Assignment 8: Exploratory factor analysis

DA 410

Marjorie Blanco

13.7 Use the words data of Table 5.9.

(a) Obtain principal component loadings for two factors.

You may use R to solve this part (NO built-in function). Follow the 5-steps of the example introduced in the handout to obtain principal component loadings for two factors. Make sure to show your work completely.

```
data <-read.table("Software-Files/T5_9_ESSAY.DAT")
colnames(data) <- c("student" ,"Informal words", "Informal verbs",
                    "Formal words", "Formal verbs")
data <- data[,2:5]
```

Bartlett's Test of Sphericity

```
cortest.bartlett(data)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 29.27297
##
## $p.value
## [1] 5.401014e-05
##
## $df
## [1] 6
```

Bartlett's test was statistically significant, suggesting that the observed correlation matrix among the items is not an identity matrix.

KMO

```
KMO(data)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data)
## Overall MSA =  0.5
## MSA for each item =
## Informal words Informal verbs   Formal words   Formal verbs
##           0.45           0.58           0.55           0.41
```

The overall KMO for our data is 0.5.

Step 1: Find correlation matrix R.

```
n <- nrow(data)
C <- diag(n) - matrix(1/n, n, n)
D <- diag(apply(as.matrix(data), 2, sd))
Xs <- C %*% as.matrix(data) %*% solve(D)
R <- t(Xs) %*% Xs / (n-1)
```

```
rownames(R) <- colnames(R) <- colnames(data)
```

R=

	Informal words	Informal verbs	Formal words	Formal verbs
Informal words	1.0000000	0.7660725	0.5953551	0.2173378
Informal verbs	0.7660725	1.0000000	0.5600505	0.4427548
Formal words	0.5953551	0.5600505	1.0000000	0.7202028
Formal verbs	0.2173378	0.4427548	0.7202028	1.0000000

```
# Calculate the correlation matrix
res.cor <- cor(data)
```

R=

	Informal words	Informal verbs	Formal words	Formal verbs
Informal words	1.0000000	0.7660725	0.5953551	0.2173378
Informal verbs	0.7660725	1.0000000	0.5600505	0.4427548
Formal words	0.5953551	0.5600505	1.0000000	0.7202028
Formal verbs	0.2173378	0.4427548	0.7202028	1.0000000

```
n <- nrow(data)
C <- diag(n) - matrix(1/n, n, n)
Xc <- C %*% as.matrix(data)
S <- t(Xc) %*% Xc / (n-1)
```

S=

	Informal words	Informal verbs	Formal words	Formal verbs
Informal words	1405.78095	153.70952	804.7667	43.10476
Informal verbs	153.70952	28.63810	108.0524	12.53333
Formal words	804.76667	108.05238	1299.7810	137.34762
Formal verbs	43.10476	12.53333	137.3476	27.98095

```
res.cov <- cov(data)
```

S=

	Informal words	Informal verbs	Formal words	Formal verbs
Informal words	1405.78095	153.70952	804.7667	43.10476
Informal verbs	153.70952	28.63810	108.0524	12.53333
Formal words	804.76667	108.05238	1299.7810	137.34762
Formal verbs	43.10476	12.53333	137.3476	27.98095

Step 2: Find the eigenvalue D and eigenvectors C of R.

```r
# Then use that correlation matrix to calculate eigenvalues
res.eig <- eigen(res.cor, symmetric = FALSE)
res.eig
```

```
## eigen() decomposition
## $values
## [1] 2.6657459 0.8993358 0.3276382 0.1072801
##
## $vectors
##            [,1]       [,2]       [,3]       [,4]
## [1,] 0.4914201  0.5642628 -0.3208298  0.5806737
## [2,] 0.5241023  0.3441774  0.6604771 -0.4130722
## [3,] 0.5409202 -0.2848265 -0.6020434 -0.5136371
## [4,] 0.4372967 -0.6942790  0.3136590  0.4778769
```

Step 3: Find C_1 and D_1

```r
c.1 <- res.eig$vectors[,1:2]
d.1 <- diag(res.eig$values[1:2])
```

C_1=

0.4914201	0.5642628
0.5241023	0.3441774
0.5409202	-0.2848265
0.4372967	-0.6942790

D_1=

2.665746	0.0000000
0.000000	0.8993358

Step 4: Find $C_1 D_1^{1/2}$

```r
l <- as.data.frame(c.1 %*% sqrt(d.1))
```

$C_1 D_1^{1/2}$=

V1	V2
0.8023471	0.5351091
0.8557077	0.3263949
0.8831664	-0.2701104
0.7139792	-0.6584078

Step 5: Obtain loadings

```r
l[,3] <- l[,1]^2 + l[,2]^2
l[,4] <- 1 - l[,3]
```

```r
prop <- res.eig$values[1:2]/sum(res.eig$values)
cumprop <- c(prop[1], sum(prop))
cumulative.proportion <- 0
prop <- c()
cumulative <- c()
for (i in res.eig$values) {
  proportion <- i / dim(data)[2]
  cumulative.proportion <- cumulative.proportion + proportion
  prop <- append(prop, proportion)
  cumulative <- append(cumulative, cumulative.proportion)
}
data.frame(cbind(prop, cumulative))
```

```
##         prop cumulative
## 1 0.66643647  0.6664365
## 2 0.22483396  0.8912704
## 3 0.08190955  0.9731800
## 4 0.02682002  1.0000000
```

```r
factors <- t(t(res.eig$vectors[,1:2]) * sqrt(res.eig$values[1:2]))
round(factors, 2)
```

```
##      [,1]  [,2]
## [1,] 0.80  0.54
## [2,] 0.86  0.33
## [3,] 0.88 -0.27
## [4,] 0.71 -0.66
```

```r
h2 <- rowSums(factors^2)
h2
```

```
## [1] 0.9301027 0.8387693 0.8529425 0.9432671
```

```r
u2 <- 1 - h2
u2
```

```
## [1] 0.06989729 0.16123067 0.14705746 0.05673285
```

```r
com <- rowSums(factors^2)^2 / rowSums(factors^4)
com
```

```
## [1] 1.742660 1.284950 1.185457 1.987011
```

```r
mean(com)
```

```
## [1] 1.55002
```

	PC1	PC2	h2	u2	com
Informal words	0.802	0.535	0.930	0.070	1.743
Informal verbs	0.856	0.326	0.839	0.161	1.285
Formal words	0.883	-0.270	0.853	0.147	1.185
Formal verbs	0.714	-0.658	0.943	0.057	1.987

Variance accounted for 2.6657459, 0.8993358 3.5650817

Proportion accounted for 0.6664365, 0.224834, 0.0819095, 0.02682 1

Cumulative proportion 0.6664365, 0.8912704 1

The first two factors account for ($2.6657 + 0.8993)/4 = 1$ of the total sample variance. 100 % of the variance explained by two factors is very high.

Informal words, Informal words, Formal words all have high factor loadings around 0.8 on the first factor (PC1).

Decide how many factors to retain.

For the probe data set, I recomend to retain 1 factor (PC1)

Method 1: Choose m equal to the number of factors necessary for the variance accounted for to achieve a predetermined percentage

An appropriate threshold percentage should be selected prior to starting the process. If we want to explain at least 70% of variance then we would select $PC1$ and $PC2$

Method 2: Choose m equal to the number of eigenvalues greater than the average eigenvalue.

Eigenvalues for $PC1$ is 1. In the probe data, retaining only $PC1$ is recomended.

Method 3: Scree plot

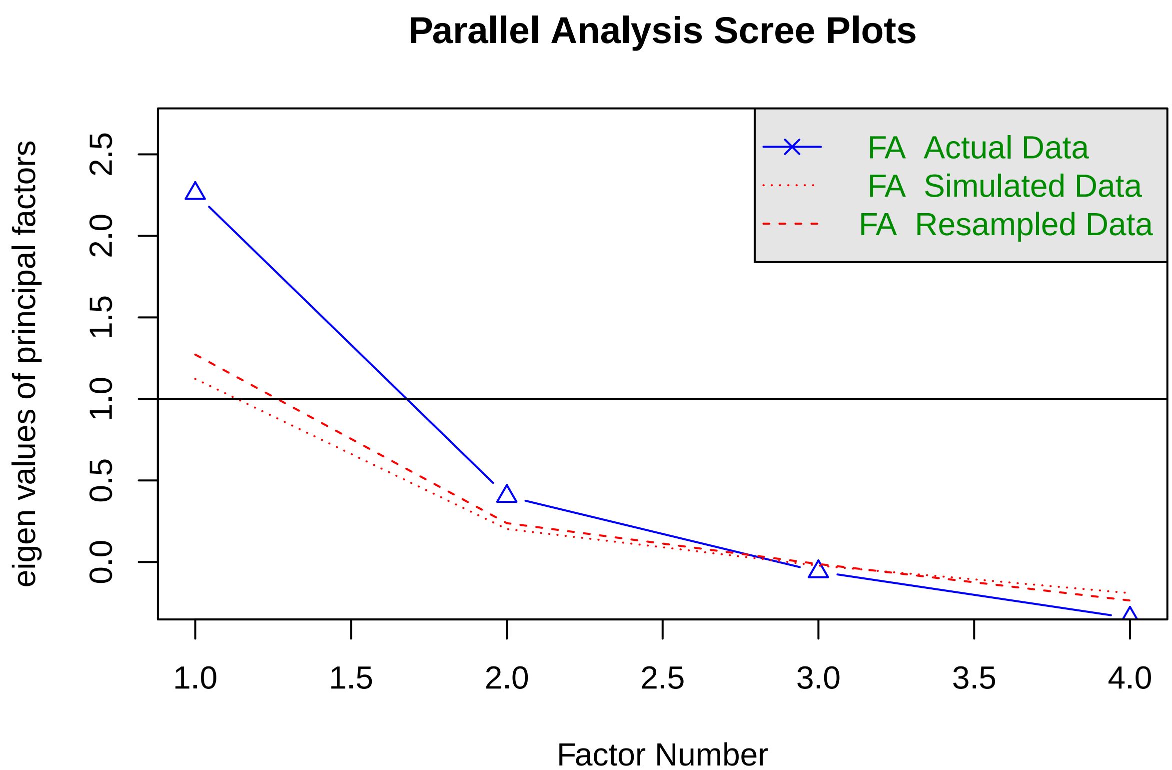
The number of points after point of inflexion. For this plot, retaining $PC1$ is recomended.

```
fa <- principal(data, nfactors = 2, rotate = 'none')
fa
```

```
## Principal Components Analysis
## Call: principal(r = data, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PC1   PC2   h2    u2 com
## Informal words 0.80 -0.54 0.93 0.070 1.7
## Informal verbs 0.86 -0.33 0.84 0.161 1.3
## Formal words   0.88  0.27 0.85 0.147 1.2
## Formal verbs   0.71  0.66 0.94 0.057 2.0
##
##                        PC1  PC2
## SS loadings           2.67 0.90
## Proportion Var        0.67 0.22
## Cumulative Var        0.67 0.89
## Proportion Explained  0.75 0.25
## Cumulative Proportion 0.75 1.00
```

```
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  0.95  with prob <  NA
##
## Fit based upon off diagonal values = 0.98
```

```r
parallel <- fa.parallel(data, fm = 'minres', fa = 'fa')
```




```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  NA
```