Computer lab 2 block 2

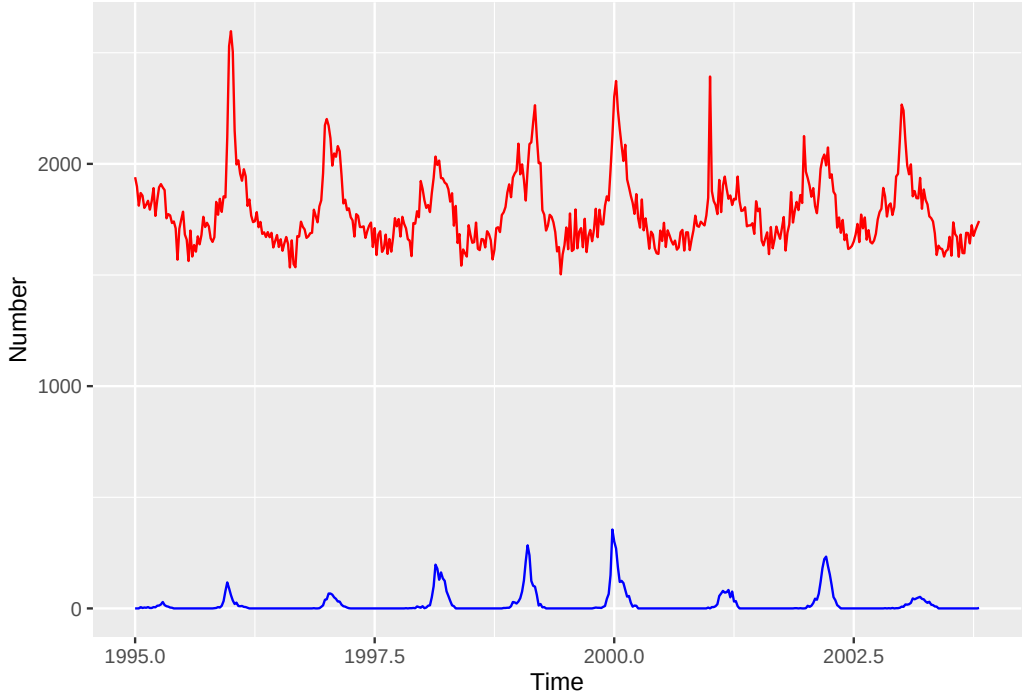
Zuxiang Li

12/10/2019

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

The Excel document influenza.xlsx contains weekly data on the mortality and the number of laboratory-confirmed cases of influenza in Sweden. In addition, there is information about population-weighted temperature anomalies (temperature deficits).

1. Use time series plots to visually inspect how the mortality and influenza number vary with time (use Time as X axis). By using this plot, comment how the amounts of influenza cases are related to mortality rates.



The red line in the plot represents mortality, and the blue line represents influenza varies with time. We can observe from the plot that every time influenza breaks out, there must be a drastic increase in mortality.

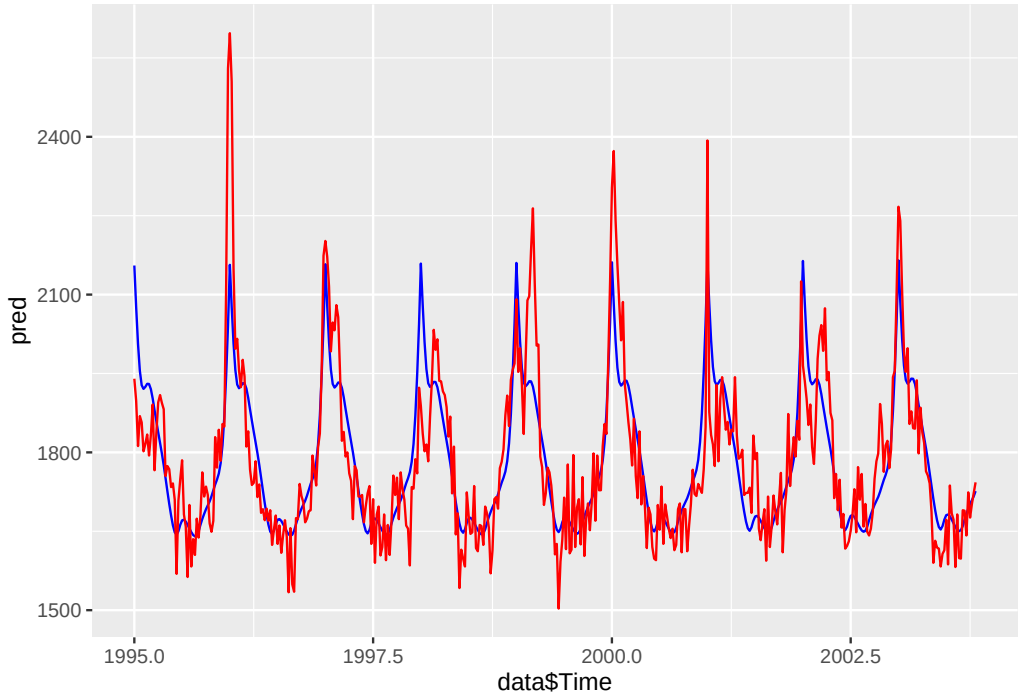
2. Use gam() function from mgcv package to fit a GAM model in which Mortality is normally distributed and modelled as a linear function of Year and spline function of Week, and make sure that the model parameters are selected by the generalized cross-validation. Report the underlying probabilistic model.

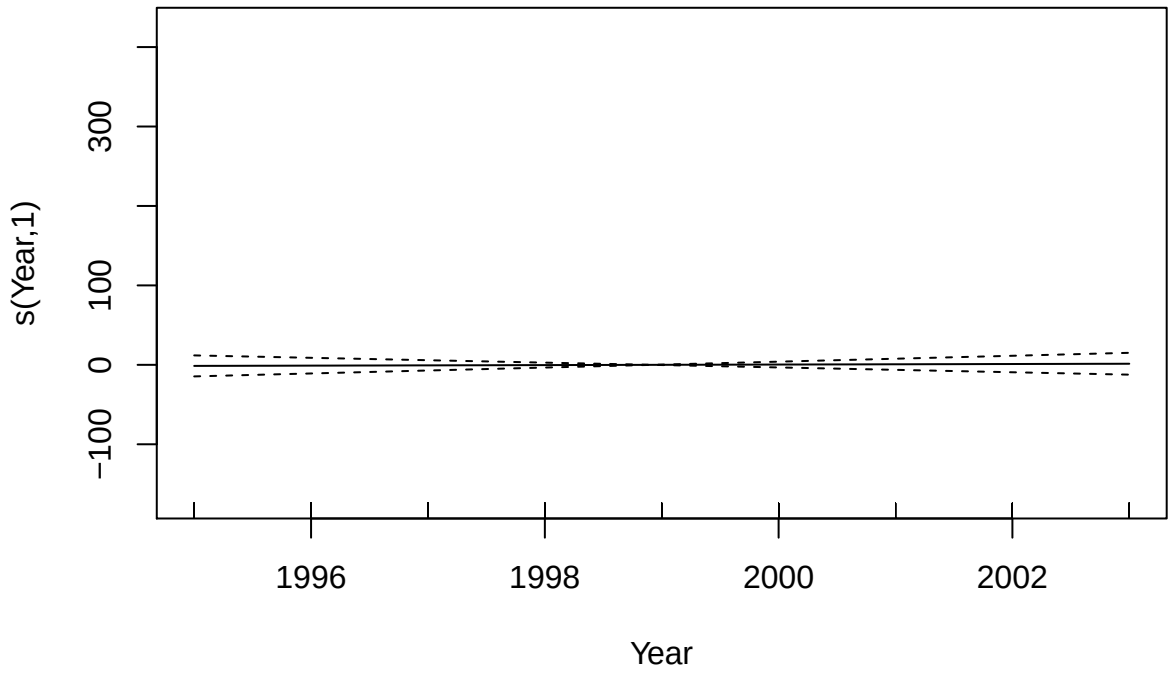
```
##
## Family: gaussian
```

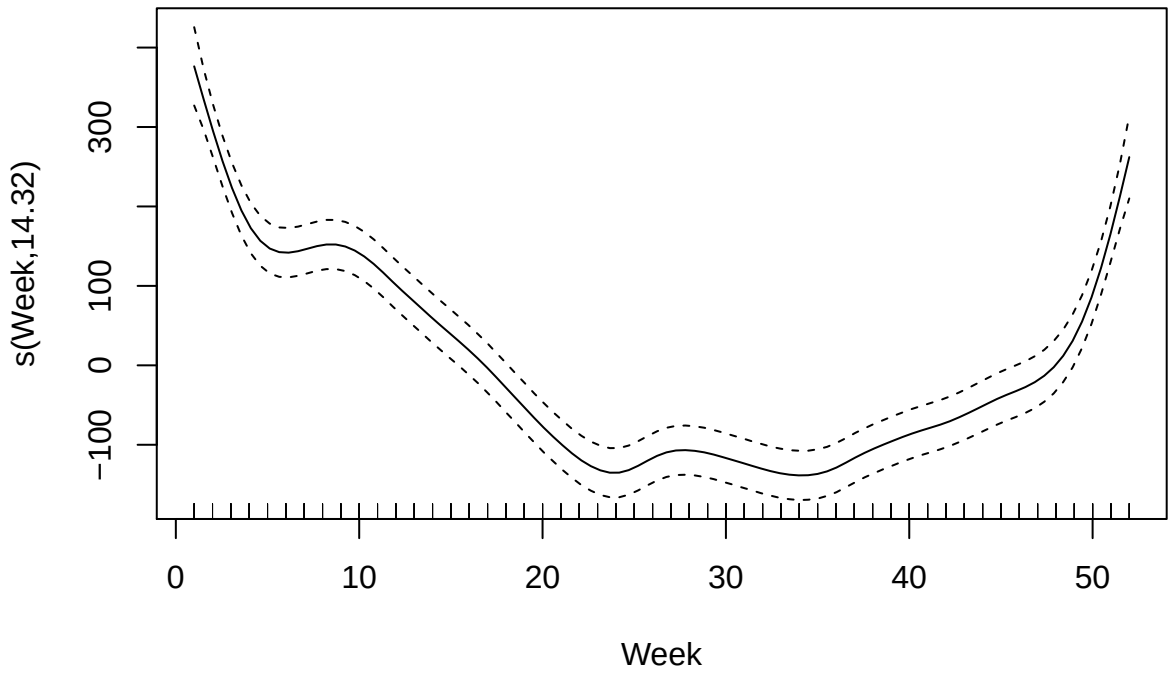
```
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(data$Week)))
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -680.598   3367.760  -0.202    0.840
## Year           1.233      1.685   0.732    0.465
##
## Approximate 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
```

The model here we build is accroding to $Mortality = Year + s(Week)$, Week is the spline function.

3. Plot predicted and observed mortality against time for the fitted model and comment on the quality of the fit. Investigate the output of the GAM model and report which terms appear to be significant in the model. Is there a trend in mortality change from one year to another? Plot the spline component and interpret the plot.

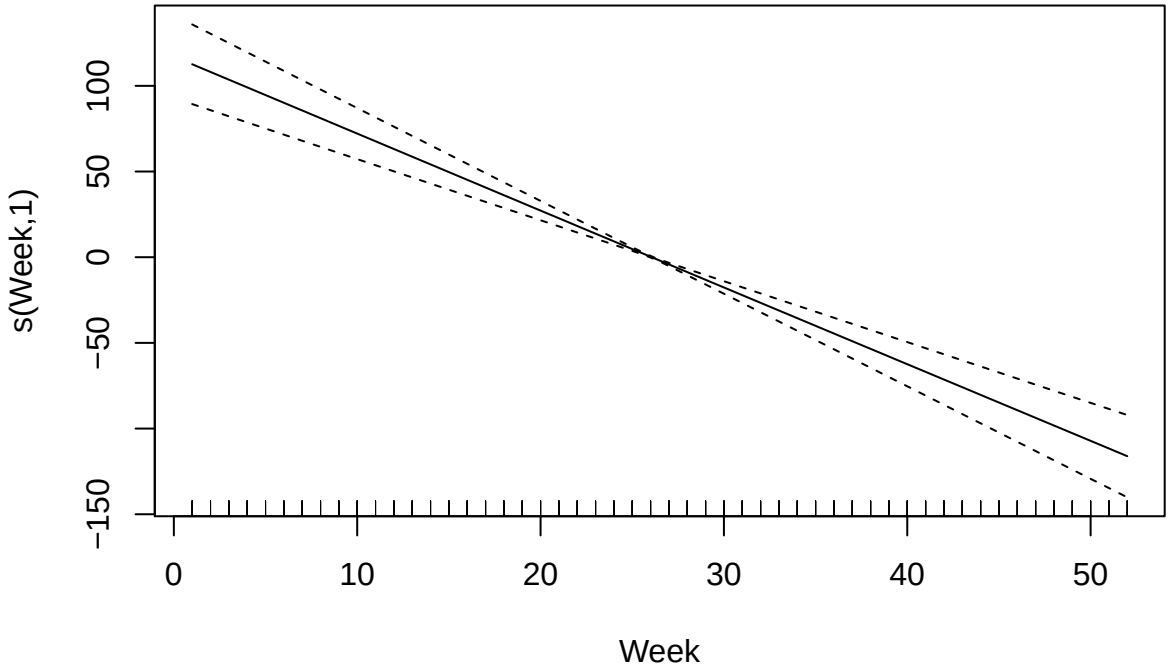


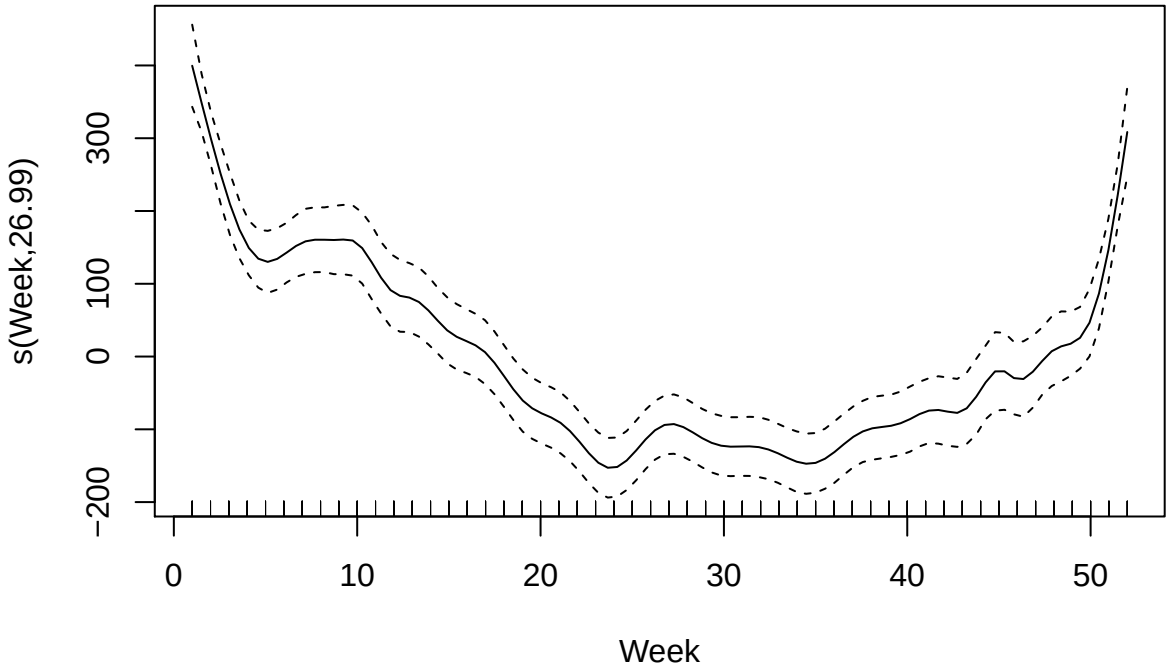




The red line is observed mortality and the blue line represents predict values. From the spline plot, we can observe that the value perform a stright line, which means there is not a trend for mortality between years, however, there is a strong connection between week and number of mortality, in the begining and end of the year, which indicate the winter of a year, the value maintains at a relatively high level.

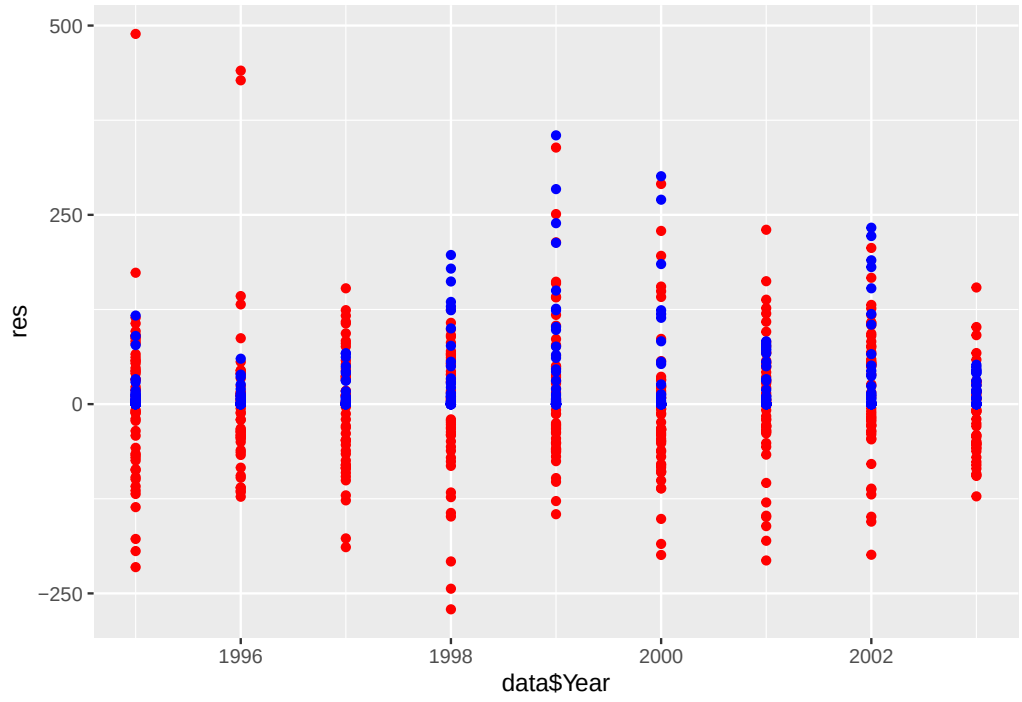
4. Examine how the penalty factor of the spline function in the GAM model from step 2 influences the estimated deviance of the model. Make plots of the predicted and observed mortality against time for cases of very high and very low penalty factors. What is the relation of the penalty factor to the degrees of freedom? Do your results confirm this relationship?





In the previous model, we got the sp value which is the penalty factor for spline function. The sp value for week is 1.131922e-04, with the optimal value, we choose 1000 as very high and 0.00000001 as very low values for sp. With the higher penalty, the estimated degrees of freedom will get closer to 1, this means the model is underfitted. Otherwise, the degrees of freedom gets higher with lower penalty, the model tends to be overfitted.

5. Use the model obtained in step 2 and plot the residuals and the influenza values against time (in one plot). Is the temporal pattern in the residuals correlated to the outbreaks of influenza?

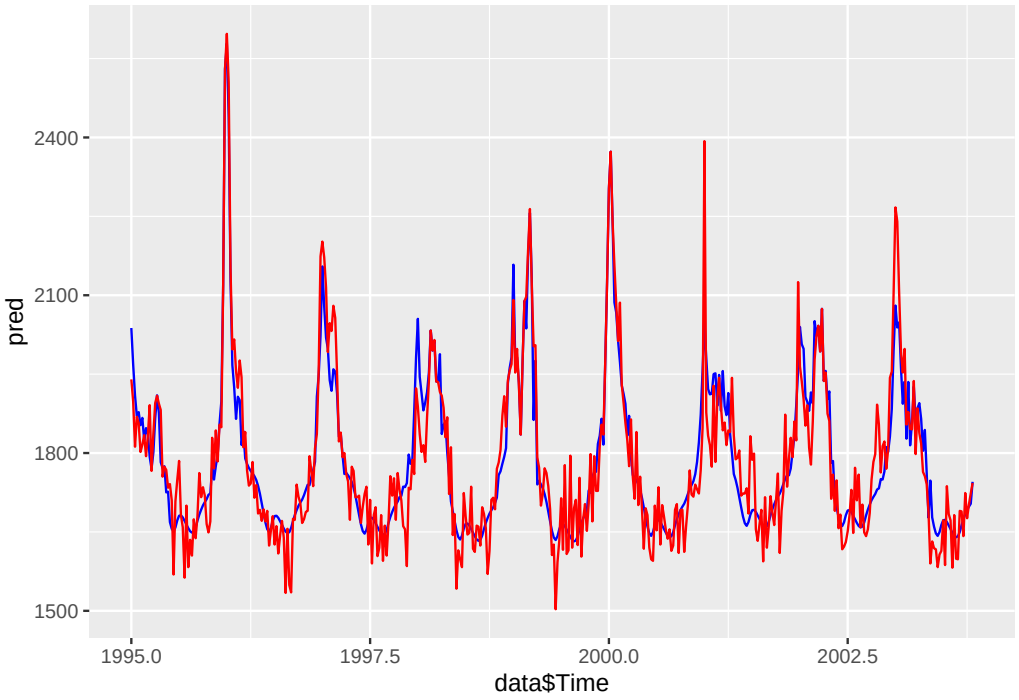


There is no obvious correlation between residuals and influenza

6. Fit a GAM model in R in which mortality is be modelled as an additive function of the spline functions of year, week, and the number of confirmed cases of influenza. Use the output of this GAM function to conclude whether or not the mortality is influenced by the outbreaks of influenza. Provide the plot of the original and fitted Mortality against Time and comment whether the model seems to be better than the previous GAM models.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Year, k = length(unique(data$Year))) + s(Week,
##     k = length(unique(data$Week))) + s(Influenza, k = length(unique(data$Influenza)))
##
## 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
```



Since we took influenza into consideration, the model fits better compare to the previous one.

Assignment 2. High-dimensional methods

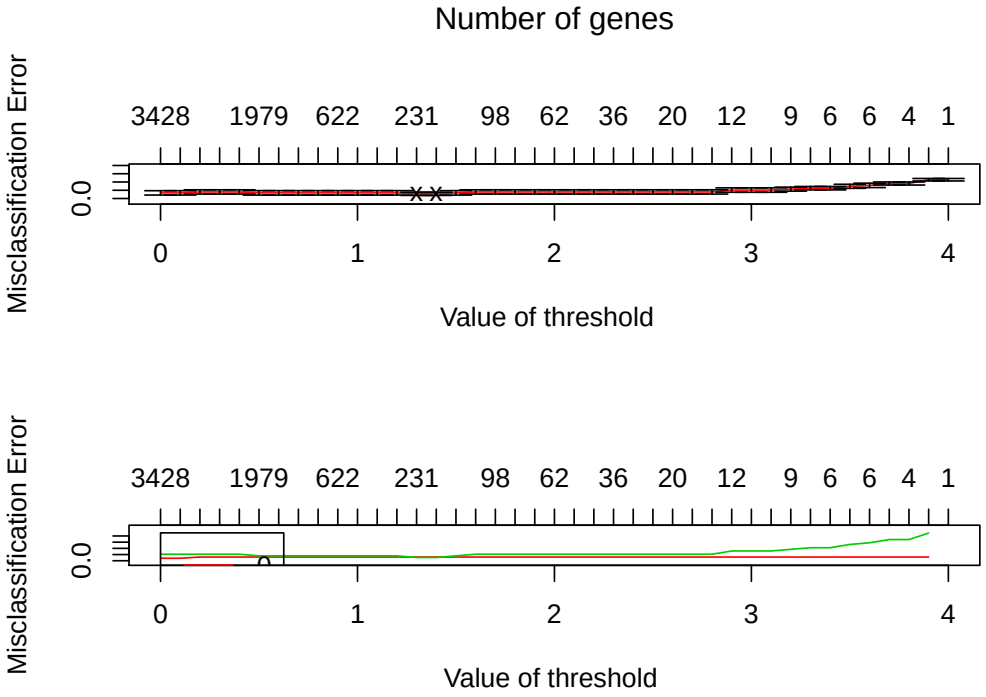
The data file data.csv contains information about 64 e-mails which were manually collected from DBWorld mailing list. They were classified as: 'announces of conferences' (1) and 'everything else' (0) (variable Conference)

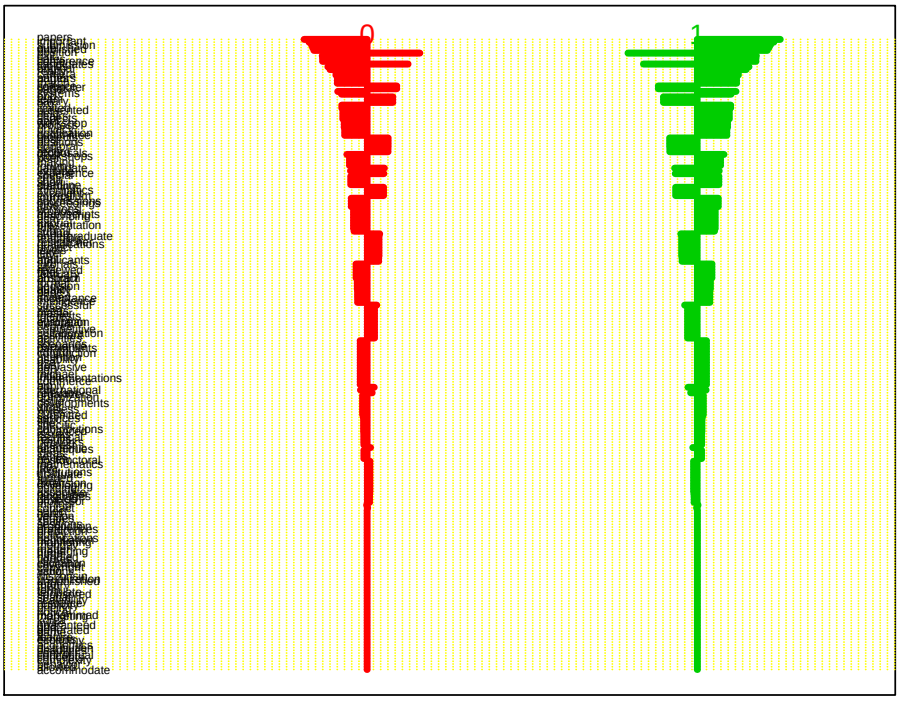
1. Divide data into training and test sets (70/30) without scaling. Perform nearest shrunken centroid classification of training data in which the threshold is chosen by cross-validation. Provide a centroid plot and interpret it. How many features were selected by the method? List the names of the 10 most contributing features and comment whether it is reasonable that they have strong effect on the discrimination between the conference mails and other mails? Report the test error.

```
## 1234567891011121314151617181920212223242526272829303132333435363738394041
```

```
## 12Fold 1 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 2 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 3 :1234567891011121314151617181920212223242526272829303132333435363738394041
```

```
## Fold 4 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 5 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 6 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 7 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 8 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 9 :12345678910111213141516171819202122232425262728293031323334353637383940 41
## Fold 10 :12345678910111213141516171819202122232425262728293031323334353637383940 41
```



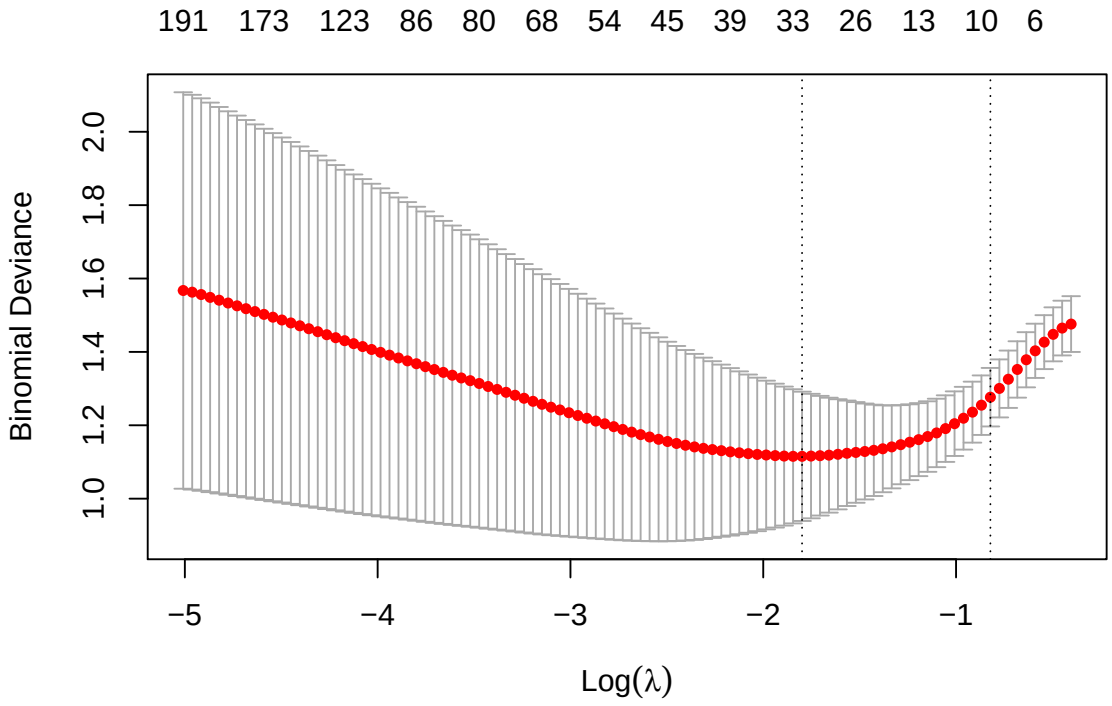


We get the threshold is 1.3 with using cross-validation. From the centroid plot we can see the contribution

of each word made to the result(conference or not). There are 693 features selected in total. The 10 most contributing features are “papers”, “important”, “submission”, “due”, “published”, “position”, “call”, “conference”, “dates”, “candidates”. It’s clear that these word have a strong connection to conference. The test error is 5%.

2. Compute the test error and the number of the contributing features for the following methods fitted to the training data:

a. Elastic net with the binomial response and = 0.5 in which penalty is selected by the cross-validation




```
## 10 %
```

b. Support vector machine with “vanilladot” kernel.

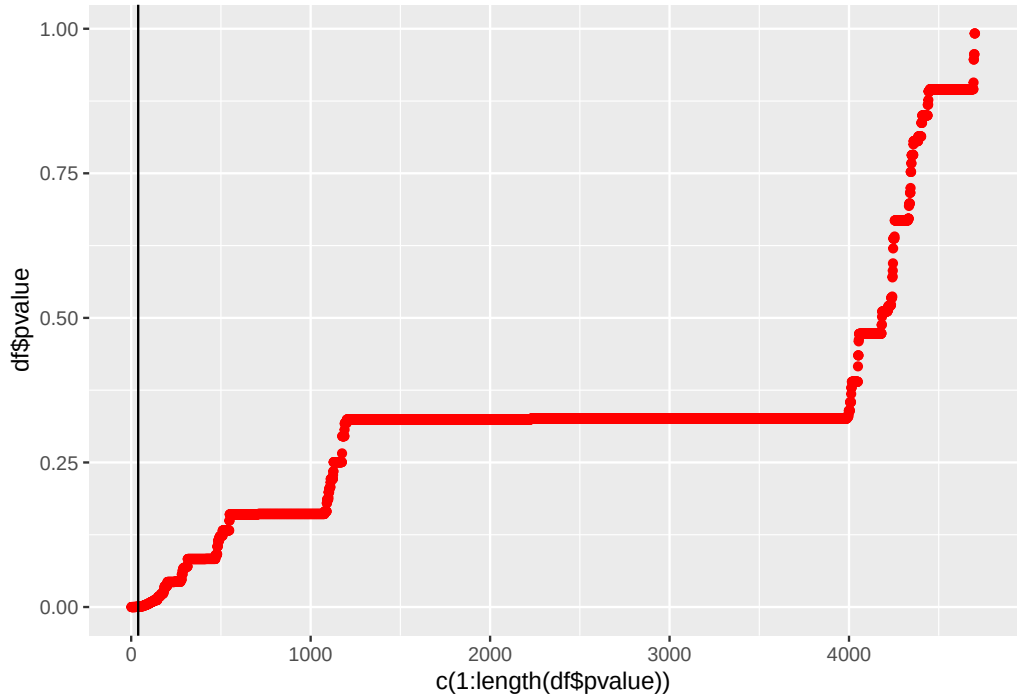
```
##  Setting default kernel parameters
```

```
## 5 %
```

Compare the results of these models with the results of the nearest shrunken centroids (make a comparative table). Which model would you prefer and why?

Error rate for Elastic net is 10% and for SVM is 5%. In this case we prefer to use SVM since it ignores the effect of high-dimensional data and it provides the lowest misclassfication rate.

3. Implement Benjamini-Hochberg method for the original data, and use t.test() for computing p-values. Which features correspond to the rejected hypotheses? Interpret the result.




```
## [1] 39
```

```
##              name       pvalue
## 3036        papers 1.116910e-10
## 4060    submission 7.949969e-10
## 3187      position 8.219362e-09
## 3364     published 1.835157e-07
## 2049     important 3.040833e-07
## 596           call 3.983540e-07
## 869     conference 5.091970e-07
## 607     candidates 8.612259e-07
## 1045         dates 1.398619e-06
## 3035         paper 1.398619e-06
## 4282        topics 5.068373e-06
## 2463       limited 7.907976e-06
## 606      candidate 1.190607e-05
## 599         camera 2.099119e-05
## 3433         ready 2.099119e-05
## 389        authors 2.154461e-05
## 3125           phd 3.382671e-05
## 3312      projects 3.499123e-05
## 2974           org 3.742010e-05
## 681         chairs 5.860175e-05
## 1262           due 6.488781e-05
## 2990      original 6.488781e-05
## 2889  notification 6.882210e-05
## 3671        salary 7.971981e-05
## 3458        record 9.090038e-05
```

```
## 3891        skills 9.090038e-05
## 1891          held 1.529174e-04
## 4177          team 1.757570e-04
## 3022         pages 2.007353e-04
## 4628      workshop 2.007353e-04
## 810      committee 2.117020e-04
## 3285   proceedings 2.117020e-04
## 272          apply 2.166414e-04
## 4039        strong 2.246309e-04
## 2175 international 2.295684e-04
## 1088        degree 3.762328e-04
## 1477     excellent 3.762328e-04
## 3191          post 3.762328e-04
## 3243     presented 3.765147e-04
```

There are 39 features correspond to the rejected hypotheses with $\alpha = 0.05$.

Appendix

Assignment 1

1.

```
data<-read_xlsx("data/influenza.xlsx")

ggplot(data=data)+geom_line(aes(x=Time,y=Mortality),col="red")+
  geom_line(aes(x=Time,y=Influenza),col="blue")+ylab("Number")
```

2.

```
res=gam(Mortality~Year+s(Week,k=length(unique(data$Week))),method ="GCV.Cp" ,data=data)
```

3.

```
pred=predict.gam(res,data)

ggplot()+geom_line(aes(x=data$Time,y=pred),col="blue")+geom_line(aes(x=data$Time,y=data$Mortality),col=

res=gam(Mortality~Year+s(Year,k=length(unique(data$Year)))+s(Week,k=length(unique(data$Week))),method =
plot(res)
```

4.

```
res=gam(Mortality~Year+s(Week,k=length(unique(data$Week)),sp=1000),method ="GCV.Cp" ,data=data)
plot(res)

res=gam(Mortality~Year+s(Week,k=length(unique(data$Week)),sp=0.00000001),method ="GCV.Cp" ,data=data)
plot(res)
```

5.

```
res=gam(Mortality~Year+s(Week,k=length(unique(data$Week))),method ="GCV.Cp" ,data=data)
pred=predict.gam(res,data)
res=data$Mortality-pred

ggplot()+geom_point(aes(x=data$Year,y=res),col="red")+geom_point(aes(x=data$Year,y=data$Influenza),col=
```

6.

```
res=gam(Mortality~s(Year,k=length(unique(data$Year)))+s(Week,k=length(unique(data$Week)))+s(Influenza,k=
summary(res)
pred=predict.gam(res,data)

ggplot()+geom_line(aes(x=data$Time,y=pred),col="blue")+geom_line(aes(x=data$Time,y=data$Mortality),col=
```

Assignment 2

1.

```
data<-read.csv2("data/data.csv",check.names = FALSE)
names(data)<-iconv(names(data),to="ASCII")
RNGversion("3.5.1")

n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.7))
train=data[id,]
test=data[-id,]

x<-t(train[,-4703])

y<-train[[4703]]


x_test<-t(test[,-4703])

y_test<-test[[4703]]

my_data<-list(x=x,y=as.factor(y),geneid=as.character(1:nrow(x)),genenames=rownames(x))
my_data_test<-list(x=x_test,y=as.factor(y_test),geneid=as.character(1:nrow(x_test)),genenames=rownames(

mod<-pamr.train(my_data,threshold = seq(0,4,0.1))
cvmodel<-pamr.cv(mod,my_data)

thr<-cvmodel$threshold[which.min(cvmodel$error)]
pamr.plotcv(cvmodel)
pred<-pamr.predict(mod,my_data_test$x,threshold = thr,type="class")
pamr.plotcen(mod,my_data,thr)
```

2.

```
library(glmnet)

x<-train[,-4703]
y<-train[[4703]]

x_test<-test[,-4703]
y_test<-test[[4703]]

mod<-cv.glmnet(as.matrix(x),y,alpha=0.5,family="binomial")
penalty_min<-mod$lambda.min
real_mod<-glmnet(as.matrix(x),y,alpha=0.5,lambda = penalty_min,family="binomial")
pred<-predict(real_mod,as.matrix(x_test),type="class")
cft<-table(pred,y_test)
mis_rate<-1-(cft[1,1]+cft[2,2])/sum(cft)
cat(mis_rate*100,"%")

fit<-ksvm(as.matrix(x),y,data=train,kernel="vanilladot",type="C-svc",scale=FALSE)
```

```
pred<-predict(fit,x_test,type="response")

cft<-table(pred,y_test)

mis_rate<-1-(cft[1,1]+cft[2,2])/sum(cft)
cat(mis_rate*100,"%")
```

3.

```
x=as.matrix(data[,-4703])
y=as.factor(data[[4703]])

df<-data.frame(name=c(),pvalue=c())

for(i in 1:ncol(x)){
  tmpv<-t.test(x[,i]~y,alternative="two.sided",conf.level=0.95)$p.value
  tdf<-data.frame(name=colnames(x)[i],pvalue=tmpv)
  df<-rbind(df,tdf)
}
df<-df[order(df$pvalue),]

a=0.05
max_i=1
for(i in 1:length(df$pvalue)){
  if(df$pvalue[i]<=a*i/length(df$pvalue)){
    max_i=i
  }
}
ggplot()+geom_point(aes(x=c(1:length(df$pvalue)),y=df$pvalue),col="red")+geom_vline(xintercept = 39)
print(max_i)

df[1:39,]
```