A5P1

Dawu Liu

In this assignment, principal component is written as **PC** sometimes for short.

Also, here are the keys for temperature data that will be analyzed below:

x1 = maximum daily air temperature

x2 = minimum daily air temperature

x3 = integrated area under daily air temperature curve

x4 = maximimum daily soil temperature

x5 = mininum daily soil temperature

x6 = integrated area under soil temperature

x7 = maximum daily relative humidity

x8 = minimum daily relative humidity

x9 = integrated area under daily humidity

x10 = total wind, miles per day

x11 = evaporation

(a)

Sample covariance matrix **S**

```
##            x1       x2        x3       x4       x5        x6       x7        x8
## x1    55.6812  16.4899  117.5536  26.9768  10.3836   97.7295  -1.5585  -42.3565
## x2    16.4899  10.8638   61.6725  14.0251   8.2271   56.4783  -0.6580  -10.7971
## x3   117.5536  61.6725  402.6995  92.9498  43.4570  365.2657  -3.8618 -106.8947
## x4    26.9768  14.0251   92.9498  25.6638  10.5952   92.8106  -0.6309  -26.6696
## x5    10.3836   8.2271   43.4570  10.5952  13.4401   59.9251  -0.5367    8.4618
## x6    97.7295  56.4783  365.2657  92.8106  59.9251  438.2536  -1.0237  -59.9449
## x7    -1.5585  -0.6580   -3.8618  -0.6309  -0.5367   -1.0237   1.4517    1.9024
## x8   -42.3565 -10.7971 -106.8947 -26.6696   8.4618  -59.9449   1.9024  106.1971
## x9  -128.2725 -44.4048 -387.7237 -94.3285  -4.3623 -313.4696   9.9700  271.6493
## x10 -209.0957  14.9314 -294.5517 -67.9314 224.5773  386.8188 -26.3831  597.6686
## x11   61.3671  25.9874  201.8696  53.6126  17.5802  217.0614  -3.2401  -96.8541
##             x9        x10       x11
## x1  -128.2725  -209.0957   61.3671
## x2   -44.4048    14.9314   25.9874
## x3  -387.7237  -294.5517  201.8696
## x4   -94.3285   -67.9314   53.6126
## x5    -4.3623   224.5773   17.5802
## x6  -313.4696   386.8188  217.0614
## x7     9.9700   -26.3831   -3.2401
```

```
## x8   271.6493   597.6686  -96.8541
## x9   885.6290   970.9671 -355.3913
## x10  970.9671 22227.1580   94.5879
## x11 -355.3913    94.5879  214.0604
```

Sample correlation matrix **R**

```
##          x1      x2      x3      x4      x5      x6      x7      x8      x9
## x1   1.0000  0.6705  0.7850  0.7136  0.3796  0.6256 -0.1733 -0.5508 -0.5776
## x2   0.6705  1.0000  0.9324  0.8400  0.6809  0.8185 -0.1657 -0.3179 -0.4527
## x3   0.7850  0.9324  1.0000  0.9143  0.5907  0.8695 -0.1597 -0.5169 -0.6492
## x4   0.7136  0.8400  0.9143  1.0000  0.5705  0.8751 -0.1034 -0.5109 -0.6257
## x5   0.3796  0.6809  0.5907  0.5705  1.0000  0.7808 -0.1215  0.2240 -0.0400
## x6   0.6256  0.8185  0.8695  0.8751  0.7808  1.0000 -0.0406 -0.2779 -0.5032
## x7  -0.1733 -0.1657 -0.1597 -0.1034 -0.1215 -0.0406  1.0000  0.1532  0.2781
## x8  -0.5508 -0.3179 -0.5169 -0.5109  0.2240 -0.2779  0.1532  1.0000  0.8858
## x9  -0.5776 -0.4527 -0.6492 -0.6257 -0.0400 -0.5032  0.2781  0.8858  1.0000
## x10 -0.1880  0.0304 -0.0985 -0.0899  0.4109  0.1239 -0.1469  0.3890  0.2188
## x11  0.5621  0.5389  0.6876  0.7233  0.3278  0.7087 -0.1838 -0.6424 -0.8162
##         x10     x11
## x1  -0.1880  0.5621
## x2   0.0304  0.5389
## x3  -0.0985  0.6876
## x4  -0.0899  0.7233
## x5   0.4109  0.3278
## x6   0.1239  0.7087
## x7  -0.1469 -0.1838
## x8   0.3890 -0.6424
## x9   0.2188 -0.8162
## x10  1.0000  0.0434
## x11  0.0434  1.0000
```

(b) and (c)

PCA using covariance matrix **S**

i. The eigenvalues are:

```
##  [1] 22303.4976  1590.6789   358.0457    63.3665    29.3270    17.1149
##  [7]    12.7478     2.8330     1.9069     0.8769     0.7028
```

The first eigenvalue is massively larger compared to the rest, and it accounts for the most of the total variance. (table shown in ii)

ii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  22303.4976          91.4786                      91.4786
## Dim.2   1590.6789           6.5242                      98.0029
## Dim.3    358.0457           1.4685                      99.4714
## Dim.4     63.3665           0.2599                      99.7313
```

```
## Dim.5     29.3270          0.1203                99.8516
## Dim.6     17.1149          0.0702                99.9218
## Dim.7     12.7478          0.0523                99.9741
## Dim.8      2.8330          0.0116                99.9857
## Dim.9      1.9069          0.0078                99.9935
## Dim.10     0.8769          0.0036                99.9971
## Dim.11     0.7028          0.0029               100.0000
```

This method suggests to keep **1** principal component which gives 91.5% of the total variance.

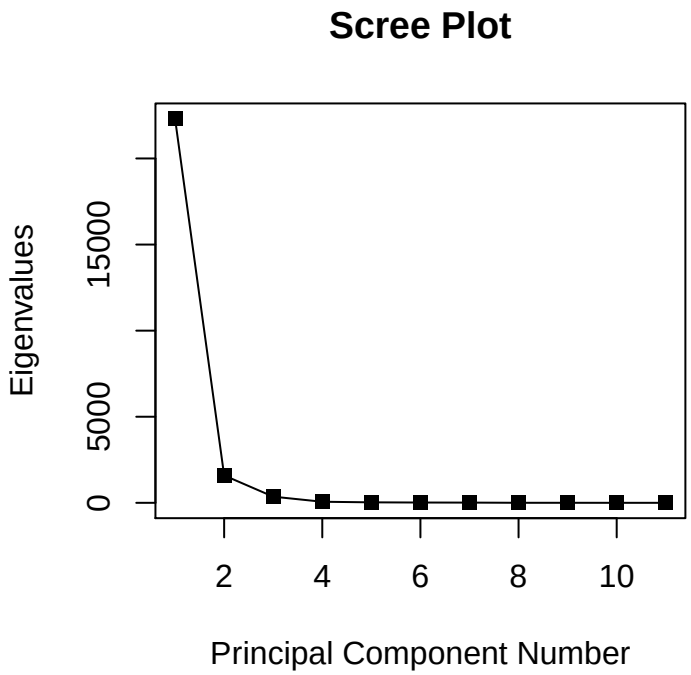
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 2216.463
```

This method also suggests to keep **1** principal component.

Criteria3, scree plot



The "bend" occurs at PC2, indicating from PC2 and on, the the eigenvalues are relatively small. This method also suggests to keep **1** principal component.

Overall, **1** principal component is retained.

But in order to make a scatter plot, we will use 2 principal components.

iii.

The eigenvectors for the principal components:

$(\tilde{\hat{e}_1})^T =$

```
##  [1]  0.0097 -0.0006  0.0141  0.0033 -0.0101 -0.0167  0.0012 -0.0275 -0.0456
## [10] -0.9982 -0.0034
```

$(\tilde{\hat{e}_2})^T =$

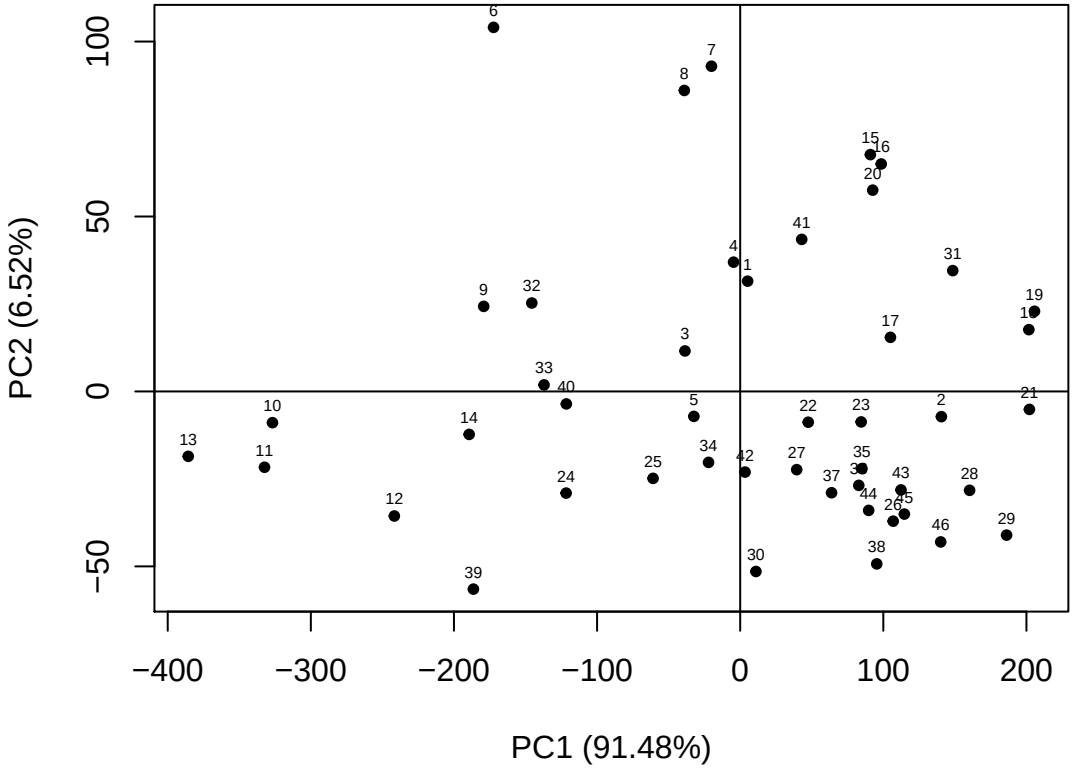
```
##  [1] -0.1331 -0.0608 -0.4397 -0.1078 -0.0398 -0.4290  0.0072  0.1844  0.6657
## [10] -0.0346 -0.3311
```

iv.

Almost all of PC1 depends on x_{10}, representing total wind (miles per day). For PC2, we see x_3, x_6, x_9, and x_{11} have significantly large magnitudes than the rest, means the majority of PC2 depends on integrated areas under the temperature/humidity curves and the evaporation. Also, PC2 shows the contrast between integrated area under daily humidity on one hand, and integrated area under daily air temperature, soil temperature, and evaporation on the other hand.

v.

Scatter plot for PC2 vs PC1



The density of the plot becomes higher as the PC1 value increases and PC2 value decreases. About half of the data points are grouped toward the right bottom portion of the plot.

PCA using sample correlation matrix **R**

i. The eigenvalues are:

```
##  [1] 6.0202 2.1193 1.1303 0.7600 0.3554 0.2593 0.1221 0.1105 0.0598 0.0422
## [11] 0.0209
```

The first two eigenvalues are relatively larger compared to the rest, and they account for the majority of the total variance. (table shown in ii)

ii.

Criteria 1, eigenvalues and their cumulative proportions table

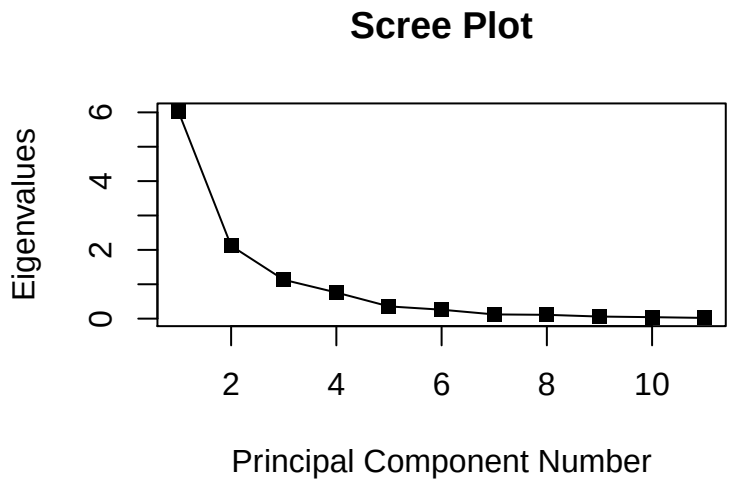
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1      6.0202          54.7295                       54.7295
## Dim.2      2.1193          19.2667                       73.9962
## Dim.3      1.1303          10.2754                       84.2716
## Dim.4      0.7600           6.9092                       91.1808
## Dim.5      0.3554           3.2305                       94.4113
## Dim.6      0.2593           2.3577                       96.7690
## Dim.7      0.1221           1.1098                       97.8788
## Dim.8      0.1105           1.0044                       98.8832
## Dim.9      0.0598           0.5437                       99.4269
## Dim.10     0.0422           0.3835                       99.8104
## Dim.11     0.0209           0.1896                      100.0000
```

This method suggests to keep **3** principal component which gives 84.3% of the total variance.

Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

Since three eigenvalues are larger than the mean of 1, this method suggests to keep **3** principal component.

Criteria3, scree plot



The "bend" occurs at PC2, indicating from PC2 and on, the the eigenvalues are relatively small. This method suggests to keep **1** principal component.

Overall, **3** principal components are retained.

iii.

The eigenvectors for the principal components:

$(\tilde{\hat{e}_1})^T =$

```
##  [1]  0.3304  0.3542  0.3923  0.3820  0.2323  0.3621 -0.0884 -0.2501 -0.3111
## [10] -0.0243  0.3357
```

$(\tilde{\hat{e}_2})^T =$

```
##  [1]  0.0787 -0.1928 -0.0518 -0.0474 -0.5303 -0.2361 -0.0213 -0.5023 -0.3595
## [10] -0.4685  0.1153
```

$(\tilde{\hat{e}_3})^T =$

```
##  [1]  0.0880  0.1071  0.1105  0.1334  0.0154  0.1198  0.7946  0.0826  0.2136
## [10] -0.4669 -0.1853
```

iv.

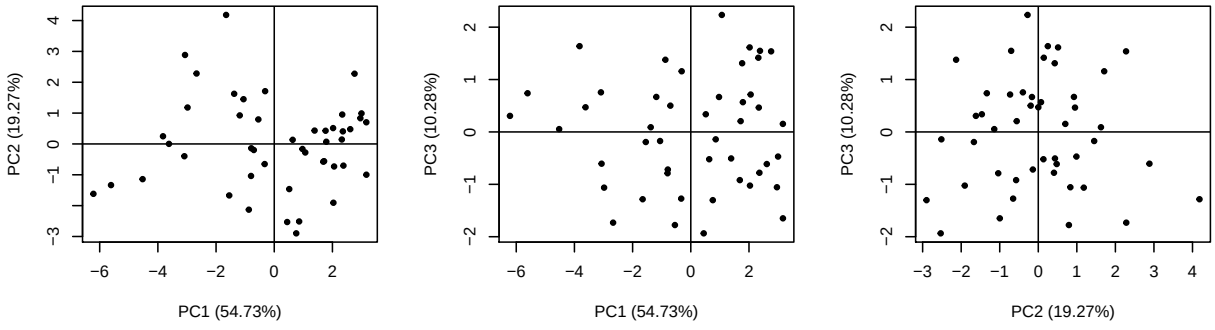
PC1 depends on all the variables other than total wind(miles per day) and maximum daily relative humidity almost evenly. Coefficients on humidity have opposite signs compared to temperature, showing contrast relationship.

The majority of PC2 depends on x_5, x_8, and x_{10} have relatively large magnitudes than the rest, means PC2 primarily interprets the minimum daily soil temperature, minimum daily relative humidity, and total wind(miles per day).

The majority of PC3 depends on maximum daily relative humidity, followed by a significant amount of total wind(miles per day) but with opposite sign, indicating there is contrast between those two.

v.

Scatter plot for the principal components



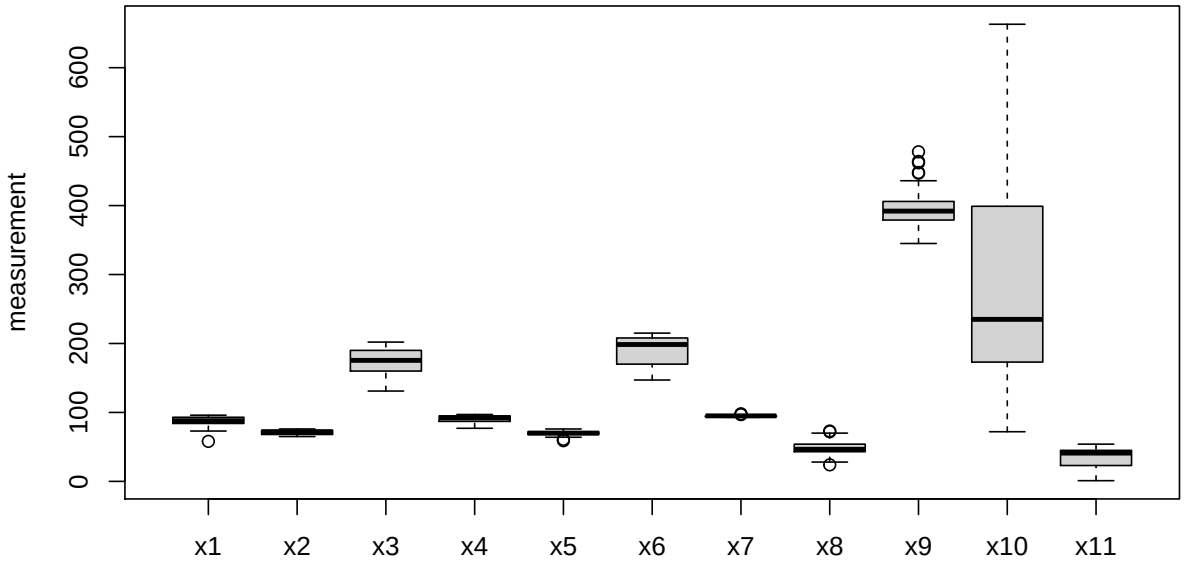
PC2 VS PC1 plot: Data points are more gathered toward to the right side of the plot, data points becomes more scattered as they decrease in PC1 values.

PC3 VS PC1 plot: Data points are nearly evenly distributed across the plot, with some sparse points on the left, it's hard to get any relationships here.

PC2 VS PC3: It appears there are two small groups on the plots, one is slight above the intersection of the axes and one is slightly below the origin.

(d)

The boxplot of the data:



Correlation matrix **R** is better here. In the boxplot of the data above, the range for the values of x_{10} are significantly larger than the rest of the variables, causing a scaling issue. In the **S** PCA, all three criteria suggest to keep only one principal component, and that principal component only really tells us about the total wind(miles per day). Therefore, we can not obtain information about the relationships between the other variables from it. But in the **R** PCA, we are able to obtain those information.

(e)

Interpretation of the principal components:

For **S**:

PC1 is a component of total wind(miles per day), higher PC1 values means less wind(negative sign).
PC2 shows the contrast between integrated area under daily humidity on one hand, and integrated area under daily air temperature/soil temperature, and evaporation on the other hand.

For **R**:

PC1 represents the contrast between humidity and temperature.
PC2 mainly represents the minimum temperature/humidity values and the total wind (negative signs), higher PC2 values indicates lower minimum temperature/humidity, and less wind. Showing dependencies between those variables.
PC3 shows the contrast between maximum daily relative humidity and total wind.

Code used to solve the questions(graphs are hidden):

```
rm(list = ls())
library(readxl)
library(factoextra)
X <- read_excel("C:/Users/John/Desktop/STAT 445/Data/temperaturedata-clean.xlsx")
S <- cov(X)
round(S,4)
```

```
##            x1       x2        x3       x4       x5        x6       x7        x8
## x1    55.6812  16.4899  117.5536  26.9768  10.3836   97.7295  -1.5585  -42.3565
## x2    16.4899  10.8638   61.6725  14.0251   8.2271   56.4783  -0.6580  -10.7971
## x3   117.5536  61.6725  402.6995  92.9498  43.4570  365.2657  -3.8618 -106.8947
## x4    26.9768  14.0251   92.9498  25.6638  10.5952   92.8106  -0.6309  -26.6696
## x5    10.3836   8.2271   43.4570  10.5952  13.4401   59.9251  -0.5367    8.4618
## x6    97.7295  56.4783  365.2657  92.8106  59.9251  438.2536  -1.0237  -59.9449
## x7    -1.5585  -0.6580   -3.8618  -0.6309  -0.5367   -1.0237   1.4517    1.9024
## x8   -42.3565 -10.7971 -106.8947 -26.6696   8.4618  -59.9449   1.9024  106.1971
## x9  -128.2725 -44.4048 -387.7237 -94.3285  -4.3623 -313.4696   9.9700  271.6493
## x10 -209.0957  14.9314 -294.5517 -67.9314 224.5773  386.8188 -26.3831  597.6686
## x11   61.3671  25.9874  201.8696  53.6126  17.5802  217.0614  -3.2401  -96.8541
##             x9        x10       x11
## x1  -128.2725  -209.0957   61.3671
## x2   -44.4048    14.9314   25.9874
## x3  -387.7237  -294.5517  201.8696
## x4   -94.3285   -67.9314   53.6126
## x5    -4.3623   224.5773   17.5802
## x6  -313.4696   386.8188  217.0614
## x7     9.9700   -26.3831   -3.2401
## x8   271.6493   597.6686  -96.8541
## x9   885.6290   970.9671 -355.3913
## x10  970.9671 22227.1580   94.5879
## x11 -355.3913    94.5879  214.0604
```

```
R <- cor(X)
round(R,4)
```

```
##           x1      x2      x3      x4      x5      x6      x7      x8      x9
## x1   1.0000  0.6705  0.7850  0.7136  0.3796  0.6256 -0.1733 -0.5508 -0.5776
## x2   0.6705  1.0000  0.9324  0.8400  0.6809  0.8185 -0.1657 -0.3179 -0.4527
## x3   0.7850  0.9324  1.0000  0.9143  0.5907  0.8695 -0.1597 -0.5169 -0.6492
## x4   0.7136  0.8400  0.9143  1.0000  0.5705  0.8751 -0.1034 -0.5109 -0.6257
## x5   0.3796  0.6809  0.5907  0.5705  1.0000  0.7808 -0.1215  0.2240 -0.0400
## x6   0.6256  0.8185  0.8695  0.8751  0.7808  1.0000 -0.0406 -0.2779 -0.5032
## x7  -0.1733 -0.1657 -0.1597 -0.1034 -0.1215 -0.0406  1.0000  0.1532  0.2781
## x8  -0.5508 -0.3179 -0.5169 -0.5109  0.2240 -0.2779  0.1532  1.0000  0.8858
## x9  -0.5776 -0.4527 -0.6492 -0.6257 -0.0400 -0.5032  0.2781  0.8858  1.0000
## x10 -0.1880  0.0304 -0.0985 -0.0899  0.4109  0.1239 -0.1469  0.3890  0.2188
## x11  0.5621  0.5389  0.6876  0.7233  0.3278  0.7087 -0.1838 -0.6424 -0.8162
##         x10     x11
## x1  -0.1880  0.5621
## x2   0.0304  0.5389
## x3  -0.0985  0.6876
```

```
## x4  -0.0899  0.7233
## x5   0.4109  0.3278
## x6   0.1239  0.7087
## x7  -0.1469 -0.1838
## x8   0.3890 -0.6424
## x9   0.2188 -0.8162
## x10  1.0000  0.0434
## x11  0.0434  1.0000
```

```
# S analysis
S_pr = prcomp(X, center = TRUE, scale. = FALSE)
S_eigen_table = get_eigenvalue(S_pr)
round(S_eigen_table[,1], 4)
```

```
##  [1] 22303.4976  1590.6789   358.0457    63.3665    29.3270    17.1149
##  [7]    12.7478     2.8330     1.9069     0.8769     0.7028
```

```
round(S_eigen_table,4)
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  22303.4976          91.4786                       91.4786
## Dim.2   1590.6789           6.5242                       98.0029
## Dim.3    358.0457           1.4685                       99.4714
## Dim.4     63.3665           0.2599                       99.7313
## Dim.5     29.3270           0.1203                       99.8516
## Dim.6     17.1149           0.0702                       99.9218
## Dim.7     12.7478           0.0523                       99.9741
## Dim.8      2.8330           0.0116                       99.9857
## Dim.9      1.9069           0.0078                       99.9935
## Dim.10     0.8769           0.0036                       99.9971
## Dim.11     0.7028           0.0029                      100.0000
```

```
mean(S_eigen_table[,1])
```

```
## [1] 2216.463
```

```
plot(S_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
```

```
e_S1 <- as.vector(S_pr$rotation[,1])
e_S2 <- as.vector(S_pr$rotation[,2])
round(e_S1, 4)
```

```
##  [1]  0.0097 -0.0006  0.0141  0.0033 -0.0101 -0.0167  0.0012 -0.0275 -0.0456
## [10] -0.9982 -0.0034
```

```
round(e_S2, 4)
```

```
##  [1] -0.1331 -0.0608 -0.4397 -0.1078 -0.0398 -0.4290  0.0072  0.1844  0.6657
## [10] -0.0346 -0.3311
```

```r
plot(S_pr$x[,1], S_pr$x[,2], xlab = "PC1 (91.48%)", ylab = "PC2 (6.52%)", pch = 20)
text(S_pr$x[,1], S_pr$x[,2], pos = 3, offset = 0.3, cex = 0.5)
abline(v = 0, h = 0)
```

```r
# R analysis
R_pr = prcomp(X, center = TRUE, scale. = TRUE)
R_eigen_table = get_eigenvalue(R_pr)
round(R_eigen_table[,1], 4)
```

```
##  [1] 6.0202 2.1193 1.1303 0.7600 0.3554 0.2593 0.1221 0.1105 0.0598 0.0422
## [11] 0.0209
```

```r
round(R_eigen_table,4)
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1      6.0202          54.7295                       54.7295
## Dim.2      2.1193          19.2667                       73.9962
## Dim.3      1.1303          10.2754                       84.2716
## Dim.4      0.7600           6.9092                       91.1808
## Dim.5      0.3554           3.2305                       94.4113
## Dim.6      0.2593           2.3577                       96.7690
## Dim.7      0.1221           1.1098                       97.8788
## Dim.8      0.1105           1.0044                       98.8832
## Dim.9      0.0598           0.5437                       99.4269
## Dim.10     0.0422           0.3835                       99.8104
## Dim.11     0.0209           0.1896                      100.0000
```

```r
plot(R_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
```

```r
e_R1 <- as.vector(R_pr$rotation[,1])
e_R2 <- as.vector(R_pr$rotation[,2])
e_R3 <- as.vector(R_pr$rotation[,3])
round(e_R1, 4)
```

```
##  [1]  0.3304  0.3542  0.3923  0.3820  0.2323  0.3621 -0.0884 -0.2501 -0.3111
## [10] -0.0243  0.3357
```

```r
round(e_R2, 4)
```

```
##  [1]  0.0787 -0.1928 -0.0518 -0.0474 -0.5303 -0.2361 -0.0213 -0.5023 -0.3595
## [10] -0.4685  0.1153
```

```r
round(e_R3, 4)
```

```
##  [1]  0.0880  0.1071  0.1105  0.1334  0.0154  0.1198  0.7946  0.0826  0.2136
## [10] -0.4669 -0.1853
```

```r
par(mfrow=c(1, 3))
plot(R_pr$x[,1], R_pr$x[,2], xlab = "PC1 (54.73%)", ylab = "PC2 (19.27%)", pch = 20)
abline(v = 0, h = 0)

plot(R_pr$x[,1], R_pr$x[,3], xlab = "PC1 (54.73%)", ylab = "PC3 (10.28%)", pch = 20)
abline(v = 0, h = 0)

plot(R_pr$x[,2], R_pr$x[,3], xlab = "PC2 (19.27%)", ylab = "PC3 (10.28%)", pch = 20)
abline(v = 0, h = 0)
```