Multivariate Statistical Methods

Assignment 1

Allan Gholmi, Emma Wallentinsson, Rasmus Holm

2017-11-23

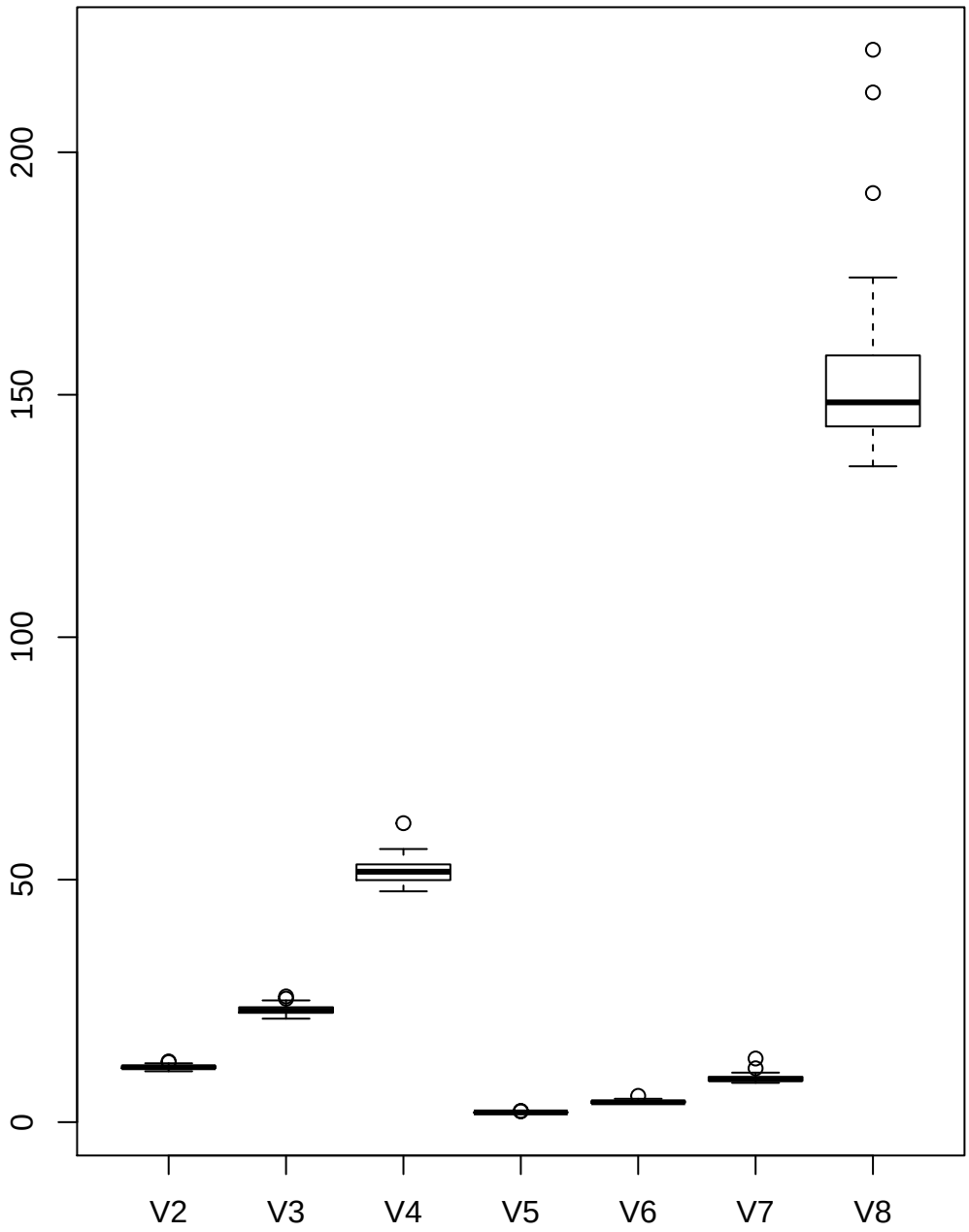
Question 1

a)

```
data <- read.table("../data/T1-9.dat")
numeric_data <- data[, -1]

summary(numeric_data)
#>        V2              V3              V4              V5
#>  Min.   :10.49   Min.   :21.34   Min.   :47.60   Min.   :1.890
#>  1st Qu.:11.12   1st Qu.:22.57   1st Qu.:49.97   1st Qu.:1.970
#>  Median :11.32   Median :22.98   Median :51.65   Median :2.005
#>  Mean   :11.36   Mean   :23.12   Mean   :51.99   Mean   :2.022
#>  3rd Qu.:11.57   3rd Qu.:23.61   3rd Qu.:53.12   3rd Qu.:2.070
#>  Max.   :12.52   Max.   :25.91   Max.   :61.65   Max.   :2.290
#>        V6              V7              V8
#>  Min.   :3.840   Min.   : 8.100   Min.   :135.2
#>  1st Qu.:4.003   1st Qu.: 8.543   1st Qu.:143.5
#>  Median :4.100   Median : 8.845   Median :148.4
#>  Mean   :4.189   Mean   : 9.081   Mean   :153.6
#>  3rd Qu.:4.338   3rd Qu.: 9.325   3rd Qu.:157.7
#>  Max.   :5.420   Max.   :13.120   Max.   :221.1
apply(numeric_data, 2, sd)
#>          V2          V3          V4          V5          V6          V7
#>  0.39410116  0.92902547  2.59720188  0.08687304  0.27236502  0.81532689
#>          V8
#> 16.43989508
```

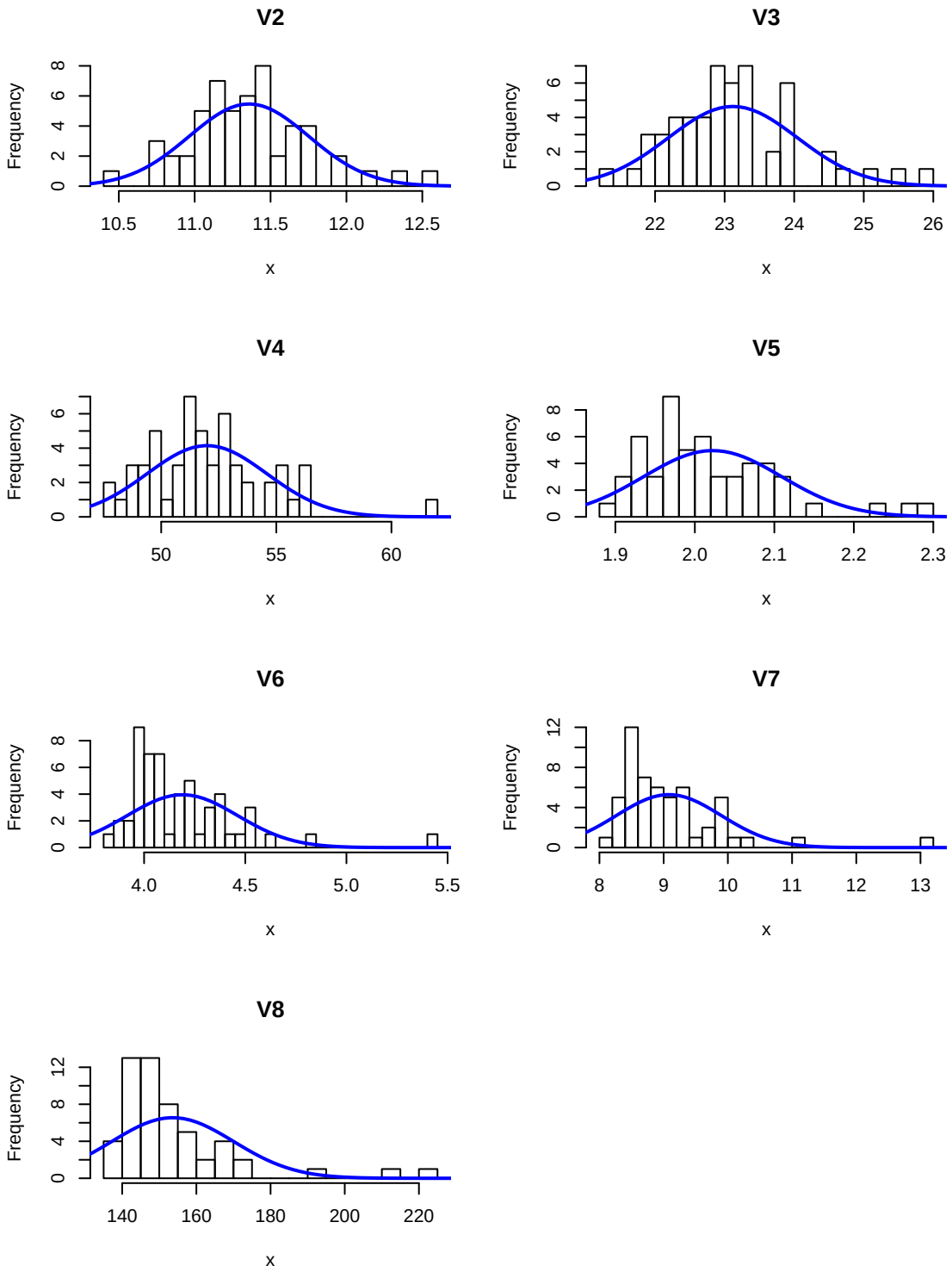
b)

```
boxplot(numeric_data)
```




```
old <- par(mfrow=c(4, 2))
for (col in names(numeric_data)) {
    x <- numeric_data[, col]
    h <- hist(x, breaks=25, main=col)
    offset <- (max(x) - min(x)) / 2
```

```
    xfit <- seq(min(x) - offset, max(x) + offset, length = 100)
    yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
    yfit <- yfit * diff(h$mids[1:2]) * length(x)
    lines(xfit, yfit, col="blue", lwd=2)
}
par(old)
```


V2
Frequency
0 2 4 6 8
10.5 11.0 11.5 12.0 12.5
x
V3
Frequency
0 2 4 6
22 23 24 25 26
x
V4
Frequency
0 2 4 6
50 55 60
x
V5
Frequency
0 2 4 6 8
1.9 2.0 2.1 2.2 2.3
x
V6
Frequency
0 2 4 6 8
4.0 4.5 5.0 5.5
x
V7
Frequency
0 4 8 12
8 9 10 11 12 13
x
V8
Frequency
0 4 8 12
140 160 180 200 220
x

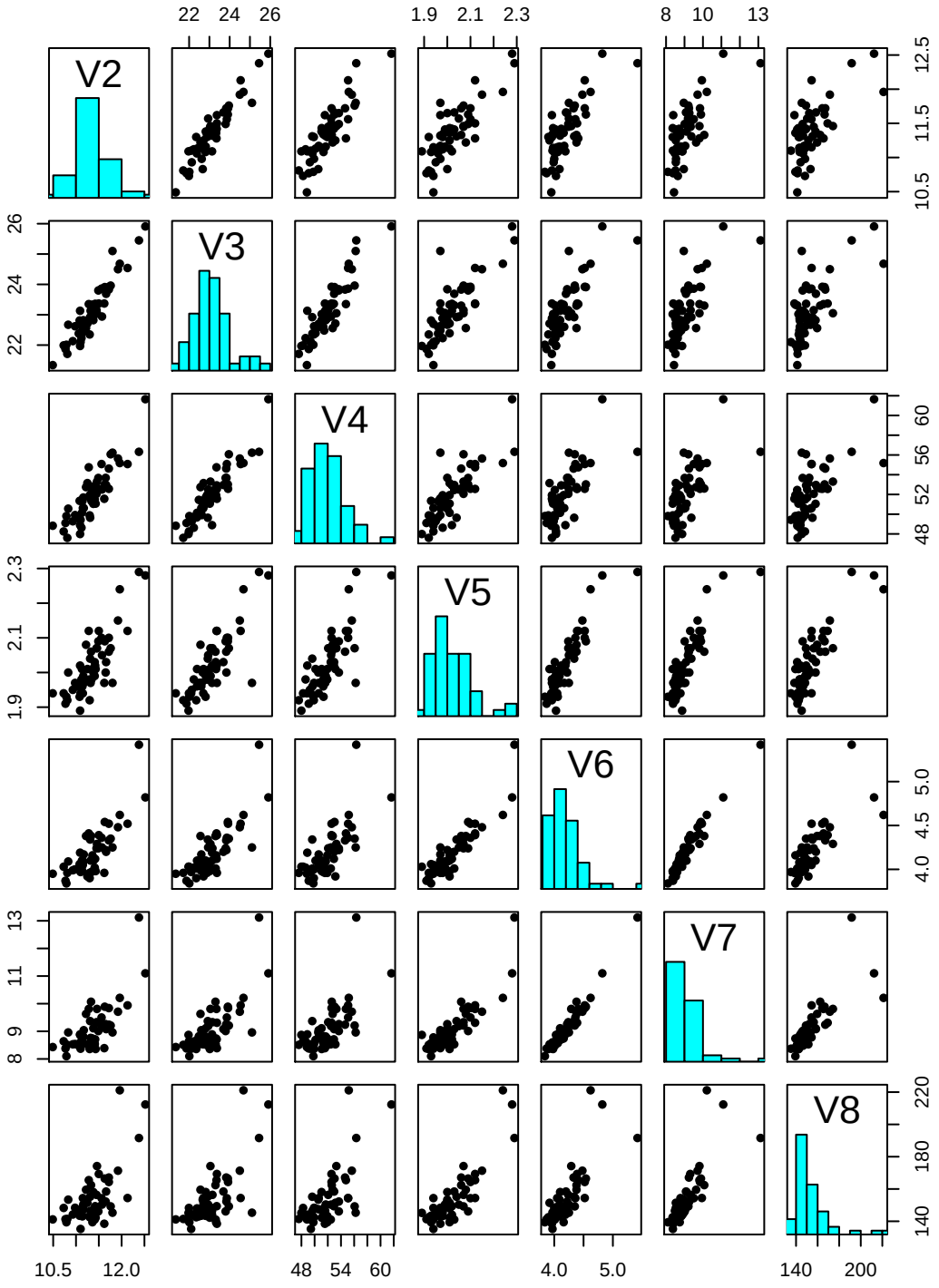
Question 2

a)

```
covariance_mat <- cov(numeric_data)
correlation_mat <- cor(numeric_data)
```

b)

```
panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y / max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

pairs(numeric_data, pch=16, diag.panel=panel.hist)
```


V2
V3
V4
V5
V6
V7
V8
22 24 26
1.9 2.1 2.3
8 10 13
10.5 11.5 12.5
22 24 26
48 52 56 60
1.9 2.1 2.3
4.0 4.5 5.0
8 9 11 13
140 180 220
10.5 12.0
48 54 60
4.0 5.0
140 200

c)

Question 3

a)

b)

```
X <- as.matrix(numeric_data)

means <- colMeans(X)
X_central <- X - rep(1, nrow(X)) %*% t(means)

edist_sq <- X_central %*% t(X_central)
country_edist <- diag(edist_sq)

edist_extreme_countries <- data[order(country_edist, decreasing=TRUE), 1][1:5]
as.character(edist_extreme_countries)
#> [1] "PNG" "COK" "SAM" "BER" "GBR"
```

c)

```
V_inv <- diag(1 / apply(X, 2, var))
edist_central_sq <- X_central %*% V_inv %*% t(X_central)
country_central_edist <- diag(edist_central_sq)

central_edist_extreme_countries <- data[order(country_central_edist, decreasing=TRUE), 1][1:5]
as.character(central_edist_extreme_countries)
#> [1] "SAM" "COK" "PNG" "USA" "SIN"
```

d)

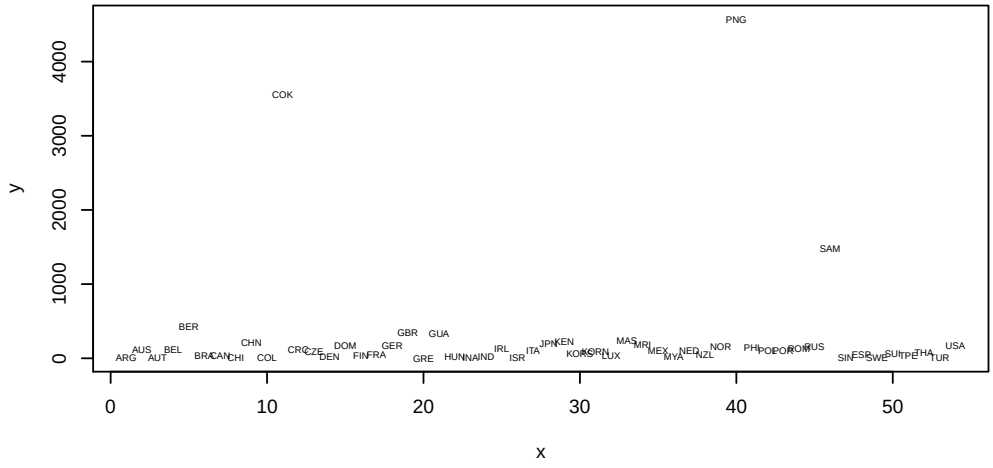
```
mdist_sq <- X_central %*% solve(covariance_mat) %*% t(X_central)
country_mdist <- diag(mdist_sq)

mdist_extreme_countries <- data[order(country_mdist, decreasing=TRUE), 1][1:5]
as.character(mdist_extreme_countries)
#> [1] "SAM"  "PNG"  "KORN" "COK"  "MEX"
```

e)

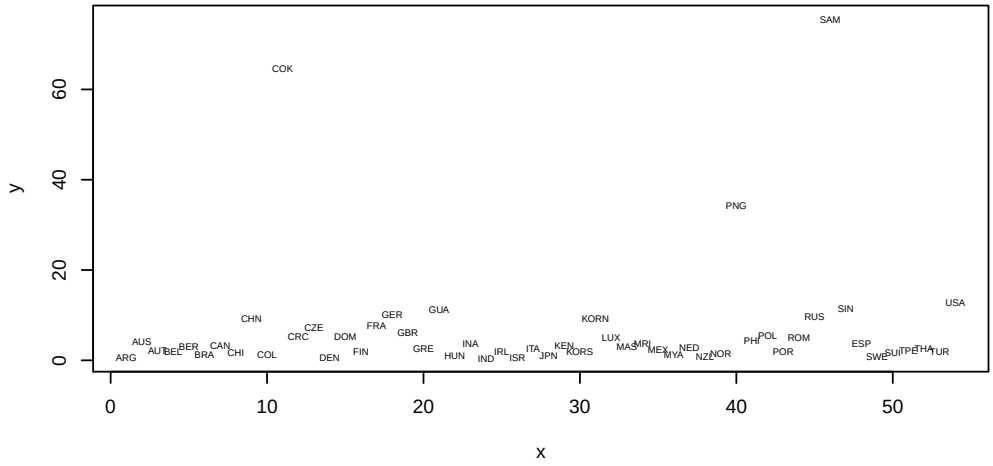
```
countries <- as.character(data$V1)

x <- 1:length(countries)

old <- par(mfrow=c(3, 1))

y <- country_edist
```

```r
plot(x, y, main="Squared Euclidean Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)

y <- country_central_edist

plot(x, y, main="Squared Central Euclidean Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)

y <- country_mdist

plot(x, y, main="Mahalanobis Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)
```

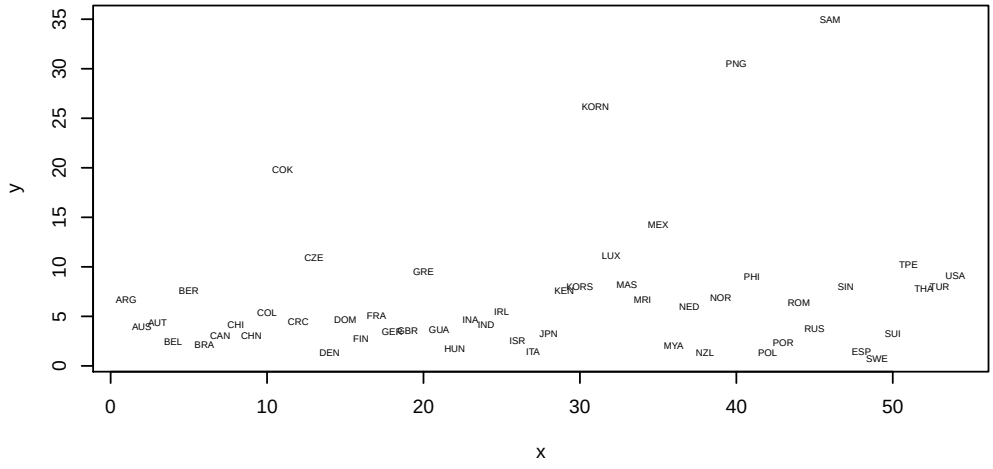
Squared Euclidean Distance



Squared Central Euclidean Distance



Mahalanobis Distance



```
par(old)
```