# Lecture 11: Penalized regression

## Statistical Learning (BST 263)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

(Figures from *An Introduction to Statistical Learning*, James et al., 2013)

# Outline

Penalized regression / Regularization

Subset selection

- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods

- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Outline

Penalized regression / Regularization

Subset selection

- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods

- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Least-squares

- We have seen the least-squares (maximum likelihood) approach to fitting linear models like

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^{\mathsf{T}} \beta + \varepsilon_i.$$

- Linear models have a number of advantages:
  - Interpretability
  - Simple and easy to use
  - Low flexibility helps control variance and improve performance

- But least-squares fitting has several disadvantages:
  - Variance can still be high, especially when $p$ is large
  - Collinear predictors cause difficulties
  - Cannot be used when $p > n$
  - Requires extra steps to do variable selection

# Penalized regression

- Penalized regression improves upon least-squares.
- Basic idea: Constrain or “shrink” parameter estimates.
- Intuition: Keep the parameter estimates from being too wild.

- Penalization is also known as *regularization*.

- Regularization reduces variance and increases bias.
- Test performance can be improved by regularizing an appropriate amount (due to bias-variance tradeoff).

# Penalized regression

- In least-squares, we minimize the objective function

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} (y_i - x_i^{\mathtt{T}}\beta)^2.$$

- Meanwhile, in *penalized regression*, we minimize

$$F(\beta) = \mathrm{RSS}(\beta) + \lambda\,\mathrm{Penalty}(\beta).$$

- $\lambda \geq 0$, and the penalty function can take various forms.
- More generally, in *penalized maximum likelihood*, we minimize

$$F(\beta) = -\ell(\beta) + \lambda\,\mathrm{Penalty}(\beta)$$

where $\ell(\beta)$ is the log-likelihood function.

# Examples of penalty functions

Common choices of penalty:

- Ridge: $\text{Penalty}(\beta) = \sum_{j=1}^{p} \beta_j^2$.
- Lasso: $\text{Penalty}(\beta) = \sum_{j=1}^{p} |\beta_j|$.
- Naive elastic net: $\text{Penalty}(\beta) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|$ .
- Best subset: $\text{Penalty}(\beta) = \sum_{j=1}^{p} \text{I}(\beta_j \neq 0)$.

Some choices of penalty enable variable selection (lasso, elastic net, best subset), but others do not (e.g., ridge).

# Duality

- Under some conditions, this is equivalent to minimizing $-\ell(\beta)$ subject to the constraint that

$$\mathrm{Penalty}(\beta) \leq s,$$

  for some $s$ that depends on $\lambda$.

- The details are beyond the scope of this course, but look up *Lagrangian duality* if you are curious to learn more.

# Outline

Penalized regression / Regularization

Subset selection
- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods
- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Best subset selection

- *Best subset selection* (for subsets of size $\leq k$) minimizes $\mathrm{RSS}(\beta)$ subject to the constraint that

  $$\sum_{j=1}^{p} \mathrm{I}(\beta_j \neq 0) \leq k,$$

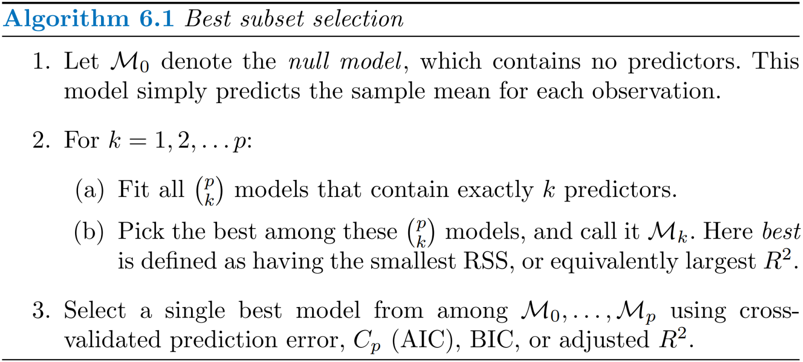
  that is, the $\#$ of nonzero coefficients is constrained to be $\leq k$.

- How to implement? Find least-squares fit $\hat{\beta}$ on all $\binom{p}{k}$ subsets of size $k$, and pick the subset with the smallest $\mathrm{RSS}(\hat{\beta})$.

- The choice of $k$ is made using cross-validation or some other model selection criterion.

- By duality, best subset selection is equivalent to minimizing $\mathrm{RSS}(\beta) + \lambda \sum_{j=1}^{p} \mathrm{I}(\beta_j \neq 0)$ for some choice of $\lambda$.

# Best subset selection



**Algorithm 6.1** *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

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

   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest $R^2$.

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 $R^2$.

# Best subset selection

Credit example: Predict balance (credit card debt) from income, credit limit, credit rating, # of cards, student/non-student, etc.

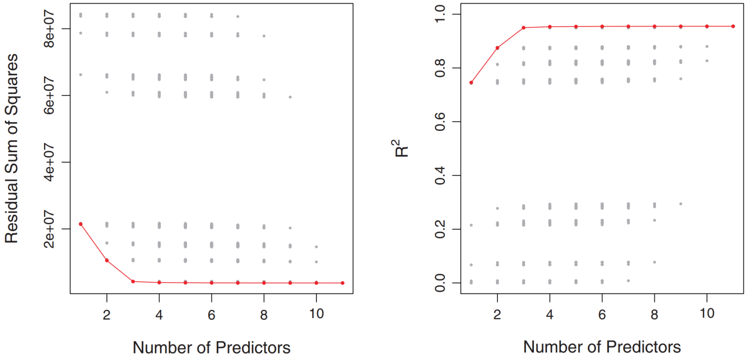


**FIGURE 6.1.** *For each possible model containing a subset of the ten predictors in the* Credit *data set, the RSS and $R^2$ are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and $R^2$. Though the data set contains only ten predictors, the $x$-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

# Forward/backward stepwise selection

- Best subset selection is computationally expensive as $p$ grows.
- Stepwise selection is a "greedy" approach to speed things up.

- *Forward stepwise selection*: Sequentially add the best predictor (i.e., greatest decrease in $\mathrm{RSS}(\hat{\beta})$ if added).

- *Backward stepwise selection*: Sequentially remove the worst predictor (i.e., least increase in $\mathrm{RSS}(\hat{\beta})$ if removed).

- Forward and backward stepwise require fitting $1 + p(p+1)/2$ models, compared to $2^p$ for best subset selection over all $k$.

- Advantages of forward stepwise (compared to backward):
  - Can speed up by stopping after $k$ predictors have been added.
  - Can handle $p > n$ with no problem — just stop early.

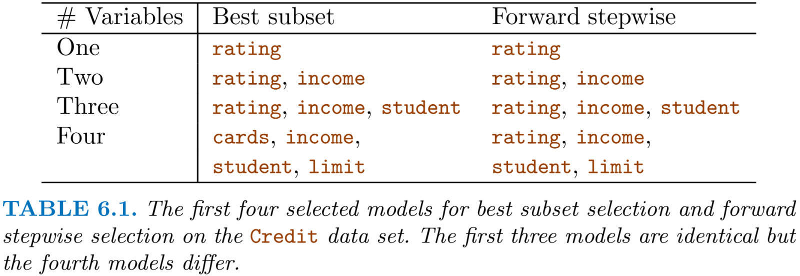
# Forward stepwise selection

**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 $R^2$.

# Comparison: Best subset vs Forward stepwise

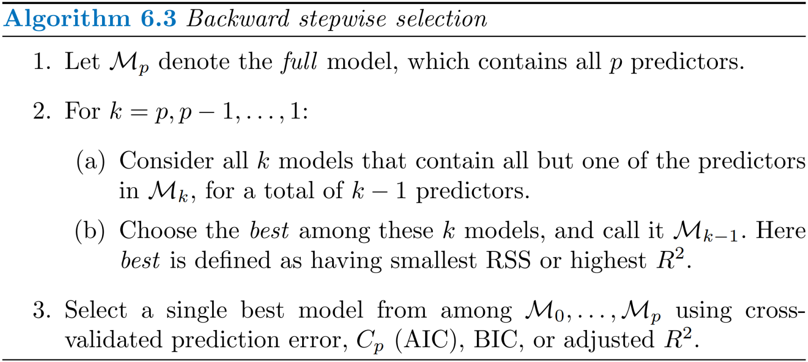
Credit example: Predict balance (credit card debt) from income, credit limit, credit rating, # of cards, student/non-student, etc.



| # Variables | Best subset | Forward stepwise |
|---|---|---|
| One | `rating` | `rating` |
| Two | `rating`, `income` | `rating`, `income` |
| Three | `rating`, `income`, `student` | `rating`, `income`, `student` |
| Four | `cards`, `income`,<br>`student`, `limit` | `rating`, `income`,<br>`student`, `limit` |

**TABLE 6.1.** *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*

# Backward stepwise selection



**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 $R^2$.

# Outline

Penalized regression / Regularization

Subset selection

- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods

- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Model selection

- Two general approaches to model selection:
  1. Directly estimate test performance (e.g., via cross-validation).
  2. Implicitly estimate test performance using an approximation (e.g., AIC, BIC).

- Approach 2 typically involves an adjustment to the training error (or log-likelihood) to account for model complexity.

- Pros/Cons: Approach 1 is more reliable. Approach 2 is faster.

- Bayesian inference is a third approach.
  - Similar to Approach 2.
  - BIC is an approximation to the Bayesian approach.

# Akaike information criterion (AIC)

- AIC is applicable to likelihood-based models.

- Goal: Choose among models $k = 1, \ldots, K$.
- Suppose model $k$ has $d_k$ parameters. Define

  $$\mathrm{AIC}_k = 2d_k - 2\ell_k$$

  where $\ell_k$ is the maximum log-likelihood for model $k$, i.e.,

  $$\ell_k = \log p(\text{data} \mid \hat{\theta}_k)$$

  where $\hat{\theta}_k$ is the MLE for model $k$.
- The AIC approach: Choose the model with the smallest $\mathrm{AIC}_k$.

- For least-squares regression, AIC is proportional to Mallow's $C_p$, so they are equivalent for least-squares model selection.

# Bayesian information criterion (BIC)

- BIC is similar to AIC, but with $\log(n)d_k$ instead of $2d_k$.
- Suppose model $k$ has $d_k$ parameters. Define:

$$\mathrm{BIC}_k = \log(n)d_k - 2\ell_k$$

  where $\ell_k$ is the maximum log-likelihood for model $k$.
- The BIC approach: Choose the model with the smallest $\mathrm{BIC}_k$.
- BIC is derived from an asymptotic approximation to the marginal likelihood $p(\text{data}|k)$ under a Bayesian model.

# Comparison: AIC versus BIC

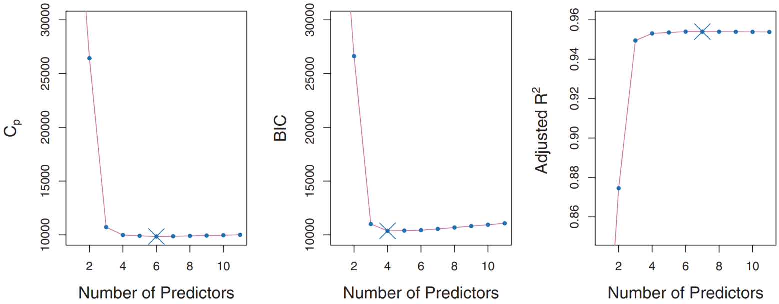


**FIGURE 6.2.** *$C_p$, BIC, and adjusted $R^2$ are shown for the best models of each size for the* Credit *data set (the lower frontier in Figure 6.1). $C_p$ and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

AIC is proportional to $C_p$ for least-squares regression.

# AIC versus BIC

What is the difference between AIC and BIC?

- BIC penalizes complexity more strongly than AIC, since $\log(n) > 2$ whenever $n > 7$.
- AIC tries to pick the model with the best expected test performance.
- BIC tries to pick the true model, assuming one of the models is true.
- AIC is asymptotically equivalent to LOO-CV in some cases.

# AIC/BIC versus cross-validation

- AIC and BIC are faster. Cross-validation is more reliable.
- Defining the “number of parameters” $d_k$ appropriately can be nontrivial.
- Also, when $d_k$ is large, the assumptions underlying AIC and BIC typically break down.
- My recommendation: AIC and BIC are often fine for a quick assessment, especially for low-dimensional models. But if you want something you can really trust, go with cross-validation.

# BIC versus cross-validation

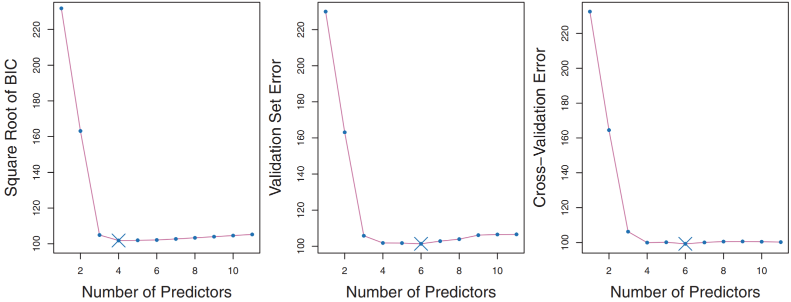


**FIGURE 6.3.** *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11*. The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

# Outline

Penalized regression / Regularization

Subset selection
- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods
- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Ridge regression

- *Ridge regression* minimizes

$$F(\beta) = \sum_{i=1}^{n} (y_i - x_i^{\mathtt{T}}\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- $\lambda \geq 0$ is a tuning parameter that acts as a “flexibility knob”.
- When $\lambda = 0$, ridge regression is the same as least-squares.
- As $\lambda$ increases, the coefficient estimates are pulled toward $0$. This is called *shrinkage*.

- $\lambda$ can be chosen using cross-validation.

- The name “ridge” comes from the original usage of the method, but it’s not really semantically meaningful anymore.

# Ridge regression path

The "path" is a plot of the coefficient estimates versus $\lambda$.

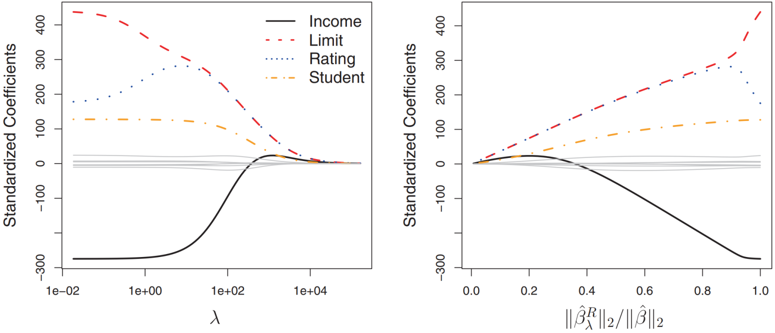


**FIGURE 6.4.** *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* $\lambda$ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

# Standardization

- In least-squares, it is not necessary to standardize the predictors or the outcomes before fitting the model.
- However, standardizing is important in penalized regression.
  - The reason is that the same penalty factor $\lambda$ is applied to all coefficients $\beta_j$ equally.
- Recommended to standardize to zero mean, unit variance:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n}\sum_{i'=1}^{n}(x_{i'j} - \bar{x}_j)^2}} \text{ and } \tilde{y}_i = \frac{y_i - \bar{y}}{\sqrt{\frac{1}{n}\sum_{i'=1}^{n}(y_{i'} - \bar{y})^2}}$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Usually, the intercept $\beta_0$ is not penalized, so the estimated intercept is $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ when using standardization.
- The interpretation of $\beta_j$ in the original units can be obtained by inverting these standardizations after estimation.

# Ridge versus least-squares

- Advantages of ridge versus least-squares:
  - Can improve test performance by reducing variance.
  - Collinearity handled more gracefully.
  - Can use when $p > n$.

- Disadvantages of ridge versus least-squares:
  - Need to choose $\lambda$, but that is not so bad.
  - Bias is increased, so $\lambda$ needs to be chosen well.

- Bias-variance tradeoff:
  - As $\lambda$ increases, bias increases and variance decreases.
  - Pay a little bit in increased bias for a big reduction in variance.

# Comparison: Ridge versus least-squares

Simulation example with $p = 45$ and $n = 50$. If $p$ is relatively large, ridge can help a lot.

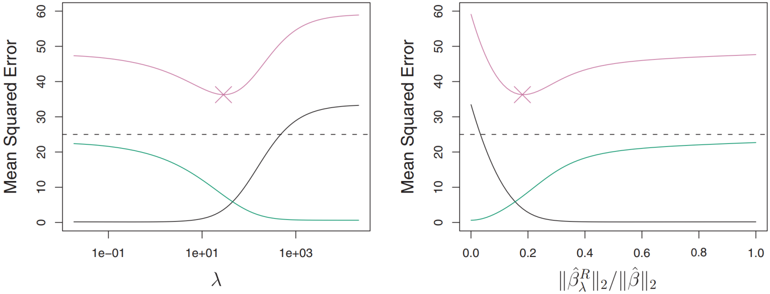


**FIGURE 6.5.** *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* $\lambda$ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$*. The horizontal dashed lines indicate the minimum possible MSE.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

# More on ridge regression

- Ridge regression can be solved in closed form:

$$\hat{\beta}^{\text{ridge}} = (A^{\mathtt{T}}A + \lambda I)^{-1}A^{\mathtt{T}}y$$

  where $A = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathtt{T}}$ is the design matrix and $y = (y_1, \ldots, y_n)^{\mathtt{T}}$, just as before in least-squares.

- Ridge regression is also called $\ell_2$ *regularization*, since the penalty function is the square of the $\ell_2$ norm of $\beta$,

$$\ell_2 \text{ norm} = \|\beta\|_2 = \Big( \sum_{j=1}^{p} |\beta_j|^2 \Big)^{1/2}.$$

- Ridge is also roughly equivalent to randomly perturbing the predictors by multiplying times i.i.d. $\mathcal{N}(1, s^2)$ random vars.

# The lasso (Tibshirani, 1996)

- lasso $=$ **l**east **a**bsolute **s**hrinkage and **s**election **o**perator
- The *lasso* minimizes

$$F(\beta) = \sum_{i=1}^{n} (y_i - x_i^{\mathtt{T}} \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- $\lambda \geq 0$ plays a similar role as in ridge regression.

- Convex optimization problem — fast solvers exist (glmnet).
- Lasso can be viewed as a "convex relaxation" of best subset selection.

- Lasso is also called $\ell_1$ *regularization*, since the penalty function is the $\ell_1$ norm of $\beta$,

$$\ell_1 \text{ norm} = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

# Lasso example

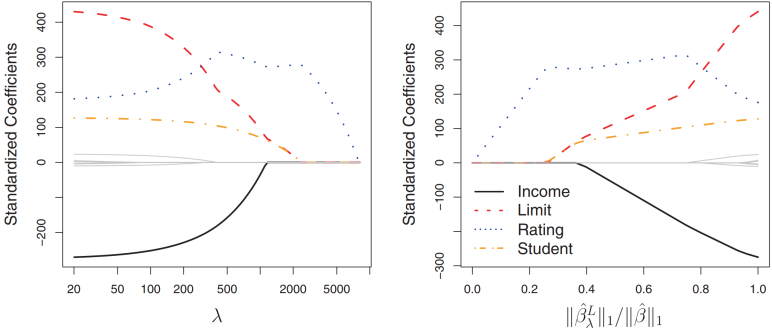


**FIGURE 6.6.** *The standardized lasso coefficients on the* Credit *data set are shown as a function of* $\lambda$ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

# Lasso versus ridge

- Lasso yields a "sparse" $\hat{\beta}$ (i.e., many zeros).
- Ridge yields a "dense" $\hat{\beta}$ (i.e., no zeros).

- Advantages of lasso versus ridge:
  - Lasso performs *variable selection*, which improves:
    - interpretability
    - computational efficiency of predictions
    - test performance, sometimes!
- Disadvantages of lasso versus ridge:
  - More complicated to solve (but can use packages like glmnet)

- Test performance can be better or worse, depending on the problem.
- Lasso does well if only a small number of predictors are needed.

# Why does lasso yield sparsity?

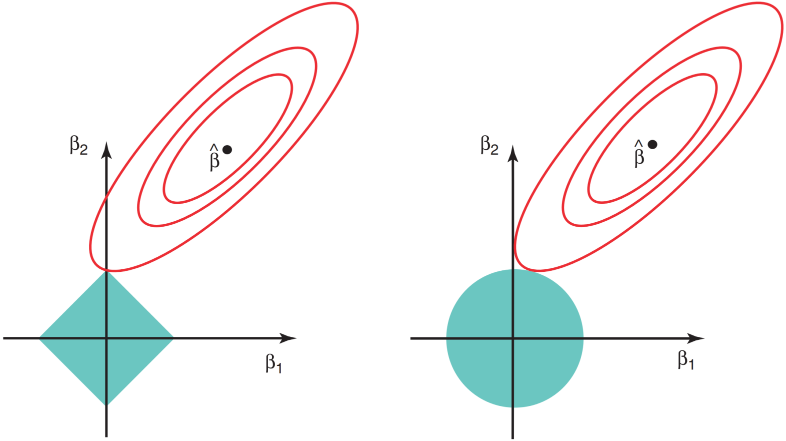


**FIGURE 6.7.** *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right)*. The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.*

# Test performance: Lasso versus ridge

If the true $\beta$ is dense, then ridge tends to be better:

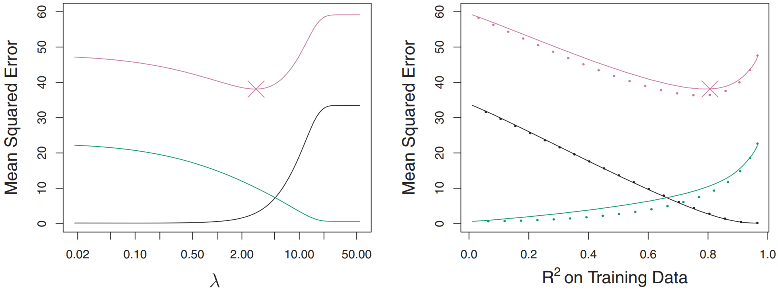


**FIGURE 6.8.** Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their $R^2$ on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

# Test performance: Lasso versus ridge

If the true $\beta$ is sparse, then lasso tends to be better:

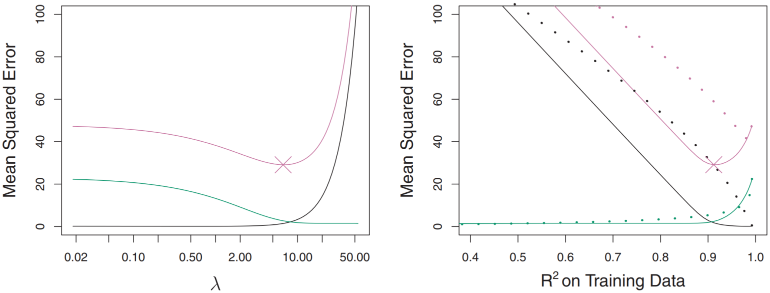


**FIGURE 6.9.** Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their $R^2$ on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

# Intuition: Least-squares vs ridge vs lasso

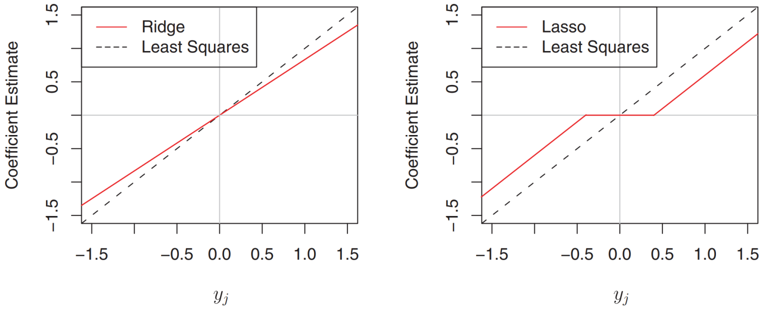


**FIGURE 6.10.** *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

# The elastic net (Zou and Hastie, 2005)

- The elastic net interpolates between ridge and lasso.

- *Naive elastic net* uses a penalty of the form

$$\text{Penalty}(\beta) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|.$$

- Elastic net is a slight modification of this.

- $0 \leq \alpha \leq 1$ controls the mix of ridge and lasso.
- $\lambda \geq 0$ plays a similar role as in ridge and lasso.

- Convex optimization problem — fast solvers exist (glmnet).
- Elastic net captures nice aspects of both ridge and lasso.

# Outline

Penalized regression / Regularization

Subset selection

- Best subset selection
- Stepwise selection

Model selection

Shrinkage methods

- Ridge regression
- Lasso
- Elastic net

Tuning parameter selection

# Tuning parameter selection

- Basic idea: Use cross-validation for each $\lambda$ in a grid.
- Choose the $\lambda$ with the smallest CV-estimated test error.
- Some packages (e.g., glmnet) will do this for you in a computationally efficient way.

# Tuning parameter selection

Example: Ridge regression on Credit data

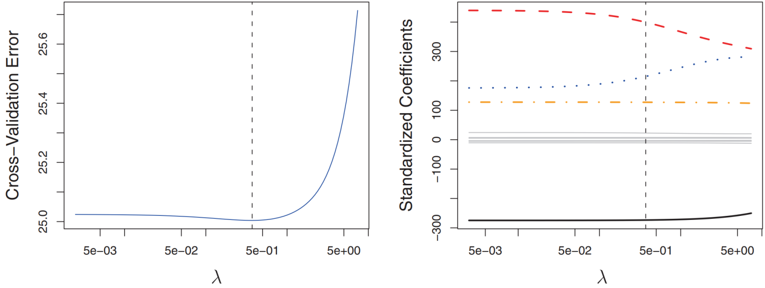


**FIGURE 6.12.** Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* $\lambda$. Right: *The coefficient estimates as a function of* $\lambda$. *The vertical dashed lines indicate the value of* $\lambda$ *selected by cross-validation.*

# Tuning parameter selection

Example: Lasso on simulated data with $p = 45$ and $n = 50$, where the true $\beta$ is sparse (2 nonzero coefficients)

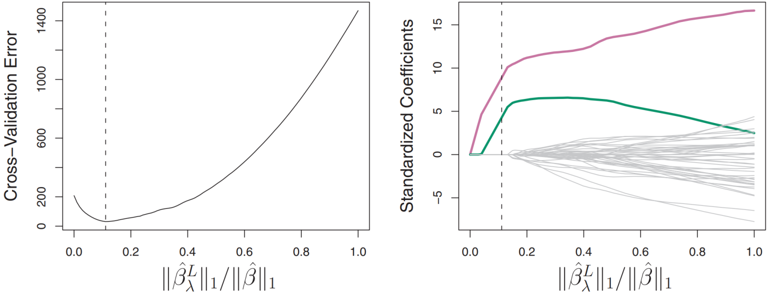


**FIGURE 6.13.** Left*: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right*: The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

# References

- James G, Witten D, Hastie T, and Tibshirani R (2013). *An Introduction to Statistical Learning*, Springer.
- Tibshirani R (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, Series B. Wiley. 58(1): 267-88.
- Zou H, and Hastie T (2005). *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society, Series B: 67:301-320.