MAS8404 Statistical Learning Project

Abdullah Turki H Alshadadi, 190582184

Introduction

This is a report on the Wisconsin `BreastCancer` dataset. The goal is to build a classifier that is able to identify if a tissue sample has benign or malignant cancer tumor using only the 9 cytological characteristics in the dataset.

Exploratory Data Analysis

Before proceeding with the exploratory data analysis, the `BreastCancer` dataset contains `NA` row values. Due to the limted time given, these rows `NA` values will simply be removed reducing the dataset from 699 to 683.

The distribution of tissue sample of benign or malignant is 444 and 239 respectively. The data contain more benign tissue samples, 65.01%, than malignant tissue samples, 34.99%, by 2.17x. This could effect the generalisation of the classifier as it might understand the cytological characteristics that constitutes a benign tissue tumor much better than a malignant tissue. Therefore, an out-of-sample K-fold cross validation could be used to help mitigate bias model evaluation by better assessing the imbalanced class values of the tissue samples through ensuring each fold has balanced representation of both classes (that is benign and malignant).

Table 1: Distribution benign and malignant tissue samples in `BreastCancer` dataset

benign	malignant
444	239

Plotting a scatterplot matrix (See Appendix) shows that there are clear divide of benign and malignant samples, where lower cytological characteristics values tends to represent benign tissues whereas larger values shows a malignant values.

Without considering the “Mitoses” variable, benign tissue sample variables tend to range in the mean values of 1.26 to 2.96 whereas malignant tissue sample variables mean range from 5.33 to 7.63. This conveys a great separation of the classes which hints that modelling classifiers would likely have great accuracy identifying that tissue samples with lower cytological characteristics tends to be benign and higher tends to be malignant.

Although there are outliers like benign cytological characteristics containing a high value of 10 and malignant containing a low value of 1, it would unlikely affect the accuracy of the classifiers model. This is because benign tissue sample variables show lower spread in the standard deviation, ranging from 0.86 to 1.67 indicating that the outliers have little effect on the means of the dataset; on the other hand, even though malignant tissue samples contain higher spread with standard deviation ranging from 2.28 to 3.35, this is likely due to the larger range of mean values of the malignant variable, 5.33 to 7.63, as larger values tends to indicates higher chance of a malignant tissue tumor anyways. This is further supported by comparing the benign tissue medians ranging from 1 to 3 whereas the malignant tissue medians ranging from 5 to 10, showing the divide between lower cytological characteristics classifying as benign and higher classifying as malignant.

For the “Mitoses” variables, the mean values of benign and malignant tissues are much closer together, 1.07 and 2.54. This shows that it would likely be harder to distinguish between benign tissue samples and malignant tissue samples using that variable.

Table 2: Summary of benign tissue sample

Variables	Min	Max	Median	Mean	SD
Cl.thickness	1	8	3	2.96	1.67
Cell.size	1	9	1	1.31	0.86
Cell.shape	1	8	1	1.41	0.96
Marg.adhesion	1	10	1	1.35	0.92
Epith.c.size	1	10	2	2.11	0.88
Bare.nuclei	1	10	1	1.35	1.18
Bl.cromatin	1	7	2	2.08	1.06
Normal.nucleoli	1	8	1	1.26	0.95
Mitoses	1	8	1	1.07	0.51

Table 3: Summary of malignant tissue sample

Variables	Min	Max	Median	Mean	SD
Cl.thickness	1	10	8	7.19	2.44
Cell.size	1	10	6	6.58	2.72
Cell.shape	1	10	6	6.56	2.57
Marg.adhesion	1	10	5	5.59	3.20
Epith.c.size	1	10	5	5.33	2.44
Bare.nuclei	1	10	10	7.63	3.12
Bl.cromatin	1	10	7	5.97	2.28
Normal.nucleoli	1	10	6	5.86	3.35
Mitoses	1	9	1	2.54	2.40

Investigating further, the “Cell.shape” and “Cell.size” are highly correlated with a correlation of 90.72% showing that one of the variables is redundant to other as both represent similar cytological characteristics (See Appendix for the correlation heatmap).

Modelling

For the classifier models, 3 models will be built:-

1) Logistic Regression with BIC subset selection
2) Logistic Regression with LASSO regularisation
3) Linear Discriminant Analysis (LDA)

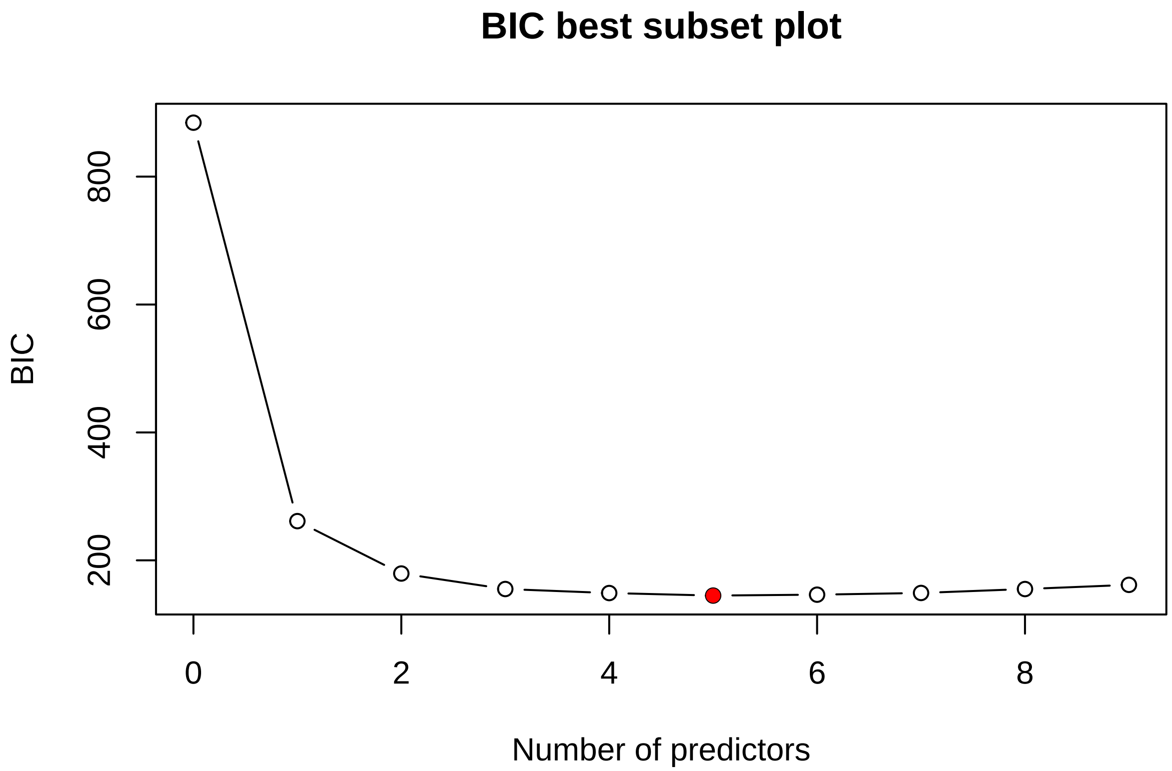
This section of the report examines how each classifier model behaves when inputted with the `BreastCancer` dataset, in terms of what variables where dropped and what are the most significant coefficients predicted variable in each model.

In the next section, “Cross Validation and Determining Best Model”, the models will be evaluated under the K-fold cross validation to determine which model is the “best” in identifying benign tissue samples and malignant tissue samples.

BIC Subset Selection

Using the BIC subset selection, it identified that the best subset of columns are "Cl.thickness", "Marg.adhesion", "Bare.nuclei", "Bl.cromatin", "Normal.nucleoli", a total of 5 out the 9 explanatory variables. These variables are the best coefficients to determining if a tissue sample is benign or malignant.

Running the Logistic Regression with the selected best subset by BIC had show that all of the selected explanatory variables are highly significant as all have have p-value of basically a zero.



```
## BIC
## BICq equivalent for q in (0.121688553587467, 0.668928807899912)
## Best Model:
##                   Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)     -10.1305998 1.09454253 -9.255556 2.131182e-20
## Cl.thickness      0.7412901 0.13188526  5.620720 1.901632e-08
## Marg.adhesion     0.3951547 0.11592178  3.408804 6.524829e-04
## Bare.nuclei       0.4473292 0.08797213  5.084896 3.678267e-07
## Bl.cromatin       0.5528700 0.15018648  3.681224 2.321174e-04
## Normal.nucleoli   0.3341920 0.09781468  3.416583 6.341227e-04
```

LASSO Regluarisation

Before running the Logistic Regression with LASSO regularisation, it is important to first determine the best lambda value through grid search to retrieve the optimal tuning parameter of lambda to get the best result of LASSO regularisation.

When selecting the optimal lambda variable that provides the lowest rate of misclassification, none of the coefficient has been dropped showing that in LASSO regularisation that all of the explanatory variables are significant enough for performing classification of tissue samples as benign or malignant in logistic regression.

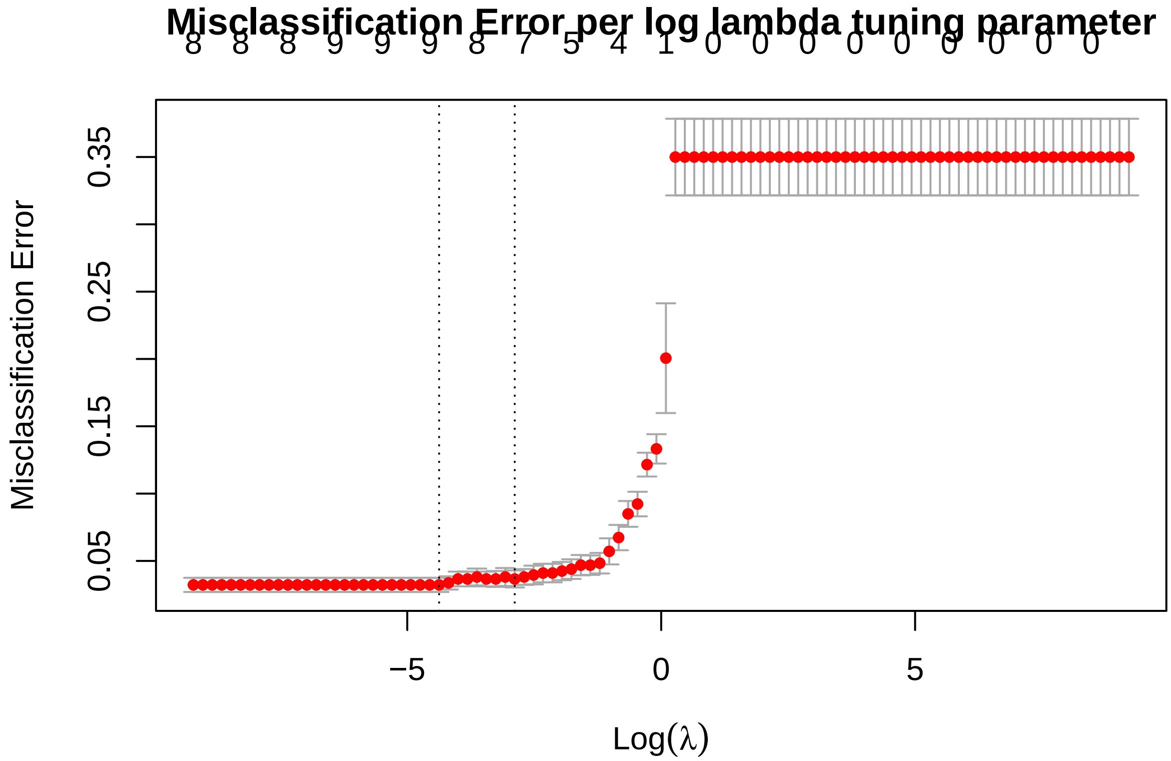


Table 4: Coefficients under the optimal tuning parameter

	s1
(Intercept)	-8.0869301
Cl.thickness	0.4582907
Cell.size	0.0868482
Cell.shape	0.2658567
Marg.adhesion	0.2252095
Epith.c.size	0.0424739
Bare.nuclei	0.3591685
Bl.cromatin	0.3128878
Normal.nucleoli	0.1867224
Mitoses	0.0942446

Linear Discriminant Analysis (LDA)

After running the LDA, investigating the histogram of the groups shows that there are an excellent separation of the data that represents benign (group 0) and malignant (group 1). This is more evident when looking at the group means for each explanatory variables, where benign tissue samples cytological characteristics are lower than malignant.

This makes logical sense because, as discuss in the "Exploratory Data Analysis", the cytological charateristics of benign tissue samples are generally lower than malignant tissue samples.

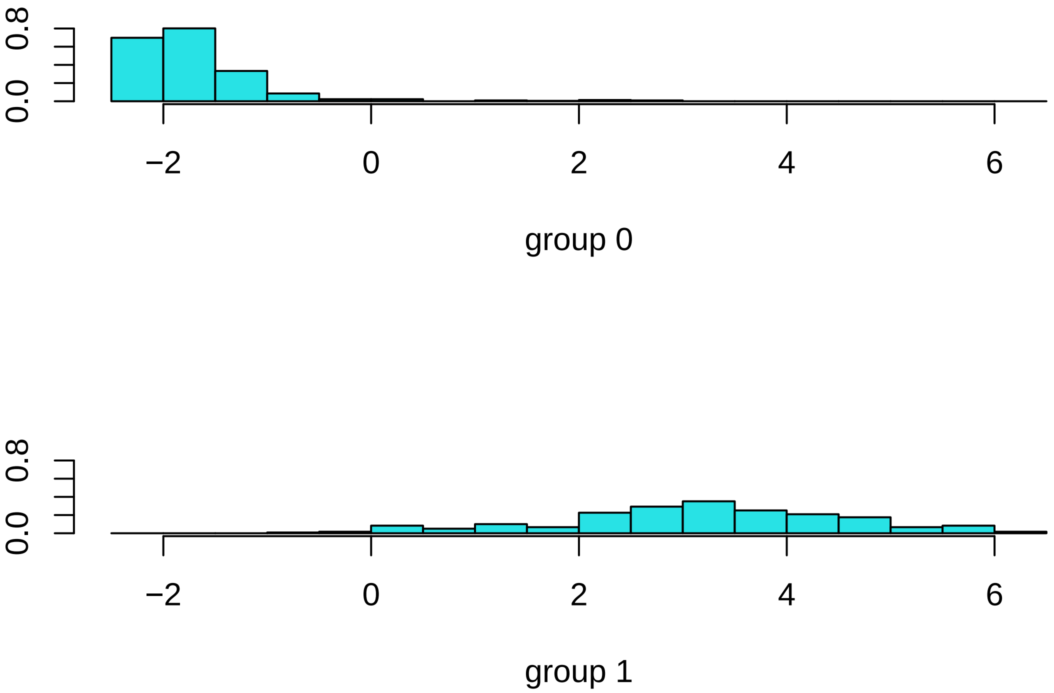


Table 5: LDA Group Means (rows 0 = benign, 1 = malignant)

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
0	2.963964	1.306306	1.414414	1.346847	2.108108	1.346847	2.083333	1.261261	1.065315
1	7.188284	6.577406	6.560670	5.585774	5.326360	7.627615	5.974895	5.857741	2.543933

Cross Validation Evaluation and Determining Best Model

As discuss under “Exploratory Data Analysis” and “Modelling”, K-fold cross validation will be used to evaluate each of the implemented models, specifically a K-fold of 10 will be used for better evaluation of the models.

Results

Table 6: Accuracy of Classification per model

models	train_accuracy_rate	train_error_rate	test_accuracy_rate	test_error_rate
Log Reg with BIC	91.57	8.43	90.46	9.54
Log Reg with LASSO	96.79	3.21	95.95	4.05
LDA	95.24	4.76	95.33	4.67

Table 7: Test set on Logistic Regression with BIC subset selection under K-fold 10 Cross Validation

	Predicted Benign	Predicted Malignant
Ground-Truth Bengin	96.40	3.60
Ground-Truth Malignant	15.48	84.52

Table 8: Test set on Logistic Regression with LASSO under K-fold 10 Cross Validation

	Predicted Benign	Predicted Malignant
Ground-Truth Bengin	97.75	2.25
Ground-Truth Malignant	5.86	94.14

Table 9: Test set on Linear Discriminant Analysis (LDA) under K-fold 10 Cross Validation

	Predicted Benign	Predicted Malignant
Ground-Truth Bengin	98.20	1.80
Ground-Truth Malignant	7.53	92.47

The Logistic Regression with BIC scores the highest test error rate of all the models with a rate of 9.54%. It also has the highest False Positives and False Negative of 15.48% and 3.6% respectively.

The Logistic regression with LASSO has the least test error rate of all the models with a rate of 4.05%. It also has the least False Positives and the second least False Negatives of 5.86% and 2.25% respectively.

The LDA has the second least test error rate of all the models with a rate of 4.67%. It also has the second least False Positives and the least False Negatives of 7.53% and 1.8% respectively.

Determining the Best Model

Although the Logistic Regression with BIC model achieved the worst metrics between the other models, it still has achieved a respectable 90.46% test accuracy with only 3 to 4 explanatory variables out 9 (See Appendix, Cross Validation Results). Furthermore, in the context of medical diagnosis of the tissue samples, the lower False Negative (3.6%) can be argued to be more important than the relatively high False Positives (15.48%) as falsely informing patients that they do not have a malignant tissue can lead to unchecked treatment and may lead to fatality compare to falsely informing patients that they do have a malignant as it means the patients will receive unnecessary treatment but with less risk of fatality. Therefore, this model could be used if speed is priority in diagnosis.

For the most accurate model, Logistic Regression with LASSO has the highest test accuracy rate of all models with a rate of 95.95% whereas the LDA model follows closely with test accuracy rate of 95.33%. However, even though it has better False Positive rate (5.86% compare to LDA model 7.53%), it has a worse False Negative rate of 2.25% compare to LDA models' False Negative rate of 1.8%. As discussed earlier, having better False Negative is much preferred than False Positives as miss diagnosing patients with benign but they actually have malignant tissues can lead to fatality due to unchecked treatment.

In summary, if the scenario values speed with relatively high accuracy then Logistic Regression with BIC is the best model otherwise if the scenario values absolute accuracy with little miss diagnosis of malignant tissues then the LDA model is the best model.

Appendix

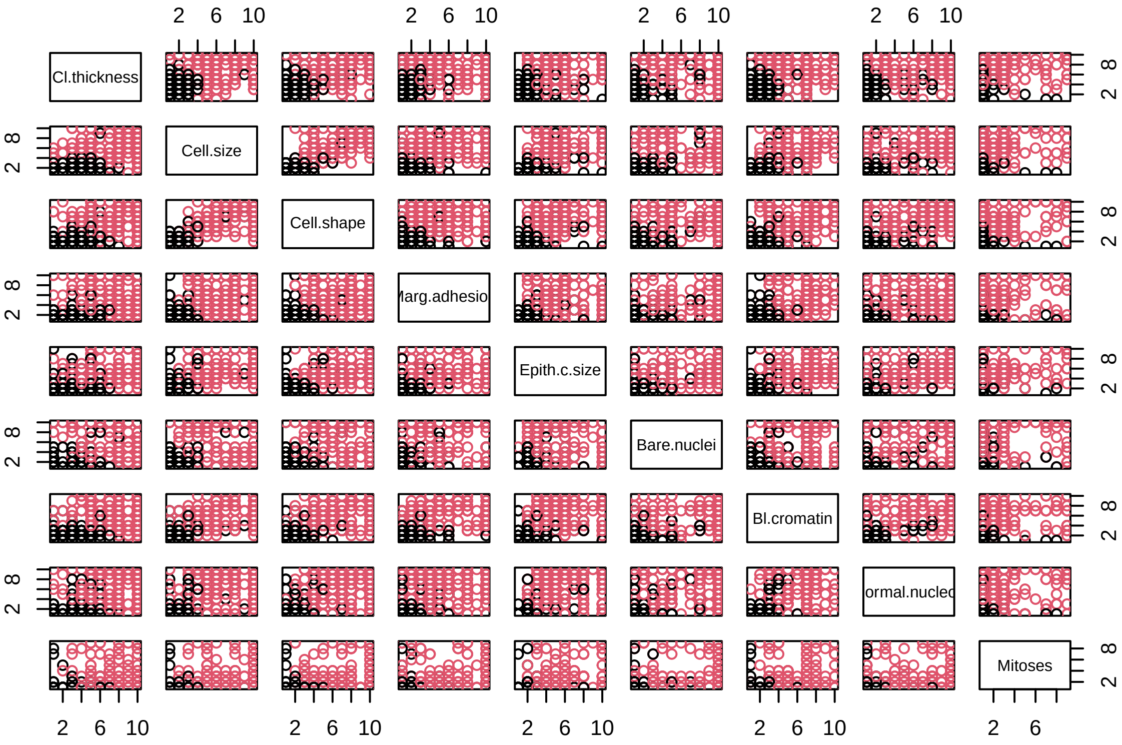


Figure 1: Scatterplot matirx of `BreastCancer` dataset (benign = black, malignant = red)

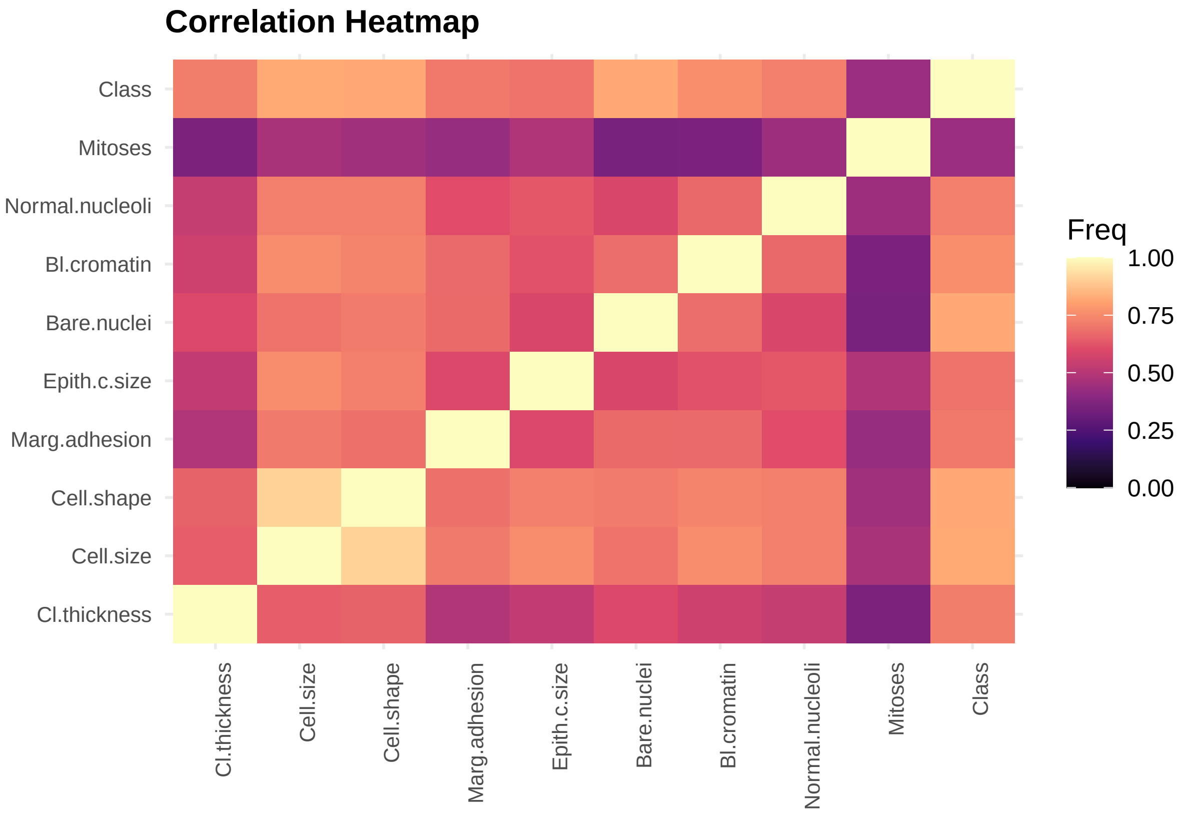


Figure 2: Correlation heatmap for the `BreastCancer` dataset

Cross Validation Results

```
## [[1]]
## BIC
## BICq equivalent for q in (0.341970469041527, 0.91865967136686)
## Best Model:
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      0.44858660 0.10148464  4.420242 1.167362e-05
## Marg.adhesion    0.10852084 0.02688207  4.036923 6.106474e-05
## Epith.c.size     0.24378731 0.03349755  7.277765 1.050940e-12
## Bl.cromatin     -0.09290854 0.03345031 -2.777510 5.646411e-03
## Normal.nucleoli  0.11416090 0.02573478  4.436055 1.087346e-05
##
## [[2]]
## BIC
## BICq equivalent for q in (0.0957661255559992, 0.831813204293244)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.41926204 0.09968141 4.206020 2.982576e-05
## Marg.adhesion   0.08451994 0.02553942 3.309392 9.891706e-04
## Epith.c.size    0.19406142 0.03428445 5.660334 2.311023e-08
## Normal.nucleoli 0.10129140 0.02536327 3.993625 7.287556e-05
##
## [[3]]
## BIC
## BICq equivalent for q in (0.366368931834001, 0.792084685741506)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.39368536 0.10137691 3.883383 1.142438e-04
## Marg.adhesion   0.07318969 0.02669981 2.741206 6.300520e-03
## Epith.c.size    0.22090592 0.03488841 6.331785 4.696661e-10
## Normal.nucleoli 0.09419744 0.02512536 3.749097 1.943278e-04
##
## [[4]]
## BIC
## BICq equivalent for q in (0.109474122515012, 0.644027108233034)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.43122685 0.10275232 4.196760 3.109721e-05
## Marg.adhesion   0.09233091 0.02660242 3.470771 5.555533e-04
## Epith.c.size    0.20384882 0.03522940 5.786327 1.149839e-08
## Normal.nucleoli 0.08535173 0.02617073 3.261343 1.170574e-03
##
## [[5]]
## BIC
## BICq equivalent for q in (0.0734433322319442, 0.900974518603936)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.41747283 0.10125165 4.123121 4.262805e-05
## Marg.adhesion   0.12295119 0.02725060 4.511871 7.728921e-06
## Epith.c.size    0.17923758 0.03492295 5.132372 3.866387e-07
## Normal.nucleoli 0.08819632 0.02599672 3.392594 7.377707e-04
##
## [[6]]
## BIC
## BICq equivalent for q in (0.198120429960181, 0.870641179233009)
## Best Model:
##                  Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.3595261 0.09946145 3.614728 3.256977e-04
## Marg.adhesion   0.1025741 0.02724533 3.764834 1.829140e-04
## Epith.c.size    0.2219006 0.03605329 6.154795 1.368890e-09
## Normal.nucleoli 0.0781535 0.02574266 3.035953 2.500868e-03
##
## [[7]]
## BIC
## BICq equivalent for q in (0.199426470605596, 0.838437242999495)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.41778636 0.10124110 4.126648 4.186562e-05
## Marg.adhesion   0.07957815 0.02620770 3.036441 2.494857e-03
## Epith.c.size    0.21838868 0.03482297 6.271397 6.731383e-10
## Normal.nucleoli 0.08306399 0.02543315 3.265973 1.151321e-03
##
## [[8]]
## BIC
## BICq equivalent for q in (0.405573457109976, 0.82303748910125)
## Best Model:
##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.38966105 0.10107382 3.855213 1.277561e-04
## Marg.adhesion   0.07170026 0.02673754 2.681633 7.522422e-03
## Epith.c.size    0.23159394 0.03592232 6.447076 2.301631e-10
## Normal.nucleoli 0.08984780 0.02586752 3.473383 5.499191e-04
##
## [[9]]
## BIC
## BICq equivalent for q in (0.363577485327038, 0.723525778138038)
## Best Model:
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    0.5452893 0.10823576  5.037978 6.217350e-07
## Cell.size      0.1260589 0.03503516  3.598067 3.467686e-04
## Marg.adhesion  0.1015073 0.02862603  3.545979 4.213512e-04
## Epith.c.size   0.2161256 0.03811863  5.669817 2.213706e-08
## Bl.cromatin   -0.1036206 0.03579516 -2.894820 3.930352e-03
##
## [[10]]
## BIC
## BICq equivalent for q in (0.0122127742418605, 0.687965304572225)
## Best Model:
##                  Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)     0.4202233 0.09866969 4.258890 2.380483e-05
## Marg.adhesion   0.1034882 0.02606206 3.970835 8.020591e-05
## Epith.c.size    0.1767754 0.03404840 5.191884 2.844280e-07
## Normal.nucleoli 0.1001664 0.02561559 3.910368 1.025515e-04
```

```
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -8.35347581
## Cl.thickness     0.45496510
## Cell.size        0.09142666
## Cell.shape       0.40999632
## Marg.adhesion    0.22660411
## Epith.c.size     .
## Bare.nuclei      0.29275679
## Bl.cromatin      0.34573567
## Normal.nucleoli  0.22061075
## Mitoses          .
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                           s1
## (Intercept)     -8.652743149
## Cl.thickness     0.500077044
## Cell.size        0.046652255
## Cell.shape       0.224060069
## Marg.adhesion    0.293180275
## Epith.c.size     0.005552757
## Bare.nuclei      0.402850280
## Bl.cromatin      0.338136248
## Normal.nucleoli  0.231421781
## Mitoses          0.262518649
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -7.48687262
## Cl.thickness     0.40539531
## Cell.size        0.18556152
## Cell.shape       0.28119890
## Marg.adhesion    0.16381120
## Epith.c.size     0.07861898
## Bare.nuclei      0.36498039
## Bl.cromatin      0.18268650
## Normal.nucleoli  0.18550062
## Mitoses          0.01109922
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -9.31105930
## Cl.thickness     0.55512056
## Cell.size        0.02752166
## Cell.shape       0.20569959
## Marg.adhesion    0.31816359
## Epith.c.size     0.09315774
## Bare.nuclei      0.42361686
## Bl.cromatin      0.38876021
## Normal.nucleoli  0.23940510
## Mitoses          0.21788915
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)     -8.9350319
## Cl.thickness     0.4359338
## Cell.size        0.1154163
## Cell.shape       0.3427896
## Marg.adhesion    0.2050466
## Epith.c.size     0.1287800
## Bare.nuclei      0.3477625
## Bl.cromatin      0.4391851
## Normal.nucleoli  0.1429208
## Mitoses          0.1592258
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -8.50517362
## Cl.thickness     0.47515939
## Cell.size        0.06897314
## Cell.shape       0.24381111
## Marg.adhesion    0.23506436
## Epith.c.size     0.10441778
## Bare.nuclei      0.36035775
## Bl.cromatin      0.34133455
## Normal.nucleoli  0.16357010
## Mitoses          0.18169346
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)     -8.8129794
## Cl.thickness     0.4993586
## Cell.size        0.0275321
## Cell.shape       0.2513003
## Marg.adhesion    0.2581297
## Epith.c.size     .        
## Bare.nuclei      0.3756788
## Bl.cromatin      0.4155207
## Normal.nucleoli  0.2323655
## Mitoses          0.2676339
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -8.07474748
## Cl.thickness     0.43471080
## Cell.size        0.03426199
## Cell.shape       0.26538350
## Marg.adhesion    0.29212254
## Epith.c.size     0.06766758
## Bare.nuclei      0.37202355
## Bl.cromatin      0.30126343
## Normal.nucleoli  0.16061365
## Mitoses          0.14978819
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -8.27523111
## Cl.thickness     0.47423177
## Cell.size        0.06344735
## Cell.shape       0.27972218
## Marg.adhesion    0.17407871
## Epith.c.size     0.03604048
## Bare.nuclei      0.37079809
## Bl.cromatin      0.33233686
## Normal.nucleoli  0.17085735
## Mitoses          0.18312475
## [1] "Fold Index: 11"
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -9.61036349
## Cl.thickness     0.54511006
## Cell.size        .
## Cell.shape       0.31100655
## Marg.adhesion    0.34787629
## Epith.c.size     0.07351312
## Bare.nuclei      0.36084059
## Bl.cromatin      0.41758819
## Normal.nucleoli  0.19151721
## Mitoses          0.30928606
```