Analysis of High Dimensional data HW1

Omkar Kulkarni, Thomas Verschueren, Alok Kumar

March 6, 2016

1. Introduction

We have data on avarage daily rainfall (mm/day) for the 365 days in the year and for 35 Canadian cities.

By doing Functional Data Analysis we reduce the dimension from p=365 to m+1, with m the degree of a polynomial. Next we perform Multi Dimensional Scaling on the avarage daily rainfall functions of the Canadian cities. This way we obtain a reduction from m+1 to k=2.

2. Analysis

2.1 Functional Data Analysis

First we read the data

```r
setwd("C:/Users/ThomasV/Downloads/minerva files/MASTAT/sem2/HighDim/[HW1]")
#install.packages("ldr")
#library(ldr) # needed for the bf function for generation of Fourier basis

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)
```

For the transformation 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 overfit.

```r
# 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)
#dim(phi)

# 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)
  #summary(m.city)
  # 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 obtained the $\boldsymbol{\Theta}$ matrix of n = 35 rows (cities) by m+1 = 11 columns (regression coefficients of the polynomial fit for each city).

2.2 Multidimensional Scaling of Functions

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.

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

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

```
# SVD
X <- theta
n <- nrow (X)
H <- diag (n) -1/n* matrix (1, ncol=n, nrow=n)
X[,] <- H %*% as.matrix (X)  # column centering of data
#colMeans(X)
round(colMeans(X), digits=10) # check column centered!
```

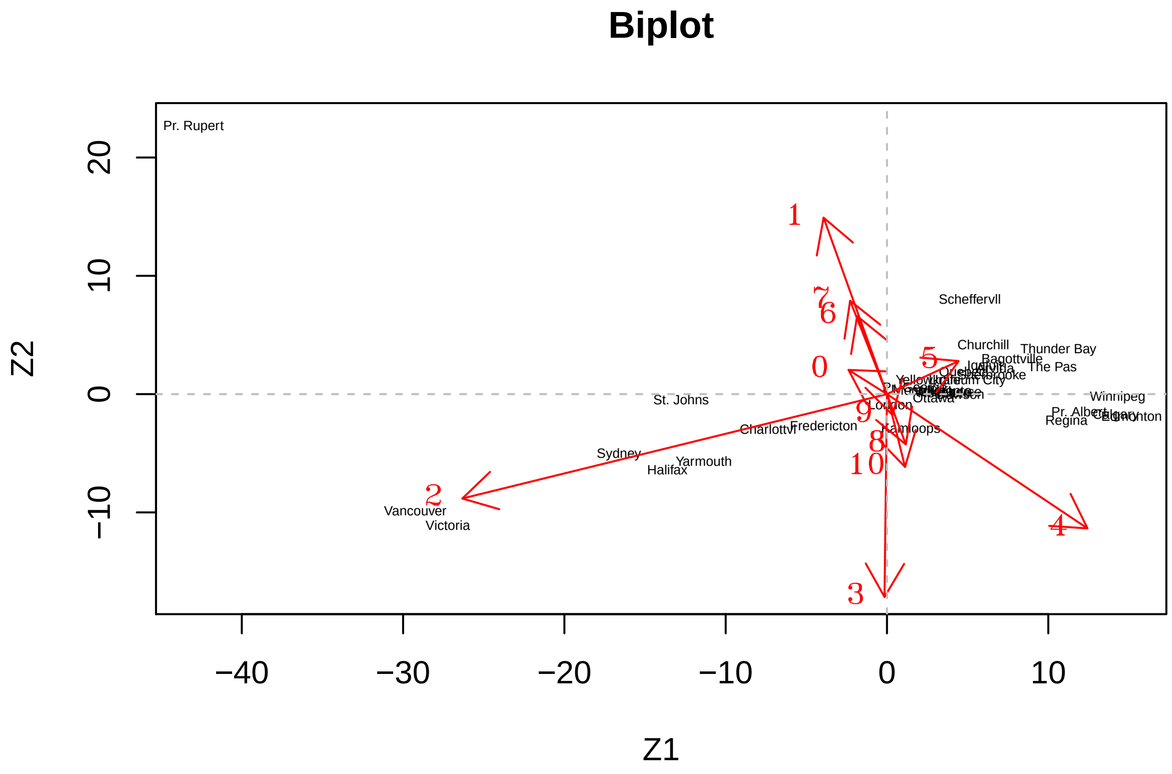
```
##  [1] 0 0 0 0 0 0 0 0 0 0 0
```

```
X.svd <- svd(X)
#To show a 2 dimensional plot. k = 2
k <-2
Uk <- X.svd$u[ ,1:k]
Dk <- diag(X.svd$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 <- X.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')
```



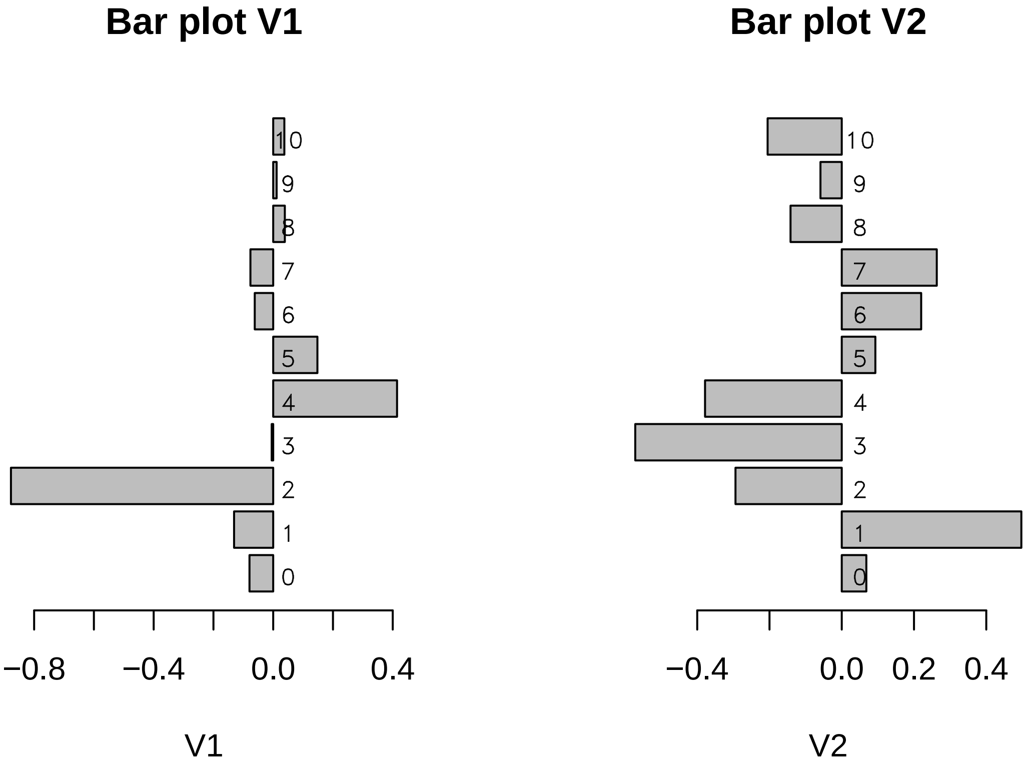
2.3 Interpretations

The MDS resulted in a biplot. We expect cities located closely to each other on the biplot show similar rainfall patterns and similar polynomial precipitation functions and cities located far away from each other on the biplot show dissimilar rainfall patterns and dissimilar polynomial precipitation functions. 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. Cities with high (low) values for projections on V-tilda "i" (i=O,. . . ,m+1) will have high (low) estimated regression coefficients for the polynomial function of degree i.

The bar plots show the elements in V1 (1st dimension of Zk with k=2) and V2 (2nd dimension of Zk with k=2). For instance high V-tilda “2” values will lead to low Z1 coordinates and low Z2 coordinates. Vancouver and Victoria (which have a high V-tilda “2” distance when projected on the V-tilda “2” vector) indeed have low Z1 coordinates and low Z2 coordinates.

```
# Barplots with elements of Vk
par(mfrow=c(1,2))
barplot(Vk[,1], xlab="V1", names.arg=NA, horiz=TRUE, main = "Bar plot V1")
for(i in 1:(m+1))
{   text(0.05, -0.6+i*1.2,rownames(Vk)[i], cex=0.8, vfont=c("sans serif","bold"))}

barplot(Vk[,2], xlab="V2", names.arg=NA, horiz=TRUE, main = "Bar plot V2")
for(i in 1:(m+1))
{ text(0.05, -0.6+i*1.2, rownames(Vk)[i], cex=0.8, vfont=c("sans serif","bold")) }
```



```
par(mfrow=c(1,1))
```

To give a meaningful interpretation of the biplot, we combine the (Z1,Z2) coordinates of the cities with the metadata for every city and we create some extra variables out of the available data for each city: total yearly precipitation (mm); total precipitation per season (mm); season with the highest total precipitation. With these data we perform a linear regression to see if there is a relation between the (Z1,Z2) coordinates of the cities and the metadata (and the extra created variables).

```
## 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)
# 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
```

Most cities (26) have a maximal rainfall in summer. None of the cities has a maximal rainfall in spring.

First we perform a linear regression on dependent variable Z1 with independent variables region (categorical variable with base “Arctic”), (centered) total yearly rainfall (mm) and a categorical variable which indicates the season with the highest rainfall (dummy variables for autumn and winter, base=summer). Since “region”" and “province”" seem to be quite highly correlated, (eg. region “pacific” = province “British Colombia”) we decided to implement only region in the regression model.

```
## Interpretation V1
#cor.test(as.numeric(data.cities$region), as.numeric(data.cities$province))
m.V1<-lm(V1 ~ region + total.rain.centered + season.max.rain , data=data.cities)
summary(m.V1)
```

```
##
## Call:
## lm(formula = V1 ~ region + total.rain.centered + season.max.rain,
##     data = data.cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2119 -3.0080  0.1298  2.8714  6.6555
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.685034   2.973879  -0.567 0.575492
## regionAtlantic          8.772788   3.381968   2.594 0.014924 *
## regionContinental       6.453013   2.901366   2.224 0.034379 *
## regionPacific          -4.819604   3.922978  -1.229 0.229464
## total.rain.centered    -0.011394   0.002871  -3.969 0.000457 ***
## season.max.rainautumn -13.236500   2.720577  -4.865 4.01e-05 ***
## season.max.rainwinter -20.061755   5.084461  -3.946 0.000486 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.452 on 28 degrees of freedom
## Multiple R-squared:  0.9039, Adjusted R-squared:  0.8833
## F-statistic: 43.91 on 6 and 28 DF,  p-value: 5.673e-13
```

```
#the residuals are normally distributed
#qqnorm(m.V1$residuals)
#qqline(m.V1$residuals)
```

The regression results show a highly significant ($p<0.001$) negative relationship between Z1 and total yearly rainfall. Pr. Rupert has a very high total yearly rainfall which results in a very low Z1 coordinate. Cities which have their highest rainfall in autumn and winter (eg Pr. Rupert, Vancouver, Victoria,…) will have a significantly ($p<0.001$) lower Z1 coordinate compared to cities which have their highest rainfall in summer (eg London, Thunder Bay) . Regions "Atlantic" and "Pacific" show a very low significant relationship ($p<0.05$) with Z1 which disappears after a Bonferroni correction. We verified visually (by a qqplot) that the residuals follow a normal distribution.

Next we performed a regression on the dependent variable Z2 with the same independent variables as for Z1.

```
###interpretation V2
m.V2<-lm(V2 ~ region + total.rain.centered + season.max.rain , data=data.cities)
summary(m.V2)
```

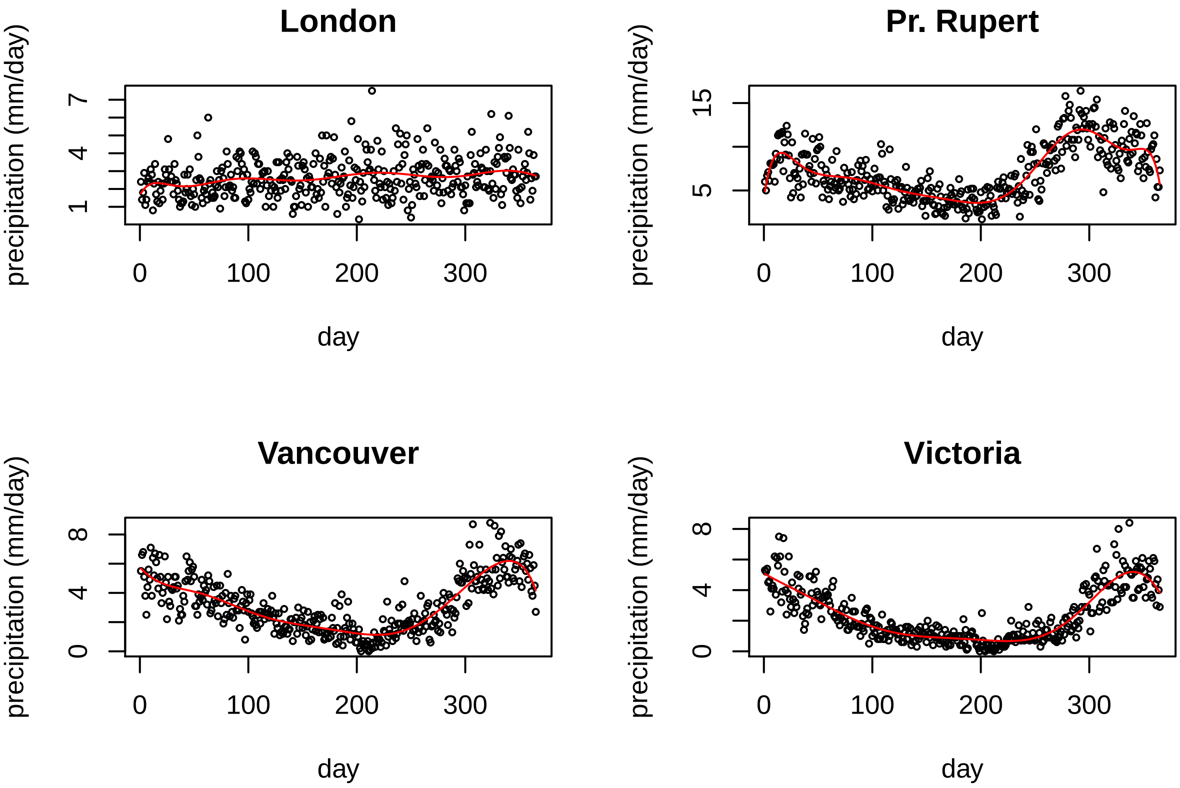
```
##
## Call:
## lm(formula = V2 ~ region + total.rain.centered + season.max.rain,
##     data = data.cities)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.537 -2.082  0.000  1.484  8.258
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.523434   2.165954   3.935 0.000500 ***
## regionAtlantic         -8.819674   2.463177  -3.581 0.001278 **
## regionContinental      -2.508765   2.113142  -1.187 0.245116
## regionPacific          -5.413601   2.857209  -1.895 0.068500 .
## total.rain.centered     0.014624   0.002091   6.994 1.32e-07 ***
## season.max.rainautumn -11.725256   1.981468  -5.917 2.29e-06 ***
## season.max.rainwinter -15.049850   3.703147  -4.064 0.000354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.242 on 28 degrees of freedom
## Multiple R-squared:  0.7059, Adjusted R-squared:  0.6429
## F-statistic:  11.2 on 6 and 28 DF,  p-value: 2.285e-06
```

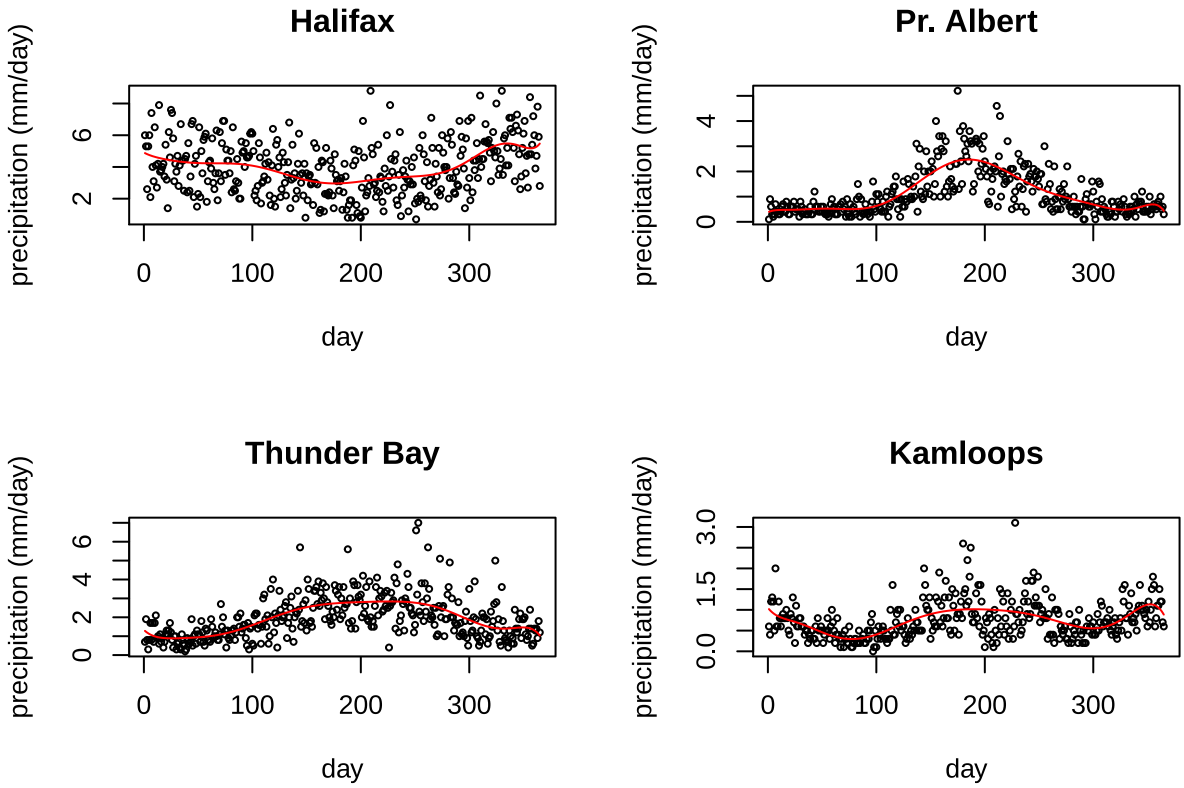
```
#the residuals are normally distributed
#qqnorm(m.V2$residuals)
#qqline(m.V2$residuals)
```

The regression results show a highly significant (p<0.001) positive relationship between Z2 and total yearly rainfall. Pr. Rupert has a very high total yearly rainfall which results in a very high Z2 coordinate. Cities which have their highest rainfall in autumn and winter (eg Pr. Rupert, Vancouver, Victoria,…) will have a sigificantly (p<0.001) lower Z2 coordinate compared to cities which have their highest rainfall in summer (eg London, Thunder Bay). "Atlantic" regions have a significantly (p<0.01) lower Z2 coordinate compared to "Arctic" cities. We verified visually (by a qqplot) that the residuals follow a normal distribution.

Next a selection of scatter plots of cities with "characteristical" yearly rainfall patterns is plotted which will be interpreted based on the biplot and the regression results. Below the plots, a table shows the total yearly rainfall (mm); the season with the highest rainfall; and the total rainfall (mm) for each season (spring, summer, autumn and winter) for each city.

```
# Selection of scatter plots of cities with "characteristical" yearly rainfall patterns
par(mfrow=c(2,2))
city_plot = c("London","Pr. Rupert", "Vancouver", "Victoria",
              "Halifax", "Pr. Albert", "Thunder Bay", "Kamloops")
for (i in city_plot)
{
  m.city<-lm(da[,i]~phi)
  plot(1:365,da[,i], main=i, cex=0.5, xlab="day", ylab="precipitation (mm/day)")
  lines(1:365,m.city$fitted.values,type="l", col=2)
}
```


London
precipitation (mm/day)
1
4
7
0
100
200
300
day
Pr. Rupert
precipitation (mm/day)
5
15
0
100
200
300
day
Vancouver
precipitation (mm/day)
0
4
8
0
100
200
300
day
Victoria
precipitation (mm/day)
0
4
8
0
100
200
300
day



```
par(mfrow=c(1,1))
rm(i)
# informative table (for each city): total yearly rainfall (mm); season with the highest
# rainfall; and total rainfall (mm) for each season (spring, summer, autumn and winter)
data.cities2 <- data.cities[,c("total.rain","season.max.rain","total.spring",
                               "total.summer","total.autumn","total.winter")]
colnames(data.cities2) <- c("total.rain","max.season","spring","summer","autumn","winter")
data.cities2
```

```
##              total.rain max.season spring summer autumn winter
## St. Johns        1480.8     autumn  315.6  313.5  430.8  420.9
## Halifax          1455.0     autumn  330.6  290.0  424.0  410.4
## Sydney           1474.4     autumn  327.9  285.4  432.9  428.2
## Yarmouth         1262.5     autumn  289.8  249.4  373.7  349.6
## Charlottvl       1201.7     autumn  272.4  265.7  351.5  312.1
## Fredericton      1127.9     autumn  258.6  269.2  316.5  283.6
## Scheffervll       801.2     summer  166.6  279.4  224.7  130.5
## Arvida            896.6     summer  193.5  299.9  236.8  166.4
## Bagottville       931.4     summer  209.1  311.4  240.7  170.2
## Quebec           1208.6     summer  274.0  364.1  313.4  257.1
## Sherbrooke       1109.3     summer  258.6  346.6  280.8  223.3
## Montreal          940.8     summer  220.5  272.8  250.3  197.2
## Ottawa            912.7     summer  218.0  265.7  240.7  188.3
## Toronto           782.6     summer  195.0  238.2  200.9  148.5
## London            958.0     summer  232.9  258.5  258.2  208.4
## Thunderbay        704.3     summer  191.3  259.9  159.7   93.4
## Winnipeg         509.2     summer  156.4  218.2   81.4   53.2
## The Pas          449.1     summer  103.4  205.2   92.9   47.6
## Churchill        408.8     summer   85.9  168.3  109.3   45.3
## Regina           371.1     summer  112.7  160.9   53.6   43.9
## Pr. Albert       406.6     summer  115.7  183.6   62.0   45.3
## Uranium Cty      362.7     summer   70.8  148.5   89.9   53.5
## Edmonton         465.2     summer  115.2  235.9   57.2   56.9
## Calgary          400.5     summer  136.6  183.8   46.8   33.3
## Kamloops         271.9     summer   60.0   89.7   64.3   57.9
## Vancouver       1155.1     autumn  209.1  132.1  417.1  396.8
## Victoria         851.6     winter  119.2   75.1  321.2  336.1
## Pr. George       608.9     summer  134.3  175.3  170.9  128.4
## Pr. Rupert      2591.8     autumn  464.6  469.6  981.1  676.5
## Whitehorse       271.7     summer   42.3  117.8   67.4   44.2
## Dawson           327.7     summer   61.9  130.9   82.3   52.6
## Yellowknife      268.1     summer   46.0  104.3   77.2   40.6
## Iqaluit          414.5     summer   86.4  171.6   96.8   59.7
## Inuvik           260.1     summer   47.7  105.7   66.3   40.4
## Resolute         144.0     summer   26.9   74.1   30.9   12.1
```

First a scatterplot of London is shown. London has an average total yearly rainfall compared to the other cities. It’s not clearly visible which season has a maximal total rainfall. There are no clear visible peaks in the scatterplot.

Vancouver and Victoria have a maximal total rainfall in autumn and winter respectively. Our regression results indicated this results to lower Z1 and lower Z2 coordinates compared to cities which have a maximal total rainfall in summer. Vancouver and Victoria show similar rainfall scatterplots. In the biplot they are located close to each other.

Apparently all other cities “along” V-tilda “2” have a maximal total rainfall in autumn. In the biplot Halifax is situated between London and Vancouver/Victoria. Halifax has a maximal total rainfall in autumn and has lower Z1 and lower Z2 coordinates compared to London which has a maximal total rainfall in summer. However the peak for the maximal total rainfall in autumn is less pronounced (compared to Vancouver and Victoria). Therefore Halifax is situated in between London and Vancouver/Victoria. Similar results for Sydney, Yarmouth, St. Johns, Fredericton, Charlotville which have a maximal total rainfall in autumn and are along the V-tilda “2”.

Pr. Rupert also has a maximal total rainfall in autumn. This would normally result into lower Z1 and lower Z2 coordinates compared to London which has maximal total rainfall in summer. However Pr. Rupert seems to have a very high Z2. This is because Pr. Rupert has a very high total yearly rainfall which results in a very low Z1 coordinate and a very high Z2 coordinate. The positive impact on Z2 because of a very high total yearly rainfall seems to dominate the negative impact on Z2 because of a maximal total rainfall in autumn, which results in a very high Z2 coordinate for Pr. Rupert.

Pr. Albert (and Winnipeg, Regina, Edmonton) have a low total yearly rainfall which results in a quite high Z1 coordinate and a quite low Z2 coordinate. They all have a clear maximum total rainfall in summer. They are all located close to each other in the biplot.

Thunderbay (and cities close to Thunderbay in the biplot) also have a maximum total rainfall in summer like Pr. Albert, but their their maximum is less pronounced compared to Pr. Albert. Also their total yearly rainfall is higher compared to Pr. Albert.

Finally a scatterplot of Kamloops is shown. Kamloops has a low total yearly rainfall and the scatterplot shows multiple peaks. It’s not clearly visible which season has a maximal total rainfall. In the biplot Kamloops is located somewhat away from to the other cities with low total yearly rainfall like Pr. Albert.

3. Conclusion

We can conclude cities located closely to each other on the biplot show similar rainfall patterns. Cities located far away from each other on the biplot show dissimilar rainfall patterns. These (dis)similar patterns can be explained in terms of total yearly rainfall and maximum total rainfall in a specific season.