Predicting Breast Cancer

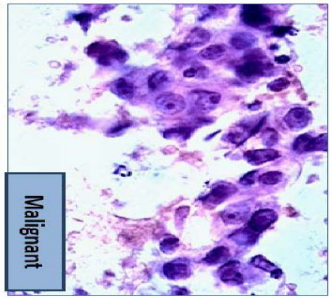
Anders Ward
Kaustubh Deshpande
Sasha Farzin-Nia

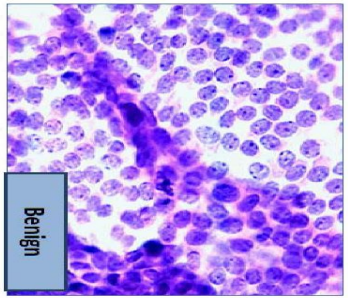
Introduction

- Breast Cancer one of the most common cancers diagnosed among women
- If a mass is discovered in the breasts, a biopsy is performed
- Up to a pathologist to determine the status of mass
- Advancement in technology and statistical methods can help provide support

Our Goal

- Dataset used is the “Breast Cancer Wisconsin (Diagnostic) Data Set”
  - Made publicly available on UCI Machine Learning Repository
  - Features of dataset are computed from the digitized image of the biopsy specimen
- Compare PCA with Lasso and Ridge Logistic Regression and a GLM

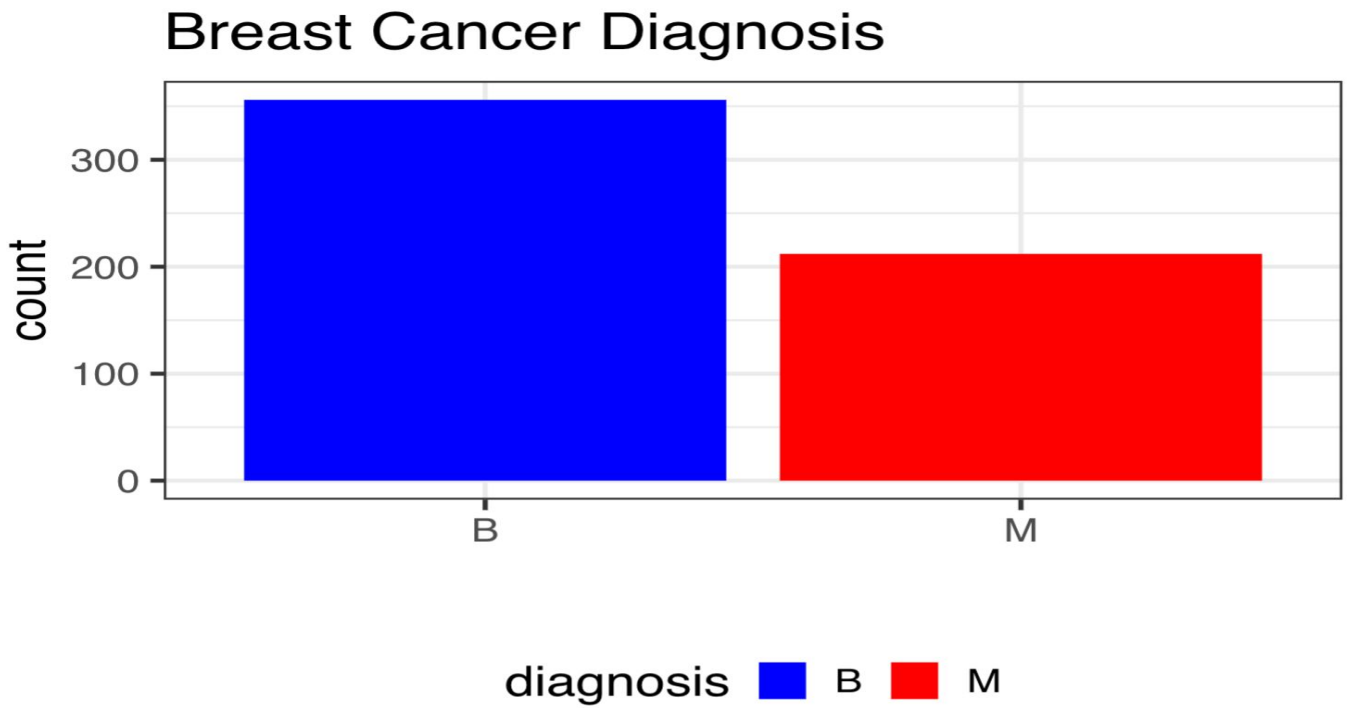




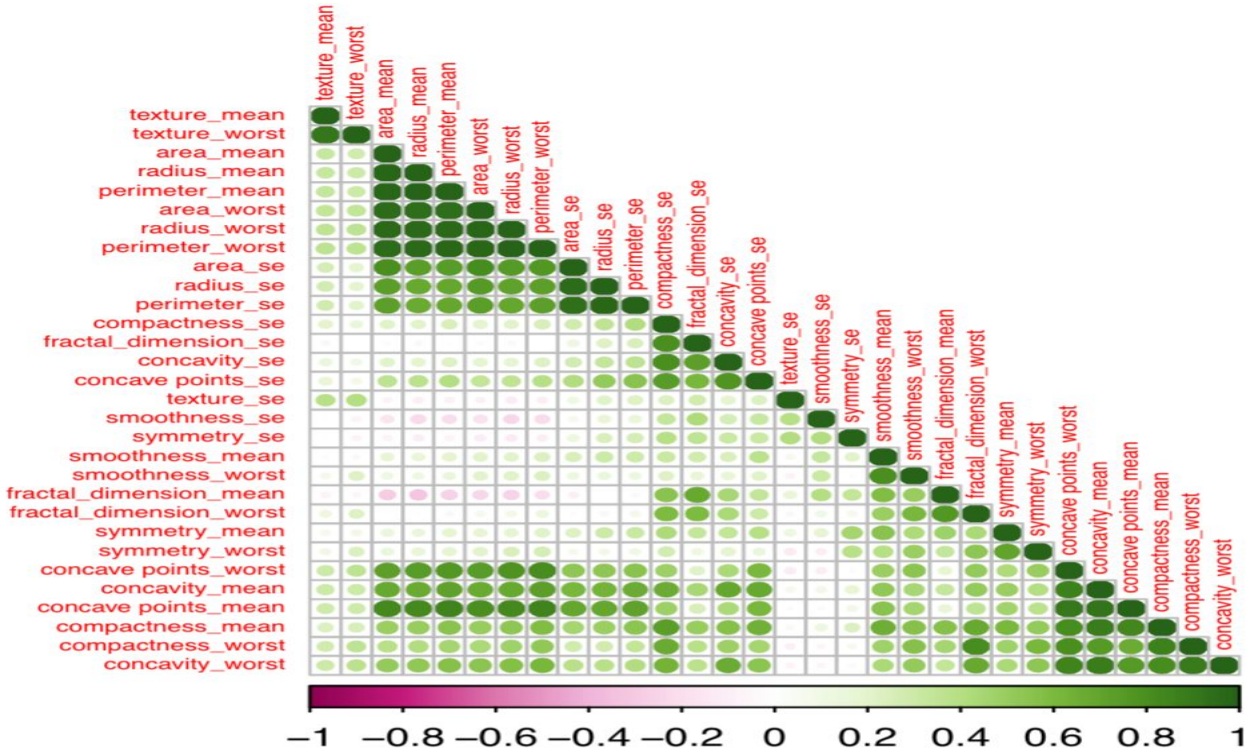
Data Structure

- 568 cases
- 30 predictors
- Most predictors were some sort of measure of the geometric shape of the mass
- Response Variable
    - Benign - Free of Cancer
    - Malignant - Cancer

Benign Vs. Malignant

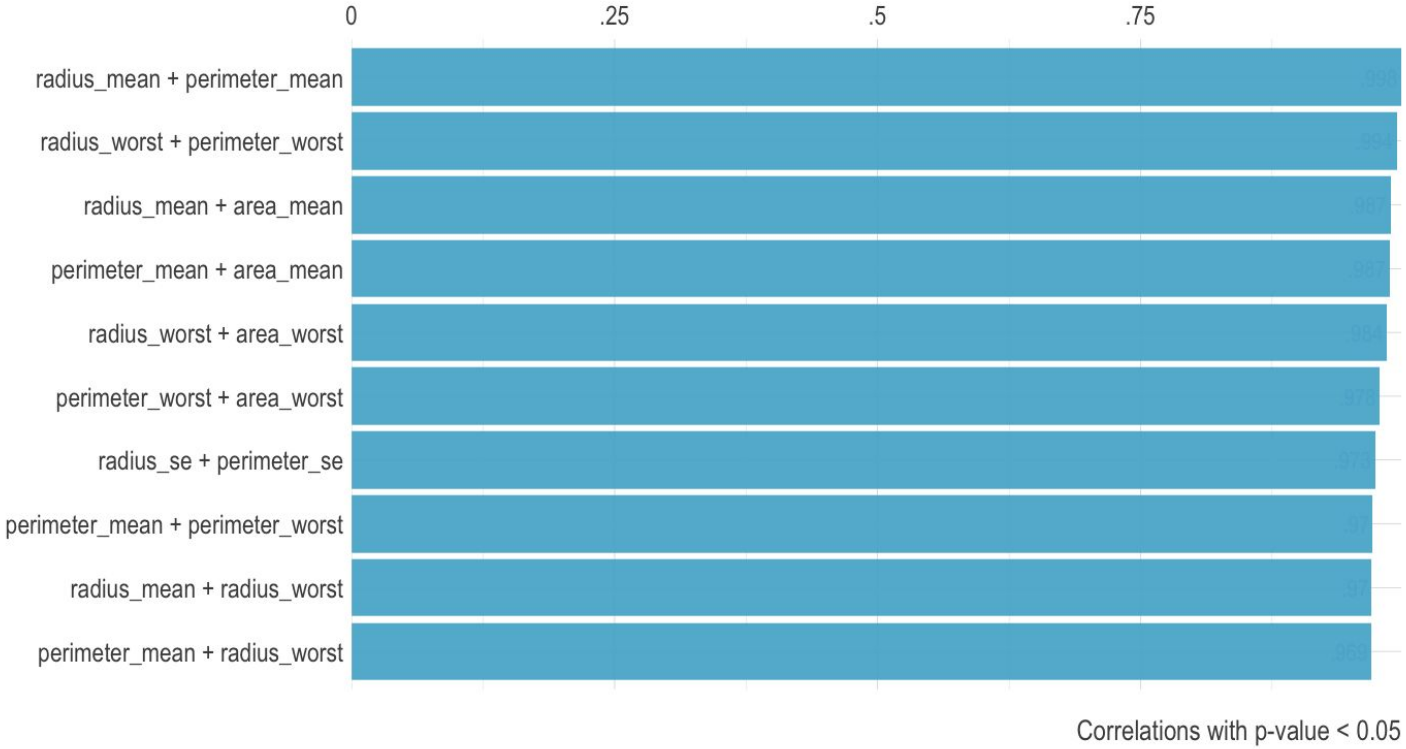


Correlation Plot

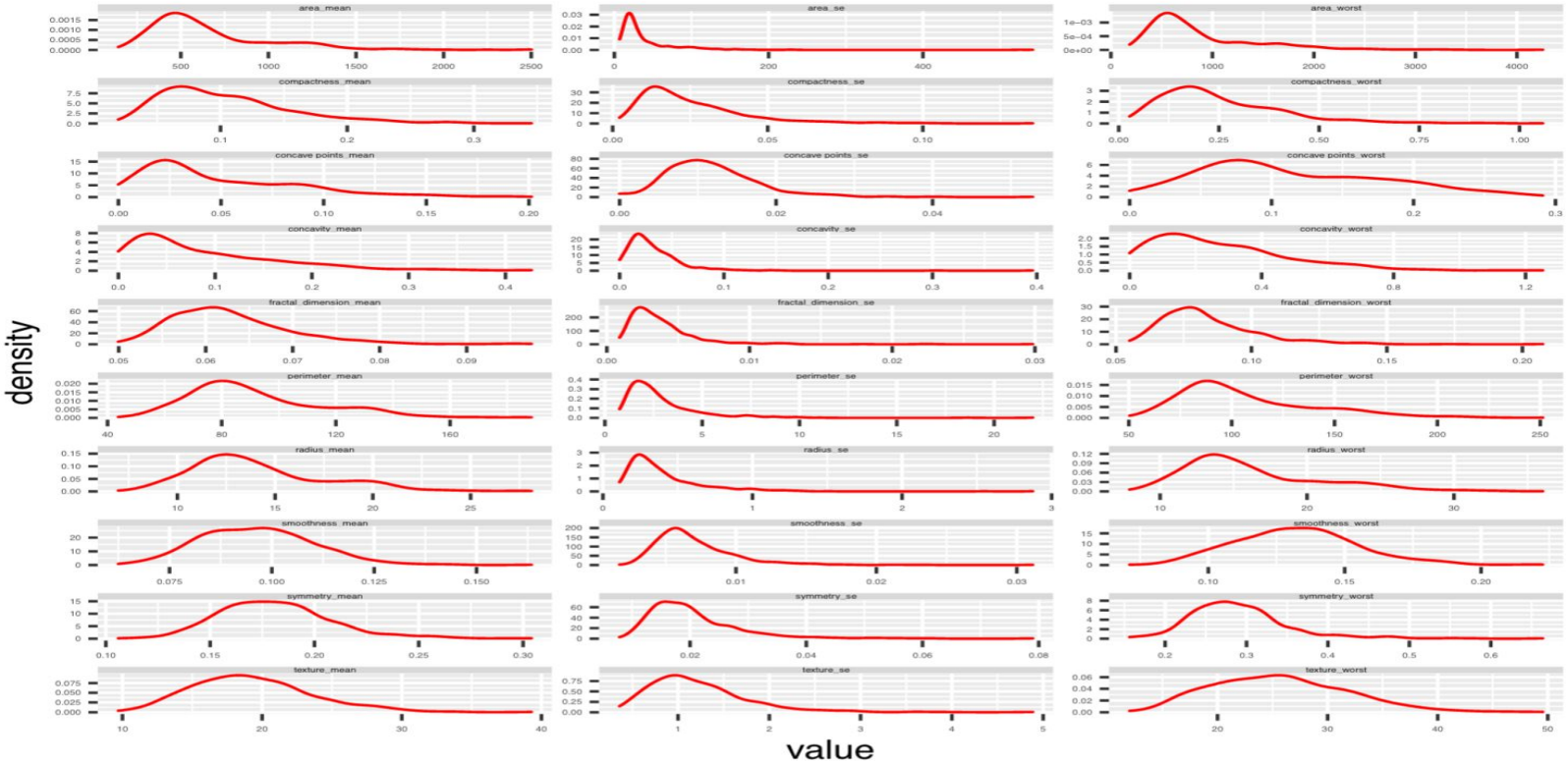


Ranked Cross-Correlations

10 most relevant



Distribution Plots



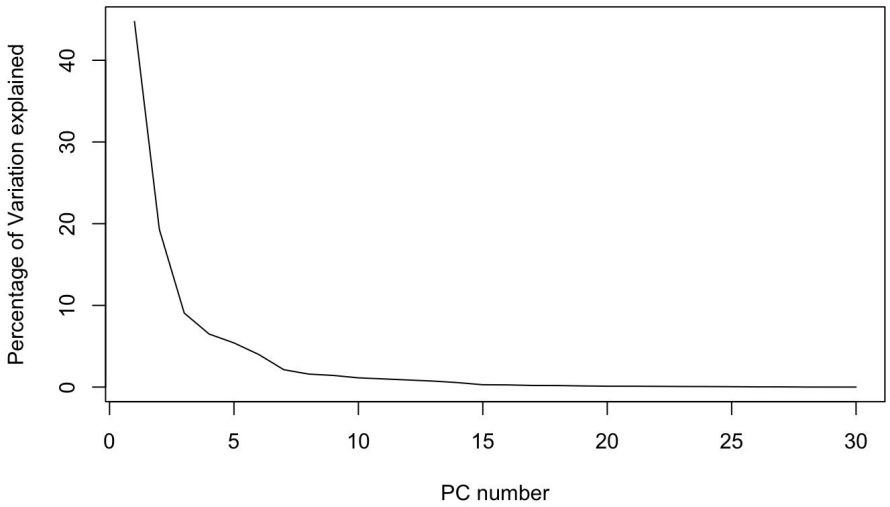
Logistic GLM Model

- Before we do any of PCA, Ridge, Lasso, we try a logistic GLM model
  - 80:20 split - used same data for all models
  - Removed predictors with 90%+ collinearity
  - Diagnosis ~ remaining predictors

- Accuracy: 94.7%
  - Benchmark model

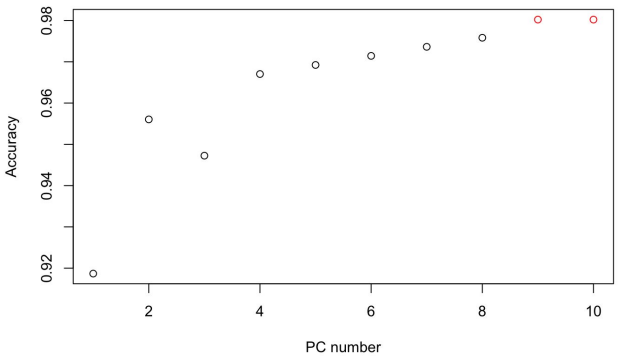
Principal Component Analysis

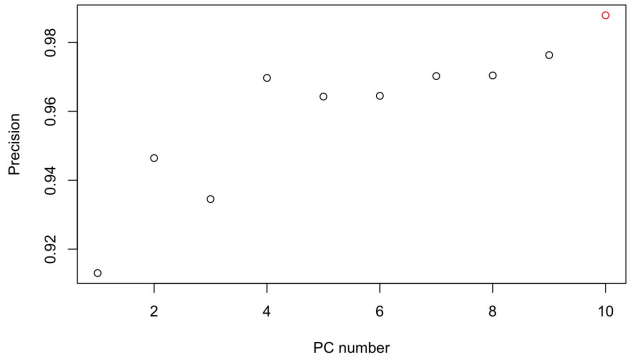
- Our dataset had strong hints of multicollinearity so we attempted to use PCA first.
- PCA takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables.
- First 10 PCs contain majority of Variation. Specifically :-
-
    - PC1- 44.8%
    - PC2 - 19.3%
    - PC3 - 9.1%
    - PC4 - 6.5%
    - PC5 - 5.4%
    - PC6 - 4.0%
    - PC7 - 2.1%
    - PC8 - 1.6%
    - PC9 - 1.4%
    - PC10 - 1.1%

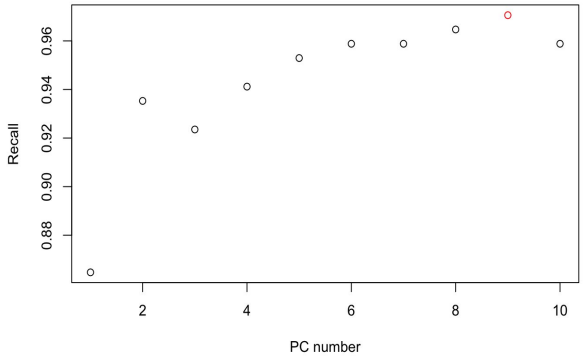


PC selection

- Graphs below demonstrate that using 9 PCs provides considerable increase in Accuracy, Precision and Recall.







Results on Test Set

Confusion Matrix for Predictions on Test Set (80-20 split)
Accuracy : **98.23%**

	B	M
B	69	2
M	0	42

Ridge and Lasso

- Good at dealing with multicollinearity

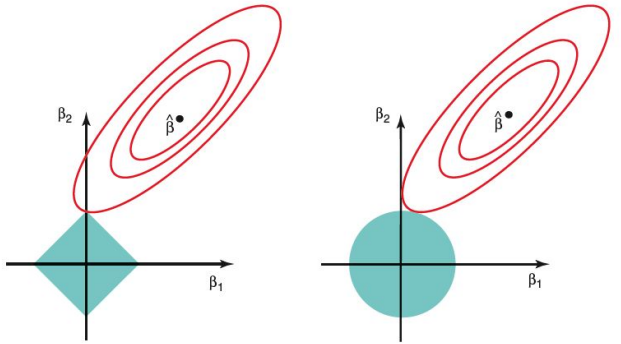
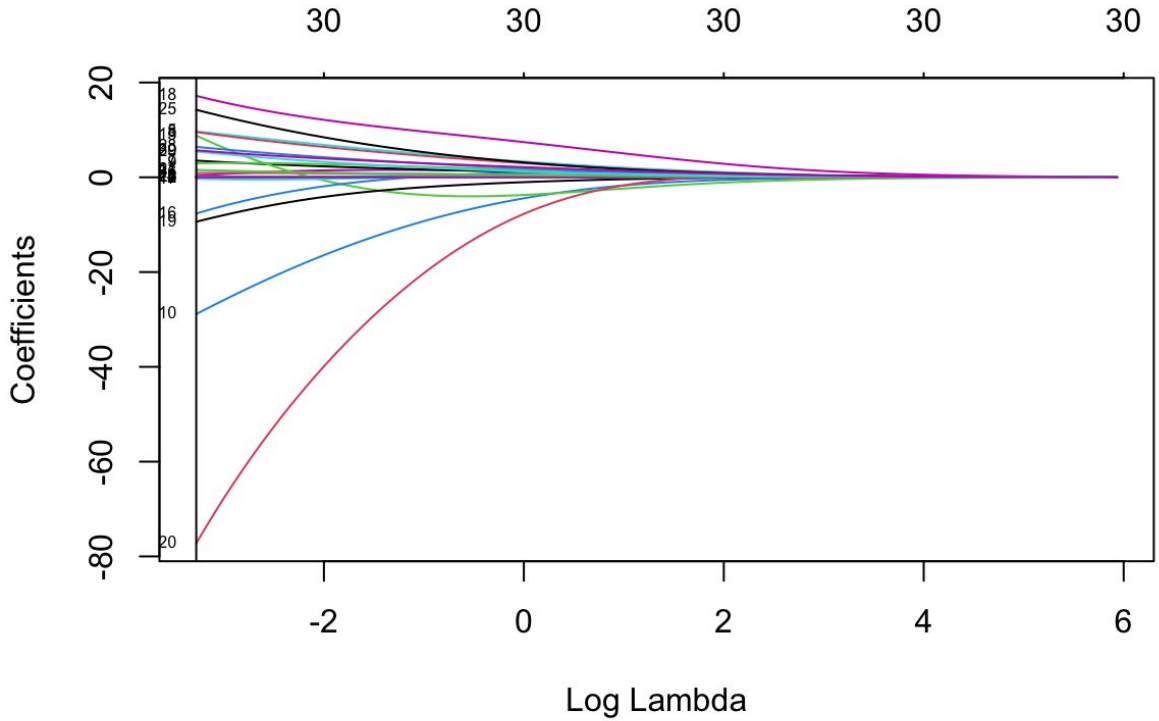
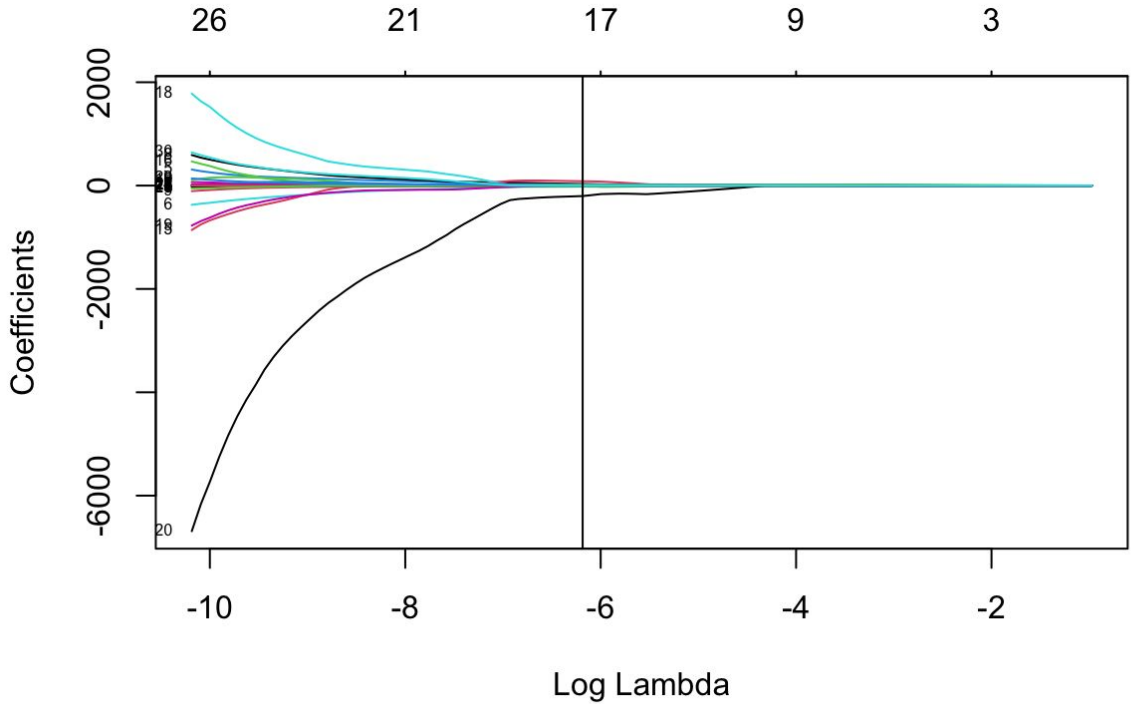


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Ridge Lambda Value



Lasso Lambda Value



Lasso Coefficients

perimeter mean	.
area mean	.
smoothness mean	.
compactness mean	-7.97
concavity mean	4.09
concave points mean	30.84
symmetry mean	.
fractal dimension mean	.
radius se	10.82
texture se	-1.05
perimeter se	.
area se	.
smoothness se	84.37
compactness se	-20.38
concavity se	.
concave points se	.

symmetry se	.
fractal dimension se	-202.33
radius worst	0.31
texture worst	0.28
perimeter worst	0
area worst	0
smoothness worst	22.87
compactness worst	.
concavity worst	3.82
concave points worst	19.5
symmetry worst	9.17
fractal dimension worst	.

Testing

- Same testing data as before

Ridge Accuracy = 100%

Lasso Accuracy = 100%

- More rigorous testing needed
  - Lasso or Ridge?
- 100% accuracy may have been due to randomness
- Not enough data to prove better than PCA

4 Fold Cross Validation

- Try to understand whether ridge or lasso is preferrable
- 4 “separate” tests
  - More chances to make mistakes
- We keep the same lambda values, they are already our “optimal” choice

Four Fold Cross Validation Results

Overall
Ridge: 97.2% accuracy
Lasso: 98.1% accuracy

- Same model, new training/testing splits
- Lambda values used for the models were the same as shown earlier
- The two models **appear** to have performed nearly identically

Fold Number	Ridge	Lasso
1	96.5% 5 incorrect	97.2% 4 incorrect
2	99.3% 1 incorrect	100% 0 incorrect
3	96.5% 5 incorrect	97.9% 3 incorrect
4	96.5% 5 incorrect	97.2% 4 incorrect

Confusion Matrix

Actual

Predicted	**Ridge**	Benign	Malignant
	Benign	354	14
	Malignant	2	198

Actual

Predicted	**Lasso**	Benign	Malignant
	Benign	352	7
	Malignant	4	205

Future Work

- Directly compare PCA with Lasso
  - Different packages make this a challenge
- Larger dataset
- Neural net
- Compare link function
- Minimize type 2 error (false negatives)
  - Don’t want to label a cancerous tumor benign
  - Even at the cost of lower accuracy