Correspondence Analysis (CA)

Alireza Sheikh-Zadeh, PhD

Correspondence analysis provides a graphical method of exploring the relationship between categorical variables in a contingency table.

In summary:

- We are looking for similarity (or short distance) between row variables and column variables.
- In the CA, the distance between column variables of a contingency table is defined as the weighted sum of squared difference between proportions (probabilities) in the columns. Same process for finding the distance between row variables.

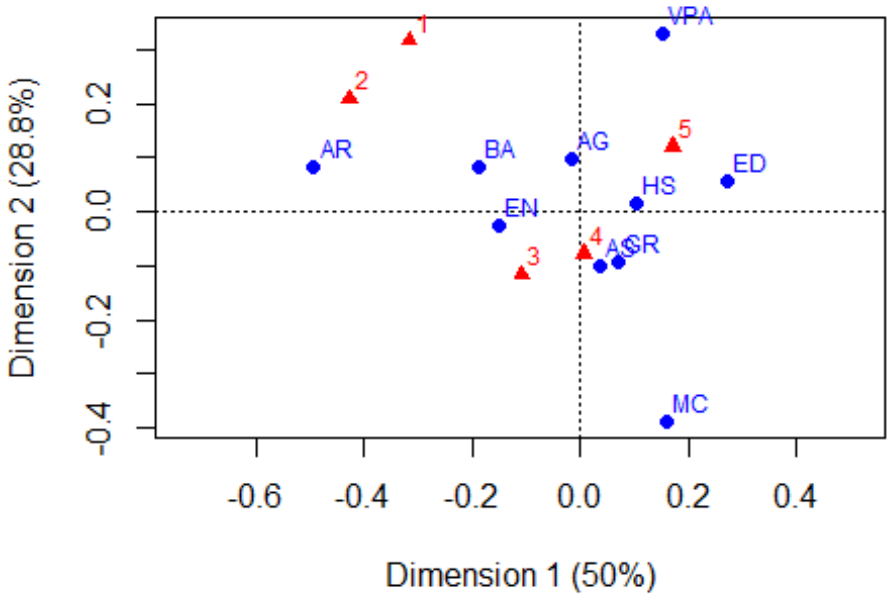
```
> pet <- c("dog", "cat", "NA",  "NA",  "NA",  "NA",  "dog", "NA",  "cat",
"NA")
> col <- c("BA", "BA", "EN", "EN", "BA", "EN", "EN", "BA", "BA", "BA")
> data <- data.frame(pet, col)

> table(data)
     col
pet   BA EN
  cat  2  0
  dog  1  1
  NA   3  3
```

(This process is explained in Video Lecture 15)

- In the CA, column points that are close together represent column categories with similar profiles (conditional distributions) across columns. Same interpretation for row points.
- If the angle between the row points A and column points B is less than 90-degree angle P(A|B)>P(A), which means the likelihood of A's occurrence given B is larger than the likelihood of A's occurrence in the entire sample space.
- If the angle between the row points A and column points, B is more than 90-degree angle P(A|B)<P(A), which means the likelihood of A's occurrence given B is less than the likelihood of A's occurrence in the entire sample space.

Example: TTU College Graduation Survey (2002)

```
grad <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/pgs.csv", header
= T)
nrow(grad)

## [1] 2002

ncol(grad)

## [1] 124

attach(grad)

#make a contingency table for college vs general rating
tbl = table(COL, FacTeaching)
tbl

##       FacTeaching
## COL      1   2   3   4   5
##   AG     4  15  26  78  56
##   AR     3   4   6  16   4
##   AS    12  24 124 290 171
##   BA     9  28  44 116  66
##   DUAL   0   0   2   0   0
##   ED     3   6  26 113  93
##   EN     5  36  65 168  86
##   GR     0   3   8  27  15
##   HS     1   5  17  41  33
##   MC     0   0   3  25   6
##   VPA    4   7  10  37  44

library(ca)

grad.ca <- ca(tbl <- tbl[-5, ]) # removing dual
plot(grad.ca)
```




```
#Lets focus on the conditional distribution of ratings among specific college
s: AS, BA, GR, and VPA
prop.table(tbl, 1)[c(3, 4, 7, 10), ]
```

```
##      FacTeaching
## COL        1      2      3      4      5
##   AS  0.0193 0.0386 0.1997 0.4670 0.2754
##   BA  0.0342 0.1065 0.1673 0.4411 0.2510
##   GR  0.0000 0.0566 0.1509 0.5094 0.2830
##   VPA 0.0392 0.0686 0.0980 0.3627 0.4314
```

How to interpret?

- The distribution of AS and GR among ratings is more similar than the distribution of VPA among ratings.