Stat 897 Spring 2017 Data Analysis Assignment 6

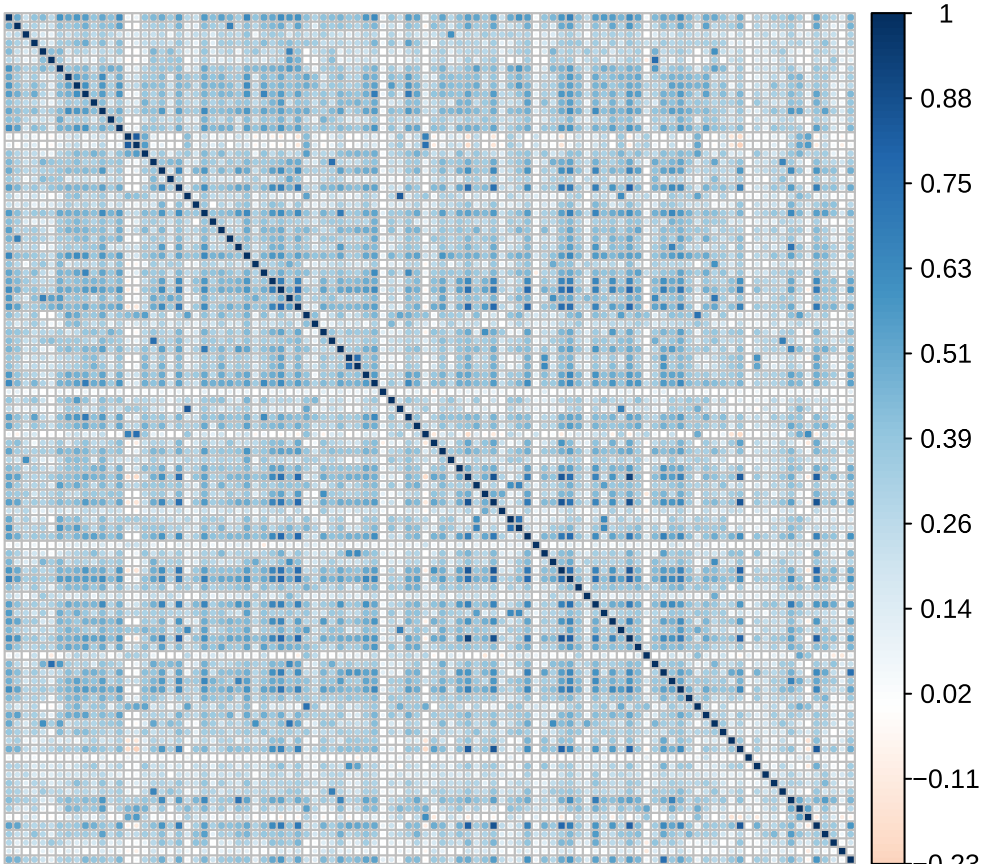
Penn State

Due October 8, 2017

In this assignment we will again use the stock return data (spreturns) from DAA 5 part two. For the purposes of this exercise you may assume the mean returns are zero if it helps.

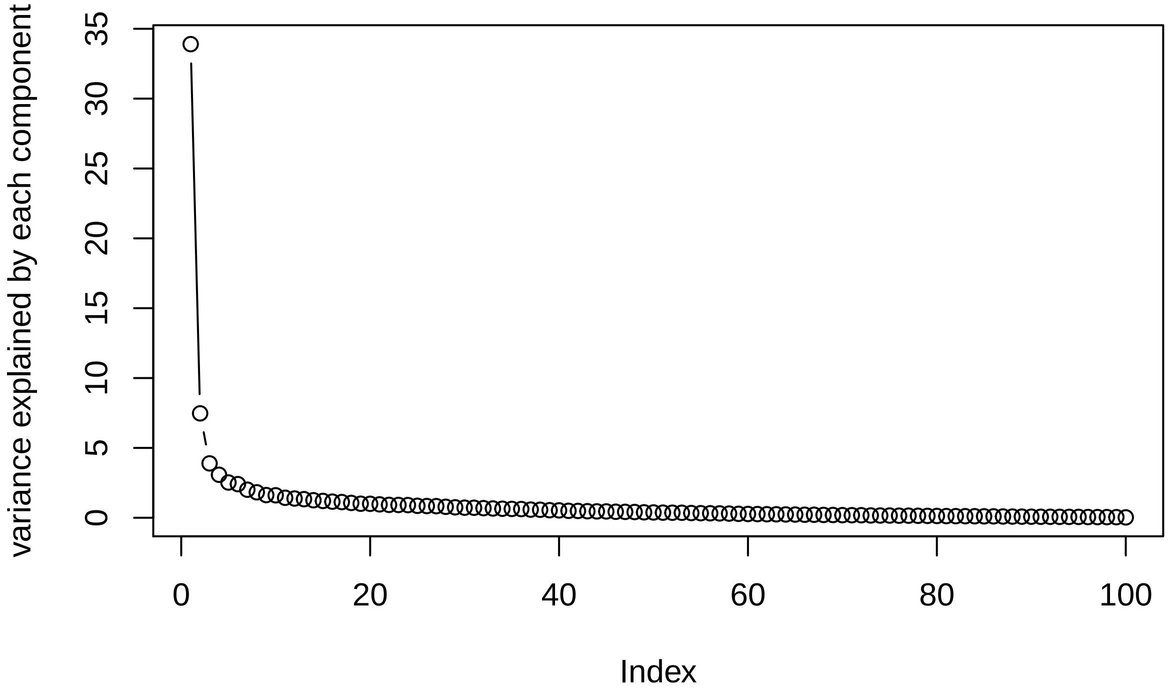
a) Restrict to the first 100 columns (first 100 stocks) in the data. Compute the covariance matrix of these stocks. Using the `corrplot` package, produce a visual of the *correlation* matrix. Do not print the covariance matrix. (I will take points off if you do because printing out a 100 x 100 matrix really isn't helpful, visuals are much better!)

```
suppressWarnings(library(corrplot))
load("../spreturns.Rda")
spreturns100 = spreturns[ ,1:100]
### covariance matrix
cov100 = var(spreturns100)
### correlation matrix
cor100 = cor(spreturns100)
### this is the plot for the correlation matrix NOT the covariance matrix
corrplot(cor100, is.corr = F, tl.pos = 'n')
```



b) Perform PCA on the stocks, and plot the variance explained. Do you see any natural "knee"?

```
pca.spre = prcomp(spreturns100, scale = TRUE)
plot(pca.spre$sdev ^ 2, type = 'b', ylab = 'variance explained by each component')
```



We do see a 'knee' shape after the first 3 principal components.

c) How would you interpret the factor or first principal component? How would you describe the covariance matrix that results from keeping only the projection onto the first or first and second components?

```
# percent variance explained by first component
pca.spre$sdev[1] ^ 2
```

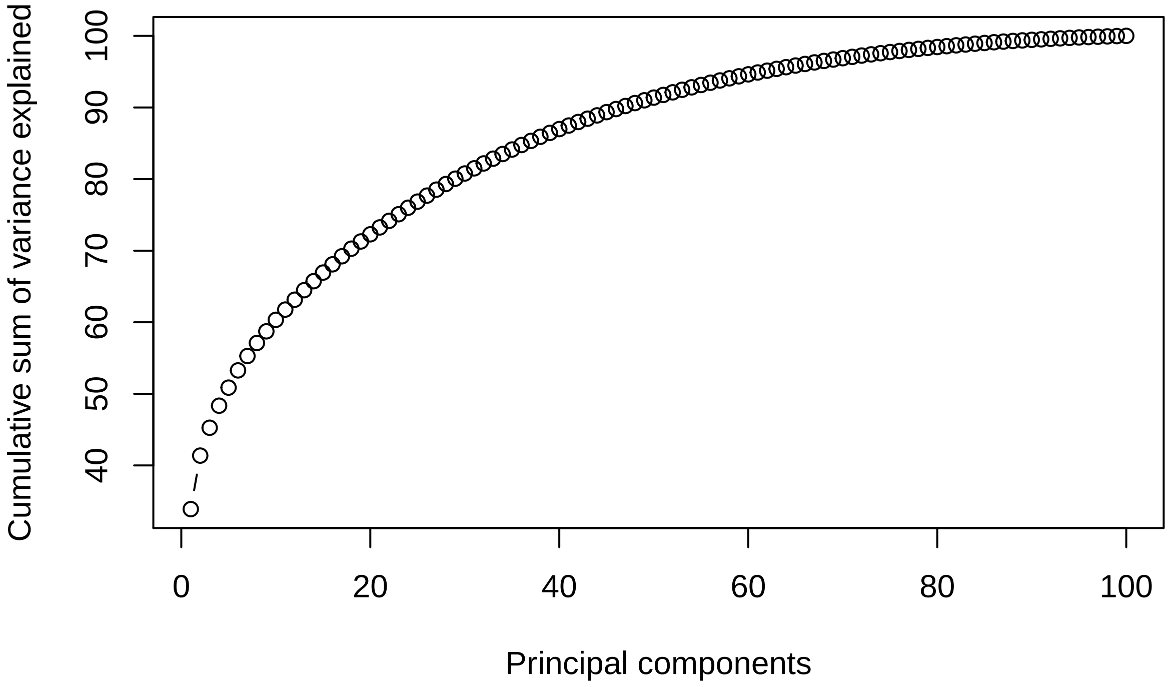
```
## [1] 33.90063
```

```
# percent variance explained by first and second component
sum(pca.spre$sdev[1:2] ^ 2)
```

```
## [1] 41.3735
```

The first principal component is the linear combination of variables that captures the most variance. In this problem, the first principal component captures 33.9% of the total variance and the first two components captures about 41.4% of the total variance.

```
# note that we can visualize the cumulative sum of the variance explained also
plot(cumsum(pca.spre$sdev ^ 2), type = "b", ylab = "Cumulative sum of variance explained",
     xlab = "Principal components")
```



```
# and we can check that all of them sum to 100 percent
sum(pca.spre$sdev ^ 2)
```

```
## [1] 100
```

d) Relate the covariance matrix method for PCA (eigendecomposition) to the direct SVD on the given 252 by 100 data matrix X.

The direct SVD decomposition on X gives us $X = UDV^\top$, where U, V are orthogonal matrix and D is a diagonal matrix. Under the assumption that **the mean returns are zero**, we have $S = \frac{1}{n-1}X^\top X$, where S is the sample covariance matrix. Thus $S = \frac{1}{n-1}X^\top X = \frac{1}{n-1}VDU^\top UDV^\top = \frac{1}{n-1}VD^2V^\top$, which is the eigen decompostion of S and the columns in V are the corresponding eigen vectors.