Assignment: Canadian Weather - PMHD - Group 6

The data

The data is loaded.

```
load("CanadianWeather.rda")
da<-CanadianWeather[[1]]
da<-da[,,"Precipitation.mm"] # precipitation data
head(da)
```

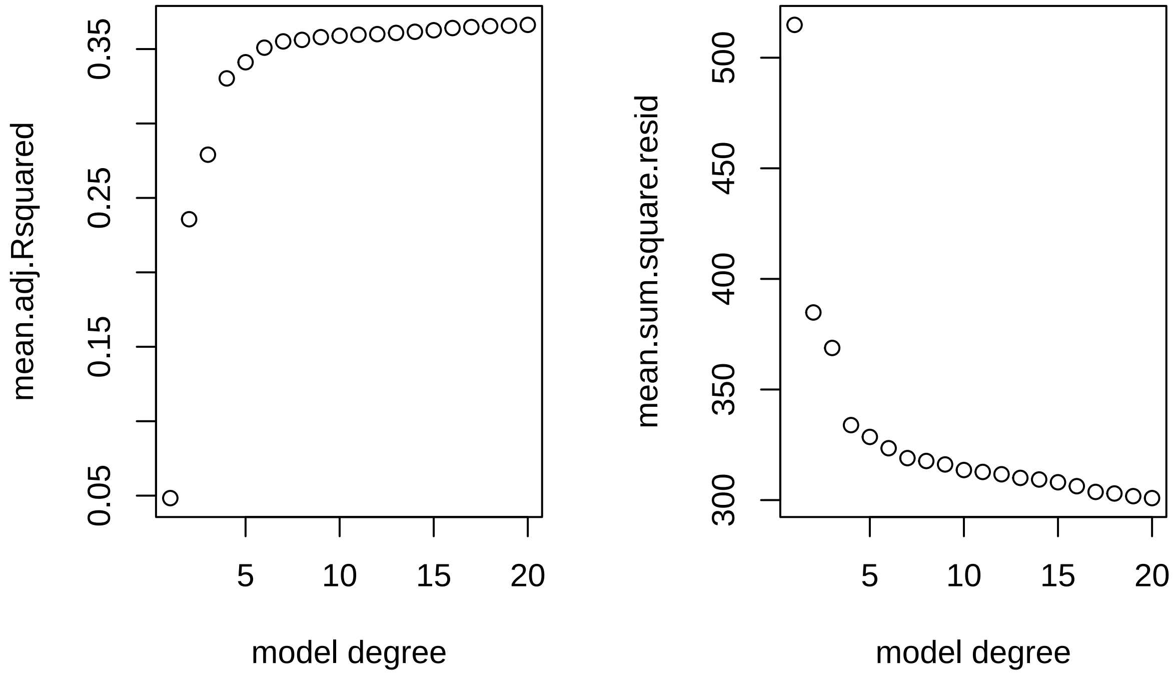
```
##       St. Johns Halifax Sydney Yarmouth Charlottvl Fredericton Scheffervll
## jan01       5.2     6.0    5.3      5.6        4.6         4.0         1.1
## jan02       5.8     5.3    5.2      3.7        4.4         3.2         1.3
## jan03       3.9     2.6    2.1      2.8        2.3         3.3         1.2
## jan04       4.3     5.3    5.0      5.3        4.8         3.3         1.3
## jan05       6.2     6.0    7.3      3.8        5.1         2.7         1.0
## jan06       3.4     2.1    2.2      2.4        1.5         0.8         1.3
##       Arvida Bagottville Quebec Sherbrooke Montreal Ottawa Toronto London
## jan01    2.6         3.0    4.1        2.9      2.9    2.5     1.8    2.4
## jan02    1.2         1.8    2.3        2.9      1.2    1.1     0.9    1.4
## jan03    2.1         1.3    2.6        1.9      1.4    1.3     0.9    1.8
## jan04    2.3         2.5    4.3        2.9      3.6    3.1     1.5    2.9
## jan05    1.7         2.1    2.3        2.1      1.6    1.3     0.8    1.1
## jan06    2.0         1.6    1.5        0.8      1.1    1.3     1.0    1.4
##       Thunder Bay Winnipeg The Pas Churchill Regina Pr. Albert Uranium City
## jan01         0.7      0.5     0.5       0.5    0.2        0.1          0.3
## jan02         1.9      0.6     0.9       0.6    0.3        0.9          0.4
## jan03         0.8      0.3     0.7       0.5    0.6        0.6          1.3
## jan04         0.3      0.5     0.5       0.4    0.3        0.3          0.6
## jan05         0.8      0.4     0.2       0.4    0.8        0.2          0.8
## jan06         1.7      0.7     0.9       0.2    0.5        0.3          1.1
##       Edmonton Calgary Kamloops Vancouver Victoria Pr. George Pr. Rupert
## jan01      0.4     0.3      0.6       5.5      5.3        2.2        6.0
## jan02      0.8     0.1      0.4       6.6      5.2        1.9        5.0
## jan03      1.1     0.3      1.2       6.8      5.4        1.9        6.7
## jan04      1.1     0.6      1.3       5.1      4.5        1.8        7.1
## jan05      1.0     1.0      1.2       3.8      4.6        1.1        6.1
## jan06      0.8     0.2      0.5       2.5      2.6        1.2        8.1
##       Whitehorse Dawson Yellowknife Iqaluit Inuvik Resolute
## jan01        0.5    0.9         0.6     1.1    0.8      0.1
## jan02        0.8    0.6         0.7     0.9    0.9      0.1
## jan03        1.1    0.8         0.3     0.8    0.8      0.0
## jan04        0.2    0.8         0.5     0.7    0.4      0.2
## jan05        0.6    1.0         0.7     0.9    0.8      0.2
## jan06        0.7    1.0         0.5     0.2    0.4      0.2
```

```
days<-1:365
days.range<-diff(range(days))
days<-(days-min(days))/days.range # rescaling to [0,1]
city.names <- colnames(da)
n.cities<-ncol(da)
MetaData<-data.frame(city=colnames(da), region=CanadianWeather$region,
                     province=CanadianWeather$province)
head(MetaData)
```

```
##                    city   region      province
## St. Johns     St. Johns Atlantic  Newfoundland
## Halifax         Halifax Atlantic   Nova Scotia
## Sydney           Sydney Atlantic   Nova Scotia
## Yarmouth       Yarmouth Atlantic   Nova Scotia
## Charlottvl   Charlottvl Atlantic       Ontario
## Fredericton Fredericton Atlantic New Brunswick
```

A polynomial fitting is performed. But first,the optimal degree d for the polynomial function is determined.

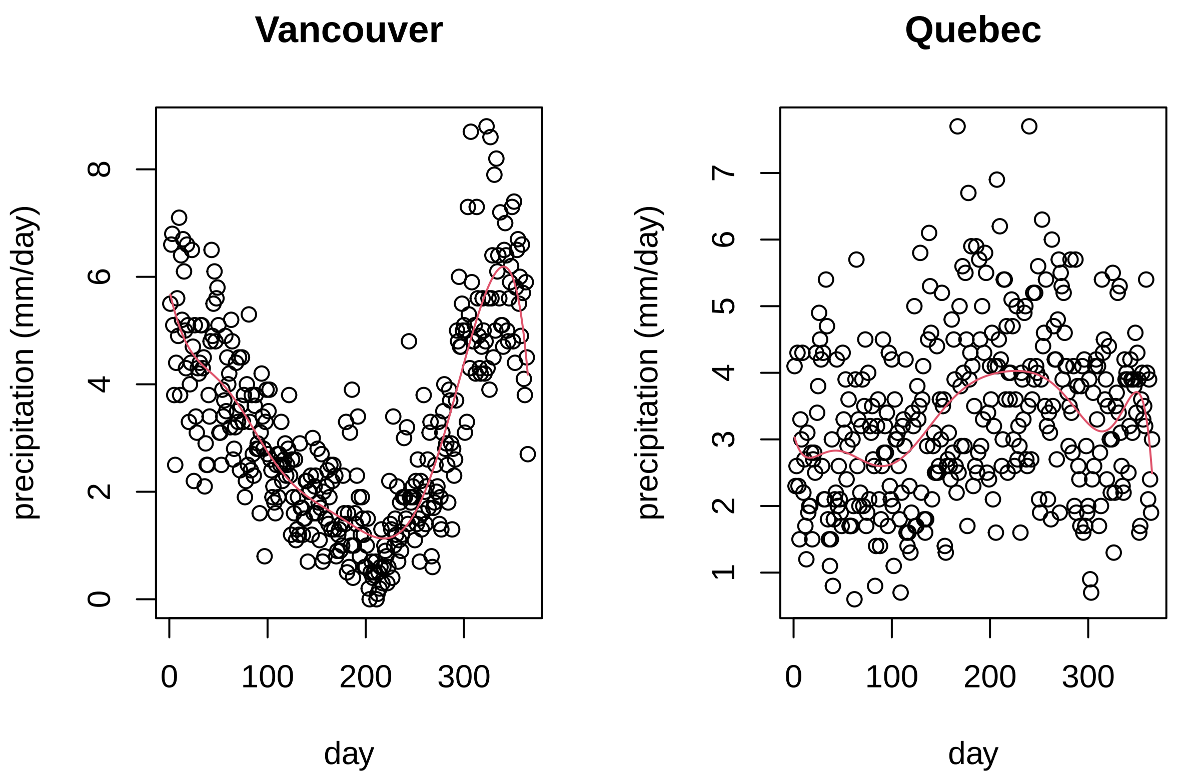
```
# Choose optimal m
max.d <- 20
mean.adj.Rsquared <- rep(0,max.d)
mean.sum.square.resid <- rep(0,max.d)
for (d in 1:max.d){
  cities.adj.Rsquared <- rep(0,length(city.names))
  sum.square.resid <- rep(0,length(city.names))
  phi<-poly(days,degree=d)
  for (city in 1:length(city.names)){
    m <- lm(da[,city.names[city]]~phi)
    cities.adj.Rsquared[city] <- summary(m)$adj.r.squared
    sum.square.resid[city] <- deviance(m)
  }
  mean.adj.Rsquared[d] <- mean(cities.adj.Rsquared)
  mean.sum.square.resid[d] <- mean(sum.square.resid)
}
par(mfrow=c(1,2))
plot(1:max.d,mean.adj.Rsquared,xlab = "model degree")
plot(1:max.d,mean.sum.square.resid,xlab = "model degree")
```



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

We choose to work with a degree of 10. We plot the fit for Vancouver and Quebec

```
phi<-poly(days,degree=10)
par(mfrow=c(1,2))
# estimation of the theta parameters for Vancouver
m.Vancouver<- lm(da[,'Vancouver']~phi)
# plot of fitted function
plot(1:365,da[,'Vancouver'],main="Vancouver", xlab="day", ylab="precipitation (mm/day)")
lines(1:365,m.Vancouver$fitted.values,type="l", col=2)

# estimation of the theta parameters for Quebec
m.Quebec<- lm(da[,'Quebec']~phi)
# plot of fitted function
plot(1:365,da[,'Quebec'],main="Quebec", xlab="day", ylab="precipitation (mm/day)")
lines(1:365,m.Quebec$fitted.values,type="l", col=2)
```



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

Then the matrix with the cities and the corresponding parameters is generated.

```r
d<-10
parameters <- data.frame(matrix(0,length(city.names),d+1))
for (city in 1:length(city.names)){
  m <- lm(da[,city.names[city]]~phi)
  parameters[city,] <- m$coefficients
}
colnames(parameters)<- attr(m$coefficients, "names")
rownames(parameters)<-city.names
dim(parameters)
```

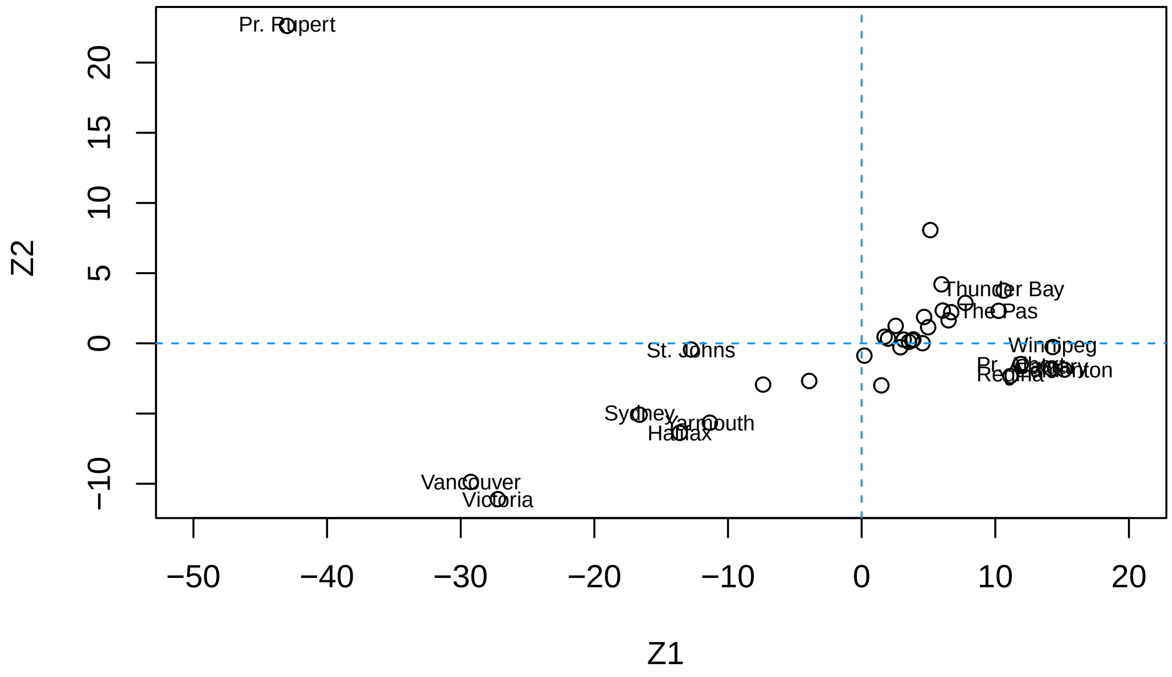
```
## [1] 35 11
```

The MSD is performed. First we apply the column centering on the parameters matrix, then we obtain the SVD of this matrix.

```r
# column centring of the matrix and applying SVD
parameters.mean<-colMeans(parameters)
parameters<-scale(parameters,center = TRUE, scale = FALSE)
parameters.svd<-svd(parameters)
```

Here we construct the scores matrix Zk with k=2, then we plot these scores (each city has a score).

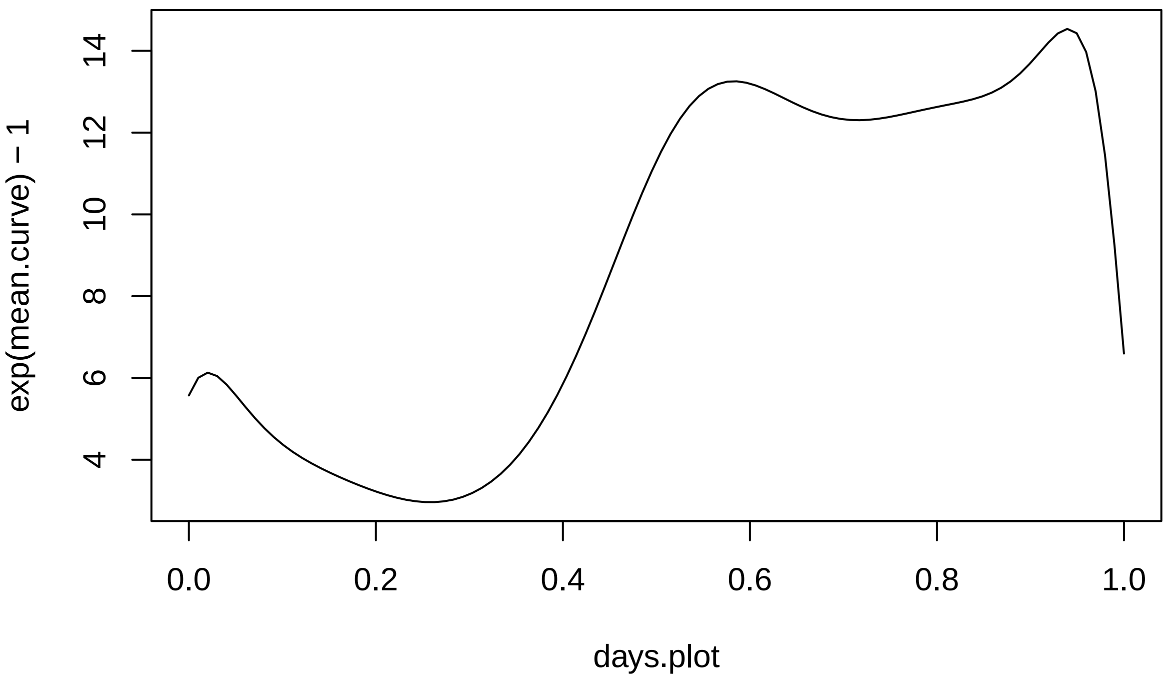
```
k <-2
Uk <-parameters.svd$u[ ,1:k]
Dk <- diag(parameters.svd$d [1:k])
Zk <-Uk%*%Dk
rownames(Zk)<- rownames(parameters)

plot (Zk , xlab =" Z1", ylab =" Z2",xlim=c(-50,20))
ind.label<-which(abs(Zk[,1])>10)
text(Zk[ind.label,1],Zk[ind.label,2],rownames(Zk)[ind.label],cex =0.7,)
abline(v=0,lty=2,col=4)
abline(h=0,lty=2,col=4)
```



from this graph it can be seen that the origin ((0, 0)) corresponds to the average precipitation/day function. The plot shows also that there are some cities with positive scores in the first dimension(e.g. Edmonton,Winnipeg, and Scheffervll). The city of Pr. Rupert has a large positive score in the second dimension but a large negative score in the first dimension. To better understand what large scores mean, we will back transform the SVD to the original function space.

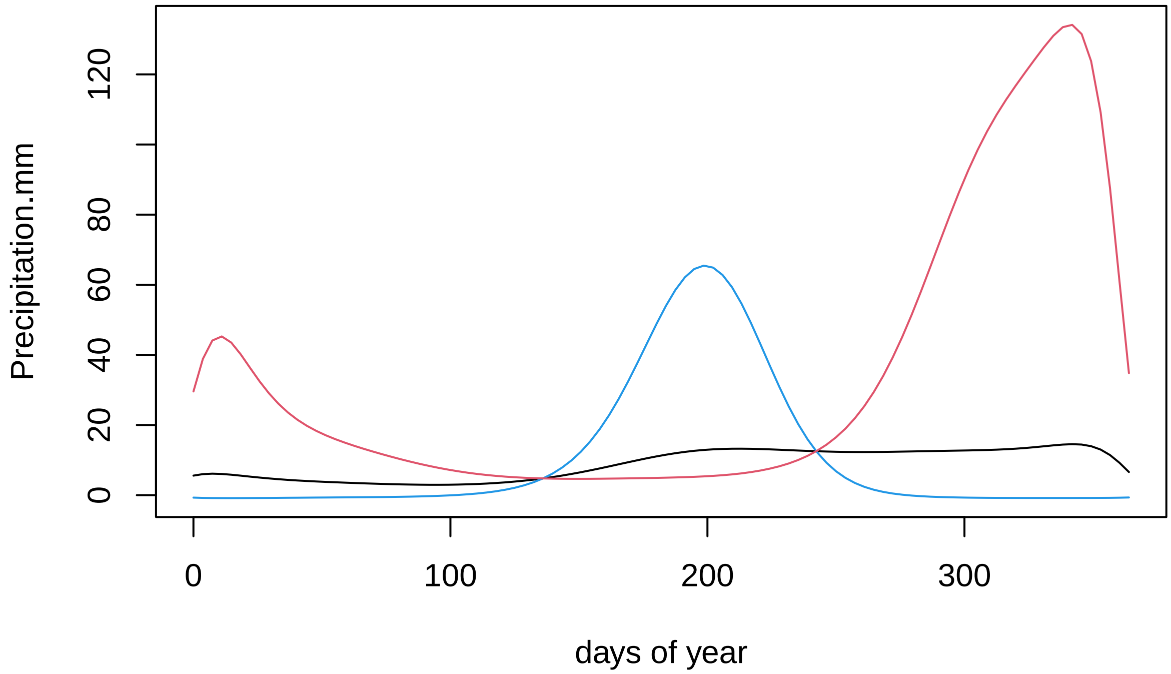
```
# right signular vectors (loadings)
V<-parameters.svd$v[,1:2]
# a vector with many points in the [0,1] interval.
days.plot<-seq(0,1,length.out = 100)
# evaluate the polynomial basis functions at all these points
phi.plot<-poly(days.plot,degree=10)
# construct the design matrix for the linear model
X<-cbind(1,phi.plot)
```

```
# product of this X matrix with the V matrix
XV<-X %*% V
# X times the vector with the column means of the
# original parameters matrix gives an estimate of the
# average precipitation/day function.
mean.curve<-X%*%parameters.mean
plot(days.plot,exp(mean.curve)-1, type="l")
```



From the score plot we see that the scores in the first dimension vary between −40 and +20. Now we illustrated how changing the z-score in the first dimension will show the effect of this score and how it varies from the average curve.

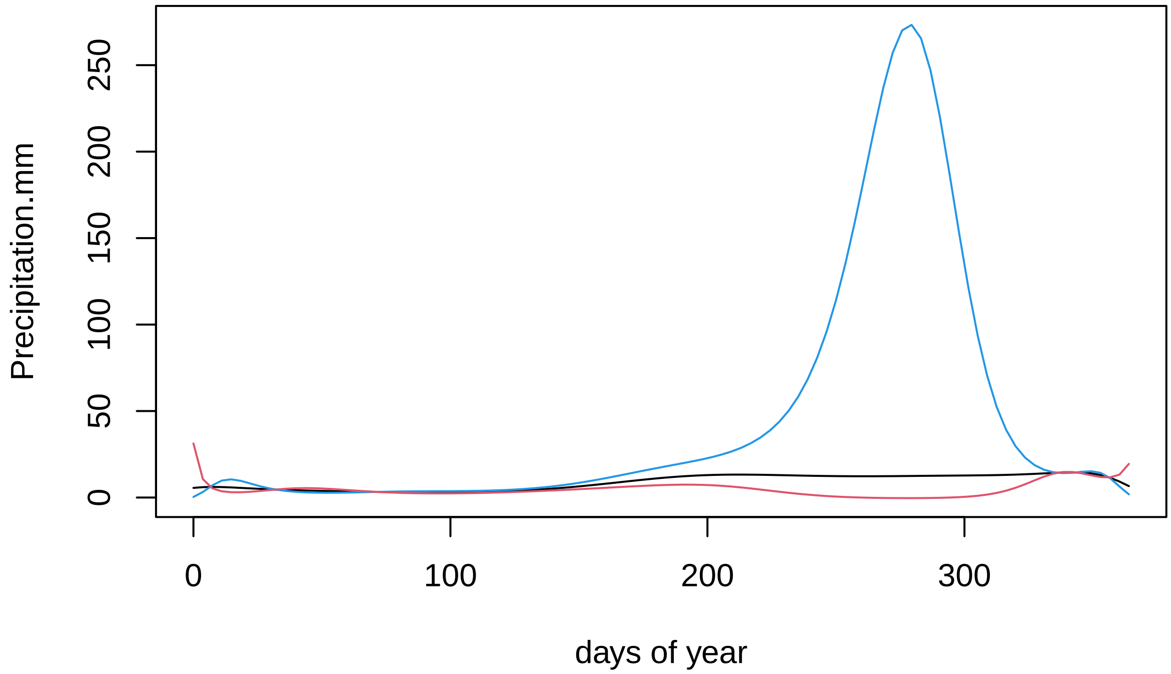
```
rng<-exp(range(mean.curve-XV[,1]*10,mean.curve+XV[,1]*20))-1
plot(days.plot*days.range,exp(mean.curve)-1,type="l",ylim=rng,
     xlab="days of year", ylab="Precipitation.mm")
lines(days.plot*days.range,exp(mean.curve+XV[,1]*20)-1,col=4)
lines(days.plot*days.range,exp(mean.curve-XV[,1]*10)-1,col=2)
```



From the graph we may conclude that cities that have large negative score in the first dimension tends to have high precipitation amount in the period between the end of the year and the beginning of the new year. On the other hand, cities with positive score have their high precipitation in the middle of the year. Therefore, we may interpret the first dimension as a dimension related to the period of the highest precipitation amount for each city. looking at the score graph and combining it with this interpretation, we may conclude that particularly Pr. Rupert,Vancouver,Sydney have high precipitation at the end and the beginning of the year, while cities as Winnipeg, Pr.Alpert, and Regina have high precipitation in the middle of the year.

Now We repeat the procedure for the second dimension.

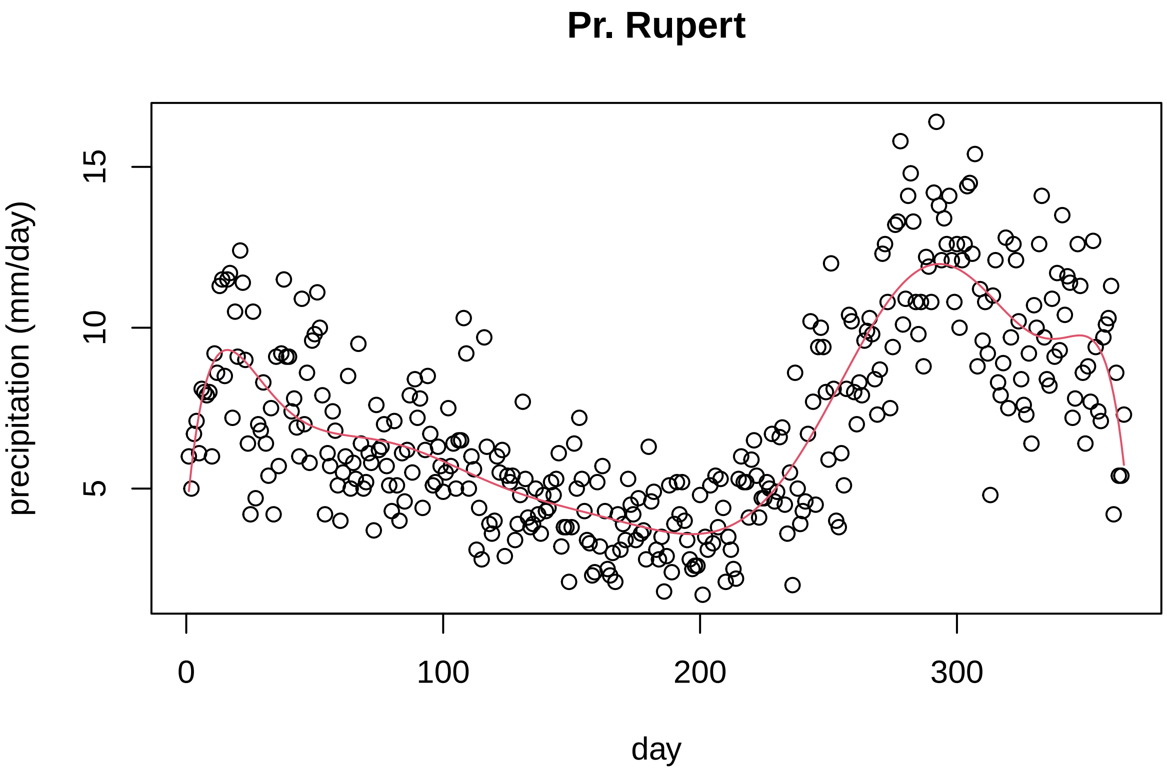
```
rng<-exp(range(mean.curve-XV[,2]*10,mean.curve+XV[,2]*10))-1
plot(days.plot*days.range,exp(mean.curve)-1,type="l",ylim=rng,
     xlab="days of year", ylab="Precipitation.mm")
lines(days.plot*days.range,exp(mean.curve+XV[,2]*10)-1,col=4)
lines(days.plot*days.range,exp(mean.curve-XV[,2]*10)-1,col=2)
```



The graph may allow us to conclude that cities with large positive scores in the second dimension, have high overall amount of precipitation. While cities with large negative scores have lower overall amount of precipitation. From the score plot we can see that Pr.Rupert have the largest positive score in the second dimension, and therefore it may has the highest overall amount of precipitation.

Plotting for Pr Rupert

```
m.PrRupert<- lm(da[,"Pr. Rupert"]~phi)
plot(1:365,da[,"Pr. Rupert"],main="Pr. Rupert", xlab="day", ylab="precipitation (mm/day)")
lines(1:365,m.PrRupert$fitted.values,type="l", col=2)
```


Pr. Rupert
precipitation (mm/day)
5
10
15
0
100
200
300
day