Homework 1

ST557

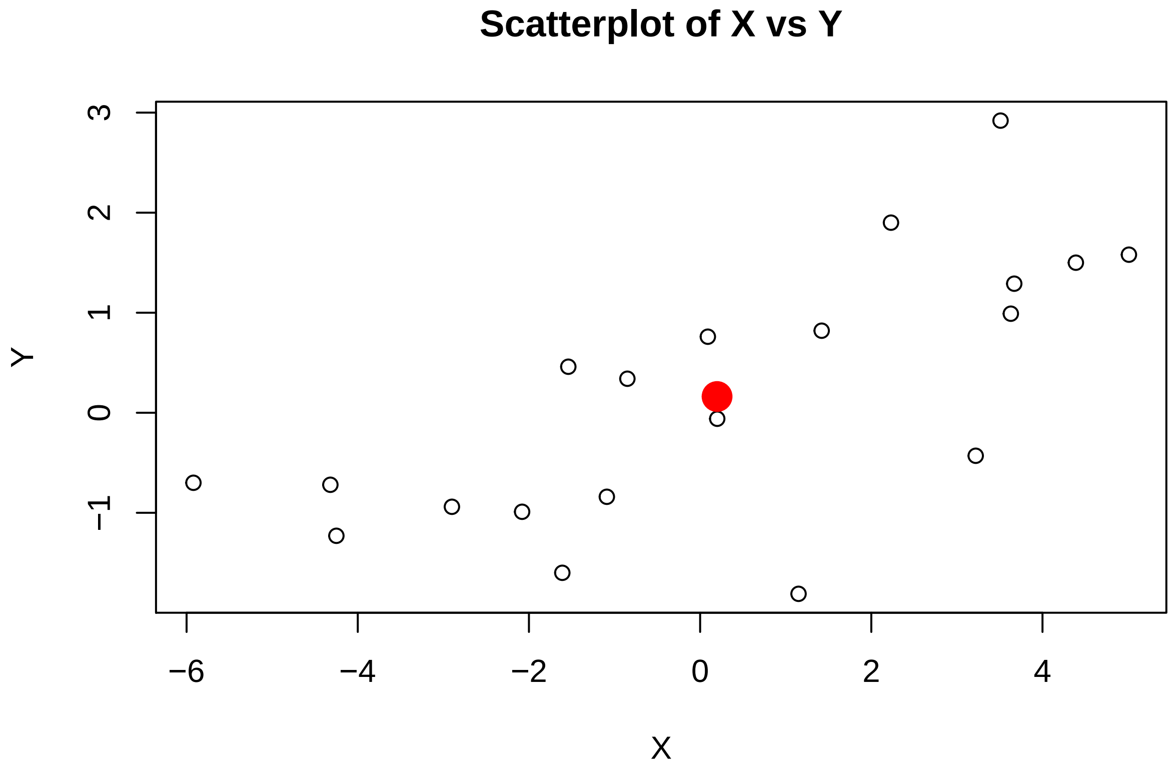
Nick Sun

October 8, 2019

Question 1

```r
data1 <- read_csv("HW1-1.csv")
sampvector <- apply(data1, 2, mean)

plot(data1$X,
     data1$Y,
     xlab = "X",
     ylab = "Y",
     main = "Scatterplot of X vs Y")
points(sampvector[1], sampvector[2], col = 2, cex = 2, pch = 19)
```



```r
cov(data1)
```

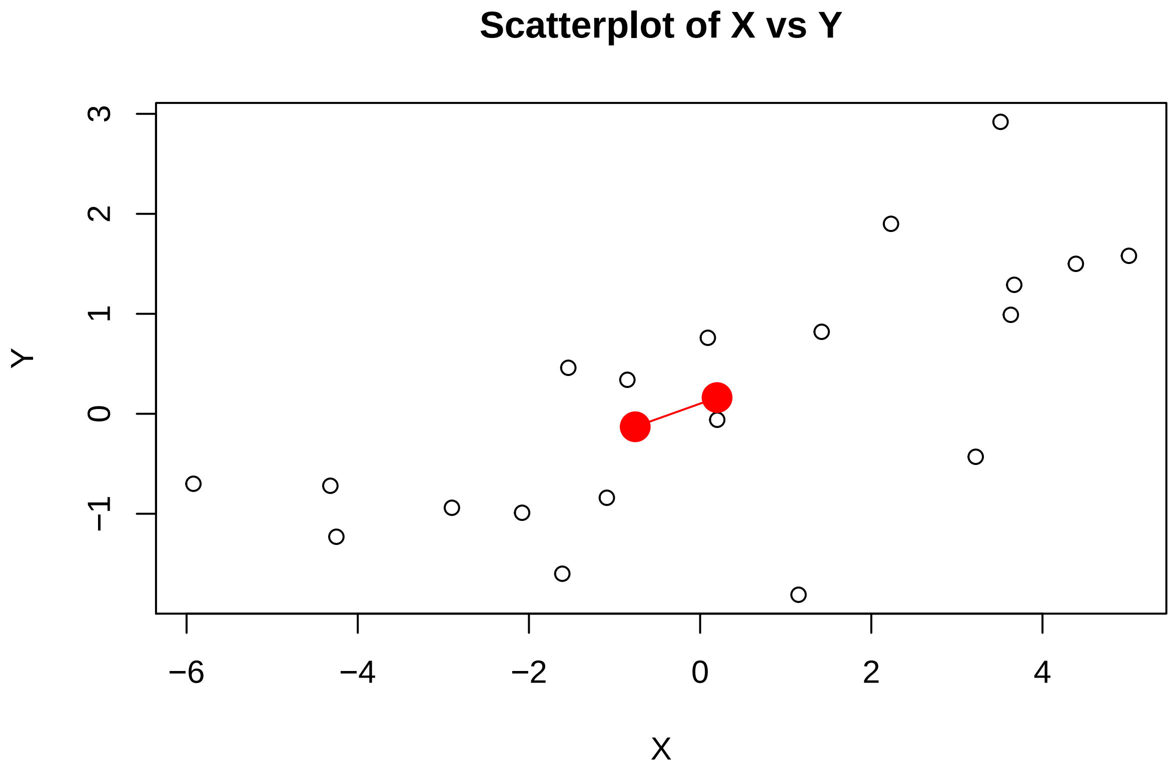
```
##           X        Y
## X 10.140227 2.852078
## Y  2.852078 1.668133
```

```
eigen(cov(data1))
```

```
## eigen() decomposition
## $values
## [1] 11.0108859  0.7974741
##
## $vectors
##            [,1]       [,2]
## [1,] -0.9564274  0.2919702
## [2,] -0.2919702 -0.9564274
```

```
eigenvectors <- eigen(cov(data1))$vectors
samp_plus_eig <- sampvector + eigenvectors[c(1, 2)]

plot(data1$X,
     data1$Y,
     xlab = "X",
     ylab = "Y",
     main = "Scatterplot of X vs Y")
points(sampvector[1], sampvector[2], col = 2, cex = 2, pch = 19)
points(samp_plus_eig[1], samp_plus_eig[2], col = 2, cex = 2, pch = 19)
lines(rbind(sampvector,
            samp_plus_eig),
      col = 2)
```



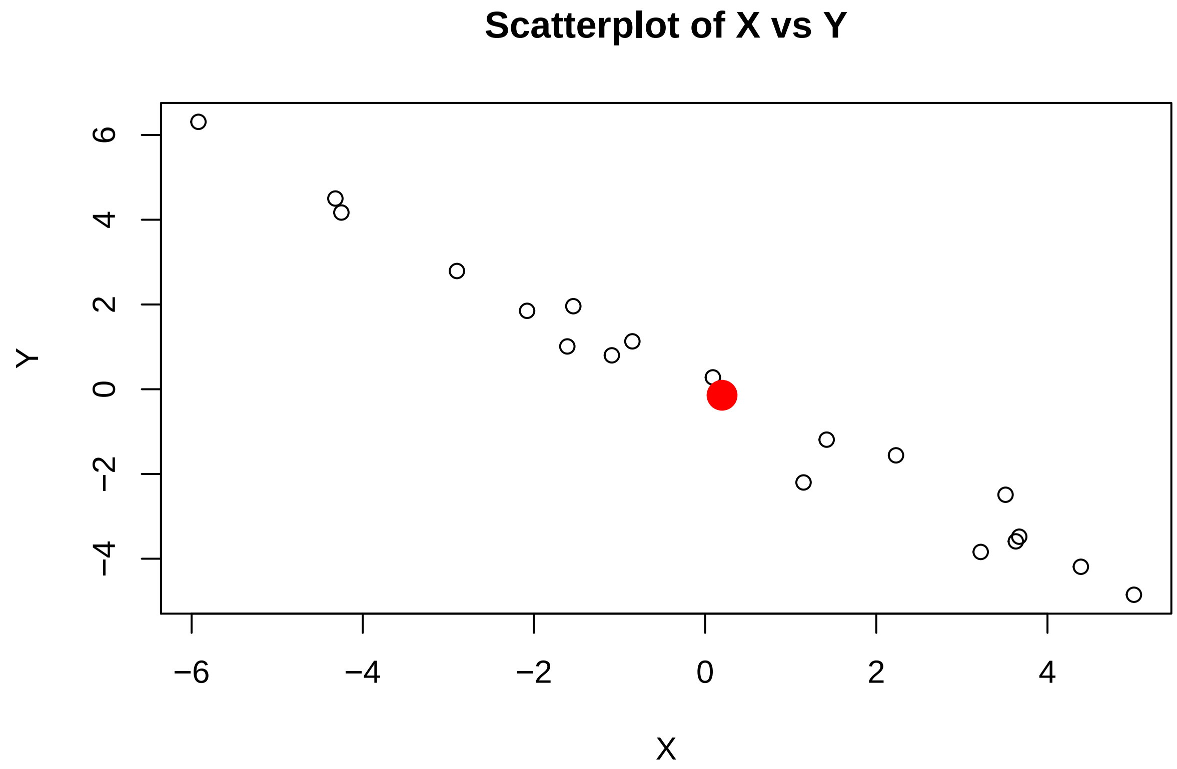
```
# lines(list(x = c(sampvector[1],
#                  samp_plus_eig[1]),
#            y = c(sampvector[2],
#                  samp_plus_eig[2])),
#       col = 2
# )
```

This eigenvector describes the axis of largest variance in our dataset. In fact, it can be shown that in principal component analysis, the direction first principal component is the first eigenvector of the covariance matrix which is what we have displayed here!

Question 2

```
data2 <- read_csv("HW1-2.csv")
sampvector <- apply(data2, 2, mean)

plot(data2$X,
     data2$Y,
     xlab = "X",
     ylab = "Y",
     main = "Scatterplot of X vs Y")
points(sampvector[1], sampvector[2], col = 2, cex = 2, pch = 19)
```



```r
cov(data2)
```

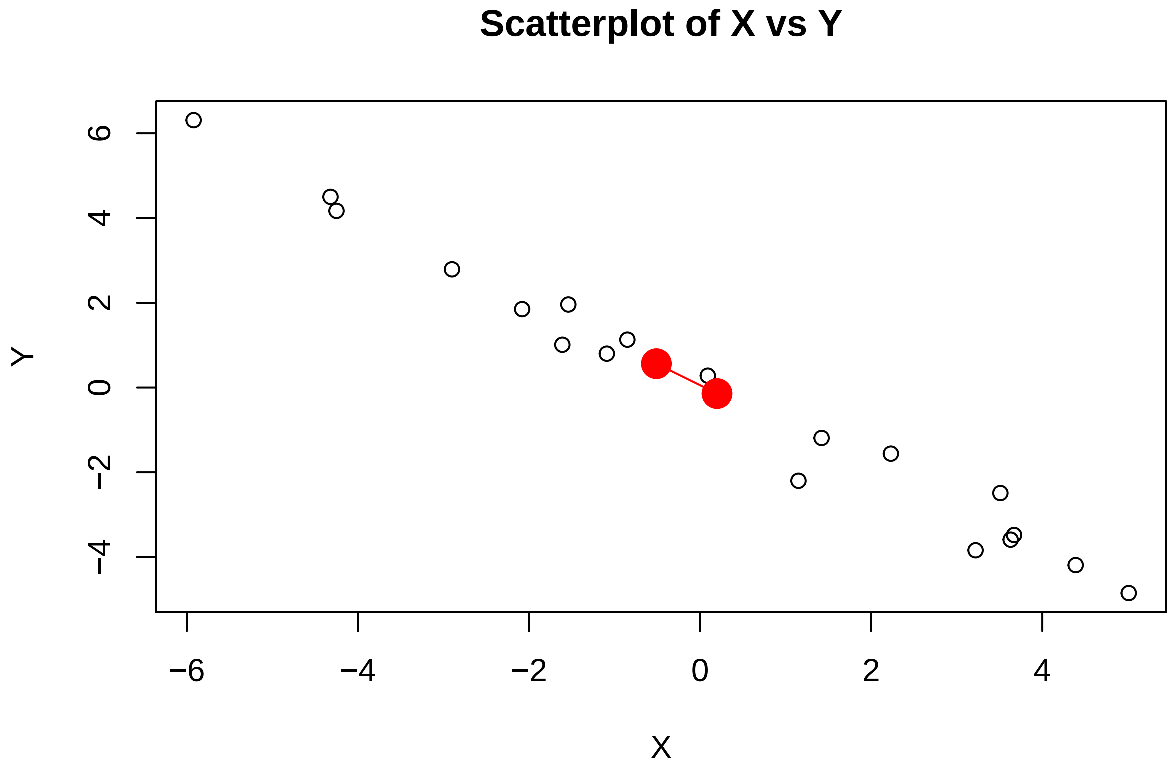
```
##           X         Y
## X 10.140227 -9.983968
## Y -9.983968 10.046978
```

```r
eigen(cov(data2))
```

```
## eigen() decomposition
## $values
## [1] 20.0776798  0.1095252
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7087559 -0.7054538
## [2,]  0.7054538 -0.7087559
```

```r
eigenvectors <- eigen(cov(data2))$vectors
samp_plus_eig <- sampvector + eigenvectors[c(1, 2)]

plot(data2$X,
     data2$Y,
     xlab = "X",
     ylab = "Y",
     main = "Scatterplot of X vs Y")
points(sampvector[1], sampvector[2], col = 2, cex = 2, pch = 19)
points(samp_plus_eig[1], samp_plus_eig[2], col = 2, cex = 2, pch = 19)
lines(rbind(sampvector,
            samp_plus_eig),
      col = 2)
```



Again, not that the eigenvector describes the "direction" of the axis with the greatest variance in the dataset! In this problem, this relationship is even more clear than in the first problem.

Question 3

Part a.

```
A = cbind(
  c(5.125, 3.875, 2.125, -1.125, 0),
  c(3.875, 5.125, -1.125, 2.125, 0),
  c(2.125, -1.125, 5.125, 3.875, 0),
  c(-1.125, 2.125, 3.875, 5.125, 0),
  c(0, 0, 0, 0, -3)
)

eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 10.0  8.0  4.5 -2.0 -3.0
##
## $vectors
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.5 -0.5  0.5  0.5    0
## [2,]  0.5 -0.5 -0.5 -0.5    0
## [3,]  0.5  0.5  0.5 -0.5    0
## [4,]  0.5  0.5 -0.5  0.5    0
## [5,]  0.0  0.0  0.0  0.0    1
```

Part b.

Not positive definite since there are negative eigenvalues.

An example of a vector that x such that $x^T Ax < 0$ is one of the eigenvectors that correspond to a negative eigenvalue. For example, we can take the fourth eigenvector above which corresponds with an eigenvalue of -2:

```
x <- eigen(A)$vectors[,4]
t(x)%*%A%*%x
```

```
##      [,1]
## [1,]   -2
```

Part c.

Let $x = 4v_1 + 2v_5$. Find Ax in terms of $v_1, v_5, \lambda_1, \lambda_5$.

Then

$$\begin{aligned} Ax &= A(4v_1 + 2v_5) \\ &= 4Av_1 + 2Av_5 \\ &= 4\lambda_1 v_1 + 2\lambda_5 v_5 \end{aligned}$$

Question 4

Part a.

Here is the sample mean vector.

```
iris <- read_csv("IrisData.csv")
apply(iris, 2, mean)[-5]
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

Part b.

Here is the sample mean vector for each individual species.

```
aggregate(iris[1:4], by = list(iris$Species), mean, na.rm = TRUE)
```

```
##   Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1       1        5.006       3.428        1.462       0.246
## 2       2        5.936       2.770        4.260       1.326
## 3       3        6.588       2.974        5.552       2.026
```

Part c.

Here is the sample covariance matrix.

```
cor(iris[,-5])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Petal length and width are **highly** correlated with each other . Petal length and width are also highly correlated with sepal length.

Part d.

Here are the sample covariance matrices for the 3 species.

```
cor(subset(iris, Species == "1")[,-5])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

In species 1, sepal length is correlated with sepal width but that is about it.

```
cor(subset(iris, Species == "2")[,-5])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.5259107    0.7540490   0.5464611
## Sepal.Width     0.5259107   1.0000000    0.5605221   0.6639987
## Petal.Length    0.7540490   0.5605221    1.0000000   0.7866681
## Petal.Width     0.5464611   0.6639987    0.7866681   1.0000000
```

In species 2, petal length is correlated with sepal length and petal length is correlated with petal width.

```
cor(subset(iris, Species == "3")[,-5])
```

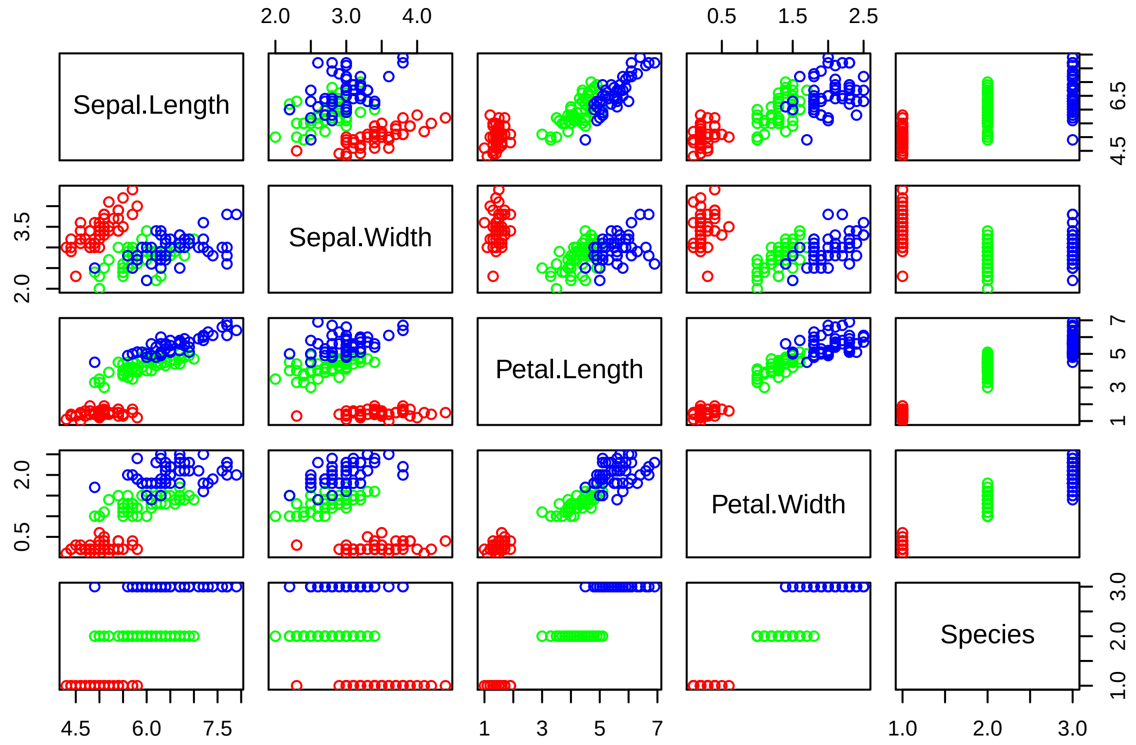
```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.4572278    0.8642247   0.2811077
## Sepal.Width     0.4572278   1.0000000    0.4010446   0.5377280
## Petal.Length    0.8642247   0.4010446    1.0000000   0.3221082
## Petal.Width     0.2811077   0.5377280    0.3221082   1.0000000
```

In species 3, petal length is highly corelated with sepal length but that is about it

The distinguishing characteristics: + In species 2 the petal length is highly correlated with sepal length, which is different than in the other two species. + In species 1, sepal length and width are highly correlated which is different than the two other species.

Part e.

```r
pairs(iris[,1:5],
      col = c("red", "green", "blue")[unclass(iris$Species)])
```



```r
# legend(x= "bottomright",
#        legend = as.vector(unique(iris$Species)),
#        fill=c("red", "green3", "blue"),
#        title = "Flower Species",
#        bty='L',
#        xpd = TRUE)
```

There appears to be distinct point clusters which relate to each of the three flower species. If we were given information on petal width, we may be able to distinguish which flower the species originated from - for example, if the petal width is very small, we would have reason to believe that the flower is from species 1.

If we were to pick two variables to receive information on in order to distinguish the flower species, I would probably go with petal length and petal width. In the scatterplot of these two variables, there appears to be the most clear clustering of the three species. High length and width correspond to species 3, low length and width corresponds to species 1, and lengths and widths in the middle correspond to species 2.

Question 5

Part a.

B is a symmetric matrix since as the product A^TA, for each element in B b_{ij}, it is both the product of the ith row of A^T and the jth column of A and the product of the jth column of A^T and the ith row of A.

Part b.

Since B is a symmetric matrix, we can use spectral decomposition. Spectral decomposition of B gives us $B = V^T\Lambda V$ where V is a matrix composed of eigenvectors and Λ is a diagonal matrix of corresponding eigenvalues.

Since B is also the product A^TA and A can be rewritten using singular value decomposition as $A = UDV^T$, we get another decomposition of B where $B = VD^2V^T$ where V is a matrix of eigenvectors and D is a diagonal matrix.

Now note that $B = VD^2V^T = V\Lambda V^T$. For this to be true, $\Lambda = D^2$. Since the diagonal elements of D^2 are nonnegative, the eigenvalues of B are nonnegative. Therefore, B is semi-positive definite.

Part c.

Note that $(X - \bar{X})^T(X - \bar{X})$ is just a symmetric matrix and multiplying by the scalar $\frac{1}{n-1}$ doesn't change the sign of any eigenvalues.

By our previous result then, S has no negative eigenvalues and is semi-positive definite.