Nicholas Anthony Small

November 1st 2020

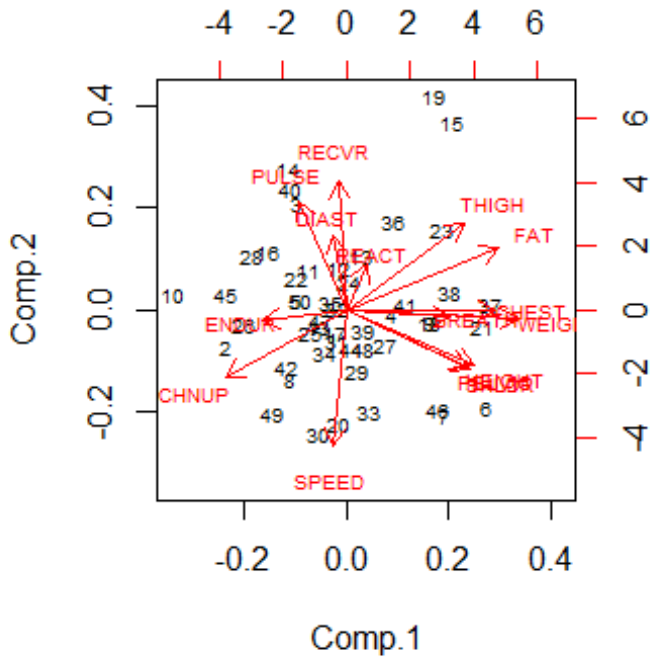
Problem 1

Use the police applicant data:

```
police <- read.csv("https://bit.ly/police_applications")
```

a) Perform principal component analysis using the correlation matrix. You do not need to do data cleaning or fixing the direction of variables in this data.

```
library(factoextra)

pc <- princomp(police, cor = T)
biplot(pc, col=c("black", "red"), cex =0.6)
```



b) What percentage of the total variance is covered by the first two principal components?

```
# 51% Percentage of the total variance is covered by the first two principal components.
summary(pc, cor = T)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.
5
## Standard deviation     2.2874533 1.5681077 1.14008912 1.11087426 1.0750293
```

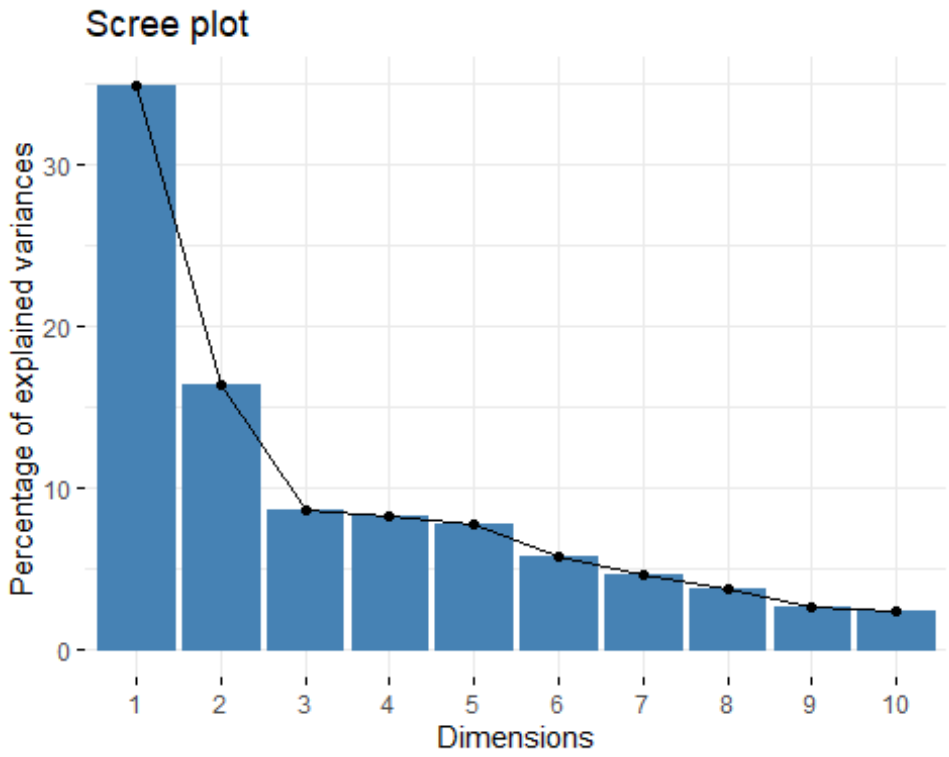
```
4
## Proportion of Variance 0.3488295 0.1639308 0.08665355 0.08226944 0.0770458
7
## Cumulative Proportion  0.3488295 0.5127603 0.59941382 0.68168326 0.7587291
3
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.
10
## Standard deviation     0.9286329 0.83740056 0.75504829 0.62158268 0.604639
75
## Proportion of Variance 0.0574906 0.04674931 0.03800653 0.02575767 0.024372
61
## Cumulative Proportion  0.8162197 0.86296905 0.90097558 0.92673324 0.951105
86
##                            Comp.11    Comp.12     Comp.13    Comp.14     Co
mp.15
## Standard deviation     0.57163001 0.43703530 0.365630031 0.20815295 0.1965
66621
## Proportion of Variance 0.02178406 0.01273332 0.008912355 0.00288851 0.0025
75896
## Cumulative Proportion  0.97288992 0.98562324 0.994535594 0.99742410 1.0000
00000
```

```
round(100*sum((pc$sdev^2)[1:2])/sum(pc$sdev^2),1)
```

```
## [1] 51.3
```

Furthermore, visualizing the eigenvalues by showing the percentage of variances vs each principal components.:

```
fviz_eig(pc)
```



c) Report the loading coefficients (eigenvector of the correlation matrix) of the first two principal components.

```
pc$loading[,1:2]
```

```
##              Comp.1        Comp.2
## REACT    0.05074485  0.1610889135
## HEIGHT   0.30938720 -0.1948652676
## WEIGHT   0.41586867 -0.0326183906
## SHLDR    0.29939671 -0.2082727762
## PELVIC   0.29314755 -0.2013196634
## CHEST    0.36054275 -0.0005222451
## THIGH    0.28379796  0.3034561646
## PULSE   -0.11804080  0.3828513742
## DIAST   -0.03411639  0.2639467180
## CHNUP   -0.29172354 -0.2346251379
## BREATH   0.25261897 -0.0262139810
## RECVR   -0.02075408  0.4537337307
## SPEED   -0.03228068 -0.4821396135
## ENDUR   -0.20471081 -0.0352845349
## FAT      0.36777130  0.2178618386
```

d) Describe what information we can extract from the first two principal components? Explain. (You need to interpret the loading of the first two PCs)

The first pc explains the size of applicants and the second one the athletic performance.

Problem 2

Use crime data :

```
##            MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY  AUTO
## ALABAMA      14.2 25.2    96.8   278.3   1135.5  1881.9 280.7
## ALASKA       10.8 51.6    96.8   284.0   1331.7  3369.8 753.3
## ARIZONA       9.5 34.2   138.2   312.3   2346.1  4467.4 439.5
## ARKANSAS      8.8 27.6    83.2   203.4    972.6  1862.1 183.4
## CALIFORNIA   11.5 49.4   287.0   358.0   2139.4  3499.8 663.5
## COLORADO      6.3 42.0   170.7   292.9   1935.2  3903.2 477.1
```

a) Perform the principal components using the correlation matrix. You do not need to do data cleaning or fixing the direction of variables in this data.

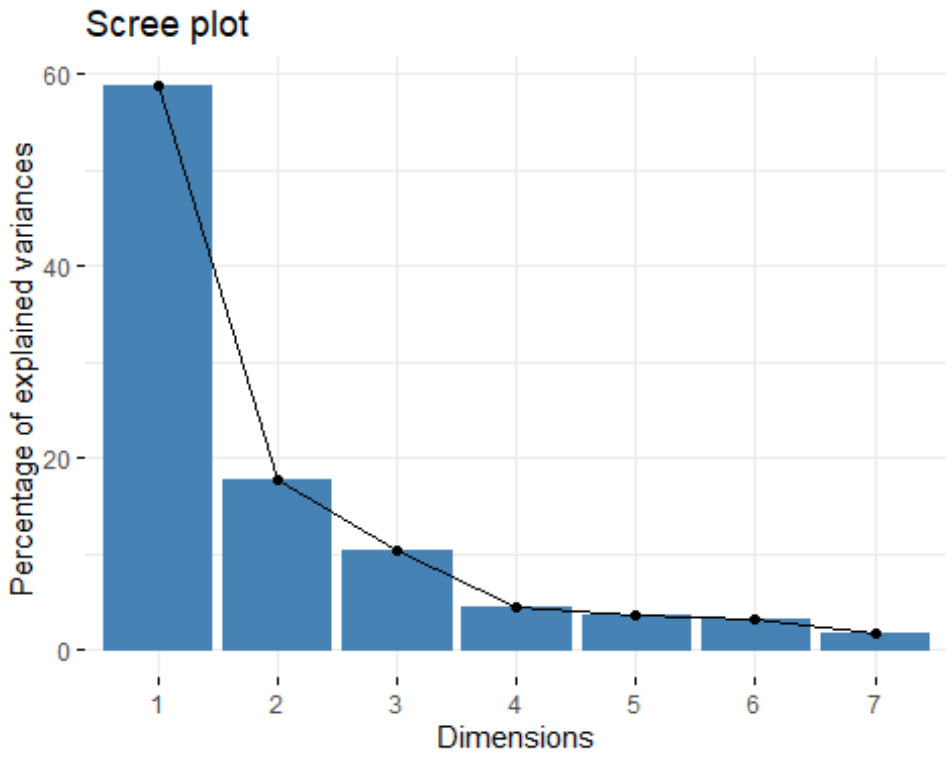
```
pc1 <- princomp(crime, cor = TRUE)
```

b) What percentage of the total variance is covered by the first two principal components?

```
# 76.5 Percentage of the total variance is covered by the first two principal components.
summary(pc1, cor = T)

## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.0285363 1.1129788 0.8519487 0.56252293 0.50791186
## Proportion of Variance 0.5878514 0.1769603 0.1036881 0.04520458 0.03685349
## Cumulative Proportion  0.5878514 0.7648116 0.8684997 0.91370429 0.95055778
##                            Comp.6     Comp.7
## Standard deviation     0.47121064 0.35221592
## Proportion of Variance 0.03171992 0.01772229
## Cumulative Proportion  0.98227771 1.00000000

round(100*sum((pc1$sdev^2)[1:2])/sum(pc1$sdev^2),1)

## [1] 76.5
```

Furthermore, visualizing the eigenvalues by showing the percentage of variances vs each principal components.:

```
fviz_eig(pc1)
```



c) Report the loading coefficients (eigenvector of the correlation matrix) of the first two principal components.

```
pc1$loading[,1:2]
```

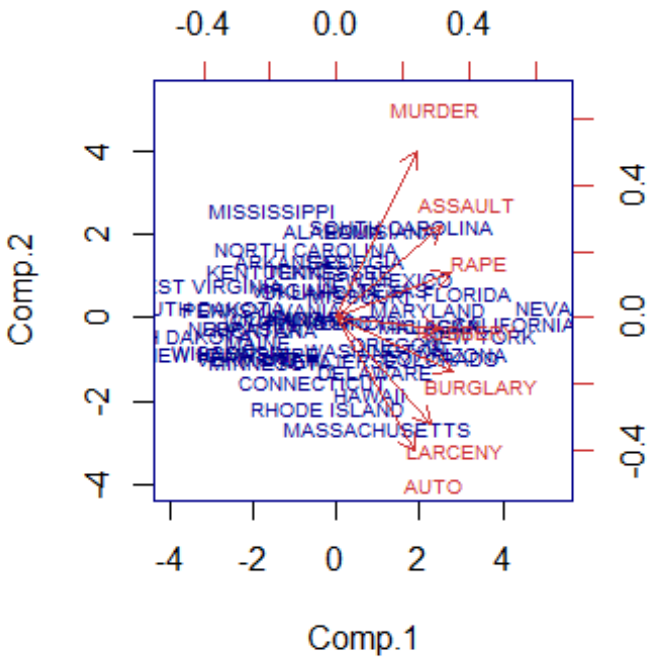
```
##             Comp.1      Comp.2
## MURDER   0.3002792  0.62917444
## RAPE     0.4317594  0.16943512
## ROBBERY  0.3968755 -0.04224698
## ASSAULT  0.3966517  0.34352815
## BURGLARY 0.4401572 -0.20334059
## LARCENY  0.3573595 -0.40231912
## AUTO     0.2951768 -0.50242093
```

d) Describe what information we can extract from the first two principal components? Explain. (You need to interpret the loading of the first two PCs)

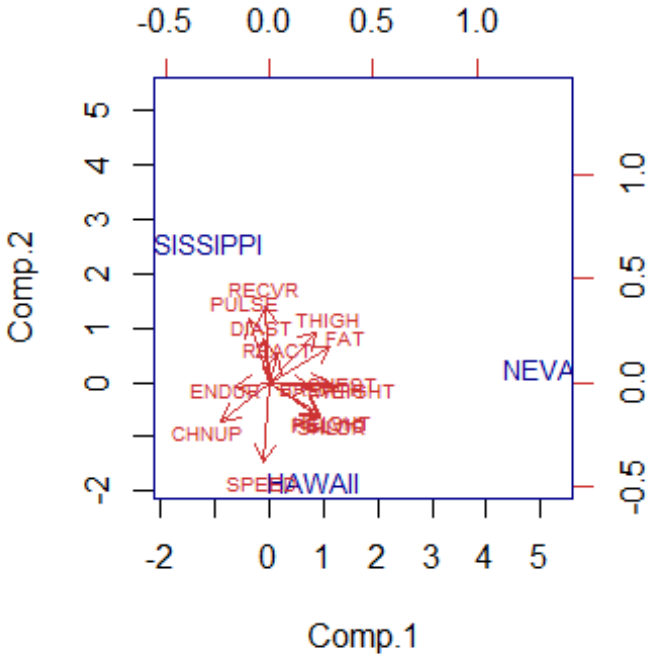
pc1 explains the overall crime, pc2 explains the specialty in violent vs non-violent crime.

e) Construct the biplot graph of the crime data. Interpret the resulting biplot graph for "MISSISSIPPI," "NEVADA," and "HAWAII." (You can validate your conclusions by looking at the actual standardized (scaled) data values.)

```
biplot(x = pc1, scale = 0, cex = 0.6, col = c("blue4", "brown3"))
```



Let's limit to the states of interest:



We can see that the Mississippi is low on all principles while Nevada is High on Larceny and Burglary while Hawaii is high on Larceny, but low on Burglary.

Problem 3

Perform canonical correlation analysis on the Husband/Wife love data set (Answered by husband and wife; a total of eight responses). In this data, columns are the rating for the following questions: Q1. What is the level of passionate love you feel for your partner? Q2. What is the level of passionate love your partner feels for you? Q3. What is the level of companionate love you feel for your partner? Q4. What is the level of companionate love your partner feels for you?

```
love <- read.csv("https://bit.ly/3onLanp", header = T)
options(digits = 3)
# Creating two sets for correlations
X <- love[, 1:4]  # Husband's responses
Y <- love[, 5:8]  # Wife's responses
#install.packages("CCA")

X; Y

library(CCA)

cca <- cc(X, Y)

cca$xcoef

##      [,1]   [,2]   [,3]   [,4]
## h1  0.256 -1.060 -1.387  0.102
## h2 -0.864  0.568  0.784  1.185
## h3 -2.775 -0.929  0.608 -3.057
## h4  2.679 -0.436  0.132  3.413

cca$ycoef

##      [,1]   [,2]   [,3]   [,4]
## w1 -0.849  0.445 -1.054 -0.243
## w2 -0.700 -0.628 -0.122  1.320
## w3 -1.744  1.714  2.506  0.582
## w4  0.821 -2.381 -1.353 -2.103
```

a) Find the linear combination of the four husband responses and the linear combination of the four wife responses, maximizing the two derived variables' correlation. (Hint: use X coefficients and write U_1 as a linear combination of X variables, then use Y coefficients and write V_1 as a linear combination of Y variables)

$$u_1 = 0.26h_1 - 0.86h_2 - 2.78h_3 + 2.68h_4$$

$$v_1 = -0.85w_1 - 0.70w_2 - 1.74w_3 + 0.821w_4$$

b) Find the correlation between U_1, V_1.

```
# The correlation between U1 and V1 is 0.572
cca$cor[1]

## [1] 0.572
```

c) What does the husband linear combination (U_1) measure? To answer, based on the coefficient of X in U_1, ask yourself, "What does the husband linear combination measure?"

The linear combination U_1 measures mostly companionate love your partner feels for you. The rest of the coefficients are negatively measured.

d) Repeat C. for the wives linear combination (V_1).

The linear combination V_1 measures negative passionate feelings with negative companion love for partner with a positive feel for companion from lover. Mostly companionate love your partner feels for you is measured.