

# Group A15 Machine Learning Lab 2 block 2

*Omkar Bhutra(omkbh878), Tejashree R Mastamardi (tejma768), Vinay Bengaluru (vinbe289)*

*11 December 2018*

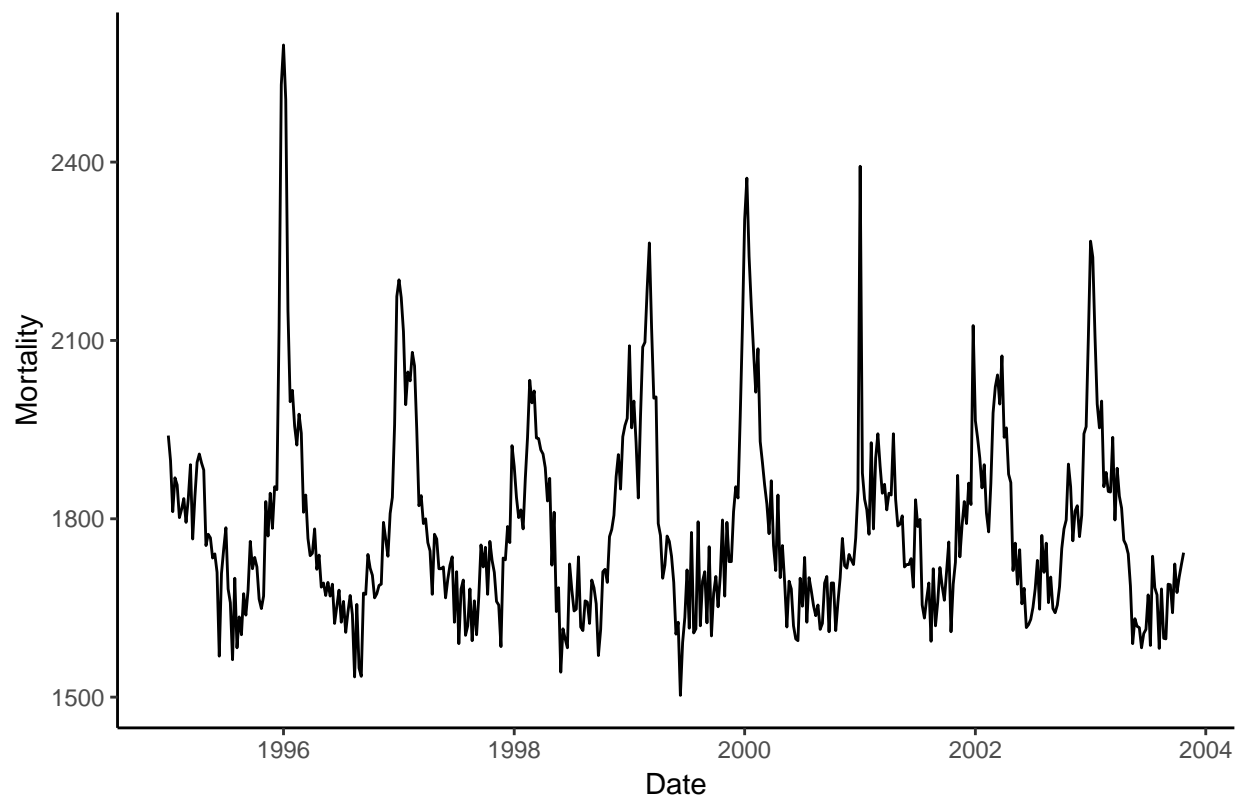
## Assignment 1.

Using GAM and GLM to examine the mortality rates

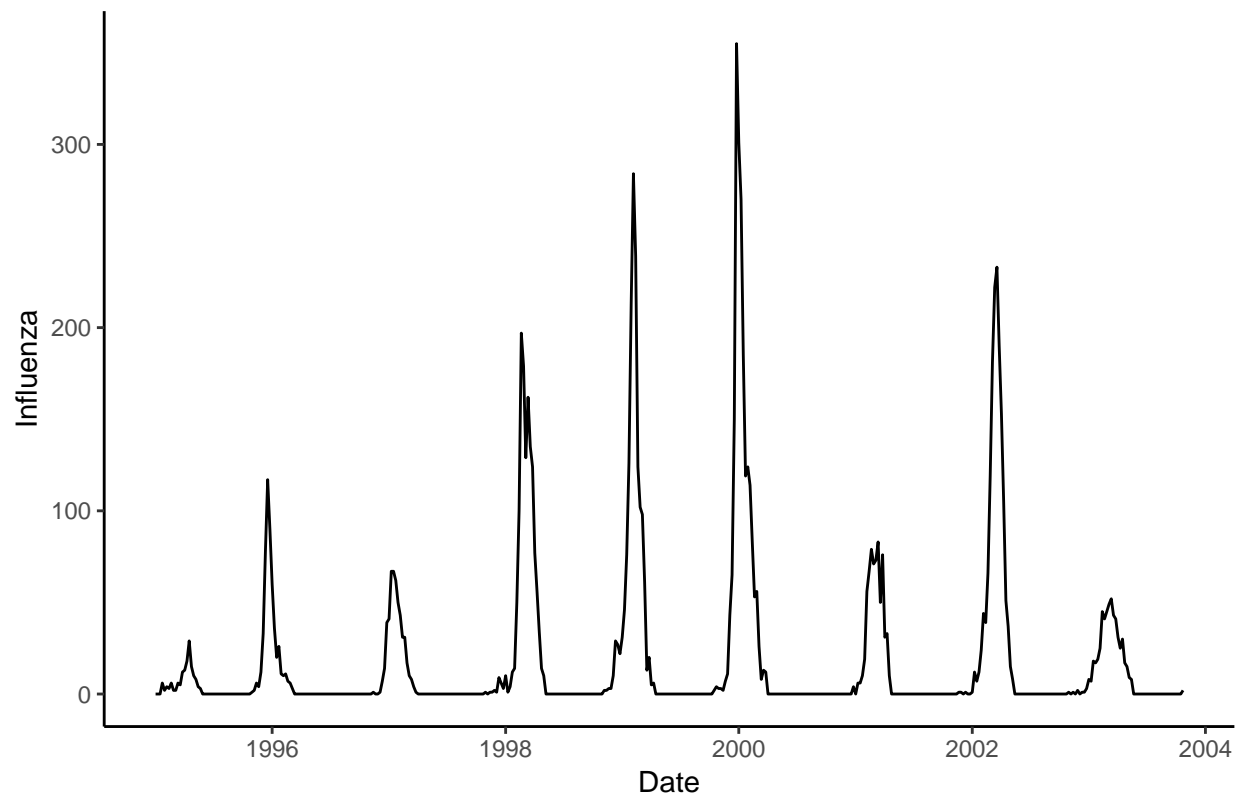
### Q1

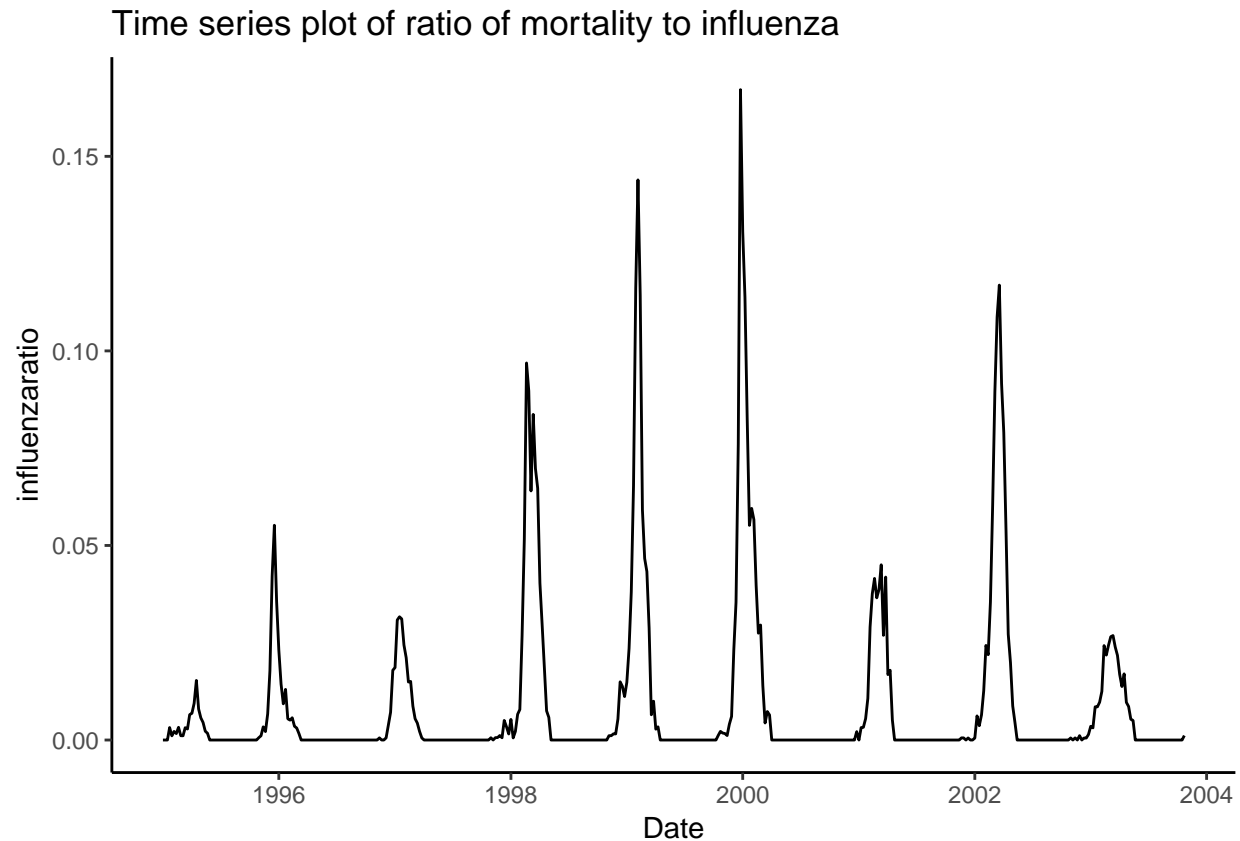
From the plots we can see that, Mortality and Influenza peaking during the same time of each year which is the 1st quarter (Jan to March) with Influenza peaking sometimes in December of the previous as well. Although, The highest mortality is in January of 1996 with 2597 deaths and the highest laboratory-confirmed cases of influenza is found in December of 1999 with 355 cases. The third plot shows the percentage of influenza cases that directly attributed to death and it confirms that the two variables are highly correlated.

Time series plot of Mortality



Time series plot of lab confirmed influenza cases





Q2

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

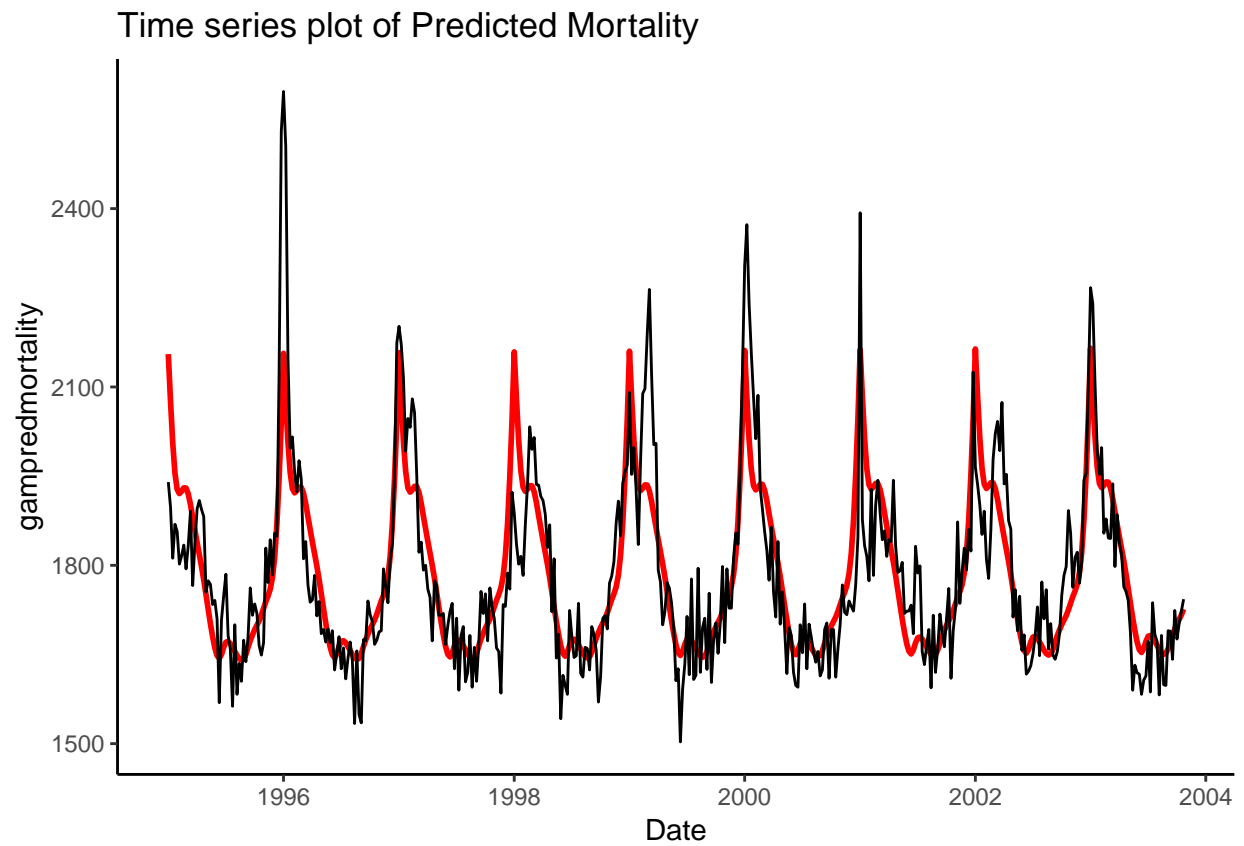
Underlying probabilistic equation of the model :

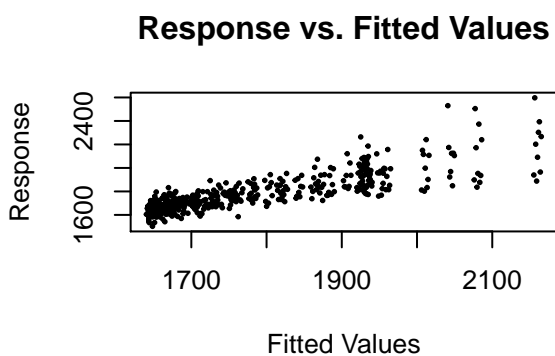
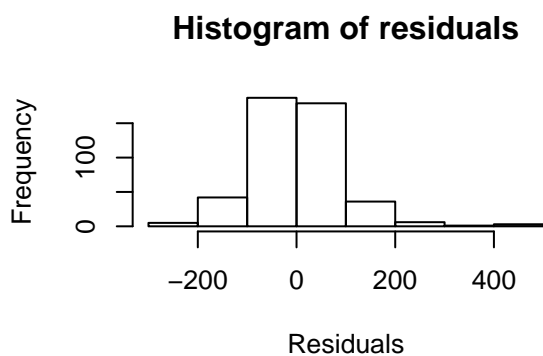
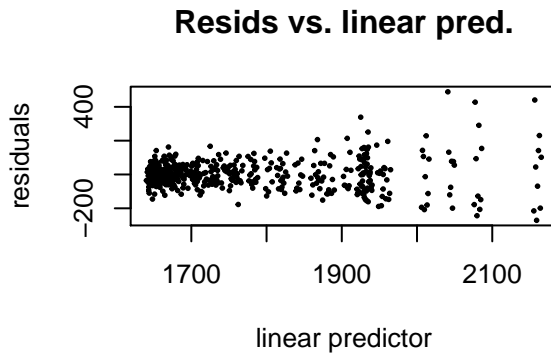
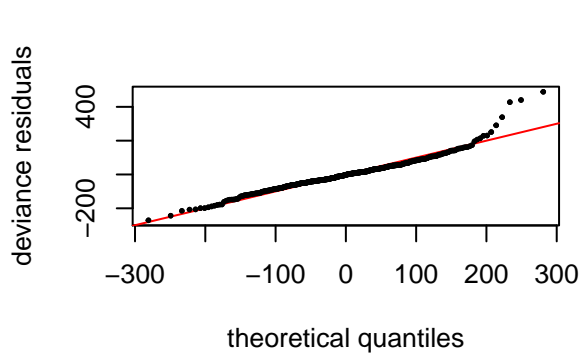
$$Mortality = N(\mu, \sigma^2)$$

$$g(\mu) = Intercept + Beta_{year} * Year + s(Week)$$

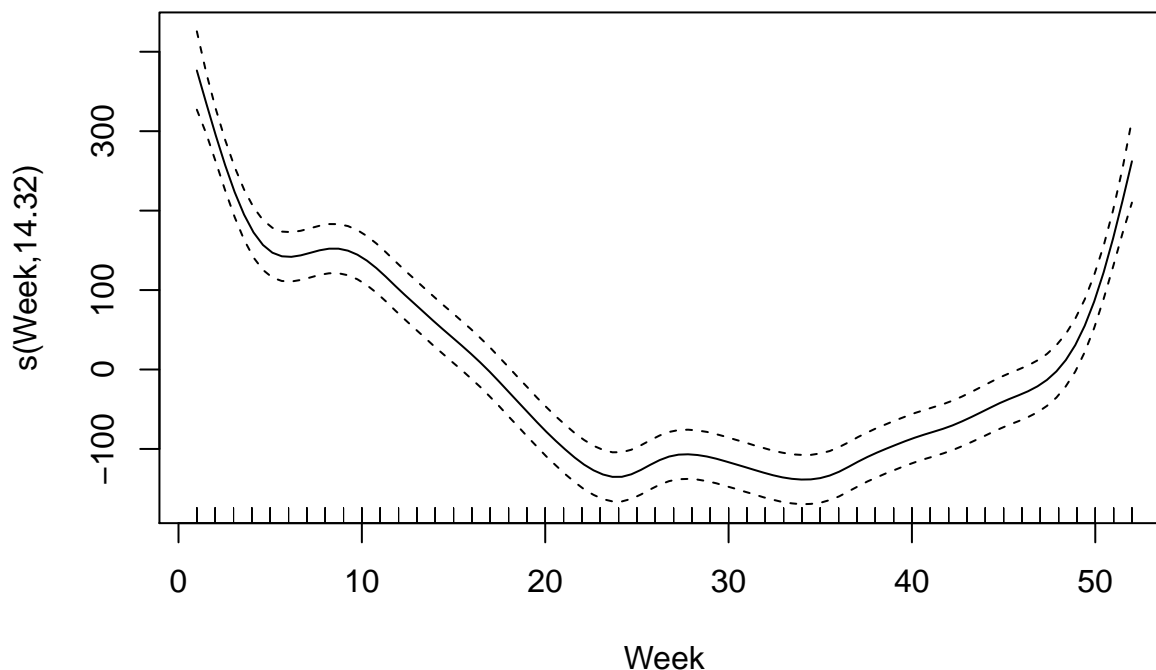
Where g is the link function, in this case it is a normal distribution

Q3





```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 9 iterations by steepest
## descent step failure.
## The RMS GCV score gradient at convergence was 0.00106719 .
## The Hessian was positive definite.
## Model rank =  52 / 53
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##           k'   edf k-index p-value
## s(Week) 51.0 14.3   1.09   0.99
```



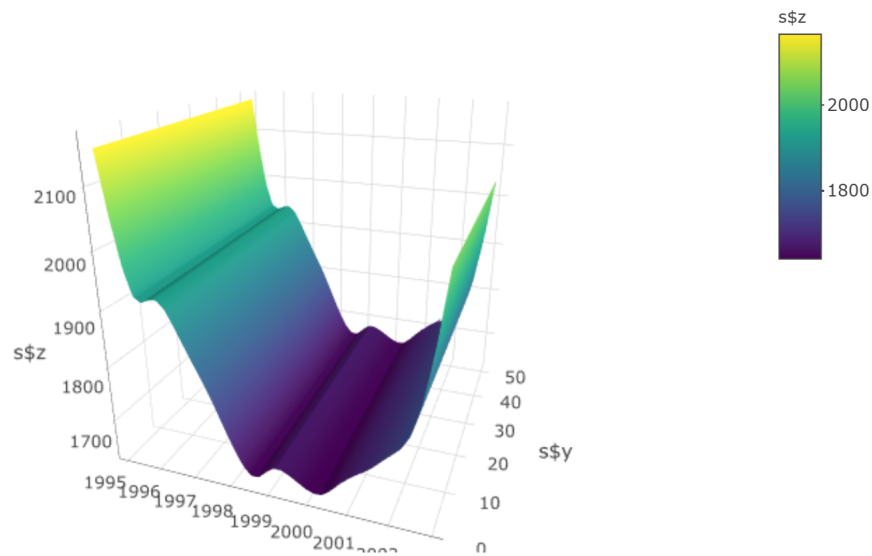
The predicted mortality fits quite well with the time (x axis) i.e the peaks and troughs match with the actual mortality value but it is a repeating function that does not capture the the mortality values in the model and hence not a very good model to predict. It is observed that the linear component of year is not significant but the spline component of Week is a significant term with a very low p value. From the plot of the spline component it is seen that mortality peaks in the winter of each year and are the least in the summer of each year.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(Influenza$Week)))
##
## Estimated degrees of freedom:
## 14.3 total = 16.32
##
## GCV score: 8708.581      rank: 52/53

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

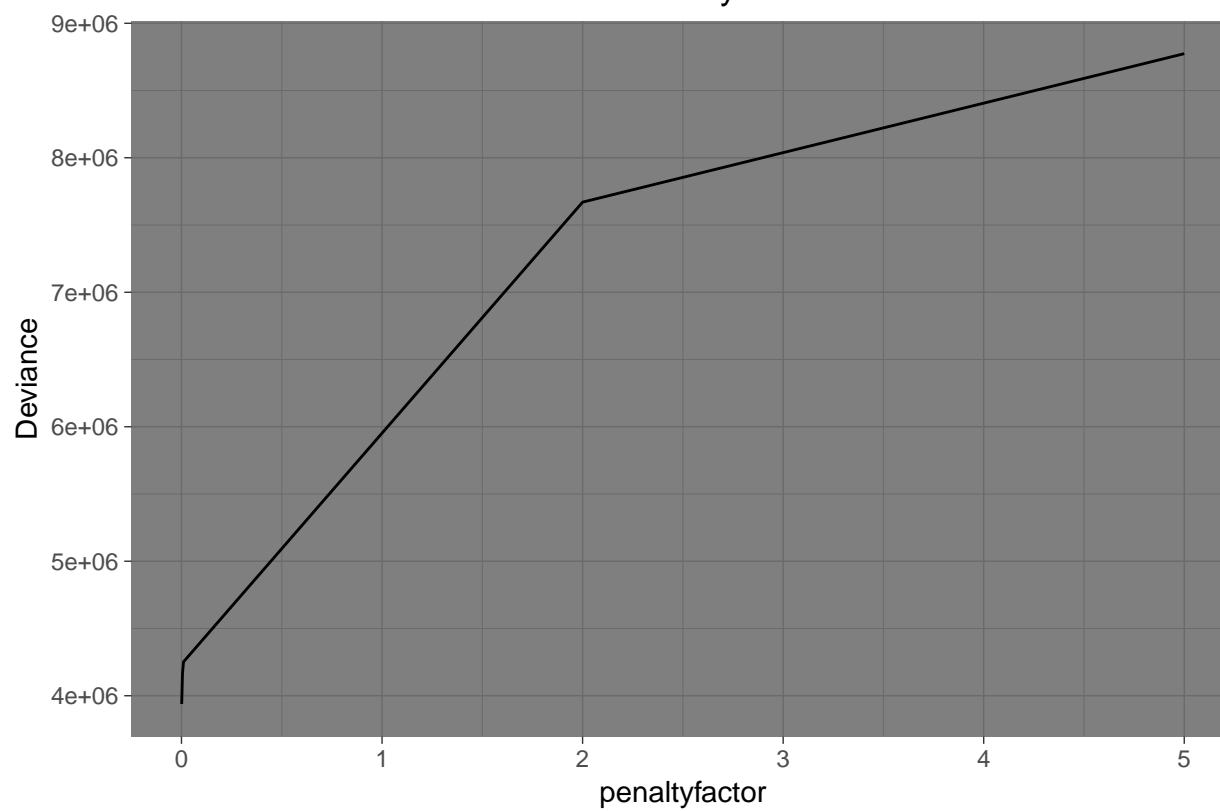
```
## Parametric coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -680.598   3367.760  -0.202   0.840
## Year         1.233     1.685    0.732   0.465
##
## Approximate significance of smooth terms:
##           edf Ref.df    F p-value
## s(Week) 14.32  17.87 53.86 <2e-16 ***
## ---
## 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

##           s(Week)
## 0.0001131932
```



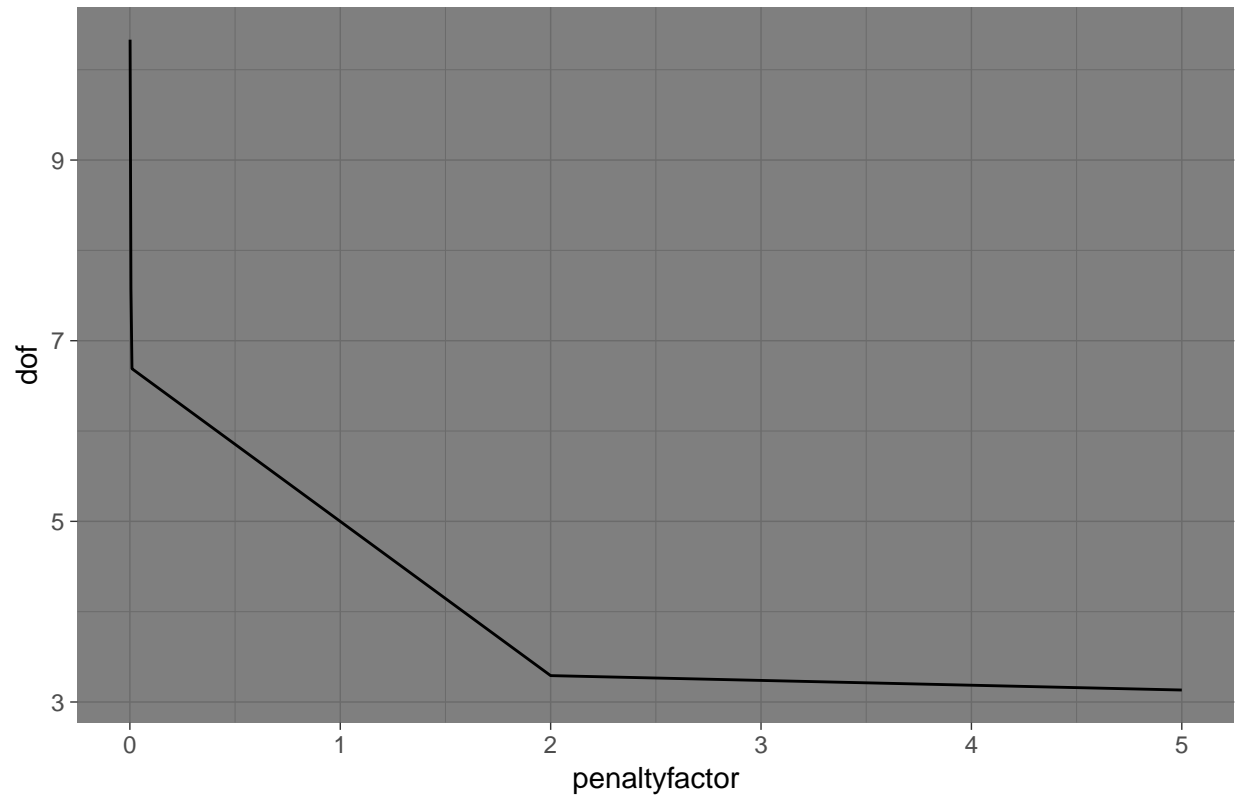
Q4

Plot of Deviances of models vs. Penalty Factors

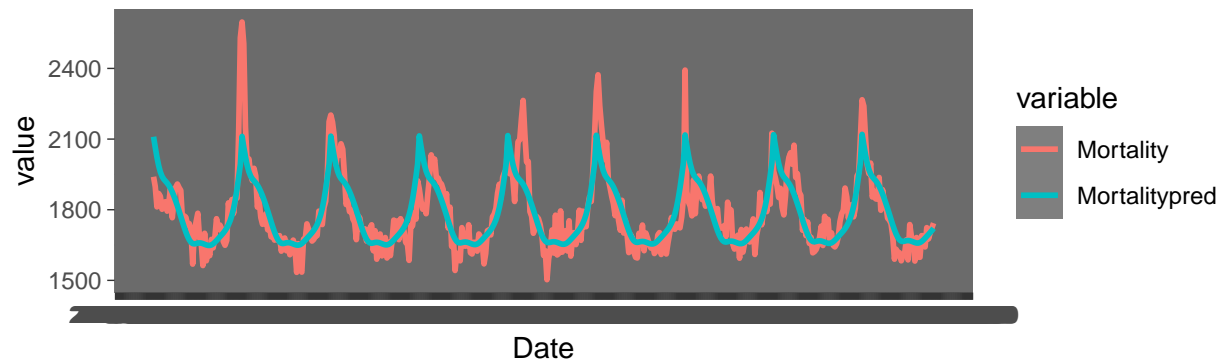




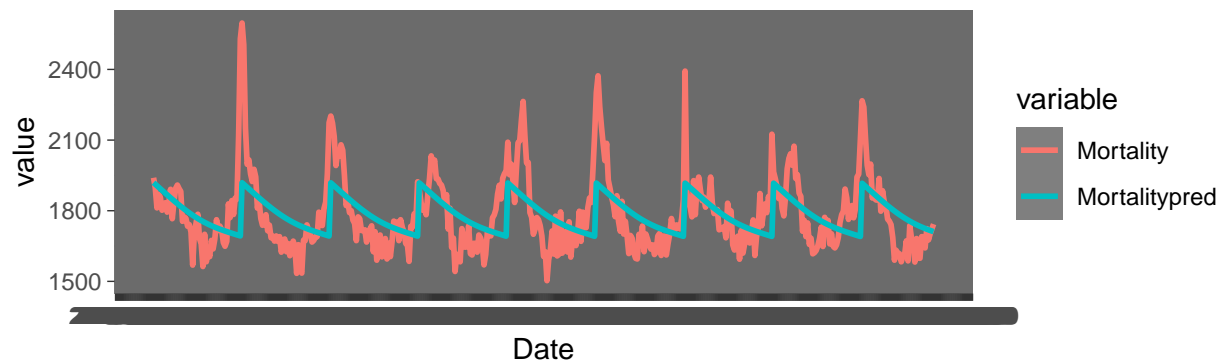
Plot of Degree of freedoms of models vs. Penalty Factors



Plot of Mortality vs. Time(Penalty factor of 0.001 is used)



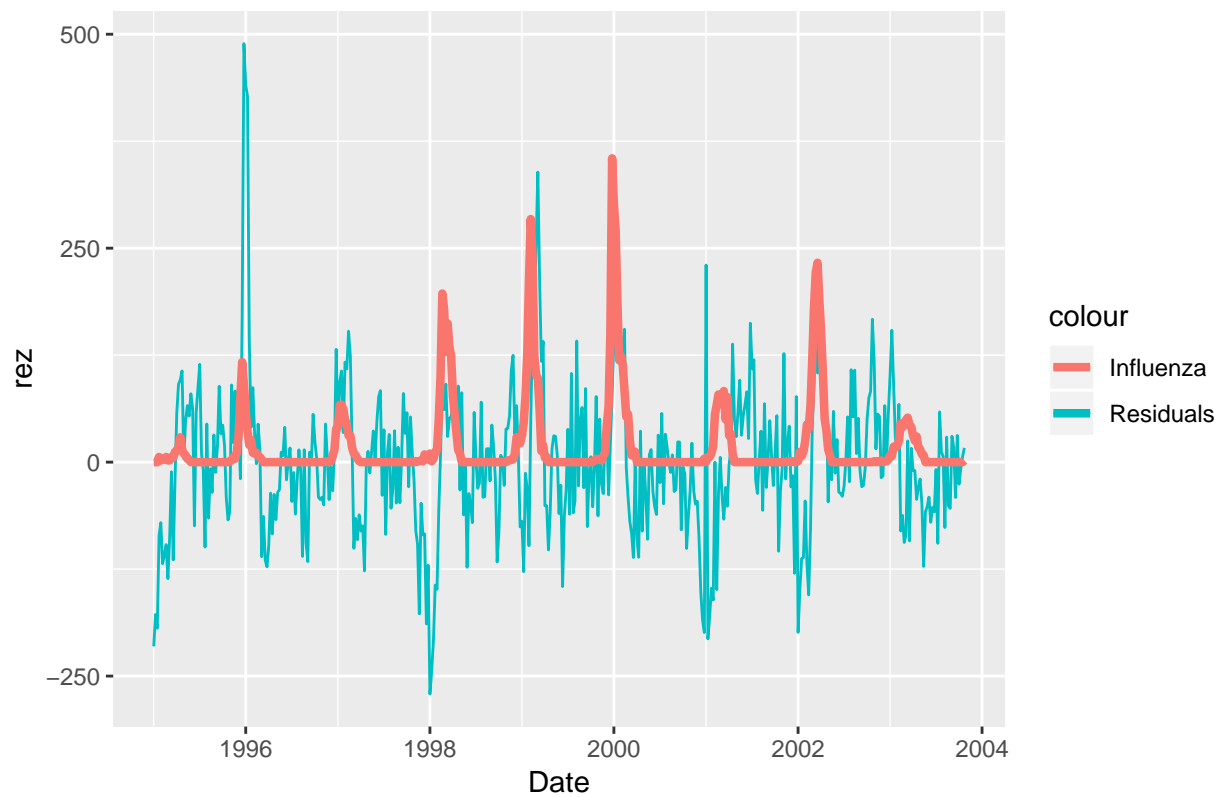
Plot of Mortality vs. Time (Penalty factor of 5 is used)



A directly proportional relationship is seen between penalty factor and deviance. Higher the penalty factor, higher is the deviance. With a higher penalty factor comes less complexity and more bias in the model. An inverse relationship is seen between penalty factor and degree's of freedom. Lower the penalty factor, Higher is the degree of freedom. yes, this is confirmed from our results.

Q5

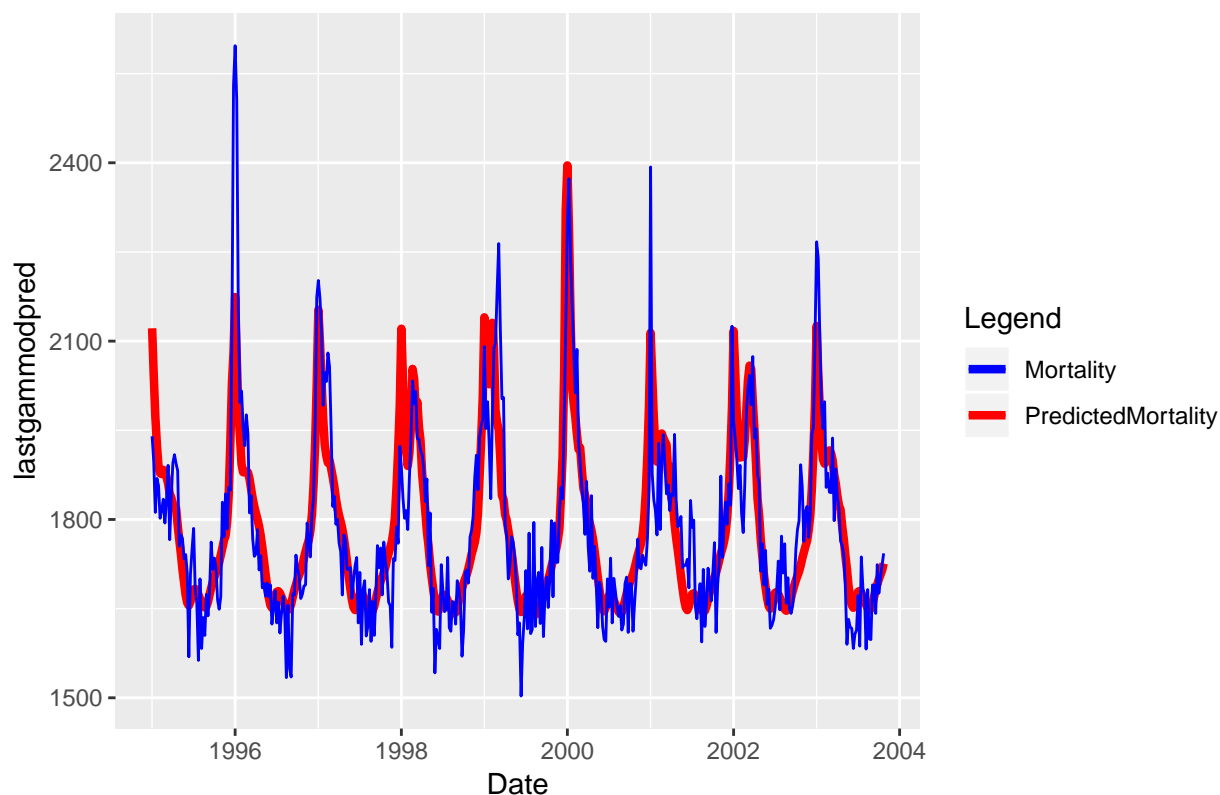
Time series plot of Residuals of model and Influenza cases



the temporal pattern in the residuals can be linked to the periodic outbreak of influenza to an extent. The Three largest outbreaks of influenza also have residuals peaking in the positive direction while it is seen that the residuals have negative troughs right before the influenza peaks that is for the last quarter of the year.

Q6

Time series plot of Predicted and Actual Mortality based on new GAM model



Yes, this Generalised Additive Model is better than the previous models as the predicted fit is good not only in the x axis but also matches the actual value peaks and troughs. It can be concluded that Mortality can be described well with non linear spline functions of Year and Week along with the linear function of Influenza. Hence, Outbreaks of Influenza in the winters have a direct effect on Mortality.

## Assignment 2.

### High-dimensional methods

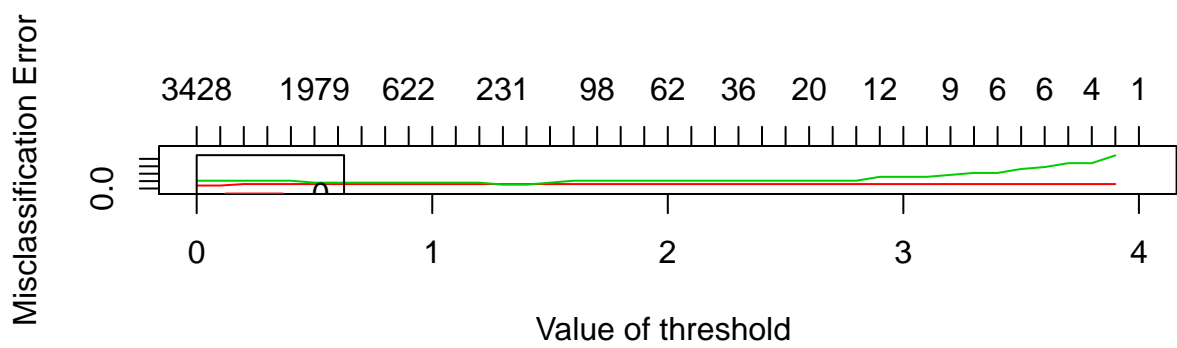
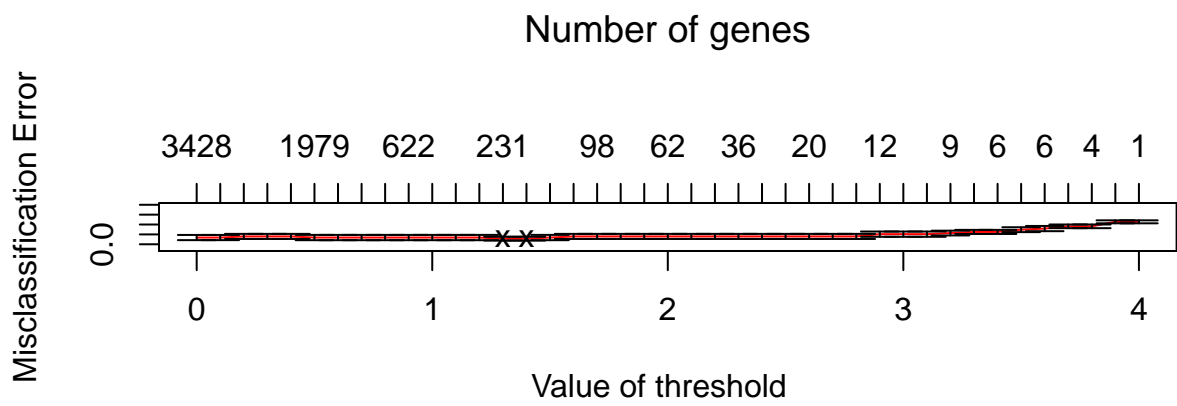
#### Q1

```
## 1234567891011121314151617181920212223242526272829303132333435363738394041
```

```
## 12Fold 1 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 2 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 3 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 4 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 5 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 6 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 7 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 8 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 9 :1234567891011121314151617181920212223242526272829303132333435363738394041
## Fold 10 :1234567891011121314151617181920212223242526272829303132333435363738394041
```

```
## Call:
```

```
## pamr.cv(fit = model, data = myemailtrain)
##      threshold nonzero errors
## 1  0.0         3428     6
## 2  0.1         3409     6
## 3  0.2         3114     7
## 4  0.3         3024     7
## 5  0.4         3000     7
## 6  0.5         1979     6
## 7  0.6          852     6
## 8  0.7          841     6
## 9  0.8          673     6
## 10 0.9          622     6
## 11 1.0          297     6
## 12 1.1          293     6
## 13 1.2          272     6
## 14 1.3          231     5
## 15 1.4          170     5
## 16 1.5          138     6
## 17 1.6          129     7
## 18 1.7           98     7
## 19 1.8           88     7
## 20 1.9           71     7
## 21 2.0           62     7
## 22 2.1           47     7
## 23 2.2           43     7
## 24 2.3           36     7
## 25 2.4           30     7
## 26 2.5           20     7
## 27 2.6           20     7
## 28 2.7           14     7
## 29 2.8           12     7
## 30 2.9           12     9
## 31 3.0           12     9
## 32 3.1           11     9
## 33 3.2            9    10
## 34 3.3            9    11
## 35 3.4            6    11
## 36 3.5            6    13
## 37 3.6            6    14
## 38 3.7            6    16
## 39 3.8            4    16
## 40 3.9            2    20
## 41 4.0            1    20
```





```

## [26,] 2463 -0.141 0.1856
## [27,] 329 -0.1349 0.1774
## [28,] 681 -0.1349 0.1774
## [29,] 1891 -0.1349 0.1774
## [30,] 3243 -0.1349 0.1774
## [31,] 283 -0.1268 0.1669
## [32,] 4628 -0.1268 0.1669
## [33,] 3286 -0.1268 0.1669
## [34,] 3274 -0.1237 0.1627
## [35,] 810 -0.1237 0.1627
## [36,] 2889 -0.1237 0.1627
## [37,] 1233 0.1141 -0.1501
## [38,] 3188 0.1141 -0.1501
## [39,] 3191 0.1141 -0.1501
## [40,] 3312 0.1141 -0.1501
## [41,] 3891 0.1133 -0.1491
## [42,] 3458 0.1133 -0.1491
## [43,] 3324 -0.11 0.1447
## [44,] 1643 -0.0946 0.1244
## [45,] 2561 -0.0946 0.1244
## [46,] 3090 -0.0946 0.1244
## [47,] 4629 -0.0946 0.1244
## [48,] 606 0.091 -0.1197
## [49,] 2058 -0.0881 0.1159
## [50,] 1501 0.0881 -0.1159
## [51,] 3952 -0.0869 0.1143
## [52,] 680 -0.0869 0.1143
## [53,] 3836 -0.0869 0.1143
## [54,] 1061 -0.0867 0.1141
## [55,] 1007 0.0864 -0.1137
## [56,] 1477 0.0864 -0.1137
## [57,] 2103 0.0864 -0.1137
## [58,] 3992 0.0864 -0.1137
## [59,] 2295 -0.084 0.1105
## [60,] 4061 -0.084 0.1105
## [61,] 2305 -0.0838 0.1103
## [62,] 3285 -0.0838 0.1103
## [63,] 92 -0.07 0.0921
## [64,] 1127 -0.07 0.0921
## [65,] 2583 -0.07 0.0921
## [66,] 3323 -0.07 0.0921
## [67,] 4500 -0.07 0.0921
## [68,] 1698 -0.07 0.0921
## [69,] 3241 -0.07 0.0921
## [70,] 4364 -0.07 0.0921
## [71,] 4062 -0.0665 0.0875
## [72,] 4039 0.0626 -0.0823
## [73,] 740 0.059 -0.0776
## [74,] 2438 0.059 -0.0776
## [75,] 2442 0.059 -0.0776
## [76,] 3311 0.059 -0.0776
## [77,] 3383 0.059 -0.0776
## [78,] 3559 0.059 -0.0776
## [79,] 4176 0.059 -0.0776

```



```

## [80,] 4402 0.059 -0.0776
## [81,] 267 0.057 -0.075
## [82,] 2553 0.057 -0.075
## [83,] 63 -0.0549 0.0723
## [84,] 1563 -0.0549 0.0723
## [85,] 1594 -0.0549 0.0723
## [86,] 3589 -0.0549 0.0723
## [87,] 3882 -0.0549 0.0723
## [88,] 4365 -0.0549 0.0723
## [89,] 3301 -0.048 0.0632
## [90,] 1636 -0.0478 0.0629
## [91,] 1072 -0.0478 0.0629
## [92,] 386 -0.0478 0.0629
## [93,] 2198 -0.0455 0.0599
## [94,] 3021 -0.0455 0.0599
## [95,] 3386 -0.0455 0.0599
## [96,] 76 -0.0452 0.0594
## [97,] 2150 -0.0452 0.0594
## [98,] 4075 0.0448 -0.0589
## [99,] 107 0.0316 -0.0416
## [100,] 336 0.0316 -0.0416
## [101,] 776 0.0316 -0.0416
## [102,] 831 0.0316 -0.0416
## [103,] 1088 0.0316 -0.0416
## [104,] 1450 0.0316 -0.0416
## [105,] 1456 0.0316 -0.0416
## [106,] 1542 0.0316 -0.0416
## [107,] 2170 0.0316 -0.0416
## [108,] 2613 0.0316 -0.0416
## [109,] 2837 0.0316 -0.0416
## [110,] 4529 0.0316 -0.0416
## [111,] 363 -0.0297 0.0391
## [112,] 879 -0.0297 0.0391
## [113,] 2433 -0.0297 0.0391
## [114,] 3051 -0.0297 0.0391
## [115,] 3514 -0.0297 0.0391
## [116,] 3711 -0.0297 0.0391
## [117,] 4449 -0.0297 0.0391
## [118,] 501 -0.0297 0.0391
## [119,] 803 -0.0297 0.0391
## [120,] 2046 -0.0297 0.0391
## [121,] 2082 -0.0297 0.0391
## [122,] 2690 -0.0297 0.0391
## [123,] 2877 -0.0297 0.0391
## [124,] 3118 -0.0297 0.0391
## [125,] 4342 -0.0297 0.0391
## [126,] 4451 -0.0297 0.0391
## [127,] 4452 -0.0297 0.0391
## [128,] 272 0.0294 -0.0386
## [129,] 2175 -0.0276 0.0364
## [130,] 3515 0.017 -0.0224
## [131,] 172 -0.0152 0.02
## [132,] 1149 -0.0152 0.02
## [133,] 2219 -0.0152 0.02

```

```

## [134,] 2964 -0.0152 0.02
## [135,] 2984 -0.0152 0.02
## [136,] 2887 -0.0152 0.02
## [137,] 4605 -0.0152 0.02
## [138,] 4064 -0.0149 0.0196
## [139,] 3800 -0.0106 0.014
## [140,] 134 -0.0091 0.0119
## [141,] 919 -0.0091 0.0119
## [142,] 3957 -0.0091 0.0119
## [143,] 4268 -0.0091 0.0119
## [144,] 4281 -0.0091 0.0119
## [145,] 2220 -0.0079 0.0104
## [146,] 2847 -0.0079 0.0104
## [147,] 3582 -0.0079 0.0104
## [148,] 4181 -0.0079 0.0104
## [149,] 2167 -0.0073 0.0096
## [150,] 67 0.0073 -0.0095
## [151,] 2005 -0.0071 0.0094
## [152,] 4185 -0.0071 0.0094
## [153,] 3588 -0.0071 0.0094
## [154,] 3794 -0.0071 0.0094
## [155,] 579 0.0038 -0.005
## [156,] 1147 0.0038 -0.005
## [157,] 1524 0.0038 -0.005
## [158,] 1591 0.0038 -0.005
## [159,] 1702 0.0038 -0.005
## [160,] 1797 0.0038 -0.005
## [161,] 2141 0.0038 -0.005
## [162,] 2251 0.0038 -0.005
## [163,] 2278 0.0038 -0.005
## [164,] 2619 0.0038 -0.005
## [165,] 3194 0.0038 -0.005
## [166,] 340 0.0038 -0.005
## [167,] 2894 0.0024 -0.0032
## [168,] 1144 0.0017 -0.0022
## [169,] 2392 0.0017 -0.0022
## [170,] 3295 0.0017 -0.0022

```

```

##      predicted
## ytest 0 1
##      0 10 0
##      1 2 8

```

```

## [1] "The misclassification rate is 0.1"

```

```

##      id 0-score 1-score
## [1,] 3036 -0.369 0.4856
## [2,] 2049 -0.3396 0.4468
## [3,] 4060 -0.3244 0.4269
## [4,] 1262 -0.3178 0.4181
## [5,] 3364 -0.31 0.4079
## [6,] 3187 0.3056 -0.4022
## [7,] 596 -0.2593 0.3412

```

```

## [8,] 869 -0.2574 0.3387
## [9,] 1045 -0.2574 0.3387
## [10,] 607 0.2344 -0.3085
## [11,] 4282 -0.2252 0.2963
## [12,] 2990 -0.2123 0.2793
## [13,] 599 -0.1765 0.2322
## [14,] 3433 -0.1765 0.2322
## [15,] 389 -0.1684 0.2216
## [16,] 2588 -0.1684 0.2216
## [17,] 3022 -0.1684 0.2216
## [18,] 850 0.1661 -0.2186
## [19,] 3725 0.1661 -0.2186
## [20,] 3035 -0.1654 0.2176
## [21,] 4129 -0.1427 0.1878
## [22,] 3125 0.1427 -0.1878
## [23,] 4177 0.1424 -0.1874
## [24,] 3671 0.1424 -0.1874
## [25,] 2974 -0.141 0.1856
## [26,] 2463 -0.141 0.1856
## [27,] 329 -0.1349 0.1774
## [28,] 681 -0.1349 0.1774
## [29,] 1891 -0.1349 0.1774
## [30,] 3243 -0.1349 0.1774
## [31,] 283 -0.1268 0.1669
## [32,] 4628 -0.1268 0.1669
## [33,] 3286 -0.1268 0.1669
## [34,] 3274 -0.1237 0.1627
## [35,] 810 -0.1237 0.1627
## [36,] 2889 -0.1237 0.1627
## [37,] 1233 0.1141 -0.1501
## [38,] 3188 0.1141 -0.1501
## [39,] 3191 0.1141 -0.1501
## [40,] 3312 0.1141 -0.1501
## [41,] 3891 0.1133 -0.1491
## [42,] 3458 0.1133 -0.1491
## [43,] 3324 -0.11 0.1447
## [44,] 1643 -0.0946 0.1244
## [45,] 2561 -0.0946 0.1244
## [46,] 3090 -0.0946 0.1244
## [47,] 4629 -0.0946 0.1244
## [48,] 606 0.091 -0.1197
## [49,] 2058 -0.0881 0.1159
## [50,] 1501 0.0881 -0.1159
## [51,] 3952 -0.0869 0.1143
## [52,] 680 -0.0869 0.1143
## [53,] 3836 -0.0869 0.1143
## [54,] 1061 -0.0867 0.1141
## [55,] 1007 0.0864 -0.1137
## [56,] 1477 0.0864 -0.1137
## [57,] 2103 0.0864 -0.1137
## [58,] 3992 0.0864 -0.1137
## [59,] 2295 -0.084 0.1105
## [60,] 4061 -0.084 0.1105
## [61,] 2305 -0.0838 0.1103

```

```

## [62,] 3285 -0.0838 0.1103
## [63,] 92 -0.07 0.0921
## [64,] 1127 -0.07 0.0921
## [65,] 2583 -0.07 0.0921
## [66,] 3323 -0.07 0.0921
## [67,] 4500 -0.07 0.0921
## [68,] 1698 -0.07 0.0921
## [69,] 3241 -0.07 0.0921
## [70,] 4364 -0.07 0.0921
## [71,] 4062 -0.0665 0.0875
## [72,] 4039 0.0626 -0.0823
## [73,] 740 0.059 -0.0776
## [74,] 2438 0.059 -0.0776
## [75,] 2442 0.059 -0.0776
## [76,] 3311 0.059 -0.0776
## [77,] 3383 0.059 -0.0776
## [78,] 3559 0.059 -0.0776
## [79,] 4176 0.059 -0.0776
## [80,] 4402 0.059 -0.0776
## [81,] 267 0.057 -0.075
## [82,] 2553 0.057 -0.075
## [83,] 63 -0.0549 0.0723
## [84,] 1563 -0.0549 0.0723
## [85,] 1594 -0.0549 0.0723
## [86,] 3589 -0.0549 0.0723
## [87,] 3882 -0.0549 0.0723
## [88,] 4365 -0.0549 0.0723
## [89,] 3301 -0.048 0.0632
## [90,] 1636 -0.0478 0.0629
## [91,] 1072 -0.0478 0.0629
## [92,] 386 -0.0478 0.0629
## [93,] 2198 -0.0455 0.0599
## [94,] 3021 -0.0455 0.0599
## [95,] 3386 -0.0455 0.0599
## [96,] 76 -0.0452 0.0594
## [97,] 2150 -0.0452 0.0594
## [98,] 4075 0.0448 -0.0589
## [99,] 107 0.0316 -0.0416
## [100,] 336 0.0316 -0.0416
## [101,] 776 0.0316 -0.0416
## [102,] 831 0.0316 -0.0416
## [103,] 1088 0.0316 -0.0416
## [104,] 1450 0.0316 -0.0416
## [105,] 1456 0.0316 -0.0416
## [106,] 1542 0.0316 -0.0416
## [107,] 2170 0.0316 -0.0416
## [108,] 2613 0.0316 -0.0416
## [109,] 2837 0.0316 -0.0416
## [110,] 4529 0.0316 -0.0416
## [111,] 363 -0.0297 0.0391
## [112,] 879 -0.0297 0.0391
## [113,] 2433 -0.0297 0.0391
## [114,] 3051 -0.0297 0.0391
## [115,] 3514 -0.0297 0.0391

```

```

## [116,] 3711 -0.0297 0.0391
## [117,] 4449 -0.0297 0.0391
## [118,] 501 -0.0297 0.0391
## [119,] 803 -0.0297 0.0391
## [120,] 2046 -0.0297 0.0391
## [121,] 2082 -0.0297 0.0391
## [122,] 2690 -0.0297 0.0391
## [123,] 2877 -0.0297 0.0391
## [124,] 3118 -0.0297 0.0391
## [125,] 4342 -0.0297 0.0391
## [126,] 4451 -0.0297 0.0391
## [127,] 4452 -0.0297 0.0391
## [128,] 272 0.0294 -0.0386
## [129,] 2175 -0.0276 0.0364
## [130,] 3515 0.017 -0.0224
## [131,] 172 -0.0152 0.02
## [132,] 1149 -0.0152 0.02
## [133,] 2219 -0.0152 0.02
## [134,] 2964 -0.0152 0.02
## [135,] 2984 -0.0152 0.02
## [136,] 2887 -0.0152 0.02
## [137,] 4605 -0.0152 0.02
## [138,] 4064 -0.0149 0.0196
## [139,] 3800 -0.0106 0.014
## [140,] 134 -0.0091 0.0119
## [141,] 919 -0.0091 0.0119
## [142,] 3957 -0.0091 0.0119
## [143,] 4268 -0.0091 0.0119
## [144,] 4281 -0.0091 0.0119
## [145,] 2220 -0.0079 0.0104
## [146,] 2847 -0.0079 0.0104
## [147,] 3582 -0.0079 0.0104
## [148,] 4181 -0.0079 0.0104
## [149,] 2167 -0.0073 0.0096
## [150,] 67 0.0073 -0.0095
## [151,] 2005 -0.0071 0.0094
## [152,] 4185 -0.0071 0.0094
## [153,] 3588 -0.0071 0.0094
## [154,] 3794 -0.0071 0.0094
## [155,] 579 0.0038 -0.005
## [156,] 1147 0.0038 -0.005
## [157,] 1524 0.0038 -0.005
## [158,] 1591 0.0038 -0.005
## [159,] 1702 0.0038 -0.005
## [160,] 1797 0.0038 -0.005
## [161,] 2141 0.0038 -0.005
## [162,] 2251 0.0038 -0.005
## [163,] 2278 0.0038 -0.005
## [164,] 2619 0.0038 -0.005
## [165,] 3194 0.0038 -0.005
## [166,] 340 0.0038 -0.005
## [167,] 2894 0.0024 -0.0032
## [168,] 1144 0.0017 -0.0022
## [169,] 2392 0.0017 -0.0022

```

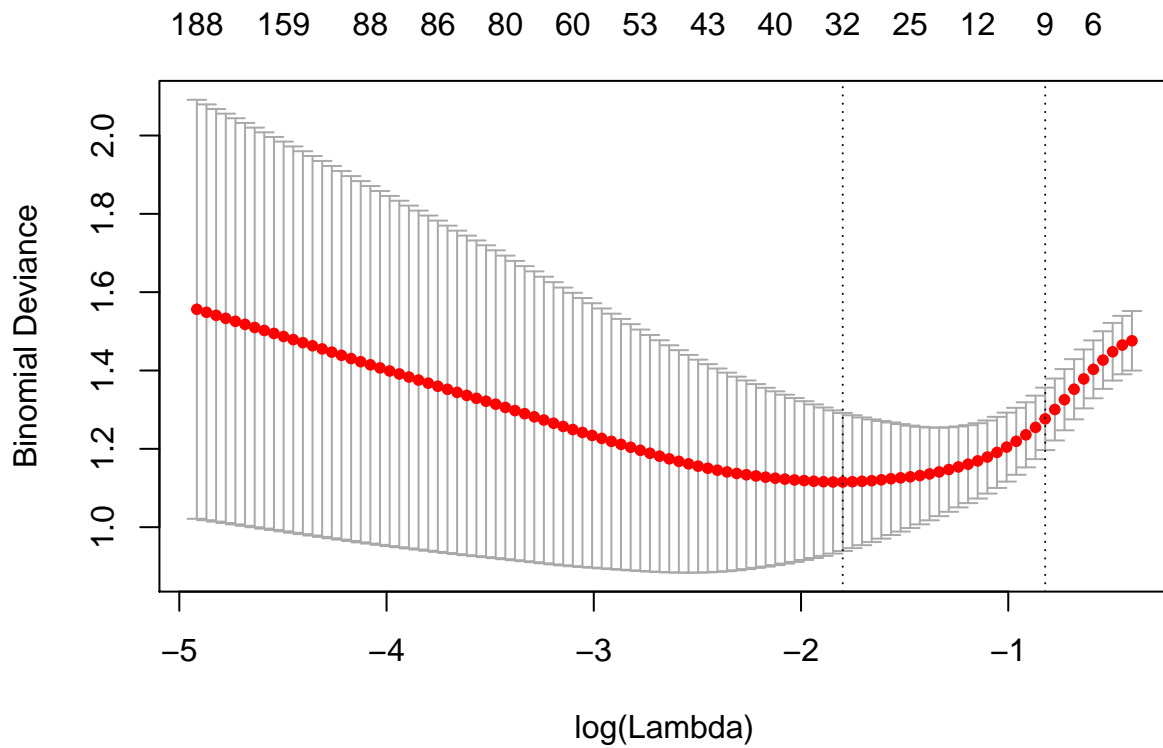
```
## [170,] 3295 0.0017 -0.0022
```

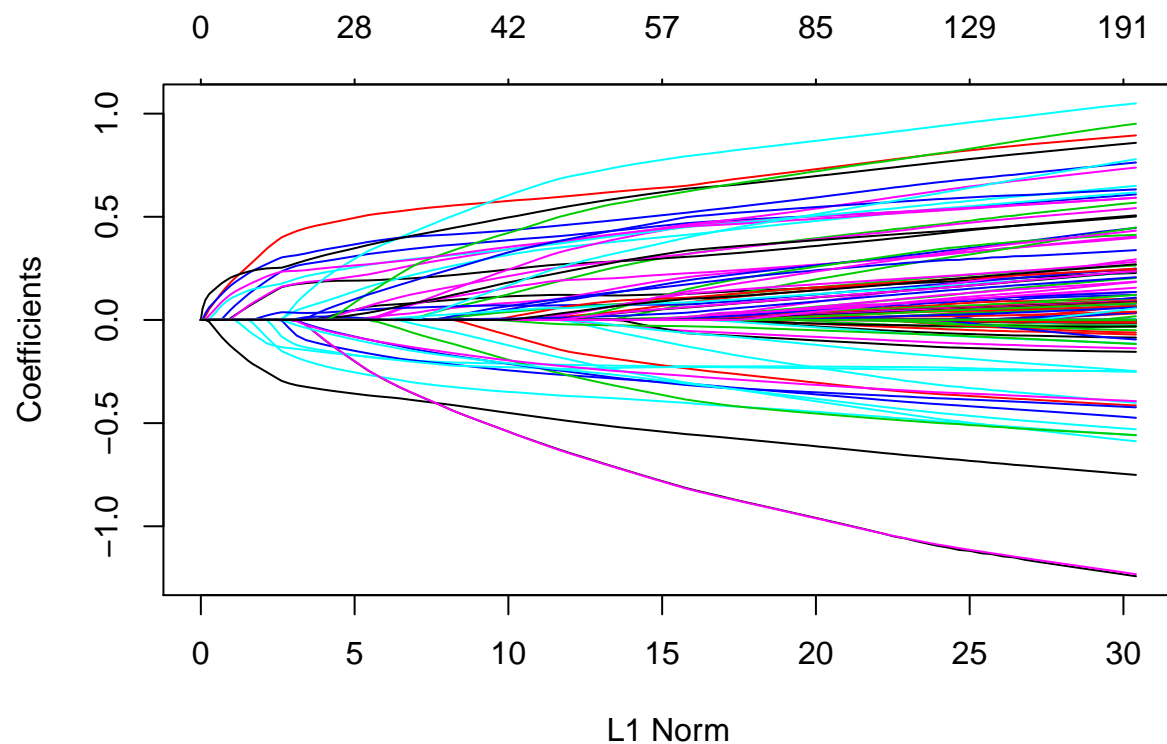
Table 1: Top 10 Important features by NSC

x
acceptance
X59â
adhere
X1st
acquiring
accessibility
agenda
aicit
X5102011
agents

From the plot generated of threshold vs misclassification error. It is observed that when the threshold value is 1.4, the misclassification error is at its lowest. 170 features were selected by this model and top 10 features are listed below. The misclassification error rate is 10%. The confusion matrix reveals that ‘everything else’ is classified with 10/10 times while ‘announces of conferences’ is classified 8/10 times.

Q2a





```
##      elasticpredict
## ytest2  0  1
##      0 10  0
##      1  2  8
```

```
## [1] "The misclassification rate is 0.1"
```

The Elastic net model has a misclassification error rate of 10%. This model selects the least number of features i.e 39 features.

## Q2b

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

```
##      Predicted svm
## Actual Test  0  1
##      0 10  0
##      1  1  9
```

```
## [1] "The misclassification rate is 0.05"
```

Table 2: Contributing features of elastic net model

x
(Intercept)
abstracts
aspects
bio
call
candidates
computer
conceptual
conference
dates
due
evaluation
exhibits
important
languages
making
manuscripts
original
papers
peer
position
process
projects
proposals
published
queries
record
relevant
scenarios
spatial
submission
team
versions



Table 3: Comparison of the models

	Nearest Shrunken Centroid Model	ElasticNet Model	SVM Model
Accuracy	90.0	90.0	95.00
Number of Features	170.0	33.0	43.00
Misclassification error rate	0.1	0.1	0.05

The SVM model can be chosen as the misclassification error rate is the least when tested on unknown data and the number of features are also close to the minimum of the three models i.e 43 features selected.

Q3

```
## [1] 0.0003765147
```

```
##          pvalue variable status Variable_name
## 1  1.116910e-10    3036  FALSE      papers
## 2  7.949969e-10    4060  FALSE    submission
## 3  8.219362e-09    3187  FALSE      position
## 4  1.835157e-07    3364  FALSE     published
## 5  3.040833e-07    2049  FALSE     important
## 6  3.983540e-07     596  FALSE        call
## 7  5.091970e-07     869  FALSE    conference
## 8  8.612259e-07     607  FALSE    candidates
## 9  1.398619e-06    1045  FALSE        dates
## 10 1.398619e-06    3035  FALSE        paper
## 11 5.068373e-06    4282  FALSE        topics
## 12 7.907976e-06    2463  FALSE      limited
## 13 1.190607e-05     606  FALSE    candidate
## 14 2.099119e-05     599  FALSE      camera
## 15 2.099119e-05    3433  FALSE      ready
## 16 2.154461e-05     389  FALSE    authors
## 17 3.382671e-05    3125  FALSE        phd
## 18 3.499123e-05    3312  FALSE    projects
## 19 3.742010e-05    2974  FALSE        org
## 20 5.860175e-05     681  FALSE    chairs
## 21 6.488781e-05    1262  FALSE        due
## 22 6.488781e-05    2990  FALSE    original
## 23 6.882210e-05    2889  FALSE    notification
## 24 7.971981e-05    3671  FALSE      salary
## 25 9.090038e-05    3458  FALSE      record
## 26 9.090038e-05    3891  FALSE    skills
## 27 1.529174e-04    1891  FALSE      held
## 28 1.757570e-04    4177  FALSE      team
## 29 2.007353e-04    3022  FALSE      pages
## 30 2.007353e-04    4628  FALSE    workshop
## 31 2.117020e-04     810  FALSE    committee
## 32 2.117020e-04    3285  FALSE    proceedings
## 33 2.166414e-04     272  FALSE      apply
## 34 2.246309e-04    4039  FALSE    strong
## 35 2.295684e-04    2175  FALSE    international
## 36 3.762328e-04    1088  FALSE      degree
## 37 3.762328e-04    1477  FALSE    excellent
```

## 38	3.762328e-04	3191	FALSE	post
## 39	3.765147e-04	3243	FALSE	presented
## 40	4.638692e-04	2588	TRUE	march
## 41	4.952306e-04	267	TRUE	applicants
## 42	5.380303e-04	3274	TRUE	privacy
## 43	5.876764e-04	4061	TRUE	submissions
## 44	6.063457e-04	1061	TRUE	deadline
## 45	7.844017e-04	1233	TRUE	doctoral
## 46	7.844017e-04	2438	TRUE	letter
## 47	7.844017e-04	3188	TRUE	positions
## 48	7.844017e-04	3383	TRUE	qualifications
## 49	8.815982e-04	1563	TRUE	february
## 50	8.815982e-04	1643	TRUE	forum
## 51	8.815982e-04	4629	TRUE	workshops
## 52	1.111221e-03	4129	TRUE	systems
## 53	1.125532e-03	329	TRUE	aspects
## 54	1.125532e-03	680	TRUE	chair
## 55	1.340994e-03	2728	TRUE	mobile
## 56	1.340994e-03	3952	TRUE	special
## 57	1.385179e-03	3324	TRUE	proposals
## 58	1.385179e-03	4451	TRUE	usa
## 59	1.408380e-03	1501	TRUE	experience
## 60	1.500237e-03	77	TRUE	accepted
## 61	1.500237e-03	2847	TRUE	networks
## 62	1.595112e-03	3725	TRUE	science
## 63	1.597427e-03	1007	TRUE	curriculum
## 64	1.597427e-03	1702	TRUE	funded
## 65	1.597427e-03	2251	TRUE	java
## 66	1.597427e-03	2442	TRUE	levels
## 67	1.597427e-03	4176	TRUE	teaching
## 68	2.321154e-03	3311	TRUE	project
## 69	2.588482e-03	283	TRUE	april
## 70	2.588482e-03	386	TRUE	author
## 71	2.588482e-03	3836	TRUE	short
## 72	2.879855e-03	92	TRUE	acm
## 73	2.879855e-03	3323	TRUE	proposal
## 74	2.879855e-03	3361	TRUE	publicity
## 75	3.186651e-03	336	TRUE	assistant
## 76	3.186651e-03	756	TRUE	closing
## 77	3.186651e-03	831	TRUE	competitive
## 78	3.186651e-03	1450	TRUE	european
## 79	3.186651e-03	1797	TRUE	graduate
## 80	3.186651e-03	2613	TRUE	master
## 81	3.186651e-03	4426	TRUE	universities
## 82	3.622243e-03	4062	TRUE	submit
## 83	3.825379e-03	2198	TRUE	invited
## 84	3.843991e-03	791	TRUE	com
## 85	3.843991e-03	3301	TRUE	program
## 86	3.917877e-03	850	TRUE	computer
## 87	3.917877e-03	3755	TRUE	security
## 88	4.791732e-03	413	TRUE	background
## 89	4.791732e-03	3992	TRUE	starting
## 90	5.156010e-03	2058	TRUE	include
## 91	5.678572e-03	2177	TRUE	internet

## 92	5.678572e-03	3090	TRUE	peer
## 93	5.832310e-03	603	TRUE	canada
## 94	5.832310e-03	1818	TRUE	grid
## 95	5.832310e-03	2986	TRUE	organizing
## 96	5.832310e-03	3216	TRUE	practitioners
## 97	5.832310e-03	4364	TRUE	tutorial
## 98	5.832310e-03	4500	TRUE	versions
## 99	6.245001e-03	107	TRUE	activities
## 100	6.245001e-03	340	TRUE	associate
## 101	6.245001e-03	1424	TRUE	equal
## 102	6.245001e-03	3194	TRUE	postdoctoral
## 103	6.245001e-03	4529	TRUE	vitae
## 104	6.939558e-03	76	TRUE	acceptance
## 105	6.939558e-03	1636	TRUE	format
## 106	6.939558e-03	3794	TRUE	series
## 107	7.927262e-03	1743	TRUE	general
## 108	7.927262e-03	2220	TRUE	issues
## 109	8.463230e-03	899	TRUE	contact
## 110	8.581177e-03	80	TRUE	access
## 111	8.600080e-03	4075	TRUE	successful
## 112	8.884208e-03	2295	TRUE	journal
## 113	8.884208e-03	3800	TRUE	services
## 114	8.898143e-03	815	TRUE	communications
## 115	8.898143e-03	1662	TRUE	france
## 116	8.898143e-03	2984	TRUE	organizers
## 117	8.898143e-03	3589	TRUE	reviewed
## 118	8.898143e-03	3882	TRUE	site
## 119	8.898143e-03	4605	TRUE	wireless
## 120	9.619705e-03	2170	TRUE	interests
## 121	9.619705e-03	4045	TRUE	students
## 122	9.619705e-03	4402	TRUE	undergraduate
## 123	1.018054e-02	67	TRUE	academic
## 124	1.018054e-02	3306	TRUE	programming
## 125	1.078220e-02	2553	TRUE	mail
## 126	1.155545e-02	803	TRUE	commerce
## 127	1.155545e-02	879	TRUE	conjunction
## 128	1.155545e-02	1291	TRUE	economics
## 129	1.155545e-02	2046	TRUE	implementations
## 130	1.155545e-02	2433	TRUE	length
## 131	1.155545e-02	2583	TRUE	manuscripts
## 132	1.155545e-02	2690	TRUE	michael
## 133	1.155545e-02	3241	TRUE	presentation
## 134	1.155545e-02	3514	TRUE	relevance
## 135	1.155545e-02	3943	TRUE	spain
## 136	1.155545e-02	4452	TRUE	usability
## 137	1.155545e-02	4606	TRUE	wisconsin
## 138	1.194587e-02	919	TRUE	contributions
## 139	1.194587e-02	3898	TRUE	smart
## 140	1.206071e-02	1139	TRUE	detailed
## 141	1.206071e-02	1372	TRUE	employer
## 142	1.206071e-02	1524	TRUE	extension
## 143	1.206071e-02	2141	TRUE	institutions
## 144	1.206071e-02	2278	TRUE	job
## 145	1.206071e-02	2770	TRUE	motivated

## 146	1.302056e-02	1498	TRUE	expected
## 147	1.459963e-02	2005	TRUE	ideas
## 148	1.666073e-02	2305	TRUE	june
## 149	1.666073e-02	3021	TRUE	page
## 150	1.666073e-02	3582	TRUE	results
## 151	1.804880e-02	172	TRUE	aims
## 152	1.804880e-02	1594	TRUE	final
## 153	1.804880e-02	2219	TRUE	issue
## 154	1.804880e-02	2964	TRUE	optimization
## 155	1.804880e-02	3040	TRUE	parallel
## 156	1.804880e-02	3242	TRUE	presentations
## 157	1.804880e-02	4365	TRUE	tutorials
## 158	1.804880e-02	4377	TRUE	ubiquitous
## 159	1.805157e-02	3286	TRUE	process
## 160	1.882562e-02	2619	TRUE	mathematics
## 161	1.882562e-02	3559	TRUE	researcher
## 162	1.882562e-02	3994	TRUE	statement
## 163	1.951854e-02	1044	TRUE	date
## 164	2.080810e-02	2961	TRUE	opportunity
## 165	2.080810e-02	3295	TRUE	professor
## 166	2.250253e-02	598	TRUE	calls
## 167	2.250253e-02	2359	TRUE	korea
## 168	2.250253e-02	2877	TRUE	non
## 169	2.250253e-02	3196	TRUE	poster
## 170	2.250253e-02	3333	TRUE	protocols
## 171	2.250253e-02	4212	TRUE	term
## 172	2.250253e-02	4435	TRUE	unpublished
## 173	2.250253e-02	4526	TRUE	visualization
## 174	2.250253e-02	4664	TRUE	yang
## 175	2.303834e-02	1591	TRUE	filled
## 176	2.303834e-02	3297	TRUE	proficiency
## 177	2.303834e-02	3991	TRUE	start
## 178	2.420131e-02	2561	TRUE	making
## 179	2.426178e-02	4427	TRUE	university
## 180	2.426178e-02	1395	TRUE	english
## 181	2.712967e-02	2167	TRUE	interest
## 182	2.934388e-02	757	TRUE	cloud
## 183	2.934388e-02	981	TRUE	cross
## 184	2.934388e-02	4281	TRUE	topic
## 185	3.052161e-02	3515	TRUE	relevant
## 186	3.573002e-02	63	TRUE	abstract
## 187	3.573002e-02	1127	TRUE	describing
## 188	3.573002e-02	1149	TRUE	developments
## 189	3.573002e-02	1766	TRUE	germany
## 190	3.573002e-02	2059	TRUE	included
## 191	3.573002e-02	2248	TRUE	japan
## 192	3.573002e-02	2643	TRUE	media
## 193	3.573002e-02	2887	TRUE	notes
## 194	3.573002e-02	3570	TRUE	resource
## 195	3.573002e-02	3732	TRUE	scope
## 196	3.573002e-02	3816	TRUE	share
## 197	3.573002e-02	4342	TRUE	trust
## 198	3.582313e-02	2735	TRUE	models
## 199	3.598686e-02	1144	TRUE	develop

##	200	3.720192e-02	4181	TRUE	technical
##	201	3.883086e-02	4280	TRUE	top
##	202	4.331307e-02	155	TRUE	agents
##	203	4.331307e-02	196	TRUE	allowed
##	204	4.331307e-02	311	TRUE	arrangements
##	205	4.331307e-02	501	TRUE	bio
##	206	4.331307e-02	856	TRUE	concepts
##	207	4.331307e-02	940	TRUE	copyright
##	208	4.331307e-02	967	TRUE	covering
##	209	4.331307e-02	1048	TRUE	david
##	210	4.331307e-02	1560	TRUE	feature
##	211	4.331307e-02	1587	TRUE	figures
##	212	4.331307e-02	1708	TRUE	fusion
##	213	4.331307e-02	1814	TRUE	green
##	214	4.331307e-02	1859	TRUE	hand
##	215	4.331307e-02	1861	TRUE	handled
##	216	4.331307e-02	2082	TRUE	india
##	217	4.331307e-02	2110	TRUE	infrastructures
##	218	4.331307e-02	2197	TRUE	invite
##	219	4.331307e-02	2274	TRUE	jin
##	220	4.331307e-02	2332	TRUE	kevin
##	221	4.331307e-02	2334	TRUE	keynote
##	222	4.331307e-02	2481	TRUE	liu
##	223	4.331307e-02	2546	TRUE	madison
##	224	4.331307e-02	2723	TRUE	mit
##	225	4.331307e-02	2810	TRUE	nanyang
##	226	4.331307e-02	2948	TRUE	ontologies
##	227	4.331307e-02	3051	TRUE	participants
##	228	4.331307e-02	3199	TRUE	posting
##	229	4.331307e-02	3259	TRUE	pricing
##	230	4.331307e-02	3591	TRUE	reviewing
##	231	4.331307e-02	3703	TRUE	scalability
##	232	4.331307e-02	3711	TRUE	scenarios
##	233	4.331307e-02	3802	TRUE	sessions
##	234	4.331307e-02	4142	TRUE	taiwan
##	235	4.331307e-02	4145	TRUE	takes
##	236	4.331307e-02	4202	TRUE	template
##	237	4.331307e-02	4303	TRUE	tracks
##	238	4.331307e-02	4423	TRUE	universite
##	239	4.331307e-02	4499	TRUE	version
##	240	4.331307e-02	4506	TRUE	vienna
##	241	4.331307e-02	4553	TRUE	wang
##	242	4.373782e-02	103	TRUE	action
##	243	4.373782e-02	147	TRUE	affirmative
##	244	4.373782e-02	275	TRUE	appointment
##	245	4.373782e-02	454	TRUE	beginning
##	246	4.373782e-02	630	TRUE	carry
##	247	4.373782e-02	806	TRUE	commission
##	248	4.373782e-02	821	TRUE	company
##	249	4.373782e-02	865	TRUE	conduct
##	250	4.373782e-02	913	TRUE	contract
##	251	4.373782e-02	1089	TRUE	degrees
##	252	4.373782e-02	1134	TRUE	desirable
##	253	4.373782e-02	1178	TRUE	directly

##	254	4.373782e-02	1230	TRUE	doc
##	255	4.373782e-02	1429	TRUE	equivalent
##	256	4.373782e-02	1651	TRUE	foundation
##	257	4.373782e-02	1660	TRUE	fp7
##	258	4.373782e-02	1913	TRUE	hiring
##	259	4.373782e-02	2115	TRUE	initially
##	260	4.373782e-02	2140	TRUE	institution
##	261	4.373782e-02	2279	TRUE	jobs
##	262	4.373782e-02	2396	TRUE	largest
##	263	4.373782e-02	2510	TRUE	looking
##	264	4.373782e-02	2663	TRUE	member
##	265	4.373782e-02	2760	TRUE	months
##	266	4.373782e-02	2924	TRUE	obtained
##	267	4.373782e-02	2945	TRUE	ongoing
##	268	4.373782e-02	3053	TRUE	participated
##	269	4.373782e-02	3339	TRUE	proven
##	270	4.373782e-02	3469	TRUE	ref
##	271	4.373782e-02	4095	TRUE	supervision
##	272	4.373782e-02	4163	TRUE	tasks
##	273	4.373782e-02	4209	TRUE	tenure
##	274	4.373782e-02	4239	TRUE	thesis
##	275	4.373782e-02	4615	TRUE	women
##	276	4.737923e-02	134	TRUE	advanced
##	277	4.737923e-02	318	TRUE	artificial
##	278	4.737923e-02	381	TRUE	australia
##	279	4.737923e-02	708	TRUE	chen
##	280	4.737923e-02	3973	TRUE	springer
##	281	4.737923e-02	4268	TRUE	title

##		pvalue	variable	status	Variable_name
##	1	1.116910e-10	3036	FALSE	papers
##	2	7.949969e-10	4060	FALSE	submission
##	3	8.219362e-09	3187	FALSE	position
##	4	1.835157e-07	3364	FALSE	published
##	5	3.040833e-07	2049	FALSE	important
##	6	3.983540e-07	596	FALSE	call
##	7	5.091970e-07	869	FALSE	conference
##	8	8.612259e-07	607	FALSE	candidates
##	9	1.398619e-06	1045	FALSE	dates
##	10	1.398619e-06	3035	FALSE	paper
##	11	5.068373e-06	4282	FALSE	topics
##	12	7.907976e-06	2463	FALSE	limited
##	13	1.190607e-05	606	FALSE	candidate
##	14	2.099119e-05	599	FALSE	camera
##	15	2.099119e-05	3433	FALSE	ready
##	16	2.154461e-05	389	FALSE	authors
##	17	3.382671e-05	3125	FALSE	phd
##	18	3.499123e-05	3312	FALSE	projects
##	19	3.742010e-05	2974	FALSE	org
##	20	5.860175e-05	681	FALSE	chairs
##	21	6.488781e-05	1262	FALSE	due
##	22	6.488781e-05	2990	FALSE	original
##	23	6.882210e-05	2889	FALSE	notification
##	24	7.971981e-05	3671	FALSE	salary

## 25	9.090038e-05	3458	FALSE	record
## 26	9.090038e-05	3891	FALSE	skills
## 27	1.529174e-04	1891	FALSE	held
## 28	1.757570e-04	4177	FALSE	team
## 29	2.007353e-04	3022	FALSE	pages
## 30	2.007353e-04	4628	FALSE	workshop
## 31	2.117020e-04	810	FALSE	committee
## 32	2.117020e-04	3285	FALSE	proceedings
## 33	2.166414e-04	272	FALSE	apply
## 34	2.246309e-04	4039	FALSE	strong
## 35	2.295684e-04	2175	FALSE	international
## 36	3.762328e-04	1088	FALSE	degree
## 37	3.762328e-04	1477	FALSE	excellent
## 38	3.762328e-04	3191	FALSE	post
## 39	3.765147e-04	3243	FALSE	presented

39 features correspond to the rejecting the null hypothesis, according to the BH rejection threshold. These contain variable names such as ‘notification’, ‘workshop’, ‘conference’, ‘candidates’, ‘published’, ‘topics’ to name a few of the 39 features. These reject that the null hypothesis that states that these features have no effect in the classification of into conference and non-conference.

From the first table , it is observed that 281 features have significant p values. Features such as ‘committee’, ‘conference’, ‘process’, ‘optimization’, ‘arrangements’ make sense in the usage.

## Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(plotly)
library(ggplot2)
library(xlsx)
library(readxl)
library(tidyr)
library(lubridate)
library(stringr)
library(mgcv)
library(gridExtra)
library(akima)
library(reshape)
library(pamr)
library(glmnet)
library(pROC)
library(kernlab)
library(e1071)
Influenza = read.xlsx("Influenza.xlsx",sheetName = "Raw data",header = TRUE)
Influenza$Date=date_decimal(Influenza$Time)
Influenza$influenzaratio<-((Influenza$Influenza)/(Influenza$Mortality))
p1<-ggplot(Influenza,aes(Date,Mortality))+geom_line(color="black")+scale_fill_brewer()+theme_classic()+
p1

p2<-ggplot(Influenza,aes(Date,Influenza))+geom_line(color="black")+scale_fill_brewer()+theme_classic()+
p2
```

```

p3<-ggplot(Influenza,aes(Date,influenzaratio))+geom_line(color="black")+scale_fill_brewer()+theme_classic()
p3

gammer<-mgcv::gam(data=Influenza, Mortality ~ Year + s(Week,k=length(unique(Influenza$Week))), method="GCV.Cp")
summary(gammer)

Influenza$gampredmortality<-mgcv::predict.gam(gammer,newdata = Influenza,type = "link")

p4<-ggplot(Influenza)+geom_line(aes(x=Date,y=gampredmortality),color="red",size=1)+geom_line(aes(x=Date,y=influenzaratio),color="black",size=1)
p4

gam.check(gammer,pch=19,cex=.3)
plot(gammer)
gammer1<-mgcv::gam(data=Influenza, Mortality ~ Year + s(Week,k=length(unique(Influenza$Week))))

s=interp(Influenza$Year, Influenza$Week, fitted(gammer1))
print(gammer1)
summary(gammer1)
gammer1$sp
#plot_ly(x=~s$x, y=~s$y, z=~s$z, type="surface")
knitr::include_graphics("surface.png")

modeldev <- NULL
for(sp in c(0.001, 0.01, 0.005, 2, 5))
{
  k=length(unique(Influenza$Week))
  gammod <- mgcv::gam(data = Influenza, Mortality~Year+s(Week, k=k, sp=sp), method = "GCV.Cp")
  temp <- cbind(gammod$deviance, gammod$fitted.values, gammod$y, Influenza$Date,
               sp, sum(influence(gammod)))
  modeldev <- rbind(temp, modeldev)
}

modeldev <- as.data.frame(modeldev)
colnames(modeldev) <- c("Deviance", "Mortalitypred", "Mortality", "Date",
                      "penaltyfactor", "dof")

modeldev$Date <- as.Date(modeldev$Date, origin = '1995-01-01')
#deviance plot
p5 <- ggplot(data=modeldev, aes(x = penaltyfactor, y = Deviance)) +geom_line() +theme_dark() +
ggtitle("Plot of Deviances of models vs. Penalty Factors")
p5
#degree of freedom plot
p6 <- ggplot(data=modeldev, aes(x = penaltyfactor, y = dof)) +geom_line() +theme_dark() +
ggtitle("Plot of Degree of freedoms of models vs. Penalty Factors")
p6

modeldevwide <- melt(modeldev[,c("Date", "penaltyfactor",
                                "Mortality", "Mortalitypred")],
                    id.vars = c("Date", "penaltyfactor"))

#predicted vs observed mortality

```



```

p7 <- ggplot(data=modeldevwide[modeldevwide$penaltyfactor == 0.001,], aes(x= Date, y = value)) +
  geom_line(aes(color = variable), size=1) +scale_fill_brewer() +theme_dark() +ggtitle("Plot of Mortali

p8 <- ggplot(data=modeldevwide[modeldevwide$penaltyfactor == 5,], aes(x= Date, y = value)) + geom_line(

grid.arrange(p7,p8,ncol=1)

Influenza$rez<-gammer$residuals
p9<-ggplot(Influenza,aes(x=Date))+geom_line(aes(y=rez,color="Residuals"))+geom_line(aes(y=Influenza,col
p9

lastgammod <- mgcv::gam(data = Influenza, Mortality~s(Year,k=length(unique(Influenza$Year)))+s(Week, k=

Influenza$lastgammodpred<-mgcv::predict.gam(lastgammod,newdata = Influenza,type = "link")

p10<-ggplot(Influenza,aes(x=Date))+geom_line(aes(y=lastgammodpred,color="PredictedMortality"),size=1.5)
p10

data<-read.csv2("data.csv",header = TRUE,sep=";")
email<-as.data.frame(data)
email$Conference<-as.factor(email$Conference)
rownames(email)=1:nrow(email)

n=dim(email)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.7))
train=email[id,]
test=email[-id,]
xtrain=t(train[, -4703])
ytrain=train[[4703]]
xtest=t(test[, -4703])
ytest=test[[4703]]
myemailtrain=list(x=xtrain,y=ytrain,geneid=as.character(1:nrow(xtrain)),genenames=rownames(xtrain))
myemailtest=list(x=xtest,y=ytest,geneid=as.character(1:nrow(xtest)),genenames=rownames(xtest))
model=pamr.train(myemailtrain,threshold = seq(0,4,0.1))

cvmodel=pamr.cv(model,myemailtrain)
print(cvmodel)

pamr.plotcv(cvmodel)

pamr.plotcen(model,myemailtrain,threshold=1.4)

a=pamr.listgenes(model,myemailtrain,threshold=1.4)
cat(paste(colnames(myemailtrain)[as.numeric(a[,1])],collapse = '\n'))

predicted <- pamr.predict(model, newx = xtest, threshold = 1.4)

contab <- table(ytest, predicted)
contab
names(dimnames(contab)) <- c("Test Actual", "Predicted by Nearest Shrunk Centroid on test")
contabres<-caret::confusionMatrix(contab)
mse1<-(1-(sum(diag(contab))/sum(contab)))

```

```

paste("The misclassificaiton rate is",mse1)

var<- as.data.frame(pamr.listgenes(model, myemailtrain, threshold = 1.4))
knitr::kable(colnames(data[,head(var$id,10)]), caption = "Top 10 Important features by NSC ")

xtrain2<-as.matrix(train[,-4703])
ytrain2<-as.matrix(train[,4703])
xtest2<-as.matrix(test[,-4703])
ytest2<-as.matrix(test[,4703])

cvmodel2<-cv.glmnet(x=xtrain2,y=ytrain2,alpha = 0.5,family="binomial")
model2<-glmnet(x=xtrain2,y=ytrain2,alpha = 0.5,family="binomial")
elasticpredict<-predict.cv.glmnet(cvmodel2, newx = xtest2, s = "lambda.min", type = "class")
elasticpredict2<-predict(model2, xtest2, type = "response")
contab22 <- table(ytest2, elasticpredict)
plot(cvmodel2)
plot(model2)
contab2 <- table(ytest2, elasticpredict)
contab2
contab2res<-caret::confusionMatrix(contab2)
mse2<-(1-(sum(diag(contab2))/sum(contab2)))
paste("The misclassificaiton rate is",mse2)
names(dimnames(contab2)) <- c("Actual Test", "Predicted by ElasticNet model")

elasticcoefs<- coef(cvmodel2, s = "lambda.min")
elasticvars <- list(name = elasticcoefs@Dimnames[[1]][elasticcoefs@i + 1])
knitr::kable(elasticvars, caption = "Contributing features of elastic net model")
set.seed(12345)
svmmmodel<- ksvm(xtrain2, ytrain2, kernel="vanilladot",scaled=FALSE)
svmpredict<- predict(svmmmodel, xtest2, type="response")

consvm<- table(ytest2, svmpredict)
names(dimnames(consvm)) <- c("Actual Test", "Predicted svm")
consvmres<-caret::confusionMatrix(consvm)
consvm
mse3<-(1-(sum(diag(consvm))/sum(consvm)))
paste("The misclassificaiton rate is",mse3)
comptab<- as.data.frame(cbind(contabres$overall[[1]]*100,
                             contab2res$overall[[1]]*100,
                             consvmres$overall[[1]] *100))
countf <- cbind(nrow(var), length(elasticcoefs@i), length(svmmmodel@coef[[1]]))
mse <- c(mse1,mse2,mse3)
comptab <- rbind(comptab, countf)
comptab <- rbind(comptab, mse)
colnames(comptab) <- c("Nearest Shrunkn Centroid Model",
                      "ElasticNet Model", "SVM Model")
rownames(comptab) <- c("Accuracy", "Number of Features","Misclassification error rate")
knitr::kable(comptab, caption = "Comparsion of the models")

set.seed(12345)
p<-c()
x<-email[,-4703]
for (i in 1:(length(email)-1)){

```

```

    x<-email[,i]
    res<-t.test(x~Conference,data=email,alternative="two.sided")
    p[i]<-res$p.value
  }

  pvalues<- data.frame(pvalue=p,variable=1:(length(email)-1))
  pvalues<- pvalues[order(pvalues$pvalue),]

  alpha<-0.05
  l<-c()
  o<-1
  for(j in 1:length(p)){
    if( pvalues$pvalue[j]< alpha*(j/nrow(pvalues)) ){
      l[o]<-j
      o<-o+1
    }
  }
  pl = pvalues$pvalue[max(l)]
  pl
  for(j in 1:nrow(pvalues)){
    if(pvalues$pvalue[j]<= pl){
      pvalues$status[j]<-FALSE
    }
    else{
      pvalues$status[j]<-TRUE
    }
  }
}

significantp<-filter(pvalues,pvalue<=0.05)
significantp<-cbind(significantp,Variable_name=colnames(email[significantp$variable]))
significantp

finalbh<-filter(pvalues,status==FALSE)
finalbh<-cbind(finalbh,Variable_name=colnames(email[finalbh$variable]))
finalbh

```