layout	title	tags	aside	sidebar	mathjax	mathjax_autoNumber
article	Parkinson Disease Diagnosis from Acoustic Features	R Logistic lasso ridge CV Logistic_Regression LASSO_Regression Ridge_Regression Cross_Validation Parkinson Disease_Diagnosis UCI Acoustic_Features glmnet data analysis	toc: true	nav: layouts	true	true

Executive Summary

The Parkinson disease has so severe effects on patients and patients' families that early detection and intervention is needed. Acoustic features of patients are important in terms of predicting whether the patient has the Parkinson disease because it's not easy for patients to control their muscle. If we predict the outbreak of this disease with acoustic features, possible patients and their families would prepare for the disease in advance. In order to improve the prediction, we compare three classification methods, such as logistic regression, lasso logistic regression, and ridge logistic regression. The lasso and ridge logistic regression has a hyperparameter that controls the prediction ability, so we tune the hyperpameter optimally. In addition to the hyperparameter, we evaluate which acoustic feature has the most contribution to the Parkinson disease detection. From the best methods, we estimate the misclassification error rate for new patients. The lower misclassification error rate the better prediction.

Data Description

This data comes from UCI Machine Learning repository. This data set is donated April 10th, 2019.

The sample size of the data set is 240, meaning that there are 240 observations. Each observation has 44 variables, most of which are acoustic features. The purpose is to predict the positiveness of the Parkinson disease from voice-related data of participants such as pitch periods, voice noise, and perturbation. Important variables will be discussed in detail. Half of the participants has the Parkinson

disease, and the other half does not. So, this dataset is balanced. We don't need to consider imbalance of the dataset.

Data Analysis

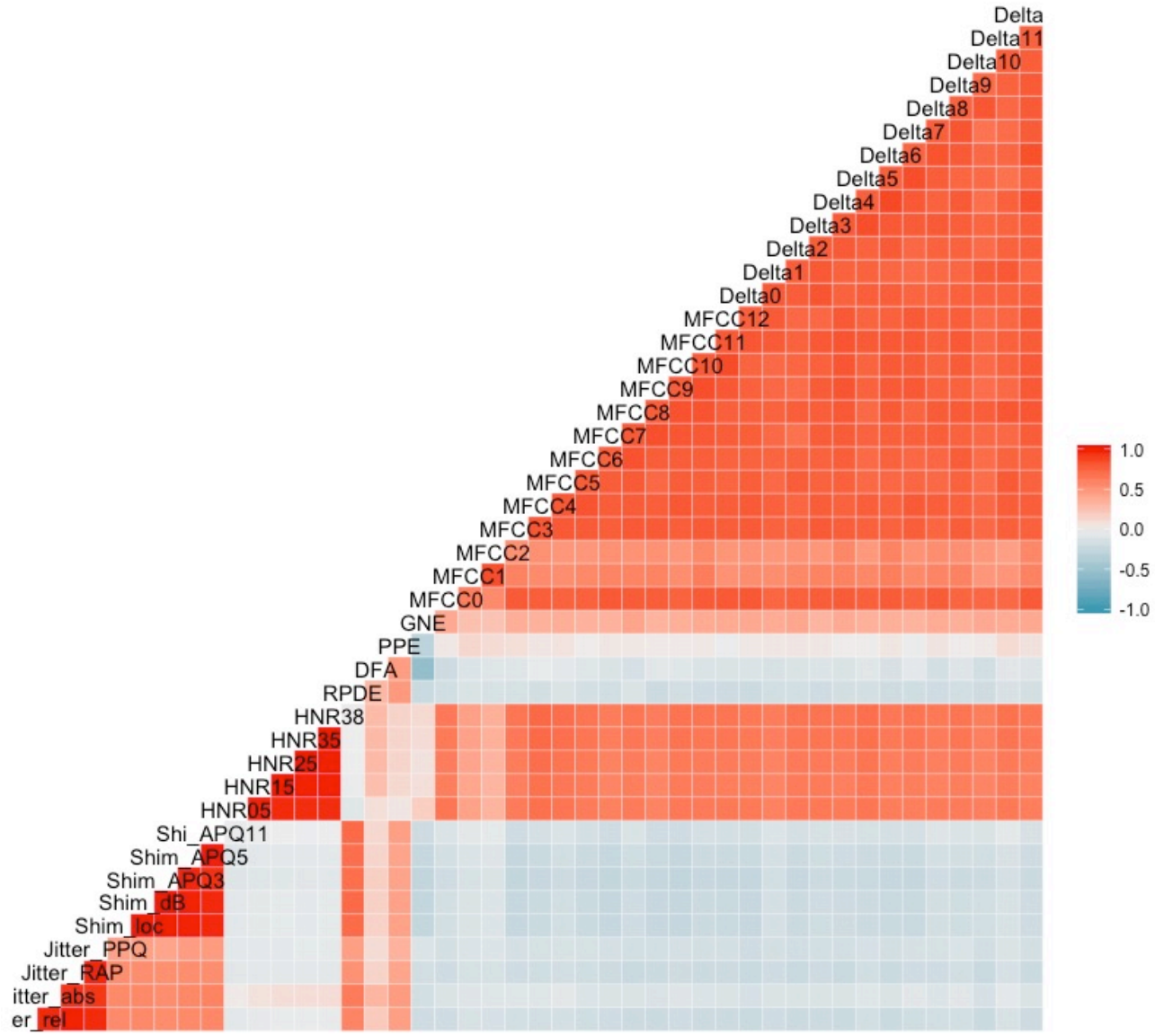
We first standardize the dataset of acoustic features. Then, this dataset randomly splits into n_{train} and n_{test} . n_{train} is 0.8n, and n_{test} is 0.2n. That is to say, n_{train} has 192 observations, and rest of the observations is for n_{test} . We can estimate the true misclassification error rate from the n_{test} . The true misclassification error rate can be computed only if we have a new dataset, so we have no choice but to estimate it. In order to compare the prediction ability, we mainly use misclassification error rates from 100 iterations of each method.

Regularization will be implemented for logistic regression. L1 norm regularization is lasso, and L2 norm regularization is ridge. Cross validation is used to tune hyperparameters of lasso and ridge as known as lambda, λ. The hyperparameters regulate coefficients of variables in the lasso and ridge regression, implying that the hyperparameters regulate the misclassification error rate. Commonly, 10-folds cross validation is known as the best; however, we will conduct an experiments to decide what number of folds is the best for reducing the misclassification error rate. In order to fix the number of folds, 15 types of folds will be tested from 3 folds to Leave-One-Out cross validation.

Then, we will decide the number of folds for lasso and ridge respectively, tune the hyperparmeters, and compare the misclassification error rates of 100 iterations in terms of logistic regression, lasso logistic regression, and ridge logistic regression. The misclassification error rates have three types: train error rate, test error rate and cross validation error rate. The train error rate comes from n_{train} , and the test error rate form n_{test} . The cross validation error rate is the minimum cross validation error rate of each type of folds.

Of course, trade-off between time and misclassification error rate will be discussed because time represents the computational complexity. For example, 5 percentage of the error rate with 10 second is of course better than 4.8 percentage with 10 minutes. Variable importance will be considered in terms of the lasso and ridge regression.

Correlation Analysis



As shown above, similar acoustic features have strong positive correlations, which is close to one. For example, Mel frequency cepstral coefficients and their derivatives(MFCC and Delta) have strong correlations. However, different acoustic features are almost uncorrelated such as Mel frequency cepstral coefficients and Pitch period entropy(PPE). Surprisingly, Harmonic-to-noise ratios are highly, positively correlated with Mel frequency cepstral coefficients and their derivatives. Negative correlations are not recognized noticeably.

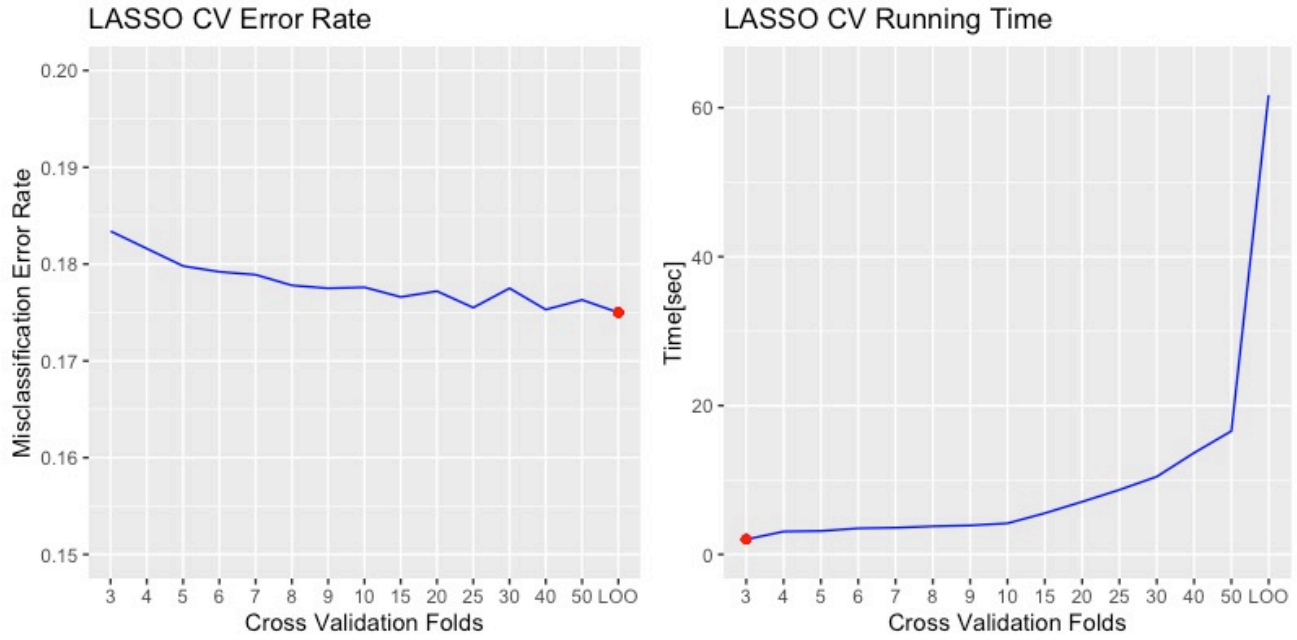
Model Building

General Comparison between Logistic, LASSO, and Ridge

Each box plot comes from 100 iterations for each algorithm. Logsitic regression is overfitted because its train error rate is under 10 percentage at average while the test error rate is over 25 percentage. That is to say, the test error rate cannot be estimated by the train error rate because the train error rate is much better than the test error rate. Unlike the logistic regression, lasso and ridge regression are not overfitted since the train error rates are close to the test error rate even though the train error rates are worse than that of the logistic regression. The test error rate is computed using the hyperparameter which is chosen by the 10 folds cross validation.

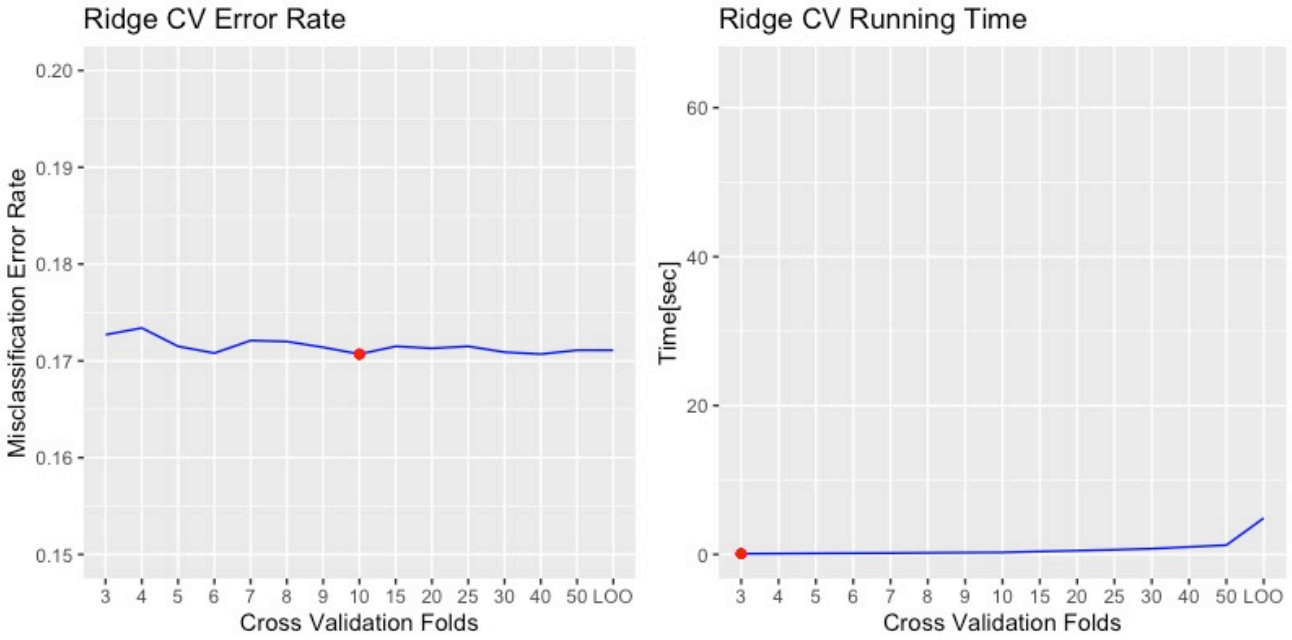
In general, 10 fold cross validation is commonly used in data analysis as used in this section. The cross validation error rate with 10 folds is also a good estimate of the test error rate for the lasso and ridge regression as we can see the similarity between the test error rate and the cross validation error rate. Therefore, figuring out what number of folds computes the best cross validation error rate is important in terms of estimating the test error rate of the lasso and ridge regression. We evaluate the number of folds in next sections.

Optimal K Fold Cross Validation for LASSO Regression



The error rates and time above are the average of 100 iterations of lasso logistic regression. The minimum value is marked as a red point. As the number of folds increase, the cross validation misclassification error rate converges to 17.5 percentage as shown. Of course, Leave-One-Out(LOO) cross validation the best tunes the hyperparameter, λ, for the lasso regression; however, it takes way more time than other folds cross validation. For example, 50 folds cross validation takes less than 20 seconds, but LOO cross validation takes three times as much as the 50 folds cross validation takes. Considering the $n_{train} = 192$, it might be said that LOO cross validation has 192 folds. Even though the LOO cross validation spends much more time, it doesn't improve the misclassification error rate remarkably. 25 folds, 40 folds, and LOO cross validation show the similar error rates each other. So, 25 folds cross validation for the lasso regression is the best choice with respect to the trade-off between time and error rates.

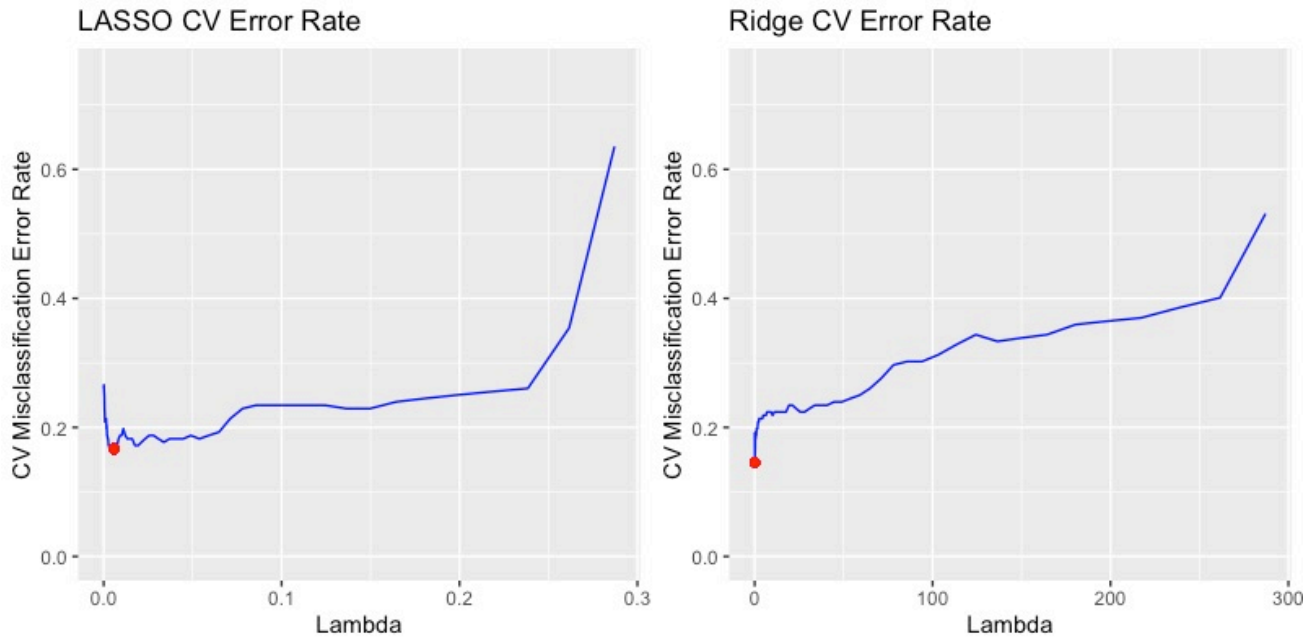
Optimal K Fold Cross Validation for Ridge Regression



When comparing the cross validation error rates of 100 iterations, 10 fold cross validation is the best in the ridge regression as shown above. Interestingly, increasing the number of folds doesn't improve the cross validation error rate. We don't need to consider the time to run the cross validation because the LOO cross validation, which takes the longest time to run, takes only less than 10 seconds.

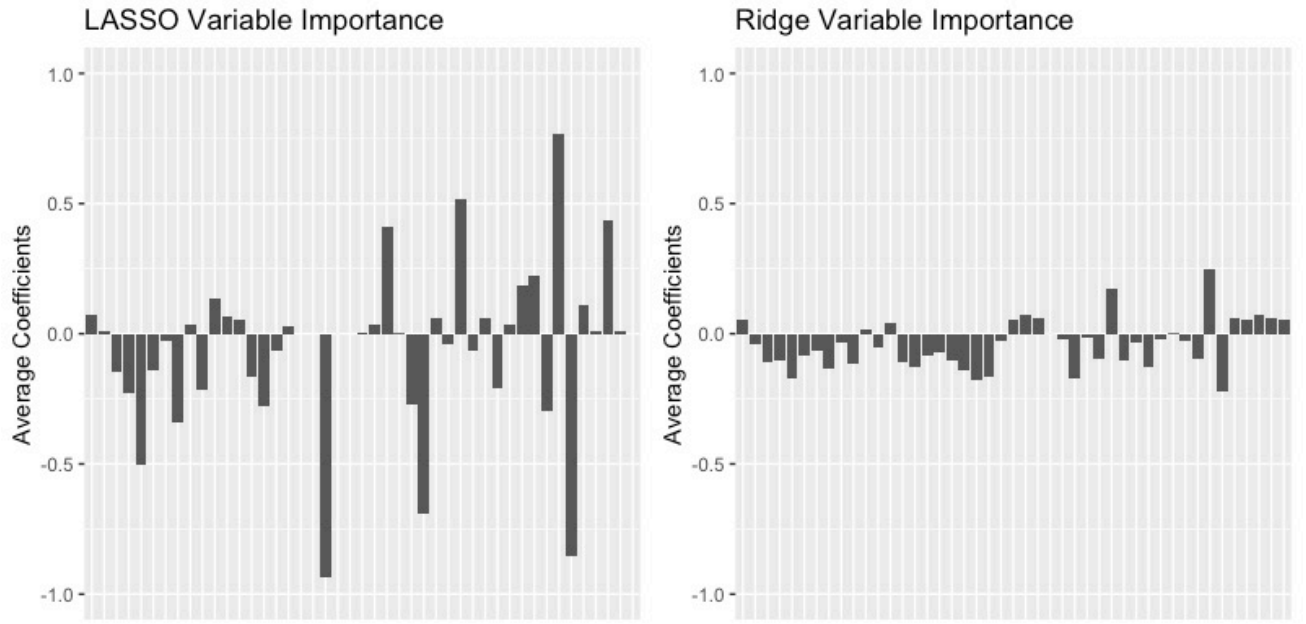
In this dataset, the ridge regression takes much shorter time to cross validate than does the lasso regression in general. For example, the LOO cross validation for the lasso regression spends more than 60 seconds, but that for the ridge regression spends less than 10 seconds. The different fold cross validation keeps this pattern in Figure 3 and Figure 4.

Optimal Hyperparameters for LASSO and Ridge Regression



The best hyperparameter for the lasso regression is 0.0058, and that for the ridge regression is 0.1272. The lowest cross validation error rate is marked as a red dot. That lowest cross validation error rate gives us the optimal hyperparameters. As the hyperparameter increases, the cross validation misclassification error rate, which is a good estimate of the test error rate, becomes worse in the both cases as shown above. The reason for that is all of variable coefficients approach to zero as the hyperparameter, λ, increases. If the hyperparameter is zero, the lasso and ridge regression is the same as the logistic regression. Of course, the range of λ for the lasso regression is from zero to one, and that for the ridge regression is from zero to positive infinity.

Variable Importance



Regarding the variable importance, there are similar patterns above. The order of variables doesn't change in the x axis. Some variables with negative coefficients in the lasso regression tend to have negative coefficients in the ridge regression. Other variables with positive coefficients in the lasso regression also do. That is to say, the sign of variable coefficients keeps consistent in the lasso and ridge regression. However, the absolute value of variable coefficients is greater in the lasso regression than in the ridge regression. Of course, the lasso regression has zero coefficient variables, but the ridge regression doesn't.

Top 5 Positive and Negative Coefficient Variables in LASSO

Top 5	**Positive**	**Top 5**	**Negative**
PPE	0.77	HNR35	-.93
MFCC2	0.52	RPDE	-.85
Shim_APQ5	0.44	MFCC10	-.69
Jitter_rel	0.41	Delta11	-.50
MFCC8	0.22	Delta3	-.34

Top 5 Positive and Negative Coefficient Variables in Ridge

Top 5	**Positive**	**Top 5**	**Negative**
PPE	0.25	RPEDE	-.22
MFCC2	0.18	HNR35	-.17
Shim_APQ5	0.08	MFCC10	-.17
Jitter_RAP	0.07	Delta11	-.17
Shim_dB	0.06	HNR38	-.16

In the tables above, Top 3 are the same in the lasso and ridge regression regardless of the sign of the coefficients of variables. Top 3 important positive variables are PPE, MFCC2, and ShimAPQ5. PPE is a new measurement of pitch variation in voice. MFCC2 is the second coefficient of MFC, the mel-frequency cepstrum (MFC). MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. ShimAPQ5 is a measurement of perturbation to the small, rapid, cycle-to-cycle changes of period in the fundamental frequency of the voice and 5 point amplitude that occur during phonation.

Pitch variation has the most positive contribution to the Parkinson disease diagnosis detection. In other words, if a patient has a big pitch variation, he or she is more likely to be diagnosed to have the Parkinson disease. The second coefficient of MFC is the second most positive contribution to the Parkinson disease

diagnosis, and the third most positive contribution is the changes of period in the fundamental frequency of voice and 5 point amplitude.

Top 3 important negative variables are HNR35, RPDE, and MFCC10. HNR35 is a harmonic-to-noise ratio in the frequency band of 0-3500 Hz. RPDE is an acronym of Recurrence period density entropy for determining the periodicity, or repetitiveness of a signal. MFCC10 is the tenth coefficient of MFC previously explained in the important positive variables.

The harmonic-to-noise ratio in 0-3500 Hz has the most negative contribution to the Parkinson disease diagnosis detection. That is to say, if a patient has a high value of harmonic-to-noise ratio in the frequency range, the patient is less likely to have the Parkinson disease. The repetitiveness of a voice signal is the second most negative contribution, and the tenth coefficient of MFC is the third. Interestingly, the second coefficient of MFC is the second most positive important variable while the tenth coefficient of MFC is the third most negative important variable. So, MFC, a representation of the short-term power spectrum of a sound, has positive and negative contribution to the Parkinson disease diagnosis.

In the ridge regression, the order of Top 3 important variables of negative coefficients changes. The repetitiveness of a voice signal is the most negative one and the harmonic-to-noise ratio is the second most negative.

Best Method between Logistic, LASSO, and Ridge Regression

Confusion Matrix for Ridge Regression

	True Negative	True Positive
Predicted Negative	102	17
Predicted Positive	18	103

The best method for this dataset is the ridge logistic regression with $\lambda = 0.1272$ chosen by 10 fold cross validation because the ridge regression shows the lowest test error rate with the optimal lambda between the logistic, lasso, and ridge regression in Figure 7. In addition, the ridge regression is the least overfitted between three methods, implying that the train error rate is very close to the test error rate.

The logistic regression is the most problematic one since it's overfitted severely, and the lasso logistic regression is also more overfitted than the ridge logistic regression.

Model Fit

Instead of $n_{train} = 0.8n = 192$, When we use the full dataset of $n = 240$ for the ridge regression with $\lambda = 0.1272$, the train error rate is 0.146. The train error rate of 0.146 is underestimated the true misclassification error rate, implying that the true misclassification error rate from a new dataset of new patients definitely expected to be higher than the train error rate. However, since the ridge regression is less overfitted than other methods and the test error rate from $n_{test} = 0.2n = 48$ is just slightly higher than the train error rate from $n_{train} = 0.8n = 192$, the true misclassification error rate of the new dataset wouldn't be much worse than the train error rate from the full dataset of n = 240.

As the dataset is balanced, the confusion matrix of the ridge regression is also balanced in Table 3. That means that the prediction have false negative and false positive evenly.

Conclusion

The ridge logistic regression with $\lambda = 0.1272$ is the best method for this dataset since the ridge regression is not overfitted too much and has the best estimate of the true misclassification error rate. In other words, the train error rate of 0.146 from the full dataset can be a good estimate for the true misclassification error rate when we face a new dataset. For example, when we diagnose 100 new patients with this ridge logistic regression model, the misclassification error rate would be around 15 percentage. In addition to the prediction accuracy, this model doesn't spend too much time to cross validate and train the model because we already consider the trade-off between time and error rate in the section of 3.2.2 and 3.2.3. In particular, this dataset requires less than 5 second to cross validate with 10 folds.

We tune the hyperparameter, λ. When we tune the hyperparameter of the ridge logistic regression, we implement the cross validation method. However, we eclectically choose 10-fold cross validation. We compare 15 types of cross validation folds from 3 fold to LOO cross validation by taking the average of

100 iterations of cross validation error rates. Of 15 types of the folds, 10-fold cross validation is the best one that gives the best hyperparameter, $\lambda = 0.1272$.