eNote 5

CA, Correspondence Analysis in R

Indhold

5.1 Reading material

We have shared 3 documents with you on Campusnet on this topic:

- A description of CA from the NTSYS software (allthough we do not need this software) - the Lebart data (Lebart et. al, 1984) is decsribed here.
- A CA description by Dianne Phillips (Social Research Update, Univ. Surrey)
- The paper: Nenadic and Greenacre (2007). Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software* 20(3), 1–13.

In the latter we will focus on the simple CA, and you may skip everything else. Even though this paper is almost 8 years old, the ca package was updated by the end of 2014. One may check the 'Package NEWS' to see the improvements since then. (The basic stuff did not change)

5.2 Smoke data example from the ca package

In this section we include the main stuff from the analysis of the smoke data included in the package: (remember to install the ca-package first)

```
library(ca)
data(smoke)
smoke


   none light medium heavy
SM    4     2      3     2
JM    4     3      7     4
SE   25    10     12     4
JE   18    24     33    13
SC   10     6      7     2
```

Note how the smoke data is a standard 2-way frequency table, which could be an example for a standard χ^2-statistic analysis in an introductory statistics class, see e.g. Chapter 7 in:

http://introstat.compute.dtu.dk/enote/afsnit/NUID178/

(Sec 7.5 on contingency tables). The basic analysis:

```
# The basic analysis:
casmoke <- ca(smoke)
casmoke



 Principal inertias (eigenvalues):
           1        2        3
Value      0.074759 0.010017 0.000414
Percentage 87.76%   11.76%   0.49%


 Rows:
                SM        JM        SE       JE        SC
Mass      0.056995  0.093264  0.264249 0.455959  0.129534
ChiDist   0.216559  0.356921  0.380779 0.240025  0.216169
Inertia   0.002673  0.011881  0.038314 0.026269  0.006053
Dim. 1   -0.240539  0.947105 -1.391973 0.851989 -0.735456
Dim. 2   -1.935708 -2.430958 -0.106508 0.576944  0.788435
```

```
 Columns:
             none    light   medium     heavy
Mass     0.316062 0.233161 0.321244  0.129534
ChiDist  0.394490 0.173996 0.198127  0.355109
Inertia  0.049186 0.007059 0.012610  0.016335
Dim. 1  -1.438471 0.363746 0.718017  1.074445
Dim. 2  -0.304659 1.409433 0.073528 -1.975960
```

Content of result object:

```
# Content of result object:
names(casmoke)


 [1] "sv"          "nd"          "rownames"    "rowmass"     "rowdist"
 [6] "rowinertia"  "rowcoord"    "rowsup"      "colnames"    "colmass"
[11] "coldist"     "colinertia"  "colcoord"    "colsup"      "call"
```

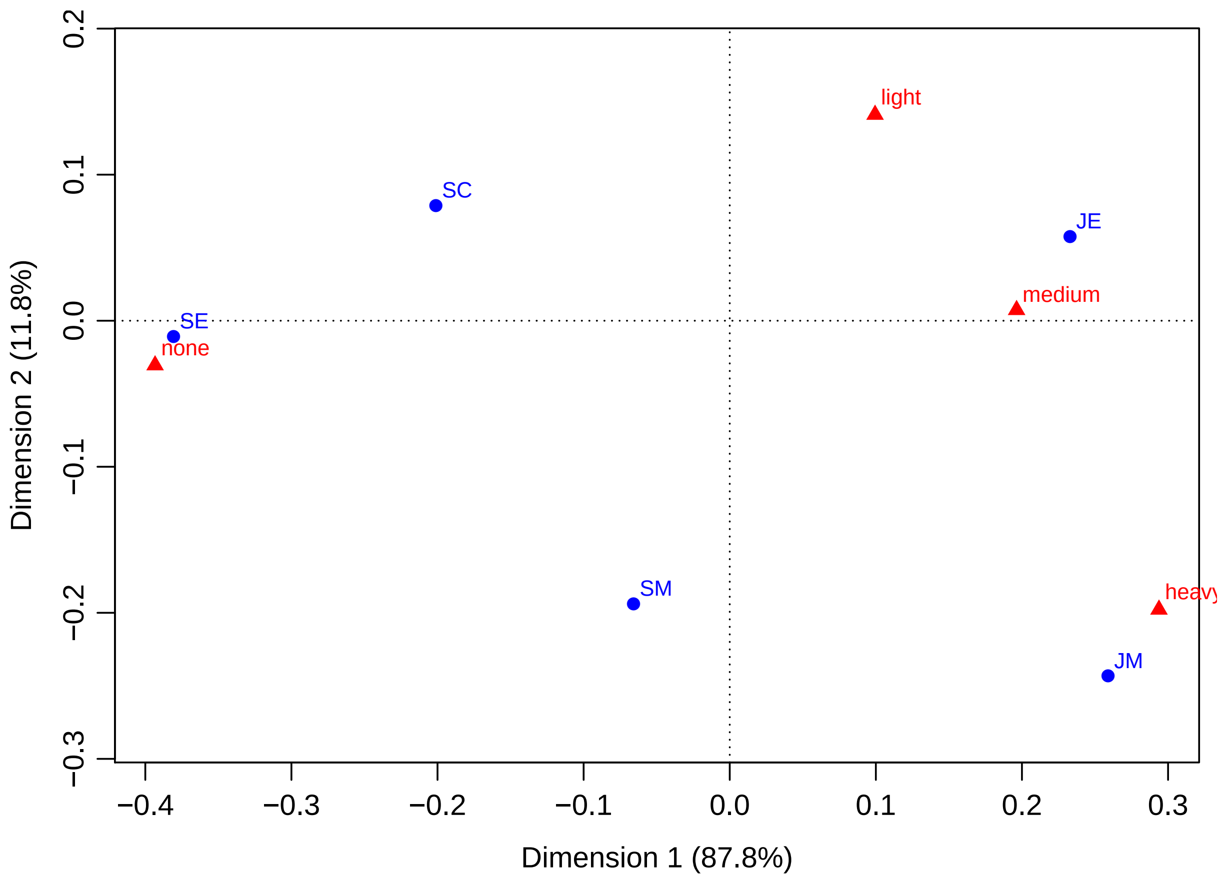
```
# E.g. row standard coordinates:
casmoke$rowcoord


         Dim1       Dim2       Dim3
SM -0.2405388 -1.9357079  3.4903231
JM  0.9471047 -2.4309584 -1.6573725
SE -1.3919733 -0.1065076 -0.2535221
JE  0.8519895  0.5769437  0.1625337
SC -0.7354557  0.7884353 -0.3973677
```

```
# Summary:
summary(casmoke)



Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.074759  87.8  87.8  **********************
 2      0.010017  11.8  99.5  ***
 3      0.000414   0.5 100.0
```

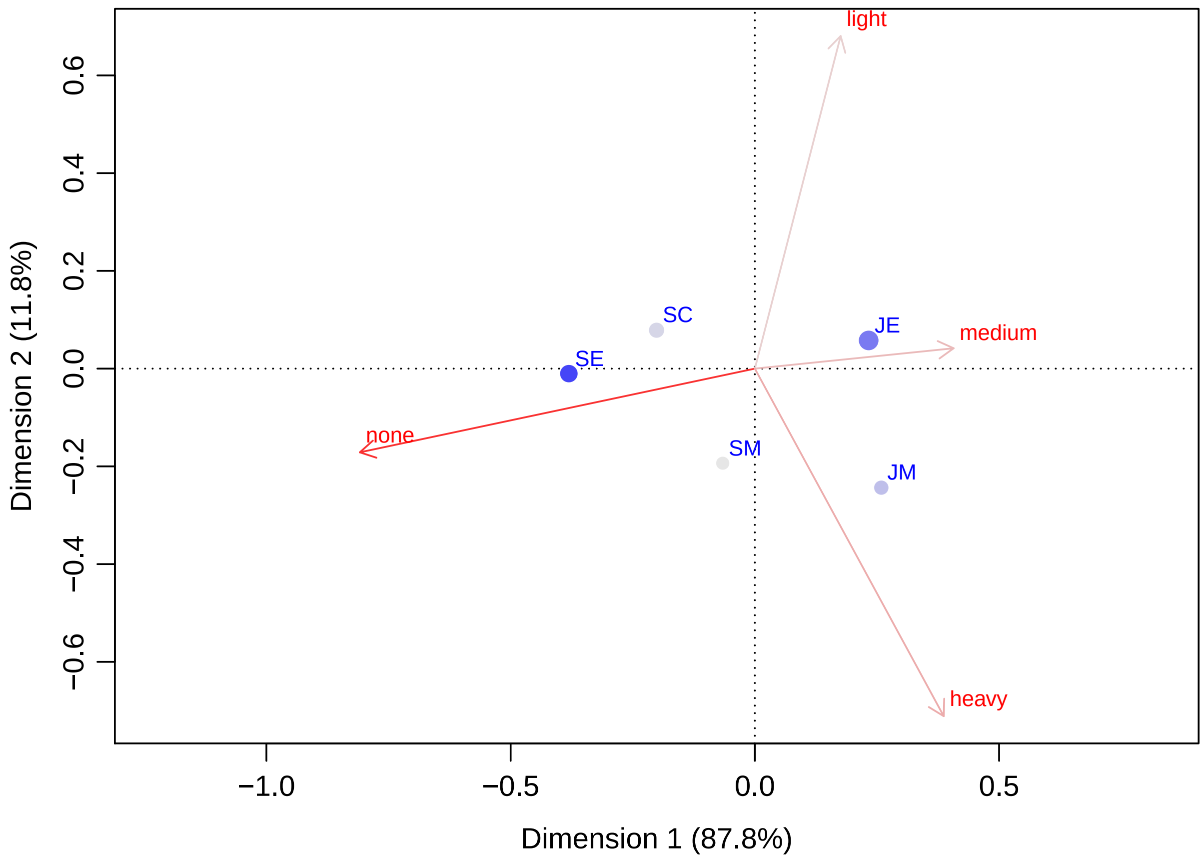
```
        -------- -----
 Total: 0.085190 100.0


Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   SM |   57  893   31 |  -66  92   3 | -194 800 214 |
2 |   JM |   93  991  139 |  259 526  84 | -243 465 551 |
3 |   SE |  264 1000  450 | -381 999 512 |  -11   1   3 |
4 |   JE |  456 1000  308 |  233 942 331 |   58  58 152 |
5 |   SC |  130  999   71 | -201 865  70 |   79 133  81 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | none |  316 1000  577 | -393 994 654 |  -30   6  29 |
2 | lght |  233  984   83 |   99 327  31 |  141 657 463 |
3 | medm |  321  983  148 |  196 982 166 |    7   1   2 |
4 | hevy |  130  995  192 |  294 684 150 | -198 310 506 |
```

Plotting of results (default):

```
# Plotting of results (default):
plot(casmoke)
```




```
# Standard CA biplot:
plot(casmoke, mass = TRUE, contrib = "absolute",
     map = "rowgreen", arrows = c(FALSE, TRUE))
```



3D-plotting is available by: (will create an additional plot window for interactive 3D spinning and zooming)

```
plot3d.ca(ca(smoke, nd=3))
```

5.3 CA, PCA and χ^2-statistics

Let us try to understand the link to PCA and standard χ^2-statistic based analysis. The latter would give:

```
# Chi-square analysis
chisqresults <- chisq.test(smoke)


Warning in chisq.test(smoke): Chi-squared approximation may be incorrect
```

```
chisqresults


Pearson's Chi-squared test

data:  smoke
X-squared = 16.4416, df = 12, p-value = 0.1718
```

Actually as there is no (apparently) significant deviation from the independence hypothesis, there is not a strong reason to move on with the analysis of these data. BUT generally, the CA is a way to have a look at the structures in this classic analysis that makes the χ^2-statistic significant.

The χ^2-statistic is the sum of the squared pearson contributions:

```
chisqresults$observed


   none light medium heavy
SM    4     2      3     2
JM    4     3      7     4
SE   25    10     12     4
JE   18    24     33    13
SC   10     6      7     2


chisqresults$expected


        none     light    medium     heavy
SM  3.476684  2.564767  3.533679  1.424870
JM  5.689119  4.196891  5.782383  2.331606
SE 16.119171 11.891192 16.383420  6.606218
JE 27.813472 20.518135 28.269430 11.398964
SC  7.901554  5.829016  8.031088  3.238342


sum((chisqresults$observed-chisqresults$expected)^2/chisqresults$expected)


[1] 16.44164
```

We see that some of the expected frequencies are smaller than 5, so the classic χ^2-distribution would be questionable to us here, as also warned about by R.

A standard recommendation in basic statistics would be to look at the signed root-contributions to the statistic for interpretations - to identify the row-column combinations that are the reason for the statistic to become large:

```
nphis <- (chisqresults$observed-chisqresults$expected)/sqrt(chisqresults$expected)
nphis
```

```
         none       light     medium      heavy
SM  0.2806606 -0.35265110 -0.2839006  0.4818124
JM -0.7081704 -0.58423937  0.5063573  1.0926246
SE  2.2119849 -0.54843210 -1.0829559 -1.0139913
JE -1.8607802  0.76867548  0.8897233  0.4742076
SC  0.7465200  0.07082051 -0.3638384 -0.6881439
```

```
 # Check:
sum(nphis^2)
```

```
[1] 16.44164
```

Note that the so-called *total inertia* in CA is simply the χ^2-statistic divided by the total number of observations in the table:

```
sum(smoke)
```

```
[1] 193
```

```
sum(nphis^2)/sum(smoke)
```

```
[1] 0.08518986
```

which can be found as the *total inertia* in the CA output above. So let's define the inertia contributions as:

```
phis <- nphis/sqrt(sum(smoke))
# Check:
sum(phis^2)


[1] 0.08518986


phis


          none        light      medium       heavy
SM  0.02020239 -0.025384382 -0.02043562  0.03468162
JM -0.05097522 -0.042054470  0.03644840  0.07864884
SE  0.15922216 -0.039477006 -0.07795287 -0.07298869
JE -0.13394189  0.055330472  0.06404368  0.03413421
SC  0.05373569  0.005097772 -0.02618966 -0.04953368
```

So these values, simply the signed root χ^2-contributions (relative to $\sqrt{n}$) are the individual contributions to the Inertia expressed in bullit point 7 in the Nenadic and Greenacre (2007) paper. And also the actual S-matrix defined in bullit point 1 as the basic numbers that are analyzed by CA.

The CA computations are simply the PCA of this matrix: (without any further centering nor scaling)

```
pcaofphis <- prcomp(phis, center=FALSE)

# Percentage of explained variation in the PCA:
round(100*pcaofphis$sdev^2/sum(pcaofphis$sdev^2),2)


[1] 87.76 11.76  0.49  0.00
```

If we compare with the CA, we see that this is ECAXTLY what comes out of this.

Try to compare the CA biplot with the contributions to the χ^2-statistics! See how the largest contributions could be identified in the plot. These are the combinations of row level and column level that mostly deviates from independence - either positively or negatively.

5.4 Exercises

|||| Exercise 1 Lebart data

Use the questionnaire results from Lebart et al. (1984) to see the different applications of correspondence analysis. The rows are 23 occupations and columns represent 15 advantages of these occupations according to the questionnaire (data file Lebart.txt in ascii text format):

```
options(width=90)
Lebart <- read.table("Lebart.txt", sep = " ", header = TRUE, row.names = 1)
Lebart


      VARI FREE HUMA SCHE SALA SECU COMP INTE NEAR ATMO SOCI AUTO LIKE NONE OUTD
FARM1    4  189    0    3    2    2    9    3   12    2    1    4   11   12    8
FARM2    1   13    3   10   17   12    4    1    8    3    5    1    9   11    0
ENER     1    9    1    0    4   13    0    2    2    0    2    1    4    6    1
STEE     5    5    2    9   18    5    3    2    6    5    4    0    2   22    0
CHEM     2    7    1    4   15    5    2    1    6    1    2    2    3    5    0
WOOD     2    5    0    4    1    0    3    0    2    1    1    1    1    3    0
AUT      2    3    1    8   16   17    1    8    7    2    4    3    6   24    0
TEXT     3   18    0    6   16    5    4    4   13    4    2    3    6   26    0
PHAR     3    7    3    6    6    0    0    2    6    3    3    0    2    8    0
MANU     0   18    1   12   31    7    0    8   19   11    3    2   10   26    0
CONS     7   63    2    9   31    9    4    6    9   10    3    4   14   35    2
FOOD     2   43   16    7    6    4    7    1    8    2    0    1    6    7    0
SBUS     8   95   23   15   15    2   13    7    9    5    2    3   13   18    1
MBUS     5   32    9    9   17    4    5    4    7    4    3    0    8   18    0
TELE     1    7    2   11    3   14    2    6    3    1    1    2    1    5    0
SO.S     4   10   10    8    2    1    6    4    2    3    1    0    3    1    0
HE.S     3   31   16   15   11   19    5   19   10    2    3    7   24    5    0
TEAC     2   33   27   31    9   18   27   24    3    4   43    8   18   11    1
TRAN     2   19    2   12   12   21    0    1    4    5    5    1    3   13    0
BANK     8   12    4    8   13   21    2   10    4    2    5    6    3   10    0
DOME     0    8    0    4    5    2    7    1    5    7    2    1    2   11    1
O.SE     8   35   14   13   16   10    6   25    6    4   10    9   11   14    0
PRIV     3   26    9    3   12    5    8    8    4    4    2    3   10    8    0
```

a) How much of the variation is explained by the first axes? How many axes are needed to explain most of the variance in the data (80

b) Which row and column factors are most deviating from the expected?

c) Make a biplot in graphics, with both objects and variables. Which occupations are related to which advantages? (to see names on biplots, use Options – Plot Options and add labels to both object and variable points)

|||| Exercise 2 EU data. (for presentationby student group)

: Use correspondence analysis to analyze the EU data (EU.txt) from EU Government conference 1996:

Row factors (Variables):

1. INT: The fundamental opinion on more integration in EU (General)
2. EXP: The fundamental opinion on expansion of EU (Institutions)
3. VOT: New placement of votes in the ministerial council (Institutions)
4. RUL: New rules on the chairmanship of EU (Institutions)
5. POW: More power to the European Commission (Institutions)
6. STR: Enforcement of the European Parliament (Institutions)
7. BES: Enhanced use of the consiliarity principle (for EU Parliament) (Institutions)
8. SUB: Enhancement of the subsidiarity principle? (Institutions)
9. BUD: Treatment of budget problems at the government conference? (finances)
10. HOU: More power to the European parliament in the "household" budget=? (finances)
11. MAJ: Enhanced use of majority decisions? (First column)
12. OMU: European Monitary Union (ØMU) at the government conference? (First column)
13. COM: New treaty based competences to EU (new areas)? (First column)
14. ENV: Fortification of the environment and the social dimension? (First column)
15. FOR: Gradual movement to majority decisions in foreign policy? (Second column)

16. MRX: Election of a mister X to represent EU in foreign policy? (Second column)
17. FIN: Financing of the common foreign policy over the EU budget? (Second column)
18. WEU: Merging of EU and the Western Union? (Defence)
19. OVE: Transfer of between-state to the "over-state" collaboration? (Third column)

FEW: Fewer EU commisarians? (Institutions) has not been included, but can be included if you wish so!

```
EU <- read.table("EU.txt", sep = " ", header = TRUE, row.names = 1)
EU


    B DK D GR E F IRL I L NL A P SF S GB EK EP
INT 6  3 4  4 4 4   4 4 4  4 4 4  2 2  0  4  4
EXP 6  6 6  6 2 3   2 4 4  4 4 2  4 4  6  4  4
VOT 2  3 4  2 6 6   2 6 2  3 2 4  2 2  4  4  3
RUL 3  3 3  3 3 4   2 3 3  4 2 3  2 2  4  3  2
POW 2  2 2  2 2 4   0 2 2  2 2 2  2 2  4  0  2
STR 4  3 4  3 3 2   2 4 2  4 4 2  2 2  2  3  6
BES 4  3 4  4 3 3   3 4 4  4 4 4  3 4  2  4  4
SUB 2  4 4  2 2 2   3 3 4  4 4 3  4 4  4  3  3
BUD 3  2 2  4 4 2   4 2 2  2 2 4  2 2  2  4  4
HOU 3  3 2  3 3 2   2 2 3  2 3 2  2 3  3  4  6
MAJ 6  4 4  4 3 4   4 4 4  3 4 4  3 4  2  4  4
OMU 2  2 0  2 3 2   2 3 2  2 2 2  2 3  2  3  3
COM 4  2 2  4 4 4   2 4 4  2 4 4  2 4  2  4  4
ENV 4  4 3  3 3 3   4 3 4  3 6 3  4 6  0  4  4
FOR 4  4 6  2 4 3   2 4 4  4 4 4  2 4  0  4  4
MRX 3  4 3  2 3 6   3 4 2  2 3 2  2 3  2  2  2
FIN 4  4 4  4 4 4   4 4 4  4 4 4  4 4  2  4  4
WEU 4  2 4  4 4 2   4 4 4  4 2 4  2 2  2  4  4
OVE 4  2 3  3 2 2   2 4 4  3 4 4  2 2  0  4  4
```

a) How much variance is explained by the first three CA axes? How many CA axes are necessary to explain the most important information in the data?

b) How many axes can maximally be extracted from these data? (Please note that the "first" CA axis is always of eigenvalue one (1), as you loose one degree of freedom as you make column and row summations, so this is completely disregarded)

c) Which countries are most "extreme"?

d) Are there any groupings of the countries?

e) Are the any correspondence between certain countries and certain variables?

f) If time permits please analyze the same data in PCA. What are the differences between CA and PCA?