Assignment04 P01

Dawu Liu

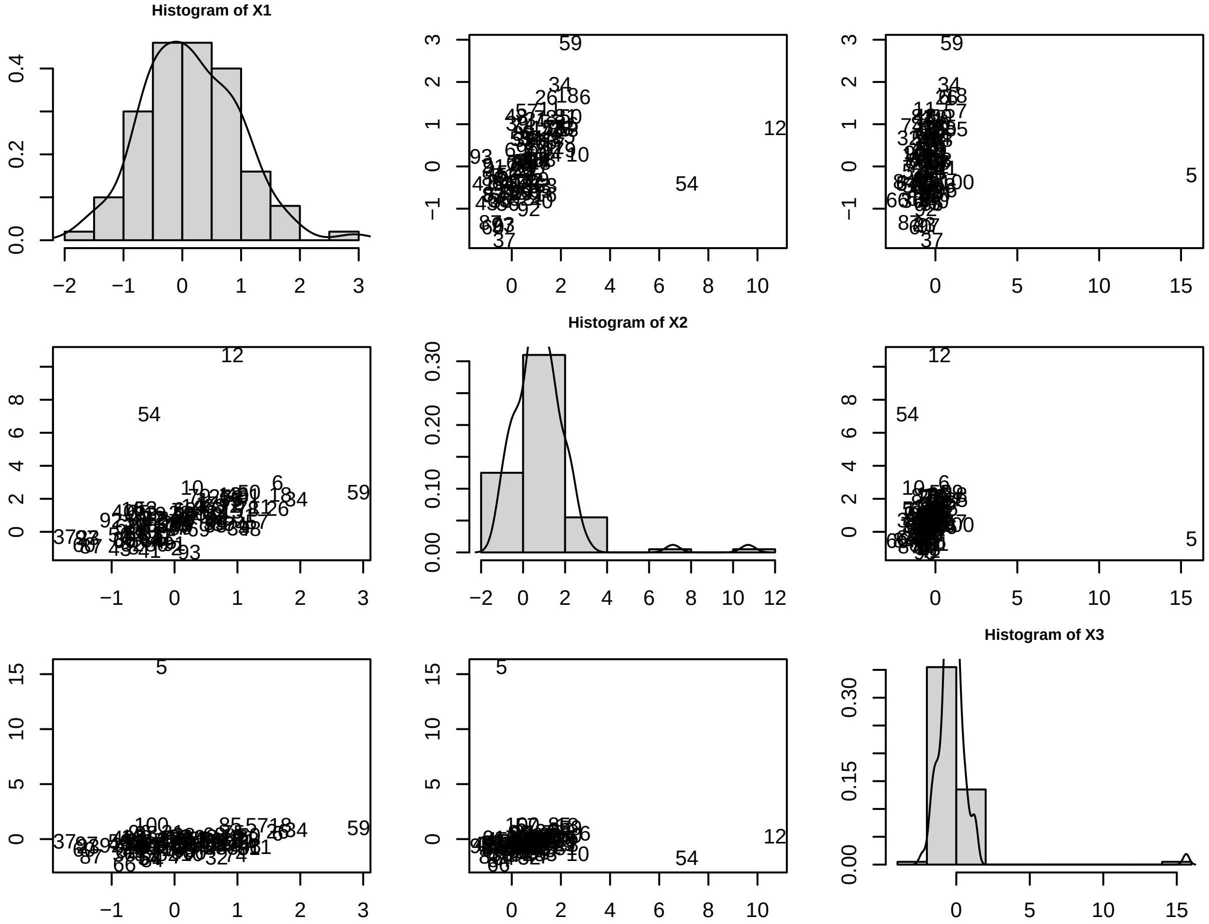
(a) Sample covariance matrix:

$$S = \begin{bmatrix} 0.65821889919388 & 0.54853765690655 & 0.175868141491043 \\ 0.54853765690655 & 2.32605135623547 & 0.0219115575942205 \\ 0.175868141491043 & 0.0219115575942205 & 3.06394221761802 \end{bmatrix} \tag{1}$$

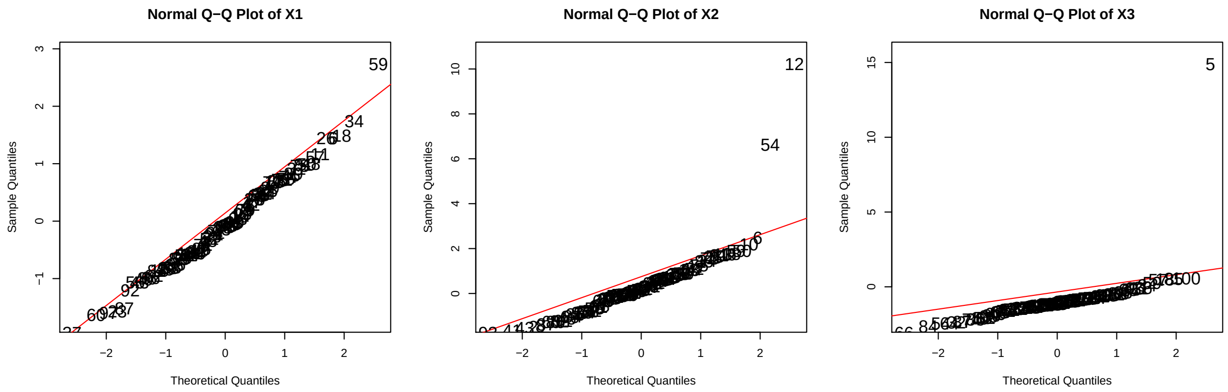
mean vector:

$$\tilde{\bar{x}} = \begin{pmatrix} 0.1441416 \\ 0.8633031 \\ -0.1925616 \end{pmatrix} \tag{2}$$

(b) Matrix scatter plot with histograms being the diagonals, the numbers shown on the plot are the row numbers (observations).

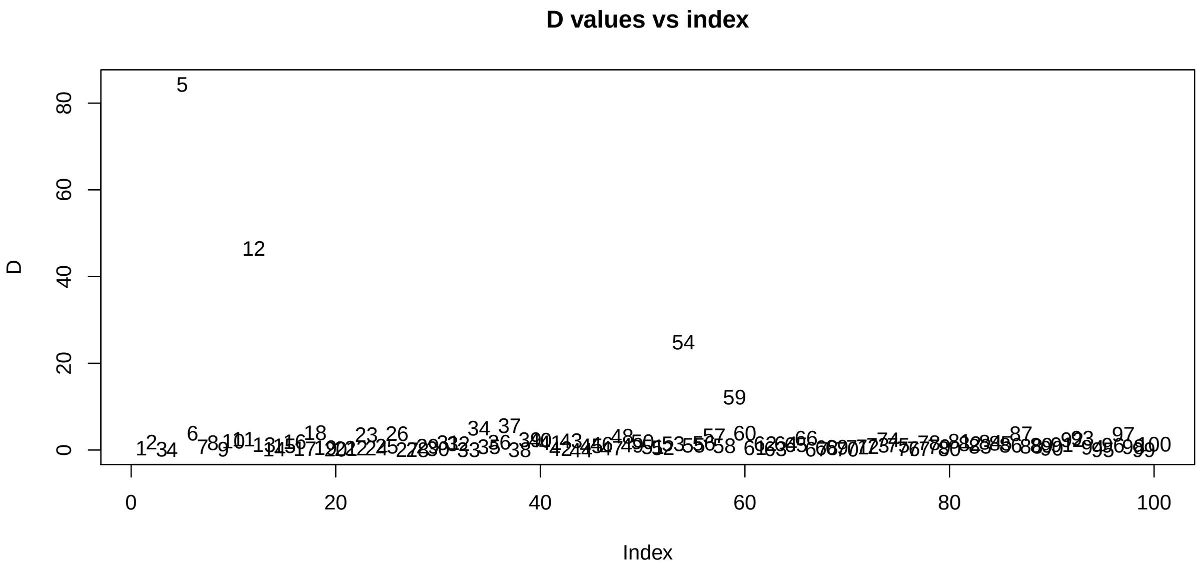


Univariate q-q plots:



(c) Statistical distance vector $\tilde{D}$ and it's values versus index plot: $\tilde{D} =$

```
##   [1]  0.493  1.816  0.214  0.102 84.314  3.857  0.747  1.692  0.222  1.977
##  [11]  2.540 46.501  1.210  0.229  0.926  1.891  0.367  3.969  0.495  0.167
##  [21]  0.315  0.423  3.511  0.482  0.864  3.720  0.113  0.112  0.913  0.334
##  [31]  1.681  1.489  0.081  5.056  0.707  1.827  5.612  0.087  2.409  2.252
##  [41]  1.747  0.323  2.142  0.033  0.948  1.361  0.401  3.212  1.148  1.917
##  [51]  0.741  0.512  1.387 24.914  1.024  1.606  3.221  1.008 12.166  3.891
##  [61]  0.556  1.395  0.293  1.409  1.258  2.762  0.222  0.475  0.583  0.171
##  [71]  0.733  0.819  1.075  2.467  1.102  0.300  0.336  1.644  0.730  0.347
##  [81]  1.985  1.469  0.922  1.765  1.526  1.144  3.784  0.903  1.129  0.429
##  [91]  1.189  2.356  2.695  0.687  0.093  1.080  3.624  0.793  0.035  1.294
```



(d) The 1st, 2nd, and 3rd variables are represented by X_1, X_2, and X_3 respectively.
The scatterplot indicates the outliers are observation 12 and 54 in X_1 vs X_2; 5 in X_1 vs X_3; 5, 12 and 54 in X_2 vs X_3. Even 59 appears to be a bit far away from the cloud, it still follows the linear trend. By looking at the q-q plots, 12 and 54 are significantly above the straight line in X_2, so as 5 for X_3. The $\tilde{D}$ values vs index plot shows that observation 5, 12, and 54 have the top three largest distances. Thus we conclude the outliers are observation **5, 12, 54**.

(e) New sample covariance matrix:

$$S = \begin{bmatrix} 0.668136852725533 & 0.517193086605832 & 0.227261815320189 \\ 0.517193086605832 & 0.937622838632426 & 0.263446058804402 \\ 0.227261815320189 & 0.263446058804402 & 0.499434430158578 \end{bmatrix} \tag{3}$$

New mean vector:

$$\tilde{\tilde{x}} = \begin{pmatrix} 0.1453851 \\ 0.7103047 \\ -0.3447489 \end{pmatrix} \tag{4}$$

Subtract the new covariance matrix from the old covariance matrix, we get:

$$\text{difference, old - new} = \begin{bmatrix} -0.0099 & 0.0313 & -0.0514 \\ 0.0313 & 1.3884 & -0.2415 \\ -0.0514 & -0.2415 & 2.5645 \end{bmatrix} \tag{5}$$

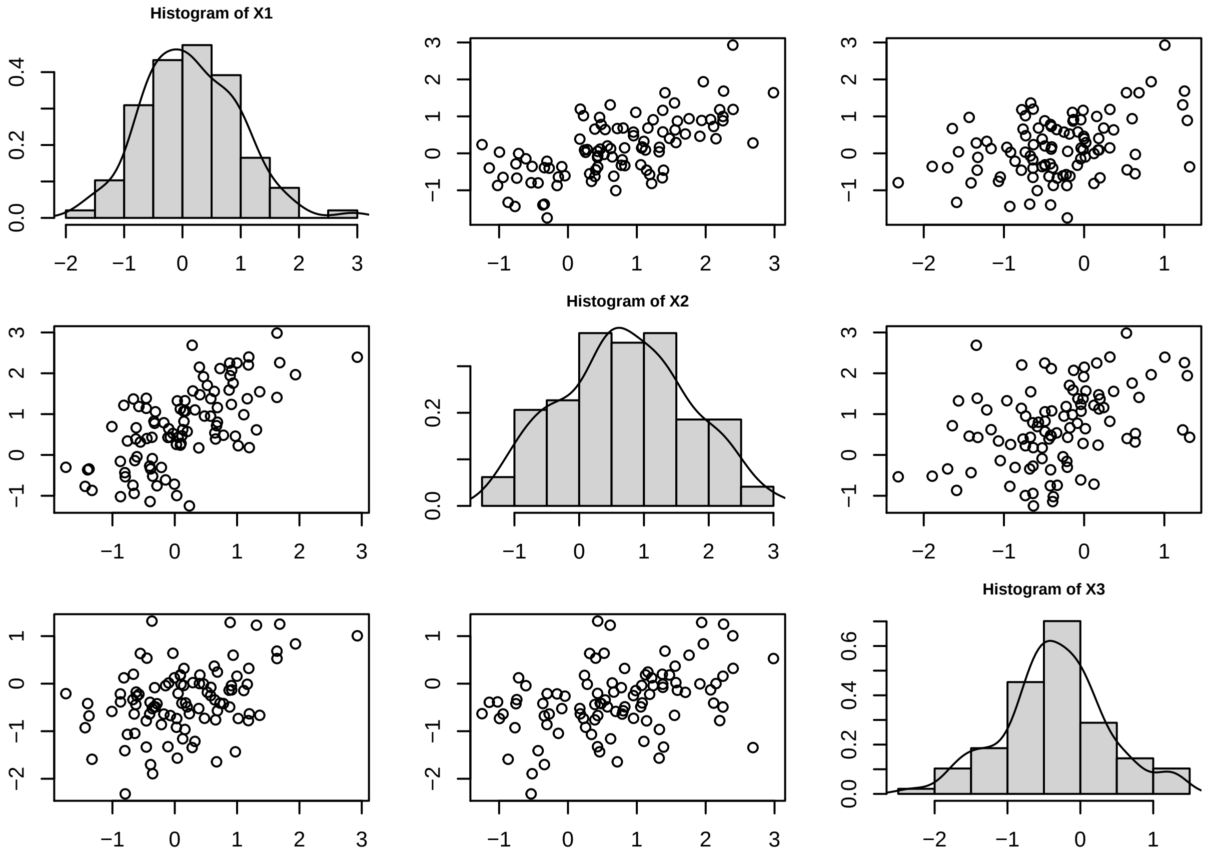
Removing the outliers affects the covariances across all three variables. The biggest covariance change is between the 2nd and 3rd variables. It also significantly decreases the variances of the 2nd and 3rd variables themselves.

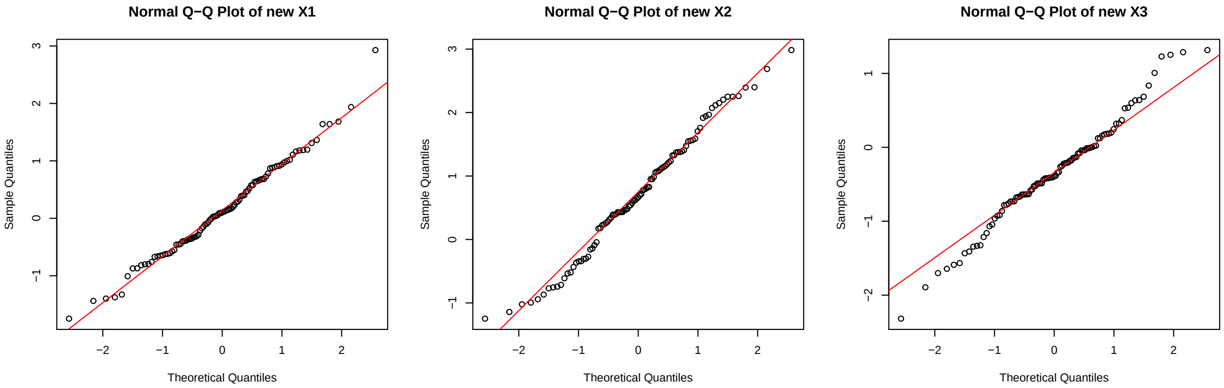
Subtract the new mean vector from the old mean vector, we get:

$$\text{difference, old - new} = \begin{pmatrix} -0.0012 \\ 0.1530 \\ 0.1522 \end{pmatrix} \tag{6}$$

There is a significant decrease in the mean of the 2nd and 3rd variables (by 0.15).

(f) Matrix scatter plot and univariate q-q plots for the new data set:





This evidence is fairly consistent with a normal distribution. The histograms of all three variables appear to be normal. The data points in matrix scatter plots for each of all three variables appear to fit in an ellipse pretty well, suggesting each pair of the the data is bivariate normal. Also, in the q-q plots, the points in all three variables appear to fit the straight lines well, only the third variable being slightly light tailed.

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

```
rm(list = ls())
library(MESS)
X <- read.table("C:/Users/John/Desktop/STAT 445/Data/assignment4_data1.txt", sep = ",")
#a
S <- cov(X) ; x_bar <- colMeans(X)
S; x_bar
```

```
##           V1         V2         V3
## V1 0.6582189 0.54853766 0.17586814
## V2 0.5485377 2.32605136 0.02191156
## V3 0.1758681 0.02191156 3.06394222
```

```
##         V1         V2         V3
##  0.1441416  0.8633031 -0.1925616
```

```
#b
par(mfcol=c(3, 3), mar=c(2,2,2,2))
for(i in 1:3){
  for(j in 1:3) {
    if(i != j){
      plot(X[[i]],X[[j]],type="n", xlab = paste0("X", i), ylab = paste0("X", j))
      text(X[[i]],X[[j]], cex = 0.7, pos=2)}
    else{
      hist(X[[i]], xlab = paste0("X", i),
           main = paste0("Histogram of X", i), prob=TRUE,cex.main=0.8)
      lines(density(X[[i]]))}
  }
}
```

```
par(mfrow=c(1, 3), mar=c(2,2,2,2))
q <- vector("list", length = 3)
for (i in 1:3) {
  q[[i]]=qqnorm(X[[i]],type="n", main = paste0("Normal Q-Q Plot of X",i));qqline(X[[i]],col="red")
  text(q[[i]]$x, q[[i]]$y,  pos = 1)
}
```

```
#c
D <- c()
for (i in 1:nrow(X)) {
  x_i <- t(X[i,])
  D[i]=t(x_i-x_bar)%*%solve(S)%*%(x_i-x_bar)
}
round(D,3)
```

```
##   [1]  0.493  1.816  0.214  0.102 84.314  3.857  0.747  1.692  0.222  1.977
##  [11]  2.540 46.501  1.210  0.229  0.926  1.891  0.367  3.969  0.495  0.167
##  [21]  0.315  0.423  3.511  0.482  0.864  3.720  0.113  0.112  0.913  0.334
##  [31]  1.681  1.489  0.081  5.056  0.707  1.827  5.612  0.087  2.409  2.252
##  [41]  1.747  0.323  2.142  0.033  0.948  1.361  0.401  3.212  1.148  1.917
##  [51]  0.741  0.512  1.387 24.914  1.024  1.606  3.221  1.008 12.166  3.891
##  [61]  0.556  1.395  0.293  1.409  1.258  2.762  0.222  0.475  0.583  0.171
##  [71]  0.733  0.819  1.075  2.467  1.102  0.300  0.336  1.644  0.730  0.347
##  [81]  1.985  1.469  0.922  1.765  1.526  1.144  3.784  0.903  1.129  0.429
##  [91]  1.189  2.356  2.695  0.687  0.093  1.080  3.624  0.793  0.035  1.294
```

```
plot(D, main = "D values vs index", type="n");text(D)

#e
new_X <- X[-c(5,12,54),]
S_new = cov(new_X)
S_new; round(S-S_new,4)
```

```
##           V1        V2        V3
## V1 0.6681369 0.5171931 0.2272618
## V2 0.5171931 0.9376228 0.2634461
## V3 0.2272618 0.2634461 0.4994344

##         V1      V2      V3
## V1 -0.0099  0.0313 -0.0514
## V2  0.0313  1.3884 -0.2415
## V3 -0.0514 -0.2415  2.5645
```

```
new_x_bar = colMeans(new_X)
new_x_bar; round(x_bar-new_x_bar,4)
```

```
##        V1         V2         V3
##  0.1453851  0.7103047 -0.3447489

##      V1      V2      V3
## -0.0012  0.1530  0.1522
```

```
#f
par(mfcol=c(3, 3), mar=c(2,2,2,2))
```

```r
for(i in 1:3){
  for(j in 1:3) {
    if(i != j){
      plot(new_X[[i]],new_X[[j]],xlab = paste0("X", i), ylab = paste0("X", j))
    }
    else{
      hist(new_X[[i]], xlab = paste0("X", i),
           main = paste0("Histogram of X", i), prob=TRUE,cex.main=0.8)
      lines(density(X[[i]]))}
  }
}
```

```r
par(mfrow=c(1, 3), mar=c(2,2,2,2))
for (i in 1:3) {
  qqnorm(new_X[[i]], main = paste0("Normal Q-Q Plot of new X",i));qqline(X[[i]],col="red")
}
```