P-MHD: Multivariate Methods

Dam, Nsoh Tanih(1850242) | Juachon, Maria Joanna (1849785)|Kamau, Njeri (1747614)|Kirezi, Beatrice|Okafor, Chinenye Innocent(185739)

5 June 2019

Overview

Because of its great geographical extent, Canada has a wide variety of climates, and this can be readily seen in differences in precipitation patterns across the country. The objective of this project is to determine the differences and similarities in these patterns of selected cities across Canada (e.g. is there overall more rainfall in Western Canada, or less rain in summer?) The data available contains the average daily rainfall (mm/day) for the 365 days in the year for 35 cities. Extra information about the cities, the region they belong to and their coordinates is also available and is looked into.

Data Collection

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

```
str(da)

##  num [1:365, 1:35] 5.2 5.8 3.9 4.3 6.2 3.4 3.7 7.1 4.9 3.6 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:365] "jan01" "jan02" "jan03" "jan04" ...
##   ..$ : chr [1:35] "St. Johns" "Halifax" "Sydney" "Yarmouth" ...
```

The data set also contains extra information about the cities. For example, the regions, provinces and the coordinates are included.

```
str(MetaData)

## 'data.frame':    35 obs. of  5 variables:
##  $ city             : Factor w/ 35 levels "Arvida","Bagottville",..: 24 9
25 34 4 8 22 1 2 19 ...
##  $ region           : Factor w/ 4 levels "Arctic","Atlantic",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ province         : Factor w/ 12 levels "Alberta","British Colombia",..:
5 7 7 7 9 4 10 10 10 10 ...
##  $ coord.N.latitude : num  47.3 44.4 46.1 43.5 42.5 ...
##  $ coord.W.longitude: num  52.4 63.4 60.1 66.1 80.2 ...
```

Data Preparation

Functional Data Analysis

Functional Data Analysis (FDA) is used to first convert the 365 data entries for each city to a single function. Thus each city will have its set of q parameter estimates, and thus an $35 \times q$ data matrix can be constructed. These parameter estimates form now the input for the MDS.

The statistical model for $Y_i(t_{ij})$,

$$Y_i(t_{ij}) = \sum_{k=0}^{m} \theta_{ik}\, \phi_k(t_{ij}) + \varepsilon_{ij}$$

or

$$Y_i(t_{ij}) = \sum_{k=0}^{m} \theta_{ik}\, x_{ijk} + \varepsilon_{ij}.$$

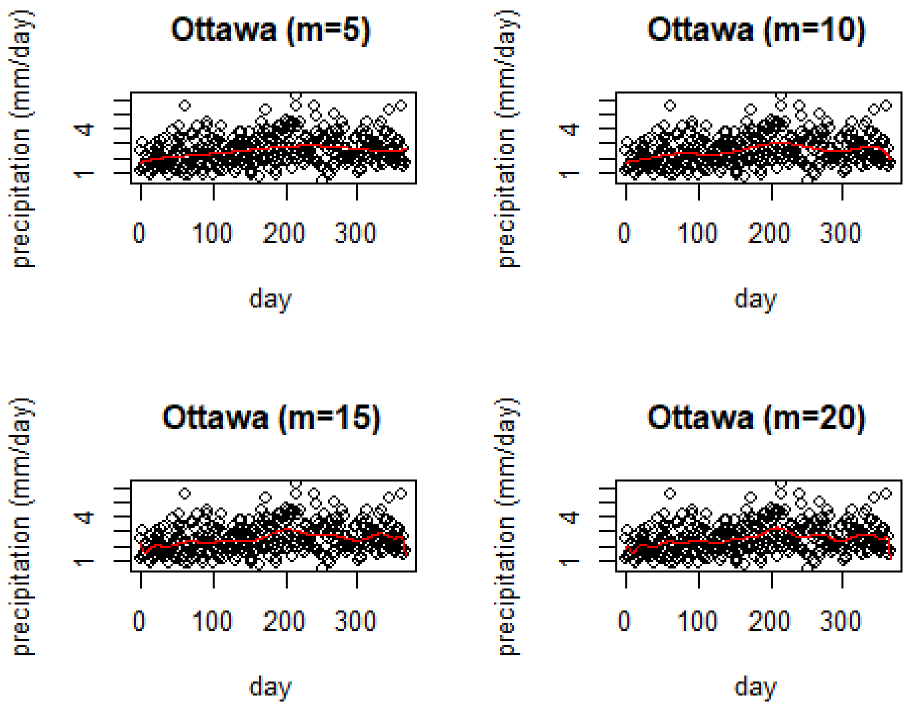
Let $Y_i(t)$ denote the outcome of observation $i = 1, \dots, 35$ (here: average daily rainfall) at time $t \in [1,365]$. For observation i we have data on times $t_{ij}, j = 1, \dots, p_i$. For a given i, this has the structure of a linear regression model with outcomes $Y_i(t_{ij}), j = 1, \dots, n$, and $q = m + 1$ regressors x_{ijk}.

The statistical model for city i in matrix notation is given as:

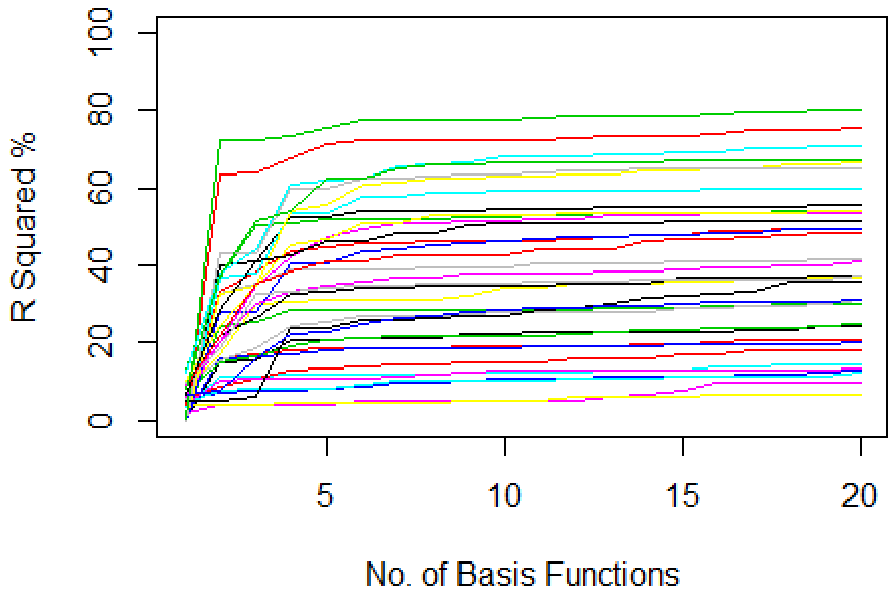
$$\mathbf{Y}_i = \boldsymbol{\theta}_i^t \mathbf{X}_i + \boldsymbol{\varepsilon}_i$$

with $\mathbf{Y}_i$ the vector with the outcomes of observation i (one for each day t_{ij}), $\boldsymbol{\theta}_i$ the vector with the θ_{ik} (one for each basis function k), $\mathbf{X}_i$ the matrix with the x_{ijk} (days j in the rows, basis function index k in columns), and $\boldsymbol{\varepsilon}_i$ the vector with the i.i.d. error terms.

The parameters θ_{ik} can be estimated by means of least squares and these can be used to plot the fitted function. One example is shown below, for Ottawa as the capital city of Canada.



The choice of m is based on the comparison of RSquared for each fitted function with m from 1 to 20, for each city. The largest m which shows significant contribution taking into account all of the cities is chosen. The plot below shows that there has steep/increase until m=18.



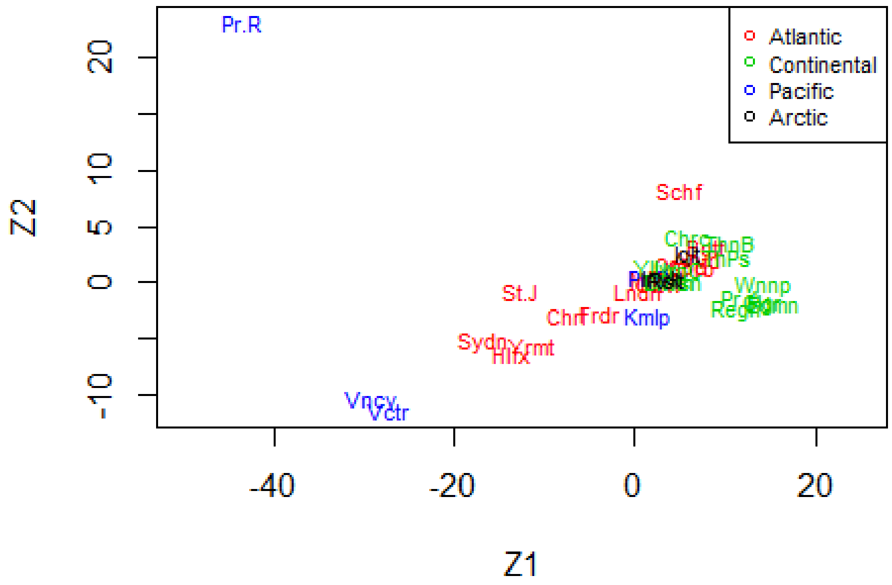
Multidimensional Scaling

The estimates for all cities are collected into a single new $35 \times (18 + 1)$ data matrix $\boldsymbol{\Theta}$ which contains all information on the shape of the precipitation functions. After column centering, the truncated SVD of $\boldsymbol{\Theta}$ is given as:

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

```r
n<-nrow(theta)
H<-diag(n)-(1/n)*matrix(1,ncol=n,nrow=n)
theta[,]<-H%*%as.matrix(theta)  #column centering
theta.svd<-svd(theta)
k<-2
Uk<-theta.svd$u[,1:k]
Dk<-diag(theta.svd$d[1:k])
Vk<-theta.svd$v[,1:k]
thetak<-Uk%*%Dk%*%t(Vk)
Zk<-Uk%*%Dk

plot(Zk,type="n",xlab="Z1",ylab="Z2", xlim=c(-50,25))
text(Zk,abbr,cex=0.7,col=as.integer(MetaData$region))
legend("topright",
legend=unique(MetaData$region),pch=1,col=unique(MetaData$region), cex=0.7)
```



Comparing the above plot to the geographical map of Canada, it can be generally seen that the patterns of cities on the precipitation MDS plot loosely correspond to their geographical locations. This is mostly evident by the fact that cities that are geographically closer together seem to be clustered more closely together, as shown by the regional color coding on the plot. However, there is some mixture within regions as well, mainly visible by clustering between coastal cities and those that are more inland. The first grouping of cities on the plot, those with lower scores on the Z1 axis, are mostly all coastal. These include Pr. Rupert, Victoria, Vancouver, Charlottesville, Yarmouth, Sydney, Halifax and St. John’s. The furthest cluster along the Z1 axis includes cities that are the furthest inland geographically, including Winnipeg, Calgary, Regina, Pr. Albert and Edmonton. This suggests that geographic factors play a major role in explaining precipitation patterns between the cities, specifically their proximity to the coastline. The cities that lie within the arctic region seem to display a similar rainfall pattern to the non-coastal cities, despite their proximity to a coastline. This is to be expected however, as this region is known to exhibit low precipitation patterns. The variation on the Z2 axis is fairly minimal. However, it can be seen that the closer to zero a city is on this axis, the more inland it is (farther from the North Pacific Ocean), and thus has lower precipitation. In general, cities with lower Z1 scores and very high or very low Z2 scores are located along the pacific coastline in Western canada. These cities seem to experience the highest precipitation among all cities in Canada all year round.

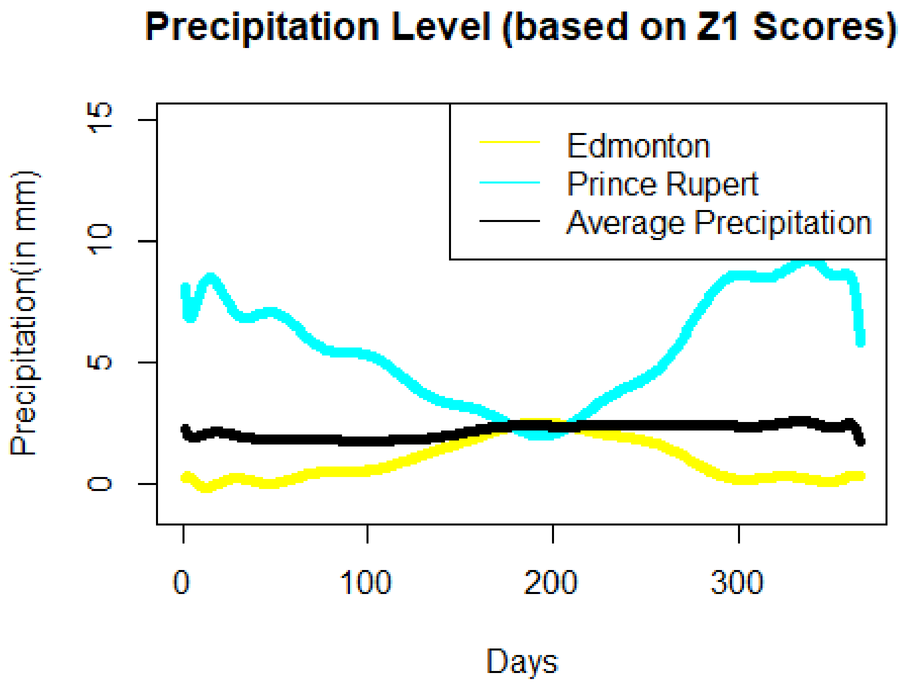
Biplots

After substituting $\boldsymbol{\Theta}$ with its truncated SVD (after k terms) and adding column means as part of backtransformation, we get the model fit:

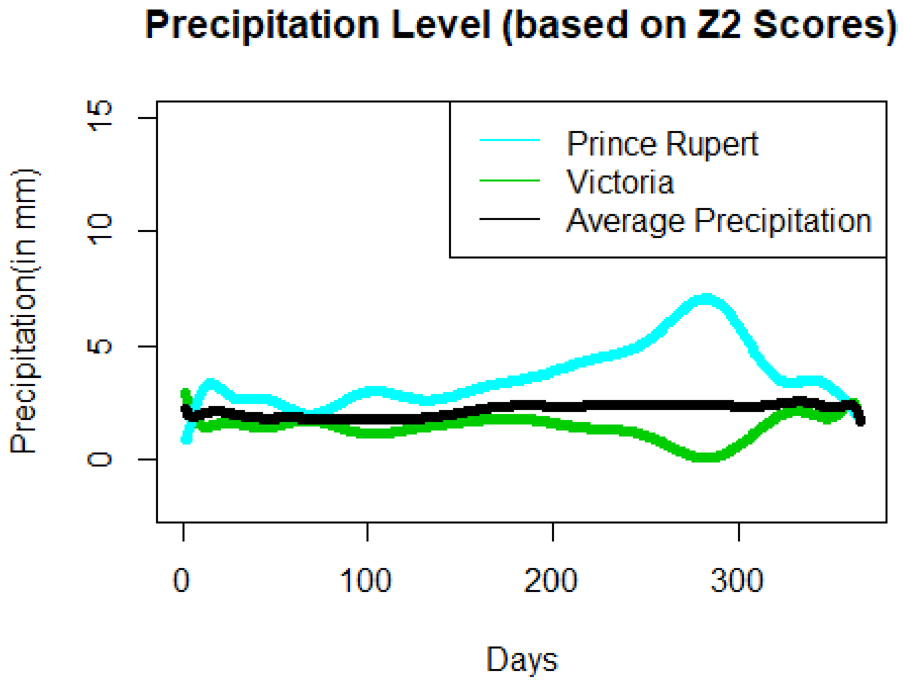
$$\hat{\mathbf{Y}}_k = \mathbf{\Theta}_k \mathbf{X}^t + \bar{Y} = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^t \mathbf{X}^t + \bar{Y} = \mathbf{Z}_k \mathbf{V}_k^t \mathbf{X}^t + \bar{Y}.$$

The latter expression relates an approximate model fit ($\hat{\mathbf{Y}}_k$) for the complete data set of 35 cities to the scores of the SVD ($\mathbf{Z}_k$) and the loadings ($\mathbf{V}_k$). The relation is established through the matrix $\mathbf{X}$ which was used to relate the basis functions to the θ-parameters.

```
t<-as.data.frame(t(apply(theta,2,mean))) #mean of original theta 1x(18+1)
matrix
x_orig<-as.data.frame(cbind((matrix(1,nrow=365,ncol=1)),as.matrix(phi)))
#Xphi with intercept 365x(18+1) matrix
tX<-as.matrix(t)%*%t(as.matrix(x_orig)) #newmean data to replace ybar for
smooth curve
Zmax1<-(max(Zk[,1])) %*% t(Vk[,1]) %*% t(x_orig)+tX #Z1 is at maximum
Zmin1<-(min(Zk[,1])) %*% t(Vk[,1]) %*% t(x_orig)+tX #Z1 is at minimum
Zmax2<-(max(Zk[,2])) %*% t(Vk[,2]) %*% t(x_orig)+tX #Z2 is at maximum
Zmin2<-(min(Zk[,2])) %*% t(Vk[,2]) %*% t(x_orig)+tX #Z2 is at minimum
#Plot function
plot(days, tX,type='l', ylim=c(-1,15),ylab="Precipitation(in mm)",
xlab="Days", main="Precipitation Level (based on Z1 Scores)")
lines(days, Zmaxi, col=j) #color j refers to the jth city with maximum Zi
score, change i and j
lines(days, Zmini), col=k) #color k refers to the kth city with maximum Zi
score, change i and j
```



The above plot seems to show that Z1 picks up the differences in variability of precipitation between the different seasons of the year. In general, cities with high Z1 scores have low precipitation. Particularly, we see the highest variability as well as the highest levels of precipitation in the winter months, and very little variability during summer. Prince Rupert shows the greatest changes in precipitation over the year, with levels peaking over the winter season and falling just below average in summer. Edmonton shows a more consistent pattern across the year and is closer to average levels, however it exhibits slightly lower than average precipitation during winter.



Based on the plot above, the Z2 axis shows the most amount of variability in precipitation around the end of summer and beginning of the autumn season. Around this time the two cities shown seem to exhibit opposing rainfall patterns. Prince Rupert shows the highest level of precipitation around the month of September, which is quite above average. However, Victoria's rainfall dips at this point and falls below the average level.

As can be seen on the MDS plot shown previously, cities in the arctic region of the country have Z1 and Z2 scores close to zero. This means that they exhibit very little variation in precipitation across the year. Cities that are more inland show lower levels of rainfall in general. Further, Prince Rupert experiences significant seasonal variation in precipitation over the course of the year, and cities along the western pacific coastline exhibit the greatest amount of rainfall in the country.