Statistics 452: Statistical Learning and Prediction

Chapter 10, part 1.5: Introduction to Multiple Correspondence Analysis

Brad McNeney

2018-11-14

Multiple Correspondence Analysis (MCA)

- An exploratory analysis methodology for multivariate datasets with categorical variables.
- In basic form, it is PCA on dummy variables that represent the categorical variables.
- Illustrate with the health utilities index (HUI) variables from the Canadian Community Health Survey - Healthy Aging

HUI Data

- Read from the Stat 652 Project folder.

```
hui <- read.csv("../../Project652/HUI.csv")
names(hui)
```

```
## [1] "DHHGAGE" "DHH_SEX" "HUIDCOG" "HUIGDEX" "HUIDEMO" "HUIGHER" "HUIGMOB"
## [8] "HUIGSPE" "HUIGVIS"
```

Summaries

```
summary(hui)
```

```
##             DHHGAGE       DHH_SEX                    HUIDCOG
##  55 TO 59 YEARS:3085   FEMALE:11385   COG. ATT. LEVE 1:13949
##  60 TO 64 YEARS:2982   MALE  : 8615   COG. ATT. LEVE 2:  496
##  85 AND OLDER  :2602                  COG. ATT. LEVE 3: 3764
##  65 TO 69 YEARS:2595                  COG. ATT. LEVE 4: 1268
##  70 TO 74 YEARS:1958                  COG. ATT. LEVE 5:  429
##  75 TO 79 YEARS:1928                  COG. ATT. LEVE 6:   71
##  (Other)       :4850                  NOT STATED      :   23
##             HUIGDEX                   HUIDEMO                   HUIGHER
##  LIM. HANDS/F   :  252   EMOT. ATT. LEV.1:14912   NO PROBLEMS    :17335
##  NOT STATED     :   10   EMOT. ATT. LEV.2: 4067   NOT STATED     :  296
##  USE OF HANDS/F.:19738   EMOT. ATT. LEV.3:  749   PROB./CORR.    : 1579
##                          EMOT. ATT. LEV.4:  183   PROB./NOT CORR.:  790
##                          EMOT. ATT. LEV.5:   39
##                          NOT STATED      :   50
##
##             HUIGMOB                   HUIGSPE                   HUIGVIS
##  NEED MECH. SUPP: 1580   NO PROBLEMS    :19837   NO PROBLEMS     : 4210
##  NO AID REQUIRED:  322   NOT STATED     :   11   NOT STATED      :  142
##  NO PROBLEMS    :17496   PARTIAL/NOT UND.:  152   VISUAL P. UNCOR.:  658
##  NOT STATED     :   16                            VISUAL PROB. COR:14990
##  REQUIRES HELP  :  586
##
```

Remove records with missing values

- I will consider the response `NOT STATED` to be missing data.
  - Remove subjects with any missing data

```
recode_ns <- function(x) {
  x[x=="NOT STATED"] <- NA
  x <- droplevels(x)
  x
}
for(i in 1:ncol(hui)) {
  hui[,i] <- recode_ns(hui[,i])
}
hui <- na.omit(hui)
dim(hui)
```

```
## [1] 19523     9
```

HUIDCOG

- Cognitive function (our focus) with levels:

1. Able to remember most things, think clearly and solve day to day problems
2. Able to remember most things, but have a little difficulty when trying to think and solve day to day problems
3. Somewhat forgetful, but able to think clearly and solve day to day problems
4. Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems
5. Very forgetful, and have great difficulty when trying to think or solve day to day problems
6. Unable to remember anything at all, and unable to think or solve day to day problems

```r
levels(hui$HUIDCOG) <- as.character(1:6)
table(hui$HUIDCOG)
```

```
##
##     1     2     3     4     5     6
## 13708   481  3664  1212   400    58
```

Pairwise summaries

- Relationship between HUIDCOG and others

```
tt <- xtabs(~DHHGAGE+HUIDCOG,data=hui)
tt
```

```
##                     HUIDCOG
## DHHGAGE                1    2    3    4    5    6
##   45 TO 49 YEARS    1158   34  273   76   29    1
##   50 TO 54 YEARS    1307   45  279   83   32    2
##   55 TO 59 YEARS    2297   62  515  126   52    1
##   60 TO 64 YEARS    2201   55  511  127   41    6
##   65 TO 69 YEARS    1907   52  444  105   25    3
##   70 TO 74 YEARS    1351   38  383  112   21    4
##   75 TO 79 YEARS    1236   51  397  138   39    5
##   80 TO 84 YEARS     891   36  328  131   36    9
##   85 AND OLDER      1360  108  534  314  125   27
```

Pairwise summaries, cont.

- Age distributions for each cognitive level
  - Proportions of column variable for each row (level of `HUIDCOG`).

```
round(prop.table(tt,margin=2),2)
```

```
##                  HUIDCOG
## DHHGAGE             1    2    3    4    5    6
##   45 TO 49 YEARS 0.08 0.07 0.07 0.06 0.07 0.02
##   50 TO 54 YEARS 0.10 0.09 0.08 0.07 0.08 0.03
##   55 TO 59 YEARS 0.17 0.13 0.14 0.10 0.13 0.02
##   60 TO 64 YEARS 0.16 0.11 0.14 0.10 0.10 0.10
##   65 TO 69 YEARS 0.14 0.11 0.12 0.09 0.06 0.05
##   70 TO 74 YEARS 0.10 0.08 0.10 0.09 0.05 0.07
##   75 TO 79 YEARS 0.09 0.11 0.11 0.11 0.10 0.09
##   80 TO 84 YEARS 0.06 0.07 0.09 0.11 0.09 0.16
##   85 AND OLDER   0.10 0.22 0.15 0.26 0.31 0.47
```

Pairwise summaries, cont.

- Relationship between HUIDCOG and HUIDEX.

```
tt <- xtabs(~HUIGDEX+HUIDCOG,data=hui)
tt
```

```
##                   HUIDCOG
## HUIGDEX                 1     2     3     4     5     6
##   LIM. HANDS/F        106    11    45    31    28    12
##   USE OF HANDS/F. 13602   470  3619  1181   372    46
```

Pairwise summaries, cont.

```
round(prop.table(tt,margin=2),2)
```

```
##                   HUIDCOG
## HUIGDEX                1    2    3    4    5    6
##   LIM. HANDS/F      0.01 0.02 0.01 0.03 0.07 0.21
##   USE OF HANDS/F.   0.99 0.98 0.99 0.97 0.93 0.79
```

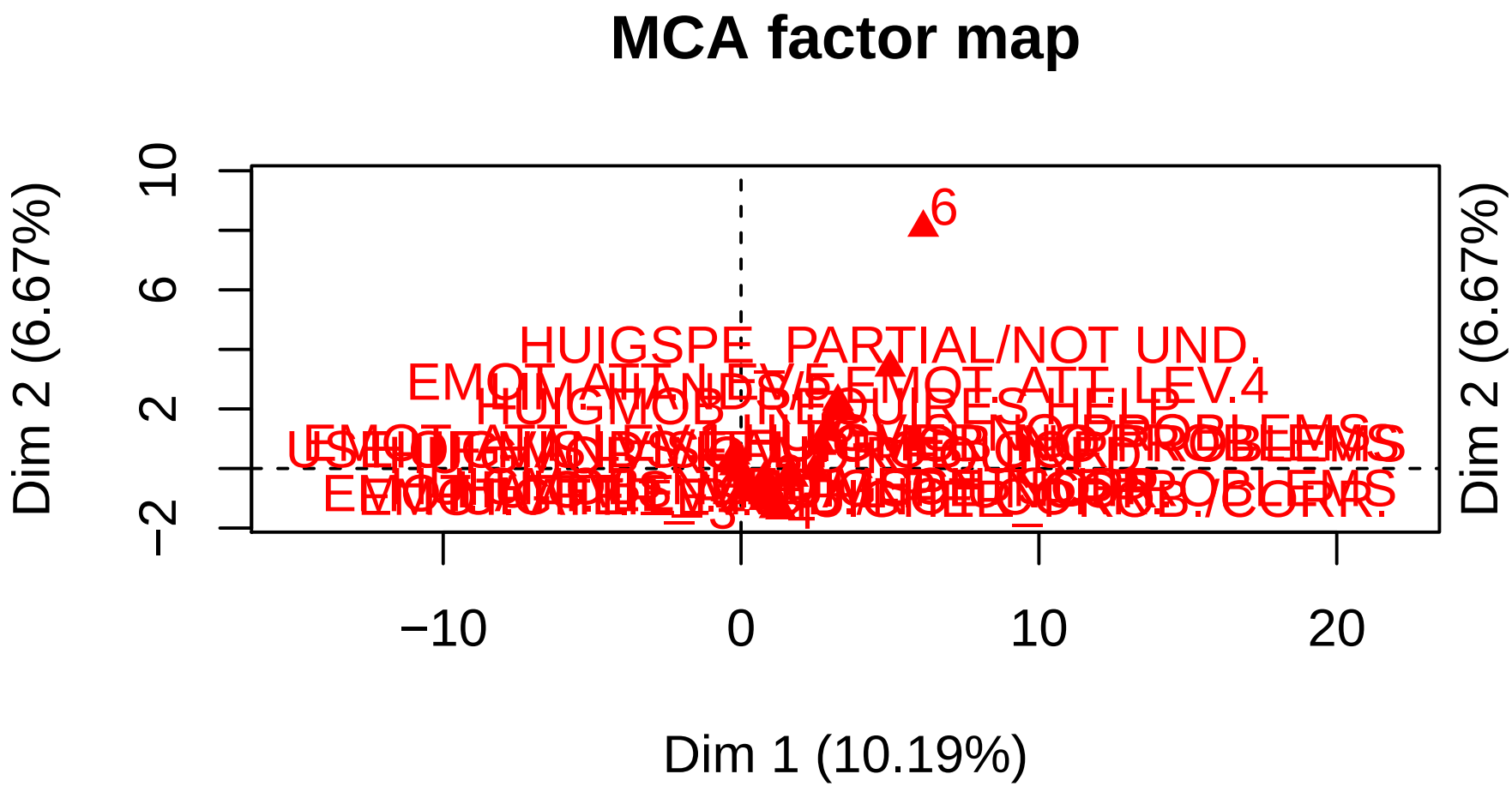
- And so on . . .

Correspondence Analysis (CA)

- Write the categorical variables as dummy variables and do a biplot of the result; e.g., for the HUIs.

```
library(dplyr)
hsub <- select(hui,starts_with("HUI"))
X <- model.matrix(~.,data=hsub)[,-1]
X <- scale(X)
pp <- prcomp(X)
# biplot(pp) -- too many points, too messy
```

```
library(FactoMineR)
res.mca <- MCA(hsub)
```



```r
plot(res.mca,invisible=c("ind"),cex=0.5)
```