Project 1 - Shrinkage

2NN: Bjerke, Martin & Spremic, Mina

2/15/2021

Introduction

The goal of this project is to apply the shrinkage methods to a dataset of the groups choice. As one of the main methods, of the first part of the course, has been lasso, that is often applied in cases when the number of covariates is much larger than number of observations, in addition to there being very few observations. Hence, we have chosen a dataset satisfying these criteria.

Dataset

The dataset in question is a Gastrointestinal Lesions in Regular Colonoscopy dataset, which can be accessed at https://archive.ics.uci.edu/ml/datasets/Gastrointestinal+Lesions+in+Regular+Colonoscopy.

This dataset contains a response variable, which is one of the three types of lesions (growth): hyperplasic, adenoma and serates, with the firts one being benign and the latter two malignant. The rest of the covariates, 699 of them, have been extracted from video, more specificly from a colonoscopy (a camera being sent through the bowels), and represent different aspects of it, through rotation, texture, colour, contrast etc. They are divided roughly in three groups, which are textural features, color features and shape features, all having multiple subgroups.

This dataset had to be labeled all over again, as the .txt file including the data did not have a header. Relabeling code is available in the git-repo. In this dataset, majority of the 699 covariates have not been easy to decipher, both due to their large number and the lack of information on them, both on the webpage of the dataset, but also in the paper that the dataset is connected to. Most of the information about the variables have been references to other papers which do analysis and describe the process of extracting some of these variables from video. We will include references to some of these articles in the end of the report. The dataset is connected to a published paper Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy.

The names of the covariates, which have been used as labels are the following:

Group	Subgroup	Number of covariates
Type of light used for recording*		1
2D textural features		422
	AHT - Autocorrelation Homogeneous Texture (Invariant Gabor Texture)	166
	Rotational Invariant LBP	256
2D color features		76
	Color naming	16
	Discriminative Color	13
	Hue	7
	Opponent	7
	Color gray-level	

Group	Subgroup	Number of covariates
3D shape features	co-occurence matrix	33
		200
	shapeDNA	100
	KPCA	100

*Types of the light are 1=WL (White Light), 2=NBI(Narrow Band Imaging)

We wish to specify that this is a classification problem. Provided in the dataset are three clases of the response. We have chosen not to use all of the three classes in our problem, and have instead merged two of the classes. Instead of there being classes: hyperplasic, adenoma and serates, we have malignant and benign instead.
An observations made is that some of the covariates had only zero entries, but due to lack of documentation on the cause of this, we have chosen to keep all the covariates in the dataset. Finally, it is worth mentioning that there actually are only 76 unique cases in the dataset, that have been measures twice, using different light.

Analysis

The problem at hand is a classification problem. As such, we have chosen to apply logistic regression, lasso and group lasso. We attampt using logistic regression, hoping it can serve as sort of reference. The reason behind choosing lasso is due to the large number of covariates in the dataset, which we hope could be shrunk. Group lasso is chosen due to a natural grouping of the variables due to the structuring of the dataset.

We show a contigency table, showing the proportion of the number of positive and negative responses from the dataset and present some simple summary of variables representing each of the main groups. Due to the large number of covariates, and inability to deem some of them more important than others, due to lack of domain knowledge, we conclude the exploratory analysis with that. If the reader is interested in all of the covariates, we suggest they inspect the covaraites on their own initiative.

```
table(test_ds$type_of_lesion, dnn=c("Benign Malignant"))
```

```
## Benign Malignant
##   0   1
##  42 110
```

```
summary(test_ds[,c(1,2,3,169,425,441,454,461,468,501,601)])
```

```
##  type_of_lesion   type_of_light_1_WL_2_NBL Textural_feature_AHT1
##  Min.   :0.0000   Min.   :1.0              Min.   : 65.18
##  1st Qu.:0.0000   1st Qu.:1.0              1st Qu.:107.46
##  Median :1.0000   Median :1.5              Median :128.84
##  Mean   :0.7237   Mean   :1.5              Mean   :128.24
##  3rd Qu.:1.0000   3rd Qu.:2.0              3rd Qu.:147.23
##  Max.   :1.0000   Max.   :2.0              Max.   :204.53
##  Rot_invariant_LBP1 Color_naming1   Discriminative_color1      Hue1
##  Min.   :  82.0     Min.   : 44.0   Min.   :0             Min.   :0
##  1st Qu.: 400.5     1st Qu.: 77.5   1st Qu.:0             1st Qu.:0
##  Median : 820.0     Median :111.5   Median :0             Median :0
##  Mean   : 993.8     Mean   :118.5   Mean   :0             Mean   :0
##  3rd Qu.:1479.2     3rd Qu.:149.2   3rd Qu.:0             3rd Qu.:0
##  Max.   :3906.0     Max.   :279.0   Max.   :0             Max.   :0
##    Opponent1 Col_gray_lvl_co-occurr_mx1   shapeDNA1     KPCA1
##  Min.   :0   Min.   : 72.73              Min.   :0   Min.   :0.2977
```

```
##  1st Qu.:0   1st Qu.:142.13            1st Qu.:0   1st Qu.:0.4865
##  Median :0   Median :168.01            Median :0   Median :0.6080
##  Mean   :0   Mean   :167.47            Mean   :0   Mean   :0.5898
##  3rd Qu.:0   3rd Qu.:198.54            3rd Qu.:0   3rd Qu.:0.6892
##  Max.   :0   Max.   :253.03            Max.   :0   Max.   :0.9049
```

Logistic regression

Unfortunately, the logistic regression did not manage to run, and we have recieved an error: *does not converge*, indicating that we most probably have multicollinearity issues and the issue with number of covariates being too large comapared to the number of observations. We have attempted to surpass this by increasing the number of iterations, but even though it converges, it does not provide meaningful results, with very many coefficients being set to NA.

Lasso

We proceed applying the lasso, as implemented in the glmnet package.

Cross-validation is performed using the cv.lasso function from glmnet package in order to find the optimal lambda. We plot the lasso-plot, visualising at which λs the coefficients get shrunk to zero, and how many of them have not been shrunken.

Additionally we show the value minimum λ as well as one standard deviation λ.

```
## [1] "The lamda giving the smallest CV error:  0.0159523691820344"
```

```
## [1] "The one standard deviation lambda:  0.0510358904388196"
```

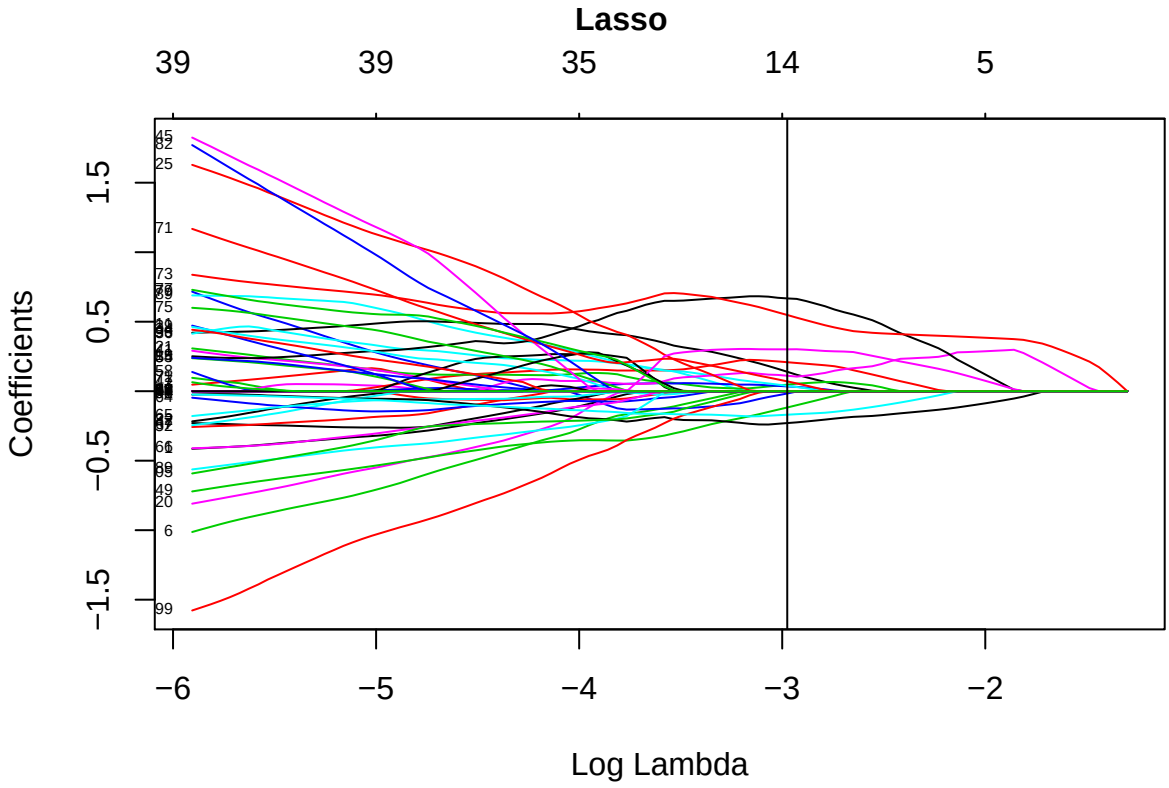

Figure 1: The plot shows the number of coefficients that have not been shrunken, for the optimal lambda - one standard deviation.

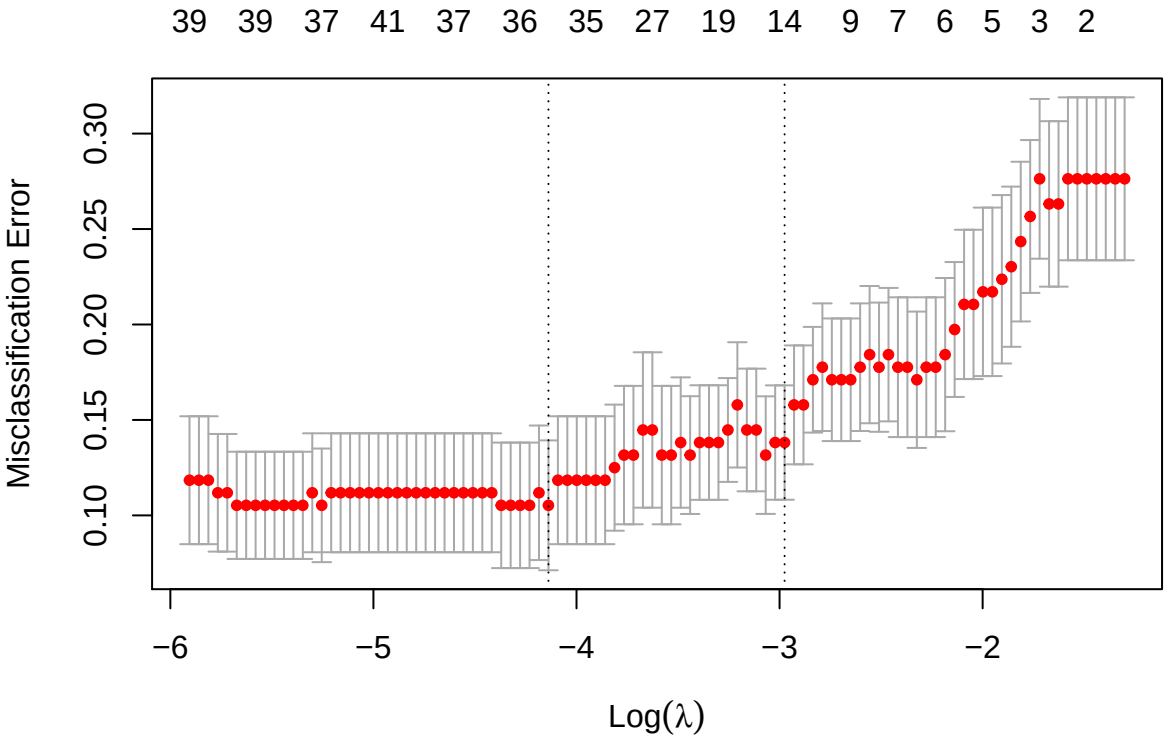

Figure 2: The plot shows the log-lambda and the corresponding misclassification error. Additionally on the top we see how many covariates minimum lambda and one standard deviation lambda give.

We can see the covariates picked with the lasso method. These are: 6 different textural feature AHT number 9, 98, 110, 120, 132, 135, rotationally invariante LBP number 4 and 52. It shows discriminative_color 10 is non-zero. In addition three of the covariates realted to the level of grayscale are included as well, namely 7, 25 and 32. Finally shapeDNA number 72 and KPCA number 2 is included. It is rather difficult to say whether these results were as expected, considering the lack of knowledge about the covariates. We could perhaps speculate that, for example, grey level scale indicates some sort of contrast in the video/pictures, giving certain indication of edges and features more pronounced for malignant lesions.

```
##       [,1]                          [,2]
##  [1,] "Intercept"                   "1.48187289885271"
##  [2,] "Textural_feature_AHT9"       "-0.0165054433060385"
##  [3,] "Textural_feature_AHT98"      "0.139713077485207"
##  [4,] "Textural_feature_AHT110"     "0.208600738760856"
##  [5,] "Textural_feature_AHT120"     "0.0469315398978425"
##  [6,] "Textural_feature_AHT132"     "0.0345788814397196"
##  [7,] "Textural_feature_AHT135"     "0.0388000771717282"
##  [8,] "Rot_invariant_LBP4"          "0.111325378543639"
##  [9,] "Rot_invariant_LBP52"         "0.667335547128893"
## [10,] "Discriminative_color10"      "-0.117155757241041"
## [11,] "Col_gray_lvl_co-occurr_mx7"  "0.548221658456"
## [12,] "Col_gray_lvl_co-occurr_mx25" "0.301191346888653"
## [13,] "Col_gray_lvl_co-occurr_mx32" "-0.229364332180344"
## [14,] "shapeDNA72"                  "0.0705669560002838"
## [15,] "KPCA2"                       "-0.169205707520923"
```

Group Lasso

We proceed to apply the group lasso to our dataset. The groups chosen are structured such that the covariates belonging to the group of variables presented in the begining, are in the same group in this analysis as well. This has been the only sensible decision of a group split, taken into considerationg the number of covariates and the difficulties concerning decyphering their concrete meaning.

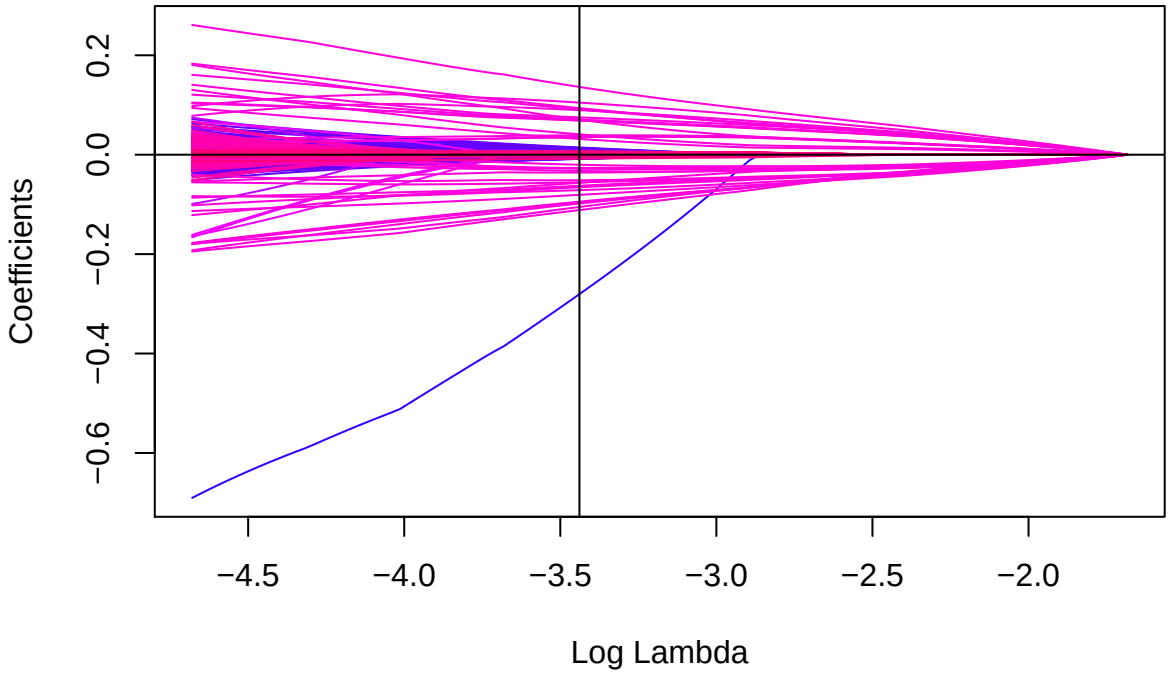

Figure 3: The plot shows different groups of coefficients, and lambdas at which they are being shrunken. The vertical line indicates the one standard deviation lambda obtained through cross validation.

```
##       [,1]
## [1,] "type_of_light_1_WL_2_NBL"
## [2,] "Textural_feature_AHT1"
## [3,] "`Col_gray_lvl_co-occurr_mx1`"
## [4,] "KPCA1"
```

The covariate groups chosen by the group lasso are as given above. If we compare it to the ones chosen by the lasso, we observe that the groups of covariates chosen by the group lasso are represented among the covariates which have not bene shrunken by the lasso . The difference between the two is that the group lasso picks the covariate indicating type of light but not the rotational invariance, discriminateive color and shapeDNA.

Prediction

We wish to point out that due to a low number of observations, we do not believe splitting the dataset into training and test set is feasible. We believe we would not obtain a realistic representation of the fraction of malignant lesions in population with so few observations in both train and test set. Hence we will not be including any prediction, but will do some inference in the next section

Inference

For the inference section we choose to do bootstrapping on the lasso, in order to infere whether the covariates that have not been shrunken, have an indication of actually being significant.

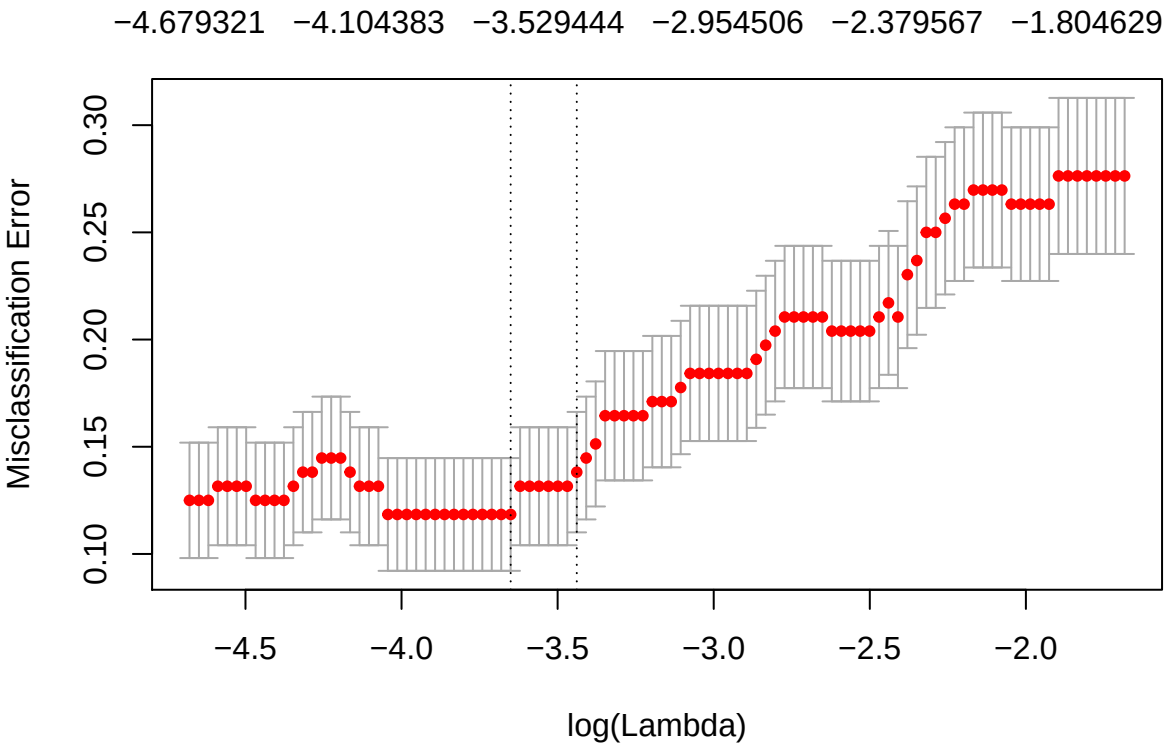

Figure 4: The plot shows the log-lambda and the corresponding misclassification error. Additionally on the top we see how many groups of covariates minimum lambda and one standard deviation lambda give.

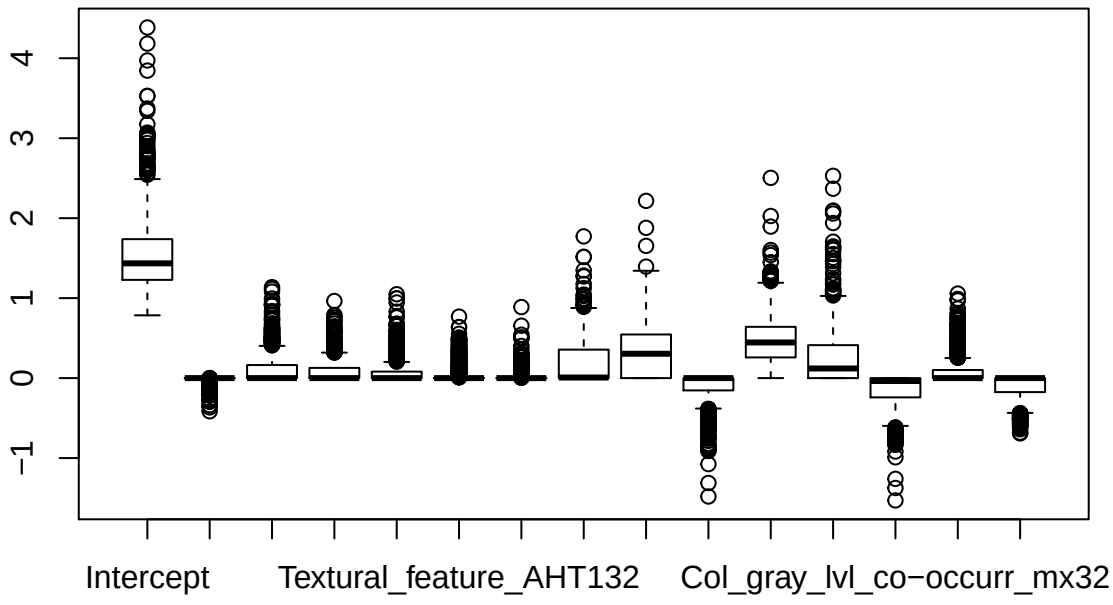

Figure 5: Boxplot for coefficients for 14 covaraites including the intercept.

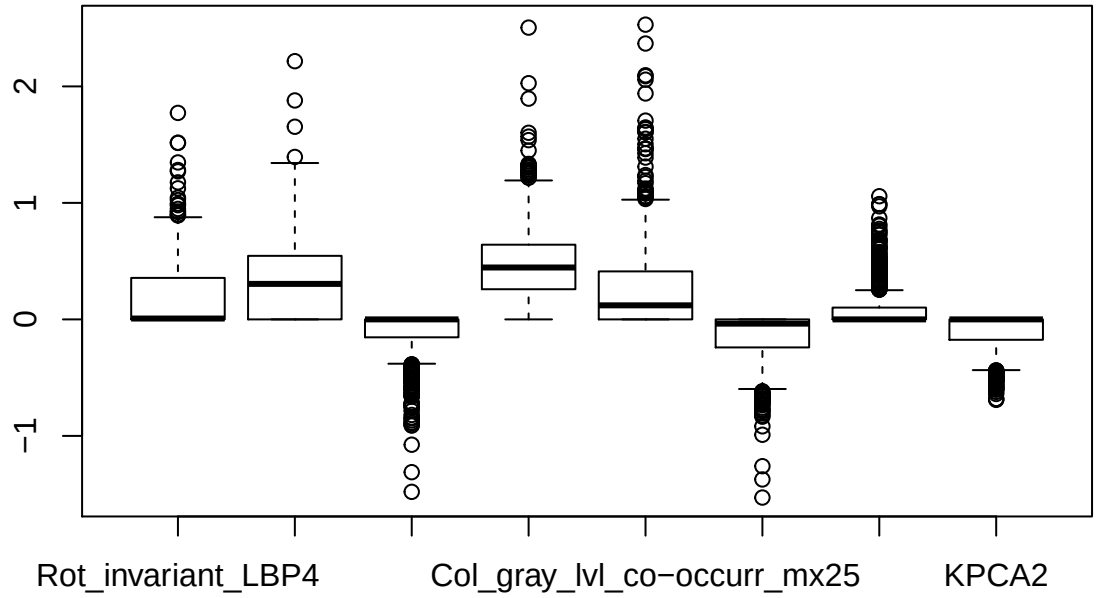

Figure 6: Boxplot showing the latter 8 coefficients, making it easier to see the coefficient which do not have median at zero. Due to the long covariate names, we provide these here, in order of appearance in the boxplot: Rot_invariant_LBP4, Rot_invariant_LBP52, Discriminative_color10, Col_gray_lvl_co-occurr_mx7, Col_gray_lvl_co-occurr_mx25, Col_gray_lvl_co-occurr_mx32, shapeDNA72, KPCA2.

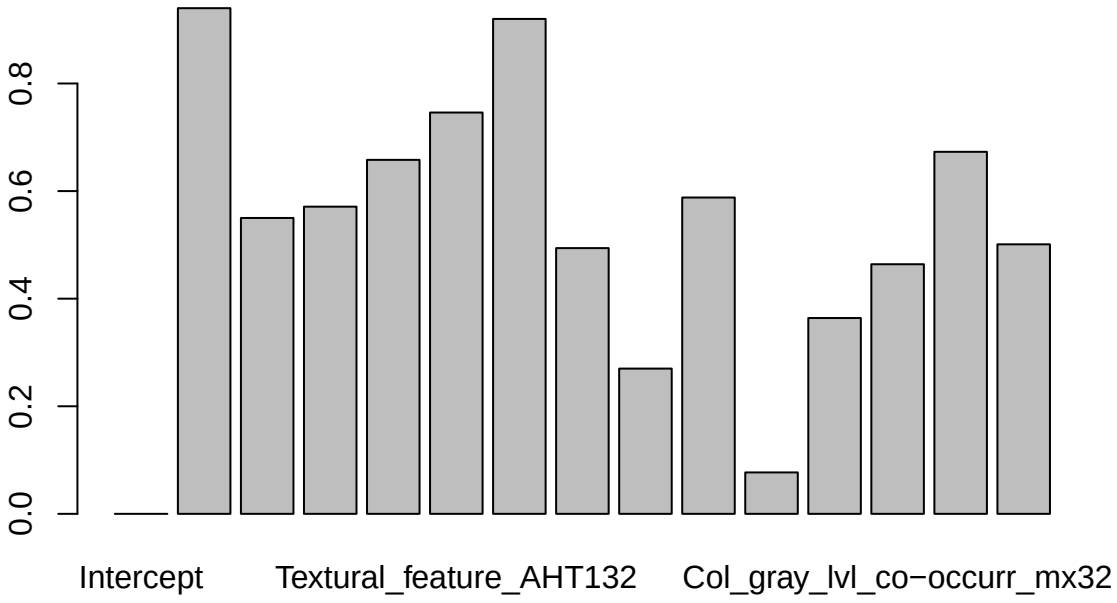

Figure 7: Barplot showing the proportion of time each of the covariates is zero.

From the boxplots we observe that majority of the boxes have median at zero, or are very close to zero. This is further confirmed throught the barplot. In the barplot we can see that almost all of the coefficients are zero at 40% of the time. The only coefficient that is noticeably different, and is almost never zero, is one of the coefficients for the gray-level contrast, followed by one of the coefficients for rotational invariance and another gray-level contrast coefficient.

Discussion and concluding words

After the analysis has been conducted, and inference has been performed, the results were slightly suprising. Eventhough we did not know exactly what to expect due to the large number of covariates, most of which have been extracted from videos, as mentioned previously, it was rather suprising only one of the coefficients was almost never zero, while only 3 were non-zero less than 40% of the time. We could speculate whether the results would have been different have we had opted out for a classification problem with three classes.

References

[1]: "Gastrointestinal Lesions in Regular Colonoscopy Data Set" - https://archive.ics.uci.edu/ml/datasets/Gastrointestinal+Lesions+in+Regular+Colonoscopy
[2]: "Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy" https://pubmed.ncbi.nlm.nih.gov/28005009/
[3]: R. Nava, G. Cristo bbal, and B. Escalante-Ramiırez, "Invariant texture analysis through local binary patterns," CoRR, vol. abs/1111.7271, 2011.
[4]: F. Riaz, F. B. Silva, M. Dinis-Ribeiro, and M. T. Coimbra, "Invariant gabor texture descriptors for classification of gastroenterology images," IEEE Trans. Biomed. Engineering, vol. 59, no. 10, pp. 2893–2904, 2012.
[5]: M. Reuter, F.-E. Wolter, and N. Peinecke, "Laplace–beltrami spectra as shape-dnaof surfaces and solids," Computer-Aided Design, vol. 38, no. 4, pp. 342–366, 2006.
[6]:C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.
[7]: grplasso package https://cran.r-project.org/web/packages/grplasso/grplasso.pdf
[8]: gglasso package https://cran.r-project.org/web/packages/gglasso/gglasso.pdf
[9]: glmnet package https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
[10]: matrix package https://cran.r-project.org/web/packages/Matrix/Matrix.pdf