Chapter 6: Linear Model Selection and Regularization

Thiago G. Martins | NTNU & Verizon

Spring 2020

Recap

- So far, we have not made assumptions about f.
  - But from now on we assume f to be linear on the coefs.
- Assumptions:
  - (x_i^T, y_i) is independent from (x_j^T, y_j), $\forall i \neq j$.
  - The design matrix has full rank, $rank(\boldsymbol{X}) = p + 1, n >> (p + 1)$
  - $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$

Standard Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

or in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

- Least Square Fitting: Minimize the RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Standard Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

or in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Least Square Fitting: Minimize the RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

Recommended exercise 1

1. Show that the least square estimator of a standard linear model is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

2. Show that the maximum likelihood estimator is equal to the least square estimator for the standard linear model.

- Balance: average credit card debt

Recommended exercise 2

Write R code to create a similar representation of the Credit data figure of the previous slide. That is, try to recreate a similar plot in R.

Introduction

Objective of the module

Improve linear models **prediction accuracy** and/or **model interpretability** by replacing least square fitting with some alternative fitting procedures.

Prediction accuracy …

… when using standard linear models

Assuming true relationship is approx. linear: **low bias.**

- n >> p: **low variance**
- n not much larger than p: **high variance**
- n < p: multiple solutions available, **infinite variance**, model cannot be used.

Prediction accuracy …

… when using standard linear models

Assuming true relationship is approx. linear: **low bias**.

- n >> p: **low variance**
- n not much larger than p: **high variance**
- n < p: multiple solutions available, **infinite variance**, model cannot be used.

By constraining or shrinking the estimated coefficients:

- often substantially reduce the variance at the cost of a negligible increase in bias.
- better generalization for out of sample prediction

Model Interpretability

- Some or many of the variables might be irrelevant wrt the response variable
- Some of the discussed approaches lead to automatically performing feature/variable selection.

Outline

We will cover the following alternatives to using least squares to fit linear models

- **Subset Selection**: Identifying a subset of the p predictors that we believe to be related to the response.

Outline

We will cover the following alternatives to using least squares to fit linear models

- **Subset Selection**: Identifying a subset of the p predictors that we believe to be related to the response.
- **Shrinkage:** fitting a model involving all p predictors with the estimated coefficients shrunken towards zero relative to the least squares estimates.

Outline

We will cover the following alternatives to using least squares to fit linear models

- **Subset Selection**: Identifying a subset of the p predictors that we believe to be related to the response.
- **Shrinkage:** fitting a model involving all p predictors with the estimated coefficients shrunken towards zero relative to the least squares estimates.
- **Dimension Reduction**: This approach involves projecting the p predictors into a M-dimensional subspace, where $M < p$.

Subset Selection

Subset Selection

Identifying a subset of the p predictors that we believe to be related to the response.

Subset Selection

Identifying a subset of the p predictors that we believe to be related to the response.

Outline:

- Best subset selection
- Stepwise model selection

Best Subset Selection

1. Fit a least square regression for each possible combination of the p predictors.

2. Look at all the resulting models and pick the best.

Best Subset Selection

1. Fit a least square regression for each possible combination of the p predictors.

2. Look at all the resulting models and pick the best.

Number of models considered:

$$\binom{p}{1} + \binom{p}{2} + \ldots + \binom{p}{2} = 2^p$$

- Step 2 identifies the best model (on the training data) for each subset size
  - Reduces the problem from 2^p to $p + 1$ models to select from
- Step 3 choose among the $p + 1$ using the test error
  - Otherwise we would always choose the model with all parameters

- The red frontier tracks the best model for a given number of predictors, according to RSS and R^2.

Recommended exercise 3

1. For the Credit Dataset, pick the best model using Best Subset Selection according to C_p, BIC and Adjusted R^2
   - Hint: Use the `regsubsets()` of the `leaps` library, similar to what was done in Lab 1 of the book.
2. For the Credit Dataset, pick the best model using Best Subset Selection according to a 10-fold CV
   - Hint: Use the output obtained in the previous step and build your own CV function to pick the best model.
3. Compare the result obtained in Step 1 and Step 2.

Best Subset Selection (Drawbacks)

- Does not scale well -> the number of models to consider explode as p increases
  - $p = 10$ leads to approx. 1000 possibilities
  - $p = 20$ leads to over 1 million possibilities

Best Subset Selection (Drawbacks)

- Does not scale well -> the number of models to consider explode as p increases
    - $p = 10$ leads to approx. 1000 possibilities
    - $p = 20$ leads to over 1 million possibilities
- Large search space might lead to overfitting on training data

Stepwise selection

Add and/or remove one predictor at a time.

Stepwise selection

Add and/or remove one predictor at a time.

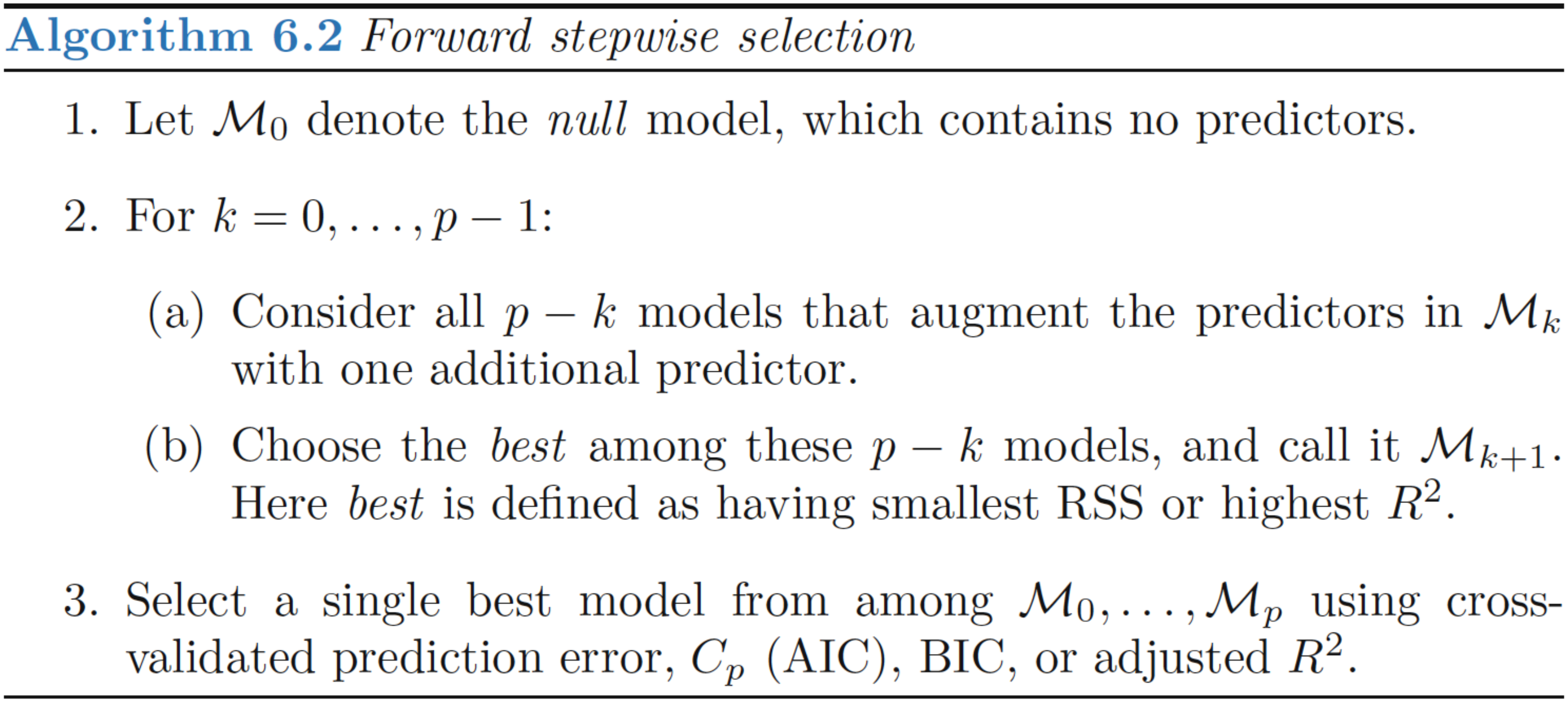
Methods outline:

- Forward Stepwise Selection
- Backward Stepwise Selection
- Hybrid approaches

Forward Stepwise Selection

- Starts with a model containing no predictors, $\mathcal{M}_0$
- Adds predictors to the model, one at time, until all of the predictors are in the model
  - $\mathcal{M}_1$, $\mathcal{M}_2$, …, $\mathcal{M}_p$
- Select the best model among $\mathcal{M}_0$, $\mathcal{M}_1$, …, $\mathcal{M}_p$

Forward Stepwise Selection (Algorithm)



Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.

2. For $k = 0, \ldots, p-1$:

   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.

   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Stepwise Selection (About the Algorithm)

- Goes from fitting 2^p models to $1 + \sum_{k=0}^{p-1}(p - k) = 1 + p(p + 1)/2$ models

Forward Stepwise Selection (About the Algorithm)

- Goes from fitting 2^p models to $1 + \sum_{k=0}^{p-1}(p - k) = 1 + p(p + 1)/2$ models
- It is a guided search, we don't choose $1 + p(p + 1)/2$ models to consider at random.

Forward Stepwise Selection (About the Algorithm)

- Goes from fitting 2^p models to $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models
- It is a guided search, we don't choose $1 + p(p+1)/2$ models to consider at random.
- Not guaranteed to yield the best model containing a subset of the p predictors.

Forward Stepwise Selection (About the Algorithm)

- Goes from fitting 2^p models to $1 + \sum_{k=0}^{p-1}(p - k) = 1 + p(p + 1)/2$ models
- It is a guided search, we don't choose $1 + p(p + 1)/2$ models to consider at random.
- Not guaranteed to yield the best model containing a subset of the p predictors.

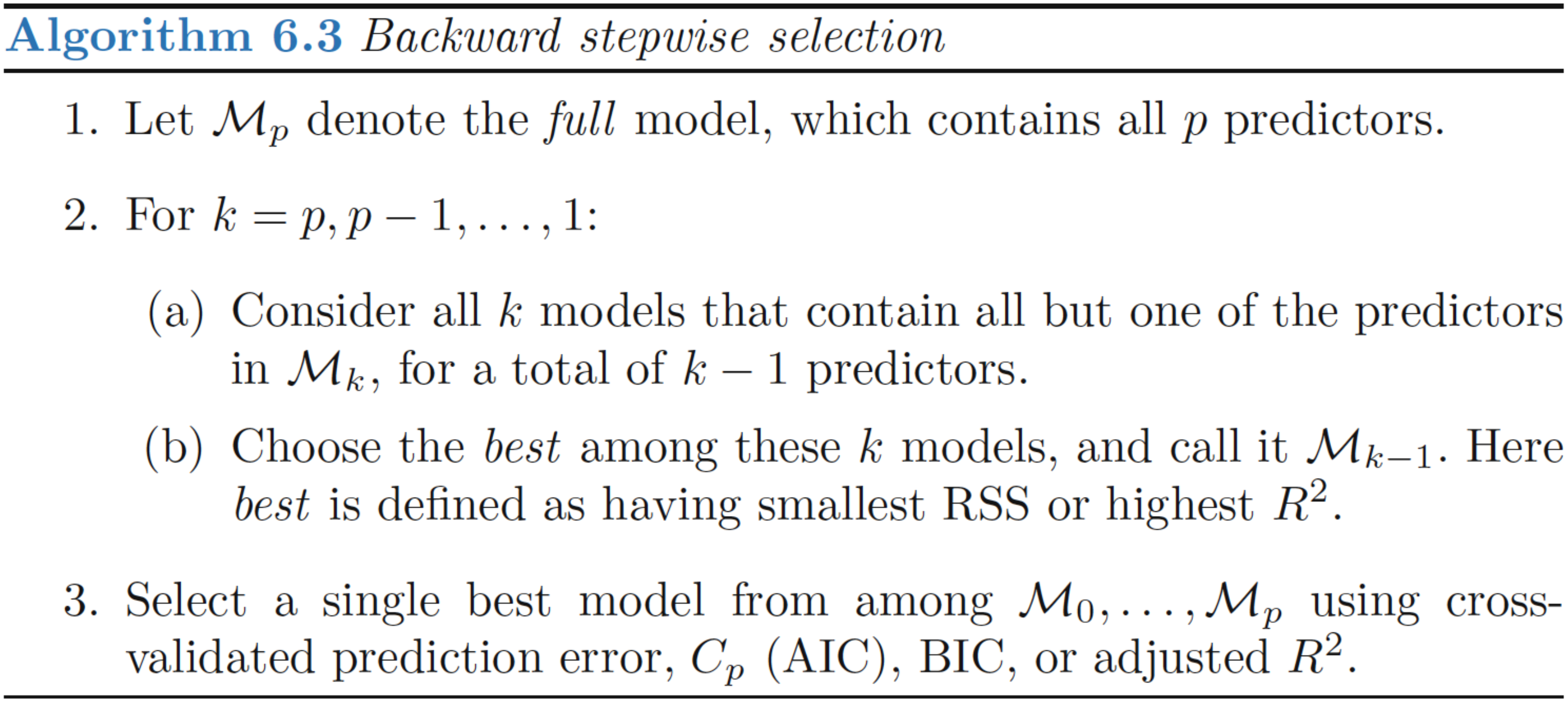
- Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$
    - By limiting the algorithm to submodels $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$ only

- The first three models are identical but the fourth models differ.

Backward Stepwise Selection

- Starts with a model containing all predictors, $\mathcal{M}_p$.
- Iteratively removes the least useful predictor, one-at-a-time, until all the predictors have been removed.
    - $\mathcal{M}_{p-1}$, $\mathcal{M}_{p-2}$, …, $\mathcal{M}_0$
- Select the best model among $\mathcal{M}_0$, $\mathcal{M}_1$, …, $\mathcal{M}_p$

Backward Stepwise Selection (Algorithm)



Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward Stepwise Selection (About the Algorithm)

- Similar properties to the Forward algorithm
  - Search $1 + p(p+1)/2$ models instead of 2^p models
  - It is a guided search, we don't choose $1 + p(p+1)/2$ models to consider at random.
  - Not guaranteed to yield the best model containing a subset of the p predictors.

Backward Stepwise Selection (About the Algorithm)

- Similar properties to the Forward algorithm
  - Search $1 + p(p+1)/2$ models instead of 2^p models
  - It is a guided search, we don’t choose $1 + p(p+1)/2$ models to consider at random.
  - Not guaranteed to yield the best model containing a subset of the p predictors.

- However, Backward selection requires that the number of samples n is larger than the number of variables p
  - So that the full model can be fit.

Hybrid Approach

- Similarly to forward selection, variables are added to the model sequentially.
- However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.
- Better model space exploration while retaining computational advantages of stepwise selection.

Recommended exercise 4

1. Select the best model for the Credit Data using Forward, Backward and Hybrid (sequential replacement) Stepwise Selection.
   - Hint: Use the `regsubsets()` of the `leaps` library
2. Compare with the results obtained with Best Subset Selection.

Shrinkage Methods

Shrinkage Methods

- fit a model containing all p predictors
    - using a technique that constrains (or regularizes) the coefficient estimates
    - or equivalently, that shrinks the coefficient estimates towards zero.

Shrinkage Methods

- fit a model containing all p predictors
  - using a technique that constrains (or regularizes) the coefficient estimates
  - or equivalently, that shrinks the coefficient estimates towards zero.
- Reduce the number of effective parameters
  - While retaining the ability to capture the most interesting aspects of the problem.

Shrinkage Methods

- fit a model containing all p predictors
    - using a technique that constrains (or regularizes) the coefficient estimates
    - or equivalently, that shrinks the coefficient estimates towards zero.
- Reduce the number of effective parameters
    - While retaining the ability to capture the most interesting aspects of the problem.
- The two best-known techniques for shrinking the regression coefficients towards zero are:
    - the ridge regression.
    - the lasso.

Ridge regression

The ridge regression coefs β^R are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

with $\lambda > 0$ being a tuning parameter.

- Intercept is the mean value of the response when the covariates are set to zero

- multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$.

Ridge regression

- Ridge regression are not scale-invariant
    - The standard least square are scale-invariant
    - β^R will not only depend on λ but also on the scaling of the jth predictor

Ridge regression

- Ridge regression are not scale-invariant
  - The standard least square are scale-invariant
  - β^R will not only depend on λ but also on the scaling of the jth predictor
  - Apply Ridge regression after standardizing the predictors

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

The standardized ridge regression coefficients are displayed for the Credit data set.

Ridge regression (Effectiveness)

- Why does it work?
  - As λ increase, the flexibility of the fit decreases.
  - Leading to a decrease variance but increased bias

Ridge regression (Effectiveness)

- Why does it work?
  - As λ increase, the flexibility of the fit decreases.
  - Leading to a decrease variance but increased bias
- MSE is a function of the variance and the squared bias
  - Need to find sweet spot (see next Fig.)

Ridge regression (Effectiveness)

- Why does it work?
  - As λ increase, the flexibility of the fit decreases.
  - Leading to a decrease variance but increased bias
- MSE is a function of the variance and the squared bias
  - Need to find sweet spot (see next Fig.)
- Therefore, ridge regression works best for the cases where
  - The relationship between covariates and response is close to linear (low bias)
  - And the least square estimates have high variance (high p in relation to n)

- Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set.
- The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

Ridge regression (Computationally efficient)

- The computations required to solve β_λ^R, simultaneously for all values of λ, are almost identical to those for fitting a model using least squares.
  - See (Friedman, Hastie, and Tibshirani 2010) and the references therein.

Ridge regression (Disadvantages)

- Unlike previous methods, ridge regression will include all p predictors in the final model.
    - The penalty λ will shrink all of the coefficients towards zero.
    - But it will not set any of them exactly to zero (unless $\lambda = \infty$).

Ridge regression (Disadvantages)

- Unlike previous methods, ridge regression will include all p predictors in the final model.
  - The penalty λ will shrink all of the coefficients towards zero.
  - But it will not set any of them exactly to zero (unless $\lambda = \infty$).
- This may not be a problem for prediction accuracy, but makes model interpretation hard for large p.

Recommended exercise 5

1. Apply Ridge regression to the Credit Dataset.
2. Compare the results with the standard linear regression.

Lasso

- The Lasso regression coefs β^L are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

with $\lambda > 0$ being a tuning parameter.

Lasso

- The Lasso regression coefs β^L are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

with $\lambda > 0$ being a tuning parameter.

- Lasso also shrinks the coefficients towards zero
- In addition, the l_1 penalty has the effect of forcing some of the coefficients to be exactly zero when λ is large enough

Lasso

- The Lasso regression coefs β^L are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

with $\lambda > 0$ being a tuning parameter.

- Lasso also shrinks the coefficients towards zero
- In addition, the l_1 penalty has the effect of forcing some of the coefficients to be exactly zero when λ is large enough
- A geometric explanation will be presented in a future slide.

The standardized lasso coefficients are displayed for the Credit data set.

- grey lines represent unrelated predictors
- Minimum CV error points to only the two real predictors have coefs != zero
- least-square estimate assign high value to one of the two predictors
    - Many unrelated predictors have non-zero values

Ridge and Lasso (Different formulations)

- Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

- Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

- The red ellipses are the contours of the RSS.
- The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$
- the explanation holds for $p > 2$, just harder to visualize

Comparison between Ridge and Lasso

- Neither is universally better than the other
- One expects lasso to perform better for cases where a relatively small number of predictors have coefs that are very small or zero

Comparison between Ridge and Lasso

- Neither is universally better than the other
- One expects lasso to perform better for cases where a relatively small number of predictors have coefs that are very small or zero
- One expects ridge to be better when the response is a function of many predictors, all with roughly equal size

Comparison between Ridge and Lasso

- Neither is universally better than the other
- One expects lasso to perform better for cases where a relatively small number of predictors have coefs that are very small or zero
- One expects ridge to be better when the response is a function of many predictors, all with roughly equal size
- Hard to know a priori, techniques such as CV required

Recommended exercise 6

1. Apply Lasso regression to the Credit Dataset.
2. Compare the results with the standard linear regression and the Ridge regression.

Left: Ridge regression is the posterior mode for β under a Gaussian prior. Right: The lasso is the posterior mode for β under a double-exponential prior.

Selecting λ

- Pick λ for which the cross-validation error is smallest.
- re-fit using all of the available observations and the selected value of λ.

References

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1). NIH Public Access: 1.