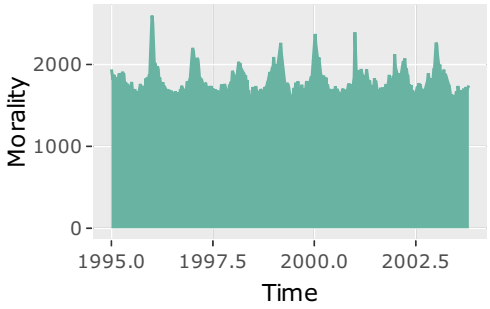
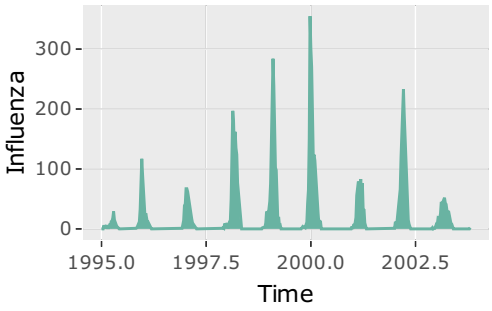
# Machine learning lab2 block2

Karthikeyan Devarajan- Karde799

## Assignment 1: Using GAM and GLM to examine the mortality rates.

### 1. Time Series Plot



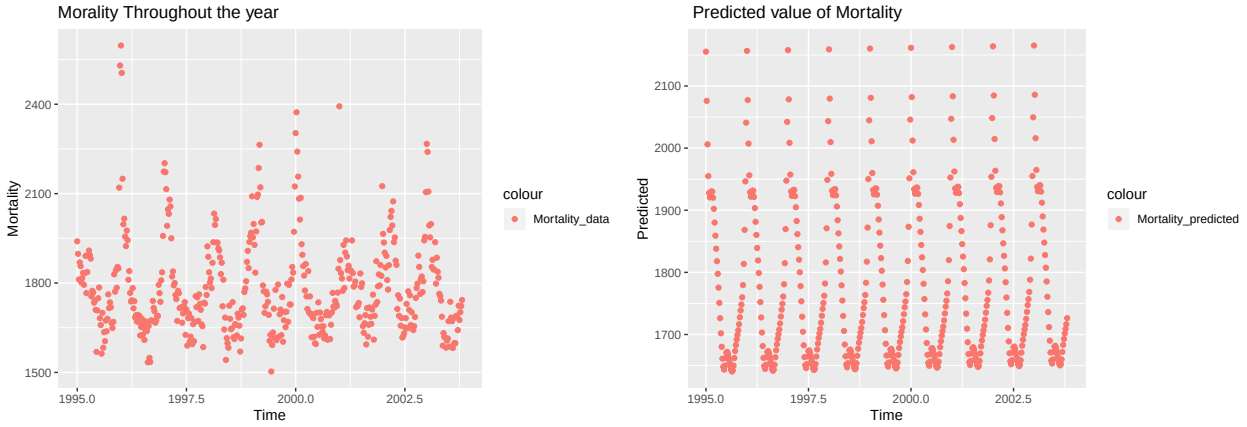



The Mortality and Influenza are influenced at the same time period. The two variables are increased at the same time period.

### 2&3.GAM Model Analysis

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = 52)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -680.598   3367.760  -0.202    0.840
## Year           1.233      1.685   0.732    0.465
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Week) 14.32  17.87 53.86  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Rank: 52/53
## R-sq.(adj) =  0.677   Deviance explained = 68.8%
## GCV = 8708.6  Scale est. = 8398.9    n = 459
```

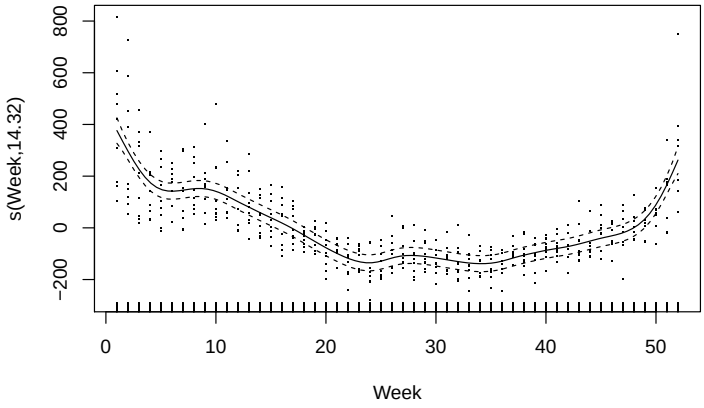



we know that the general probablistic model is $y = w_o + w_1 x_1 + s(x_2) + e$.

Probilistic model: $Mortality = -680.589 + 1.233 * Year + s(Week) + e$

The Predicted value graph is similar to the original values.Therefore, It can be said that this model is good approximation of the mortality. The increase in complexity of the spline function of variables in the model will increase the accuracy of the model. While adding the complexity of the model, the spline function should be significant on the Mortality. The p-value for the factor week is less than alpha=0.001. The variable week is significant for Mortality.

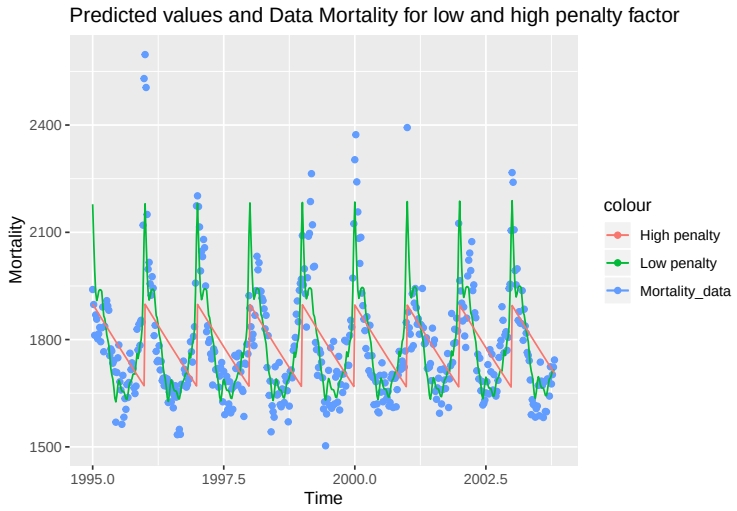
The range of Mortality value increases each year. In the starting years, the range is small and increases when year increases.



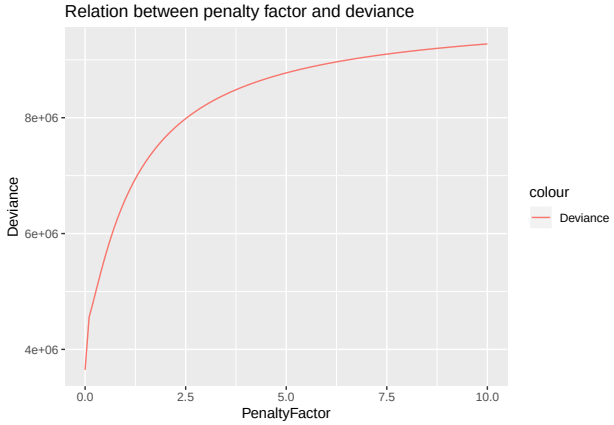

The rate of mortality is less in the middle of the year but in the rise during the initial and final weeks.
### 4. Influence of Penalty factor on spline function

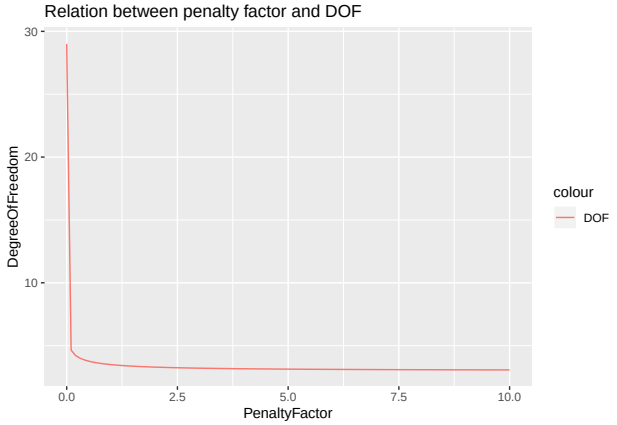
```
## The optimal penalty factor is: 0.0001131932
```

```
## The deviance at the optimal penalty factor is: 3718012
```



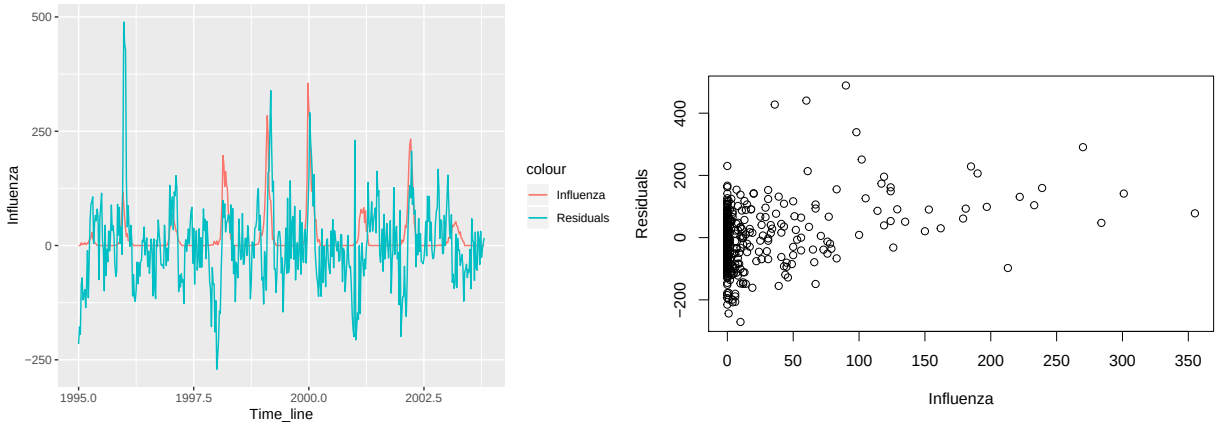

The low penalty factor i.e zero is overfitting the graph. From this, we explain that the lower penalty factor will overfit the data and minimize the error. The high penalty factor smoothen the graph. So, the high penalty factor will have more error value.




The deviance increases when the penalty factor increases wheres the degree of freedom decreases when penalty factor increases.

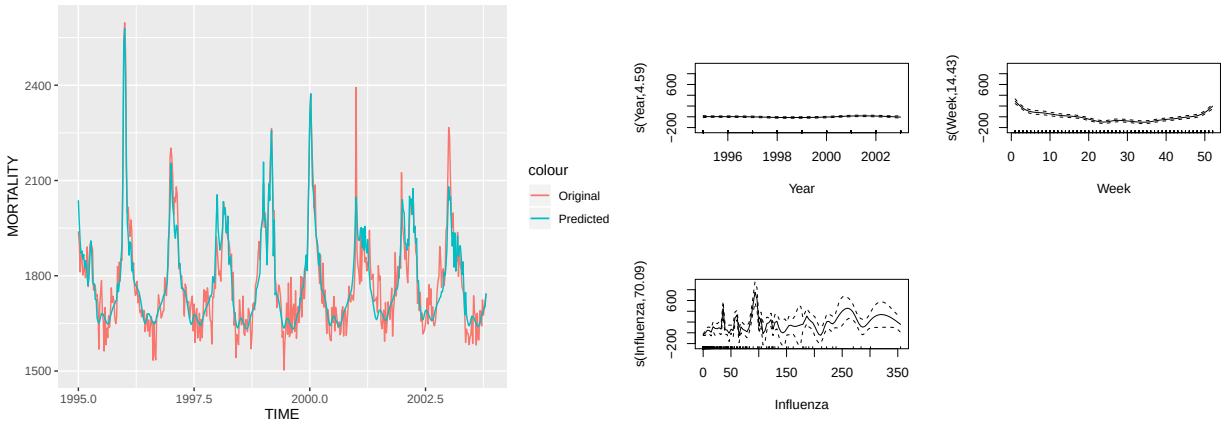
### 5. Relation between GAM and residuals




In plot 1, It can be concluded that whenever the Influenza was increasing, the residual was also increasing. The statement can be supported from plot 2.

### 6.GAM for multipe spline function

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Year, k = 9) + s(Week, k = 52) + s(Influenza, k = 85)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1783.765      3.198   557.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df      F p-value
## s(Year)        4.587  5.592  1.500   0.178
## s(Week)       14.431 17.990 18.763  <2e-16 ***
## s(Influenza) 70.094 72.998  5.622  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 134/144
## R-sq.(adj) =  0.819   Deviance explained = 85.4%
## GCV = 5840.5  Scale est. = 4693.7    n = 459
```




The year is not significant towards Morality but whereas week and influenza have significant influence over Morality.

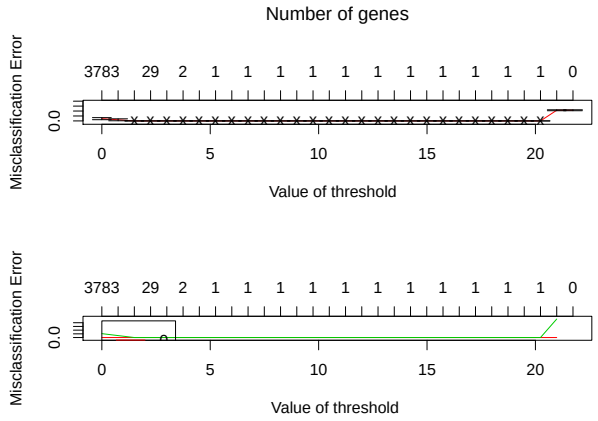
Probilistic model: $Mortality = 1783.77 + s(Year) + s(Week) + s(Influenza) + e$

## Assignment 2.High dimensional Methods

### Nearest shrunken Centroid

```
## 1234567891011121314151617181920212223242526272829 30
```

```
## 12Fold 1 :1234567891011121314151617181920212223242526272829 30
## Fold 2 :1234567891011121314151617181920212223242526272829 30
## Fold 3 :1234567891011121314151617181920212223242526272829 30
## Fold 4 :1234567891011121314151617181920212223242526272829 30
## Fold 5 :1234567891011121314151617181920212223242526272829 30
## Fold 6 :1234567891011121314151617181920212223242526272829 30
## Fold 7 :1234567891011121314151617181920212223242526272829 30
## Fold 8 :1234567891011121314151617181920212223242526272829 30
## Fold 9 :1234567891011121314151617181920212223242526272829 30
## Fold 10 :1234567891011121314151617181920212223242526272829 30
```




```
## 1
```

```
##        id   name           0-score 1-score
##   [1,] 869  conference     -2.6592 3.499
##   [2,] 3364 published      -0.2977 0.3917
##   [3,] 681  chairs         -0.2697 0.3548
##   [4,] 1891 held           -0.2697 0.3548
##   [5,] 3836 short          -0.2697 0.3548
##   [6,] 77   accepted       -0.254  0.3342
##   [7,] 1636 format         -0.254  0.3342
##   [8,] 680  chair          -0.223  0.2934
##   [9,] 389  authors        -0.2073 0.2727
##  [10,] 596  call           -0.1993 0.2622
##  [11,] 3285 proceedings    -0.1993 0.2622
##  [12,] 3324 proposals      -0.1792 0.2359
##  [13,] 3243 presented      -0.1635 0.2152
##  [14,] 3187 position       0.1603  -0.2109
##  [15,] 3036 papers         -0.1571 0.2067
##  [16,] 4628 workshop       -0.1554 0.2045
##  [17,] 810  committee      -0.1524 0.2005
##  [18,] 3022 pages          -0.1524 0.2005
##  [19,] 3323 proposal       -0.1375 0.1809
##  [20,] 3490 registration   -0.1375 0.1809
##  [21,] 4427 university     -0.1297 0.1707
##  [22,] 607  candidates     0.1294  -0.1703
##  [23,] 4282 topics         -0.128  0.1685
##  [24,] 1743 general        -0.1219 0.1604
##  [25,] 599  camera         -0.1139 0.1498
##  [26,] 3433 ready          -0.1139 0.1498
##  [27,] 3582 results        -0.1139 0.1498
##  [28,] 3188 positions      0.1011  -0.133
##  [29,] 4039 strong         0.1004  -0.1321
##  [30,] 2433 length         -0.097  0.1277
##  [31,] 3241 presentation   -0.097  0.1277
##  [32,] 4365 tutorials      -0.097  0.1277
##  [33,] 2175 international  -0.0855 0.1126
##  [34,] 4060 submission     -0.0855 0.1126
##  [35,] 2005 ideas          -0.0816 0.1073
##  [36,] 2177 internet       -0.0816 0.1073
##  [37,] 2984 organizers     -0.0816 0.1073
##  [38,] 3125 phd            0.078   -0.1027
##  [39,] 981  cross          -0.0739 0.0972
##  [40,] 3794 series         -0.0739 0.0972
##  [41,] 3383 qualifications 0.0734  -0.0966
##  [42,] 4177 team           0.0734  -0.0966
##  [43,] 3306 programming    0.0719  -0.0946
##  [44,] 2198 invited        -0.0708 0.0932
##  [45,] 2059 included       -0.057  0.075
##  [46,] 3242 presentations  -0.057  0.075
##  [47,] 4364 tutorial       -0.057  0.075
##  [48,] 879  conjunction    -0.057  0.075
##  [49,] 2487 lncs           -0.057  0.075
##  [50,] 2690 michael        -0.057  0.075
##  [51,] 2986 organizing     -0.057  0.075
##  [52,] 3216 practitioners  -0.057  0.075
##  [53,] 4606 wisconsin      -0.057  0.075
```

```
##  [54,] 663  centre         0.046   -0.0605
##  [55,] 2438 letter         0.046   -0.0605
##  [56,] 3191 post           0.046   -0.0605
##  [57,] 1477 excellent      0.046   -0.0605
##  [58,] 2442 levels         0.046   -0.0605
##  [59,] 2553 mail           0.044   -0.0579
##  [60,] 3671 salary         0.044   -0.0579
##  [61,] 3992 starting       0.044   -0.0579
##  [62,] 318  artificial     -0.0419 0.0552
##  [63,] 386  author         -0.0419 0.0552
##  [64,] 3040 parallel       -0.0419 0.0552
##  [65,] 3882 site           -0.0419 0.0552
##  [66,] 4451 usa            -0.0419 0.0552
##  [67,] 3301 program        -0.0395 0.0519
##  [68,] 1045 dates          -0.0351 0.0461
##  [69,] 1061 deadline       -0.0351 0.0461
##  [70,] 3035 paper          -0.0351 0.0461
##  [71,] 4064 submitted      -0.0351 0.0461
##  [72,] 4629 workshops      -0.0348 0.0458
##  [73,] 2150 intelligence   -0.0325 0.0428
##  [74,] 2889 notification   -0.0325 0.0428
##  [75,] 836  complex        0.0186  -0.0245
##  [76,] 1233 doctoral       0.0186  -0.0245
##  [77,] 336  assistant      0.0186  -0.0245
##  [78,] 2613 master         0.0186  -0.0245
##  [79,] 3559 researcher     0.0186  -0.0245
##  [80,] 1450 european       0.0186  -0.0245
##  [81,] 362  attendees      -0.0167 0.022
##  [82,] 1196 discuss        -0.0167 0.022
##  [83,] 1283 easychair      -0.0167 0.022
##  [84,] 1560 feature        -0.0167 0.022
##  [85,] 3055 participation  -0.0167 0.022
##  [86,] 3271 prior          -0.0167 0.022
##  [87,] 3514 relevance      -0.0167 0.022
##  [88,] 3681 san            -0.0167 0.022
##  [89,] 3816 share          -0.0167 0.022
##  [90,] 4002 steering       -0.0167 0.022
##  [91,] 4145 takes          -0.0167 0.022
##  [92,] 196  allowed        -0.0167 0.022
##  [93,] 603  canada         -0.0167 0.022
##  [94,] 1048 david          -0.0167 0.022
##  [95,] 1291 economics      -0.0167 0.022
##  [96,] 2296 journals       -0.0167 0.022
##  [97,] 2723 mit            -0.0167 0.022
##  [98,] 3361 publicity      -0.0167 0.022
##  [99,] 3051 participants   -0.0167 0.022
## [100,] 4202 template       -0.0167 0.022
## [101,] 3312 projects       0.0164  -0.0215
## [102,] 272  apply          0.0164  -0.0215
## [103,] 3705 scale          0.0164  -0.0215
## [104,] 2974 org            -0.0076 0.01
## [105,] 899  contact        0.0048  -0.0063
## [106,] 606  candidate      0.0041  -0.0053
## [107,] 708  chen           -0.0022 0.0029
```

```
## [108,] 817  community      -0.0022 0.0029
## [109,] 4281 topic          -0.0022 0.0029
## [110,] 939  copy           -0.0022 0.0029
## [111,] 3246 presenting     -0.0022 0.0029
## [112,] 3973 springer       -0.0022 0.0029
```

```
## The minimum Threshold value is: 1.498792
```

```
## Total Features Selected:  112
```

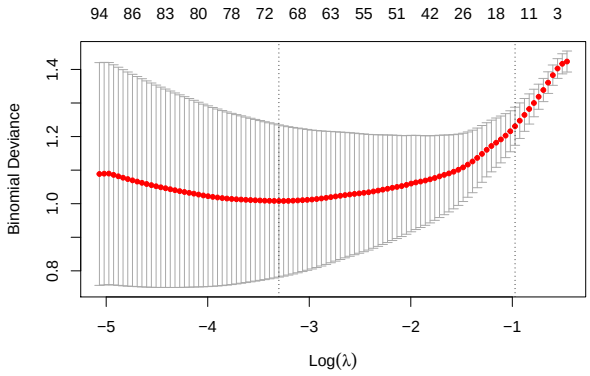
```
## Top 10 contributing features are:
##  conference published chairs held short accepted format chair authors call
```

```
## The confusion matrix is:
```

```
##      prediction1
## testy  0  1
##     0 11  0
##     1  0  9
```

```
## Misclassification rate is: 0
```

### Elastic Net with binomial response




```
## The confusion matrix is:
```

```
##      prediction2
## testy  0  1
##     0 10  1
##     1  1  8
```

```
## Misclassification rate is:  0.1
```

```
## Number of features selected:  100
```

**Support Vector Machine**

```
##  Setting default kernel parameters

## The confusion matrix is:

##       Predicted
## Actual  0  1
##      0 10  1
##      1  2  7

## Misclassification rate is:  0.15

## Number of feature selected:  43

## The comparison table is as follows:

##                       Model misClassificationrates FeaturesSelected
## 1 Nearest Shrunken Centroid                   0.00              112
## 2               Elastic Net                   0.10              100
## 3    Support Vector Machine                   0.15               43

## The Number of features selected are: 182

## The Selected parameters are listed below:

##     selected_Feature
## 1             papers
## 2         submission
## 3           position
## 4          published
## 5          important
## 6               call
## 7         conference
## 8         candidates
## 9              dates
## 10             paper
## 11            topics
## 12           limited
## 13         candidate
## 14            camera
## 15             ready
## 16           authors
## 17               phd
## 18          projects
## 19               org
## 20            chairs
## 21               due
## 22          original
## 23      notification
## 24            salary
## 25            record
```

```
## 26            skills
## 27              held
## 28              team
## 29             pages
## 30          workshop
## 31         committee
## 32       proceedings
## 33             apply
## 34            strong
## 35     international
## 36            degree
## 37         excellent
## 38              post
## 39         presented
## 40             march
## 41        applicants
## 42           privacy
## 43       submissions
## 44          deadline
## 45          doctoral
## 46            letter
## 47         positions
## 48    qualifications
## 49          february
## 50             forum
## 51         workshops
## 52           systems
## 53           aspects
## 54             chair
## 55            mobile
## 56           special
## 57         proposals
## 58               usa
## 59        experience
## 60          networks
## 61           science
## 62        curriculum
## 63            funded
## 64              java
## 65            levels
## 66          teaching
## 67           project
## 68             april
## 69            author
## 70             short
## 71          proposal
## 72         publicity
## 73         assistant
## 74           closing
## 75       competitive
## 76          european
## 77          graduate
## 78            master
## 79      universities
```

```
## 80            submit
## 81           invited
## 82               com
## 83           program
## 84          computer
## 85          security
## 86          starting
## 87           include
## 88          internet
## 89              peer
## 90            canada
## 91              grid
## 92        organizing
## 93     practitioners
## 94          tutorial
## 95          versions
## 96             equal
## 97      postdoctoral
## 98             vitae
## 99            format
## 100           series
## 101          general
## 102           issues
## 103          contact
## 104       successful
## 105          journal
## 106         services
## 107           france
## 108       organizers
## 109         reviewed
## 110             site
## 111         wireless
## 112        interests
## 113         students
## 114    undergraduate
## 115      programming
## 116             mail
## 117        economics
## 118  implementations
## 119           length
## 120     manuscripts
## 121          michael
## 122     presentation
## 123        relevance
## 124            spain
## 125        usability
## 126        wisconsin
## 127            smart
## 128         detailed
## 129         employer
## 130        extension
## 131     institutions
## 132              job
## 133        motivated
```

```
## 134         expected
## 135            ideas
## 136             june
## 137             page
## 138          results
## 139            issue
## 140     optimization
## 141         parallel
## 142    presentations
## 143        tutorials
## 144       ubiquitous
## 145          process
## 146      mathematics
## 147       researcher
## 148        statement
## 149      opportunity
## 150        professor
## 151            korea
## 152              non
## 153           poster
## 154        protocols
## 155             term
## 156     unpublished
## 157    visualization
## 158             yang
## 159      proficiency
## 160            start
## 161           making
## 162       university
## 163            topic
## 164         relevant
## 165         resource
## 166            scope
## 167            share
## 168            trust
## 169        technical
## 170              top
## 171           taiwan
## 172            takes
## 173         template
## 174           tracks
## 175       universite
## 176          version
## 177           vienna
## 178             wang
## 179            tasks
## 180           tenure
## 181           thesis
## 182            women
```

# Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(readxl)
library(ggplot2)
library(plotly)
library(dplyr)
library(hrbrthemes)
library(dygraphs)
library(mgcv)
library(pamr)
library(glmnet)
library(kernlab)

sf <- read_excel(file.choose())
sf1 <- read.csv(file.choose(),sep = ";")
M <- sf %>% ggplot(aes(x=Time, y=Mortality)) +
  geom_area(fill="#69b3a2", alpha=1) +
  geom_line(color="#69b3a2") +
  ylab("Morality")
ggplotly(M)
I <- sf %>% ggplot( aes(x=Time, y=Influenza)) +
  geom_area(fill="#69b3a2", alpha=1) +
  geom_line(color="#69b3a2") +
  ylab("Influenza")
ggplotly(I)
#2 & 3
model = gam(Mortality~Year+s(Week, k= 52), data = sf, family = gaussian, method = "GCV.Cp")
summary(model)
prediction <- predict(model,sf)
sf$Predicted <- prediction
ggplot(sf)+
  geom_point(aes(x=Time,y=Mortality,color="Mortality_data"))+
  ggtitle("Morality Throughout the year")

ggplot(sf)+
  geom_point(aes(x=Time,y=Predicted,color="Mortality_predicted"))+
  ggtitle(" Predicted value of Mortality")
plot.gam(model,residuals = TRUE)
model_optimal <- gam(Mortality~Year+s(Week, k=52, sp=model$sp),data=sf,family = "gaussian")
cat("The optimal penalty factor is:",model$sp,"\n")
cat("The deviance at the optimal penalty factor is:",model_optimal$deviance,"\n")
model_Low <-  gam(Mortality~Year+s(Week, k=52, sp=0),data=sf,family = "gaussian")
model_High <-  gam(Mortality~Year+s(Week, k=52, sp=100),data=sf,family = "gaussian")
pred_Low <- predict(model_Low,sf)
pred_High <- predict(model_High,sf)

sf$pred_Low <- pred_Low
sf$pred_high <- pred_High

ggplot(sf)+
  geom_point(aes(x=Time,y=Mortality,color="Mortality_data"))+
  geom_line(aes(x=Time,y=pred_Low,color="Low penalty"))+
```

```
  geom_line(aes(x=Time,y=pred_high,color="High penalty"))+
  xlab("Time")+
  ggtitle("Predicted values and Data Mortality for low and high penalty factor")

model_seq <- list()
dev <- numeric()
dof <- numeric()
penalty_factor <- seq(0,10,0.1)
for(sp in penalty_factor){
  model_seq <- gam(Mortality~Year+s(Week, k=52, sp=sp),data=sf,family = "gaussian")
  dev[(sp*10)+1] <- model_seq$deviance
  dof[(sp*10)+1] <- sum(model_seq$edf)
}

graph_variables <- data.frame(PenaltyFactor=penalty_factor,Deviance=dev,DegreeOfFreedom=dof)
ggplot(graph_variables)+
  geom_line(aes(x=PenaltyFactor,y=Deviance,color="Deviance"))+
  ggtitle("Relation between penalty factor and deviance")

ggplot(graph_variables)+
  geom_line(aes(x=PenaltyFactor,y=DegreeOfFreedom,color="DOF"))+
  ggtitle("Relation between penalty factor and DOF")
residual_matrix <- data.frame(Time_line=sf$Time,Influenza=sf$Influenza,Residuals=as.data.frame(model$res
ggplot(residual_matrix)+
  geom_line(aes(x=Time_line,y=Influenza,color="Influenza"))+
  geom_line(aes(x=Time_line,y=model.residuals,color="Residuals"))
plot(residual_matrix$Influenza,residual_matrix$model.residuals,xlab = "Influenza",ylab = "Residuals")
 model_mul <-gam(Mortality~s(Year,k=9)
                 +s(Week,k=52)
                 +s(Influenza,k=85),data=sf,
                 family = "gaussian",method="GCV.Cp")
summary(model_mul)
pred_mul <- predict(model_mul,sf)
final_matrix <- data.frame(TIME=sf$Time,MORTALITY=sf$Mortality,PRED_MORALITY=pred_mul)
ggplot(final_matrix)+
  geom_line(aes(x=TIME,y=MORTALITY,color="Original"))+
  geom_line(aes(x=TIME,y=PRED_MORALITY,color="Predicted"))

par(mfrow=c(2,2))
plot.gam(model_mul)
set.seed(12345)
sf1$Conference <- as.factor(sf1$Conference)
n=dim(sf1)[1]
id=sample(1:n, floor(n*0.70))
train=sf1[id,]
test=sf1[-id,]
rownames(train) <- 1:nrow(train)
trainx <- t(as.matrix(train[,-4703]))
trainy <- as.matrix(train$conference)
Train_list <- list(x=trainx,y=trainy,geneid=as.character(1:nrow(trainx)),genenames=rownames(trainx))
model1 <- pamr.train(Train_list)
model.cv1 <- pamr.cv(model1,Train_list,nfold = 10)
pamr.plotcv(model.cv1)
```

```
minimum_treshold <- model.cv1$threshold[which.min(model.cv1$error)]
model_optimal1 <- pamr.train(Train_list, threshold = minimum_treshold)
feature_selected <- pamr.listgenes(model1, Train_list, threshold = minimum_treshold,genenames=T)
cat("The minimum Threshold value is:",minimum_treshold,"\n")
No_Parameters1 <- dim(feature_selected)[1]
cat("Total Features Selected: ",No_Parameters1,"\n")
cat("Top 10 contributing features are: \n",feature_selected[1:10,"name"],"\n")
testx <- t(as.matrix(test[,-4703]))
testy <- as.matrix(test$conference)
prediction1 <- pamr.predict(model1,newx=testx,threshold = minimum_treshold,type="class")
confusion_matrix1 <- table(testy,prediction1)
cat("The confusion matrix is:\n")
confusion_matrix1
misclassification_rate1 <- 1- sum(diag(confusion_matrix1))/sum(confusion_matrix1)
cat("Misclassification rate is:",misclassification_rate1)
set.seed(12345)
trainx <- as.matrix(train[,-4703])
trainy <- as.matrix(train$Conference)
model2 <- glmnet(x=trainx,y=trainy,family = "binomial",alpha = 0.5)
model.cv2 <- cv.glmnet(x=trainx,y=trainy,family = "binomial",alpha = 0.5)
plot(model.cv2)
testx <- as.matrix(test[,-4703])
testy <- as.matrix(test$Conference)
prediction2 <- predict(model2,testx,s = model.cv2$lambda.min, type="class")
confusion_matrix2 <- table(testy,prediction2)
cat("The confusion matrix is:\n")
confusion_matrix2
misclassification_rate2 <- 1- sum(diag(confusion_matrix2))/sum(confusion_matrix2)
cat("Misclassification rate is: ",misclassification_rate2,"\n")
No_Parameters2 <- dim(coef(model2))[2]
cat("Number of features selected: ",No_Parameters2,"\n")
set.seed(12345)
model3 <- ksvm(Conference~.,data=train,kernel="vanilladot",scaled=FALSE)
prediction3 <- predict(model3,test,type="response")
confusion_matrix3 <- table(Actual=test$Conference,Predicted=prediction3)
cat("The confusion matrix is:\n")
confusion_matrix3
misclassification_rate3 <- 1- sum(diag(confusion_matrix3))/sum(confusion_matrix3)
cat("Misclassification rate is: ",misclassification_rate3,"\n")
No_Parameters3 <- length(model3@coef[[1]])
cat("Number of feature selected: ",No_Parameters3,"\n")

misclassification <-c(misclassification_rate1,misclassification_rate2,misclassification_rate3)
model_list <- c("Nearest Shrunken Centroid","Elastic Net","Support Vector Machine")
feature_selected <- c(No_Parameters1,No_Parameters2,No_Parameters3)
comparison_table <- data.frame(Model=model_list,misClassificationrates=misclassification,FeaturesSelecte
cat("The comparison table is as follows:\n")
comparison_table
p_value <- numeric(length = 4702)
test <- list()
bh <- numeric(length = 4702)
for (i in 1:4702){
  test[[i]] <- t.test(sf1[,i] ~ Conference, data = sf1)
```

```
  p_value[i] <- test[[i]]$p.value
  bh[i] <- ((0.05)*(i/4702))
}
p_data_frame <- data.frame(p_value,bh)
p_data_frame <- p_data_frame[order(p_data_frame$p_value),]
p_data_frame <- p_data_frame[which(p_data_frame$p_value <= p_data_frame$bh),]
No_of_feature = nrow(p_data_frame)
cat("The Number of features selected are:",No_of_feature,"\n")
index <- rownames(p_data_frame)
cat("The Selected parameters are listed below:\n")
data.frame(selected_Feature=colnames(sf1[,as.numeric(index)]))
```