Analysis of High Dimensional data HW2

Omkar Kulkarni, Alok Kumar, Thomas Verschueren

March 22, 2016

1. Introduction

We have data on average daily rainfall (mm/day) for the 365 days in the year and for 35 Canadian cities. Based on the SVD of $\mathbf{\Theta}$ and the MDS plot from HW1, our goal in HW2 is to interpret the SVD as a MDS and as a PCA, but this time we want to express our conclusions in terms of the original precipitation functions (i.e. rainfall as a function of time). This will require good interpretations of the two dimensions of the MDS plot (or biplot).

2. Analysis

We start with the same code from homework 1. First we read the data. For the transformation of the yearly rainfall precipitation data to functions we chose 'polynomial basis functions'. We opted for the degree of polynomial m=10 because starting from 10 we obtained nice fits for the rainfall data. An increase in degree seemed to be an over fit. We obtained the $\mathbf{\Theta}$ matrix of n = 35 rows (cities) by m+1 = 11 columns (regression coefficients of the polynomial fit for each city).

```
#setwd("C:/Users/ThomasV/Downloads/minerva files/MASTAT/sem2/HighDim/[HW2]")

load("CanadianWeather.rda")
da<-CanadianWeather[[1]]
da<-da[,,"Precipitation.mm"] # precipitation data
MetaData<-data.frame(city=colnames(da), region=CanadianWeather$region,
                     province=CanadianWeather$province, coord=CanadianWeather$coordinates)

# set m (degree of polynomials)
m <-10

# Rescaling days [1,365] to [0,1] interval
days<-1:365
days<-(days-min(days))/(diff(range(days))) # rescaling to [0,1]
phi<-poly(days,degree=m)

#### Generating theta_hat matrix ####
# loop: do for each city (row) : estimation of the m+1 theta parameters
# Optional : plot rainfall data with polynomial fit
# write theta parameters to theta matrix

theta <- matrix(NA,nrow=dim(da)[2],ncol=m+1) # initialisation of theta matrix
i=1
for (city in colnames(da)){
  m.city<-lm(da[,city]~phi)
  ## plot of fitted function
  #string_name <- paste(city," (m=",m,")") #concatenate city name with degree m of polynomial
  #plot(1:365,da[,city],main=string_name, xlab="day", ylab="precipitation (mm/day)")
```

```
  #lines(1:365,m.city$fitted.values,type="l", col=2)
  ## write theta parameters to theta matrix
  theta[i,] <- m.city$coefficients
  i=i+1
}
```

We apply a MDS to $\boldsymbol{\Theta}$, so that we can construct a 2-dimensional plot with each point representing a city. The distances between the points in the 2-dimensional MDS space are approximations of the distances between the rows of $\boldsymbol{\Theta}$, and hence can be interpreted as distances between the precipitation functions. Later in the report we will backtransform the SVD to the original function space in order to interpret these distances in terms of rainfall as a function of time.

The MDS starts from the truncated SVD of $\boldsymbol{\Theta}$,

$$\boldsymbol{\Theta}_k = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{V}_k^t$$

For the SVD we column centered the theta matrix. We stored the column means for the backtransformation of the SVD to the original function space which will be executed later in the report.

```
#### SVD ####
n <- nrow (theta)
H <- diag (n) -1/n* matrix (1, ncol=n, nrow=n)
theta.non.centered <- theta # storing the uncentered theta
theta[,] <- H %*% as.matrix (theta)  # column centering of data

## H.decentering keeps the colmeans of theta.non.centered
H.decentering <- theta.non.centered - theta
#colMeans(theta.non.centered)
#H.decentering
#dim(H.decentering) #35*11
#colMeans(theta)
#round(colMeans(theta), digits=10) # check column centered!

##SVD
theta.svd <- svd(theta)
k <-2  #To show a 2 dimensional plot. k = 2
Uk <- theta.svd$u[,1:k]
Dk <- diag(theta.svd$d[1:k])
Zk <-Uk %*% Dk
rownames(Zk) <- colnames(da)
#Zk

## Creation of the biplot (manually)
# V-tilda Vectors in the 2 dimensional space
Vk <- theta.svd$v[ ,1:k]
Vk <- as.data.frame(Vk)
rownames(Vk) <- 0:m
#Vk

# # Biplot: plot of city names in 2 dimensional space (Z1, Z2)
# plot(Zk, type ="n", xlab ="Z1", ylab ="Z2",
#      xlim =c(-43,15), ylim =c(-17,23), main = "Biplot")
# text(Zk, rownames(Zk), cex=0.45)
#
```

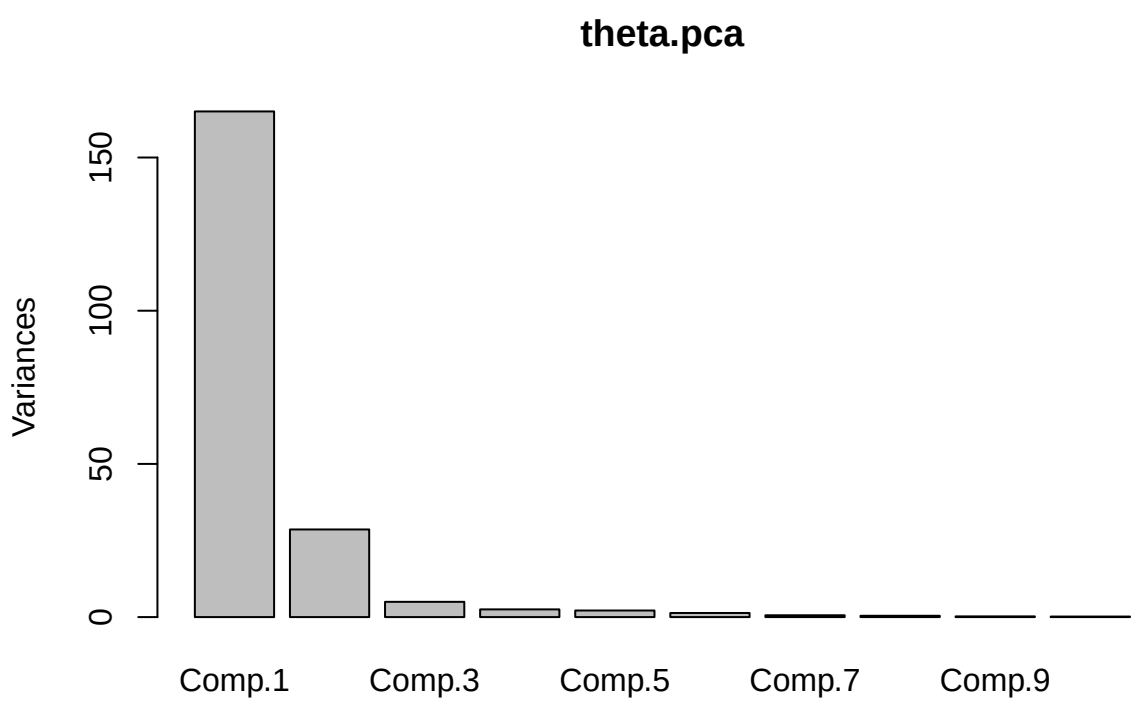
```
# # Biplot: Plot of V-tilda Vectors in the 2 dimensional space
# alpha <- 30 # rescaling to get better visualisation
# for (i in 1:17) {
#   arrows(0, 0, alpha*Vk[i,1], alpha*Vk[i,2],
#          length=0.2 , col=2)
#   text(alpha*Vk[i,1], alpha*Vk[i,2], rownames(Vk)[i],
#        col=2, pos=2, vfont=c("serif","bold"), cex=1.0)
# }
# abline (v=0 , lty=2, col ='grey')
# abline (h=0 , lty=2, col ='grey')
```

Next we performed a Principal Component Analysis on the centered theta matrix. 80% of the variance in the data is captured by the first principal component. The first 2 principal components capture 94% of the variance. In the variance barplot we notice a "knee/elbow" (significant drop in variance) after the first and after the second principal component. The first principal component already describes the differences between the cities w.r.t. the yearly rainfall pattern very well, but we decide to continue our analysis with the first 2 principal components, because they capture 94% of the variance. Adding a third principal would only add a small amount of extra captured variance. We will use the 2 principal components as our (orthogonal) dimensions in the (2-dimensional) biplot.

```
#### PCA ####
theta.pca<-princomp(theta) # theta is already centered in the SVD part
#theta.pca
##coefficients of the PCA
#theta.pca$loadings
#print(theta.pca$loadings,cutoff=0)
##how many PCs should we retain?
summary(theta.pca)
```

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     12.8469560 5.3481430 2.23036271 1.58686548
## Proportion of Variance  0.8005654 0.1387402 0.02412944 0.01221452
## Cumulative Proportion   0.8005654 0.9393056 0.96343506 0.97564958
##                            Comp.5      Comp.6      Comp.7      Comp.8
## Standard deviation     1.46675478 1.156108566 0.769242533 0.653478865
## Proportion of Variance 0.01043545 0.006483262 0.002870271 0.002071378
## Cumulative Proportion  0.98608503 0.992568296 0.995438567 0.997509946
##                             Comp.9      Comp.10      Comp.11
## Standard deviation     0.476113866 0.4181928892 0.3343336852
## Proportion of Variance 0.001099558 0.0008483003 0.0005421964
## Cumulative Proportion  0.998609503 0.9994578036 1.0000000000
```

```
screeplot(theta.pca)
```




```
##display the scores
#theta.pca$scores
## retain first 2 PCs -> Zk with k=2
Zk_PCA<-theta.pca$scores[,1:2]
colnames(Zk_PCA)[1:2]<-c("PC1","PC2")
rownames(Zk_PCA)<-colnames(da)
##display biplot with first 2 PCs
# Zk_PCA<-as.data.frame(Zk_PCA)
# plot(Zk_PCA[,1],Zk_PCA$scores[,2], type="n", xlab="PC1", ylab="PC2",
#      xlim =c(-43,15), ylim =c(-17,23), main = "Biplot")
# text(Zk_PCA[,1],Zk_PCA[,2],
#      labels=rownames(Zk_PCA), cex=0.60)
# abline(h=0)
# abline(v=0)
```

The code below verifies whether the loadings obtained from the PCA of the $\boldsymbol{\Theta}$ matrix are equal to the $\boldsymbol{V}_k$ vectors obtained from the SVD of the $\boldsymbol{\Theta}$ matrix and whether the scores obtained from the PCA of the $\boldsymbol{\Theta}$ matrix are equal to the $\boldsymbol{Z}_k = \boldsymbol{U}_k\boldsymbol{D}_k$ vectors obtained from the SVD of the $\boldsymbol{\Theta}$ matrix.

```
#### PCA vs SVD ####
##loadings PCA equal to theta.svd$v
#theta.pca$loadings
#print(theta.pca$loadings,cutoff=0)
#theta.svd$v
#Vk # (k=2) --> first 2 PC
##scores PCA equal to Zk=Uk*Dk
```

```
#theta.pca$scores
#Zk # (k=2) Zk=Uk*Dk
```

Next we perform a backtransformation of the SVD to the original function space. For the $\mathbf{\Phi}$ matrix (which should have dimension 365 by 11, since we have m+1=11 regression coefficients and 365 daily rainfall precipitation data) we have to add a column (365 by 1) of ones to our originally created (365 by 10) $\mathbf{\Phi}$ matrix in order to capture the constant of the regression. This has been verified by checking whether the fitted values obtained from the regression are equal to $\mathbf{\Theta\Phi}^t$.

For the SVD and the PCA we column centered the $\mathbf{\Theta}$ matrix (by subtracting the data with the column means). In order to obtain the original function space we have to de-center the $\boldsymbol{Z}_k\boldsymbol{V}_k^t$ by adding the column means. These column means have been stored in the SVD process before centering. It was not possible to use the inverse of the H matrix for decentering, since H is a singular (35 by 35) matrix (of rank 34).

```
#### Functional Biplot ####
# dim(phi) # 365*10
# head(phi)
constant <- rep(1,365)
phi_Mplus1<-matrix(c(constant,phi),nrow=365)
# dim(phi_Mplus1) # 365*11
# head(phi_Mplus1)
### test of phi_Mplus1 !!!
# m.city.test<-lm(da[,"Vancouver"]~phi)
# theta.test <- as.matrix(m.city.test$coefficients)
# dim(theta.test) # 11*1
# test.fit <- t(theta.test) %*% t(phi_Mplus1) # t(1*11)%*%t(365*11) --> 1*365
# dim(test.fit) # 1*365
# round(test.fit-m.city.test$fitted.values,3)

### H_invers %*% H %*% theta = H_invers %*% Zk %*% Vk
#library(Matrix)
#rankMatrix(H) # rank 34 instead of 35
#H_invers <- solve(H) # --> H singular

### Functional Biplot
## Zk : 35*2
## Vk : 11*2 --> t(Vk) : 2*11
## X : 365 *11 -> t(X) : 11*365
## Yk : (35*2)(2*11)(11*365)=(35*365)
## "de"-center Zk (Zk %*% t(Vk))+H.centering
## H.decentering : 35*11
## --> (Zk %*% t(Vk)) + H.decentering
Yk <- ((Zk %*% t(Vk))+H.decentering) %*% t(phi_Mplus1)
# dim(Yk) # 35*365
```

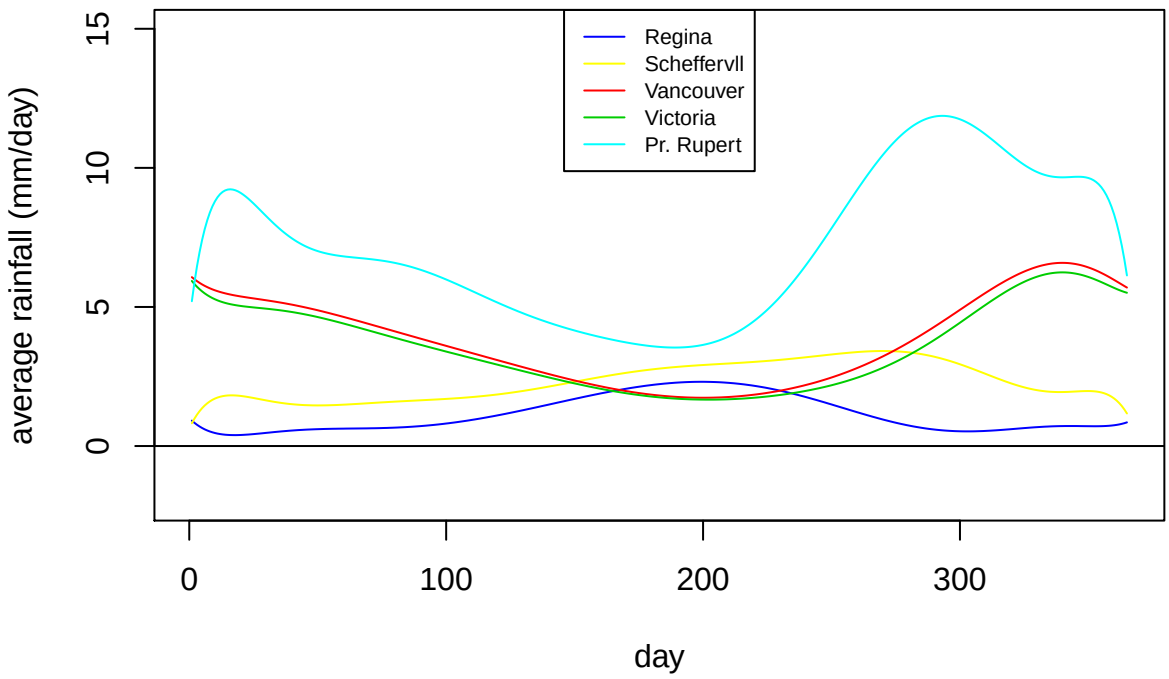
The plot below shows the yearly rainfall pattern of cities Regina, Scheffervill, Vancouver, Victoria and Pr. Rupert obtained by backtransforming the SVD to the original function space. The lines of Vancouver and Victoria are very close to each other indicating a similar yearly rainfall pattern. This is also visible in the biplot (a biplot is presented a bit further in the report). The (z_{2i1},z_{2i2}) coordinates of both cities on the biplot are very close to each other, in other words: the distance between both cities is very small, which indicates Vancouver and Victoria have similar rainfall patterns. Pr. Rupert and Regina appear to have very dissimilar rainfall patterns. As a consequence their distance on the biplot is very large. Differences in rainfall patterns are visible in terms of total amount of yearly rainfall (or average amount of daily rainfall) and in terms of where (in which season) the maximum rainfall is situated.

In the following we will talk about "total amount of yearly rainfall", but of course this is the same as speaking in terms of "average amount of daily rainfall".

```
#### plots ####
selected.cities <- c(20,7,26,27,29)
#selected.cities <- c(1:35)
MetaData$city[selected.cities]
```

```
## [1] Regina      Scheffervll Vancouver   Victoria    Pr. Rupert
## 35 Levels: Arvida Bagottville Calgary Charlottvl Churchill ... Yellowknife
```

```
plot(days*364+1,Yk[1,],type="n", ylim=c(-2,15),
      xlab="day", ylab="average rainfall (mm/day)")
for(i in selected.cities) {
      lines(days*364+1,Yk[i,],col=i)
}
legend(x="top", legend=MetaData$city[selected.cities],
      col=selected.cities, cex=0.7, lty=1)
abline(h=0)
```

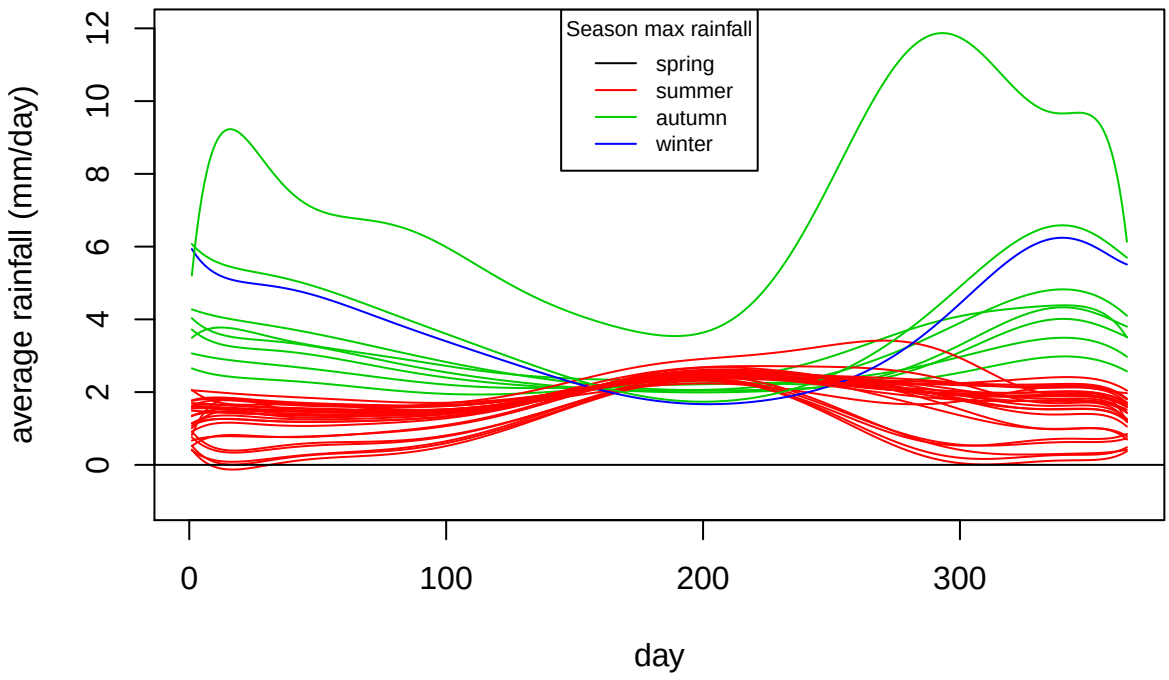


Extra variables have been created out of the available data for each city: total yearly precipitation (mm); whether a city has a high or low total yearly precipitation (with an average yearly precipitation of 794 mm as a cutoff value) ; total precipitation per season (mm); season with the highest total precipitation. Most cities (26) have a maximal rainfall in summer. None of the cities has a maximal rainfall in spring.

```r
## create database with more variables out of available info
data.cities <-MetaData
data.cities$V1 <- Zk[,1]
data.cities$V2 <- Zk[,2]
# total yearly precipitation
data.cities$total.rain <- rowSums(t(da))
data.cities$total.rain.centered <- data.cities$total.rain - mean(data.cities$total.rain)
# mean(data.cities$total.rain) #794 mm
# high or low yearly precipitation (cut off point = average)
data.cities$high.low.rain <-
  cut(data.cities$total.rain.centered,
      breaks=c(-2000,0,2000),
      labels=c("low rainfall","high rainfall"),
      right=FALSE,
      ordered_result=TRUE)
# total precipitation per season
data.cities$total.spring <-rowSums(t(da)[,80:171]) # march,21 till june,20
data.cities$total.summer <-rowSums(t(da)[,172:263]) # june,21 till sept,20
data.cities$total.autumn <-rowSums(t(da)[,264:354]) # sept,21 till dec,20
data.cities$total.winter <-rowSums(t(da)[,c(1:79,355:365)]) # dec,20 till dec,31 and
# jan,1 till march,20
# which season has maximal total precipitation ?
data.cities <- within(data.cities,{
  season.max.rain <- apply(data.cities[, c("total.spring","total.summer",
                                           "total.autumn","total.winter")],
                           1, function(x) which(x == max(x)))
  season.max.rain <- factor(season.max.rain,
                            levels=c(1,2,3,4),
                            labels=c("spring","summer","autumn","winter"))
})
summary(data.cities$season.max.rain)
```

```
## spring summer autumn winter
##      0     26      8      1
```

```r
### plots
selected.cities <- c(1:35)
#MetaData$city[selected.cities]
plot(days*364+1,Yk[1,],type="n", ylim=c(-1,12),
     xlab="day", ylab="average rainfall (mm/day)")
for(i in selected.cities) {
  lines(days*364+1,Yk[i,],col=data.cities$season.max.rain[i])
}
legend(x="top", legend=levels(data.cities$season.max.rain)[1:4],
       col=1:4, title='Season max rainfall', cex=0.7, lty=1)
abline(h=0)
```



In the figure above the daily rainfall of all cities is shown. The colors of the lines indicate in which season the cities have a maximum rainfall. We notice that cities which have a maximum rainfall in summer appear to have a smaller overall yearly amount of rainfall compared to cities which have a maximum rainfall in autumn or winter.

We checked whether this is in line with our interpretation for the two dimensions of the MDS plot (biplot).

The first dimension in the biplot is the most important one. We have learnt from the PCA the first component explains 80% of the variance between cities w.r.t. the yearly rainfall pattern. In order to understand the meaning of the first dimension, $\hat{Y}_{2i}(t)$ has been plotted in the figure below as a function of t with z_{2i1} varying between a small and a large score, and with z_{2i2} fixed at a meaningful constant. For the varying z_{2i1} we opted for lines for the minimum, the median, the maximum value and the 10th, 25th, 75th, 90th quantiles of the z_{2i1} coordinates of the 35 cities. For the fixed z_{2i2} we opted for the median value (which is equal to 0) of the z_{2i2} coordinates of the 35 cities. Other values for z_{2i2} lead to a similar interpretation.

The figure below shows that higher (lower) z_{2i1} (Z1) values appear to result into lower (higher) daily amounts of rainfall, except for days in summer. With Z1 at the 25th quantile the rainfall pattern is more or less constant. Low Z1 values appear to result into a maximum rainfall in autumn or winter, high Z1 values appear to result into a maximum rainfall in summer. When we ignore extreme values for Z1 (e.g. Pr. Rupert, which has a very low Z1, the minimum Z1 in the data) the influence of Z1 appears to be smaller.

This is in line with the figure above. Cities with a smaller (higher) overall yearly amount of rainfall with a maximum rainfall in summer (autumn or winter) must have a high (low) Z1 coordinate.

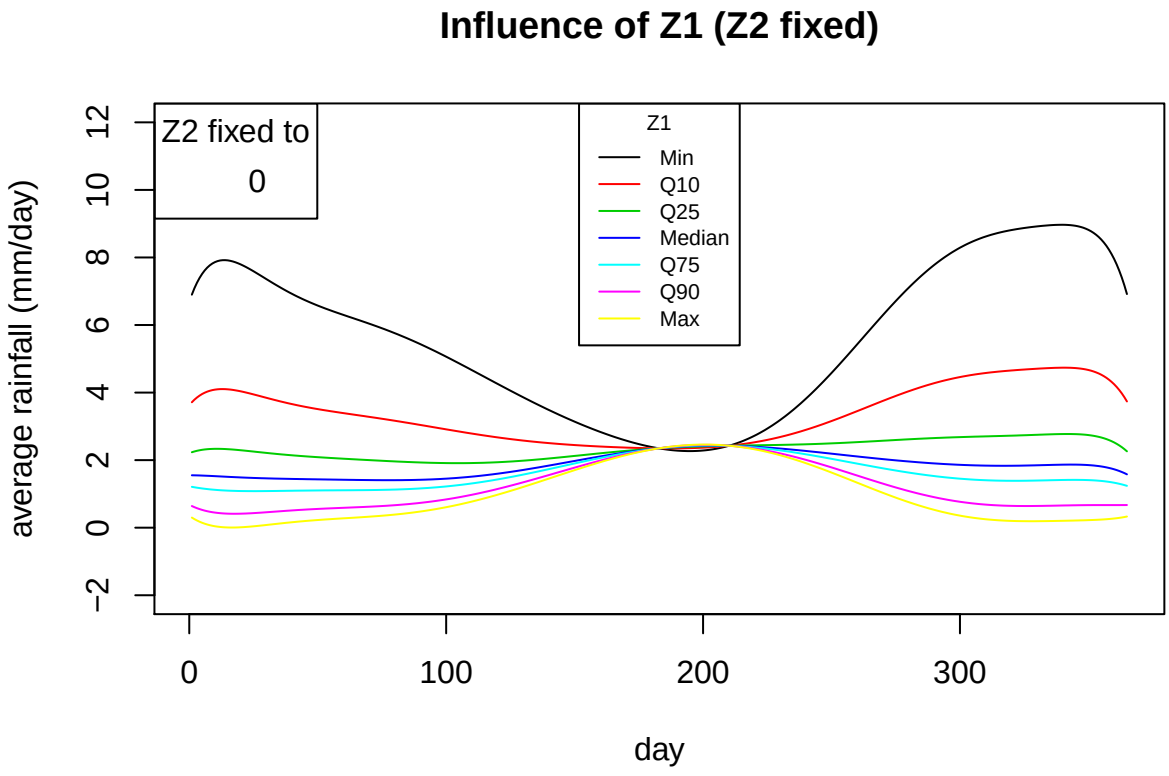
```
#### Meaning of the first dimension ####
## fixed value(s) for Z2
Z2i2.seq <- round(quantile(Zk[,2],  probs = c(0.0,0.1,0.25,0.5,0.75,0.90,1.0)))
Z2i2.val <- Z2i2.seq[c("50%")]
Z2i2.val
```

```
## 50%
##   0
```

```
## variable sequence of values for Z1
Z2i1.seq <- round(quantile(Zk[,1],  probs = c(0.0,0.1,0.25,0.5,0.75,0.90,1.0)))
Z2i1.seq
```

```
##   0%  10%  25%  50%  75%  90% 100%
##  -43  -15   -2    4    7   12   15
```

```
for(i in Z2i2.val){
  Z2i2 <-i
  color.dummy<-1 # initialize color dummy for plot
  plot(days*364+1,Yk[1,],type="n", ylim=c(-2,12),
       xlab="day", ylab="average rainfall (mm/day)", main="Influence of Z1 (Z2 fixed)")
  for (j in Z2i1.seq){
    Z2i1 <-j
    Zk.new <- matrix(c(Z2i1,Z2i2), ncol=2)
    Yk.new <- ((Zk.new %*% t(Vk))+H.decentering[1,]) %*% t(phi_Mplus1)
    lines(days*364+1,Yk.new[1,],col=color.dummy)
    color.dummy <- color.dummy + 1
  }
  legend(x="top", legend=c("Min","Q10","Q25","Median","Q75","Q90","Max"), title=paste('Z1'),
         col=1:color.dummy, cex=0.7, lty=1)
  legend(x="topleft",legend=Z2i2[1],title='Z2 fixed to')
}
```



We have learnt from the PCA that the second dimension in the biplot (the second principal component) explains 14% of the variance between cities w.r.t. the yearly rainfall pattern. This second dimension is orthogonal to the first dimension, which means there is no overlap in explained variance, resulting in a total of 94% explained variance explained by the first 2 principal components together. In order to understand the meaning of the second dimension, $\hat{Y}_{2i}(t)$ has been plotted in the figure below as a function of t with z_{2i2} varying between a small and a large score, and with z_{2i1} fixed at a meaningful constant. For the varying z_{2i2} we opted for lines for the minimum, the median, the maximum value and the 10th, 25th, 75th, 90th quantiles of the z_{2i2} coordinates of the 35 cities. For the fixed z_{2i1} we opted for the the minimum (which is equal to -43), the 10th quantile (-15), the median (4) and the maximum value (15) of the z_{2i1} coordinates of the 35 cities.

The figure below shows that higher (lower) z_{2i2} (Z2) values appear to result into slightly higher (lower) daily amounts of rainfall. The differences in daily amounts of rainfall are not as significant as for a change in z_{2i1}. Especially when we ignore extreme values for Z2 (e.g. Pr. Rupert, which has a very high Z2, the maximum Z2 in the data) the influence of Z2 appears to be small. Looking at the 4 plots we notice the first principal component has a much higher influence on the rainfall patterns. This is logical since it explains 80% of the variance between cities w.r.t. the yearly rainfall pattern.

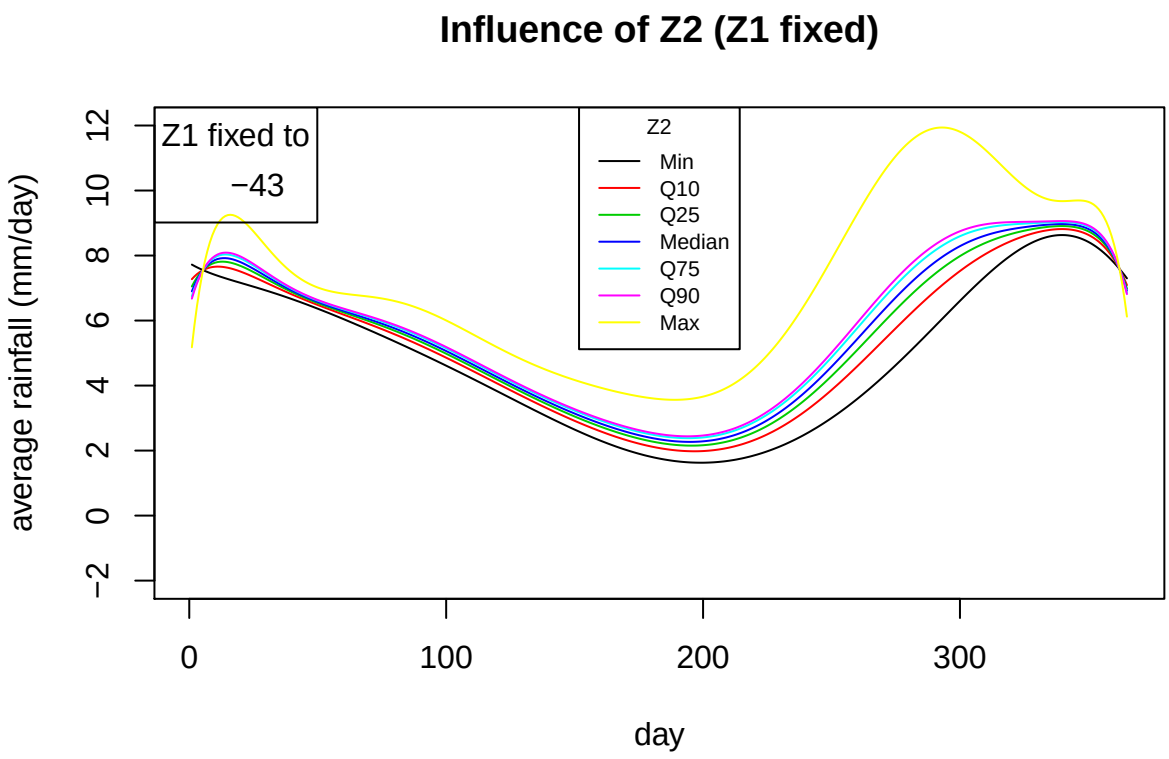
```
#### Meaning of the second dimension ####
## fixed value(s) for Z1
Z2i1.seq <- round(quantile(Zk[,1],  probs = c(0.0,0.1,0.25,0.5,0.75,0.90,1.0)))
Z2i1.val <- Z2i1.seq[c("0%","10%","50%","100%")]
Z2i1.val
```

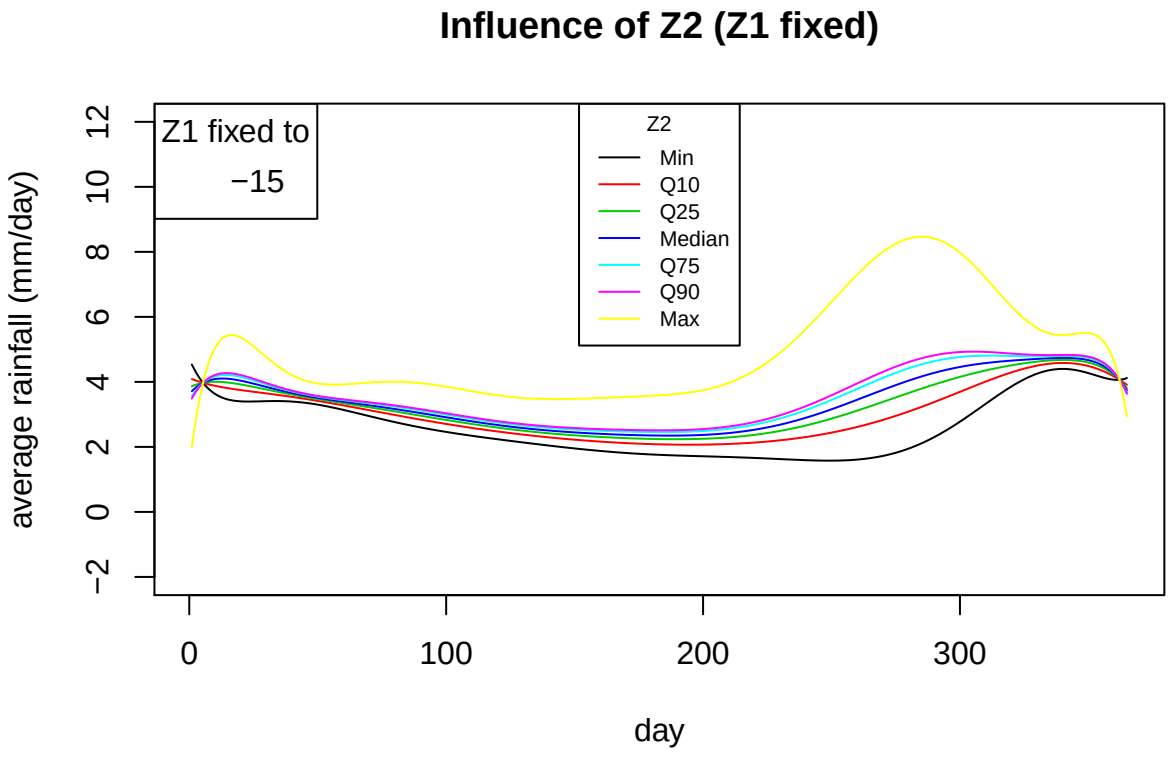
```
##   0%  10%  50% 100%
##  -43  -15    4   15
```

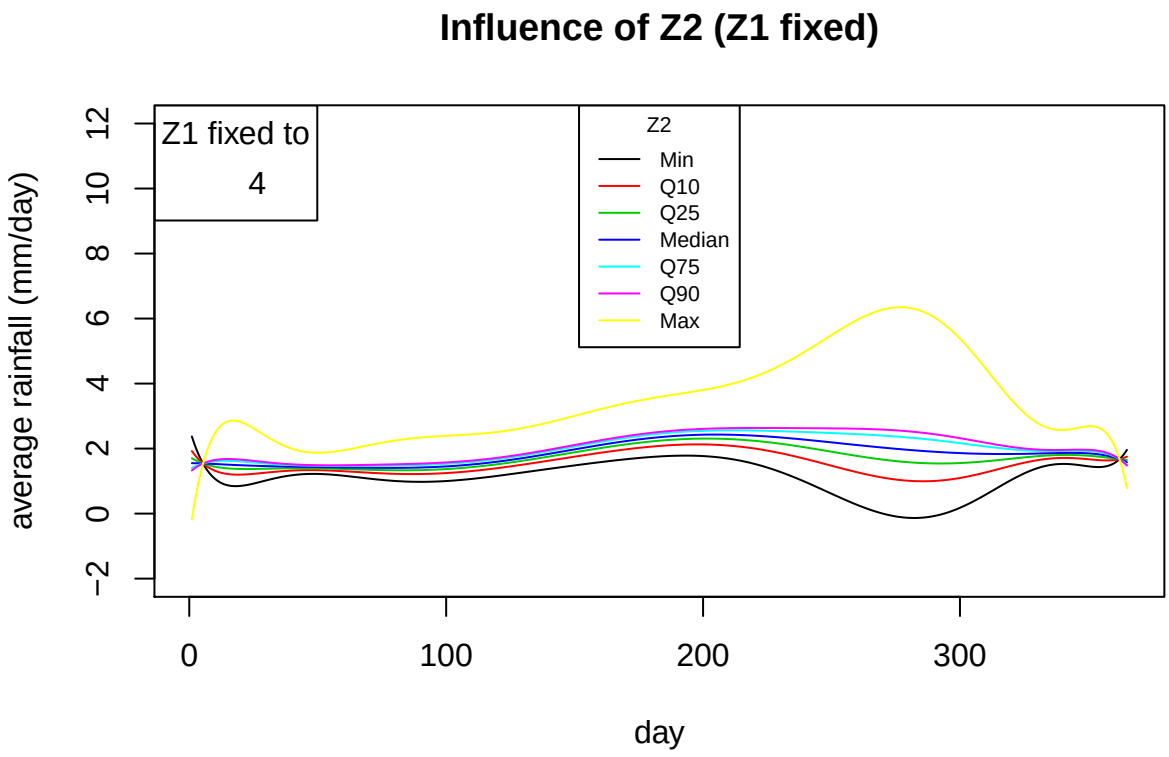
```
## variable sequence of values for Z2
Z2i2.seq <- round(quantile(Zk[,2],  probs = c(0.0,0.1,0.25,0.5,0.75,0.90,1.0)))
Z2i2.seq
```

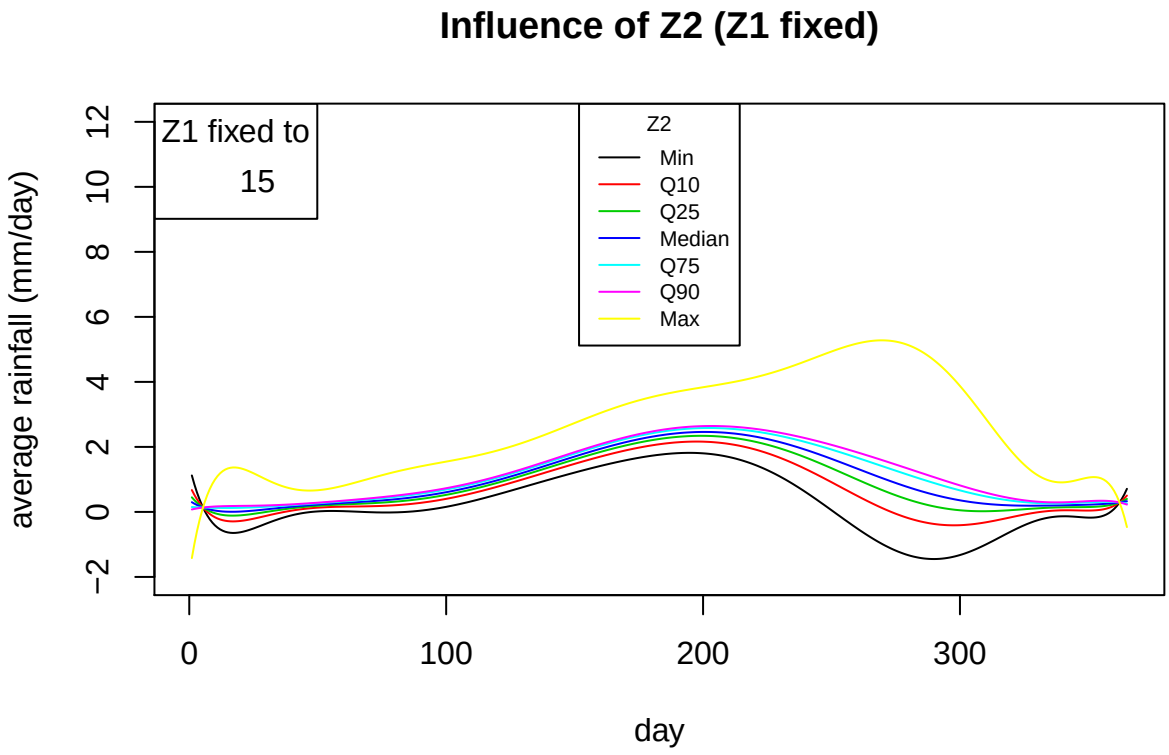
```
##   0%  10%  25%  50%  75%  90% 100%
##  -11   -5   -2    0    2    3   23
```

```
for(i in Z2i1.val){
  Z2i1 <-i
  color.dummy<-1 # initialize color dummy for plot
  plot(days*364+1,Yk[1,],type="n", ylim=c(-2,12),
       xlab="day", ylab="average rainfall (mm/day)", main="Influence of Z2 (Z1 fixed)")
  for (j in Z2i2.seq){
    Z2i2 <-j
    Zk.new <- matrix(c(Z2i1,Z2i2), ncol=2)
    Yk.new <- ((Zk.new %*% t(Vk))+H.decentering[1,]) %*% t(phi_Mplus1)
    lines(days*364+1,Yk.new[1,],col=color.dummy)
    color.dummy <- color.dummy + 1
  }
  legend(x="top", legend=c("Min","Q10","Q25","Median","Q75","Q90","Max"), title=paste('Z2'),
         col=1:color.dummy, cex=0.7, lty=1)
  legend(x="topleft",legend=Z2i1[1],title='Z1 fixed to')
}
```


Influence of Z2 (Z1 fixed)
Z1 fixed to
−43
Z2
Min
Q10
Q25
Median
Q75
Q90
Max
average rainfall (mm/day)
−2
0
2
4
6
8
10
12
0
100
200
300
day

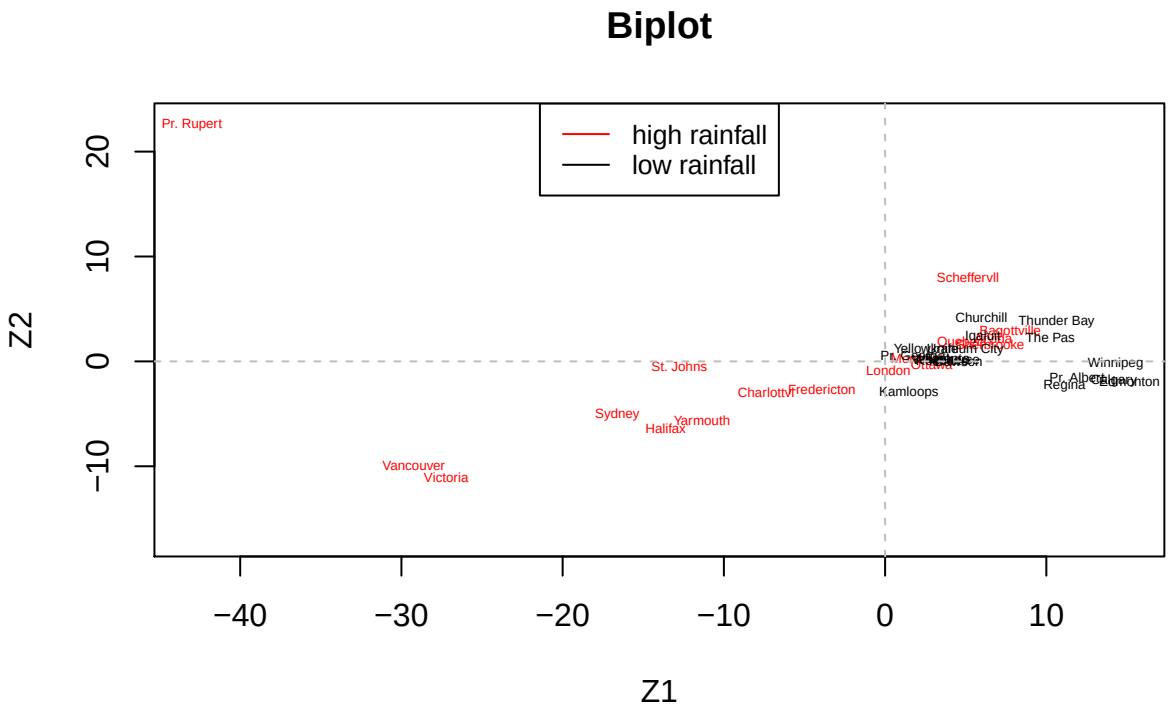

Influence of Z2 (Z1 fixed)
Z1 fixed to
−15
Z2
Min
Q10
Q25
Median
Q75
Q90
Max
average rainfall (mm/day)
−2
0
2
4
6
8
10
12
0
100
200
300
day


Influence of Z2 (Z1 fixed)
Z1 fixed to
4
Z2
Min
Q10
Q25
Median
Q75
Q90
Max
average rainfall (mm/day)
−2
0
2
4
6
8
10
12
0
100
200
300
day


Influence of Z2 (Z1 fixed)
Z1 fixed to
15
Z2
Min
Q10
Q25
Median
Q75
Q90
Max
average rainfall (mm/day)
−2
0
2
4
6
8
10
12
0
100
200
300
day

Next we created a biplot. This can be done in 2 ways. The scores of the PCA represent the coordinates $\boldsymbol{Z}_k$ of the 35 cities in the (z_{2i1}, z_{2i2}) space. The loadings of the PCA represent the coordinates of the m+1=11 $\boldsymbol{V}_k$ vectors in the (z_{2i1}, z_{2i2}) space, which can also be obtained by the SVD of $\boldsymbol{\Theta}$. The scores $\boldsymbol{Z}_k$ can be obtained by the SVD of $\boldsymbol{\Theta}$, by $\boldsymbol{Z}_k = \boldsymbol{U}_k \boldsymbol{D}_k$.
The figure below shows the biplot with the cities marked in red (black) whether the city has a high (low) total yearly amount of rainfall precipitation (with an average yearly precipitation of 794 mm as a cutoff value). Out of the interpretation of Z1 we know that cities with a low (high) Z1 tend to have a high (low) total yearly amount of rainfall. This is clearly visible in the biplot (e.g.. Pr. Rupert, Vancouver, Victoria,...). Our interpretation of Z2, which says that cities with a high (low) Z2 tend to have a high (low) total yearly amount of rainfall is visible by looking at e.g. the cities Pr.Rupert and Schefferville.

```
## Biplot: plot of city names in 2 dimensional space (Z1, Z2)
plot(Zk, type ="n", xlab ="Z1", ylab ="Z2",
     xlim =c(-43,15), ylim =c(-17,23), main = "Biplot")
text(Zk, rownames(Zk), cex=0.45, col=as.numeric(data.cities$high.low.rain))
legend("top",legend=unique(data.cities$high.low.rain),
       col=as.numeric(unique(data.cities$high.low.rain)), cex=0.8, lty=1)
abline (v=0 , lty=2, col ='grey')
abline (h=0 , lty=2, col ='grey')
```

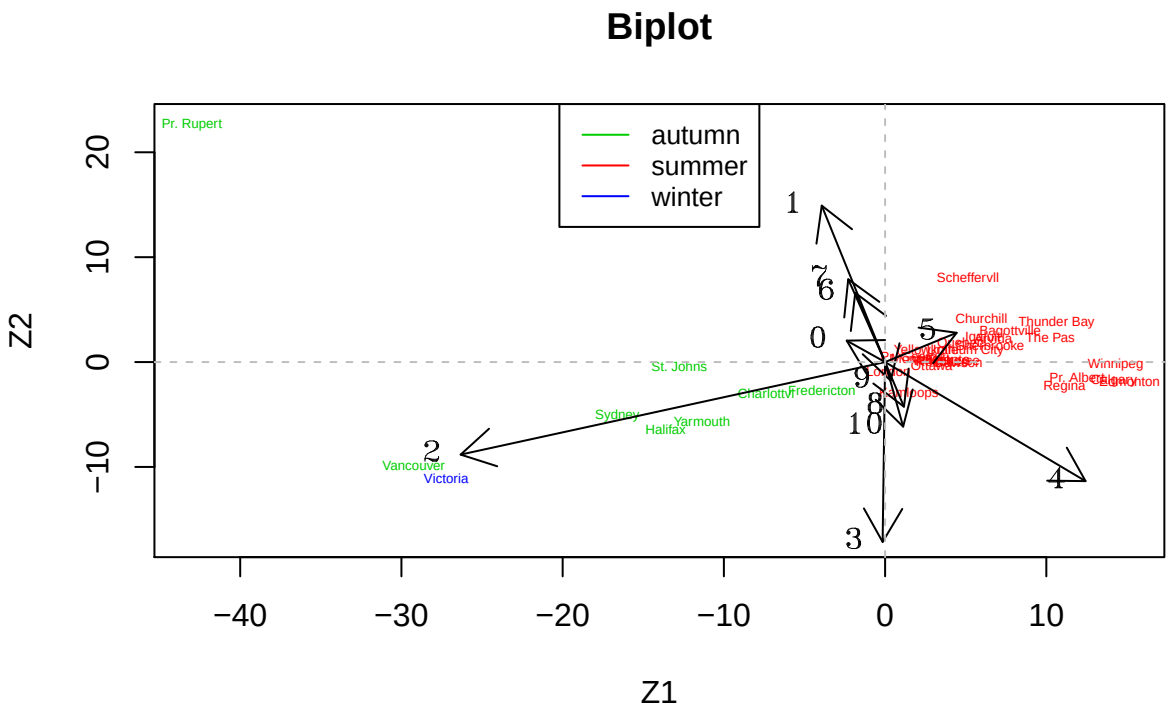


The next figure below shows the biplot with the cities marked in red, green and blue respectively whether the city has a maximum rainfall in summer, autumn or winter respectively. Out of the interpretation of Z1 we know that cities with a high (low) Z1 tend to have a maximum rainfall in summer (autumn/winter). This is clearly visible in the biplot.

The polynomial functions of degree m form a m+1 dimensional base which lead to m+1 V-tilda vectors when projected on the 2 dimensional (Z1,Z2) space. These V-tilda vectors are shown in the biplot below. We can also interpret them in terms of rainfall as a function of time. Under the biplot the loadings are shown. These

loadings represent the coordinates of the V-tilda vectors. 80% of the variance between cities w.r.t. the yearly rainfall pattern is explained by the first principal component, the coordinate of Z1. The largest loadings (in absolute value) for PC1 are the ones for m=2 (3rd row, because the 1st row represents m=0) and m=4. The loadings for PC1 of m=2 and m=4 have opposite signs, visible by V-tilda vectors pointing to the left and to the right respectively. V-tilda m=2 points toward lower values of Z1, V-tilda m=4 points to higher values of Z1. Cities which have high values after projection on V-tilda m=2, have low Z1 values, which means they have a high total amount of yearly rainfall and they are more likely to have a maximum total rainfall in autumn or winter (instead of in summer). 80% of the variance between the cities w.r.t. the yearly rainfall pattern is explained by this first PC (Z1), which means 80% of the variance is explained by the contrast between high and low yearly amount of rainfall and whether the cities have a maximum rainfall in summer or in autumn/winter.

```
## Biplot: plot of city names in 2 dimensional space (Z1, Z2)
plot(Zk, type ="n", xlab ="Z1", ylab ="Z2",
     xlim =c(-43,15), ylim =c(-17,23), main = "Biplot")
text(Zk, rownames(Zk), cex=0.45, col=as.numeric(data.cities$season.max.rain))
legend("top",legend=unique(data.cities$season.max.rain),
       col=as.numeric(unique(data.cities$season.max.rain)),
       cex=0.8, lty=1)
## Biplot: Plot of V-tilda Vectors in the 2 dimensional space
alpha <- 30 # rescaling to get better visualisation
for (i in 1:17) {
  arrows(0, 0, alpha*Vk[i,1], alpha*Vk[i,2], length=0.2, col=1)
  text(alpha*Vk[i,1], alpha*Vk[i,2], rownames(Vk)[i],
       col=1, pos=2, vfont=c("serif","bold"), cex=1.0)
}
abline (v=0 , lty=2, col ='grey')
abline (h=0 , lty=2, col ='grey')
```



```
theta.pca$loadings[,1:2]
```

```
##             Comp.1      Comp.2
##  [1,] -0.079410790  0.06810956
##  [2,] -0.131010581  0.49713689
##  [3,] -0.877610882 -0.29410478
##  [4,] -0.005131931 -0.57157698
##  [5,]  0.414428813 -0.37836321
##  [6,]  0.147959007  0.09306582
##  [7,] -0.061767736  0.21983015
##  [8,] -0.076218497  0.26305605
##  [9,]  0.038914245 -0.14174999
## [10,]  0.011443597 -0.05893241
## [11,]  0.037283522 -0.20520306
```

3. Discussion and Conclusion

The goal of this homework is to interpret the SVD as a MDS and as a PCA in terms of rainfall as a function of time. When we look at the yearly rainfall precipitation data of the cities we notice differences in terms of total amount of yearly rainfall (or average amount of daily rainfall) and in terms of where (in which season) the maximum rainfall is situated.

First we start by interpreting the SVD as a MDS. The distances between the 35 cities w.r.t. the original yearly rainfall patterns (the 35 observations in the original data matrix $\boldsymbol{X}$) are best approximated in a 2-dimensional space by the truncated $\boldsymbol{X}_k$ which can be written as $\boldsymbol{X}_k = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{V}_k^t$. Distances between cities in the MDS plot can be seen as differences in z_{2i1} and/or z_{2i2} values. As a consequence we can interpret the differences between cities in terms of rainfall as a function of time based on our interpretations of z_{2i1} and z_{2i2}. Higher (lower) z_{2i1} (Z1) values result into lower (higher) daily amounts of rainfall, except for days in summer. Low z_{2i1} values result into a maximum rainfall in autumn or winter, high z_{2i1} values result into a maximum rainfall in summer. Higher (lower) z_{2i2} (Z2) values result into slightly higher (lower) daily amounts of rainfall.

Next we interpret the SVD as a PCA. The covariances between the 35 cities w.r.t. the original yearly rainfall patterns (the 35 observations in the original data matrix $\boldsymbol{X}$) are best approximated in a 2-dimensional space by the truncated $\boldsymbol{X}_k$. 80% of the variance between cities w.r.t. the yearly rainfall pattern is explained by the first principal component. Another 14% of the variance is explained by the second principal component. Since the principal components are orthogonal (uncorrelated) there is no overlap in explained variance for the 2 PCs. Consequently together they explain 94% of the variance between cities w.r.t. the yearly rainfall pattern. The approximation error for the $\boldsymbol{X}_k$ is 6%. This is a quality measure for our SVD. Based on our interpretation of z_{2i1} we can conclude that the largest variance (80%) can be explained in terms of differences in total amount of yearly rainfall (or average amount of daily rainfall) and in terms of whether the cities have a maximum rainfall in summer or in autumn/winter. Based on our interpretation of z_{2i2} another small amount of variance (14%) can be explained in terms of differences in total amount of yearly rainfall (or average amount of daily rainfall).