UNIVERSITÄT LEIPZIG

erc European Research Council

Correspondence Analysis

Natalia Levshina ©2017

Summer School of Linguistics, Litomyšl, August 2017

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of Stuhl and Sessel in German

Introduction to CA

- CA is used to visualize and explore associations between the values of two and more categorical variables (usually represented as factors in R),
  - e.g. Do upper middle-class people prefer to play tennis and listen to opera?
  - Do languages with the Adj + N order also tend to have Num + N and Gen + N?
- Similar to MDS, CA allows to see structure in the data and identify which variables are associated and which of their values tend to co-occur.

The main idea behind CA

- CA is based on comparison of row profiles and column profiles, e.g.

	Birds	Music	Games	Total
M	20	30	50	100
F	10	70	20	100
Total	30	100	70	200

row profiles

	Birds	Music	Games	Total
M	0.2	0.3	0.5	1
F	0.1	0.7	0.2	1

column profiles

	Birds	Music	Games
M	0.67	0.3	0.71
F	0.33	0.7	0.29
Total	1	1	1

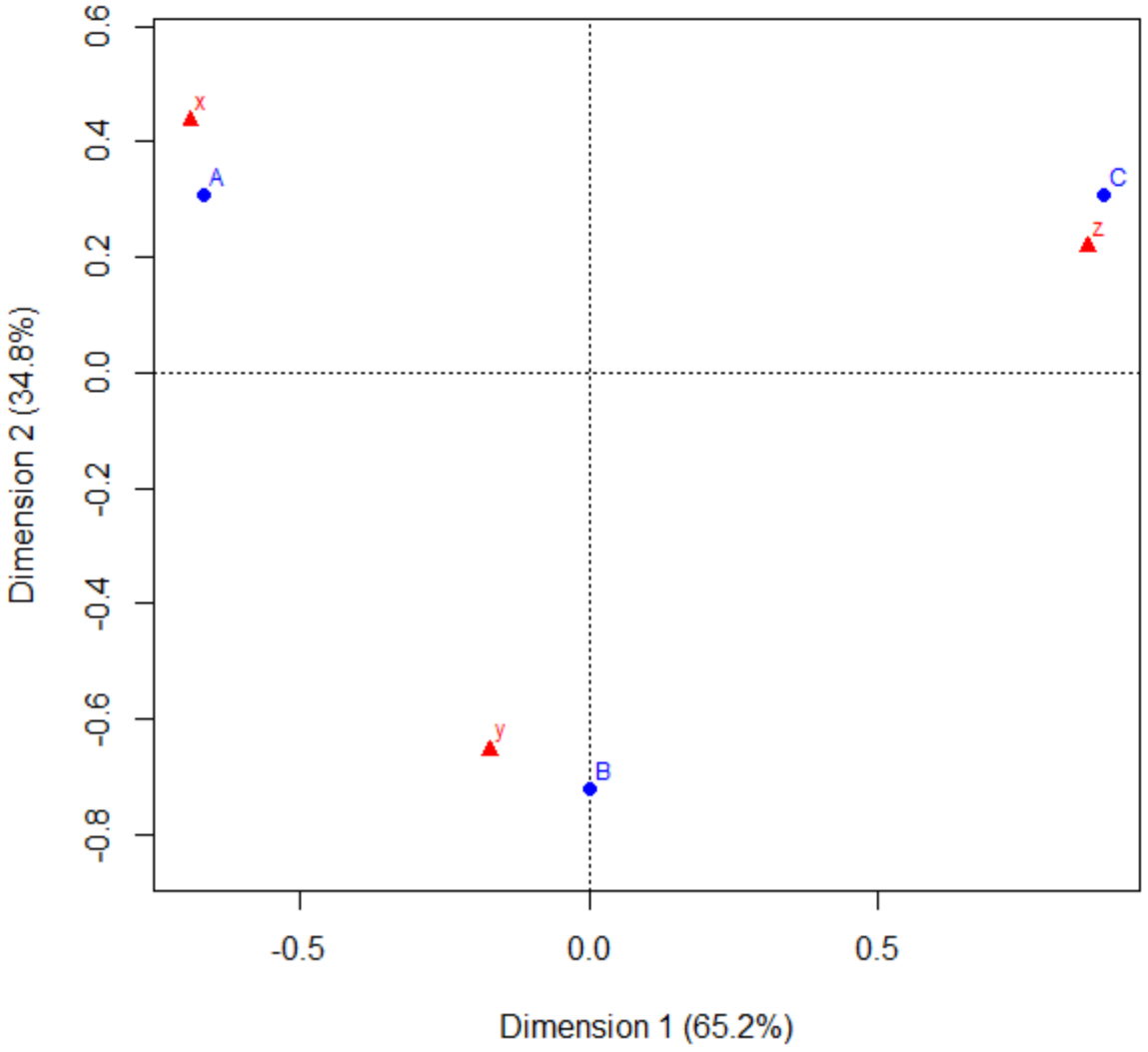
The main idea behind CA

- If two row or column profiles are similar, their labels will be closely located in a semantic map.
- If two row or column profiles are dissimilar, their labels will be located far from each other.

Strong association (all profiles are dissimilar)

	x	y	z
A	80	30	10
B	10	60	20
C	10	10	70

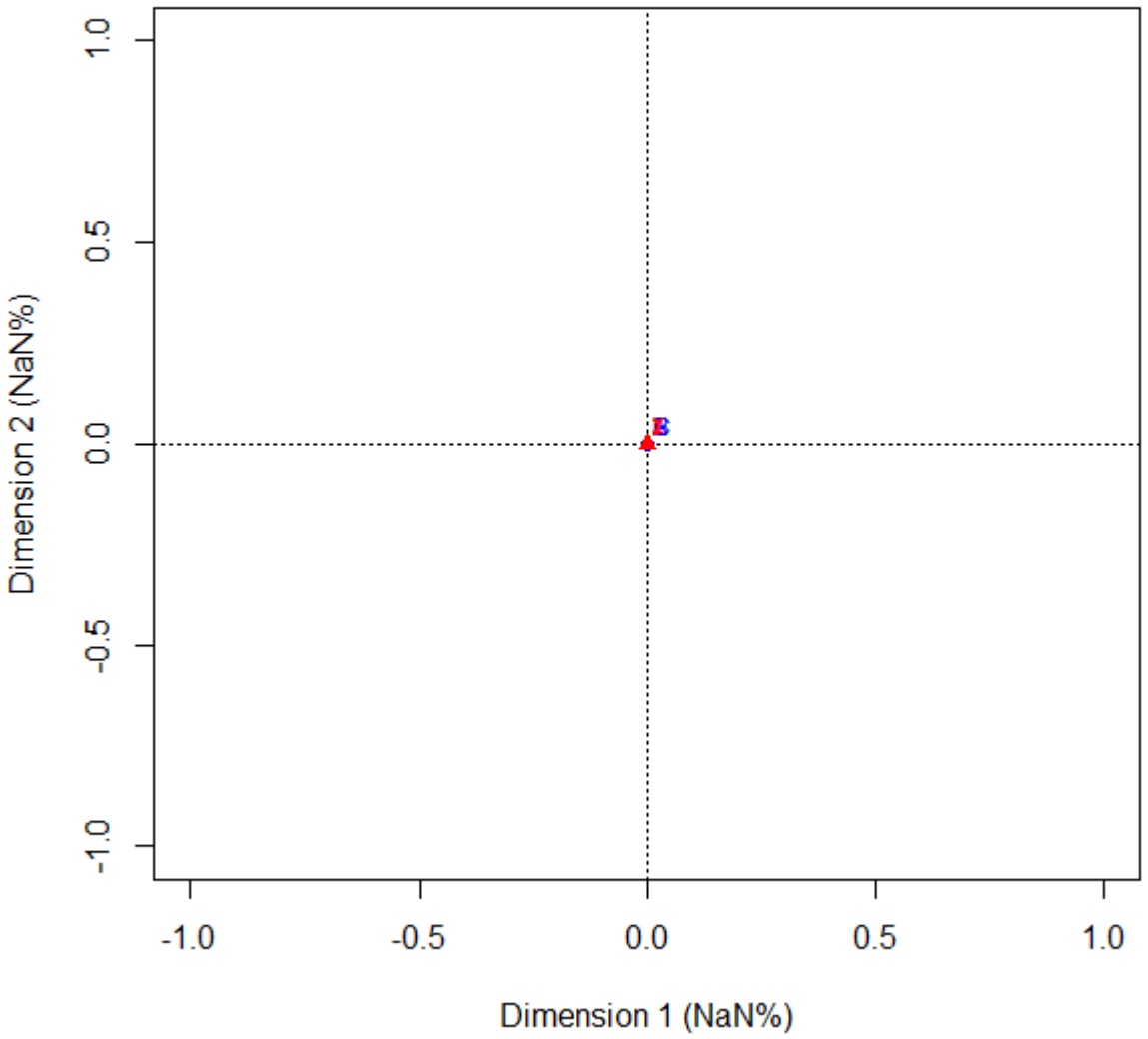
```
> chisq.test(example)

        Pearson's Chi-squared test

data:  example
X-squared = 191.67, df = 4, p-value < 2.2e-16
```


x
A
C
z
y
B
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.5
0.0
0.5
Dimension 2 (34.8%)
Dimension 1 (65.2%)

Lack of association (all profiles are similar)

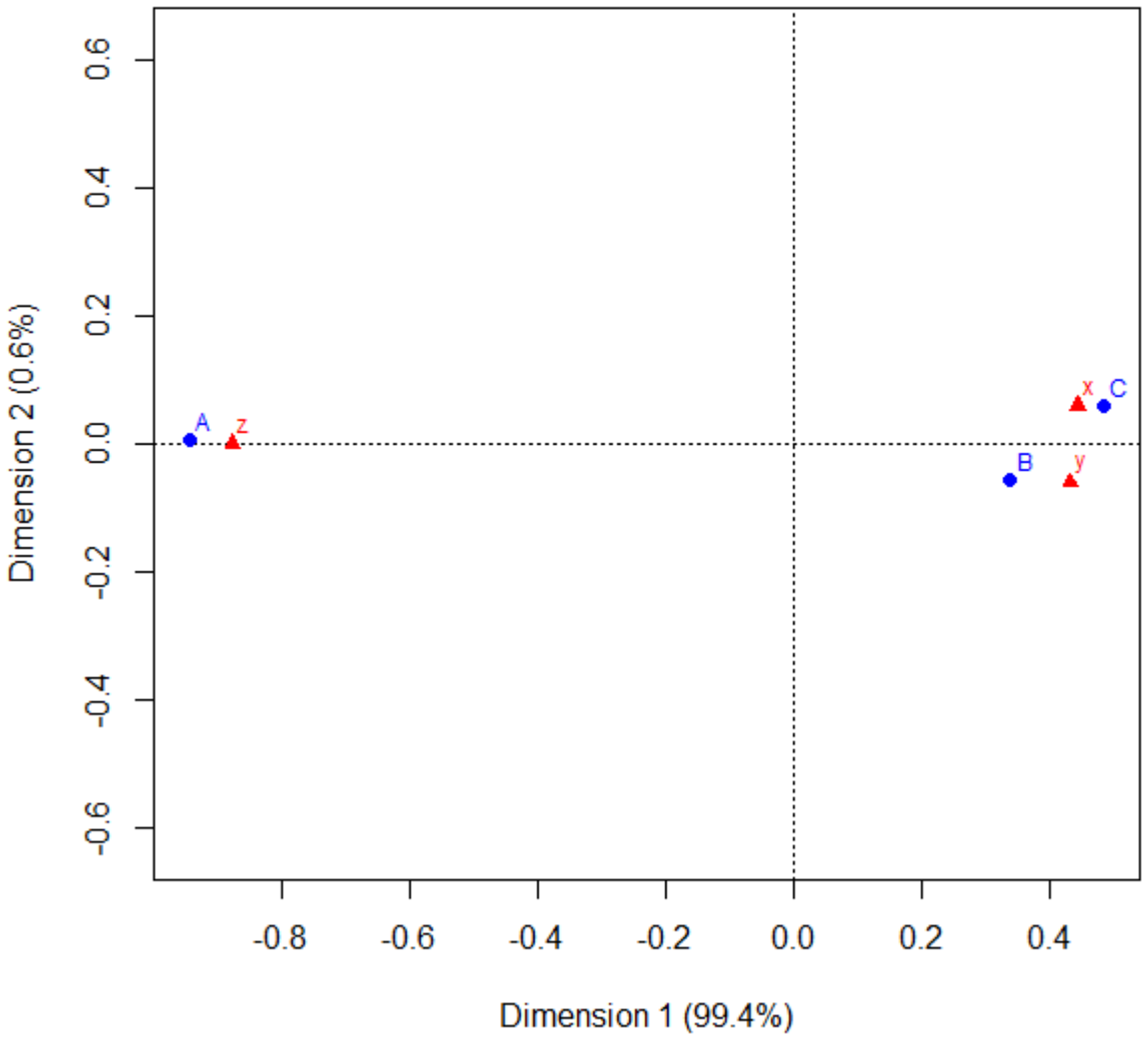
	x	y	z
A	10	10	10
B	80	80	80
C	10	10	10

```
> chisq.test(example1)

        Pearson's Chi-squared test

data:  example1
X-squared = 0, df = 4, p-value = 1
```


1.0
0.5
0.0
-0.5
-1.0
Dimension 2 (NaN%)
-1.0
-0.5
0.0
0.5
1.0
Dimension 1 (NaN%)

Some profiles are similar

	x	y	z
A	10	10	70
B	45	50	20
C	45	40	10


Dimension 2 (0.6%)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
A
z
x
C
B
y
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
Dimension 1 (99.4%)

Simple and multiple CA

- If there are two variables, which are cross-tabulated, perform a simple CA on the table with counts.
- If there are more than two variables, perform a multiple CA on the data frame with variables as columns.

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of Stuhl and Sessel in German

Verbs of communication

- announce
- assert
- babble
- blab
- chat
- chatter
- comment
- communicate
- converse
- declare
- discuss
- enunciate
- gab
- mumble
- murmur
- notify
- proclaim
- schmooze
- speak
- talk
- utter
- verbalize
- whisper
- yap

Contingency table with counts

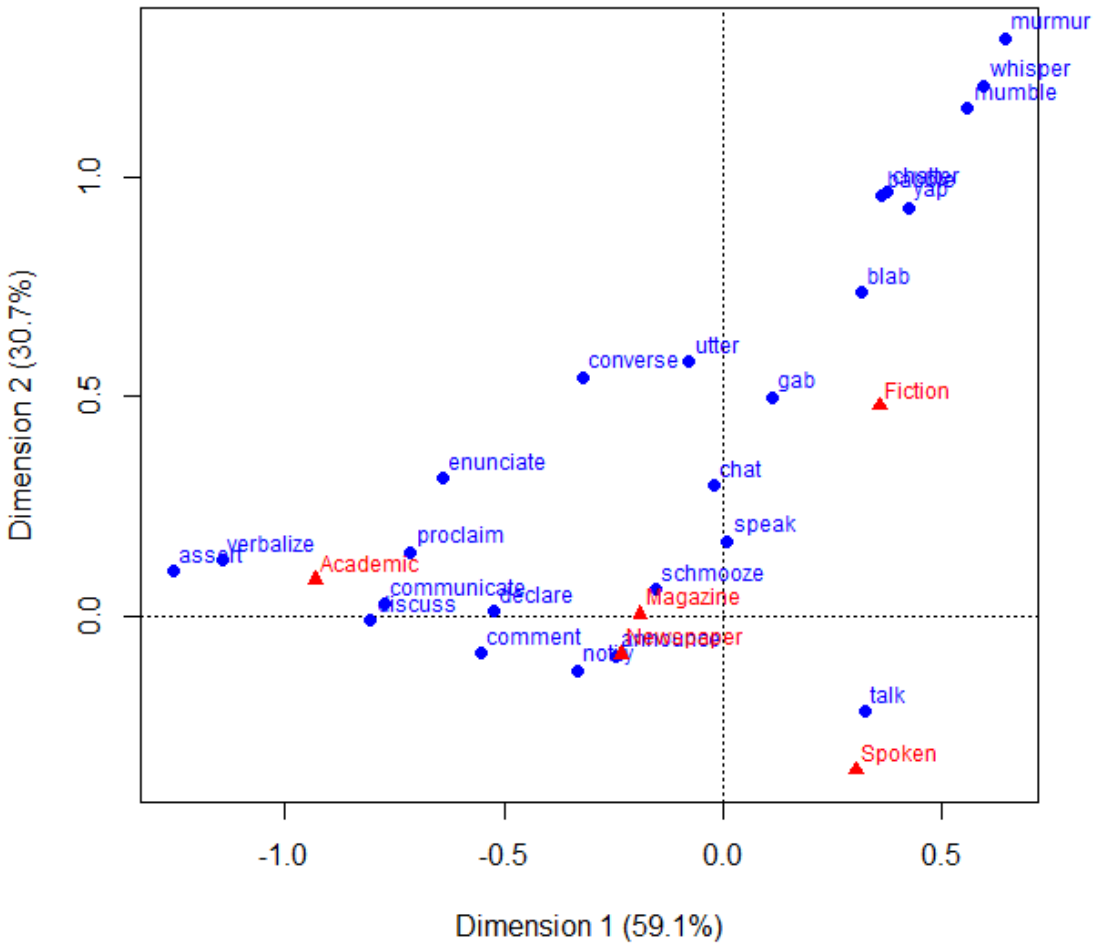
```
> head(speak)
```

	Spoken	Fiction	Magazine	Newspaper	Academic
communicate	2327	1393	2664	1825	5147
chat	684	1672	1335	1155	354
declare	3449	2762	5335	5413	5167
utter	249	1336	595	397	484
whisper	465	13668	1445	779	273
assert	577	445	2259	1654	5784

Performing a simple CA

```
> library(ca)
> speak.ca <- ca(speak)
> plot(speak.ca) #the first two dimensions, by default
> plot(speak.ca, dim = 2:3) #dimensions 2 and 3
```

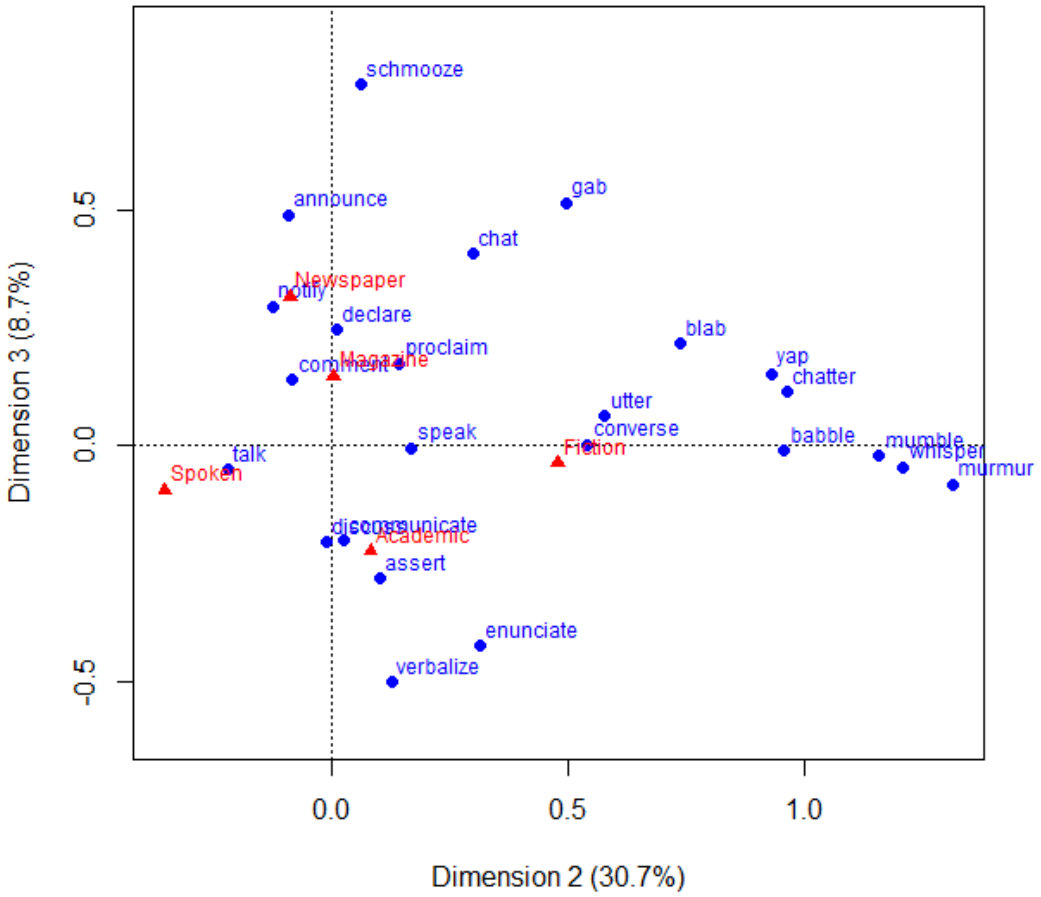
Dimensions 1 & 2



How to read a simple CA map

- If two row values are located close to each other, they have similar profiles.
  - Here, the rows are individual verbs. Their proximity means that their relative frequencies of occurrence in the subcorpora (registers) are similar.
- If two column values are close, this means that they have similar profiles, too.
  - Here, the columns are the subcorpora (registers). Their proximity means that they share similar proportions of the verbs.
- If the row and column labels are located in the same area regarding the origin, this means they co-occur frequently in the data. But the absolute distance should not be taken as a representation of association!

Dimensions 2 & 3



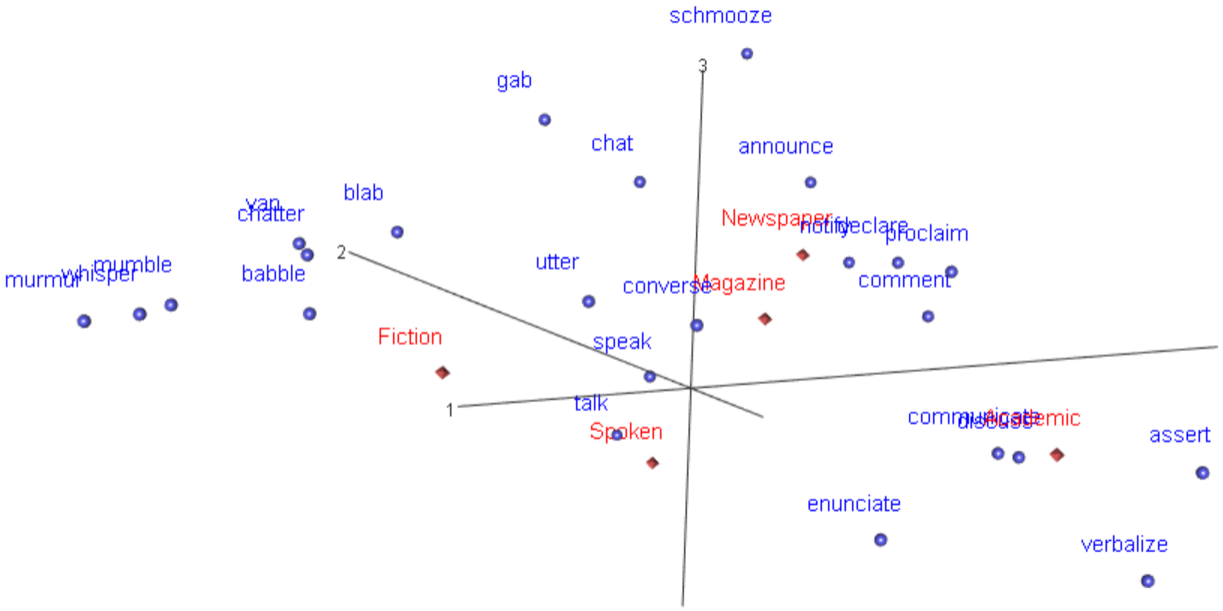
Creating an interactive 3D plot

- Important: you'll need to install package rgl first!

```
> plot3d.ca(speak.ca)
```

- The plot is interactive. You can use your mouse or touchpad to rotate the axes and zoom in/out.

Interactive 3D plot



Quality of the 3D solution

- How much information do we lose if we take the three-dimensional solution?

```
> summary(speak.ca)

Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.190546  59.1   59.1   ***************
 2      0.099111  30.7   89.9   ********
 3      0.028158   8.7   98.6   **
 4      0.004538   1.4  100.0
        -------- -----     3 dimensions explain 98.6%!
 Total: 0.322352 100.0
```

Interpretation

- The spoken subcorpus is associated with the verb *talk*.
- The verbs of manner of saying (onomatopoeic) are associated with fiction, e.g. *murmur, whisper, chatter, babble*.
- Some Latinate verbs of argumentation and verbal expression (*discuss, assert, enunciate, verbalize*) are associated with the academic prose.
- Some neutral verbs of sharing information (*notify, announce, declare, comment*) are more associated with newspapers and magazines.

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of Stuhl and Sessel in German

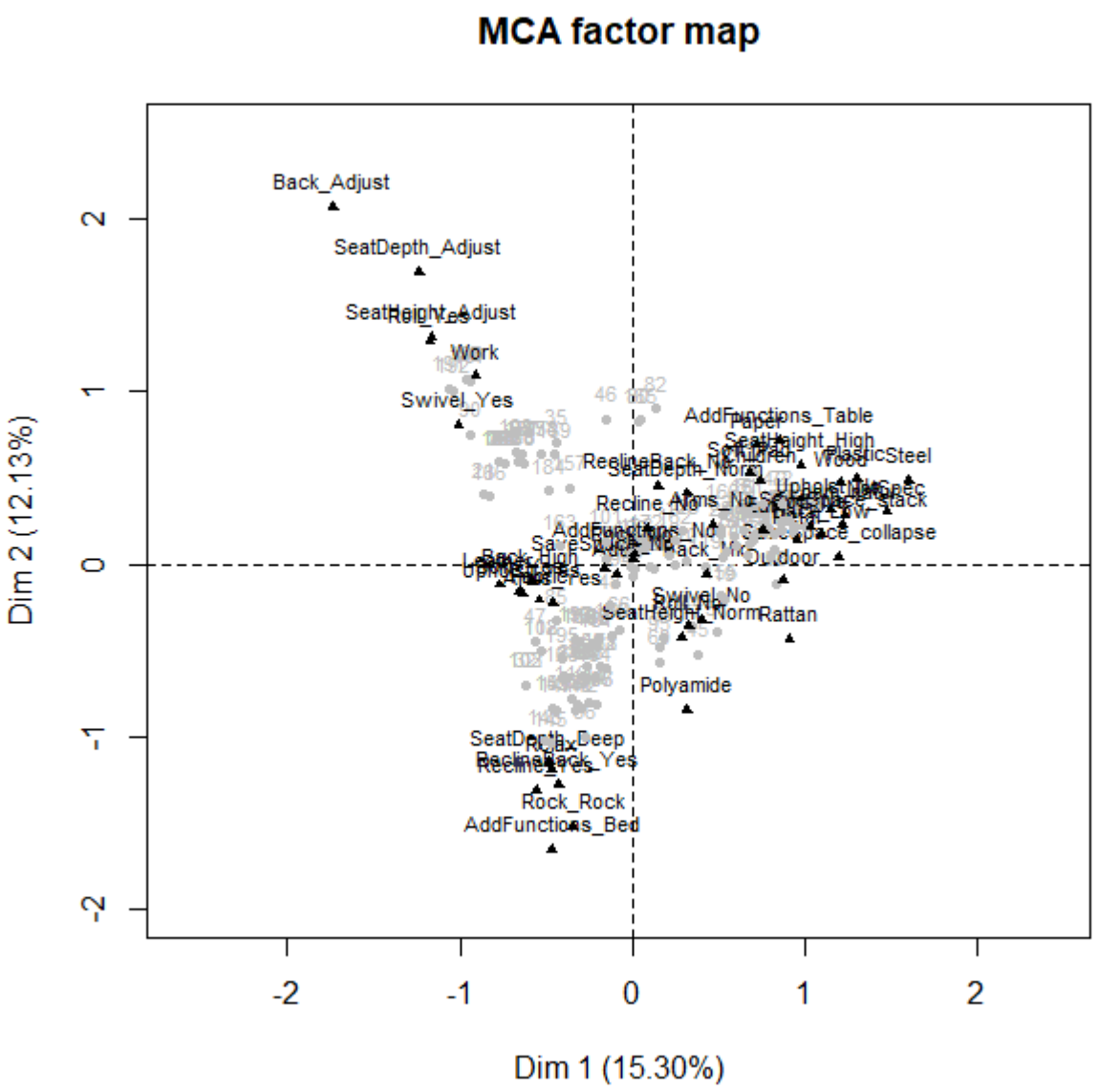
Chairs: data from online stores

```
> str(chairs)
'data.frame': 188 obs. of  19 variables:
 $ Shop        : Factor w/ 3 levels "ikea.de","Moebel-
Profi.de",..: 2 1 1 2 1 3 1 3 1 1 ...
 $ WordDE      : Factor w/ 44 levels "3-in-1-Sessel",..:
2 17 38 41 23 13 25 15 40 40 ...
 $ Category    : Factor w/ 2 levels "Sessel","Stuhl": 2
2 1 2 2 2 2 1 2 2 ...
 $ Function    : Factor w/ 5 levels "Eat","NotSpec",..:
1 1 2 1 1 5 2 4 1 1 ...
 $ Age         : Factor w/ 2 levels "Adult","Children":
1 2 1 1 2 1 1 1 1 1 ...
 $ Back        : Factor w/ 4 levels "Adjust","High",..:
3 4 4 2 2 2 4 2 4 4 ...
```

…

MCA in FactoMineR

```
> library(FactoMineR)
> chairs.ca <- MCA(chairs[, -c(1:3)], graph = FALSE) #exclude the first three columns for the moment
> plot(chairs.ca, cex = 0.7, col.var = "black", col.ind = "grey")
```

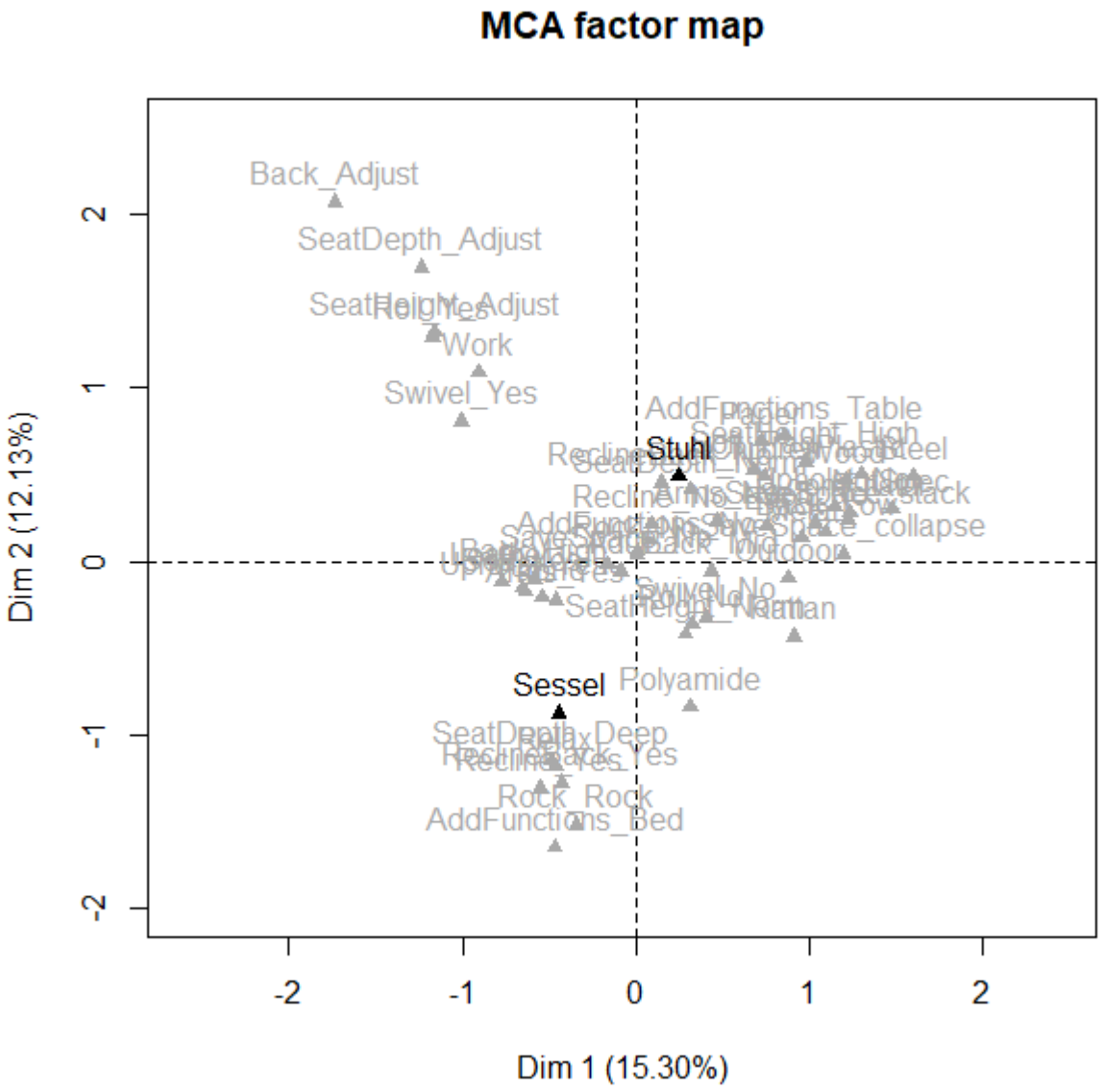

MCA factor map
Back_Adjust
SeatDepth_Adjust
SeatHeight_Adjust
Work
Swivel_Yes
AddFunctions_Table
Paper
SeatHeight_High
PlasticSteel
Wood
Recline_Arms_No
Outdoor
Swivel_No
SeatHeight_Norm
Rattan
Polyamide
Rock_Rock
AddFunctions_Bed
Dim 1 (15.30%)
Dim 2 (12.13%)
-2
-1
0
1
2

Interpretation of dimensions

```
> dimdesc(chairs.ca)
$`Dim 1`
$`Dim 1`$quali
                R2           p.value
Unholst     0.72940952 1.094774e-54
MaterialSeat0.74518860 3.215782e-48
Function    0.69158437 1.158923e-45
Soft        0.66568141 9.657154e-45
Swivel      0.40875670 5.393205e-23
Roll        0.38348403 2.728416e-21
…
```

MCA with supplementary points

```
> chairs.ca1 <- MCA(chairs[, -c(1:2)], quali.sup = 1, graph = FALSE)  #use lexical categories as supplementary points
> plot(chairs.ca1, invis = "ind", col.var = "darkgrey", col.quali.sup = "black") #make the individual points invisible
```


MCA factor map
Back_Adjust
SeatDepth_Adjust
Work
Swivel_Yes
Stuhl
Sessel
Polyamide
Rock_Rock
AddFunctions_Bed
Dim 2 (12.13%)
Dim 1 (15.30%)
-2
-1
0
1
2

How good is the solution?

```
> chairs.ca$eig
eigenvalue percentage of variance cumulative
percentage of variance
dim 1 0.3250725720 15.29753280 15.29753
dim 2 0.2576755177 12.12590671 27.42344
dim 3 0.1351901997 6.36189175  33.78533
…
```

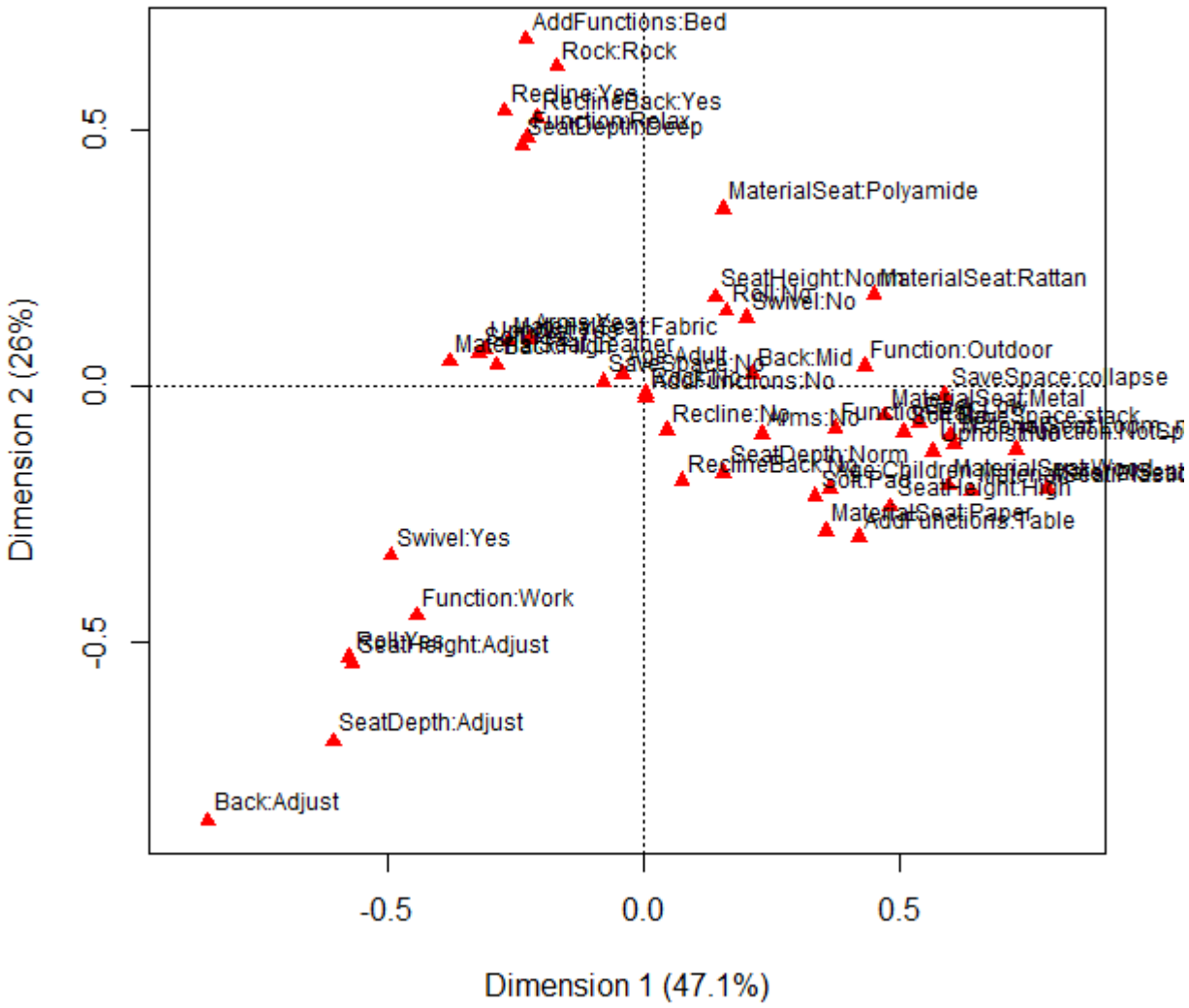
Doesn’t look too impressive. But these numbers should be taken with a grain of salt.

Adjusted MCA (Greenacre 2007)

```
> chairs.ca2 <- mjca(chairs[, -c(1:3)])
> summary(chairs.ca2)
Principal inertias (eigenvalues):
dim value % cum% scree plot
1 0.078443 47.1 47.1 ***************************
2 0.043342 26.0 73.2 ***************
3 0.006012 3.6  76.8 **
```

Not bad, after all! Is the solution similar?

```
> plot(chairs.ca2)
```


AddFunctions:Bed
Rock:Rock
Recline:Yes
ReclineBack:Yes
SeatDepth:Deep
MaterialSeat:Polyamide
SeatHeight:Norm
MaterialSeat:Rattan
Swivel:No
Back:Mid
Function:Outdoor
SaveSpace:collapse
MaterialSeat:Metal
Recline:No
SeatDepth:Norm
SeatHeight:High
MaterialSeat:Paper
AddFunctions:Table
Swivel:Yes
Function:Work
SeatHeight:Adjust
SeatDepth:Adjust
Back:Adjust
0.5
0.0
-0.5
-0.5
0.0
0.5
Dimension 2 (26%)
Dimension 1 (47.1%)

References

- Greenacre, M. 2016. *Correspondence Analysis in Practice*. 3rd edn. Boca Raton, FL: CRC Press.
- Husson, F., Lê, S., & Pagès, J. 2010. *Exploratory Multivariate Analysis by Example Using R*. Boca Raton, FL: Chapman and Hall/CRC Press.