` is shown in the output window. In this example, we get altogether three factors—three eigenvalues and three coordinates for each row (column) item.

3.4 Eigenvalues

The eigenvalues g_1, g_2, \dots, g_u give the part of total variation recovered on the first, second, ... , u -th factors. They allow to make a choice for the number of factors (or axes, in the geometrical representation) to retain.

```
[1,] EIGENVALUES AND PERCENTAGES
Contents of seig
```

```
[1,]  0.2088  89.3727  89.3727
[2,]  0.0222   9.5149  98.8876
[3,]  0.0026   1.1124 100.0000
```

We see that already the first factor explains nearly 90% of total variation in this contingency table, equal to $(0.2088 + 0.0222 + 0.0026)592 = 138.3$

3.5 Contributions

3.5.1 Global Contributions of Rows (Resp. Columns)

From the formula of Pearson's chi-square (here divided by n) one can obviously decompose the total variation across row (resp. column) items additively. This yields the global row (resp. column) contributions to total variation. In the geometrical representation of row (resp. column) profiles in a u -dimensional Euclidean space—taking the marginal row (resp. column) profile as the origin—the global contribution of a row (resp. column) is equal to the squared distance to the origin times its relative weight (say $n_{i\bullet}/n$ for row i). The squared distance itself is useful to see how a row item deviates from what is expected under independence.

```
[1,] "Row relative weights and distances to the origin"
```

```
Contents of spdai
```

```
[1,] 0.3716 0.0206
[2,] 0.1571 0.0119
[3,] 0.1081 0.0159
[4,] 0.3632 0.0228
```

```
[1,] Column relative weights and distances to the origin
```

```
Contents of spdaj
```

```
[1,] 0.1824 0.0227
[2,] 0.4831 0.0066
[3,] 0.1199 0.0146
[4,] 0.2145 0.0345
```

3.5.2 Contributions of Rows Or Columns to a Factor

It is interesting to know how much each row (or column) contributes to the variation pertaining to a given factor. These specific contributions are useful to possibly interpret the factor in terms of contrasts between row (or column) items. These contributions are usually given in percents of total variation of the factor (i.e. corresponding eigenvalues).

[1,] Coordinates of the columns

Contents of scoordj

[1,]	-0.0207	-0.0088	0.0023
[2,]	-0.0061	0.0013	-0.0020
[3,]	-0.0053	0.0131	0.0034
[4,]	0.0343	-0.0029	0.0007

[1,] Contributions of the columns

Contents of scontrj

[1,]	22.2463	37.8774	21.6330
[2,]	5.0860	2.3194	44.2838
[3,]	0.9637	55.1305	31.9125
[4,]	71.7039	4.6727	2.1706

The coordinates of the first axis show that **blond** hair color (4-th column item) is opposed to all the other hair colors on the first axis, in particular, to **black** hair color (1-st column item). The first factor can be essentially explained by a strong contrast between **blond** and **black** hair in terms of eyes color (respective contribution 71,7% and 22,2%)

The second axis (its eigenvalue 9.5% is ten times smaller than that of the first axis of 89.4%, is mainly constructed by the item of hair color **red** (55.1%) as opposed to **black** hair color (37,9%). The third factor is accounting for negligible contribution to total variation (1.1 %).

[1,] Coordinates of the rows

Contents of scoordi

[1,]	-0.0202	-0.0036	0.0009
[2,]	-0.0087	0.0069	-0.0041
[3,]	0.0066	0.0139	0.0036
[4,]	0.0225	-0.0034	-0.0002

[1,] Contributions of the rows

Contents of `scontri`

[1,]	43.1157	13.0425	6.6796
[2,]	3.4010	19.8040	61.0856
[3,]	1.3549	55.9095	31.9248
[4,]	52.1284	11.2440	0.3100

For the row items, the first axis is, solely, constructed by eye colors **dark brown** (1-st row item) and **blue** (4-th row item) (resp. contributions of 43.1% and 52.1%). Coordinates show that they are opposed in terms of hair profile. The second axis is mainly due to **green** eye color (3-rd row item).

3.5.3 Squared Correlations

The global contribution of a given row (resp. column) itself may be additively decomposed across the u factors into terms called squared correlations (by analogy with PCA) when expressed in percents of that global contribution. Squared correlations are useful to determine how well each row (resp. column) variation is recovered on a factor or on restricted number of factors (or axes in a geometrical representation). This allows to guard against illusory proximities of points (row or column profiles) in mappings.

[1,] Squared correlations of the rows

Contents of `scorri`

[1,]	0.9670	0.0311	0.0019
[2,]	0.5424	0.3363	0.1213
[3,]	0.1759	0.7726	0.0516
[4,]	0.9775	0.0224	0.0001

[1,] Squared correlations of the columns

Contents of `scorrj`

[1,]	0.8380	0.1519	0.0101
[2,]	0.8644	0.0420	0.0937
[3,]	0.1333	0.8118	0.0549
[4,]	0.9927	0.0069	0.0004

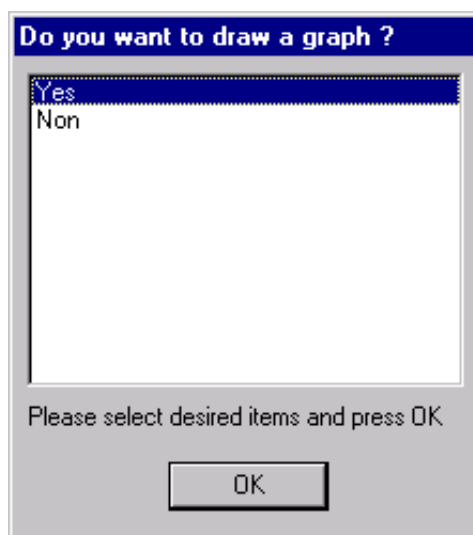
From these correlations it can be inferred, for instance, that factor 1 is exclusively specific for blond hair color.

3.6 Biplots

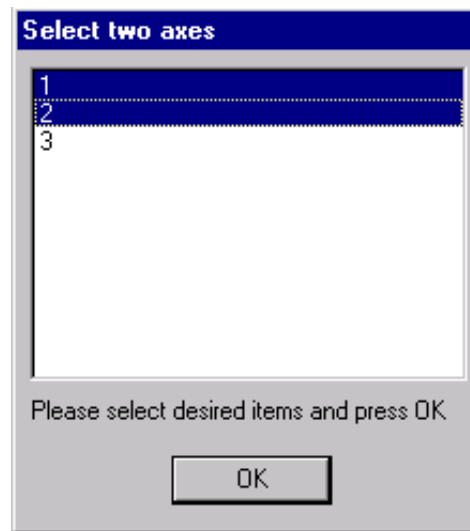
A simultaneous representation of row and column items in the same mapping has some interesting interpretational aspects. When row i and column j , say, are represented by points in the same (resp. opposite) direction with respect to the origin it means that n_{ij} is above (resp. below) the value expected according to independence (conditioned on the fact that the sum of their squared correlations on the first two factors is, for each of them, sufficiently high).

3.6.1 Asking for Graph

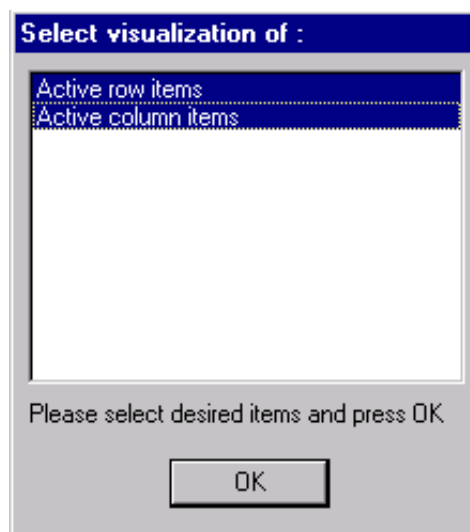
Results of the analysis can be visualized in different graphs :




We can visualize the configuration of the items in any two axes. The importance of the axes is proportional to the variation explained by this axis. It is measured by the eigenvalue. We can select any two axes for display. If $u > 5$ then the first five axes are available to choose from.



We can select different items to display in graphs :



The graph requested ( corre01.xpl) is shown in the Figure 1

The graph using the two first coordinates shows the suggestive features of simultaneous representation of row and column items in the same mapping. This allows us to interpret the proximities or distances between items of the same set with their associations to those of other item sets.

3.6.2 Supplementary Items

It is possible to project additional rows or columns onto the various factors without having these elements enter the construction of factors, as opposed to so-called active items. This may be useful for various reasons: to get some exogenous explanations of some features revealed in the data, to ignore a much too influential row or column item (in particular for items with low frequencies), to see the positions of several items forming a natural group, etc.

3.7 Brief Remark

Why is the position of the item of hair color **blond** more extreme than the eye color **blue** on the first dominant axis? Because the item of hair color **blond** is much more characterized by eye's color **blue** than the inverse fact: as can be seen from the data, 74% of **blond** people have **blue** eyes while only 44% of people with **blue** eyes have **blond** hair.

4 Example: Media

4.1 Description of the Data Set

The data set comes from a survey where 12,388 **contacts with various media** have been identified (Lebart, L., Morineau, A., and Piron, M. 1995). These contacts are crossed by **activities** (the statistical units are the media contacts). Besides, they are crossed with some supplementary variables: **sex**, **age** and **education level**.

The active data is stored in the file `m` which contains six items (columns) of media and eight activities (rows)

96	118	2	71	50	17
122	136	11	76	49	41

193	184	74	63	103	79
360	365	63	145	141	184
511	593	57	217	172	306
385	457	42	174	104	220
156	185	8	69	42	85
1474	1931	181	852	642	782

The column labels are stored in file `mctxt` as shown below

```
RADIO
TV
N_NEWS
R_NEWS
MAGAZ
TVMAG
```

The vector of row labels is stored in the file `mltxt`

```
la_Farmer
s_busin
h_manag
i_manag
empl
skil
unsk
Nowork
```

Supplementary row data are stored in the file `ms1`:


1630	1900	285	854	621	776
1667	2069	152	815	683	938
660	713	69	216	234	360
640	719	84	230	212	380
888	1000	130	429	345	466
617	774	84	391	262	263
491	761	70	402	251	245
908	1307	73	642	360	435
869	1008	107	408	336	494
901	1035	80	140	311	504
619	612	177	209	298	281

The eleven supplementary row labels are stored in the file `msltxt`:

```
MALE
FEMALE
A14-24
A25-34
A35-49
A50-64
A65+
PRIMARY
SECOND
H_TECH
UNIVER
```

4.2 Calling the Quantlet

The next code which calls the quantlet `corresp` and analyzes the dataset `m`.

```
library("stats")
corresp("m.dat", "msl.dat", "null", "MEDIA", "mltxt.dat",
        "mctxt.dat", "msltxt.dat", "null")
 corre02.xpl
```

4.3 Brief Interpretation

We obtain the following output.

```
[1,] EIGENVALUES AND PERCENTAGES
```

```
Contents of seig
```

[1,]	0.0139	62.1982	62.1982
[2,]	0.0072	32.3650	94.5632
[3,]	0.0008	3.7018	98.2650
[4,]	0.0003	1.3638	99.6288
[5,]	0.0001	0.3712	100.0000

The first two axes together account for 95% of total variation and are very dominant. This percentage gives an idea of the share of information accounted for by the first two principal axes.

Coordinates on different axes and other indices helpful for interpreting the results are shown in following output which also includes the coordinates and the squared correlations of supplementary items.

[1,] Row relative weights and distances to the origin

Contents of spdai

[1,]	0.0286	0.0032
[2,]	0.0351	0.0016
[3,]	0.0562	0.0039
[4,]	0.1015	0.0011
[5,]	0.1498	0.0009
[6,]	0.1116	0.0011
[7,]	0.0440	0.0014
[8,]	0.4732	0.0005

[1,] Coordinates of the rows

Contents of scoordi

[1,]	-0.0015	-0.0028	0.0006	0.0001	-0.0002
[2,]	-0.0006	-0.0013	0.0006	-0.0002	0.0002
[3,]	0.0039	-0.0005	0.0000	-0.0002	-0.0001
[4,]	0.0010	0.0003	0.0003	0.0002	0.0001
[5,]	-0.0001	0.0009	0.0000	0.0002	0.0000
[6,]	-0.0004	0.0009	0.0002	-0.0003	0.0000
[7,]	-0.0011	0.0009	0.0004	0.0000	-0.0002
[8,]	-0.0003	-0.0003	-0.0002	0.0000	0.0000

In the following window we remark, for instance, that the relative frequency of **national newspapers** (N NEWS) (3-rd active column item) is very small (3.54%).

[1,] Column relative weights and distances to the origin

Contents of spdaj

[1,]	0.2661	0.0005
[2,]	0.3204	0.0005
[3,]	0.0354	0.0049
[4,]	0.1346	0.0014
[5,]	0.1052	0.0015
[6,]	0.1384	0.0015

[1,] Coordinates of the columns

Contents of scoordj

[1,]	0.0001	0.0002	0.0004	0.0000	0.0000
[2,]	-0.0005	0.0000	-0.0001	-0.0001	-0.0001
[3,]	0.0049	-0.0001	-0.0002	-0.0004	0.0001
[4,]	-0.0010	-0.0010	0.0000	-0.0001	0.0001
[5,]	0.0009	-0.0012	-0.0002	0.0003	0.0000
[6,]	-0.0001	0.0015	-0.0002	0.0001	0.0001

but its distance to the origin is very high (0.049), which tells that its profile is very specific in terms of activities. As a result it contributes 74.6% as can be seen from the following output, to the construction of the first axis. Geometrically it is very close to this axis (squared correlation is 0.99).

[1,] Contributions of the columns

Contents of scontrj

[1,]	0.4287	1.8037	70.3836	0.6207	0.1489
[2,]	6.5641	0.0192	10.5160	13.2700	37.5915
[3,]	74.5877	0.0189	1.8090	18.1763	1.8723
[4,]	11.5011	22.4356	0.4460	7.5324	44.6282
[5,]	6.8233	25.6080	4.4877	50.8035	1.7592
[6,]	0.0950	50.1145	12.3576	9.5970	13.9999

[1,] Squared correlations of the columns

Contents of scorj

[1,]	0.0770	0.1685	0.7520	0.0024	0.0002
[2,]	0.8508	0.0013	0.0811	0.0377	0.0291
[3,]	0.9930	0.0001	0.0014	0.0053	0.0001
[4,]	0.4866	0.4940	0.0011	0.0070	0.0113
[5,]	0.3168	0.6186	0.0124	0.0517	0.0005
[6,]	0.0035	0.9587	0.0270	0.0077	0.0031

The first axis is highly explained by the 3-rd active row item **high manager** (h manag) in the following output window:

[1,] Contributions of the rows

Contents of scontri

[1,]	5.6928	37.9892	17.8813	1.9590	15.8850
[2,]	1.1848	9.9793	17.6701	4.7954	28.0180
[3,]	74.9579	2.8872	0.0622	5.2257	8.5732
[4,]	8.3279	1.4964	11.7552	21.4483	17.5522
[5,]	0.2675	18.9376	0.4701	20.3081	2.1711
[6,]	1.5383	15.9009	5.0508	46.0393	0.4038
[7,]	4.4054	5.4906	8.4193	0.1767	26.8961
[8,]	3.6255	7.3188	38.6910	0.0476	0.5005

[1,] Squared correlations of the rows

Contents of scorri

[1,]	0.2135	0.7414	0.0399	0.0016	0.0036
[2,]	0.1538	0.6742	0.1366	0.0137	0.0217
[3,]	0.9782	0.0196	0.0000	0.0015	0.0007
[4,]	0.8022	0.0750	0.0674	0.0453	0.0101
[5,]	0.0252	0.9289	0.0026	0.0420	0.0012
[6,]	0.1383	0.7437	0.0270	0.0907	0.0002
[7,]	0.5557	0.3604	0.0632	0.0005	0.0202
[8,]	0.3722	0.3910	0.2364	0.0001	0.0003

[1,] SUPPLEMENTARY ITEMS

[1,] Row relative weights and distances to the origin

Contents of spdsl

[1,]	0.1644	0.0006
[2,]	0.1714	0.0006
[3,]	0.0610	0.0012
[4,]	0.0614	0.0012
[5,]	0.0883	0.0004
[6,]	0.0648	0.0010
[7,]	0.0602	0.0016
[8,]	0.1010	0.0015
[9,]	0.0873	0.0004
[10,]	0.0805	0.0024
[11,]	0.0595	0.0026

The 11-th supplementary row item **university education** (UNIVER) is closely linked to factor 1, see the following output:

[1,] Squared correlations of the rows

Contents of scontrsi

[1,]	0.4813	0.1104	0.0215	0.3239	0.0629
[2,]	0.4910	0.1025	0.0213	0.3261	0.0591
[3,]	0.0150	0.5609	0.0762	0.2102	0.1377
[4,]	0.0542	0.8704	0.0100	0.0350	0.0304
[5,]	0.6140	0.1026	0.0726	0.0316	0.1791
[6,]	0.0478	0.8030	0.0011	0.1184	0.0296
[7,]	0.1438	0.5840	0.1552	0.0894	0.0275
[8,]	0.6289	0.2446	0.0209	0.1034	0.0023
[9,]	0.0002	0.6872	0.0001	0.2908	0.0218
[10,]	0.0132	0.4614	0.0187	0.1283	0.3783
[11,]	0.9882	0.0033	0.0024	0.0025	0.0037

[1,] Coordinates of the rows

Contents of scodsi

[1,]	0.0004	-0.0002	0.0001	-0.0004	0.0002
[2,]	-0.0004	0.0002	-0.0001	0.0004	-0.0002
[3,]	0.0001	0.0009	0.0003	0.0006	-0.0004
[4,]	0.0003	0.0011	0.0001	0.0002	-0.0002
[5,]	0.0003	0.0001	0.0001	0.0001	0.0001
[6,]	-0.0002	-0.0009	0.0000	-0.0003	0.0002
[7,]	-0.0006	-0.0012	-0.0006	-0.0005	0.0003
[8,]	-0.0012	-0.0007	-0.0002	-0.0005	0.0001
[9,]	0.0000	0.0004	0.0000	0.0002	0.0001
[10,]	0.0003	0.0017	0.0003	0.0009	-0.0015
[11,]	0.0026	-0.0002	0.0001	0.0001	-0.0002

It is clear in this analysis that main trait (first axis) is that the contact of **national newspapers** corresponds, in a highly significant way, to **high manager** and (or) people with **university education**.

The second axis characterizes mostly an opposition between **TV magazines** (TVMAG) (associated with **employer**, **worker**, and the **younger people**) and **magazine** (MAGAZ), and **regional newspapers** (R NEWS) associated with **farmer**, **small business** (s busin) and **older people** (A50–64, A65+). Figure 2 summarizes this set of associations.

The positions of items on Figure 2 explain a nuance interpretation on the second axis: the **employer** and **worker**, people of **middle level education** (SECOND), associated in particular with the **young** (A25–34, A14–24) (contact media such as TV magazine), are opposed to **small business** and **farmers**, who are primarily **older** (A50–64, A65+) with **less education** (PRIMARY) and contact media such as **magazine** (MAGA) and **regional newspapers** (R NEWS).

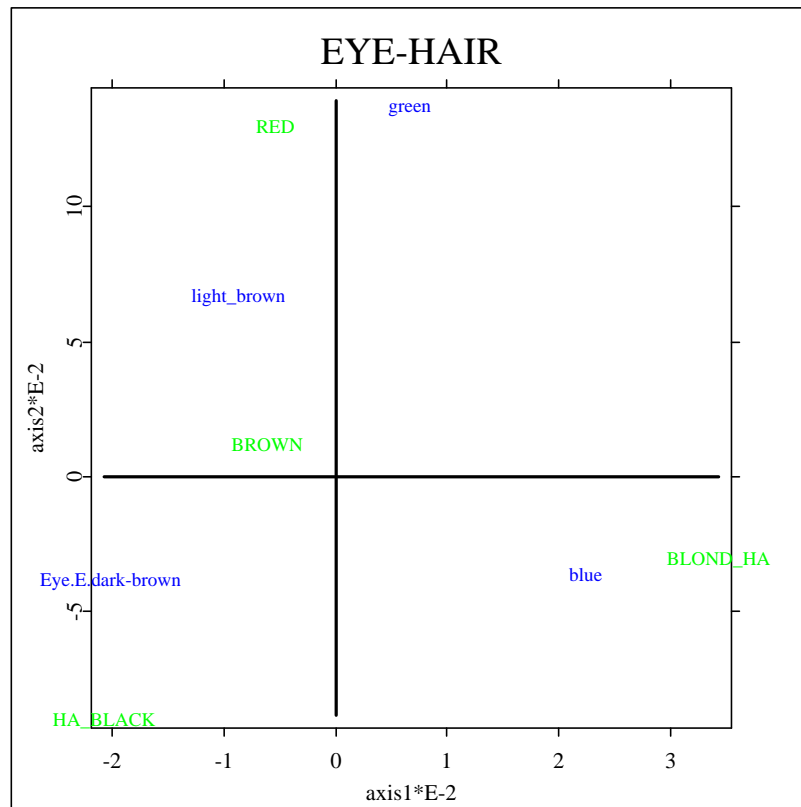


Figure 1: CA for the eye-hair example.

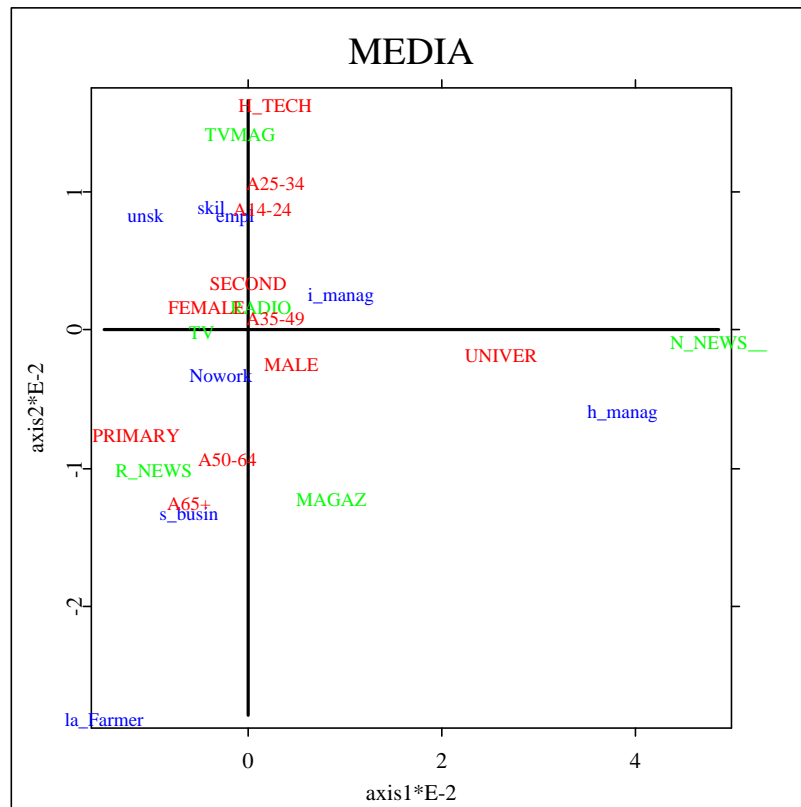


Figure 2: Biplot for media data set.

References

- Lebart, L., Morineau, A., and Piron, J. (1995). *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris.
- Härdle, W., Klinke, S., and Turlach, B.A. (1994). *XploRe: An Interactive Statistical Computing Environment*, Springer-Verlag.
- Härdle, W., Klinke, S., and Müller, M., (2000). *XploRe: Learning Guide*, Springer-Verlag.
- Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*, Editions Technip, Paris.