Machine Intelligence II: Sheet 2 (PCA)

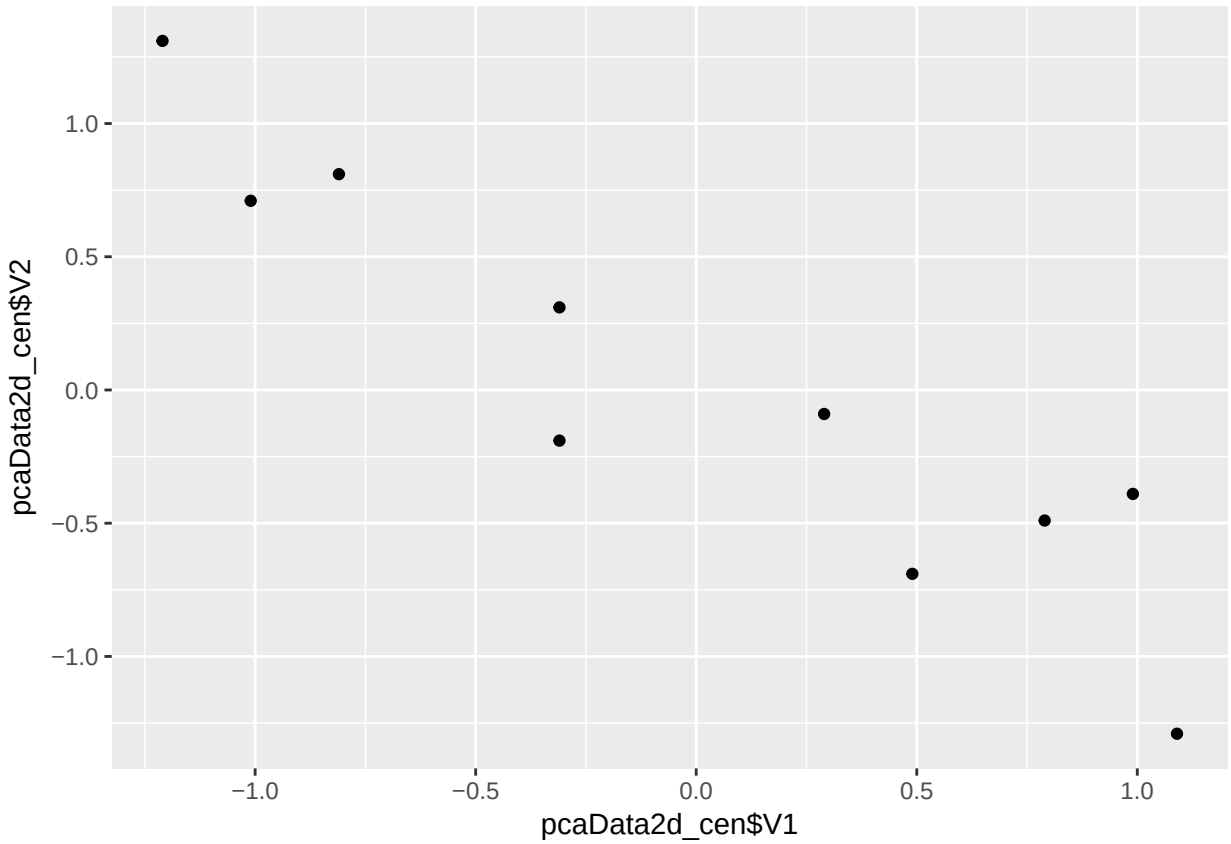
NoNames2

03.05.2017

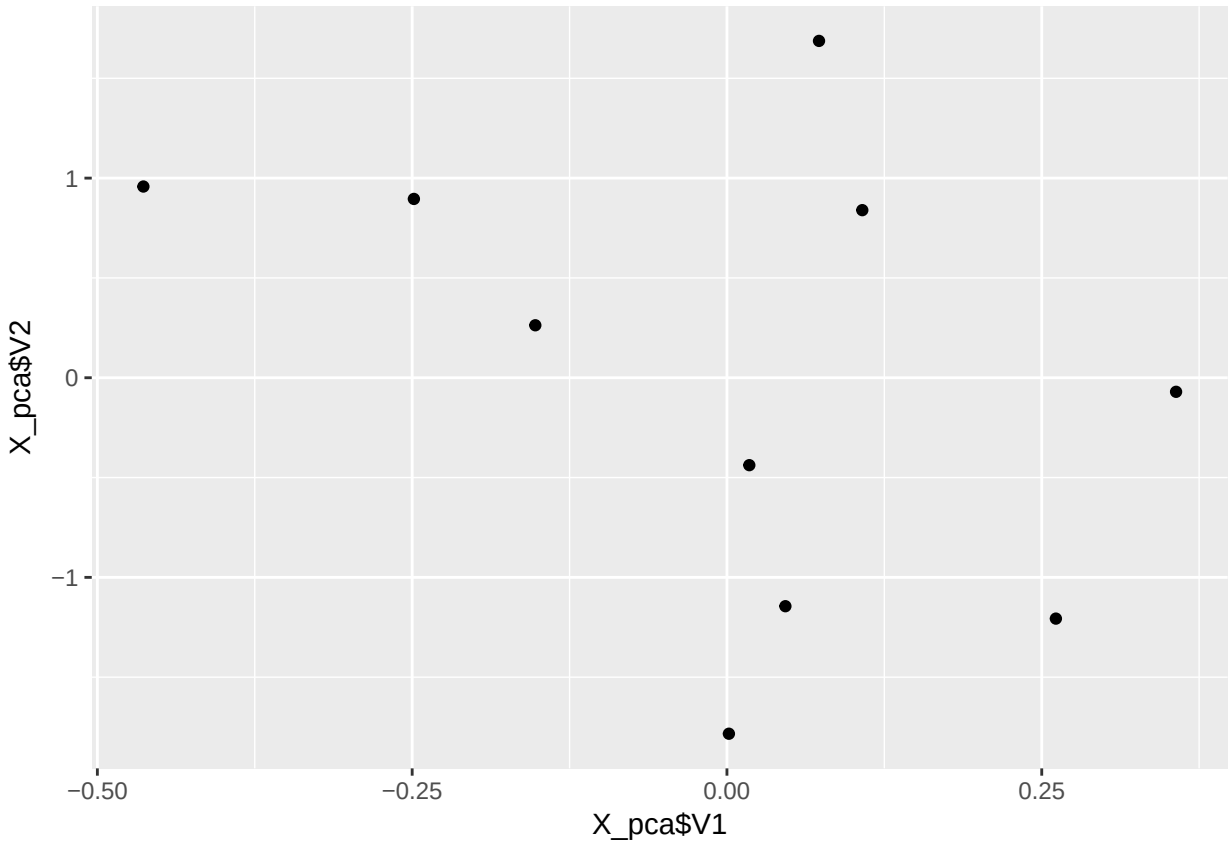
Machine Intelligence 2 Homework 1

```
# load packages
#install.packages("imager")
library(jpeg)

#install.packages("foreign")
library(foreign)

#install.packages("gridExtra")
library(gridExtra)

#install.packages("ggplot2")
library(ggplot2)

#install.packages("reshape2")
library(reshape2)

#install.packages("colorRamps")
library(colorRamps)

#setwd("~/Desktop/Uni/Statistik Master/courses/machine intelligence 2/data/natIMG.jpeg")
```

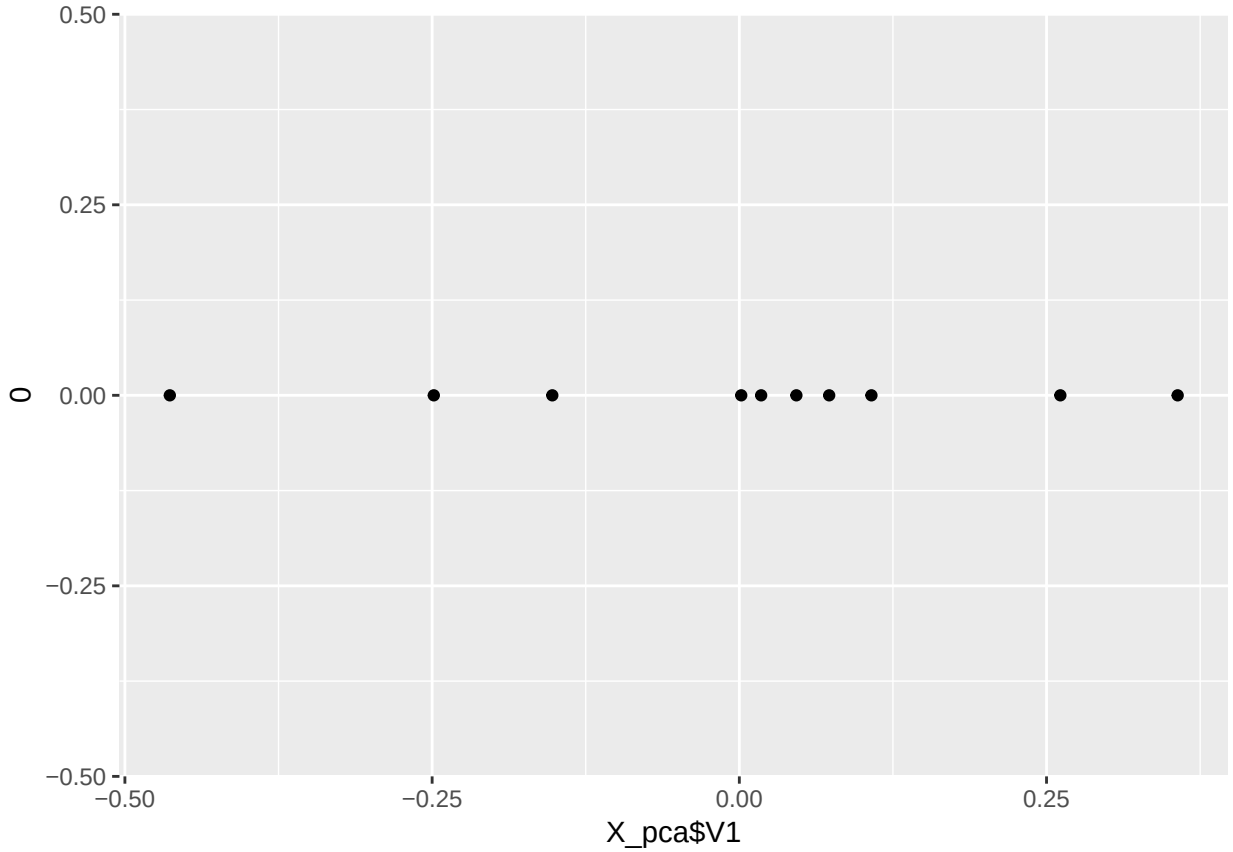
```
# 2.1 a)

# load data
pcaData2d <- read.csv("Ex2/pca-data-2d.txt", sep = "", header = FALSE)

# calcualte mean matrix
n <- nrow(pcaData2d)
M_mean <- matrix(data=1, nrow=n) %*% cbind(mean(pcaData2d[,1]),
                                           mean(pcaData2d[,2]))

# subtract mean from data set -> "Difference Matrix"
pcaData2d_cen <- pcaData2d - M_mean

# plot centered matrix
ggplot(pcaData2d_cen, aes(x=pcaData2d_cen$V1, y=pcaData2d_cen$V2)) + geom_point()
```



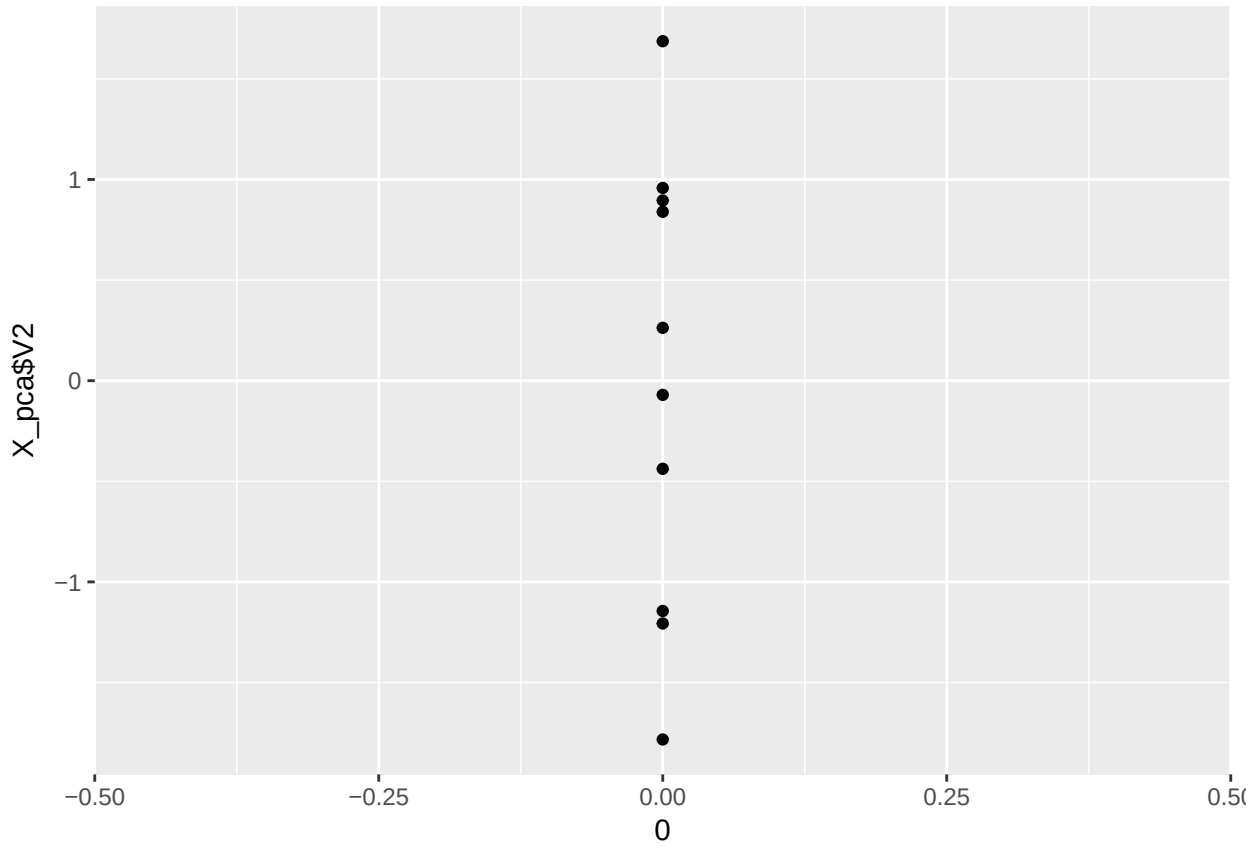

```r
# 2.1 b)

# convert df to matrix
D <- as.matrix(pcaData2d_cen)

# compute Covariance Matrix C
C <- (n-1)^-1 * t(D) %*% D

# Eigenvalues of C, Eigenvectors in P
ev <- eigen(C)
P  <- ev$vectors

# transform D
D_tr <- P %*% t(D)

# in ggplot2 plottable data frame
X_pca <- as.data.frame(t(D_tr))

# plot of the transformed set
ggplot(X_pca, aes(x=X_pca$V1, y=X_pca$V2)) + geom_point()
```



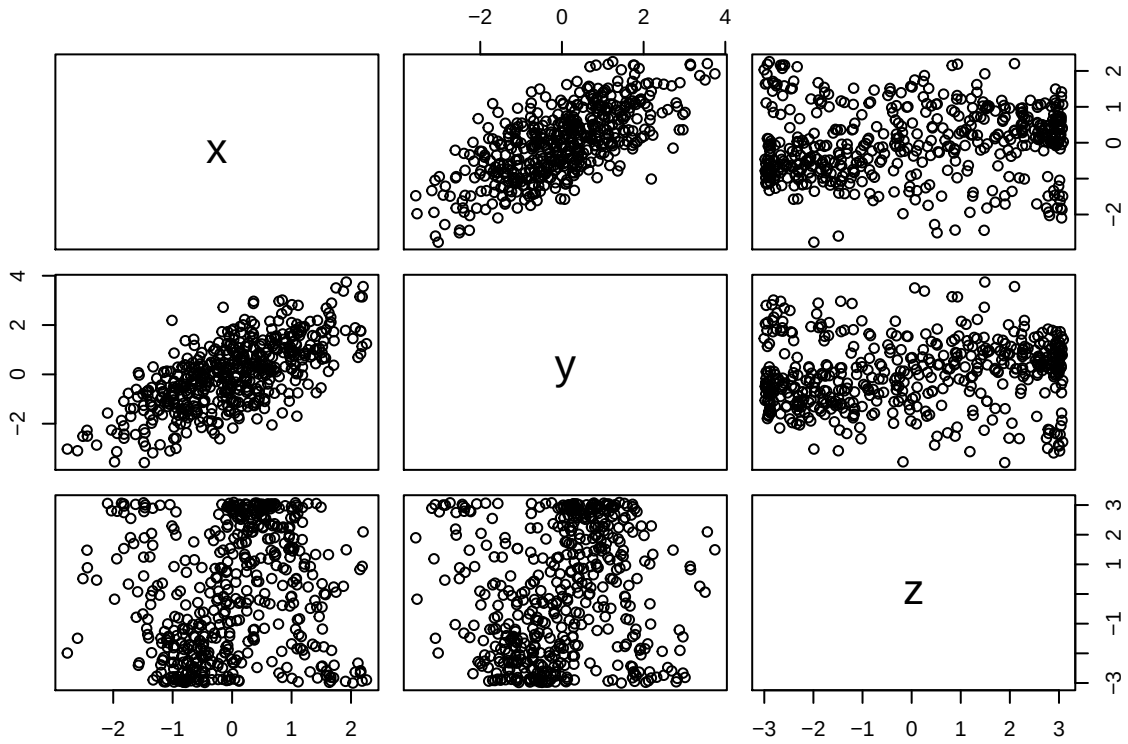

```
# 2.1 c)


# plot only PC1
ggplot(X_pca, aes(x=X_pca$V1, y=0)) + geom_point()
```



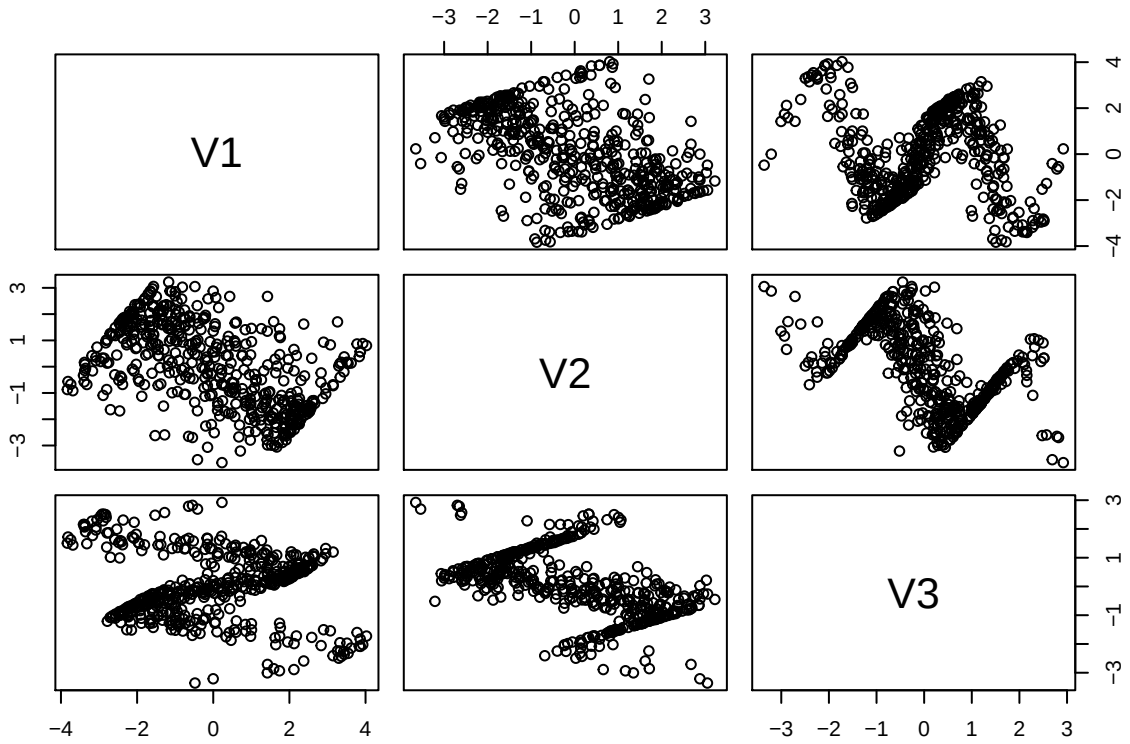

```
# plot only PC2
ggplot(X_pca, aes(x=0, y= X_pca$V2)) + geom_point()
```



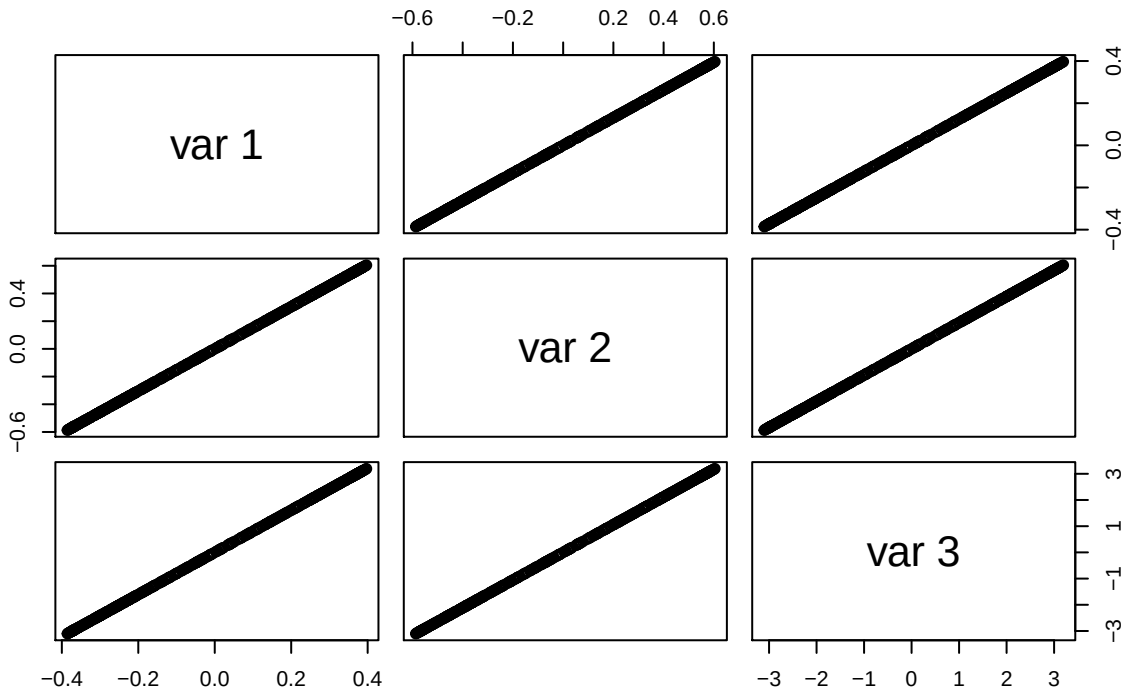

```
# 2.2 a)


# load data set
pcaData3d <- read.csv("Ex2/pca-data-3d.txt")

# centering
n_3 <- nrow(pcaData3d)
M3_mean <- matrix(data=1, nrow=n_3) %*% colMeans(pcaData3d)

# subtract mean from data set -> "Difference Matrix"
D3 <- pcaData3d - M3_mean

# scatterplot matrix
pairs(D3)
```



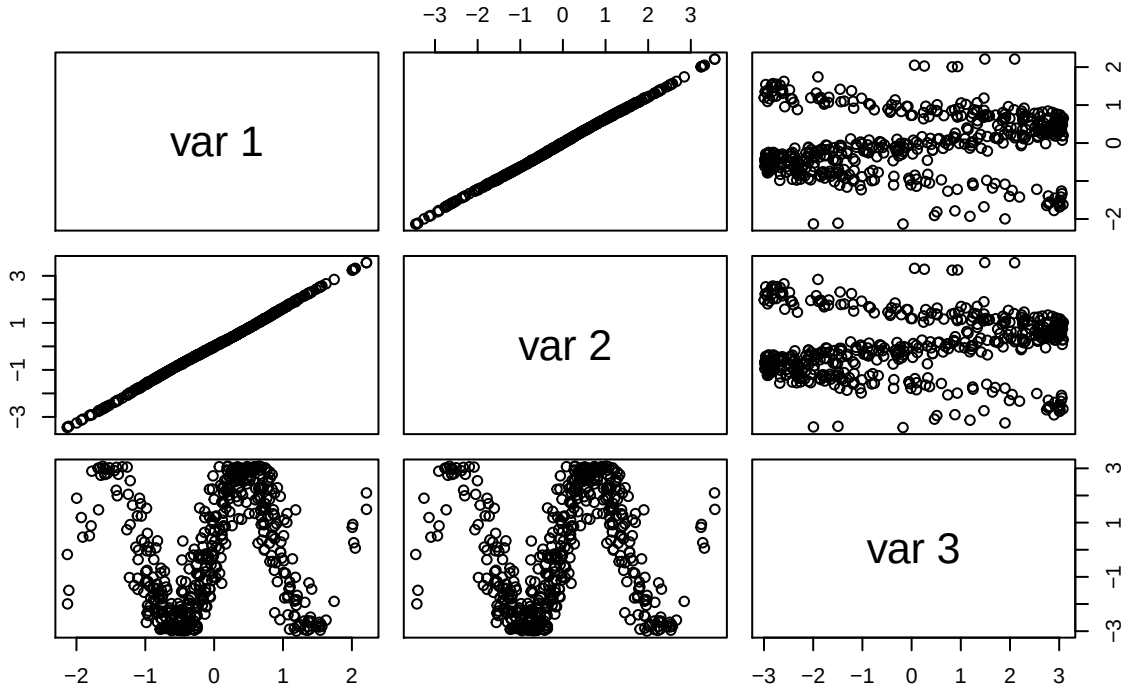

```r
# 2.2 b)

# convert df to matrix
D3 <- as.matrix(D3)

# compute Covariance Matrix C
C3 <- (n_3-1)^-1 * t(D3) %*% D3

# Eigenvalues of C3, Eigenvectors in P3
ev3 <- eigen(C3)
P3  <- ev3$vectors

# transform D3
D3 <- P3 %*% t(D3)
X_3pca <- as.data.frame(t(D3))
pairs(X_3pca)
```



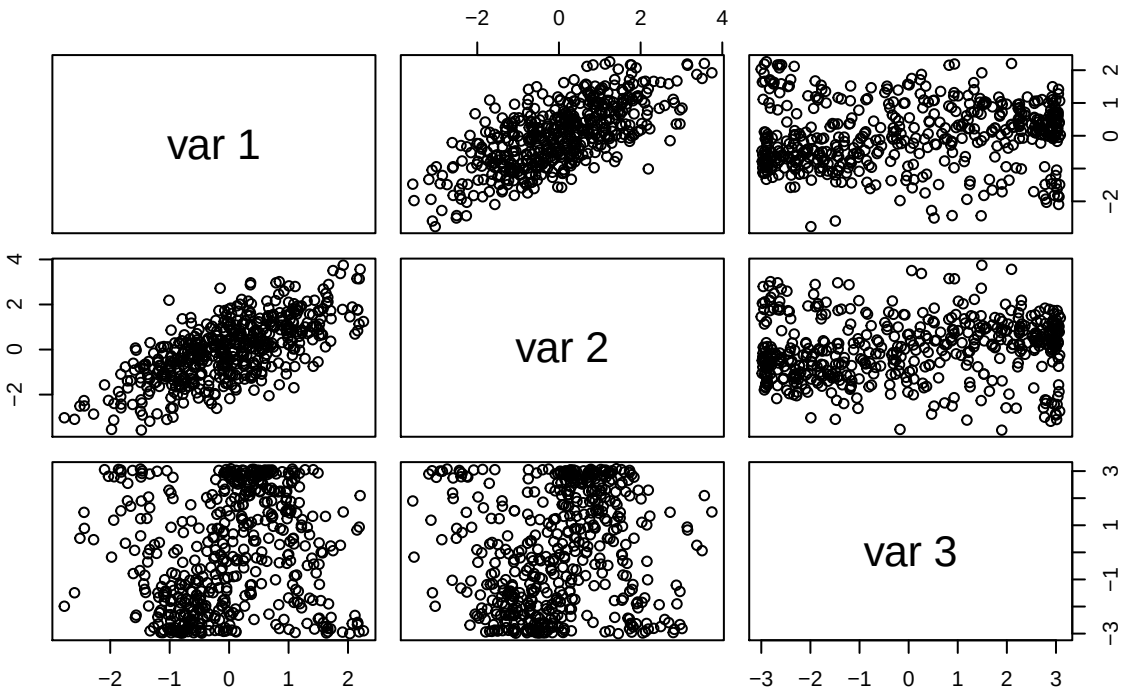

```r
# 2.2 c)


D3 <- pcaData3d - M3_mean

# convert df to matrix
D3 <- as.matrix(D3)

# compute Covariance Matrix C
C3 <- (n_3-1)^-1 * t(D3) %*% D3

eigen_vals <- eigen(C3)$values
eigen_vecs <- eigen(C3)$vectors

# compute pc scores
#pc_scores <- prcomp(D3)

pc_scores_1 <- D3 %*% eigen_vecs[, 1]
pc_scores_1_2 <- D3 %*% eigen_vecs[, 1:2]
pc_scores_1_2_3 <- D3 %*% eigen_vecs[, 1:3]


pca_reconstruction_1 <- pc_scores_1 %*% t(eigen_vecs[, 1])
pairs(pca_reconstruction_1, main = "PCs for reconstruction: 1")
```




```
pca_reconstruction_1_2 <- pc_scores_1_2 %*% t(eigen_vecs[, 1:2])
pairs(pca_reconstruction_1_2, main = "PCs for reconstruction: 1, 2")
```




```
pca_reconstruction_1_2_3 <- pc_scores_1_2_3 %*% t(eigen_vecs[, 1:3])
pairs(pca_reconstruction_1_2_3, main = "PCs for reconstruction: 1, 2, 3")
```

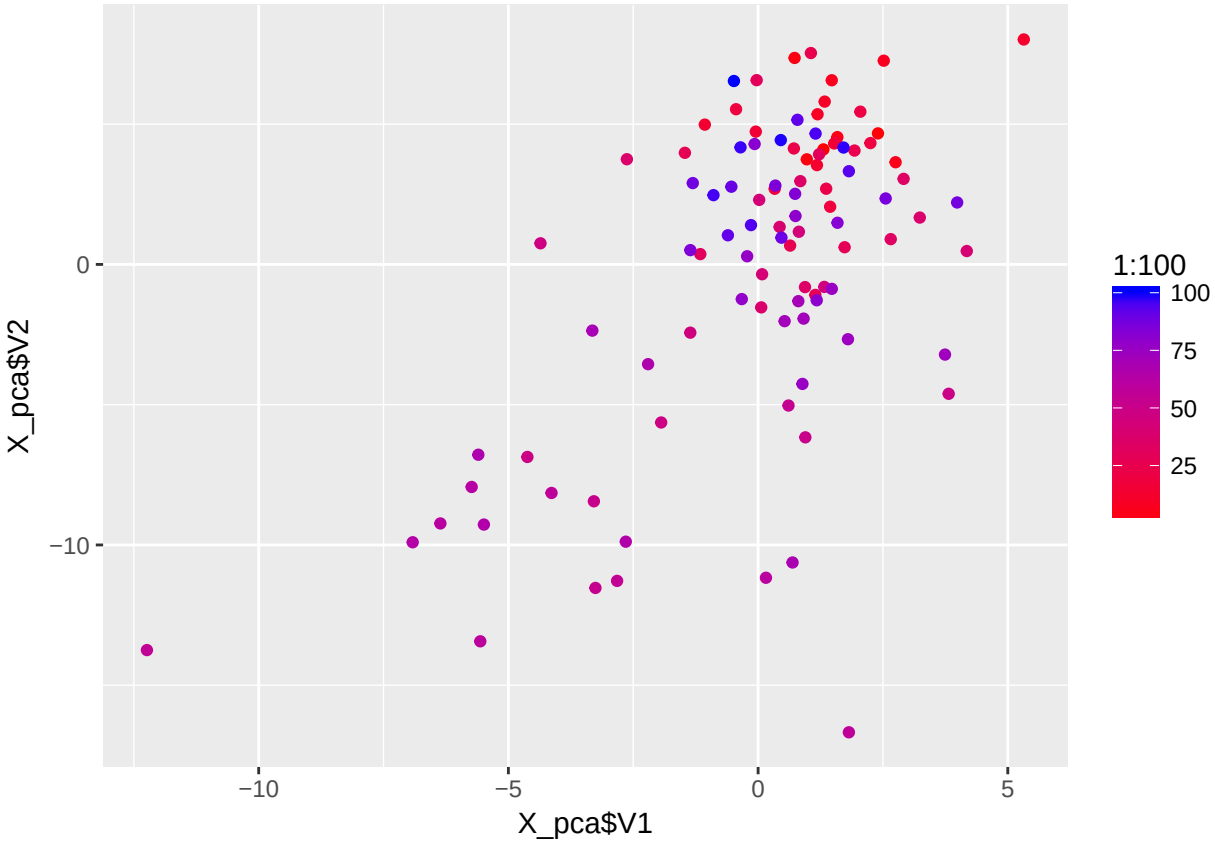



```
# for reconstruction of uncentered data, add mean veactor to each column (add M3)
```

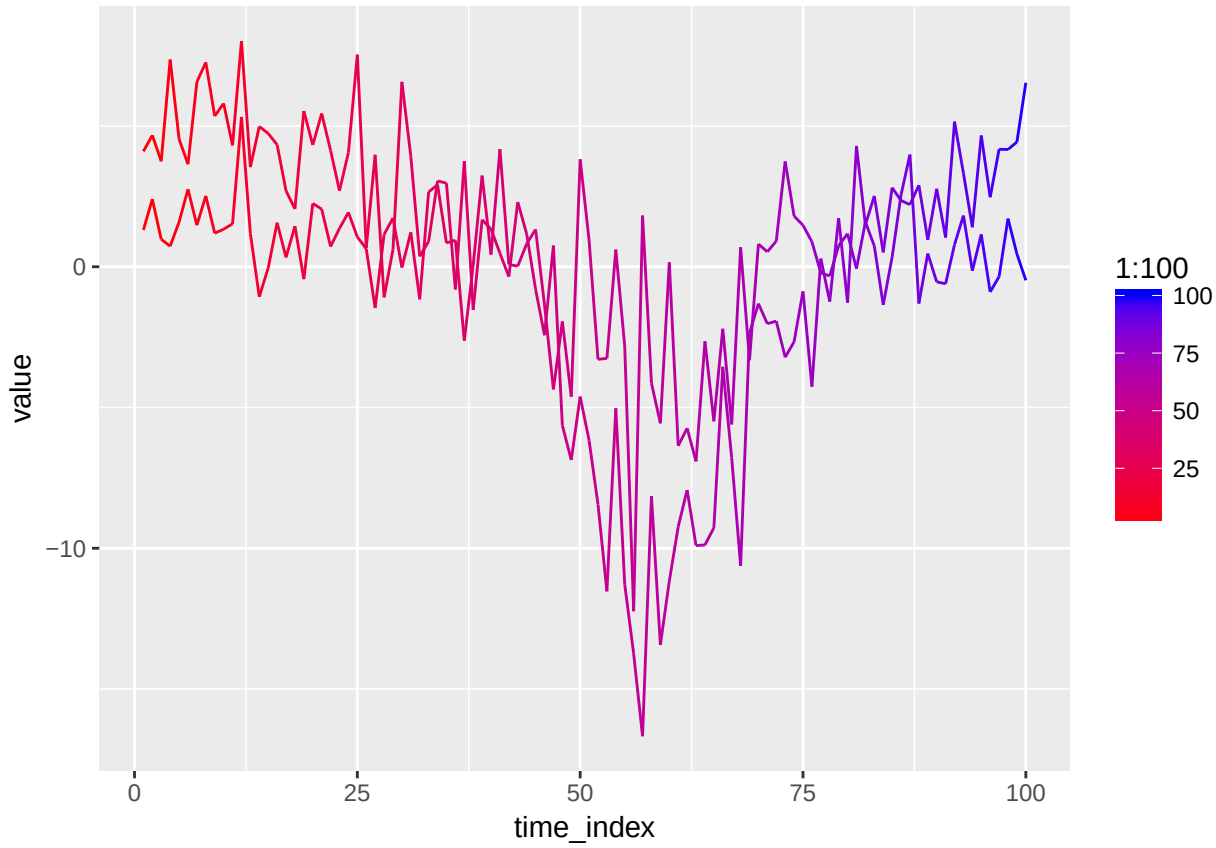
Interpretation Obviously, the first two PCs are useful here. Plotting only the first PC alone does not yield a lot of information (only about the spread in this direction). When taking into account the first two PCs, one can see the (possibly) sinusoidal trend in the data. Using all three PCs yields the original centered COV matrix that was plotted in 2.2 a). This is oubvious because due to the orthonormality of the eigenvectors eigen_vecs %*% t(eigen_vecs) will yield an identity matrix, hence, when multiplying this matrix with the centered COV matrix D3, the COV matrix remains unchanged. However, one could argue that including the third PC adds too much noise to the data and the trends are harder to detect in the scatter plot matrix. Therefore, the first two PCs should be chosen. Looking at the eigenvalues supports this.

```
# 2.3 a)

# load data
expDat <- read.csv("Ex2/expDat.txt", sep = ",", header = TRUE)

# calculate mean matrix
n <- nrow(expDat[2:ncol(expDat)])
M_mean <- matrix(data=1, nrow=n) %*% colMeans(expDat[2:ncol(expDat)])
# subtract mean from data set
expDat_cen <- expDat[2:ncol(expDat)] - M_mean

# convert df to matrix
D <- as.matrix(expDat_cen)
# compute Covariance Matrix C
C <- (n-1)^-1 * t(D) %*% D
# Eigenvalues of C, Eigenvectors in P
ev <- eigen(C)
P  <- ev$vectors
```

```r
# 2.3 b)

# transform D
D_tr <- P %*% t(D)
# in ggplot2 plottable data frame
X_pca <- as.data.frame(t(D_tr))

#scatter plot
ggplot(X_pca, aes(x=X_pca$V1, y=X_pca$V2)) + scale_colour_gradient(low = "red", high = "blue") + geom_p
```




```r
#line plot
df <- data.frame(expDat[,1], X_pca$V1, X_pca$V2)
colnames(df)[1] <- "time_index"

ggplot(df, aes(x=time_index, y=value)) +
    scale_colour_gradient(low = "red", high = "blue") +
    geom_line(aes(y=X_pca$V1, colour=1:100)) +
    geom_line(aes(y=X_pca$V2,colour=1:100))
```




```
# 2.3 c)

newDat <- apply(expDat[2:21], 2, sample)
```

```
# 2.3 d)

# shuffled dataset
# calculate mean matrix of new data set
n_new <- nrow(newDat)
M_mean_new <- matrix(data=1, nrow=n_new) %*% colMeans(newDat)
# subtract new mean from shuffled data set
Dat_cen_new <- newDat - M_mean_new
# compute Covariance Matrix C_new
C_new <- (n_new-1)^-1 * t(Dat_cen_new) %*% Dat_cen_new
```

2.3 e)

Shuffeling the data rowwise in the same sequence for all columns does not affect the resulting covariance matrix. This is easy to see without programming, since both, variances and covariances are based on element-wise sums of squared errors. For summing over a number of elements, the order is not relevant. Therefore, the variances remain unchanged even if the data was not shuffeled in the same sequence since they do not depend on the values of other variables. However, for the covariances this is important because otherwise the relationship between the features would change. But here, since the rows are shuffeled in the same way for all columns (i.e. features), the relationships (covariances) between the features remain the same as well.

** helper function for patches**

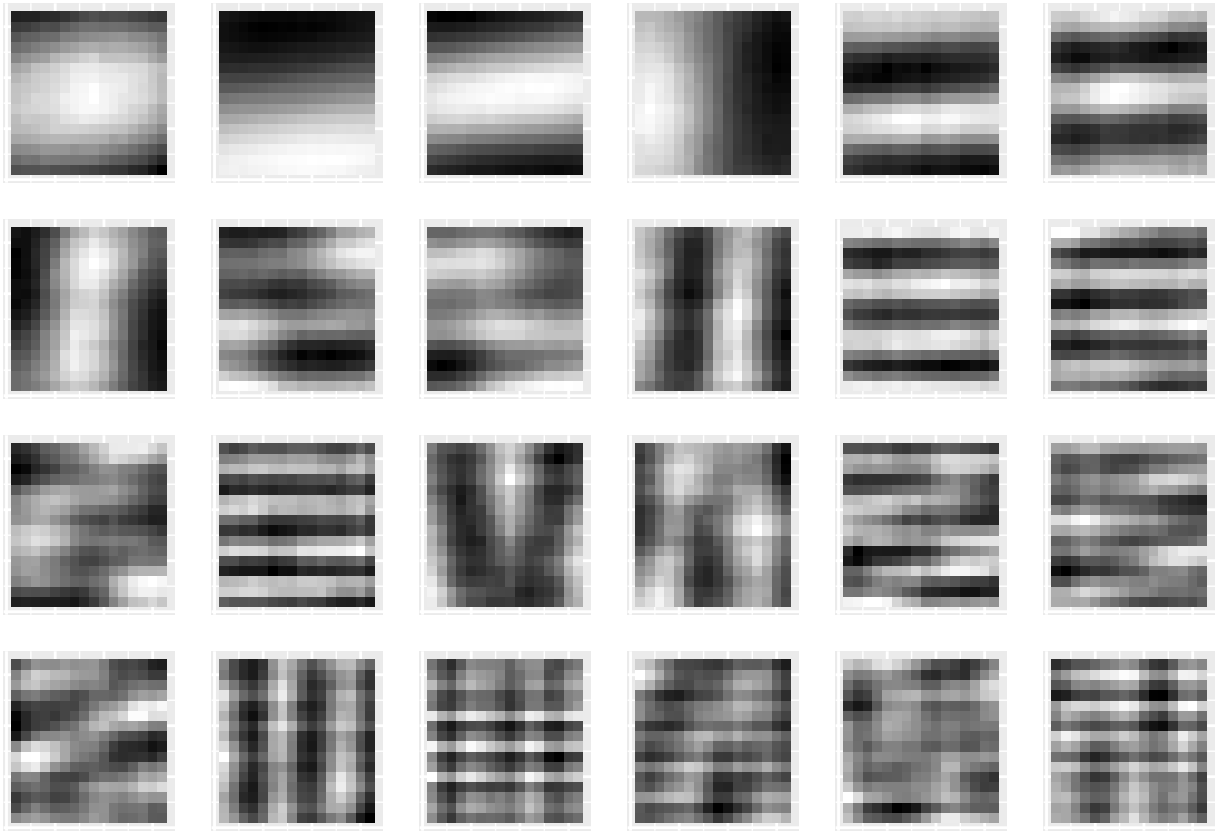
```r
get_patch <- function(matrix, h, w, return_vector = FALSE){

    h <- h-1
    w <- w-1

    n <- nrow(matrix)
    p <- ncol(matrix)

    h_sample <- sample(1:n, 1, FALSE)
    w_sample <- sample(1:p, 1, FALSE)

    if((h_sample + h) > n) {
        h_patch <- (h_sample - h):h_sample
    }else{
        h_patch <- h_sample:(h_sample + h)
    }

    if((w_sample + w) > p) {
        w_patch <- (w_sample - w):w_sample
    }else{
        w_patch <- w_sample:(w_sample + w)
    }

    if(return_vector == FALSE){
        return(matrix[h_patch, w_patch])
    }else{
        return(as.vector(matrix[h_patch, w_patch]))
    }
}
```

```r
# 2.4 a) buildings

path <- "/Users/Niko/Desktop/Uni/Statistik Master/courses/machine intelligence 2/mi2_homework/mi2_homew

f <- file.path(path, c("b1.jpg","b2.jpg", "b3.jpg", "b4.jpg", "b5.jpg", "b6.jpg", "b7.jpg", "b8.jpg", "

d <- lapply(f, readJPEG)


total_matrix_b <- matrix(nrow = 0, ncol = 256)

for(i in 1:length(d)){
    #set.seed(123)
    tmp1 <- t(replicate(500, get_patch(d[[i]], 16, 16, return_vector = TRUE), simplify = "vector"))
    #tmp2 <- replicate(10, get_patch(d[[i]], 16, 16, return_vector = TRUE))
    total_matrix_b <- rbind(total_matrix_b, tmp1)
}
```

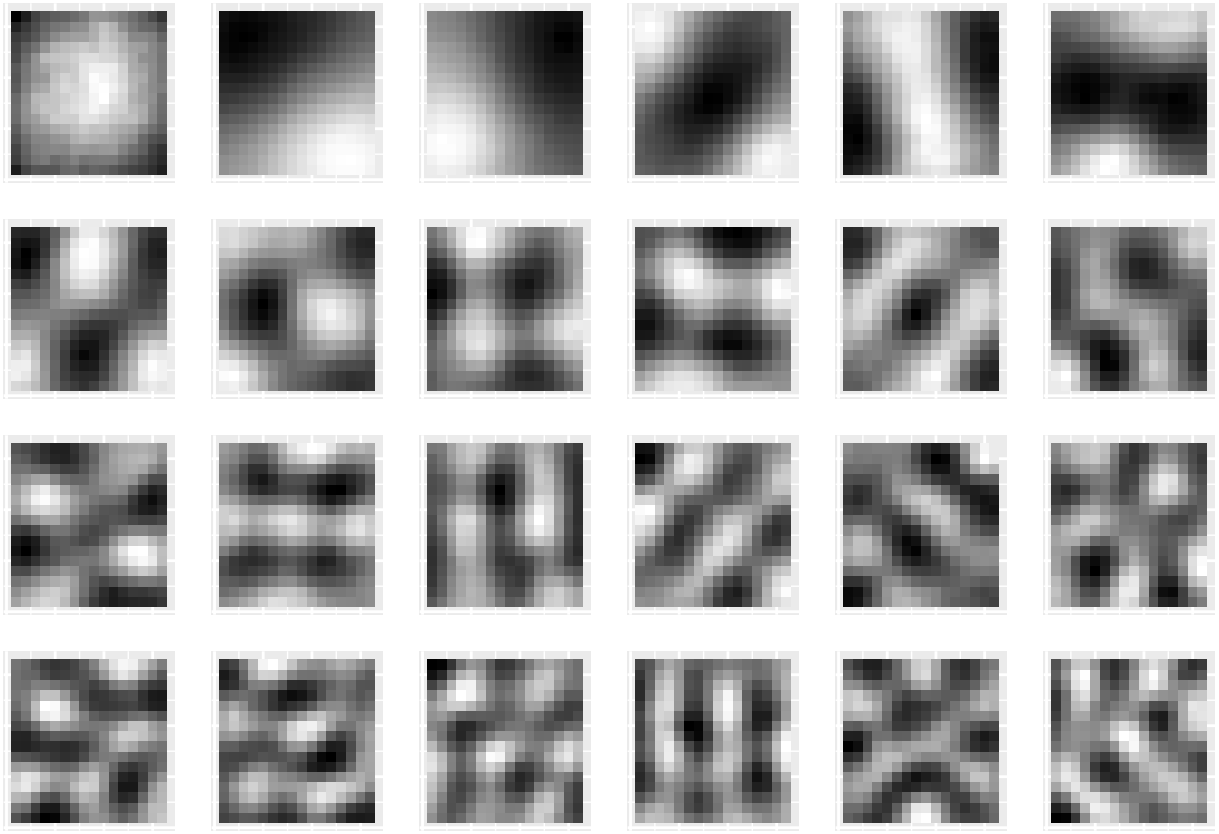
```r
# 2.4. a) Nature

path <- "/Users/Niko/Desktop/Uni/Statistik Master/courses/machine intelligence 2/mi2_homework/mi2_homew

f <- file.path(path, c("n1.jpg","n2.jpg", "n3.jpg", "n4.jpg", "n5.jpg", "n6.jpg", "n7.jpg", "n8.jpg", "
```

```r
nat <- lapply(f, readJPEG)


total_matrix_n <- matrix(nrow = 0, ncol = 256)

for(i in 1:length(nat)){
    #set.seed(123)
    tmp1 <- t(replicate(500, get_patch(nat[[i]], 16, 16, return_vector = TRUE), simplify = "vector"))
    #tmp2 <- replicate(10, get_patch(nat[[i]], 16, 16, return_vector = TRUE))
    total_matrix_n <- rbind(total_matrix_n, tmp1)
}
```

```r
# 2.4 b)

total_matrix_n_centered <- scale(total_matrix_n, center = TRUE, scale = FALSE)
total_matrix_b_centered <- scale(total_matrix_b, center = TRUE, scale = FALSE)

pcs_b <- eigen(cov(total_matrix_b_centered))$vectors
pcs_n <- eigen(cov(total_matrix_n_centered))$vectors


amount_pca <- 24
pca_patchesn <- pca_patchesb <- vector("list", length = amount_pca)

for(i in 1:24){

    pca_patchesn[[i]] <- melt(matrix(pcs_n[, i], 16, 16))
    pca_patchesb[[i]] <- melt(matrix(pcs_b[, i], 16, 16))

}


heatmap_custom <- function(matrix){
    g1 <- ggplot(data = matrix, aes(x=Var1, y=Var2, fill=value)) +
        geom_tile() +
        scale_fill_continuous(low = "white", high = "black") +
        guides(fill = FALSE) +
        theme(axis.title.x=element_blank(),
              axis.text.x=element_blank(),
              axis.ticks.x=element_blank(),
              axis.title.y=element_blank(),
              axis.text.y=element_blank(),
              axis.ticks.y=element_blank())


    return(g1)
}

plotlist_b <- lapply(pca_patchesb, heatmap_custom)
plotlist_n <- lapply(pca_patchesn, heatmap_custom)

do.call("grid.arrange", c(plotlist_b, ncol=6))
```

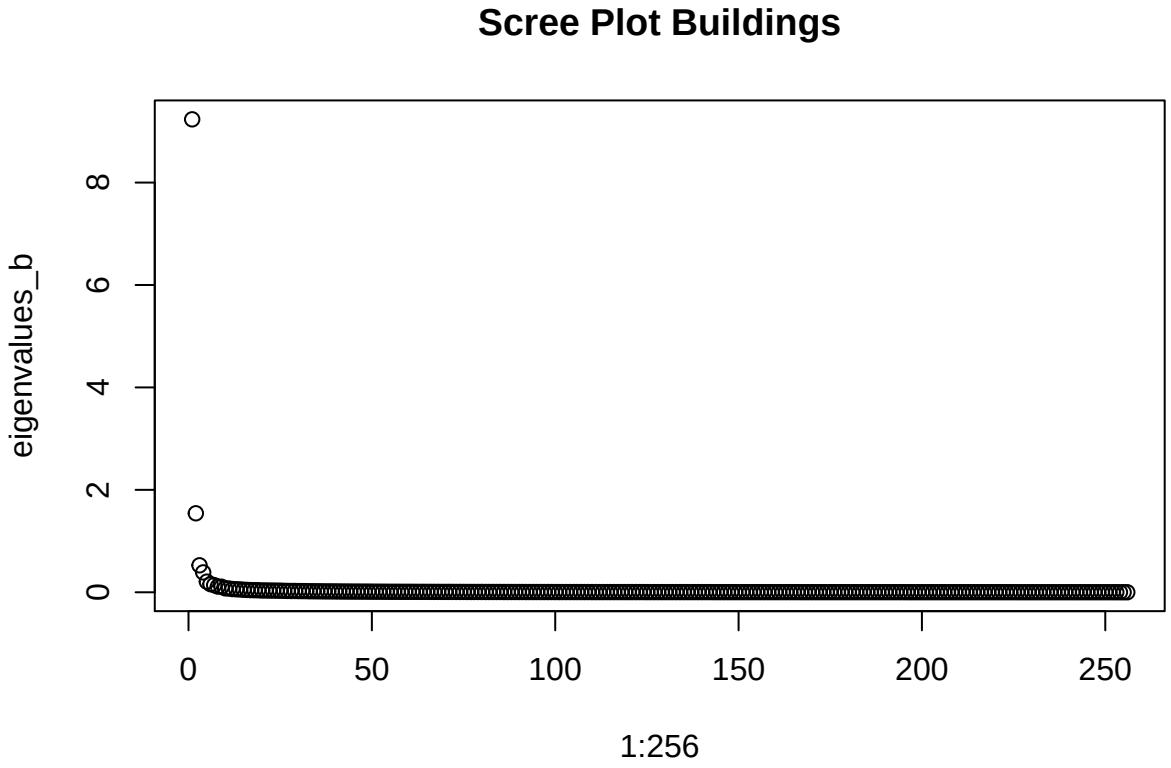


```
do.call("grid.arrange", c(plotlist_n, ncol=6))
```

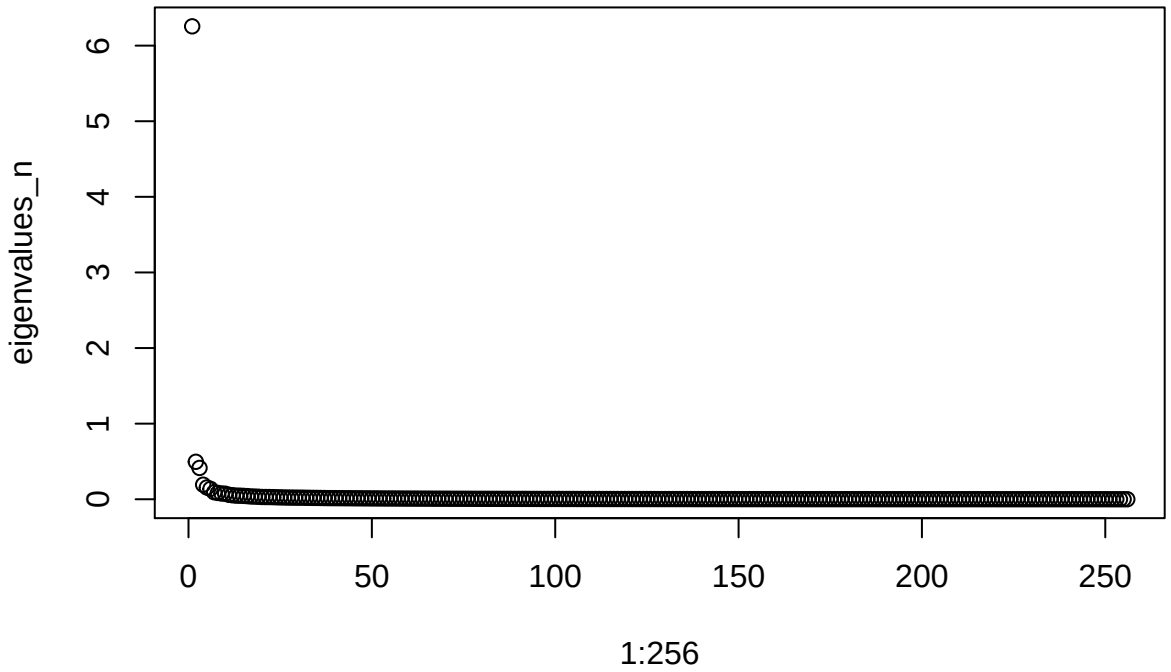


** Comparison of PCs** One can see that the PCs of both image groups are rather similar, in terms of the variance they account for (i.e. about 65% by the first PC in both cases). However, the second PC of the nature buildings accounts for about 10.7% of the variance while the second PC from the nature images only accounts for about 5.32%

```
# c)


eigenvalues_b <- eigen(cov(total_matrix_b_centered))$values
plot(1:256, eigenvalues_b, main = "Scree Plot Buildings")
```




```
cov_b <- cov(total_matrix_b)



eigenvalues_n <- eigen(cov(total_matrix_n_centered))$values
plot(1:256, eigenvalues_n, main = "Scree Plot Buildings")
```



Buildings Based on the scree plot we would choose the first 4 PCs for the building images. Because of high eigenvalue and before the elobow of the Curve. Based on the Kaiser criterion (extract PCs with eigenvalues > 1) we would extract the first two PCs.

Nature Based on the scree plot we would choose the first 3 PCs for the nature images. Because of high eigenvalue and before the elobow of the Curve. Based on the Kaiser criterion (see above) we would even extract only the first PC