

IML Term project

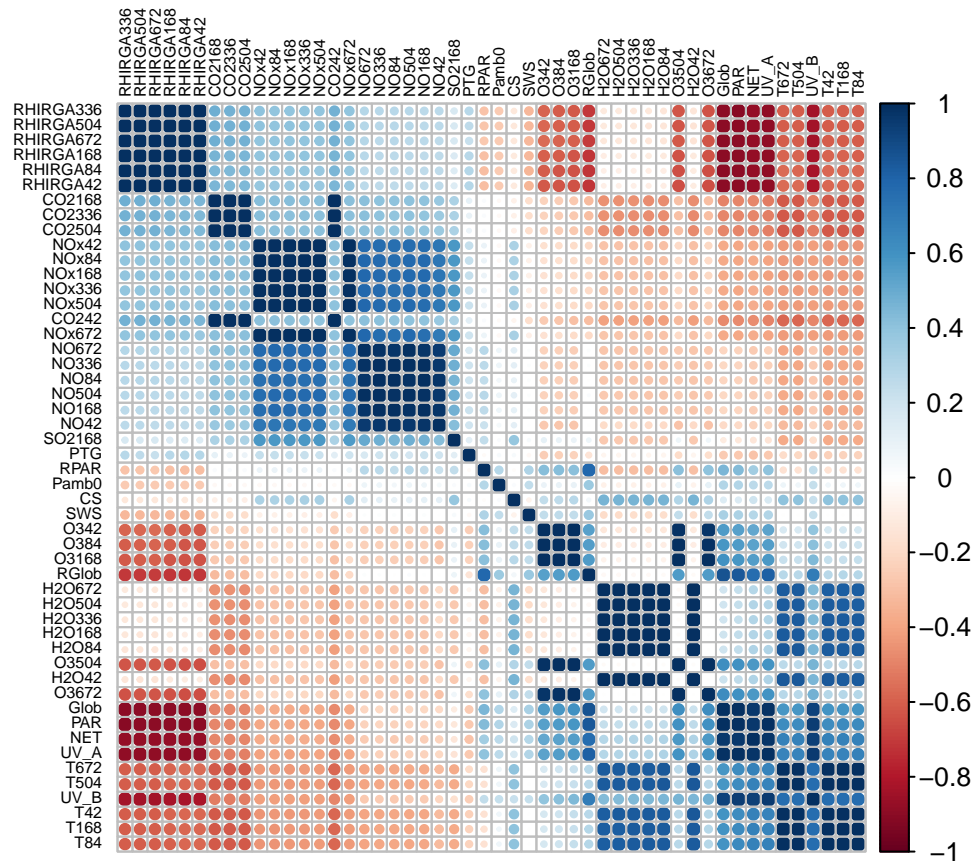
Julia Palorinne, Pyry Silomaa, Sara Sippola

11.12.2022

Preprocessing data

Correlation

Looking at the correlation matrix as it was presented in the solutions to Exercise set 1, it is clear that the measurements of the same thing at different heights correlate, as do some of the radiation-related parameters.



Correlation between measurements at different heights

Looking at the numeric values, we see that the measurements at different heights have strong positive correlation.

Table 1: Correlation (CO2)

| | CO2168 | CO2336 | CO242 | CO2504 |
|--------|-----------|-----------|-----------|-----------|
| CO2168 | 1.0000000 | 0.9996891 | 0.9939228 | 0.9988868 |
| CO2336 | 0.9996891 | 1.0000000 | 0.9928568 | 0.9996424 |
| CO242 | 0.9939228 | 0.9928568 | 1.0000000 | 0.9910311 |
| CO2504 | 0.9988868 | 0.9996424 | 0.9910311 | 1.0000000 |

Table 2: Correlation (H2O)

| | H2O168 | H2O336 | H2O42 | H2O504 | H2O672 | H2O84 |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| H2O168 | 1.0000000 | 0.9998966 | 0.9997062 | 0.9997158 | 0.9994631 | 0.9998894 |
| H2O336 | 0.9998966 | 1.0000000 | 0.9993506 | 0.9999302 | 0.9997498 | 0.9996330 |
| H2O42 | 0.9997062 | 0.9993506 | 1.0000000 | 0.9990202 | 0.9986416 | 0.9999330 |
| H2O504 | 0.9997158 | 0.9999302 | 0.9990202 | 1.0000000 | 0.9998589 | 0.9993631 |
| H2O672 | 0.9994631 | 0.9997498 | 0.9986416 | 0.9998589 | 1.0000000 | 0.9990316 |
| H2O84 | 0.9998894 | 0.9996330 | 0.9999330 | 0.9993631 | 0.9990316 | 1.0000000 |

Table 3: Correlation (NO)

| | NO168 | NO336 | NO42 | NO504 | NO672 | NO84 |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|
| NO168 | 1.0000000 | 0.9942106 | 0.9729234 | 0.9888394 | 0.9787201 | 0.9955792 |
| NO336 | 0.9942106 | 1.0000000 | 0.9681760 | 0.9962470 | 0.9889000 | 0.9912906 |
| NO42 | 0.9729234 | 0.9681760 | 1.0000000 | 0.9648354 | 0.9580581 | 0.9766436 |
| NO504 | 0.9888394 | 0.9962470 | 0.9648354 | 1.0000000 | 0.9947569 | 0.9859003 |
| NO672 | 0.9787201 | 0.9889000 | 0.9580581 | 0.9947569 | 1.0000000 | 0.9766069 |
| NO84 | 0.9955792 | 0.9912906 | 0.9766436 | 0.9859003 | 0.9766069 | 1.0000000 |

Table 4: Correlation (NOx)

| | NOx168 | NOx336 | NOx42 | NOx504 | NOx672 | NOx84 |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| NOx168 | 1.0000000 | 0.9988830 | 0.9966528 | 0.9966839 | 0.9946312 | 0.9996773 |
| NOx336 | 0.9988830 | 1.0000000 | 0.9953287 | 0.9986298 | 0.9973042 | 0.9984147 |
| NOx42 | 0.9966528 | 0.9953287 | 1.0000000 | 0.9931515 | 0.9913512 | 0.9968910 |
| NOx504 | 0.9966839 | 0.9986298 | 0.9931515 | 1.0000000 | 0.9987762 | 0.9961637 |
| NOx672 | 0.9946312 | 0.9973042 | 0.9913512 | 0.9987762 | 1.0000000 | 0.9940603 |
| NOx84 | 0.9996773 | 0.9984147 | 0.9968910 | 0.9961637 | 0.9940603 | 1.0000000 |

Table 5: Correlation (O3)

| | O3168 | O342 | O3504 | O3672 | O384 |
|-------|-----------|-----------|-----------|-----------|-----------|
| O3168 | 1.0000000 | 0.9955292 | 0.9954429 | 0.9918052 | 0.9986470 |
| O342 | 0.9955292 | 1.0000000 | 0.9843062 | 0.9789921 | 0.9987926 |
| O3504 | 0.9954429 | 0.9843062 | 1.0000000 | 0.9989553 | 0.9904563 |
| O3672 | 0.9918052 | 0.9789921 | 0.9989553 | 1.0000000 | 0.9857437 |
| O384 | 0.9986470 | 0.9987926 | 0.9904563 | 0.9857437 | 1.0000000 |

Table 6: Correlation (RHIRGA)

| | RHIRGA168 | RHIRGA336 | RHIRGA42 | RHIRGA504 | RHIRGA672 | RHIRGA84 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| RHIRGA168 | 1.0000000 | 0.9992464 | 0.9972212 | 0.9981986 | 0.9957488 | 0.9989197 |
| RHIRGA336 | 0.9992464 | 1.0000000 | 0.9950372 | 0.9994052 | 0.9973122 | 0.9970893 |
| RHIRGA42 | 0.9972212 | 0.9950372 | 1.0000000 | 0.9931951 | 0.9902415 | 0.9991006 |
| RHIRGA504 | 0.9981986 | 0.9994052 | 0.9931951 | 1.0000000 | 0.9987544 | 0.9955233 |
| RHIRGA672 | 0.9957488 | 0.9973122 | 0.9902415 | 0.9987544 | 1.0000000 | 0.9927423 |
| RHIRGA84 | 0.9989197 | 0.9970893 | 0.9991006 | 0.9955233 | 0.9927423 | 1.0000000 |

Table 7: Correlation (T)

| | T168 | T42 | T504 | T672 | T84 |
|------|-----------|-----------|-----------|-----------|-----------|
| T168 | 1.0000000 | 0.9997396 | 0.9997239 | 0.9993168 | 0.9999172 |
| T42 | 0.9997396 | 1.0000000 | 0.9993129 | 0.9987827 | 0.9998980 |
| T504 | 0.9997239 | 0.9993129 | 1.0000000 | 0.9998644 | 0.9995274 |
| T672 | 0.9993168 | 0.9987827 | 0.9998644 | 1.0000000 | 0.9990504 |
| T84 | 0.9999172 | 0.9998980 | 0.9995274 | 0.9990504 | 1.0000000 |

Measurement heights

The measurement height 16.8m is the only one with all measurements. Because of this and the correlation between the measurements at different heights, we choose to discard measurements from heights other than 16.8m.

Table 8: Measurements at different heights: What measurements have been done at particular heights?

| dm.42 | dm.84 | dm.168 | dm.336 | dm.504 | dm.672 |
|----------|----------|-----------|-----------|-----------|-----------|
| CO242 | NA | CO2168 | CO2336 | CO2504 | NA |
| H2O42 | H2O84 | H2O168 | H2O336 | H2O504 | NA |
| NO42 | NO84 | NO168 | NO336 | NO504 | NO672 |
| NOx42 | NOx84 | NOx168 | NOx336 | NOx504 | NOx672 |
| O342 | O384 | O3168 | NA | O3504 | O3672 |
| RHIRGA42 | RHIRGA84 | RHIRGA168 | RHIRGA336 | RHIRGA504 | RHIRGA672 |
| NA | NA | SO2168 | NA | NA | NA |
| T42 | T84 | T168 | NA | T504 | T672 |

vif

```
## Loading required package: carData
```

```
## [1] 3003.667
```

```
## PAR.std  
##      18
```

```
## [1] 1612.928
```

```
## UV_A.std  
##      30
```

```
## [1] 237.7665
```

```
## Glob.std  
##       4
```

```
## [1] 168.3538
```

```
## UV_B.std  
##      28
```

```
## [1] 94.19929
```

```
## RPAR.mean  
##      21
```

```
## [1] 45.30109
```

```
## NET.std  
##       6
```

```
## [1] 12.3017
```

```
## RGlob.std  
##      16
```

```
## [1] 11.65981
```

```
## CS.mean  
##      21
```

```
## [1] 7.342582
## N0x168.mean
##          7
## [1] 6.960544
## H20168.mean
##          3
```

Runnig vif() over several variables in the 16.8m restricted dataset, we come up to the following collection of coefficients

```
##      (Intercept)      C02168.mean      C02168.std      H20168.mean      H20168.std
## -1.760990e+01 -7.706613e-02  1.707145e-01 -7.413117e-01 -2.982764e-01
##      N0168.mean      N0168.std      03168.mean      03168.std      Pamb0.mean
## -5.767996e+00  4.915288e+00  8.270660e-03  3.630687e-04  2.079041e-03
##      Pamb0.std      PTG.mean      PTG.std      RHIRGA168.mean      RHIRGA168.std
##  5.275777e-01 -3.373216e+01  1.954975e+01 -6.657638e-02  1.851450e-01
##      S02168.mean      S02168.std      SWS.mean      SWS.std
## -1.942375e+00  2.594570e+00  5.644814e-02  1.811580e-02
```

Hence we keep the following parameters

```
## [1] "C02168.mean"      "C02168.std"      "H20168.mean"      "H20168.std"
## [5] "N0168.mean"      "N0168.std"      "03168.mean"      "03168.std"
## [9] "Pamb0.mean"      "Pamb0.std"      "PTG.mean"      "PTG.std"
## [13] "RHIRGA168.mean"  "RHIRGA168.std"  "S02168.mean"      "S02168.std"
## [17] "SWS.mean"      "SWS.std"      "class2"
```

PCA

PCA for all three versions of data

```
## [1] 0.3981163 0.5569929 0.6835814 0.7260025 0.7598939 0.7874098 0.8125596
## [8] 0.8346002 0.8530532 0.8686838 0.8821772 0.8942397 0.9051217 0.9157874
## [15] 0.9248316 0.9336262 0.9417328 0.9485561 0.9553166 0.9611454 0.9668515
## [22] 0.9712230 0.9745869 0.9776303 0.9801121 0.9824840 0.9844354 0.9860310
## [29] 0.9875789 0.9890640 0.9904694 0.9917631 0.9928400 0.9936991 0.9943712
## [36] 0.9949925 0.9955783 0.9961134 0.9966184 0.9970703 0.9975005 0.9978140
## [43] 0.9980826 0.9982829 0.9984779 0.9986491 0.9987938 0.9989157 0.9990305
## [50] 0.9991350 0.9992267 0.9993091 0.9993894 0.9994606 0.9995292 0.9995878
## [57] 0.9996377 0.9996800 0.9997150 0.9997457 0.9997746 0.9997990 0.9998221
## [64] 0.9998437 0.9998622 0.9998783 0.9998934 0.9999073 0.9999197 0.9999320
## [71] 0.9999411 0.9999493 0.9999558 0.9999620 0.9999667 0.9999708 0.9999743
## [78] 0.9999775 0.9999805 0.9999832 0.9999855 0.9999876 0.9999893 0.9999910
## [85] 0.9999926 0.9999940 0.9999952 0.9999962 0.9999970 0.9999978 0.9999985
## [92] 0.9999990 0.9999993 0.9999995 0.9999997 0.9999998 0.9999999 0.9999999
## [99] 1.0000000 1.0000000

## [1] 0.4309861 0.5575522 0.6537154 0.7007490 0.7412151 0.7770282 0.8055799
## [8] 0.8313561 0.8543780 0.8745338 0.8925758 0.9081481 0.9224718 0.9346085
## [15] 0.9446046 0.9531742 0.9611388 0.9675212 0.9735814 0.9786998 0.9835213
## [22] 0.9872826 0.9901902 0.9927185 0.9946737 0.9964706 0.9978635 0.9986532
## [29] 0.9990837 0.9994036 0.9996349 0.9998353 0.9999258 0.9999871 0.9999948
## [36] 1.0000000

## [1] 0.2592587 0.4142654 0.5355777 0.6267689 0.6891652 0.7479272 0.7955438
## [8] 0.8334548 0.8637669 0.8919822 0.9138447 0.9314743 0.9476770 0.9617142
```

[15] 0.9737523 0.9840080 0.9930149 1.0000000

Classifiers

Naive Bayes, LDA and QDA on the original data, 168-data and the PCA'd versions of these data sets

Naive Bayes

| Accuracy | | | | Perplexity | | | |
|-----------|-----------|-----------|--------|------------|----------|----------|---------|
| train | test | CV | LOOCV | train | test | CV | LOOCV |
| 0.8448276 | 0.7801724 | 0.8146552 | 0.8125 | 30.86972 | 82.21584 | 152.8545 | 166.084 |

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.7823276 | 0.7823276 | 36.25527 | 40.74978 |

Using PCA on the original and the 16.8m dataset.

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.7564655 | 0.7564655 | 12.14968 | 12.57788 |

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.7823276 | 0.7823276 | 36.25527 | 40.74978 |

LDA

On the original dataset and the 16.8m dataset.

| Accuracy | | Perplexity | |
|----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.875 | 0.8814655 | 1.683712 | 1.579626 |

| Accuracy | | Perplexity | |
|----------|-----------|------------|---------|
| CV | LOOCV | CV | LOOCV |
| 0.887931 | 0.8922414 | 1.329679 | 1.32923 |

Using PCA on the original and the 16.8m dataset.

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.8706897 | 0.8706897 | 1.364454 | 1.368989 |

| Accuracy | | Perplexity | |
|----------|-------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.875 | 0.875 | 1.391006 | 1.391656 |

QDA

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.8448276 | 0.8383621 | 2979.474 | 3793.874 |

| Accuracy | | Perplexity | |
|-----------|-----------|------------|---------|
| CV | LOOCV | CV | LOOCV |
| 0.8556034 | 0.8534483 | 7.745819 | 6.34633 |

Using PCA on the original and the 16.8m dataset.

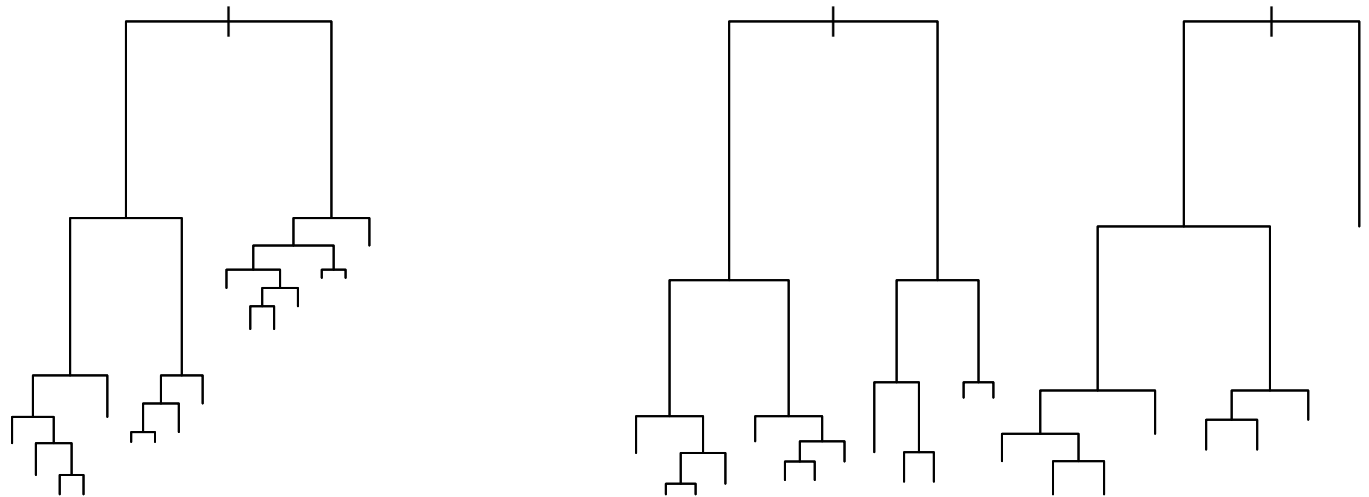
| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.8426724 | 0.8383621 | 2.06776 | 2.034851 |

| Accuracy | | Perplexity | |
|-----------|-----------|------------|----------|
| CV | LOOCV | CV | LOOCV |
| 0.8340517 | 0.8383621 | 2.024616 | 2.038691 |

Tree-based methods

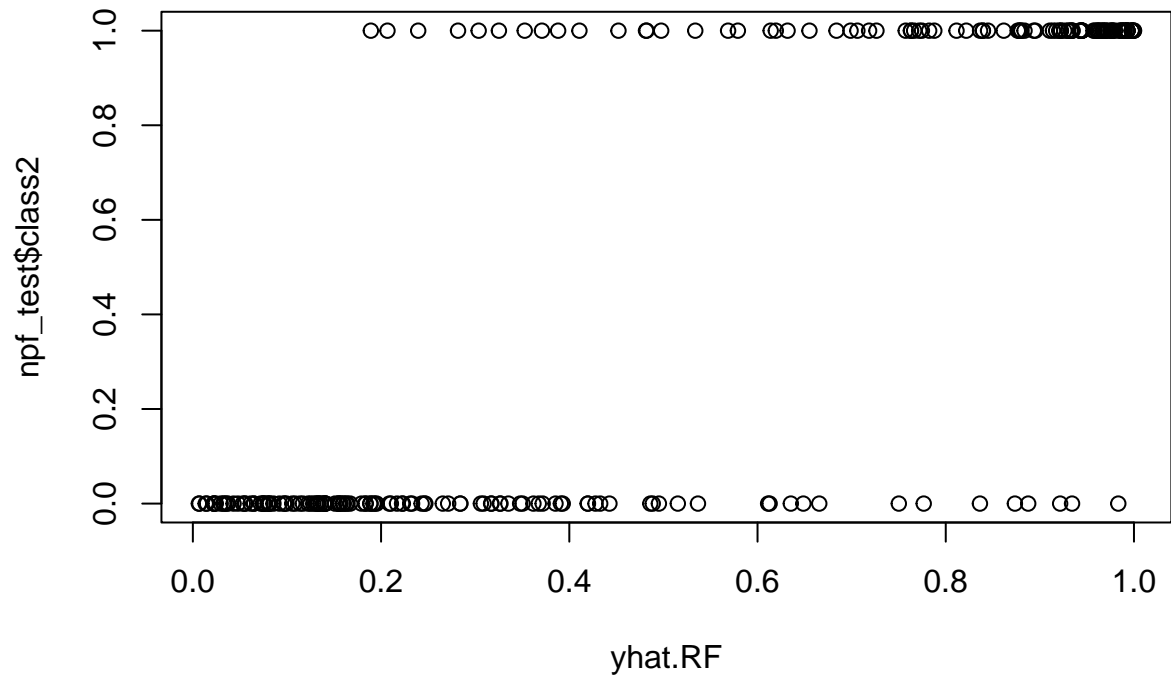
```
##
## Classification tree:
## tree(formula = class2 ~ ., data = npf.fact.train)
## Variables actually used in tree construction:
## [1] "RHIRGA84.mean" "H2O168.mean" "SWS.mean" "CS.mean"
## [5] "CS.std" "CO2168.mean" "NOx168.std" "H2O168.std"
## [9] "RHIRGA336.mean" "O3672.std" "PTG.std" "H2O504.std"
## [13] "O3168.std"
## Number of terminal nodes: 16
## Residual mean deviance: 0.1086 = 23.47 / 216
## Misclassification error rate: 0.02586 = 6 / 232

##
## Classification tree:
## tree(formula = class2 ~ ., data = npf.168.fact.train)
## Variables actually used in tree construction:
## [1] "RHIRGA168.mean" "H2O168.mean" "PTG.mean" "NO168.std"
## [5] "CO2168.mean" "Glob.std" "CS.mean" "O3168.mean"
## Number of terminal nodes: 13
## Residual mean deviance: 0.2165 = 47.41 / 219
## Misclassification error rate: 0.04741 = 11 / 232
```



| Accuracy | | | Perplexity | | |
|-----------|----|-------|------------|-----|-------|
| original | CV | LOOCV | original | CV | LOOCV |
| 0.2650862 | 0 | 0 | 27.77601 | NaN | NaN |
| 0.0862069 | 0 | 0 | 21.34815 | NaN | NaN |

Random forest tested with original and 168 dataset, and accuracies and perplexities reported. Why is mtry 12?



| Accuracy | | | Perplexity | | |
|----------|-------|-------|------------|-----|-------|
| original | CV | LOOCV | original | CV | LOOCV |
| 0.875 | 0.858 | 0.000 | 1.380 | NaN | 0.000 |
| 0.884 | 0.884 | 0.884 | 1.375 | NaN | 1.302 |

Multiple classes

Naive Bayes on the npf-dataset

```
## [1] 0.5064655
```

LDA on the npf.168-dataset

```
## [1] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [19] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [37] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [55] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [73] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [91] "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3" "3"
## [109] "3" "3" "3" "3"

## [1] 0.09051724
```

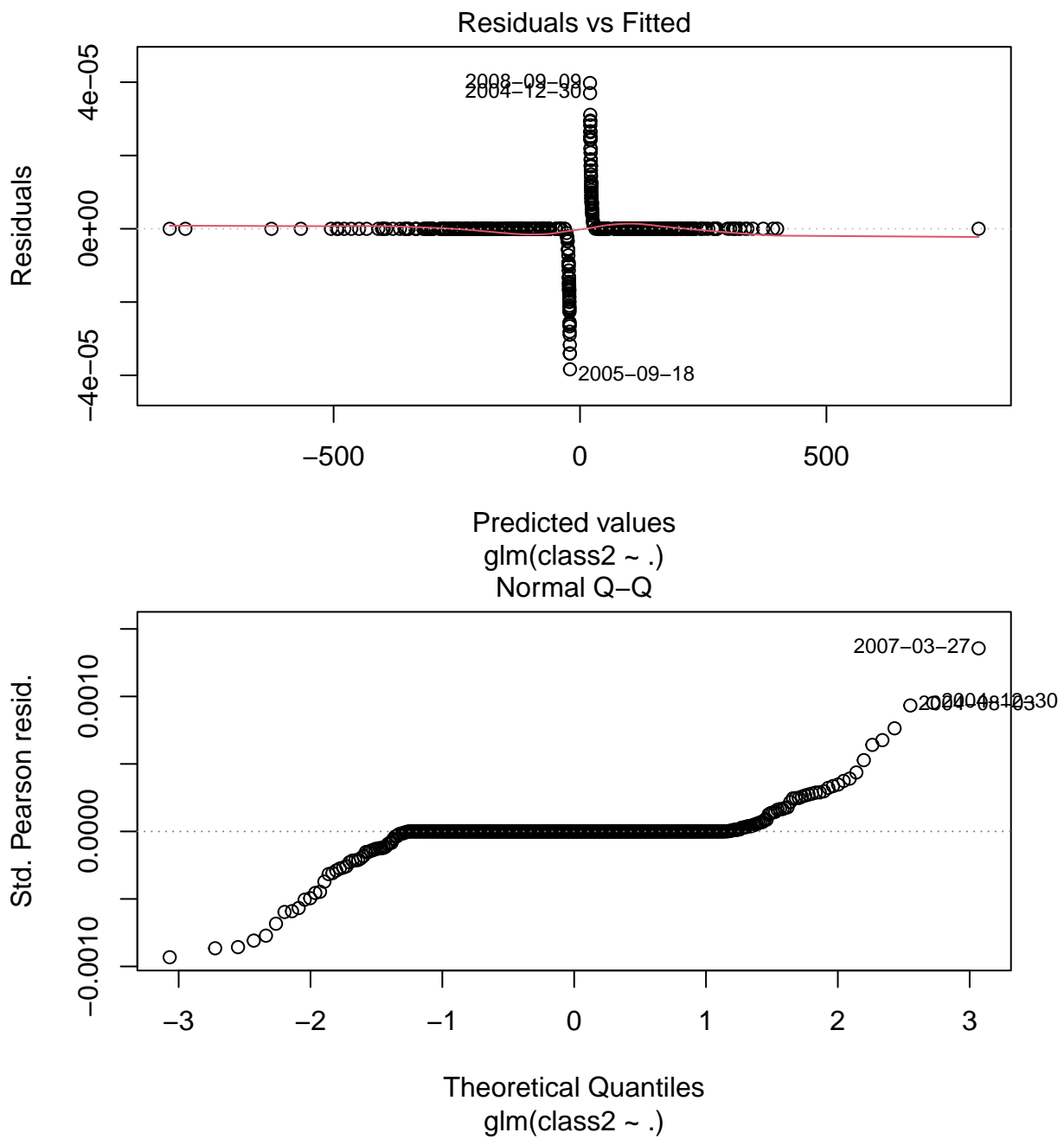
QDA on the npf.pca-dataset

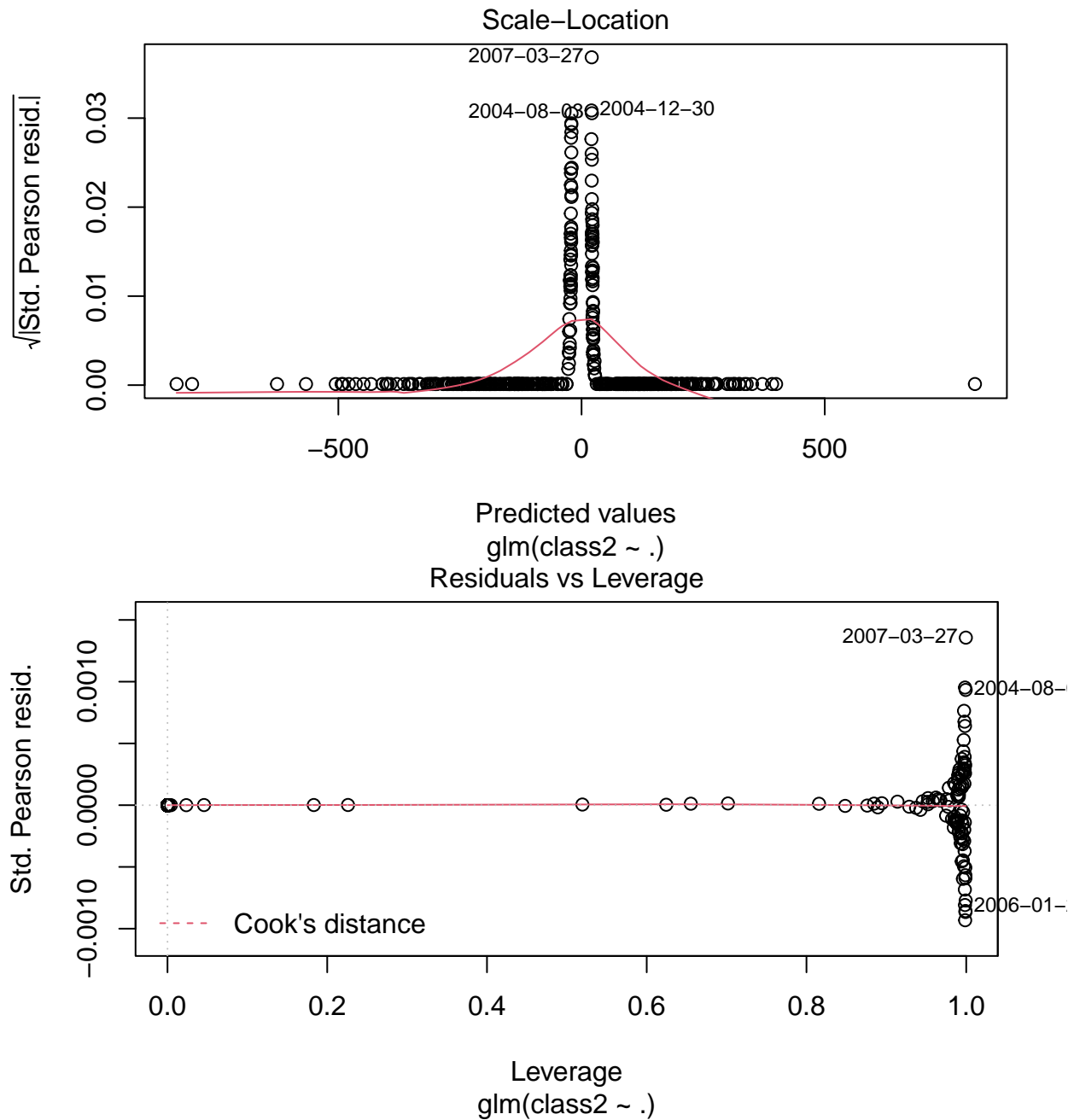
```
## [1] 0.5991379
```

QDA on the npf.168.pca-dataset

```
## [1] 0.5948276
```

GLM





GLM accuracy and perplexity.

```
## [1] 1
## [1] 0.3678794
```

Lasso / Ridge

```
## 101 x 1 sparse Matrix of class "dgCMatrix"
##          s1
## (Intercept) 1.554540e+01
## C02168.mean .
## C02168.std .
## C02336.mean .
```

```

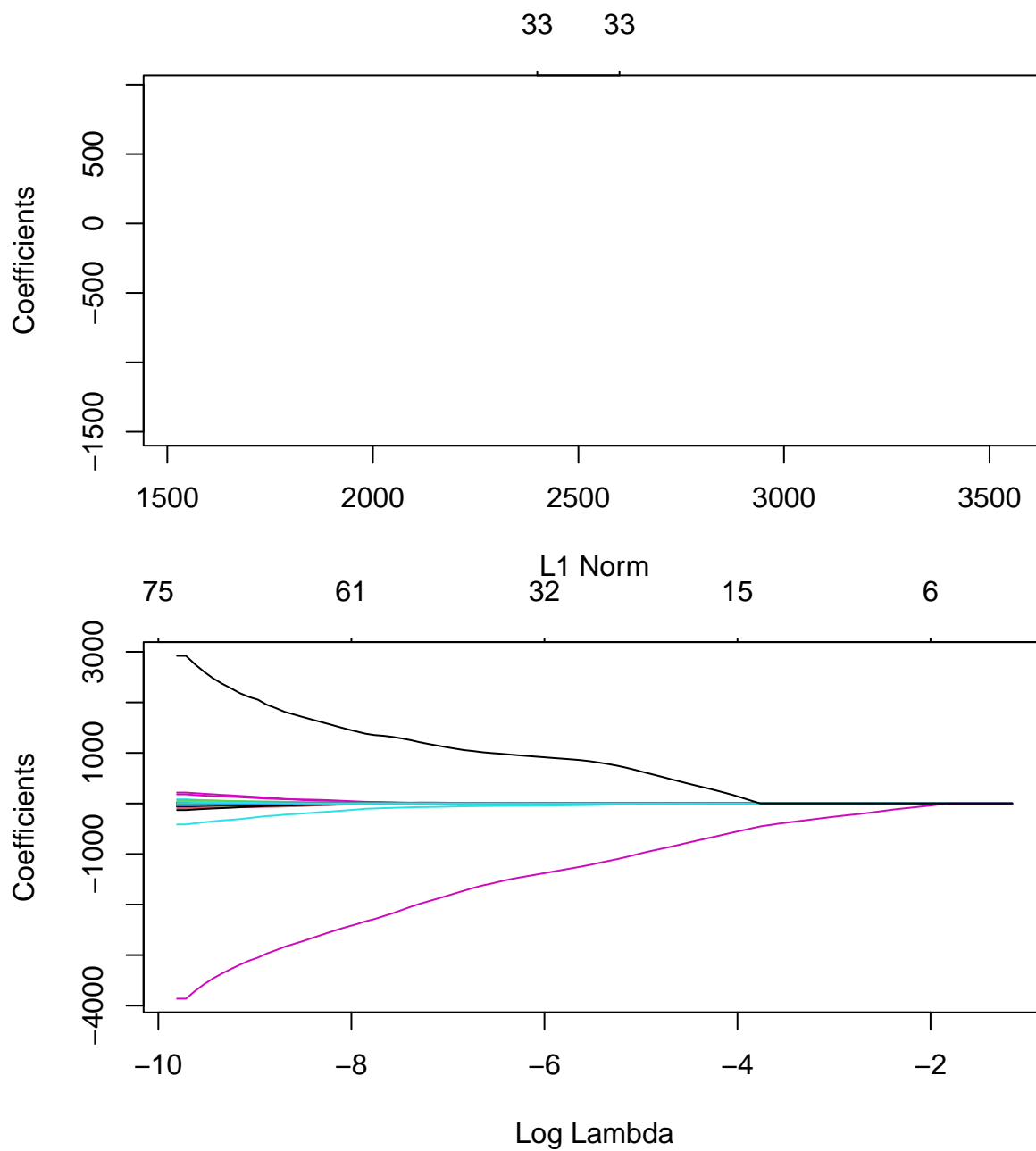
## C02336.std      .
## C0242.mean     -5.117656e-02
## C0242.std      .
## C02504.mean    .
## C02504.std     .
## Glob.mean      .
## Glob.std       .
## H20168.mean    .
## H20168.std     .
## H20336.mean    .
## H20336.std     .
## H2042.mean     -2.308488e-01
## H2042.std      .
## H20504.mean    .
## H20504.std     .
## H20672.mean    .
## H20672.std     .
## H2084.mean     .
## H2084.std      .
## NET.mean       .
## NET.std        .
## N0168.mean     .
## N0168.std      1.737770e+00
## N0336.mean     .
## N0336.std      .
## N042.mean      .
## N042.std       .
## N0504.mean     .
## N0504.std      .
## N0672.mean     .
## N0672.std      .
## N084.mean      .
## N084.std       .
## N0x168.mean    .
## N0x168.std     .
## N0x336.mean    .
## N0x336.std     .
## N0x42.mean     .
## N0x42.std      .
## N0x504.mean    .
## N0x504.std     .
## N0x672.mean    .
## N0x672.std     .
## N0x84.mean     .
## N0x84.std      .
## 03168.mean     .
## 03168.std      .
## 0342.mean      .
## 0342.std       .
## 03504.mean     .
## 03504.std      .
## 03672.mean     7.308214e-02
## 03672.std      .
## 0384.mean      .

```

```

## 0384.std      .
## Pamb0.mean    .
## Pamb0.std     1.705030e-01
## PAR.mean      .
## PAR.std       .
## PTG.mean      -5.040826e+00
## PTG.std       .
## RGlob.mean    1.107212e-03
## RGlob.std     .
## RHIRGA168.mean .
## RHIRGA168.std  3.667383e-02
## RHIRGA336.mean -6.007049e-02
## RHIRGA336.std  .
## RHIRGA42.mean .
## RHIRGA42.std  .
## RHIRGA504.mean .
## RHIRGA504.std .
## RHIRGA672.mean .
## RHIRGA672.std  4.309511e-02
## RHIRGA84.mean .
## RHIRGA84.std  .
## RPAR.mean     .
## RPAR.std      .
## S02168.mean   .
## S02168.std    3.550962e-01
## SWS.mean      8.117156e-03
## SWS.std       .
## T168.mean     .
## T168.std      4.122854e-01
## T42.mean      .
## T42.std       .
## T504.mean     .
## T504.std      .
## T672.mean     .
## T672.std      .
## T84.mean      .
## T84.std       .
## UV_A.mean     .
## UV_A.std      .
## UV_B.mean     .
## UV_B.std      .
## CS.mean       -7.302570e+02
## CS.std        3.463189e+02

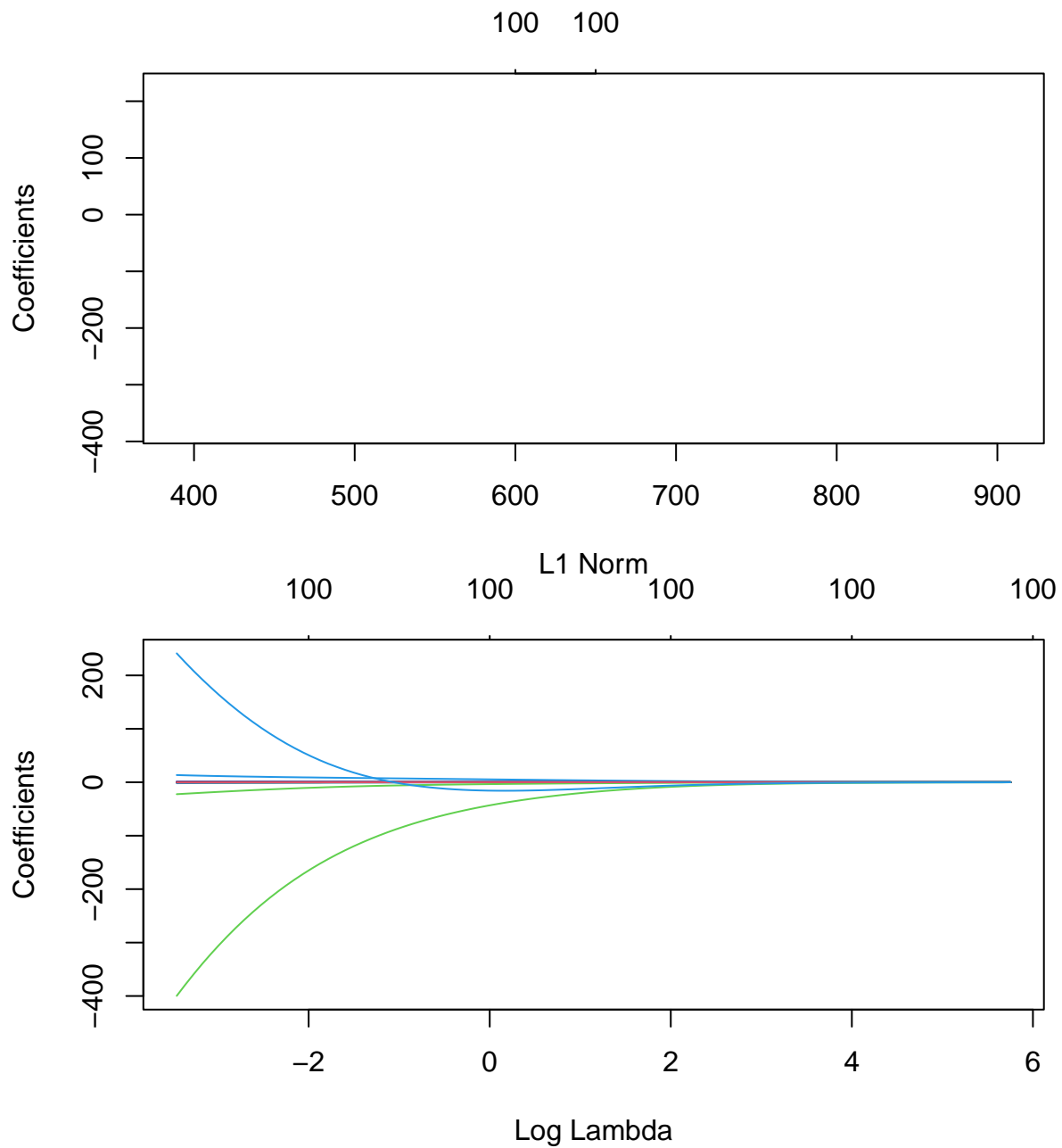
```



Lasso accuracy and perplexity.

[1] 0.9288793

[1] 0.4121674



Ridge accuracy and perplexity.

[1] 0.9094828

[1] 0.4313464

SVM

SVM accuracy and perplexity.

[1] 0.8922414

[1] 1.308934

Lasso (first version of final model)

[1] 0.4892241

[1] 1.308721

GLM (first version of actual model without usage of PCAs):

[1] 0.8900862

[1] 1.303646

SVM (first version of actual model)