Multivariate Statistical Analysis - Exercise Session 4

04.02.2022

Problem 1: Robust PCA

a) Install the package `rrcov` and read the data

Install the package `rrcov`.

```
install.packages("rrcov")
```

Import package `rrcov` and read the data.

```
library(rrcov)
wood <- read.table("wood.txt", header = TRUE, sep = "\t")
head(wood)
```

```
##         X1       X2        X3        X4       X5
## 1 23.93742 32.54228 0.7503695 0.8234348 1.160645
## 2 24.08319 35.34119 0.8314633 0.8942561 1.354641
## 3 23.38803 33.68976 0.8164181 0.8044027 1.274952
## 4 25.55386 37.34970 0.9117383 0.8044027 1.235940
## 5 24.20744 33.37664 0.7903982 0.8505827 1.231136
## 6 23.10844 33.80828 0.8064043 0.8556289 1.231136
```

The data set in this exercise is a modified version from Draper and Smith (1966) and it was used to determine the influence of anatomical factors on wood specific gravity, with five explanatory variables. The data is contaminated by replacing a few observations with outliers.

b) Plot the variables pairwise

Outliers are somewhat visible from Figure 1. Still, pairwise scatter plots are quite hard to read when data is high dimensional.

```
# Color possible outliers
color_outliers <- rep("black", nrow(wood))
color_outliers[c(10,12,13,15)] <- "grey"

# Plotting
pairs(wood, pch = 16, upper.panel = NULL, col = color_outliers)
```

c) Sample covariance and MCD

MCD based location and scatter estimates are calculated with `CovMcd` function. Parameter `alpha` controls the size of the subsets over which the determinant is minimized. Function `CovMcd` returns an S4 object. Most relevant fact about S4 objects regarding this course is that slots of S4 object are accessed with `@` not with `$`.

```
cov_regular <- cov(wood)
cov_mcd <- CovMcd(wood, alpha = 0.5)

cov_regular
```

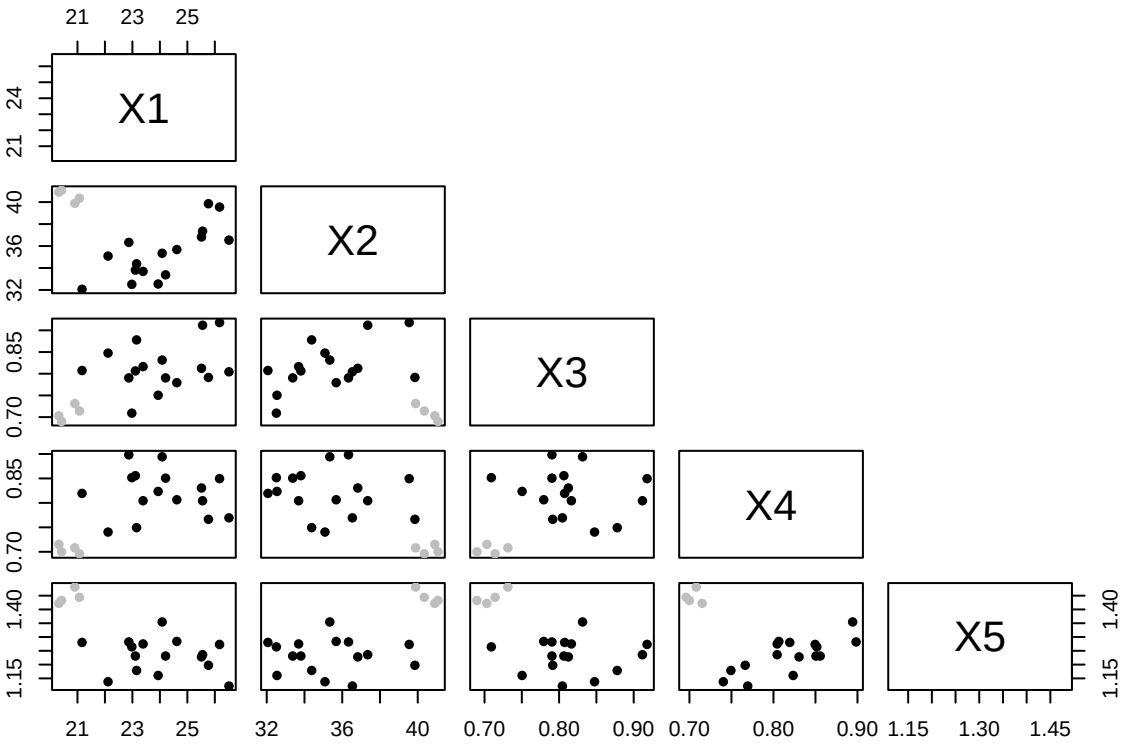

Figure 1: Pairwise scatter plots of variables.

```
##             X1          X2           X3           X4           X5
## X1  3.80227921 -0.90443869  0.078464066  0.060427001 -0.131939002
## X2 -0.90443869  9.06329355 -0.044136873 -0.112919979  0.180129369
## X3  0.07846407 -0.04413687  0.004243031  0.001595618 -0.003700091
## X4  0.06042700 -0.11291998  0.001595618  0.003984759 -0.002398752
## X5 -0.13193900  0.18012937 -0.003700091 -0.002398752  0.010638865
```

```
cov_mcd@cov
```

```
##             X1           X2           X3           X4           X5
## X1  6.49939325  8.256688895  0.097614189 -0.016889835 -0.075246561
## X2  8.25668889 14.441998803  0.270597883 -0.012891183  0.009801839
## X3  0.09761419  0.270597883  0.010140643 -0.001317422  0.000815357
## X4 -0.01688983 -0.012891183 -0.001317422  0.004319414  0.004491941
## X5 -0.07524656  0.009801839  0.000815357  0.004491941  0.010669458
```

d) Regular and robust Mahalanobis distances

Note that function `mahalanobis` returns squared mahalanobis distances.

```
maha_regular <- sqrt(mahalanobis(wood, center = colMeans(wood),
                                 cov = cov_regular))
maha_robust <- sqrt(mahalanobis(wood, center = cov_mcd@center,
                                cov = cov_mcd@cov))
n <- nrow(wood)

plot(rep(1:n, 2), c(maha_regular, maha_robust), pch = 16,
     col = c(rep("green", n), rep("blue", n)),
     xlab = "Observation", ylab = "Mahal. distance")

legend("topleft", col = c("green", "blue"), cex = 0.8,
       legend = c("Regular", "MCD"), pch = 16)
```

Figure 2 shows that with MCD estimates we can spot potential outliers, however, by using sample mean and sample covariance outliers are not visible.

e)

It seems that scales of variables are different. Additionally, we do not know if variables have the same units. Thus we perform correlation based PCA.

```
apply(wood, 2, range)
```

```
##            X1       X2        X3        X4       X5
## [1,] 20.32240 32.06244 0.6898171 0.6959214 1.122324
## [2,] 26.51415 41.07311 0.9179452 0.8983566 1.481234
```

The standard PCA is computed using princomp and the MCD PCA using PcaCov. By setting `scale = TRUE` we perform the correlation based PCA. We notice that the loadings are different implying different interpretations for the principal components.

```
pca_regular <- princomp(wood, cor = TRUE)
pca_mcd <- PcaCov(wood, scale = TRUE, cov.control = CovControlMcd(alpha = 0.5))

pca_regular$loadings[, ]
```

```
##        Comp.1      Comp.2     Comp.3      Comp.4      Comp.5
## X1  0.4675272  0.45529703 0.16766853  0.54001815  0.50437145
```

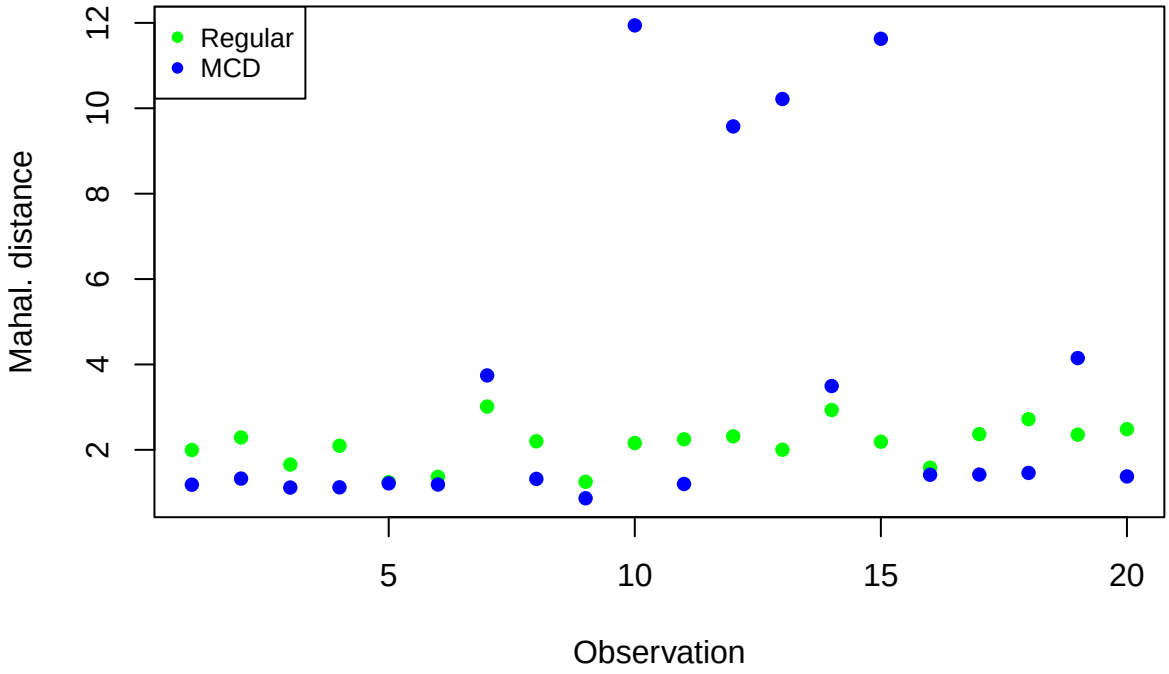

Figure 2: Regular and robust Mahalanobis distances for each observation of the `wood` data set.

```
## X2 -0.3856401  0.69640876 0.26393827  0.15298068 -0.52271409
## X3  0.4393226  0.42778998 0.04187736 -0.78798222  0.03635492
## X4  0.4357446 -0.35073854 0.72020727  0.07216355 -0.40398395
## X5 -0.4998222 -0.04120268 0.61787670 -0.24259793  0.55484656
```

```
pca_mcd@loadings
```

```
##            PC1        PC2         PC3        PC4         PC5
## X1  0.5621432 0.04329589 -0.55516799  0.1494159 -0.59294506
## X2  0.6027304 0.25366576 -0.08777107  0.2034274  0.72338319
## X3  0.4864497 0.19502242  0.66077586 -0.4693703 -0.26153331
## X4 -0.2230827 0.64331581 -0.41850979 -0.5960536  0.07712665
## X5 -0.1852343 0.69418310  0.26890437  0.6005862 -0.22535466
```

```
pairs(pca_regular$scores, pch = 16, upper.panel = NULL, col = color_outliers)
```

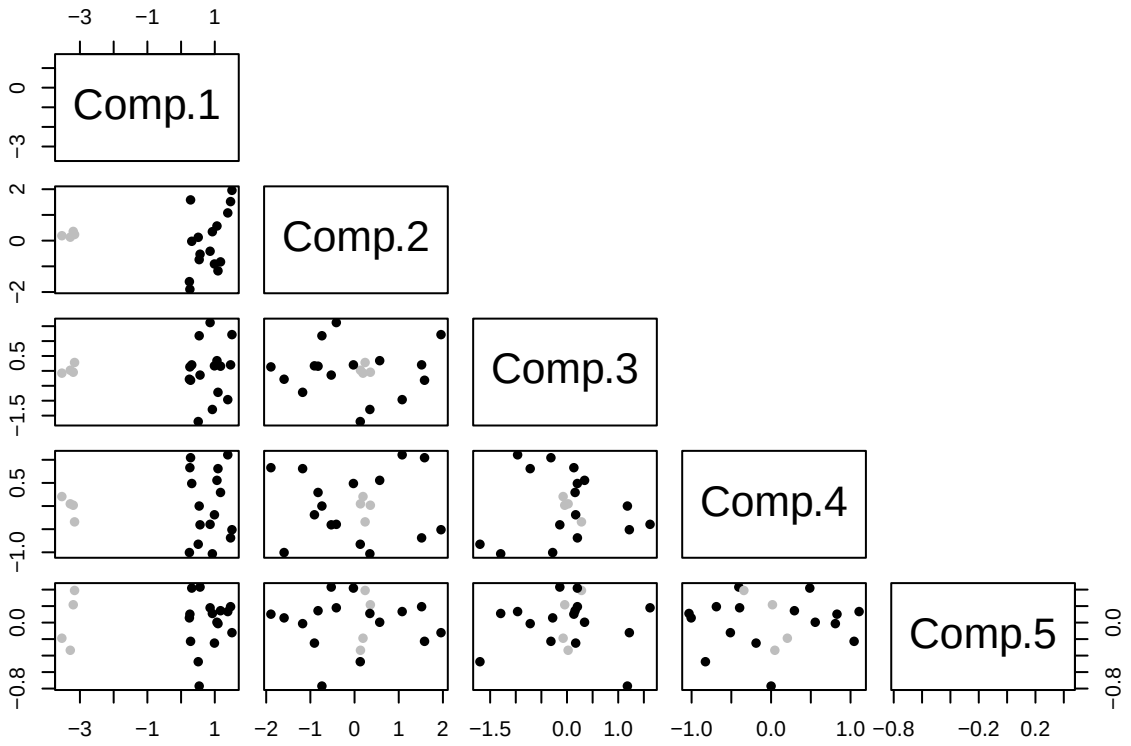

Figure 3: Pairwise scatter plots of scores for regular PCA.

```
pairs(pca_mcd@scores, pch = 16, upper.panel = NULL, col = color_outliers)
```

Problem 2: Estimators of scatter

```
library(mvtnorm)
set.seed(123)
n <- 200
```

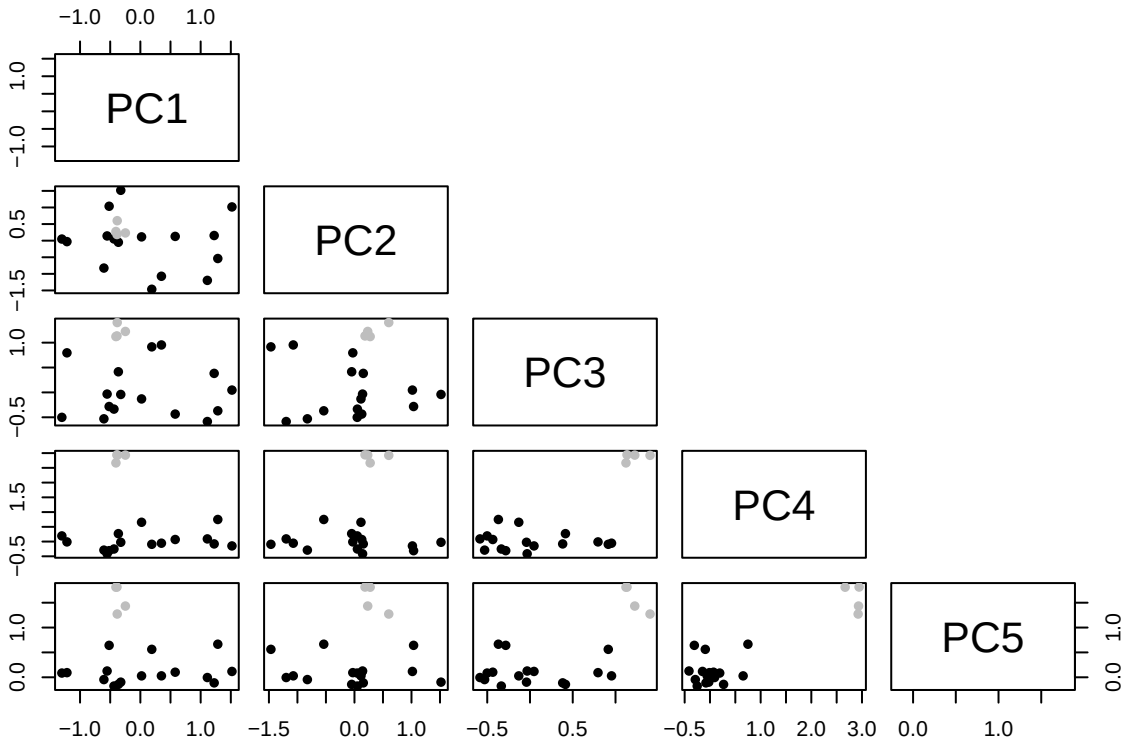

Figure 4: Pairwise scatter plots of scores for robust PCA.

a) Bivariate normal distribution

```
data1 <- rmvnorm(n, mean = c(0, 0), sigma = diag(2))
plot(data1, pch = 16)
```

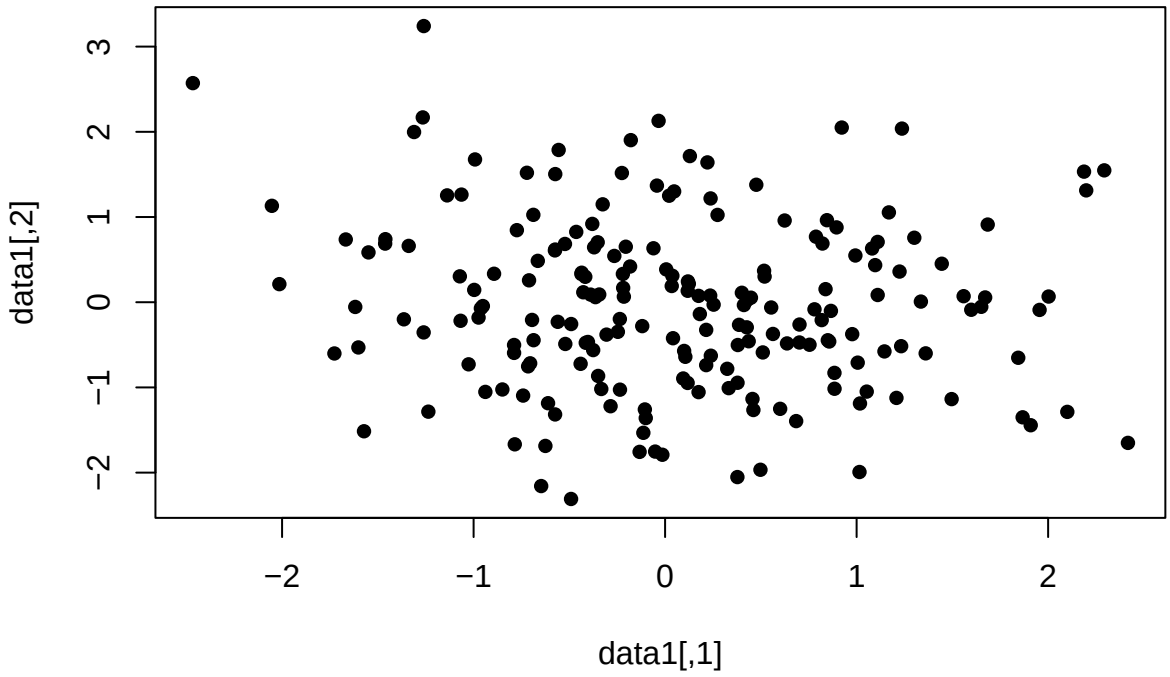

Figure 5: Scatter plot of a sample from a bivariate normal distribution.

```
cov(data1)
```

```
##            [,1]       [,2]
## [1,]  0.8861715 -0.1130047
## [2,] -0.1130047  0.9915206
```

```
CovMcd(data1, alpha = 0.5)@cov
```

```
##            [,1]       [,2]
## [1,]  0.9196207 -0.1551592
## [2,] -0.1551592  0.9580028
```

b) Bivariate t-distribution

```
data2 <- rmvt(n, df = 5, sigma = diag(2))
plot(data2, pch = 16)
```

```
cov(data2)
```

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

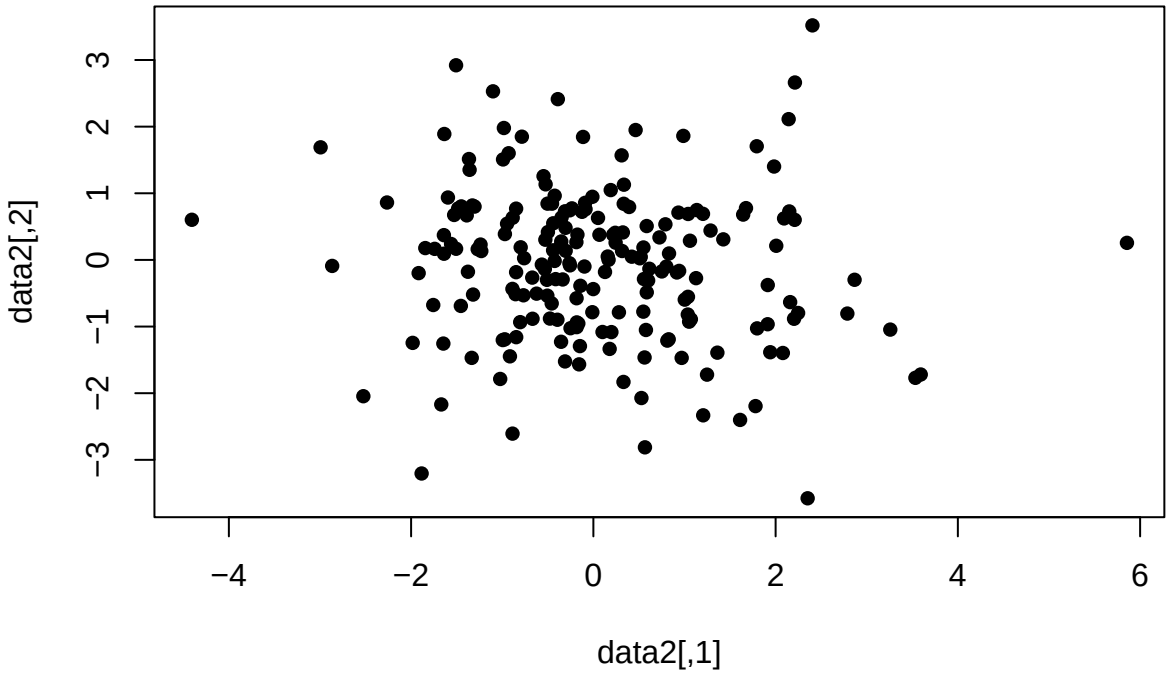

Figure 6: Scatter plot a of sample from a bivariate t-distribution.

```
## [1,]  1.7892456 -0.1774634
## [2,] -0.1774634  1.2873681
```

```
CovMcd(data2, alpha = 0.5)@cov
```

```
##            [,1]       [,2]
## [1,]  1.4039926 -0.2008512
## [2,] -0.2008512  1.2986795
```

c) Bivariate Weibull-Gamma

```
x1 <- rweibull(n, shape = 1, scale = 2)
x2 <- rgamma(n, shape = 2, scale = 1)
data3 <- cbind(x1, x2)
plot(data3, pch = 16)
```

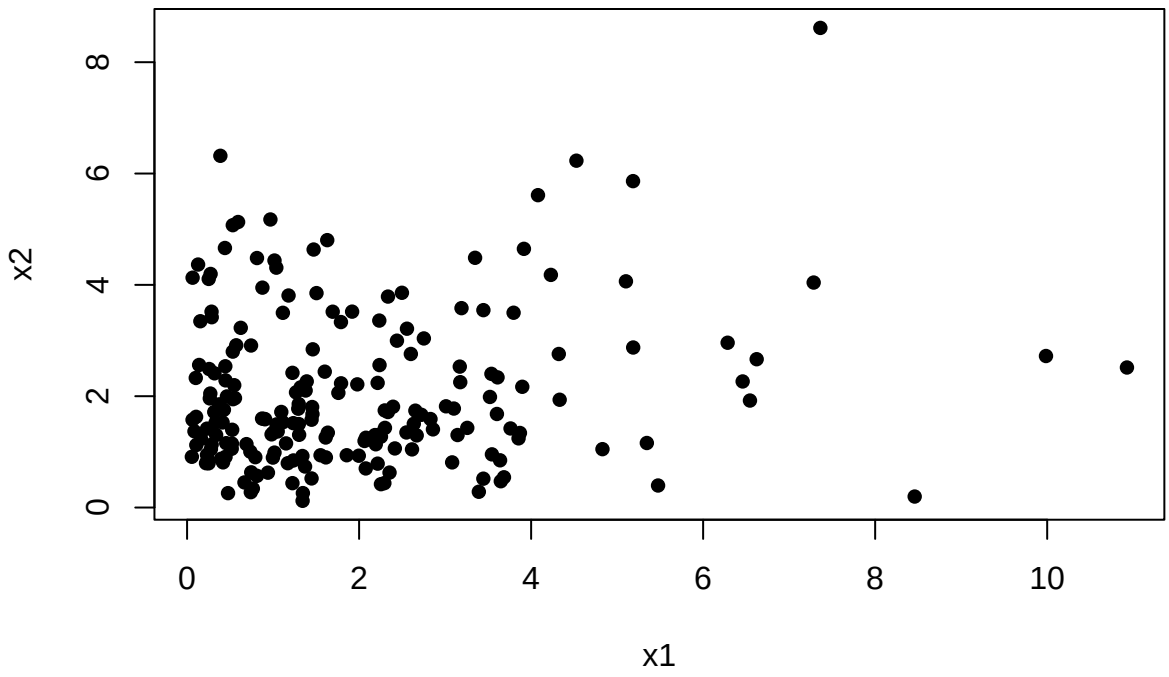

Figure 7: Scatter plot of a sample from a bivariate distribution where the first component follows the Weibull distribution and the second component follows the Gamma distribution.

```
cov(data3)
```

```
##           x1        x2
## x1 3.3691638 0.3967598
## x2 0.3967598 1.9398810
```

```
CovMcd(data3, alpha = 0.5)@cov
```

```
##            x1         x2
## x1 2.26313959 0.01216148
## x2 0.01216148 1.07616204
```

Hint for Homework 4

- Chapter 10.9 of the book "Introduction to Mathematical Statistics" by Hogg et al. has good discussion about influence function and breakdown point.