Lecture 12: Shrinkage

Reading: Section 3.4

GU4241/GR5241 Statistical Machine Learning

Linxi Liu

March 2, 2018

Issues with Least Squares

Robustness

- Least squares works only if $\mathbf{X}$ has full column rank, i.e. if $\mathbf{X}^T\mathbf{X}$ is invertible.
- If $\mathbf{X}^T\mathbf{X}$ *almost* not invertible, least squares is numerically unstable.
  Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

- Modern problems: Many dimensions/features/predictors (possibly thousands)
- Only a few of these may be important
  $\rightarrow$ need some form of feature selection
- Least squares:
  - Treats all dimensions equally
  - Relevant dimensions are averaged with irrelevant ones
  - Consequence: Signal loss

Regularity of Matrices

Regularity

A matrix which is not invertible is also called a **singular** matrix. A matrix which is invertible (not singular) is called **regular**.

In computations

Numerically, matrices can be "almost singular". Intuition:

- A singular matrix maps an entire linear subspace into a single point.
- If a matrix maps points far away from each other to points very close to each other, it almost behaves like a singular matrix.

Regularity of Symmetric Matrices

A positive semi-definite matrix A is singluar $\Leftrightarrow$ smallest EValue is 0

Illustration

If smallest EValue $\lambda_{\text{min}} > 0$ but very small (say $\lambda_{\text{min}} \approx 10^{-10}$):

- Suppose x_1, x_2 are two points in subspace spanned by ξ_{min} with $\|x_1 - x_2\| \approx 1000$.
- Image under A: $\|Ax_1 - Ax_2\| \approx 10^{-7}$

In this case

- A has an inverse, but A behaves almost like a singular matrix
- The inverse A^{-1} can map almost identical points to points with large distance, i.e.

$$\text{small change in input} \quad \rightarrow \quad \text{large change in output}$$

Consequence for Statistics

If a statistical prediction involves the inverse of an almost-singular matrix, the predictions become unreliable (high variance).

Implications for Linear Regression

Recall: Prediction in linear regression

For a point $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$, we predict the corresponding function value as

$$\hat{y}_{\text{new}} = \left\langle \hat{\beta}, \mathbf{x}_{\text{new}} \right\rangle = \mathbf{x}_{\text{new}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Effect of unstable inversion

- Suppose we choose an arbitrary training point $\mathbf{x}_i$ and make a small change to its response value y_i.
- Intuitively, that should not have a big impact on $\hat{\beta}$ or on prediction.
- If $\mathbf{X}^T\mathbf{X}$ is almost singular, a small change to y_i can prompt a huge change in $\hat{\beta}$, and hence in the predicted value $\hat{y}_{\text{new}}$.

Measuring Regularity (of Symmetric Matrices)

Symmetric matrices

Denote by λ_{max} and λ_{min} the eigenvalues of A with largest/smallest *absolute* value. If A is symmetric, then

$$A \text{ regular} \quad \Leftrightarrow \quad |\lambda_{\text{min}}| > 0 \,.$$

Idea

- We can use $|\lambda_{\text{min}}|$ as a measure of regularity:

$$\text{larger value of } \lambda_{\text{min}} \quad \leftrightarrow \quad \text{"more regular" matrix } A$$

- We need a notion of scale to determine whether $|\lambda_{\text{min}}|$ is large.
- The relevant scale is how A scales a vector. Maximal scaling coefficient: λ_{max}.

Regularity measure

$$c(A) := \frac{|\lambda_{\text{max}}|}{|\lambda_{\text{min}}|}$$

The function $c(\,.\,)$ is called the **spectral condition** ("spectral" since the set of eigenvalues is also called the "spectrum").

Ridge Regression

Objective

Ridge regression is a modification of least squares. We try to make least squares more robust if $\mathbf{X}^T\mathbf{X}$ is almost singular.

Ridge regression solution

The ridge regression solution to a linear regression problem is defined as

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} := (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{y}$$

λ is a tuning parameter.

Explanation

Recall

$\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p\times p}$ is positive definite.

Spectral shift

Suppose $\xi_1, \ldots, \xi_p$ are EVectors of $\mathbf{X}^T\mathbf{X}$ with EValues $\lambda_1, \ldots, \lambda_p$.
Then:

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})\xi_i = (\mathbf{X}^T\mathbf{X})\xi_i + \lambda\mathbb{I}\xi_i = (\lambda_i + \lambda)\xi_i$$

Hence: $(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})$ is positive definite with EValues $\lambda_1 + \lambda, \ldots, \lambda_p + \lambda$.

Conclusion

$\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}$ is a *regularized* version of $\mathbf{X}^T\mathbf{X}$.

Implications for statistics

Effect of regularization

- We deliberately distort prediction:
  - If least squares ($\lambda = 0$) predicts perfectly, the ridge regression prediction has an error that increases with λ.
  - Hence: Biased estimator, bias increases with λ.
- Spectral shift regularizes matrix $\rightarrow$ decreases variance of predictions.

Bias-variance trade-off

- We decrease the variance (improve robustness) at the price of incurring a bias.
- λ controls the trade-off between bias and variance.

Cost Function

- Linear regression solution was defined as minimizer of $L(\boldsymbol{\beta}) := \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$
- We have so far defined ridge regression only directly in terms of the estimator $\hat{\boldsymbol{\beta}}^{\text{ridge}} := (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T\mathbf{y}$.
- To analyze the method, it is helpful to understand it as an optimization problem.
- We ask: Which function L' does $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ minimize?

Ridge regression

Ridge regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

In blue, we have the RSS of the model.

In red, we have the squared ℓ_2 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_2^2$. It is called a **penalty term**.

Ridge regression

Ridge regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

In blue, we have the RSS of the model.

In red, we have the squared ℓ_2 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_2^2$. It is called a **penalty term**.

The parameter λ is a tuning parameter. It modulates the importance of fit vs. shrinkage.

Ridge regression

Ridge regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

In blue, we have the RSS of the model.

In red, we have the squared ℓ_2 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_2^2$. It is called a **penalty term**.

The parameter λ is a tuning parameter. It modulates the importance of fit vs. shrinkage.

We find an estimate $\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}$ for many values of λ and then choose it by cross-validation.

Ridge regression

Ridge regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

In blue, we have the RSS of the model.

In red, we have the squared ℓ_2 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_2^2$. It is called a **penalty term**.

The parameter λ is a tuning parameter. It modulates the importance of fit vs. shrinkage.

We find an estimate $\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}$ for many values of λ and then choose it by cross-validation. Fortunately, this is no more expensive than running a least-squares regression.

Ridge regression

In least-squares linear regression, scaling the variables has no effect on the fit of the model:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c, ie. after doing this we have the same RSS.

Ridge regression

In least-squares linear regression, scaling the variables has no effect on the fit of the model:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c, ie. after doing this we have the same RSS.

In ridge regression, this is not true.

Ridge regression

In least-squares linear regression, scaling the variables has no effect on the fit of the model:

$$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c, ie. after doing this we have the same RSS.

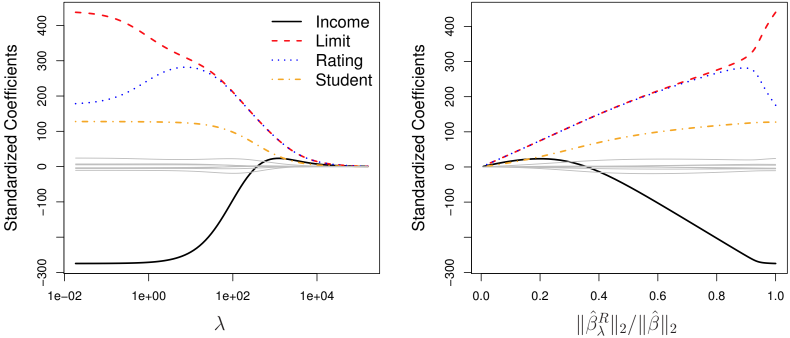
In ridge regression, this is not true.

In practice, what do we do?

- Scale each variable such that it has sample variance 1 before running the regression.
- This prevents penalizing some coefficients more than others.

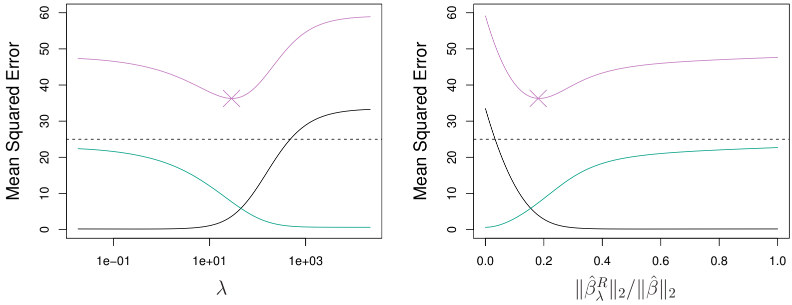
Example. Ridge regression

Ridge regression of `default` in the Credit dataset.



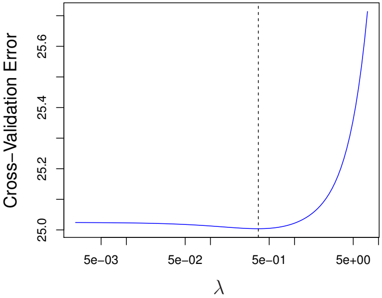
Bias-variance tradeoff

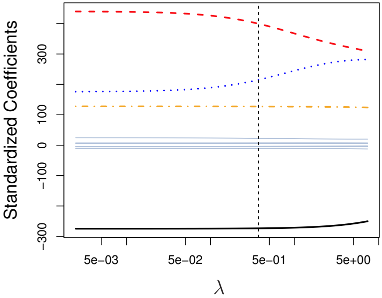
In a simulation study, we compute bias, variance, and test error as a function of λ.



Cross validation would yield an estimate of the test error.

Selecting λ by cross-validation





The Lasso

Lasso regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In blue, we have the RSS of the model.

In red, we have the ℓ_1 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_1$.

The Lasso

Lasso regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

In blue, we have the RSS of the model.

In red, we have the ℓ_1 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_1$.

Why would we use the Lasso instead of Ridge regression?

The Lasso

Lasso regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In blue, we have the RSS of the model.

In red, we have the ℓ_1 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_1$.

Why would we use the Lasso instead of Ridge regression?

- Ridge regression shrinks all the coefficients to a non-zero value.

The Lasso

Lasso regression solves the following optimization:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In blue, we have the RSS of the model.

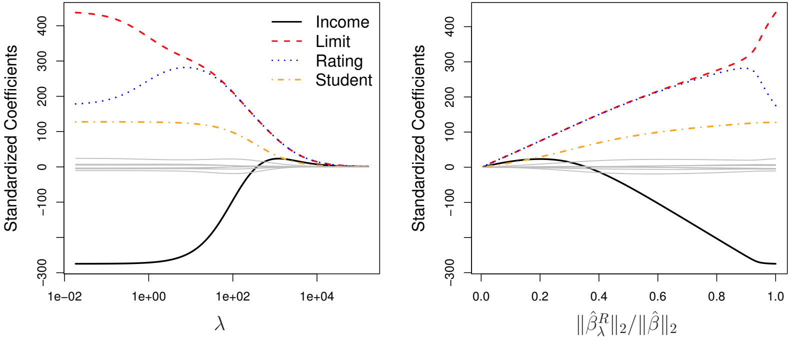
In red, we have the ℓ_1 norm of $\boldsymbol{\beta}$, or $\|\boldsymbol{\beta}\|_1$.

Why would we use the Lasso instead of Ridge regression?

- Ridge regression shrinks all the coefficients to a non-zero value.
- The Lasso shrinks some of the coefficients all the way to zero. Alternative to best subset selection or stepwise selection!

Example. Ridge regression

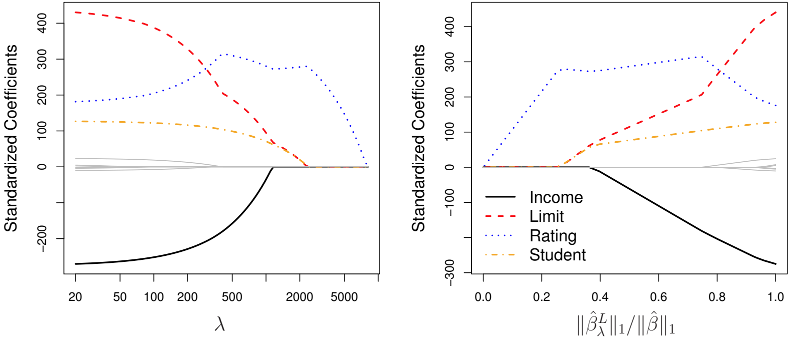
Ridge regression of `default` in the Credit dataset.



A lot of pesky small coefficients throughout the regularization path.

Example. The Lasso

Lasso regression of `default` in the Credit dataset.



Those coefficients are shrunk to zero.

An alternative formulation for regularization

- **Ridge:** for every λ, there is an s such that $\hat{\boldsymbol{\beta}}_\lambda^{\textbf{ridge}}$ solves:

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

An alternative formulation for regularization

- **Ridge:** for every λ, there is an s such that $\hat{\boldsymbol{\beta}}_\lambda^{\textsf{ridge}}$ solves:

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

- **Lasso:** for every λ, there is an s such that $\hat{\boldsymbol{\beta}}_\lambda^{\textsf{lasso}}$ solves:

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

An alternative formulation for regularization

- **Ridge:** for every λ, there is an s such that $\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}$ solves:

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

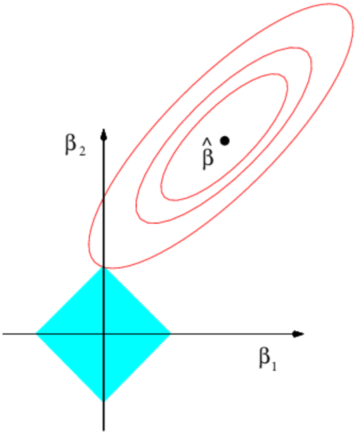
- **Lasso:** for every λ, there is an s such that $\hat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}$ solves:

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

- **Best subset:**

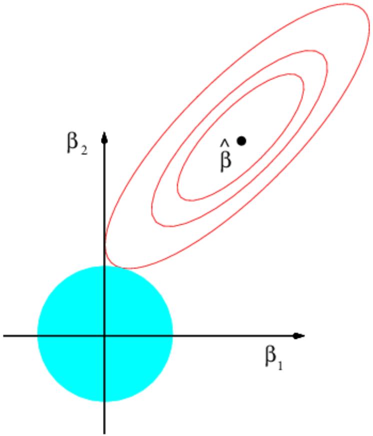
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right\} \quad \text{s.t.} \quad \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) < s.$$

Visualizing Ridge and the Lasso with 2 predictors



The Lasso

◆ : $\sum_{j=1}^{p} |\beta_j| < s$

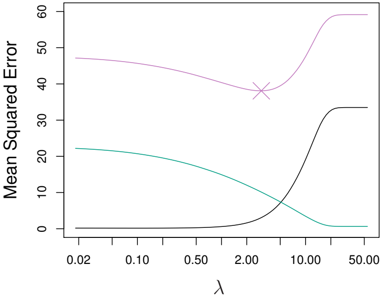


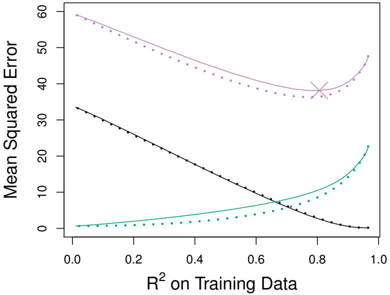
Ridge Regression

● : $\sum_{j=1}^{p} \beta_j^2 < s$

When is the Lasso better than Ridge?

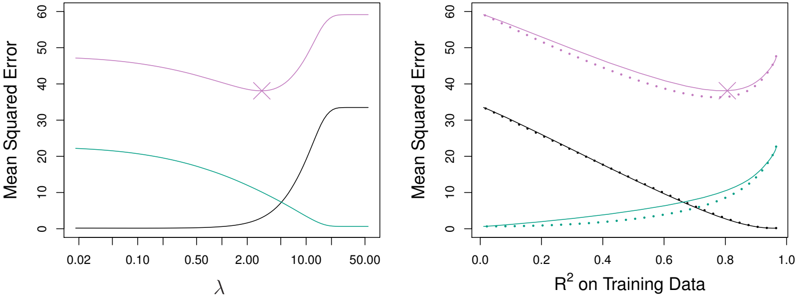
Example 1. Most of the coefficients are non-zero.





When is the Lasso better than Ridge?

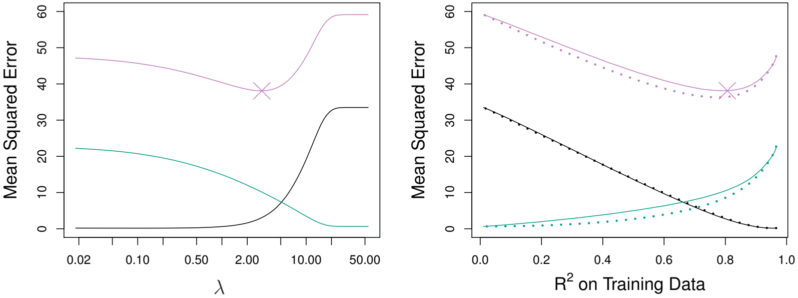
Example 1. Most of the coefficients are non-zero.



- Bias, Variance, MSE.

When is the Lasso better than Ridge?

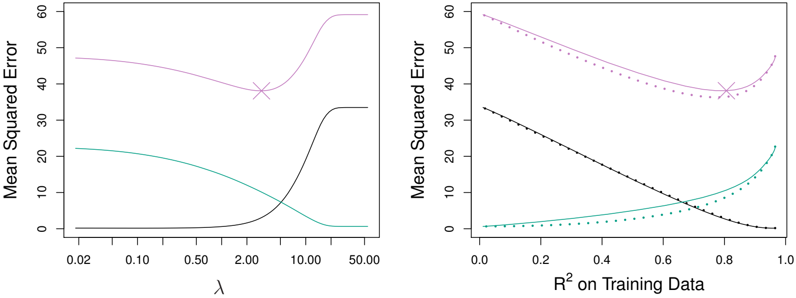
Example 1. Most of the coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge ($\cdots$).

When is the Lasso better than Ridge?

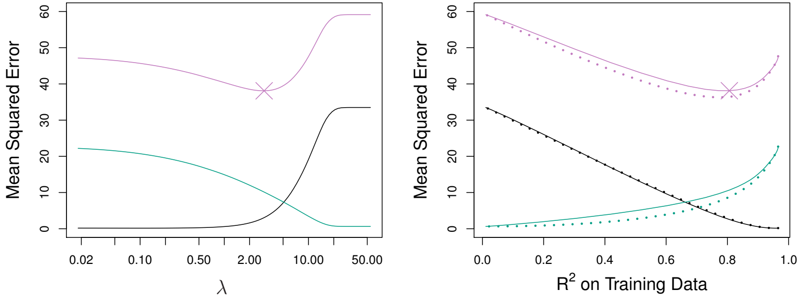
Example 1. Most of the coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge ($\cdots$).
- The bias is about the same for both methods.

When is the Lasso better than Ridge?

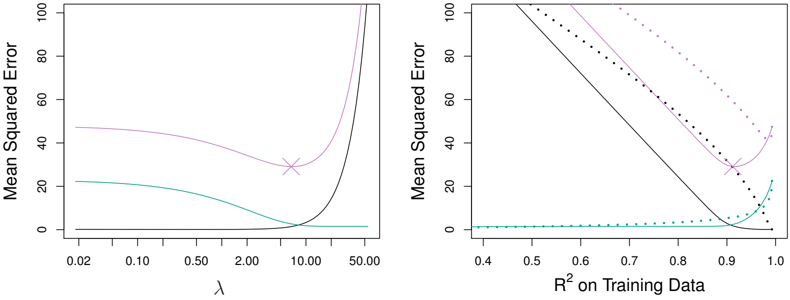
Example 1. Most of the coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge ($\cdots$).
- The bias is about the same for both methods.
- The variance of Ridge regression is smaller, so is the MSE.

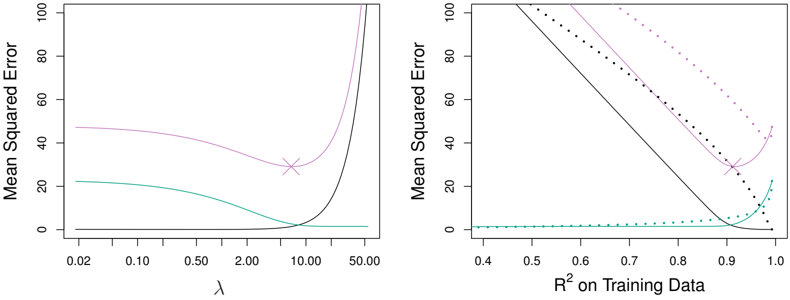
When is the Lasso better than Ridge?

Example 2. Only 2 coefficients are non-zero.



When is the Lasso better than Ridge?

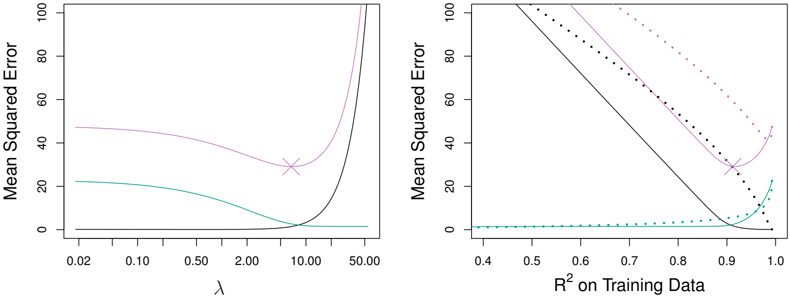
Example 2. Only 2 coefficients are non-zero.



- Bias, Variance, MSE.

When is the Lasso better than Ridge?

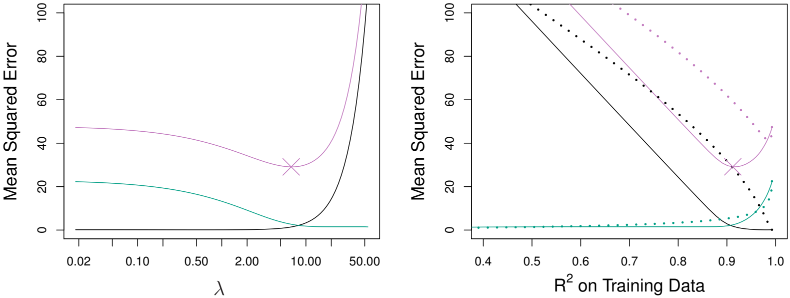
Example 2. Only 2 coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge ($\cdots$).

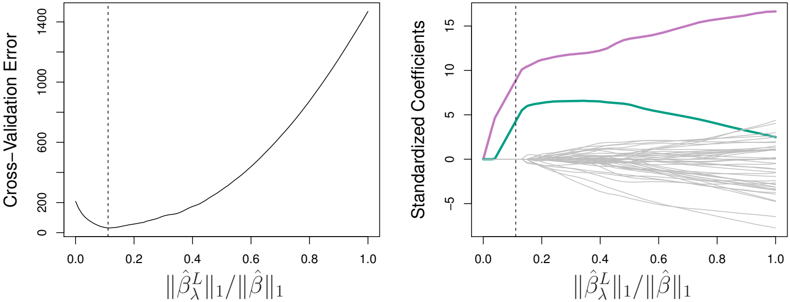
When is the Lasso better than Ridge?

Example 2. Only 2 coefficients are non-zero.



- Bias, Variance, MSE. The Lasso (—), Ridge ($\cdots$).
- The bias, variance, and MSE are lower for the Lasso.

Choosing λ by cross-validation



A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

Then, the objective function in Ridge regression can be simplified:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

Then, the objective function in Ridge regression can be simplified:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda\beta_j^2.$$

A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

Then, the objective function in Ridge regression can be simplified:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda\beta_j^2.$$

It is easy to show that

$$\hat{\beta}_j^{\text{ridge}} = \frac{y_j}{1+\lambda}.$$

A very special case

Similar story for the Lasso; the objective function is:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

A very special case

Similar story for the Lasso; the objective function is:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda|\beta_j|.$$

A very special case

Similar story for the Lasso; the objective function is:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

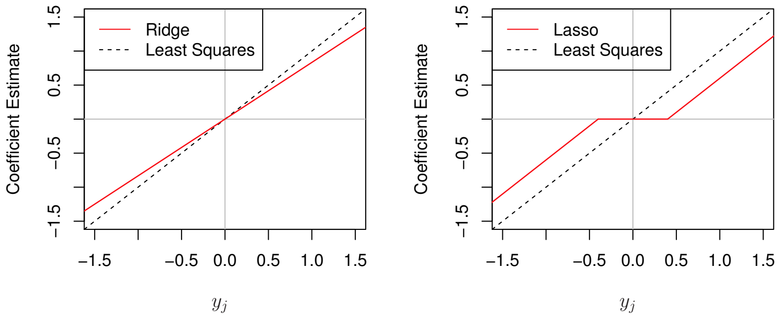
and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda|\beta_j|.$$

It is easy to show that

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| < \lambda/2. \end{cases}$$

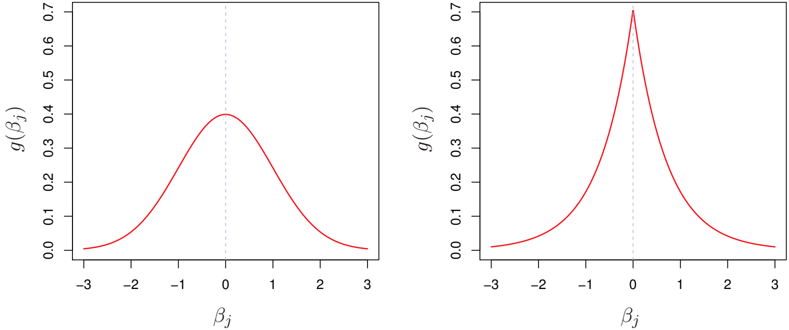
Lasso and Ridge coefficients as a function of λ



Bayesian interpretations

Ridge: $\hat{\beta}^{\text{ridge}}$ is the posterior mean, with a Normal prior on β.

Lasso: $\hat{\beta}^{\text{lasso}}$ is the posterior mode, with a Laplace prior on β.



Summary: Regression

Methods we have discussed:

- Linear regression with least squares
- Ridge regression, Lasso

Note: All of these are linear. The solutions are hyperplanes. The different methods differ only in how they *place* the hyperplane.

Summary: Regression

Ridge regression

Suppose we obtain two training samples $\mathcal{X}_1$ and $\mathcal{X}_2$ from the same distribution.

- Ideally, the linear regression solutions on both should be (nearly) identical.
- With standard linear regression, the problem may not be solvable (if $\mathbf{X}^T\mathbf{X}$ not invertible).
- Even if it is solvable, if the matrices $\mathbf{X}^T\mathbf{X}$ are close to singular (small spectral condition $c(\mathbf{X}^T\mathbf{X})$), then the two solutions can differ significantly.
- Ridge regression stabilizes the inversion of $\mathbf{X}^T\mathbf{X}$. Consequences:
  - Regression solutions for $\mathcal{X}_1$ and $\mathcal{X}_2$ will be almost identical if λ sufficiently large.
  - The price we pay is a bias that grows with λ.

Summary: Regression

Lasso

- The ℓ_1-costraint "switches off" dimensions; only some of the entries of the solution $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ are non-zero (sparse $\hat{\boldsymbol{\beta}}^{\text{lasso}}$).
- This variable selection also stabilizes $\mathbf{X}^T\mathbf{X}$, since we are effectively inverting only along those dimensions which provide sufficient information.
- No closed-form solution; use numerical optimization.

Formulation as optimization problem

Method	$f(\boldsymbol{\beta})$	Penalty	Solution method
Least squares	$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$	0	Analytic solution exists if $\mathbf{X}^T\mathbf{X}$ invertible
Ridge regression	$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$	$\|\boldsymbol{\beta}\|_2^2$	Analytic solution exists
Lasso	$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$	$\|\boldsymbol{\beta}\|_1$	Numerical optimization