Statistical Methods for Bioinformatics

II-2: Variable Selection, Ridge & Lasso

Agenda

- Framing Variable Selection and High Dimensionality
- Variable Selection
  - Best Subset Selection
  - Forward Step-wise Selection
  - Backward Step-wise Selection
- Predictor Set Size Penalized Performance Measures
- Shrinkage/regularization methods
  - Ridge
  - Lasso
- Details and interpretations of the optimization functions

High dimensional datasets are common in Bioinformatics

- It is quite common to have many measured features
  - $10^3 - 10^5$ genomic markers (GWAS)
    - Agilent Human Genome CGH[1]+SNP Microarray:

    292,097 probes for CGH and 119,091 probes to identify SNPs
  - 525 (mycoplasma) - 28,354 (Norway Spruce) expression of all genes (RNA-seq)
  - $50 - 5 \cdot 10^3$ metabolite levels (Metabolomics)
  - physiological parameters

[1] Comparative Genome Hybridisation to identify copy number changes

High dimensional datasets are a challenge

- Here we will talk about the standard linear model:
  $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$
- In general: more observations are needed to accurately fit the parameters
- As the number of parameters (p) approaches the number of observations (n)
  - The Ordinary Least Squares fit may have have little bias, but will have large variability
  - When $p > n$ Ordinary Least Squares gives no unique solution.
- The interpretability is diminished

The perennial trade-off: bias vs variance

- One wants a model that captures the regularities in training data, but also generalizes well to unseen data.

$$E(MSE) = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

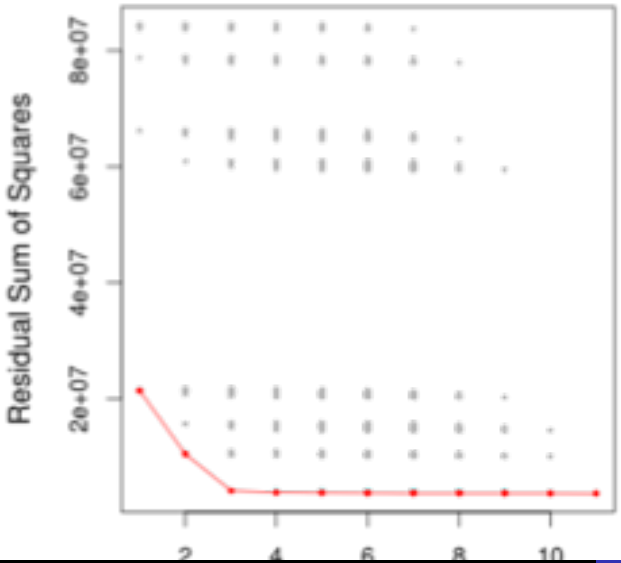
- With $Var(\hat{f}) = E((E(\hat{f}) - \hat{f})^2)$ & $Bias(\hat{f})^2 = E((f - E(\hat{f}))^2)$ and the irreducible error ϵ.
- With OLS there may be little bias, but as n approaches p variability increases.

Choose the Optimal Model

The Principle of Parsimony (Occam's Razor)

An explanation is better if it is simple.
e.g. models should be pared down until they are minimal and adequate

- Adding more predictors will always improve or match performance on a training set.
  - Estimate the test error using a validation set or a cross-validation approach.
  - How do we select the best set of features?



Linear Model Selection and Regularization

Subset Selection

- Identifying a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset
- Best subset selection, forward and backward step-wise selection

Shrinkage

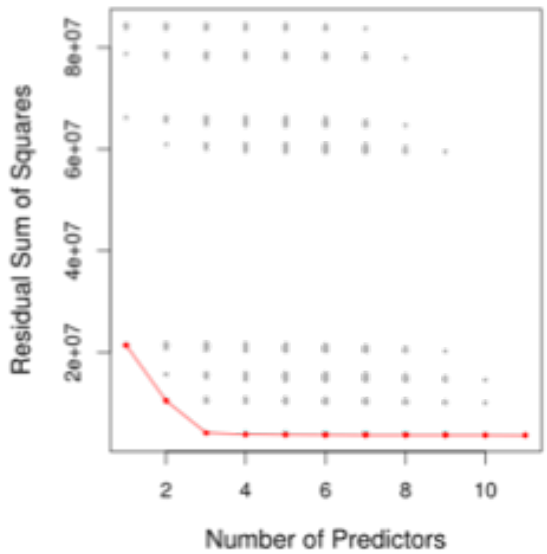
- Shrinking the estimates coefficients towards zero.
- This shrinkage reduces the variance. Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection
- E.g. Ridge regression and the Lasso

Dimension Reduction (next class)

- Involves projecting all p predictors into an M-dimensional space where $M < p$, and then fitting linear regression model
  - Principle Components Regression and Partial Least Squares

Subset selection algorithm

- In this approach, we run a model (e.g. a linear regression) for each possible combination of the X predictors
  - For k=1,2,...p:
    - Fit all $\begin{pmatrix} p \\ k \end{pmatrix}$ models that contain k predictors.
    - Pick the best among these, best is defined as having the best performance, e.g. smallest RSS, or largest R2.
  - Select a single best model across sizes taking into account the number of parameters: Cp, AIC, BIC, and adjusted R^2 ; Cross-Validated prediction error.



Refresher: definitions

Residual Sum of Squares (RSS)

$$RSS = \sum (y - \hat{y})^2$$

Total Sum of Squares (TSS)

$$TSS = \sum (y - \bar{y})^2$$

R^2 measure of fit / explained variance

$$1 - \frac{RSS}{TSS}$$

Subset Selection and Step-wise Selection

- The number of models to consider increases rapidly with the number of predictors (2^p)
  - The method is not computationally feasible for large numbers of predictors
  - The larger the search space, the likelier you have over-fitting on the training/higher variance.
- Alternatives: step-wise selection
  - Forward Step-wise Selection: Begins with the model containing no predictor, and then adds one predictor at a time that improves the model the most until no further improvement is possible
  - Backward Step-wise Selection: Begins with the model containing all predictors, and then deleting one predictor at a time that improves the model the most until no further improvement is possible

Forward Step-wise Selection

1. Start from an empty model
2. For $k = 0, \ldots, p-1$
   1. Consider all $p - k$ models that augment the predictor set with one
   2. Choose the best according to a performance measure (e.g. RSS)
3. Compare the models of different sizes using Cp, AIC, BIC, and adjusted R^2 ; Cross-Validated prediction error.

Considerations

1. Considerable less models fit ($1 + \frac{p}{2}(p+1)$ vs 2^p)
2. Works well in practice, but may be sub-optimal due to correlation and interaction between variables
   1. additions of a new variable may make already included variables "non-significant"
   2. an optimal pair or triple of parameters may be missed in the early phases by the progressive procedure

Backward Step-wise Selection

1. Start from an full model with all predictors
2. For $k = p, p-1, \ldots, 1$
   1. Consider all k models that remove one predictor from the predictor set
   2. Choose the best according to a performance measure (e.g. RSS)
3. Compare the models of different sizes using Cp, AIC, BIC, and adjusted R^2 ; Cross-Validated prediction error.

Considerations

1. Considerable less models fit than "Best Subset" ($1 + \frac{p}{2}(p+1)$ vs 2^p)
2. Works well in practice, avoids missing successful combinations between variables
   - Only possible when p<n (without further constraints and with Ordinary Least Squares)
   - No guarantee optimal solution

Predictor Set Size Penalized Performance Measures

The following measures for a linear model with Gaussian error!

- Adjusted R2

$$R_a^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- Bayesian information criterion (BIC)

$$BIC = -2 \ln \hat{L} + ln(n)d)$$

with L the maximized likelihood for the model. For a least squares *lm* fit

$$BIC = \frac{1}{n}(RSS + ln(n)d\hat{\sigma}^2)$$

d - # of predictors; n - # of observations; $\hat{\sigma^2}$ - estimate of error variance

Predictor Set Size Penalized Performance Measures

- Akaike information criterion (AIC) $AIC = 2d - 2\ln(L)$

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

- Mallow's Cp (proportional to AIC for linear regression)

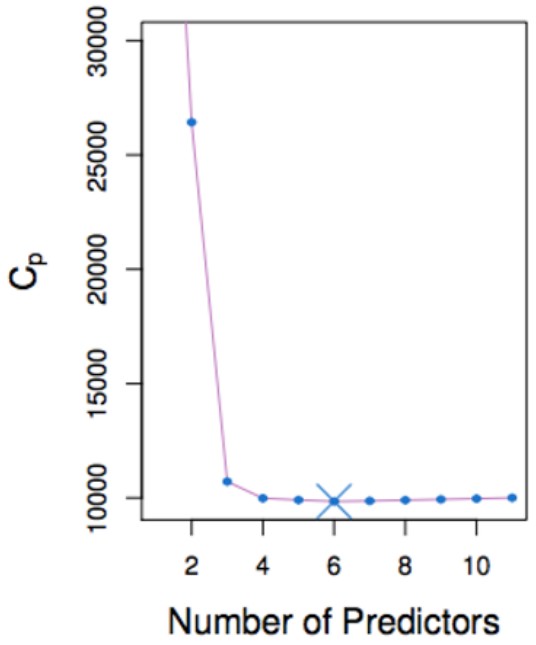
$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

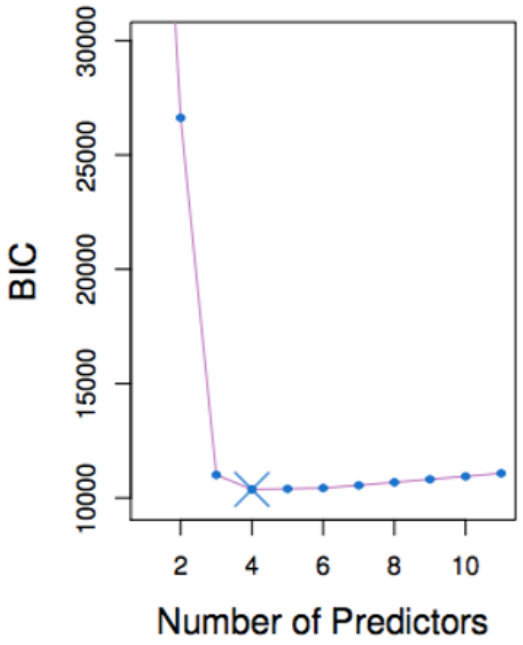
d - # of predictors; n - # of observations; $\hat{\sigma^2}$ - estimate of error variance

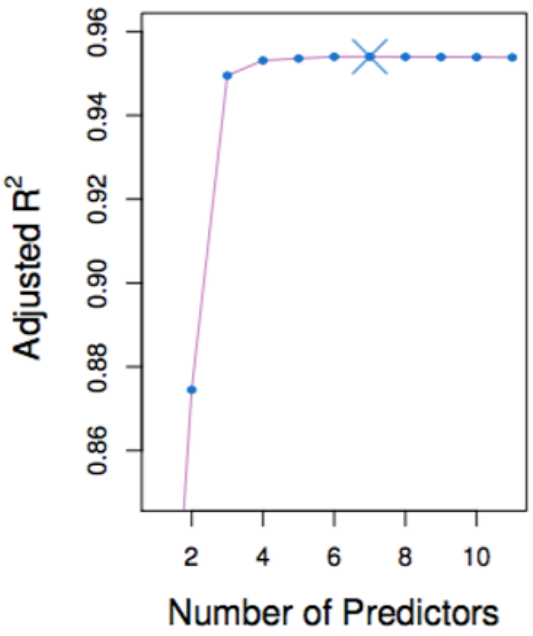
(Cross)-Validation

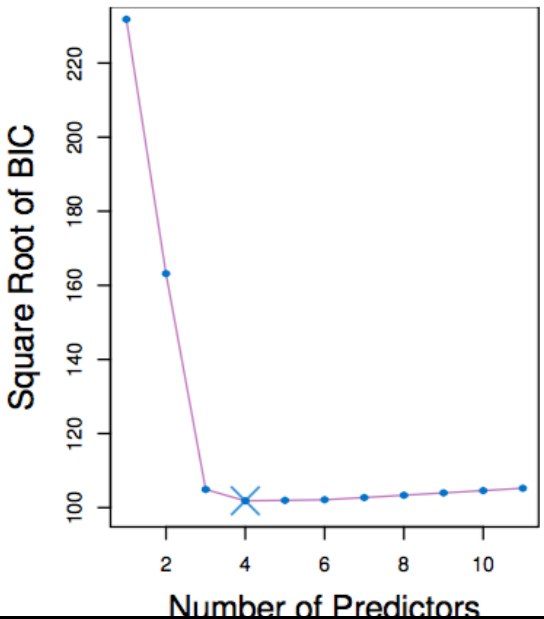
- More direct way of evaluating the effect of larger prediction models
- Validation: learn on one set, test on another normally smaller set
- Cross-Validation: similar to validation, the data set is split in training and test sets, but now repeatedly (typically 5-10 times) and the resulting performance measures are averaged.
- When comparing models for feature selection:
  - does not depend on knowing d, the fitted parameters, more difficult e.g. in regularization approaches (e.g. LASSO/Ridge, see later)
  - or $\hat{\sigma^2}$, which can be challenging to estimate, esp. with $n \sim p$

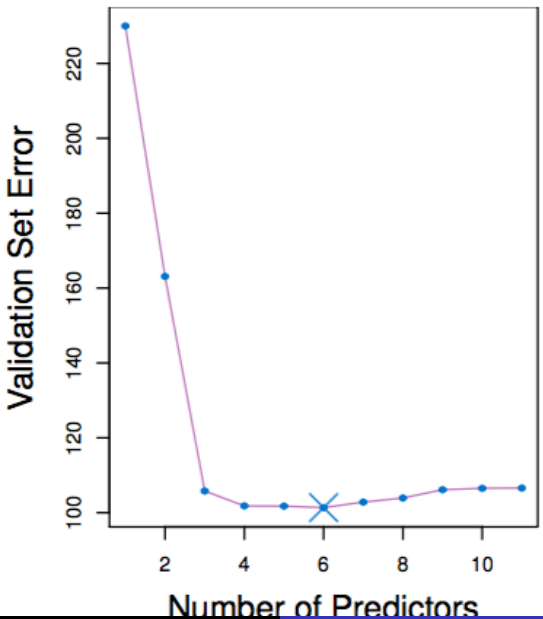
Credit data example (from the book)

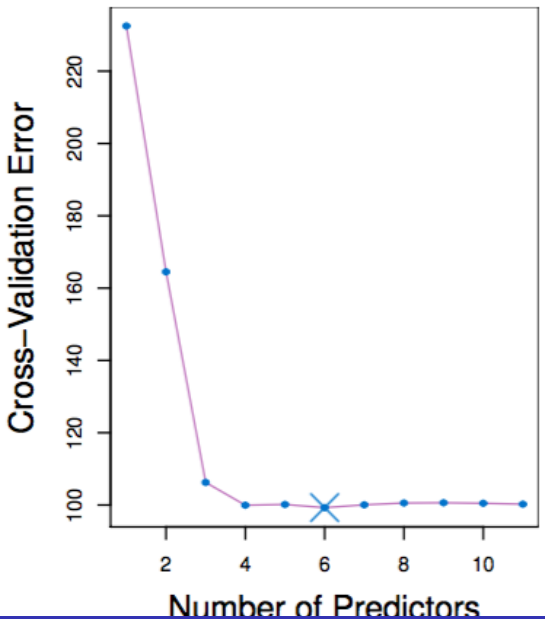












Choose the Optimal Model

- All approaches suggest that the 4-7-variable models are roughly equivalent in terms of their test errors.
- One way to select a model relies on the 1-Standard-Error Rule. Estimate the standard error of the test performance statistic for each model, and then select the -smallest- model with a score within one standard error of the lowest point on the curve. (Why?)

Shrinkage/regularization methods

- In high dimensional datasets, regression with least squares can face over-defined systems and inflated variance (as in trade-variance trade-off)
- Subset selection methods use least squares to fit a linear model with a subset of the predictors.
- An alternative approache is to fit a model containing all **p** predictors but constraining or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Shrinkage methods

- Ordinary least squares minimizes:

$$RSS = \sum_{i=1}^{n}(y_i\text{-}\beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

- Ridge regression minimizes:

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

- Lasso minimizes:

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

Shrinkage adds a penalty on coefficients

- The penalty acts against a large β vector

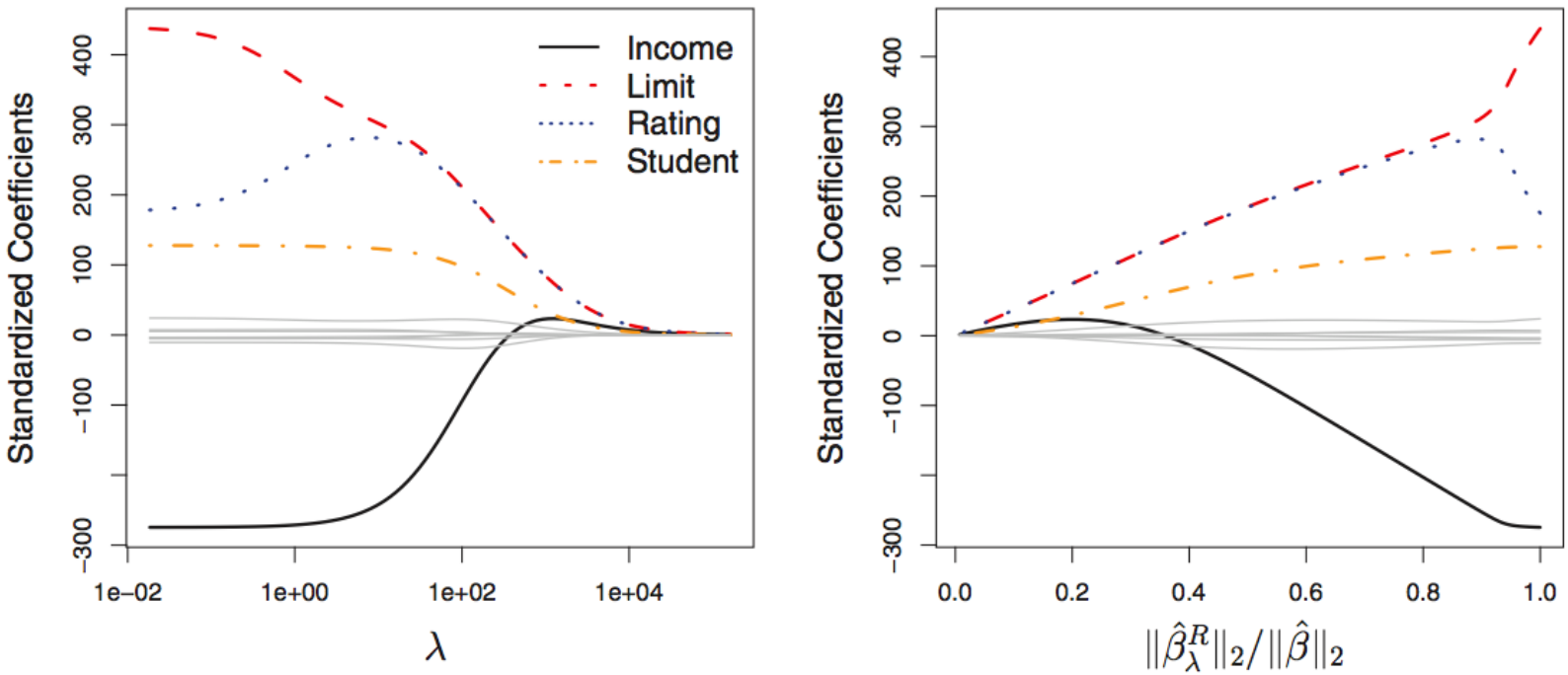


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Figure Explanation

- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we now display $\left\|\beta_\lambda^R\right\|_2 / \|\beta\|_2$ where β denotes the vector of least squares coefficient estimates, β_λ^R the reduced coefficients for a λ.
- The notation $\|\beta\|_2$ denotes the l_2 norm (pronounced "ell 2") of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

The strength of Ridge/Lasso vs Ordinary Least Squares

- Ideally the penalty in the fit reduces variance at the cost of a small increase in bias.
- In some cases variance can seriously hamper Least Squares fits. In the simulated dataset analysed below there are 45 predictors and 50 datapoints. Variance is large with n close to p. Least Squares matches the right extreme of the right panel.
- The shrinkage methods can work with n<p.

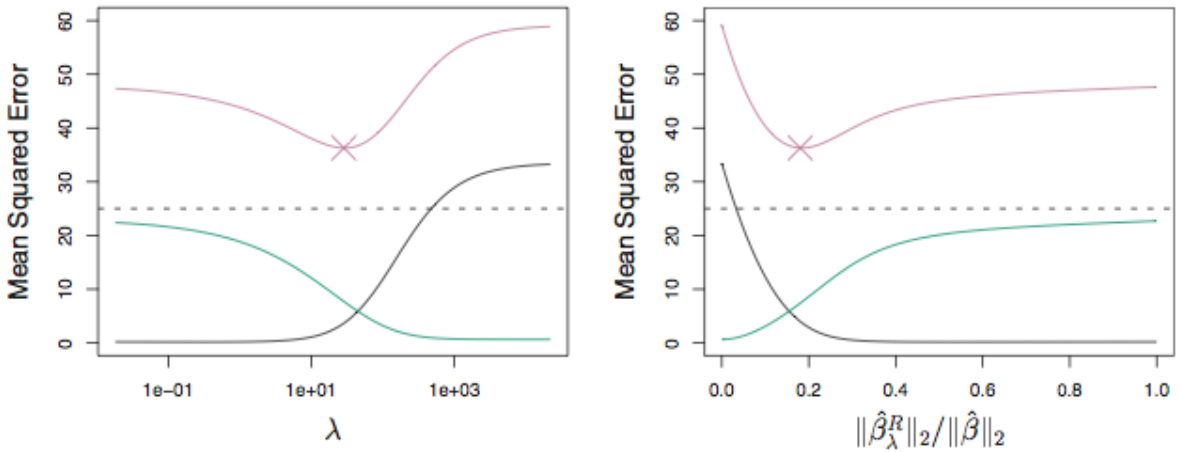


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

The strength of Ridge/Lasso vs Ordinary Least Squares

- In the simulated dataset analysed below there are 45 predictors and 50 datapoints, but only 2 predictors are associated with the response

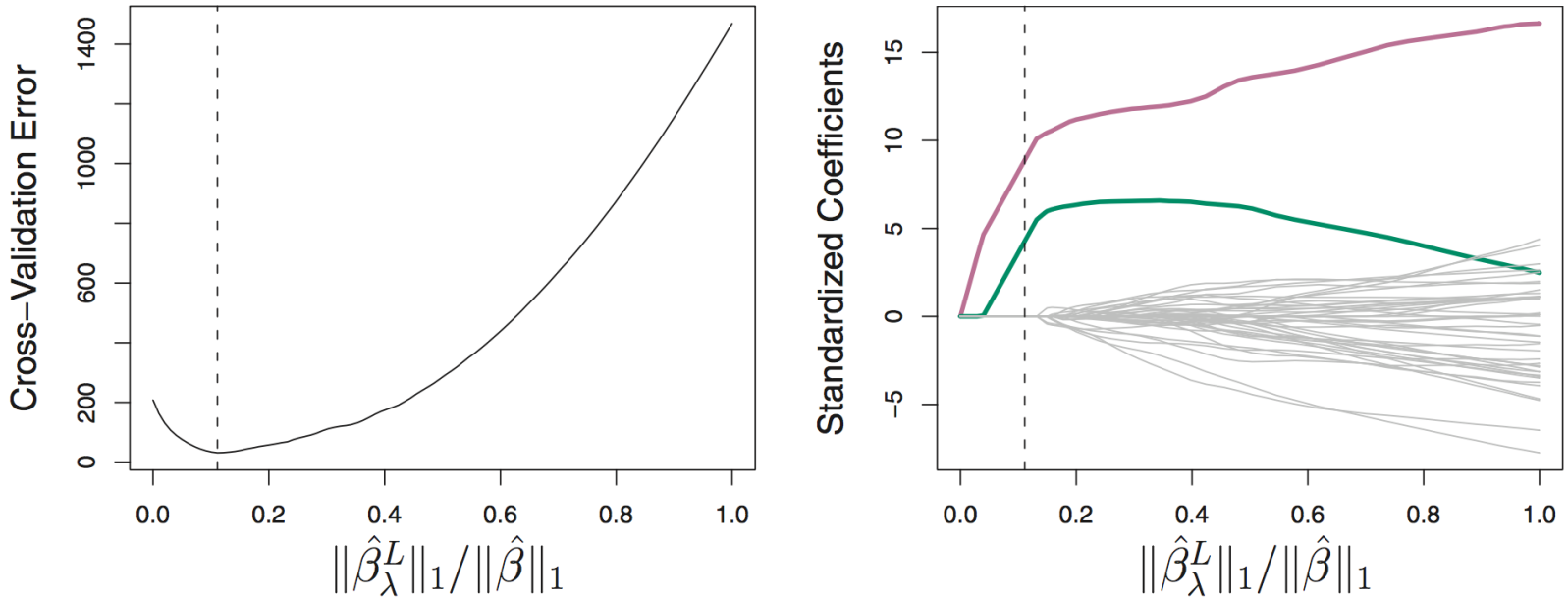


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Selecting the Tuning Parameter λ

- For optimal performance of ridge regression/lasso, finding the right λ is essential.
- Cross-validation
- Select a grid of potential values, use cross validation to estimate the error rate on test data (for each value of λ) and select the value that gives the least error rate, or use the 1-Standard-Error Rule.

Some nuts and bolts

- No penalty is imposed on the intercept.
- The penalty is formulated on the size of the coefficients, this implies that the scale of the variables is important! Therefore the variables should be standardized to make them comparable:

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

- If p is large, then using the best subset selection approach requires searching through enormous numbers of possible models
- With Ridge Regression/lasso, the computations required to solve the optimization can be very efficient. A fit, simultaneously for all values of λ, requires almost identical number of computations to those for fitting a model using least squares.

Lasso vs ridge regression

- Lasso uses a different penalty on the coefficients, which has the effect of tending to set some coefficients to zero.
- With Lasso, we can produce a model that has high predictive power and it simpler to interpret
- Automated feature selection (what could possibly go wrong).

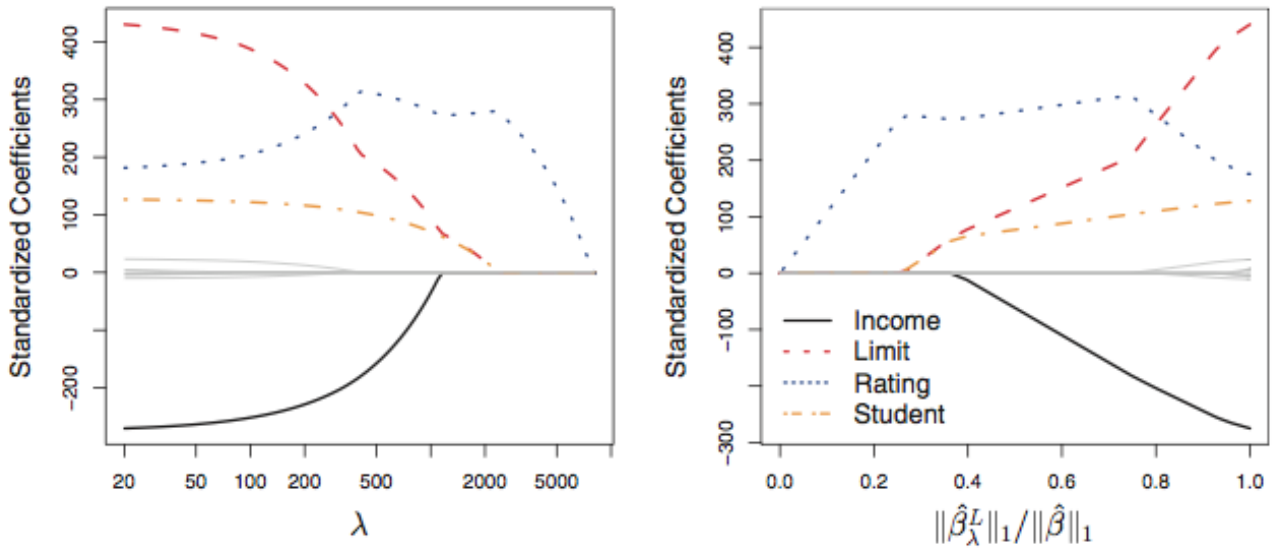


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Lasso vs ridge regression

We can rewrite the optimization problems as follows:

Definitions

Lasso:

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

Ridge:

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

In words: For every value of λ there is a value of s so that the outcome of the above defitions match our earlier definitions

Lasso vs ridge regression

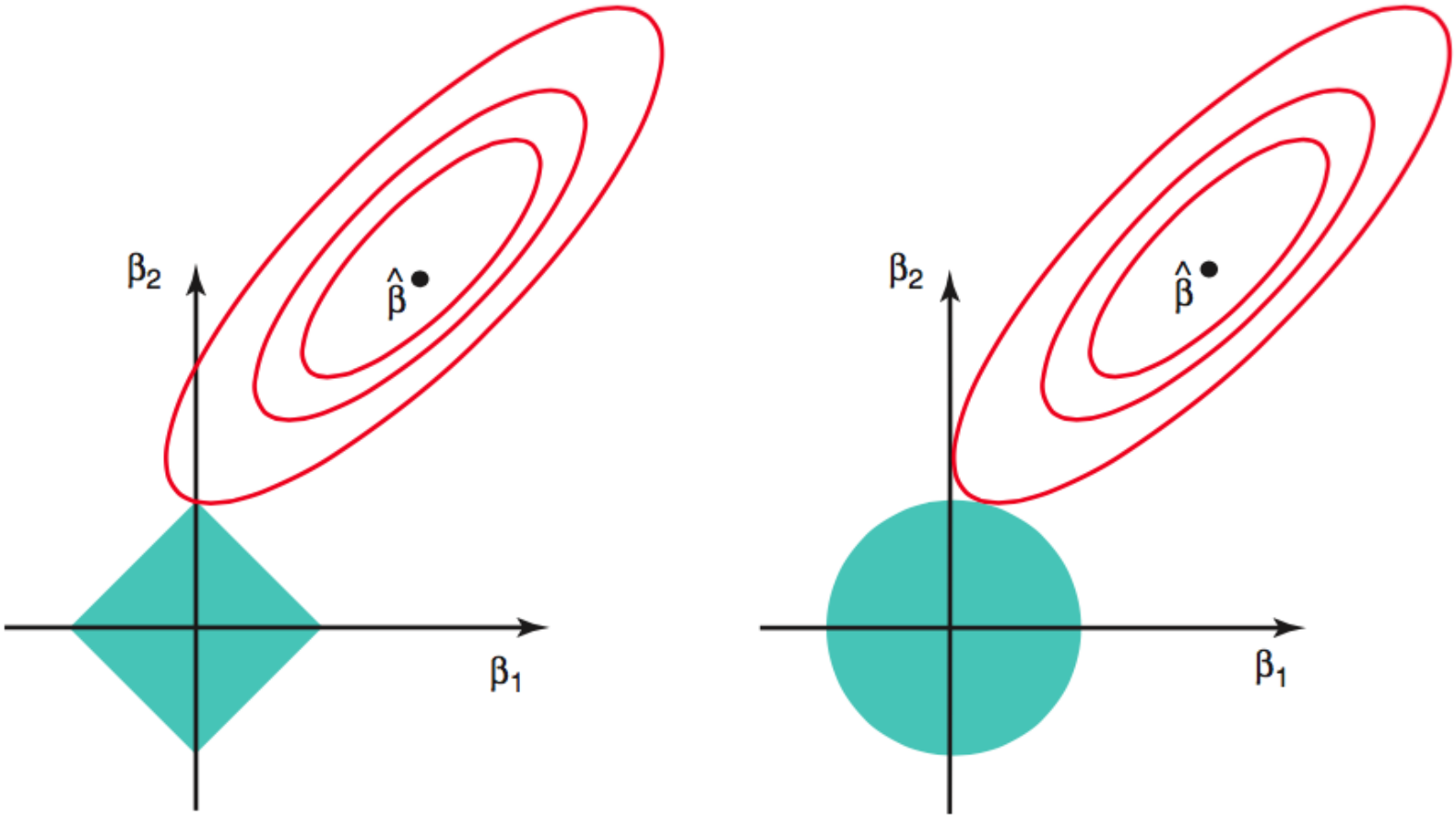
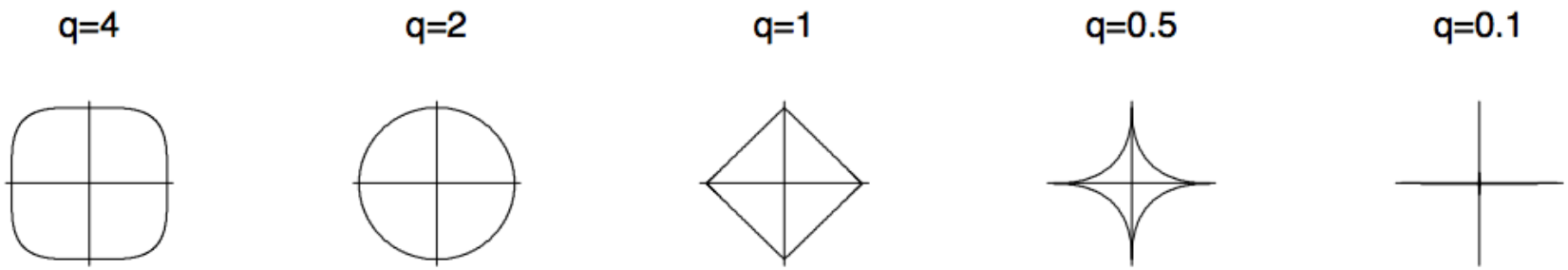


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Lasso vs ridge regression

1. Lasso produces simpler and more interpretable models
2. However predictive performance depends on the data. Are there (many) variables with no (independent) association to the response? If there are Lasso may work better, if there are not, Ridge may work better.

A generalization of the penalty



Contours of constant value of $\sum_j |\beta_j|^q$, with lasso at q=1 and ridge q=2. There are alternative forms, with as $q \to 0$ corresponding to subset selection.

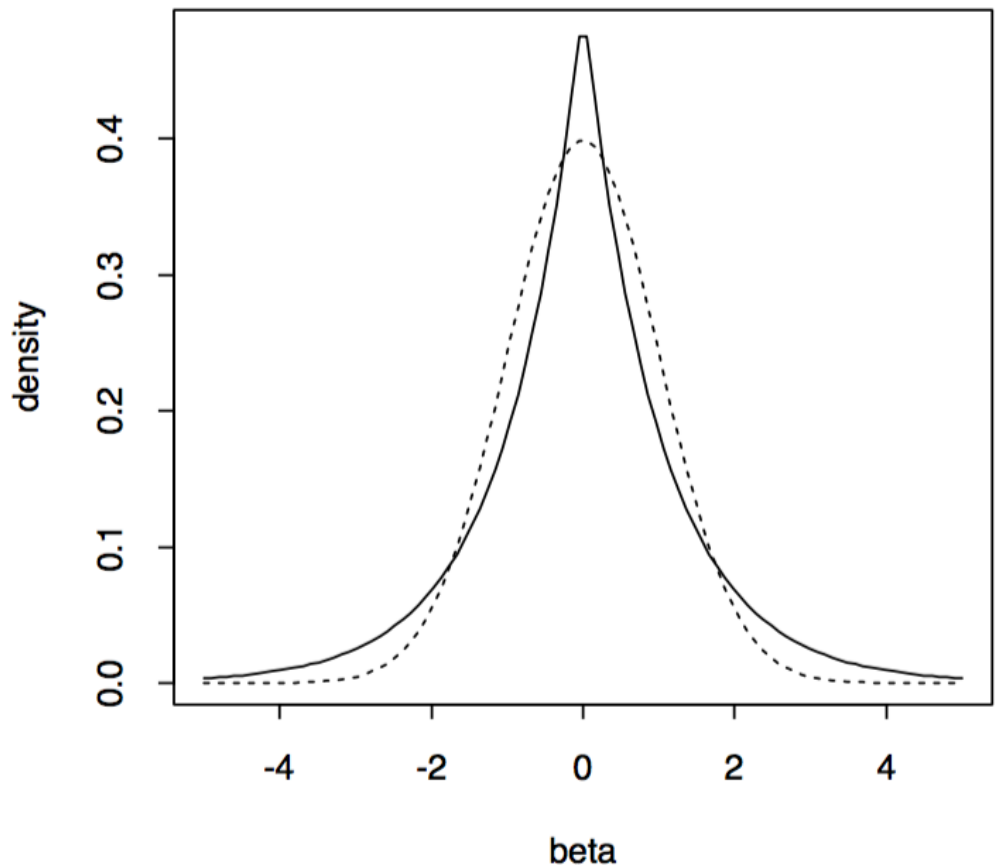
- There are also hybrid approaches: the Elastic Net combines the ridge and lasso penalties. Minimizes:

$$RSS + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

- And the penalized graphical lasso, used for estimating a sparse undirected graph.

Bayesian interpretation of Ridge Regression and the Lasso

One way of deriving the Ridge Regression and Lasso formulations is by assuming a probability distribution $\mathbf{g}$ that yields the coefficients β, $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$. The figure below shows the double exponential density (solid line), which is the implicit prior used by lasso. The dotted line shows a normal density, used by ridge regression.



Recap

- Variable Selection Approaches
- Best Subset Selection
  - Forward Step-wise Selection
  - Backward Step-wise Selection
- Measures to compare Models with different numbers of Variables
- Definition and use of Ridge and Lasso
  - How they compare to each other and other approaches
  - Interpretation of the optimization functions

To do:

Preparation for next class

- Read remainder chapter 6
- Send in any questions day before class

Exercises

- Exercise 6.8.1 and 6.8.2
- Do Lab 6.5 and 6.6
- Exercises 6.8.5, 6.8.8, 6.8.10