Lesson 12 - Correspondence Analysis

Erin M. Buchanan

04/04/2019

Language Topics Discussed

- A reanalysis of color terms and categories

Simple Correspondence Analysis

```r
library(Rling)
data(colreg)
head(colreg)
```

```
##        spoken fiction academic press
## black   20335   41118    26892 73080
## blue     4693   22093     3605 21210
## brown    1185   10914     1201 11539
## gray     1168   12140     1289  6559
## green    3860   14398     4477 26837
## orange    931    3496      474  5766
```

Basic Color Terms

- Berlin and Kay (1969) proposed a theory about our linguistic interpretations of colors, mainly that color vocabulary falls into universal categories:

$$\begin{bmatrix} \text{white} \\ \text{black} \end{bmatrix} < \begin{bmatrix} \text{red} \end{bmatrix} < \begin{bmatrix} \text{green} \\ \text{yellow} \end{bmatrix} < \begin{bmatrix} \text{blue} \end{bmatrix} < \begin{bmatrix} \text{brown} \end{bmatrix} < \begin{bmatrix} \text{purple} \\ \text{pink} \\ \text{orange} \\ \text{grey} \end{bmatrix}$$

Chi-square

- Chi-square analyses tell us if specific category frequencies are different than we might expect.
- Let's look at a simple combination to understand the math behind chi-square.

```
cs_example = colreg[1:2, 1:2]
cs_example
```

```
##       spoken fiction
## black  20335   41118
## blue    4693   22093
```

Expected values

$$E = \frac{Row * Column}{N}$$

```
cs_example_e = cs_example
rows = rowSums(cs_example)
columns = colSums(cs_example)

cs_example_e[1,1] = rows[1]*columns[1]/sum(cs_example)
cs_example_e[1,2] = rows[1]*columns[2]/sum(cs_example)
cs_example_e[2,1] = rows[2]*columns[1]/sum(cs_example)
cs_example_e[2,2] = rows[2]*columns[2]/sum(cs_example)
cs_example_e
```

```
##          spoken  fiction
## black 17430.452 44022.55
## blue   7597.548 19188.45
```

```
cs_test = chisq.test(cs_example)
cs_test$expected
```

Chi-square Formula

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

```
sum((cs_example - cs_example_e)^2 / cs_example_e)
```

```
## [1] 2225.712
```

```
cs_test$statistic
```

```
## X-squared
##  2224.946
```

Chi-Square - What Next?

- This test doesn't tell you *what* was different though, much like ANOVA
- The way to know what cells were higher/lower than expected would be to use standardized residuals

Residuals

- Residuals are (O-E)/sqrt(E), whereas standardized residuals are standardized format akin to z-scores (O-E)/ sqrt(var(residuals))

```
cs_test$residuals #(cs_example - cs_example_e) /sqrt(cs_exa
```

```
##           spoken   fiction
## black  22.00008 -13.84334
## blue  -33.32282  20.96807
```

```
cs_test$stdres
```

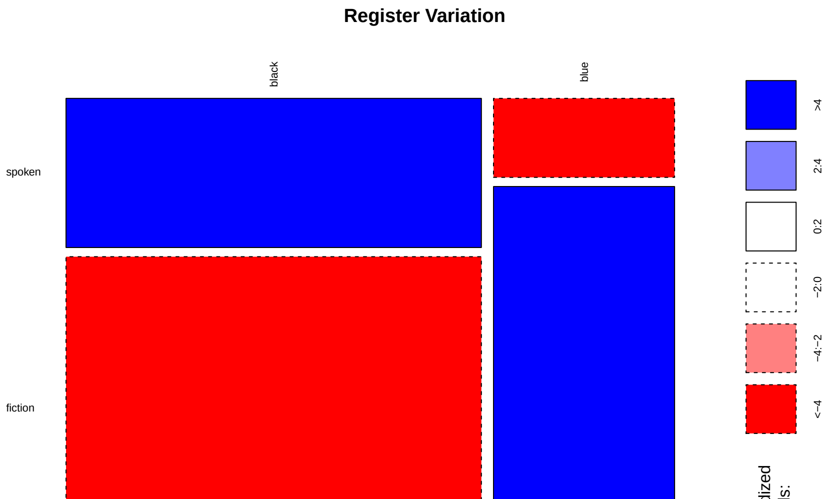
```
##           spoken   fiction
## black  47.17745 -47.17745
## blue  -47.17745  47.17745
```

Mosaic Plots

- A visualization of the standardized residuals from a chi-square type analysis
- The box size is related to the observed cell size
- Coloring is shaded based on direction and strength of the residuals

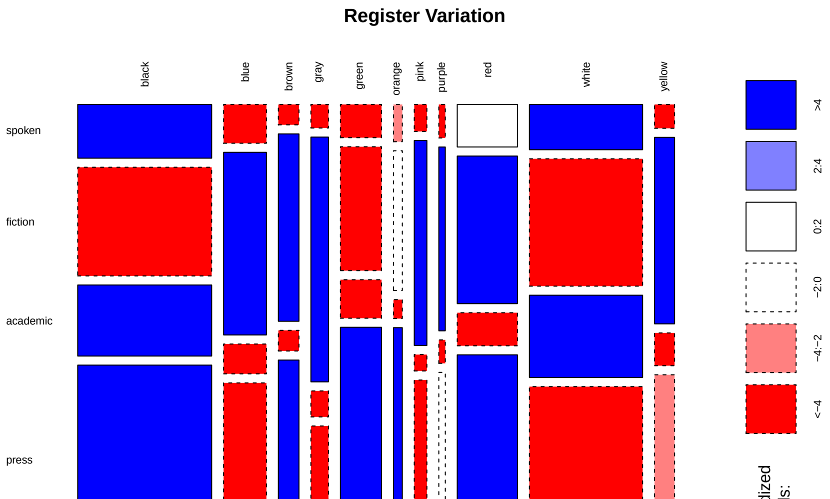
Mosaic Plots - Small Example

```
mosaicplot(colreg[1:2, 1:2], #data frame
           las = 2, #axis label style (perpendicular)
           shade = T, #color in the boxes
           main = "Register Variation")
```



Mosaic Plot - Full Data

```
mosaicplot(colreg, #data frame
           las = 2, #axis label style (perpendicular)
           shade = T, #color in the boxes
           main = "Register Variation")
```



Simple Correspondence Analysis

- Identifies systematic relationships between variables in low dimensional space
- Similar to MDS, PCA, EFA

```r
library(ca)
sca_model = ca(colreg)
```

What's in the output?

```r
summary(sca_model)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.043730  77.9  77.9  *******************
##  2      0.010787  19.2  97.1  *****
##  3      0.001650   2.9 100.0  *
##         -------- -----
##  Total: 0.056167 100.0
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  | blck |  281  980  193 | -193 961 238 |   27  19  19
## 2  | blue |   90  947   89 |  226 919 105 |  -40  28  13
## 3  | brwn |   43  957   85 |  323 949 103 |   30   8   4
```

Inertia

- Top part is the table of inertias, which explain how much variation is accounted for by each dimension
- These are similar to eigenvalues that we've seen in the last several analyses

Inertia

- Try to represent the relationship between variables in as few dimensions as possible
- Here we see that the first two dimensions capture 97% of the variance
- And the third dimension captures all the variance

Principal inertias (eigenvalues):

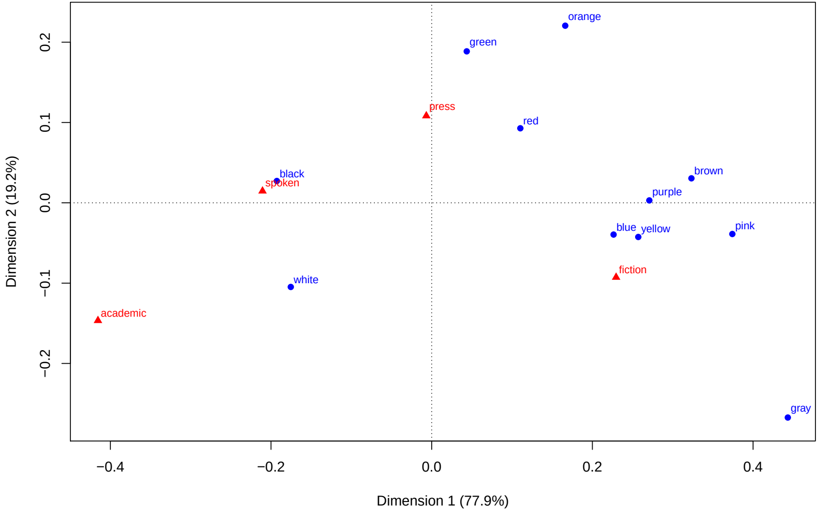
dim value % cum% scree plot
1 0.043730 77.9 77.9 *******************
2 0.010787 19.2 97.1 *****
3 0.001650 2.9 100.0 *

—— ——

Total: 0.056167 100.0

Visualize the Dimensions

```
plot(sca_model)
```

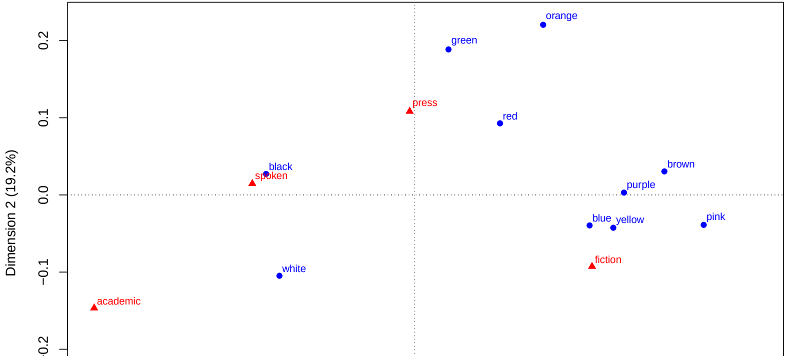


Key Differences

- How are these plots different than the ones we've been making?
  - Terms are close together if they have similar frequency counts
  - This means the rows have similar *profiles* - rather than similar relationships to a latent variable
  - The distances on the map are a representation of the χ^2 values of each row/column to the average profile

Does this match theory?

[white black] < [red] < [green yellow] < [blue] < [brown] < [purple pink orange grey]



Some other interesting notes

- Press is close to green-red because of the political orientation for these terms and proper names (Red Cross/Green Bay Packers)
- Fiction is likely close to the later color terms because of the requirement to “paint a picture” for readers
- Appears academics and spoken speech are pretty boring in their use of color terms

3D Plots

```
plot3d.ca(sca_model, #model
       labels = c(1,1)) #see both row and column labels
```

```
## Loading required namespace: rgl
```

A Quick Reminder of Categories

- What is a category?
  - Category – group or organization of related things
  - Concept – a member of a category (i.e. the thing)
  - Animals: dog, cat, bird, fish

Family Resemblance Models

- Prototype theory versus exemplar theory
  - Prototype – an abstraction that is the best example of a category
  - Prototypes are likely a combination of experienced examples, but may not exist in real world
  - Exemplar theory – we compare information to a specific stored example
  - Instantiation principle – category includes detailed information about the range of instances
- These are very similar in their ideas, but the underlying core is distinction

A Category Example

- Is there a difference between the categories for *stuhl* (chair) and *sessel* (armchair)?
- Gipper (1959) had subjects name pictures of chairs to determine their relative frequencies
- The difference appeared to be that chairs are functional, while armchairs are about comfort

The Data

- Data was coded from an online shopping place based on their text descriptions and other chair related variables

```r
data(chairs)
head(chairs)
```

```
##                 Shop                      WordDE Category Func
## 1 Moebel-Profi.de                      3D-Stuhl    Stuhl
## 2         ikea.de                   Jugendstuhl    Stuhl
## 3         ikea.de                        Sessel   Sessel  Not
## 4 Moebel-Profi.de                   Swingstuhl    Stuhl
## 5         ikea.de Kinderstuhl_mit_Sitzgurt    Stuhl
## 6       roller.de                    Drehstuhl    Stuhl
##   Soft Arms Upholst MaterialSeat SeatHeight SeatDepth Sw
## 1   No   No      No      Plastic       Norm      Norm
## 2   No   No      No         Wood       High      Norm
## 3   No  Yes      No       Rattan       Norm      Norm
## 4  Yes   No     Yes       Fabric       Norm      Norm
## 5   No  Yes      No      Plastic       High      Norm
```

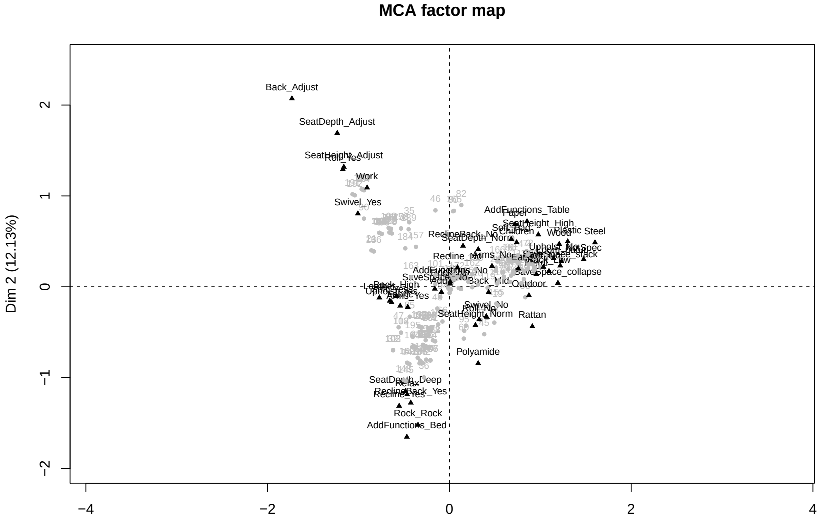
Multiple Correspondence Analysis

```
library(FactoMineR)
mca_model = MCA(chairs[ , -c(1:3)], #dataset minus the firs
                graph = FALSE)
summary(mca_model)
```

```
##
## Call:
## MCA(X = chairs[, -c(1:3)], graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   D
## Variance               0.325   0.258   0.135   0.123   0
## % of var.             15.298  12.126   6.362   5.785   5
## Cumulative % of var.  15.298  27.423  33.785  39.570  44
##                        Dim.7   Dim.8   Dim.9  Dim.10  Di
## Variance               0.090   0.086   0.082   0.073   0
## % of var.              4.244   4.056   3.843   3.419   3
## Cumulative % of var.  53.466  57.523  61.365  64.784  68
```

Plot the MCA

```
plot(mca_model, cex = .7,
     col.var = "black", #color the variable names
     col.ind = "gray") #color the indicators
```



How Useful are the Variables?

```
dimdesc(mca_model)
```

```
## $`Dim 1`
## $`Dim 1`$quali
##                     R2
## Upholst      0.72940952
## MaterialSeat 0.74518860
## Function     0.69158437
## Soft         0.66568141
## Swivel       0.40875670
## Roll         0.38348403
## SeatHeight   0.39565748
## Back         0.36654364
## Arms         0.21473392
## SeatDepth    0.20909906
## SaveSpace    0.19444992
## Age          0.06521465
## ReclineBack  0.06368029
## Recline      0.04908474
```

Interpretation

- R^2 values representing the variables association with the dimension
- p value strength of that association
- Then the `$category` section represents the directionality of the relationship
  - If this value is positive, shows on the right hand side of plot, representing a positive coefficient (and vice versa)

Overall interpretation

- First dimension seems to represent comfort chairs versus not
- Second dimensions seems to represent functionality (work versus home)
- Third is harder to understand
- Appears to separate chairs into three categories:
  - Comfortable relaxation chairs
  - Comfortable adjustable chairs for work
  - Multifunctional chairs for the house

Using the chair label

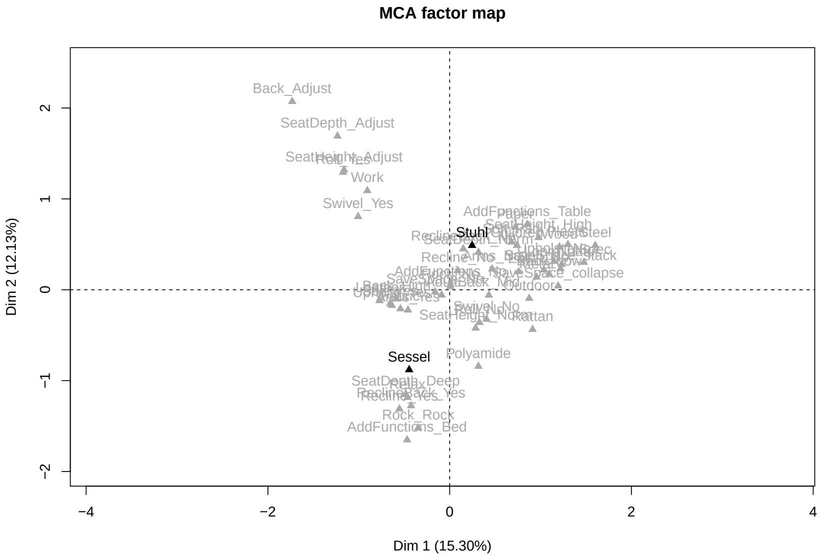
- We did not use the type of chair that is found in column 3 of our dataset
- We can map it onto our analysis using it as a supplementary variable

Running that analysis

```
mca_model2 = MCA(chairs[ , -c(1,2)],
                  quali.sup = 1, #supplemental variable
                  graph = FALSE)
```

Plot that analysis

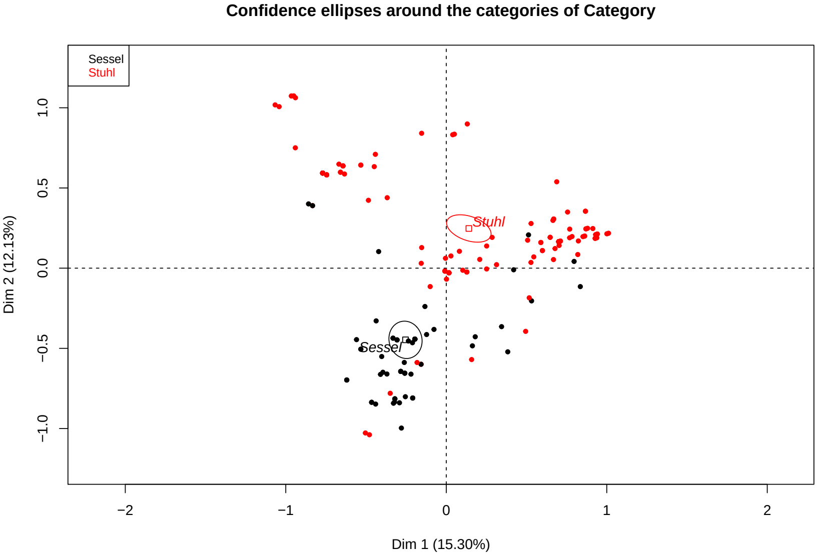
```
plot(mca_model2, invis = "ind", col.var = "darkgray", col.c
```




```
#the invis turned off the individual points
```

Examine the prototypes

```
plotellipses(mca_model2, keepvar = 1, #use column 1 to labe
             label = "quali")
```



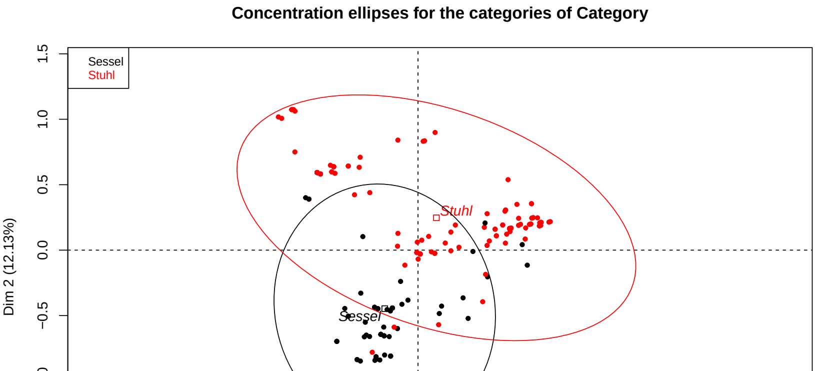
Interpretation

- These confidence ellipses do not overlap, so we could consider the prototypes distinct entities
- We can also create a more traditional 95% CI type interval

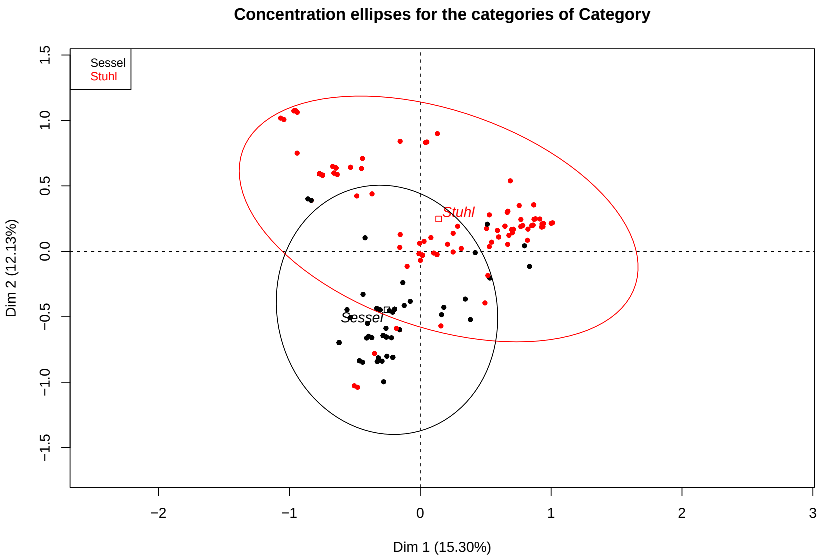
95% Ellipses

- Now you can see that the categories themselves overlap a lot, so likely a representation of the fuzzy boundaries that categories appear to have.

```
plotellipses(mca_model2,
             means = F,
             keepvar = 1, #use column 1 to label
             label = "quali")
```



The plot



Concentration ellipses for the categories of Category
Sessel
Stuhl
Stuhl
Sessel
Dim 1 (15.30%)
Dim 2 (12.13%)
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

But what about inertia?

```r
mca_model2$eig
```

```
##        eigenvalue percentage of variance
## dim 1  0.3250725720            15.29753280
## dim 2  0.2576755177            12.12590671
## dim 3  0.1351901997             6.36189175
## dim 4  0.1229322264             5.78504595
## dim 5  0.1089102792             5.12518961
## dim 6  0.0961853064             4.52636736
## dim 7  0.0901939195             4.24441974
## dim 8  0.0861985147             4.05640069
## dim 9  0.0816542710             3.84255393
## dim 10 0.0726465359             3.41866051
## dim 11 0.0706639812             3.32536382
## dim 12 0.0654187968             3.07853161
## dim 13 0.0614776588             2.89306630
## dim 14 0.0604269465             2.84362101
## dim 15 0.0545085623             2.56510882
```

Inertia part 2

```
mca_model3 = mjca(chairs[ , -c(1:3)])
summary(mca_model3)
```

```
##
## 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    *
##  4      0.004155   2.5  79.3    *
##  5      0.002451   1.5  80.8
##  6      0.001291   0.8  81.5
##  7      0.000873   0.5  82.1
##  8      0.000639   0.4  82.4
##  9      0.000417   0.3  82.7
##  10     0.000117   0.1  82.8
##  11     0.000076   0.0  82.8
##  12     0.000010   0.0  82.8
```

Further work

- From here, you could take the dimension scores `mca_model2$ind$coord` and use them to predict the categories or other variables
- This analysis would tell you how good at representing their categories each dimension does

Summary

- We applied new models to basic color terms and category groupings
- We learned how to do simple and multiple correspondence analysis