HDDA-Project23

Regina Ibragimova

2023-12-09

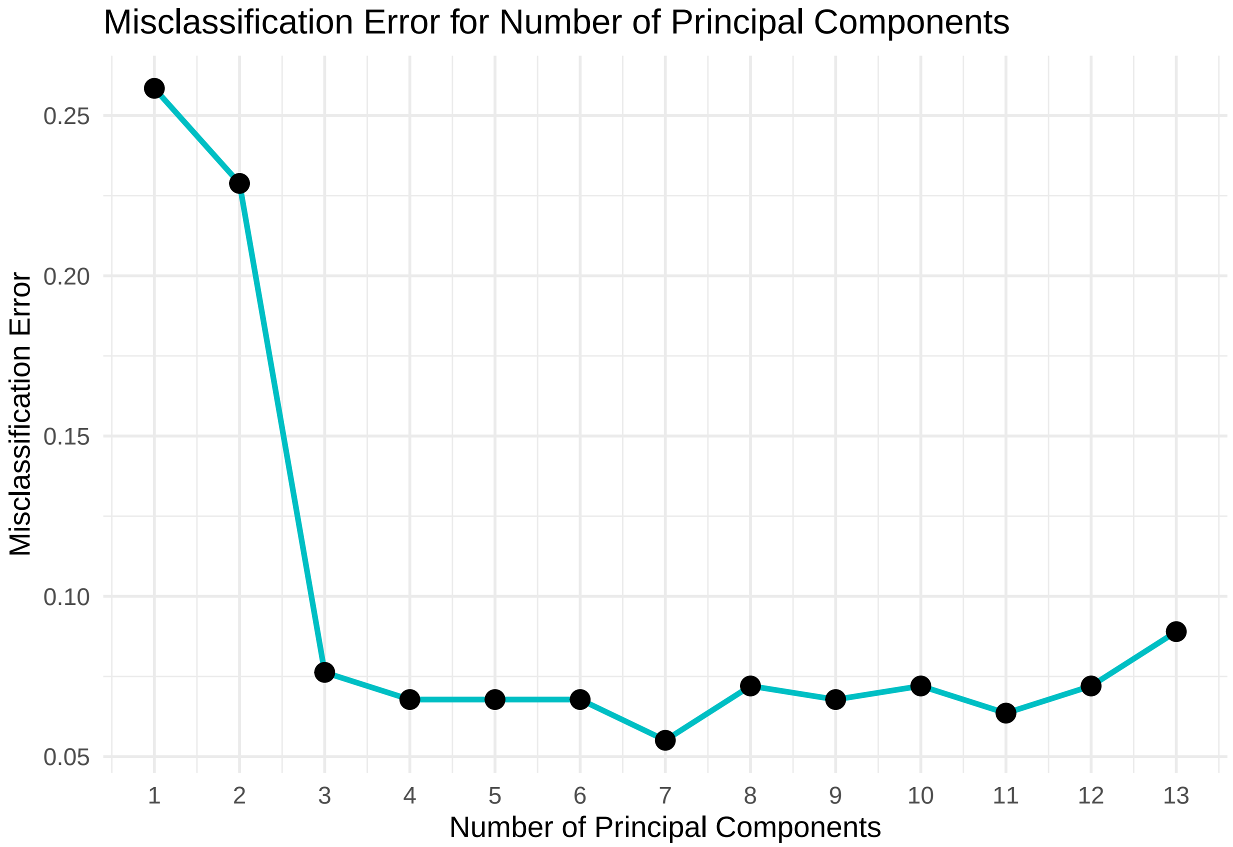
Part 1: Analysis of gene expression data

Evaluation and comparison prediction of the models

In this step, the goal was to predict the ER status using gene expression levels. First, the data was randomly split into a a training (70%) and test (30%) datasets. We evaluated Principal Component (PCR), Ridge Regression, and Lasso Regression models. The training dataset was utilized for model training, validation and hyperparameter tuning such as the number of Principal Components (PCs) for PCR and the regularization parameter λ in the Ridge and Lasso models. The test data was used for accessing the performance of models.

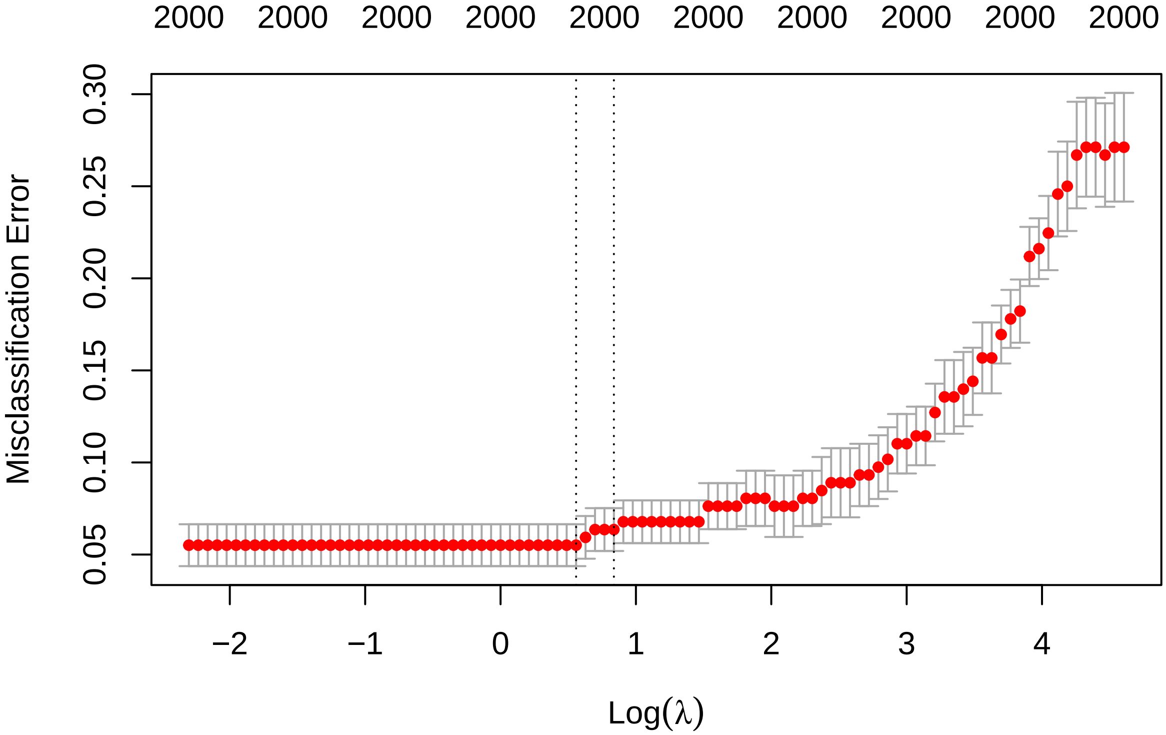
Principal Component Regression

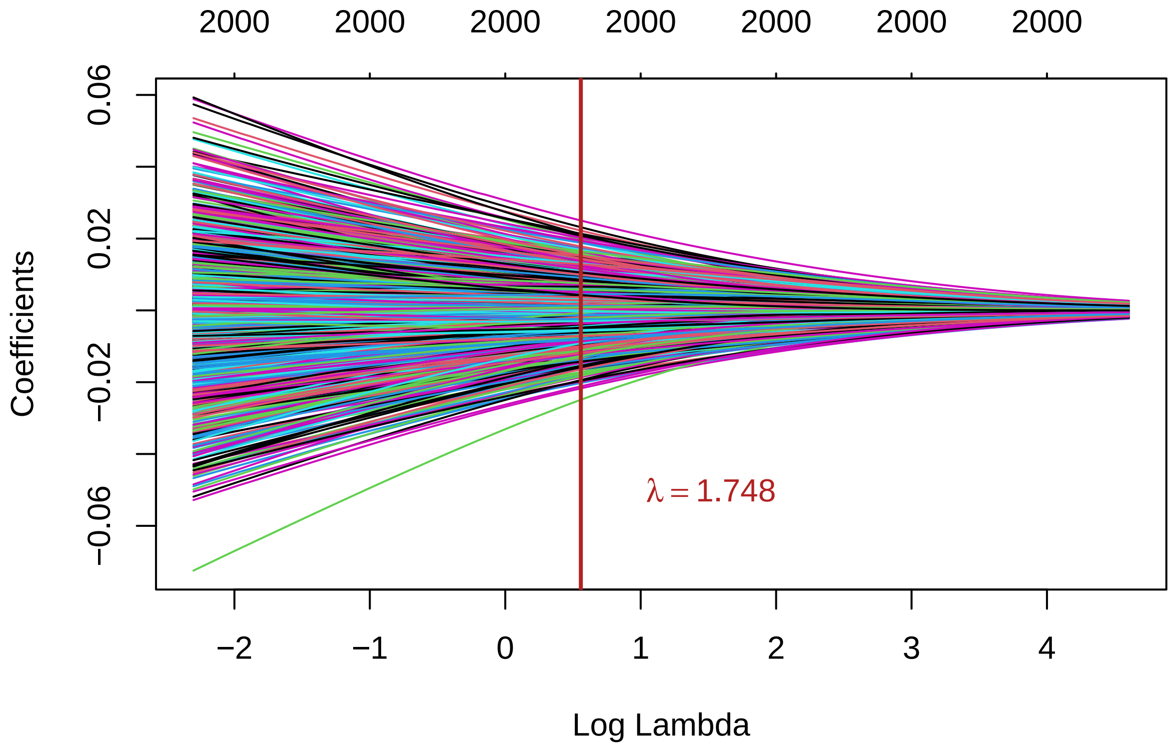
First, principal component analysis (PCA) was employed to reduce dimensionality of the gene expression data, resulting in 236 principal components (PCs). The training dataset was used to fit the model, and cross-validation was implemented to determine the optimal number of PCs, using the misclassification rate as the performance metric. A warning of 'fitted probabilities numerically 0 or 1 occurred' was observed when more than 13 PCs were used, suggesting potential overfitting of the model due to excessively good data separation. Through the cross-validation procedure, it was established that the first 7 principal components struck the best balance between capturing essential features of the gene expression data and accurately predicting the Estrogen Receptor (ER) status. The plot below visualizes the misclassification error rates resulting from the use of varying numbers of principal components in the predictive model. A sharp decline in error is observed as the number of components increases from 1 to 3, suggesting a significant gain in predictive power with the inclusion of the second and third principal component. The error decreases further, albeit at a slower rate, until it reaches a minimum at 5.5% at 7 components. Beyond this point, adding more components does not substantially improve the model's performance, and in some cases, it slightly increases the error, indicating a potential overfit or that these additional components do not contribute meaningful information for the prediction. Consequently, we decided to further restrict the model to these 7 PCs. These components were then selected for subsequent analyses. This approach significantly reduced the complexity of the model while retaining the most informative aspects of the data for predicting ER status in breast cancer patients.



Ridge Regression

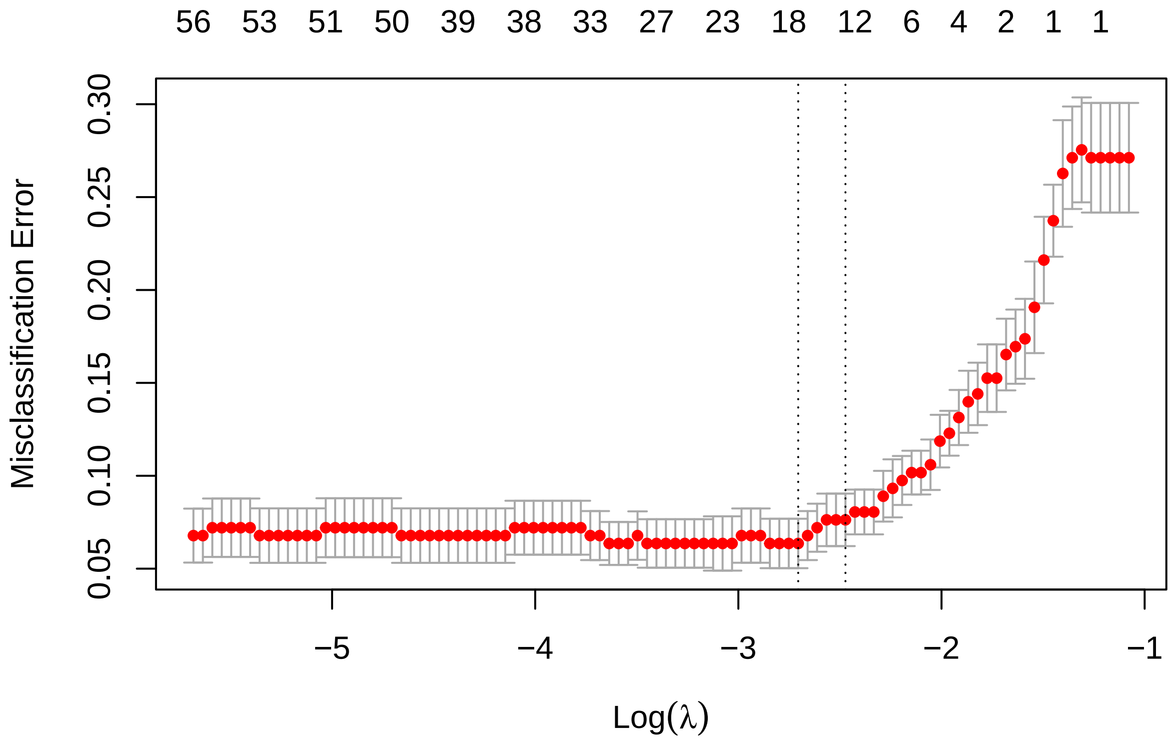
Next, Ridge Regression was applied to the training data as a regularized linear regression method to address potential overfitting. It incorporates an L_2 penalty term—squared magnitude of coefficients—into the loss function, regulated by the λ parameter. During cross-validation, the goal was to find the λ that yielded the lowest misclassification error. The optimal λ was found to be 1.748. A λ of 2.310, within one standard error of the optimal, was also considered as it corresponds to a simpler model that is expected to perform comparably to the model with the optimal lambda.

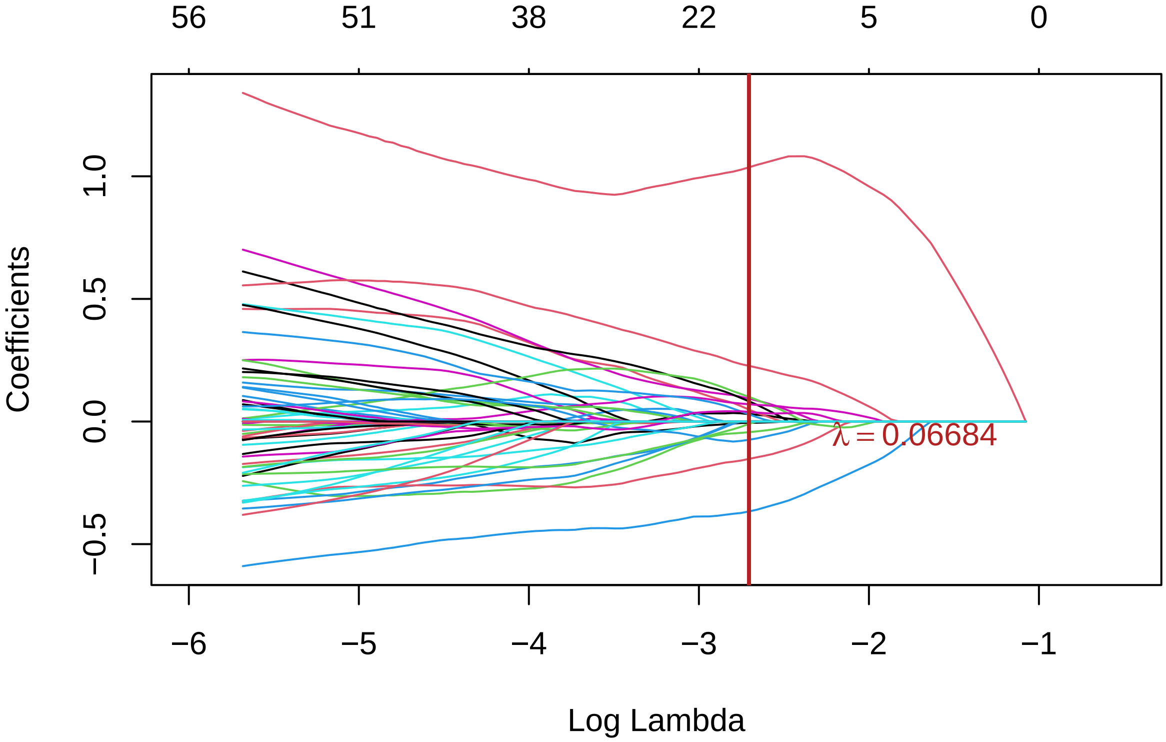

2000
2000
2000
2000
2000
2000
2000
2000
2000
2000
Misclassification Error
0.05
0.10
0.15
0.20
0.25
0.30
−2
−1
0
1
2
3
4
Log(λ)

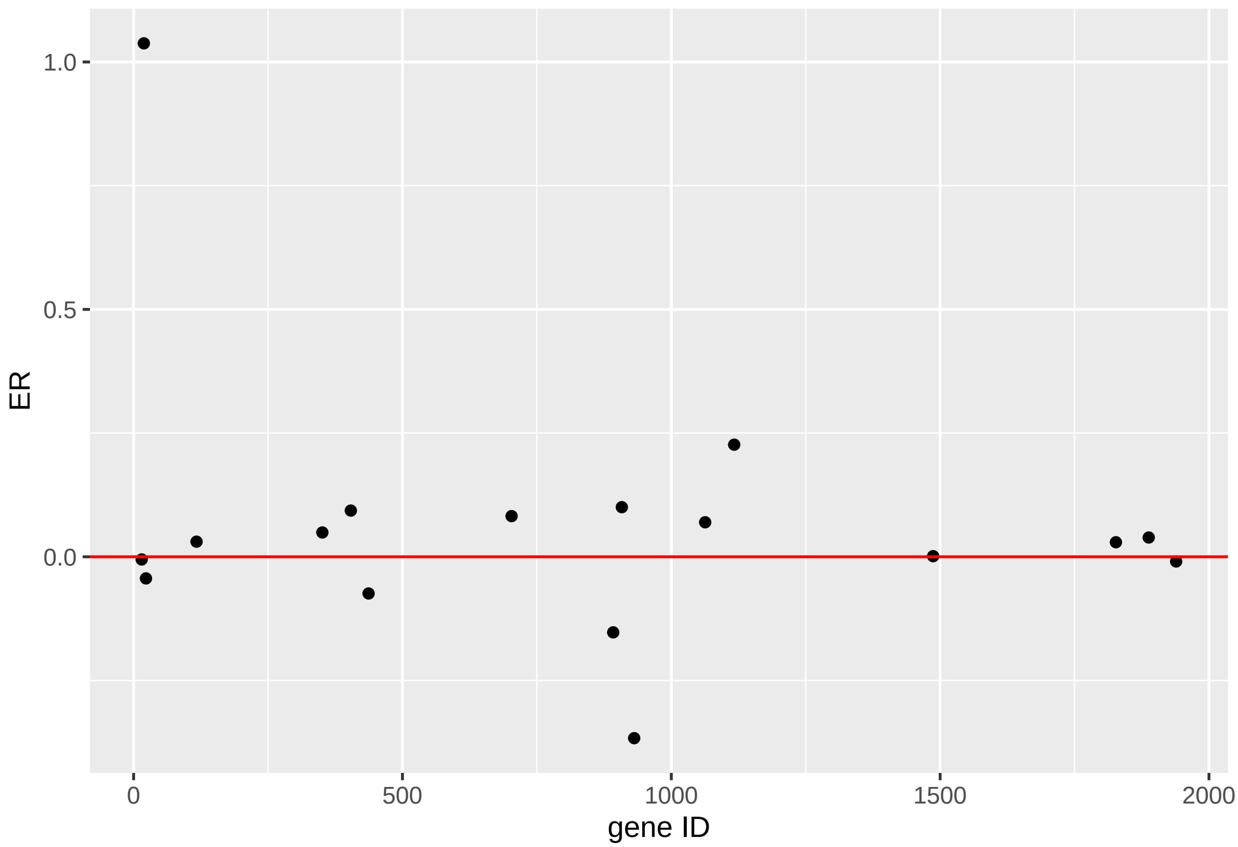


Lasso Regression

Finally, Lasso Regression, which utilizes L_1 norm penalty, was conducted using the training dataset. It can be seen at figure below, with increasing of λ the estimates shrunk towards zero, so it can be used as feature selection. Cross-validation determined the optimal λ to be 0.06684, with a larger value of λ equal to 0.08434, within one standard deviation of the optimal, also considered for comparison.


56
53
51
50
39
38
33
27
23
18
12
6
4
2
1
1
Misclassification Error
0.05
0.10
0.15
0.20
0.25
0.30
−5
−4
−3
−2
−1
Log(λ)


56
51
38
22
5
0
Coefficients
1.0
0.5
0.0
−0.5
λ = 0.06684
−6
−5
−4
−3
−2
−1
Log Lambda



The model with the optimal λ retains 17 predictors with non-zero coefficients, signifying the involvement of 17 genes. These genes are illustrated in the figure above, and their names are shown below.

```
##  [1] "Contig31197_RC" "Contig55949_RC" "Contig1970_RC"  "Contig26077_RC"
##  [5] "NM_000908"      "NM_001655"      "NM_000937"      "NM_003139"
##  [9] "Contig35702"    "Contig29543_RC" "NM_001792"      "NM_003224"
## [13] "Contig3794_RC"  "NM_004047"      "NM_002685"      "Contig52684"
## [17] "AI559539_RC"
```

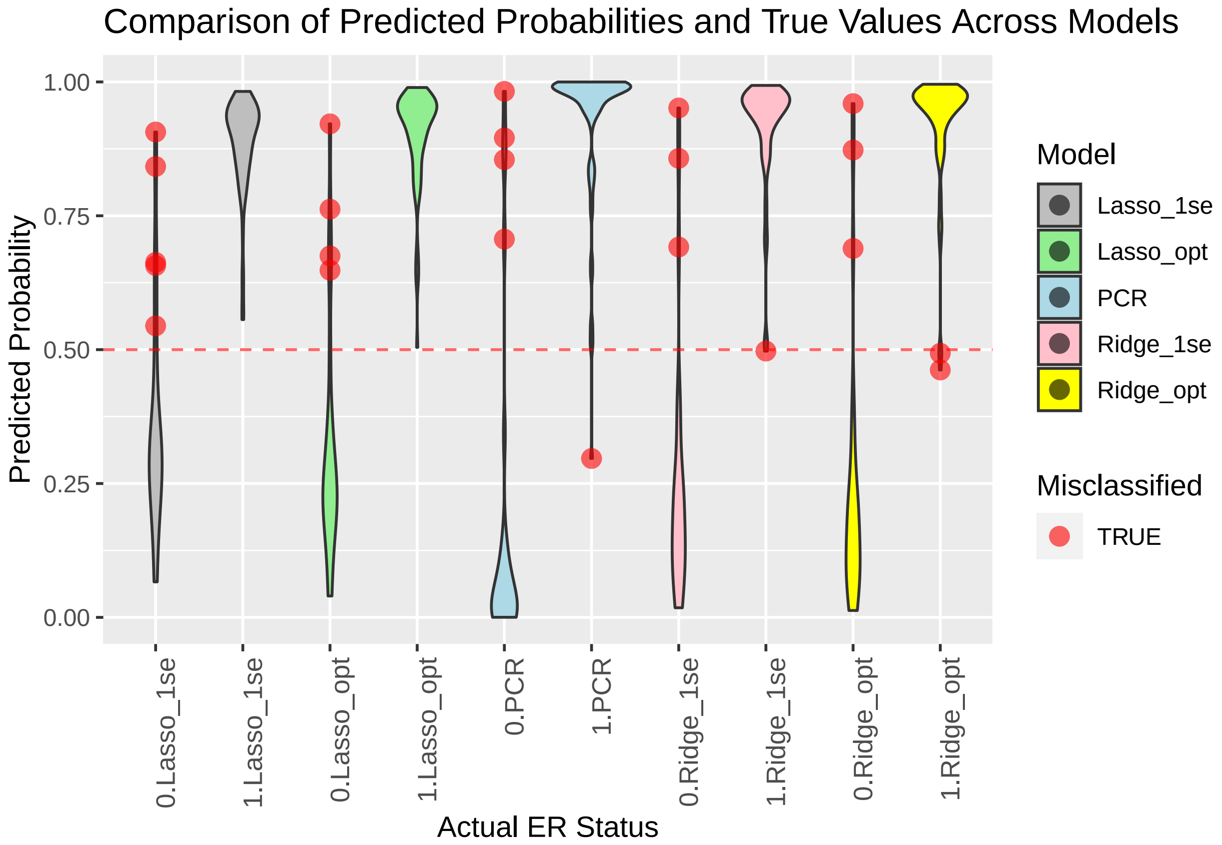
In contrast, with the λ set within one standard deviation of the optimal, the model includes only 12 genes, whose names are:

```
##  [1] "Contig55949_RC" "Contig1970_RC"  "Contig26077_RC" "NM_001655"
##  [5] "NM_000937"      "NM_003139"      "Contig35702"    "Contig29543_RC"
##  [9] "NM_001792"      "NM_003224"      "Contig3794_RC"  "Contig52684"
```

Evaluation and comparison of the models on test data

After constructing the optimal models which include PCR, Ridge Regression with optimal λ (Ridge_opt), Ridge Regression with a λ within one standard error (Ridge_1se), Lasso Regression with optimal λ (Lasso_opt), and Lasso Regression with a λ within one standard error (Lasso_1se), their performance was evaluated using the test dataset. The primary metric for this assessment was the misclassification error with a threshold being 0.5. The results indicated low misclassification errors for each model: PCR, Ridge_opt, and Lasso_1se all showed 4.5% error rate, corresponding to 5 incorrect predictions out of 101 test data points. Meanwhile, Ridge_1se and Lasso_opt demonstrated a slightly better performance at 3.9% error rate, equating to 4 misclassifications in the same 101 test data points. It is worth noting that each model presents a unique set of strengths and limitations. PCR is effective in handling multicollinearity but may lack in direct interpretability. Ridge Regression mitigates overfitting and maintains a comprehensive view by including all predictors, though this can make the model less parsimonious. Lasso Regression stands out for its feature selection capability, leading to more interpretable models but requires careful tuning to avoid excluding important variables.

To further illustrate the performance of the models, we created a violin plot that displays the predicted probabilities of ER status from different predictive models alongside the actual outcomes in test dataset. The x-axis shows the actual ER status, categorized as 0 (absence of ER) or 1 (presence of ER), and the predictive models used for each status. The y-axis represents the predicted probability for the ER status being 1. A violin shape at each category of actual ER status represents the kernel density estimation of the predicted probabilities for that category, showing the distribution and concentration of predictions made by each model. The red points represent misclassified instances, where the model's prediction did not match the actual ER status. These are plotted at the predicted probability levels for visualization. The spread of the violin plots indicates the range and density of the predicted probabilities. A narrower plot signifies less variance in predictions, while a wider plot indicates greater variance. The alignment of the violins and points across the models allows for a direct comparison of prediction distributions and misclassification rates between them. The graph also serves as a visual comparison between models, highlighting their predictive behaviors.



To sum up, the ER status can be predicted from the gene expression data with low misclassification errors, and the selection of the appropriate model depends on the specific goal of the study as well as requirements and nuances of the dataset in question.