PCA

Isabella Chittumuri

10/20/2020

12.8 Carry out a principal component analysis on all six variables of the glucose data of Table 3.9. Use both S and R. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R?

```
getwd()
```

```
## [1] "/Users/isabellachittumuri/Documents/Hunter College/Fall 2020/Stat 717/HW"
```

```
glucose <- read.table("T3_9_GLUCOSE.dat")
```

```
# Covariance of glucose
glucose_S <- var(glucose)

# PCA - covariance, glucose
glucose_pca_S <- princomp(covmat = glucose_S)
summary(glucose_pca_S, loadings = TRUE)
```

```
## Importance of components:
##                             Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     33.9408200 19.8525151 17.6291195 9.88775099 8.29674381
## Proportion of Variance  0.5570193  0.1905709  0.1502750 0.04727381 0.03328441
## Cumulative Proportion   0.5570193  0.7475902  0.8978652 0.94513900 0.97842342
##                             Comp.6
## Standard deviation     6.68003216
## Proportion of Variance 0.02157658
## Cumulative Proportion  1.00000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## V1                       0.805  0.461  0.346
## V2               -0.202  0.518 -0.295 -0.776
## V3         0.122         0.247 -0.830  0.478
## V4  0.758 -0.446  0.469
## V5  0.493        -0.844                0.154
## V6  0.412  0.878  0.147 -0.122        -0.141
```

We use S, the covariance matrix, to understand how the variables of the data set are varying from the mean with respect to each other, and to see if there is any relationship between them.

The Principal Component Analysis (PCA) is a method used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

After computing the PCA of the covariance matrix, we get six principal components. Principal components (PC) are new variables that are constructed as linear combinations of the initial variable.

We want to retain just enough components to explain a large percentage of the total variation of the original variables. The values between 70% and 90% are sufficient enough. In this case, the first and second PC account for 75% of the total variance of the observed variables. If we add the third PC, this percentage up will increase to 90%. This is understandable since the first three PCs encompass most of proportion of variance of the data.

```
# Eigen values for covariance
glucose_S_evals <- eigen(glucose_S)$values; glucose_S_evals
```

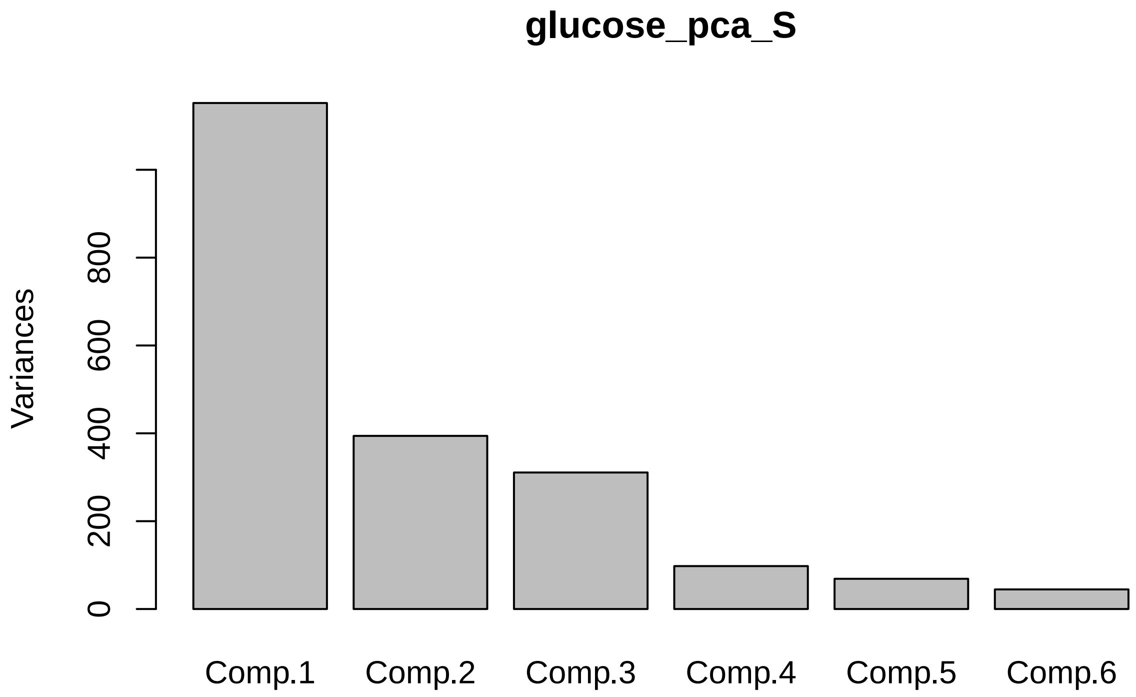
```
## [1] 1151.97926  394.12236  310.78585   97.76762   68.83596   44.62283
```

```
# Mean of eigen values
mean(glucose_S_evals)
```

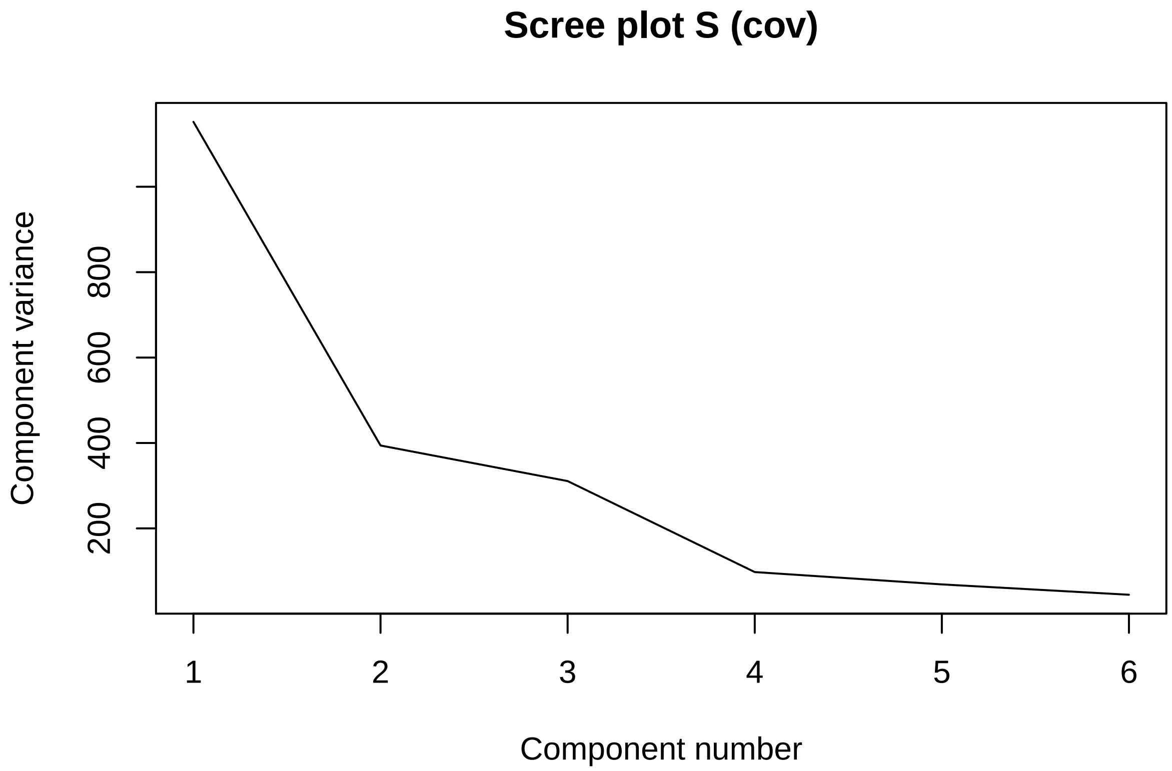
```
## [1] 344.6856
```

We can excluded the PCs whose eigenvalues are less than the average eigenvalue. In this case, we can exclude the third, fourth. , fifth and sixth PCs. This also means that only the first and second PCs are greater than the average eigenvalue.

```
# Scree diagram, covariance
plot(glucose_pca_S)
```



```
# Scree plot, covariance
plot(glucose_pca_S$sdev^2, xlab = "Component number",
     ylab = "Component variance", type = "l", main = "Scree plot S (cov)")
```



A scree plot shows how much variation each PC captures from the data. This scree plot is has a steep curve that bends quickly and flattens out. This suggests that the first two or three PCs are sufficient to describe the essence of the data, Based on our entire analysis, we should retain the first two PCs.

```
# Correlation of glucose
glucose_R <- cor(glucose)

# PCA - correlation, glucose
glucose_pca_R <- princomp(covmat = glucose_R)
summary(glucose_pca_R, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.4753986 1.0390904 0.9888386 0.9304411 0.74275606
## Proportion of Variance 0.3628002 0.1799515 0.1629670 0.1442868 0.09194776
## Cumulative Proportion  0.3628002 0.5427516 0.7057186 0.8500054 0.94195312
##                            Comp.6
## Standard deviation     0.59015361
## Proportion of Variance 0.05804688
## Cumulative Proportion  1.00000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## V1  0.336  0.176  0.497  0.713  0.141  0.285
## V2  0.258  0.843               -0.115 -0.444
## V3  0.370         0.466 -0.671  0.387  0.209
## V4  0.475 -0.329 -0.358  0.173  0.559 -0.443
## V5  0.486        -0.567        -0.223  0.618
## V6  0.471 -0.376  0.278        -0.675 -0.316
```

We use R, the correlation matrix, to understand is a method of the strength and direction of the relationships between the variables. A correlation is a function of the covariance, but a correlation matrix standardizes each of the variables, with a mean of 0 and a standard deviation of 1.

After computing the PCA of the correlation matrix, we get six principal components. In this case, the first three PCs account for 70% of the total variance of the observed variables. If we add the fourth PC, this percentage up will increase to 85%.

```
# Eigen values
glucose_R_evals <- eigen(glucose_R)$values; glucose_R_evals
```

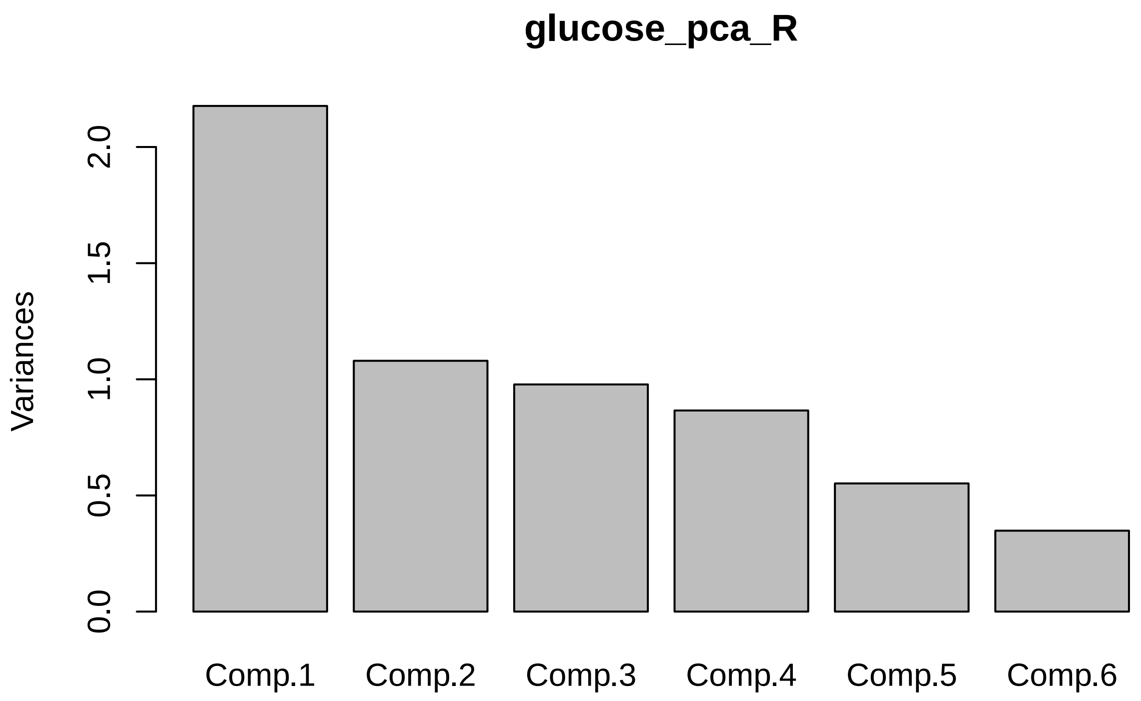
```
## [1] 2.1768009 1.0797089 0.9778017 0.8657206 0.5516866 0.3482813
```

```
# Mean of eigen values
mean(glucose_R_evals)
```

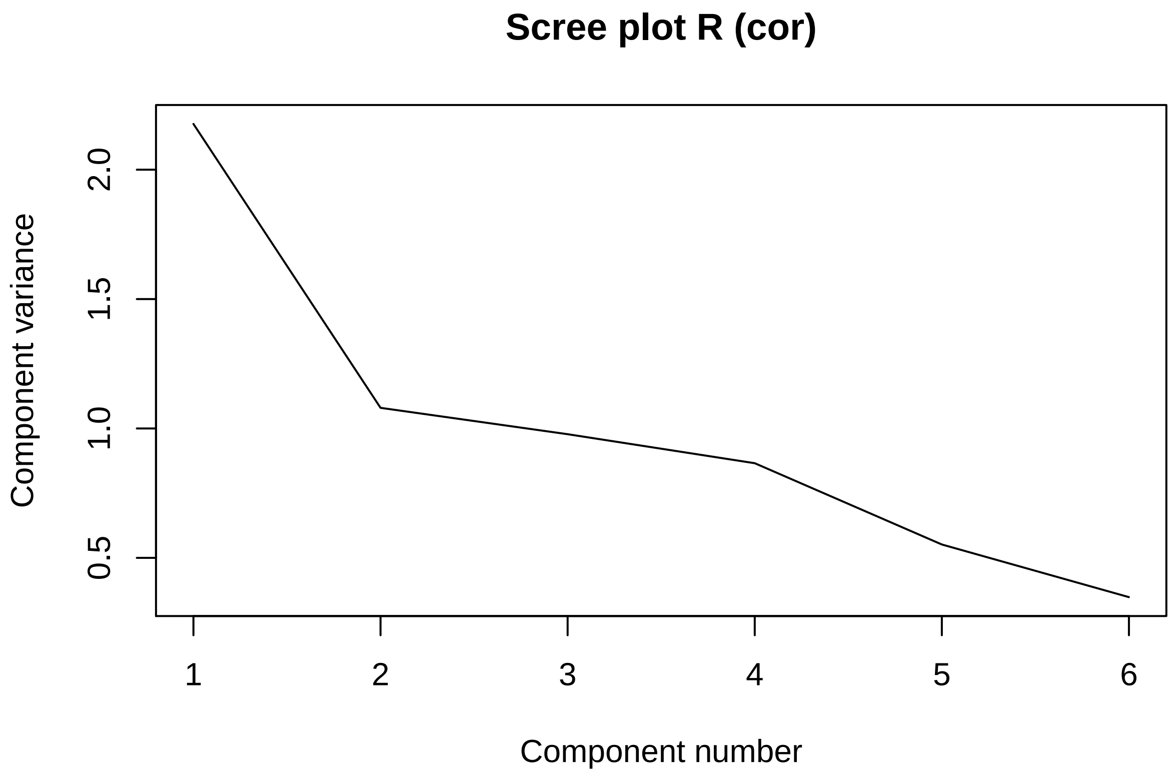
```
## [1] 1
```

Only the first three PCs are greater than the average eigenvalue, meaning that these are the PCs to include.

```
# Scree diagram, correlation
plot(glucose_pca_R)
```



```
# Scree plot, correlation
plot(glucose_pca_R$sdev^2, xlab = "Component number",
     ylab = "Component variance", type = "l", main = "Scree plot R (cor)")
```



This scree plot is doesn't have really steep curve. The line actually bends and flattens out slowly, suggesting that we would need more PCs than normal. Based on our entire analysis, we should retain the first three PCs.

In comparison, the first three components of the covariance matrix account for 90% of the variance, which that of the correlation matrix only accounts for 70% of the variance. Because a larger precent of variance is explained, it is more appropriate to use the covariance matrix S.

12.15 Carry out a principal component analysis on the temperature data of Table 7.2. Use both S and R. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R?

```
temp <- read.table("T7_2_TEMPERAT.DAT")
```

```
# Covariance of temp
temp_S <- var(temp)
```

```r
# PCA - covariance, temp
temp_pca_S <- princomp(covmat = temp_S)
summary(temp_pca_S, loadings = TRUE)
```

```
## Importance of components:
##                             Comp.1     Comp.2      Comp.3      Comp.4
## Standard deviation     149.3435555 39.8833165 18.92209647 7.960309287
## Proportion of Variance   0.9147864  0.0652423  0.01468538 0.002599002
## Cumulative Proportion    0.9147864  0.9800287  0.99471411 0.997313111
##                             Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     5.415436298 4.1370112451 3.5704058846 1.6831510247
## Proportion of Variance 0.001202856 0.0007019726 0.0005228558 0.0001161965
## Cumulative Proportion  0.998515967 0.9992179400 0.9997407958 0.9998569923
##                              Comp.9      Comp.10      Comp.11
## Standard deviation     1.380914e+00 9.364462e-01 8.383496e-01
## Proportion of Variance 7.821322e-05 3.596768e-05 2.882684e-05
## Cumulative Proportion  9.999352e-01 9.999712e-01 1.000000e+00
##
## Loadings:
##     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## V1           0.133         0.294  0.427  0.760  0.344
## V2                         0.120        -0.147               -0.212
## V3           0.440  0.429  0.568  0.162 -0.456        -0.144
## V4           0.108  0.103                      -0.178  0.946  0.205
## V5                  0.133                       0.179  0.175 -0.859  0.337
## V6           0.429  0.584 -0.414 -0.369  0.310 -0.208 -0.116
## V7                                             -0.118 -0.147  0.314  0.929
## V8          -0.184  0.243 -0.214 -0.225 -0.193  0.834  0.108  0.250
## V9          -0.666  0.602         0.362        -0.229
## V10 -0.998
## V11          0.331        -0.594  0.681 -0.238
##     Comp.11
## V1
## V2   0.946
## V3  -0.191
## V4
## V5  -0.243
## V6
## V7
## V8
## V9
## V10
## V11
```

After computing the PCA of the covariance matrix, we get eleven principal components. The first PC accounts for 92% of the total variance of the observed variables. If we add the second PC, this percentage up will increase to 98%. The rest of the PCs add very little to the cumulative proportion variance.

```r
# Eigen values for covariance
temp_S_evals <- eigen(temp_S)$values; temp_S_evals
```

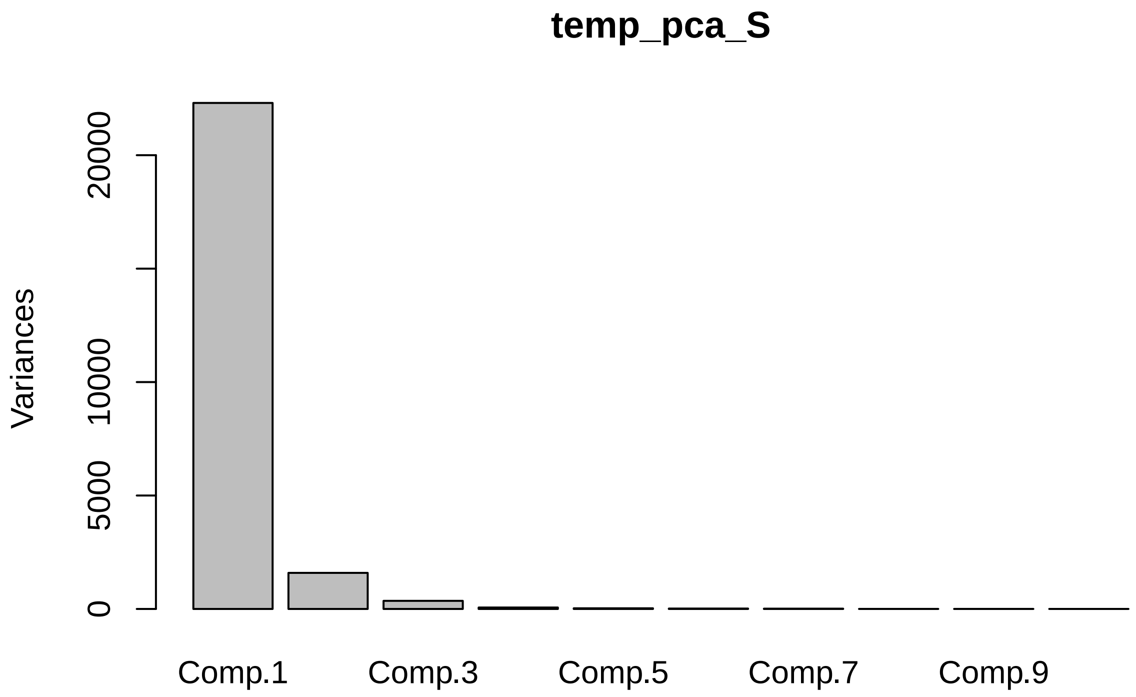
```
##  [1] 2.230350e+04 1.590679e+03 3.580457e+02 6.336652e+01 2.932695e+01
##  [6] 1.711486e+01 1.274780e+01 2.832997e+00 1.906924e+00 8.769315e-01
## [11] 7.028301e-01
```

```
# Mean of eigen values
mean(temp_S_evals)
```

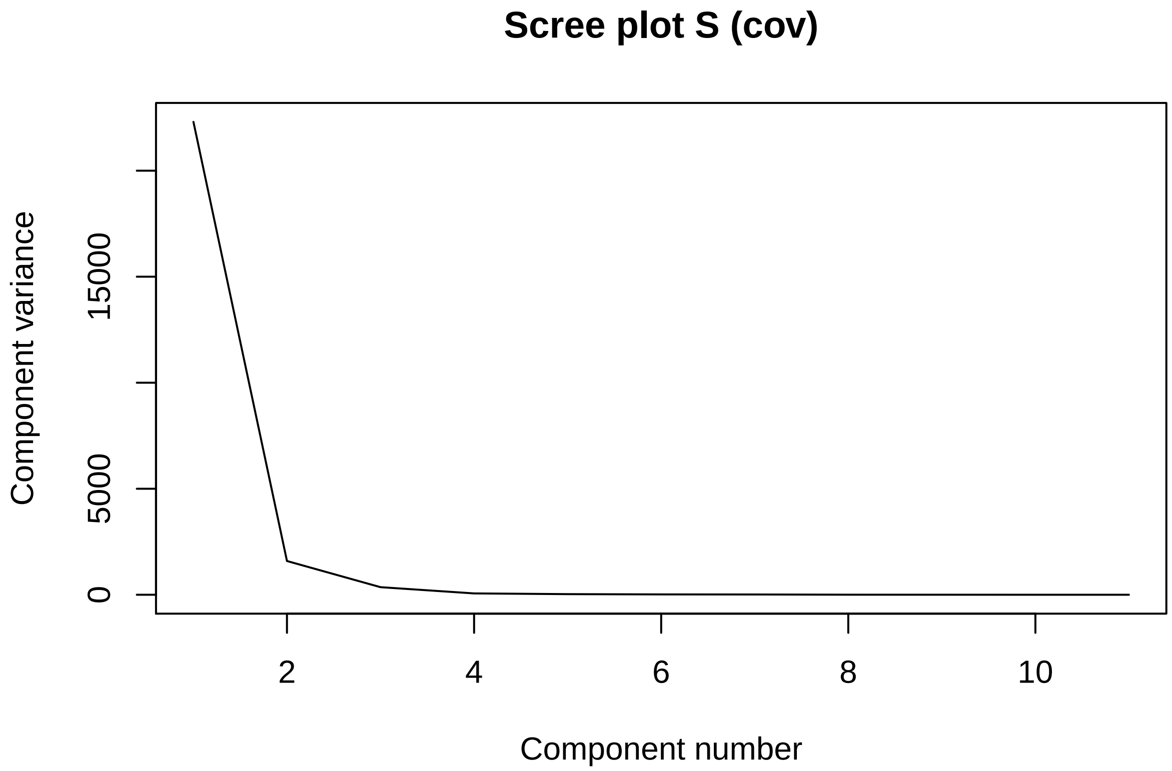
```
## [1] 2216.463
```

Similarly in the previous question, we only include in the PCs whose eigenvalues are greater than the average eigenvalue. In this case, we only include the first PC, and exclude the rest of the PCs.

```
# Scree diagram, covariance
plot(temp_pca_S)
```



```
# Scree plot, covariance
plot(temp_pca_S$sdev^2, xlab = "Component number",
     ylab = "Component variance", type = "l", main = "Scree plot S (cov)")
```



A scree plot shows how much variation each PC captures from the data. This scree plot is has one elbow at the second PC, with a steep curve that bends quickly and flattens out. This suggests that the only first PC is sufficient enough to describe the essence of the data. Based on our entire analysis, we should retain the first PC.

```
# Correlation of temp
temp_R <- cor(temp)

# PCA - correlation, temp
temp_pca_R <- princomp(covmat = temp_R)
summary(temp_pca_R, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.453619 1.4557940 1.0631514 0.87178959 0.59611694
## Proportion of Variance 0.547295 0.1926669 0.1027537 0.06909246 0.03230504
## Cumulative Proportion  0.547295 0.7399619 0.8427157 0.91180812 0.94411316
##                            Comp.6     Comp.7    Comp.8      Comp.9     Comp.10
## Standard deviation     0.50925676 0.34939323 0.3323980 0.244557333 0.205390249
## Proportion of Variance 0.02357659 0.01109778 0.0100444 0.005437117 0.003835014
## Cumulative Proportion  0.96768974 0.97878753 0.9888319 0.994269047 0.998104061
##                            Comp.11
## Standard deviation     0.144413757
## Proportion of Variance 0.001895939
## Cumulative Proportion  1.000000000
##
```

```
## Loadings:
##     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## V1   0.330                0.281  0.787  0.390
## V2   0.354 -0.193  0.107  0.230        -0.570         0.443 -0.157
## V3   0.392         0.110  0.141        -0.260         0.162  0.260
## V4   0.382         0.133        -0.167         0.665 -0.472         0.354
## V5   0.232 -0.530               -0.121  0.289 -0.450 -0.233 -0.530  0.110
## V6   0.362 -0.236  0.120 -0.135 -0.210  0.209        -0.200  0.517 -0.542
## V7                  0.795 -0.541  0.182        -0.110                0.114
## V8  -0.250 -0.502         0.151 -0.156  0.271  0.127  0.343  0.436  0.483
## V9  -0.311 -0.359  0.214  0.228  0.101         0.483  0.114 -0.294 -0.558
## V10        -0.468 -0.467 -0.500  0.438 -0.306
## V11  0.336  0.115 -0.185 -0.455 -0.187  0.382  0.262  0.562 -0.249
##     Comp.11
## V1   0.109
## V2   0.470
## V3  -0.800
## V4
## V5  -0.114
## V6   0.293
## V7
## V8
## V9  -0.146
## V10
## V11
```

After computing the PCA of the correlation matrix, we get six principal components. In this case, the first two PCs account for 74% of the total variance of the observed variables. If we add the third PC, this percentage will increase to 84%. If we add the fourth PC, this percentage will increase to 91%.

```
# Eigen values for correlation
temp_R_evals <- eigen(temp_R)$values; temp_R_evals
```

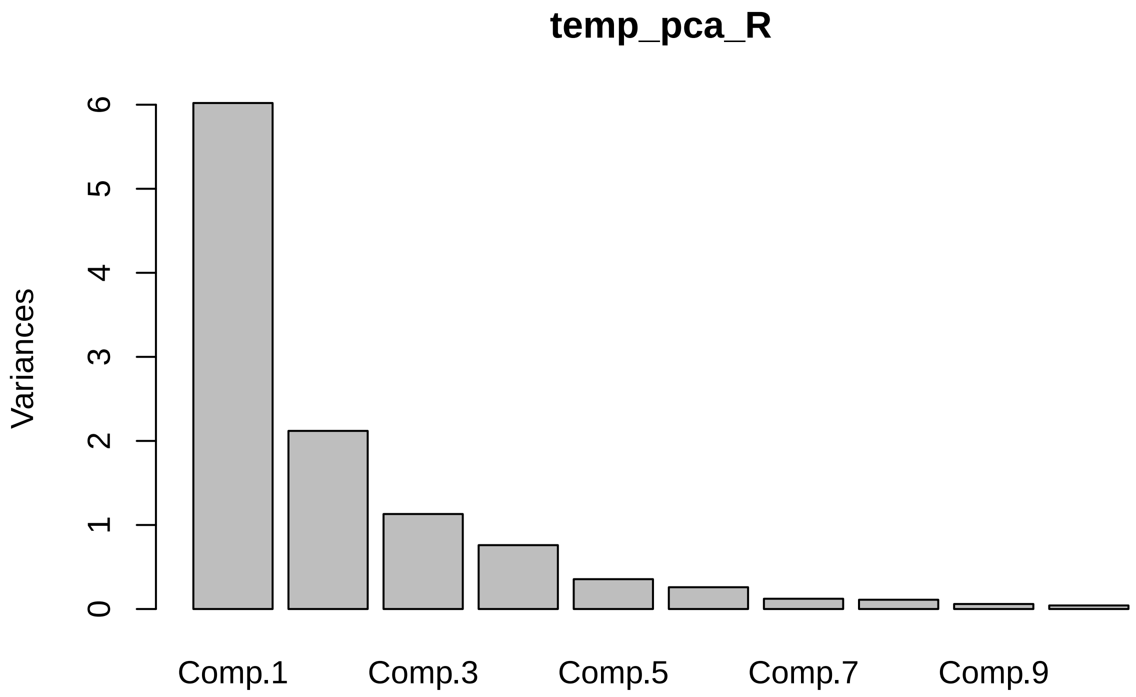
```
##  [1] 6.02024515 2.11933612 1.13029100 0.76001708 0.35535540 0.25934244
##  [7] 0.12207563 0.11048840 0.05980829 0.04218515 0.02085533
```

```
# Mean of eigen values
mean(temp_R_evals)
```

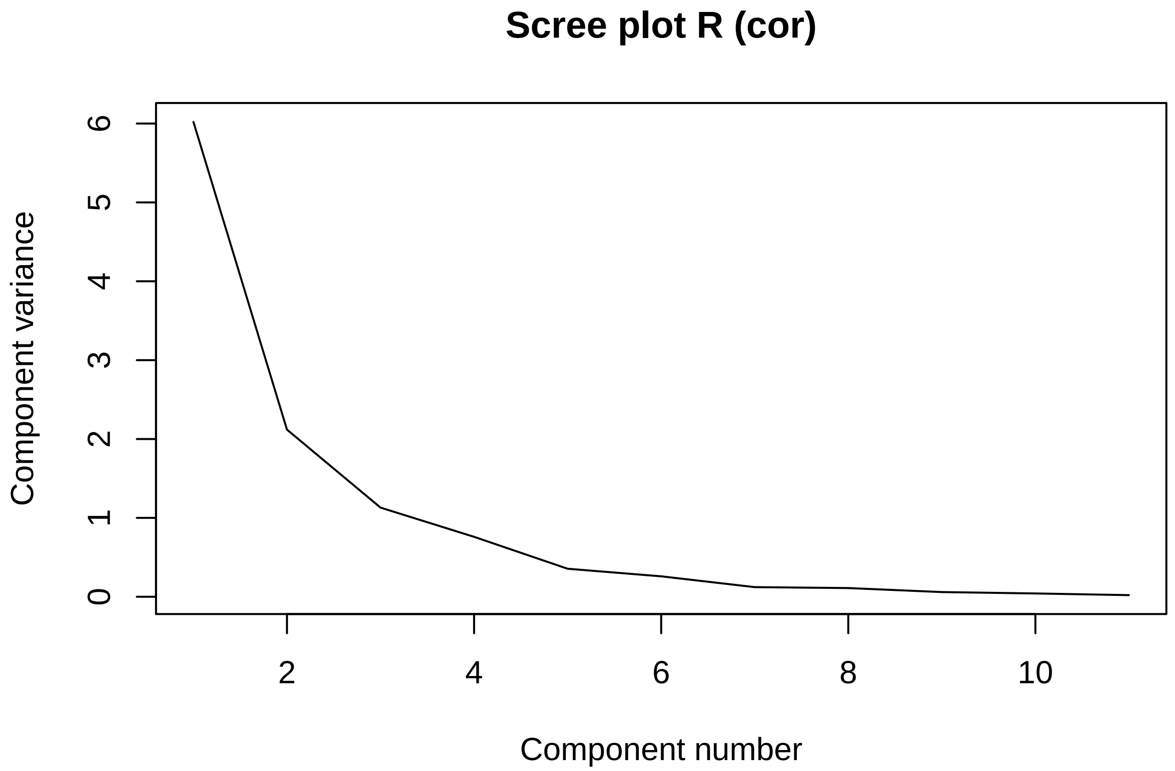
```
## [1] 1
```

Only the first three PCs are greater than the average eigenvalue, meaning that these are the PCs to include.

```
# Scree diagram, correlation
plot(temp_pca_R)
```



```r
# Scree plot, correlation
plot(temp_pca_R$sdev^2, xlab = "Component number",
     ylab = "Component variance", type = "l", main = "Scree plot R (cor)")
```



This scree plot is doesn't have really steep curve. The line bends and flattens out slowly, suggesting that we would need more PCs than normal. Based on our entire analysis, we should retain the first three PCs.

The first or second PCs would be enough for the covariance matrix with 92% or 98% of the variance accounted for. In comparison, the correlation matrix requires three PCs to account only 84% of the variance, while needing seven PCs to account for 98% of the variance. Therefore, it is more appropriate to use the covariance matrix S.