Master in Financial Engineering (EPFL)
Financial Econometrics

Part II: Machine learning and Asset Pricing
Lecture 2: Selection of variables—Shrinkage methods

Florian Pelgrin

EDHEC Business School

February - June 2019

1. Shrinkage methods
   - Overview
   - Ridge regression
   - Lasso regression
   - Comparing Ridge and Lasso
   - Improving the Lasso

2. Technical appendix

3. Shrinkage (regularization) methods

3.1. Overview

- ✓ Main idea is to constrain or to regularize the coefficient estimates, or equivalently, to shrink the coefficient estimates towards zero
- ✓ This requires introducing some constraints (e.g., L_1 versus L_2 norm)...
- ✓ ... but also to pre-process data (standardized features), to apply (generally) nonlinear optimization algorithms, and to select optimal hyperparameters.
- ✓ Interestingly some procedures (e.g., Lasso method) allow for variable selection.

3.2. Ridge regression

Definition

Ridge regression coefficient estimates, denoted $\widehat{\beta}^R$, solve

$$\widehat{\beta}^R = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

After reparametrization using centred inputs, this is equivalent to:

$$\widehat{\beta}^R = \underset{\beta}{\texttt{argmin}} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

where $\lambda \geq 0$ is a *complexity/tuning* parameter that controls the amount of shrinkage, $\beta = (\beta_1, \cdots, \beta_p)^\top$, and $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$.

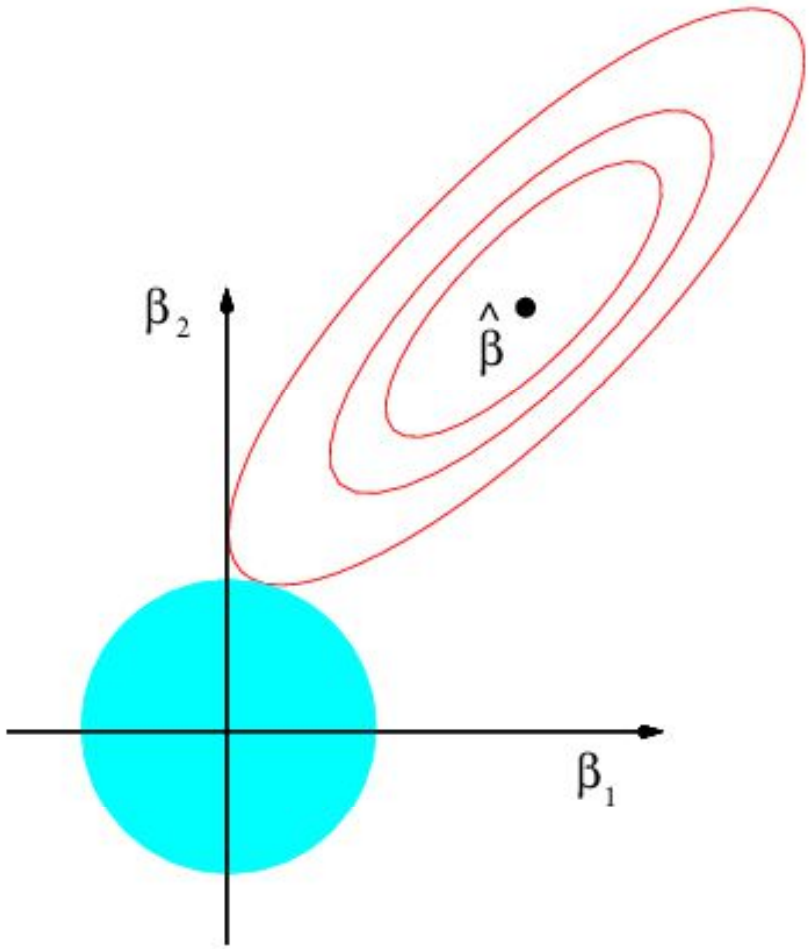

Figure: Ridge regression

Note: $\|\beta\|_2^2 = \beta_1^2 + \beta_2^2 \leq s$.
Source: The Elements of statistical learning, Hastie et al. (2001).

Remarks

- The larger the value of λ, the greater coefficients are shrunk toward zero:
  - ✓ When $\lambda = 0$: (non-unique) unconstrained ordinary least squares estimates;
  - ✓ When $\lambda \to \infty$: all coefficients $\to 0$ but none of them is exactly zero—it does not involve feature selection!
  - ✓ If λ is too high, the optimiser seeks to minimize the parameters more than it fits the data: need to find compromise value using cross-validation techniques.

Definition

Equivalently, the Ridge coefficient estimates solve

$$\widehat{\beta}^R = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2$$

$$\text{s.t.} \sum_{j=1}^{p} \beta_j^2 \leq s$$

where there is a one-to-one correspondence between the complexity parameter λ and the size of the constraint s.

Note: By imposing a size constraint, the *compensation issue* of coefficient estimates—a large positive coefficient is generally compensated by a large negative coefficient in a pair of correlated variables—is avoided in the presence of many correlated variables

Definition

The Ridge regression solutions are given by:

$$\widehat{\beta}^R = \left(X^\top X + \lambda I\right)^{-1} X^\top y$$

where I is the $p \times p$ identity matrix.

Notes:

- ✓ ℓ_2 penalty $\Rightarrow$ solution is linear w.r.t. y;
- ✓ Even if $X^\top X$ is singular, adding a positive constant makes the problem nonsingular

Key points

- When lambda is close to zero, the optimiser is mostly trying to fit the data closely whereas when lambda gets larger and larger, the optimiser is mostly trying to minimize the values of the parameters.
- The intercept is not included in the penalty.
- Solutions are **not scale-invariant** w.r.t. inputs, i.e. **inputs/features/regressors need to be standardized**.

Representations of results

- (Standardized) coefficients as a function of λ
- (Standardized) coefficients as a function of

$$\frac{\|\widehat{\beta}^R\|_2}{\|\widehat{\beta}\|_2}$$

where

  - $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$;
  - $\widehat{\beta}$ is the least squares estimate of β (if available);
  - $\widehat{\beta}^R$ is the Ridge estimate of β.

Note: As λ increases, $\|\widehat{\beta}^R\|_2$ decreases

- When $\lambda = 0$, $\frac{\|\widehat{\beta}^R\|_2}{\|\widehat{\beta}\|_2} = 1$;
- When $\lambda \to \infty$, $\frac{\|\widehat{\beta}^R\|_2}{\|\widehat{\beta}\|_2}$ "$\to$" 0.

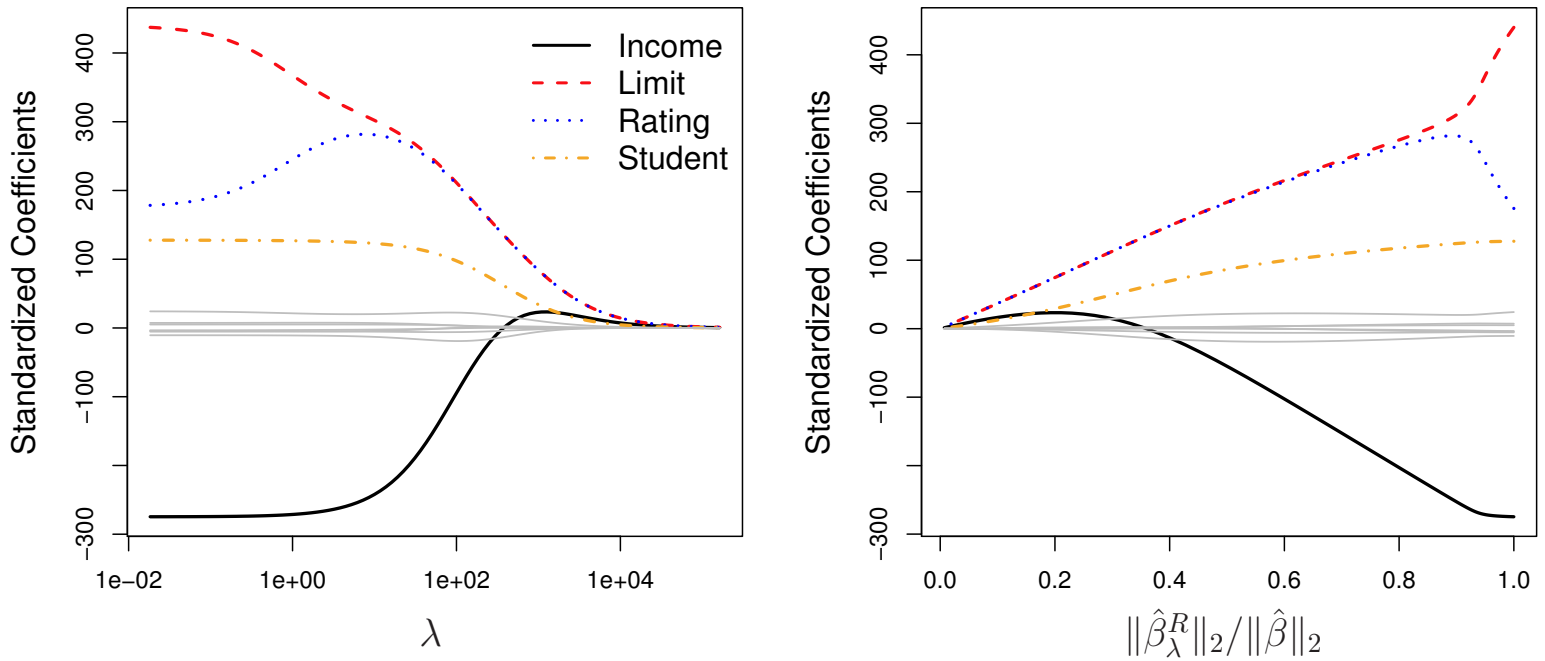

Figure: Credit data: Ridge regression

Note: Left panel—Standardized coefficients as a function of λ, Right-panel: Standardized coefficients as a function of $\frac{\|\widehat{\beta}^R\|_2}{\|\widehat{\beta}\|_2}$.

Least squares *versus* Ridge estimates

- ✓ Back to the bias-variance trade-off...Suppose p is large (and even almost as large as the sample size)
  - When $\lambda = 0$ (least squares estimation, the variance is high but the bias is small;
  - By varying λ, Ridge regression proceeds by trading off a small increase in bias for a large decrease in variance

  This means that Ridge regressions perform well when there is a high uncertainty (variance) around the least squares estimates!

3.3. Lasso regression

Definition

Lasso regression coefficient estimates, denoted $\widehat{\beta}^L$, solve

$$\widehat{\beta}^L = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

or

$$\widehat{\beta}^R = \underset{\beta}{\texttt{argmin}} \left\{ \|y - e_n\beta_0 - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where $\lambda \geq 0$ is a *complexity/tuning* parameter that controls the amount of shrinkage, $\beta = (\beta_1, \cdots, \beta_p)^\top$, $e_n = (1, \cdots, 1)^\top$, and $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

Remarks

- The role of λ is different from the one defining the Ridge regression:
  - ✓ When $\lambda = 0$: (non-unique) unconstrained ordinary least squares estimates;
  - ✓ When $\lambda \to \infty$: All estimates are exactly zero
  - ✓ Due to the nature of ℓ_1 penalty, some estimates are exactly equal to zero when the tuning parameter is sufficiently large—this is called **sparsity** and Lasso performs **variable selection**.
  - ✓ If λ is too high, the optimiser seeks to minimize the parameters more than it fits the data: need to find compromise value using cross-validation techniques.
- The intercept is not included in the penalty.
- Solutions are **not scale-invariant** w.r.t. inputs, i.e. **inputs/features/predictors need to be standardized**.

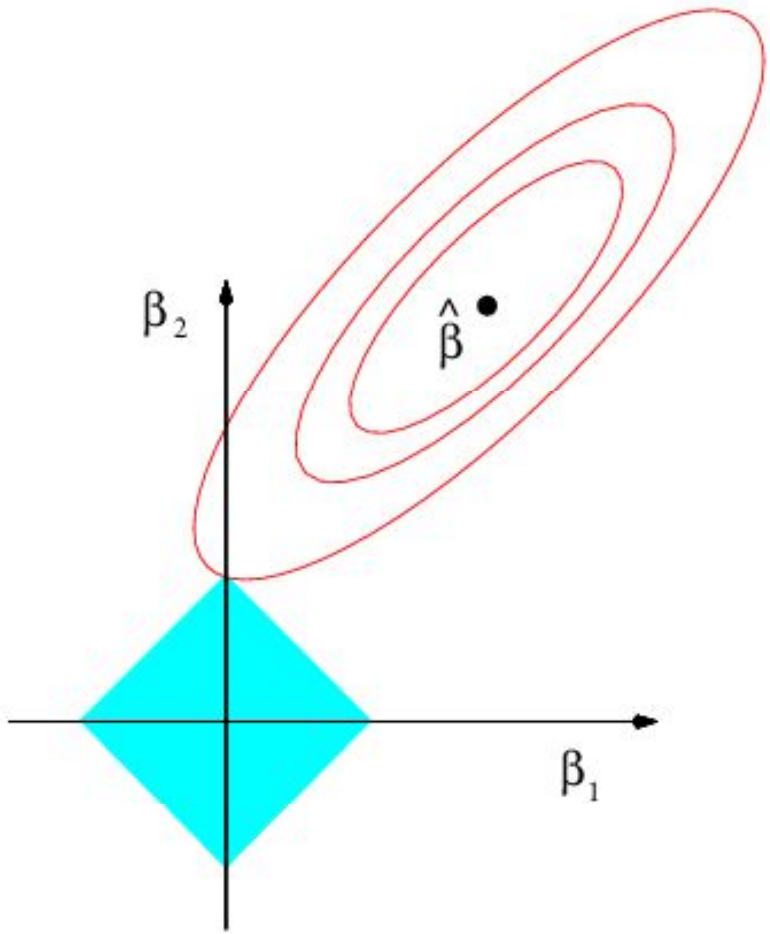

Figure: Lasso regression

Note: $\|\beta\|_1 = |\beta_1| + |\beta_2| \leq s$.
Source: The Elements of statistical learning, Hastie et al. (2001).

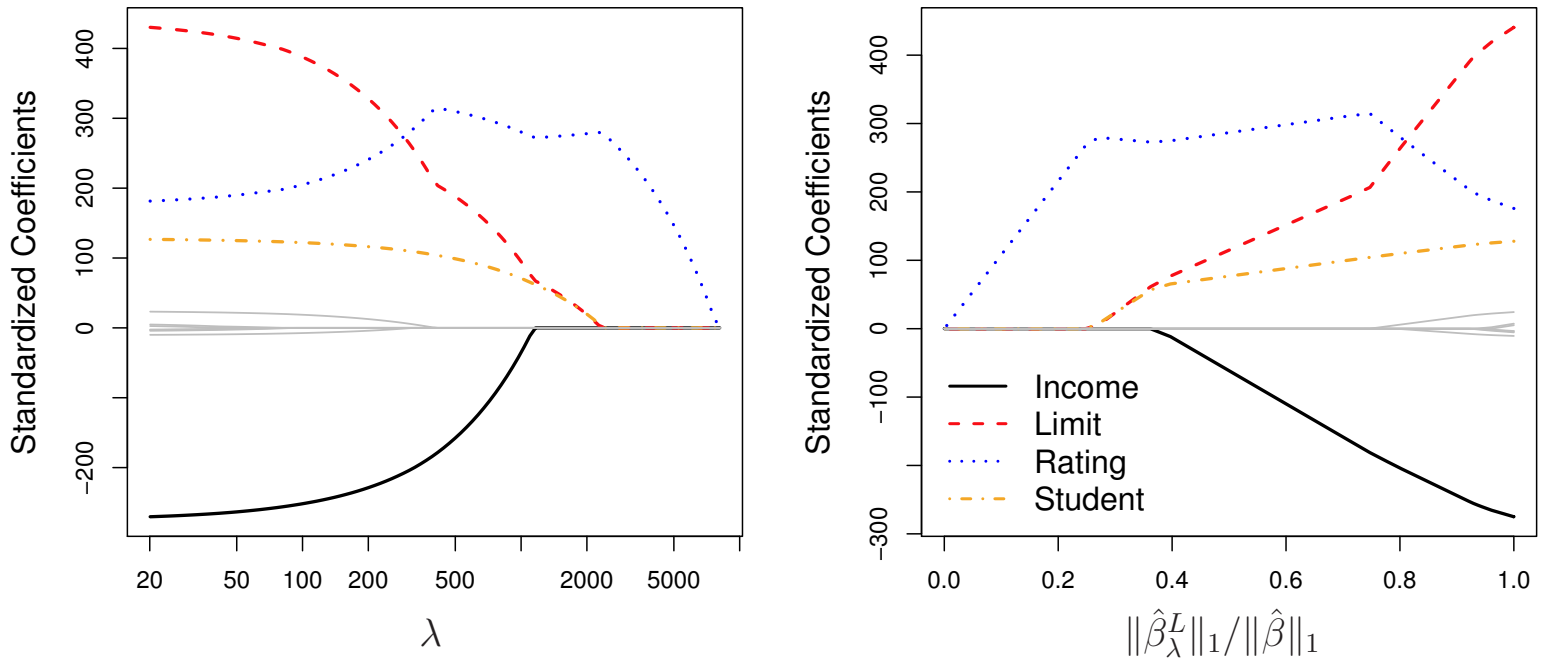

Figure: Credit data: Lasso regression

Note: Left panel—Standardized coefficients as a function of λ, Right-panel: Standardized coefficients as a function of $\frac{\|\widehat{\beta}^L\|_1}{\|\widehat{\beta}\|_1}$.

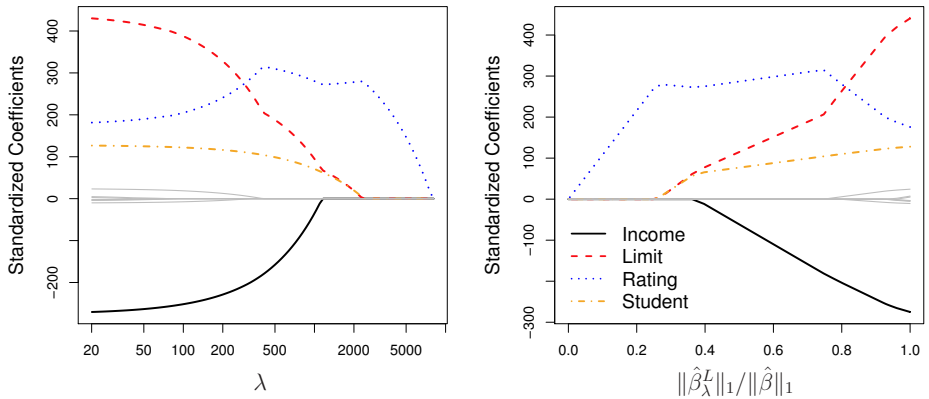

Figure: Credit data: Lasso regression

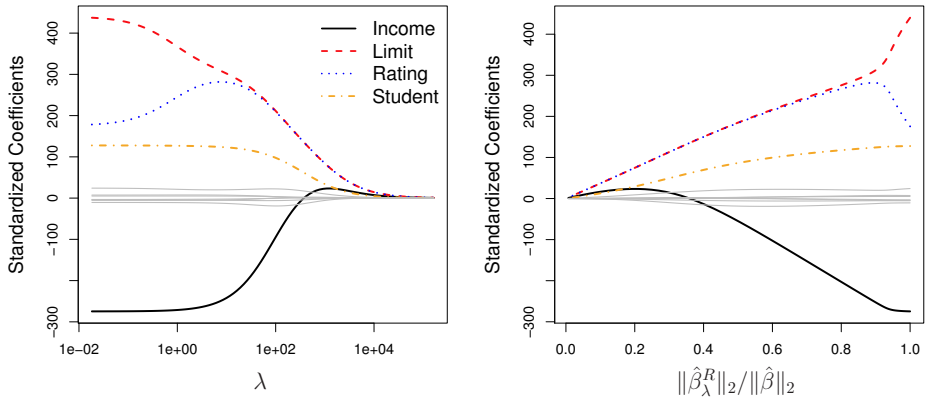

Figure: Credit data: Ridge regression

Note: Top panel—Lasso estimates, Bottom panel: Ridge estimates.

Definition

Equivalently, the Lasso coefficient estimates solve

$$\widehat{\beta}^L = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2$$

$$\text{s.t.} \sum_{j=1}^{p} |\beta_j| \leq s$$

where there is a one-to-one correspondence between the complexity parameter λ and the size of the constraint s.

Interpretation

- The inequality constraint can be interpreted as a *budget constraint*: How much are you ready to pay "s" for $\sum_{j=1}^{p} |\beta_j|$?

  - ✓ If s is large (i.e., the budget is not restrictive), then the sum can be large;
  - ✓ Especially, if s is large enough, the least squares solution will fall within this budget when $p < n$ (and there is no variable selection)
  - ✓ If s is small (i.e., the budget is restrictive), then the sum $\sum_{j=1}^{p} |\beta_j|$ must be small and variable selection is effective.

Statistical properties

- Conditions for recovering the "true" explanatory variables and some consistent estimates:
  - ✓ **Irrepresentable condition**: no high correlations between relevant and irrelevant predictors
  - ✓ **Beta-min condition**: True non-zero coefficients must be sufficiently "large".
- **Optimal** λ:
  - ✓ Cross-validation: λ is adaptively chosen to minimize the expected prediction error;
  - ✓ Optimal λ for prediction is generally greater than the optimal λ for variable selection: trade-off?

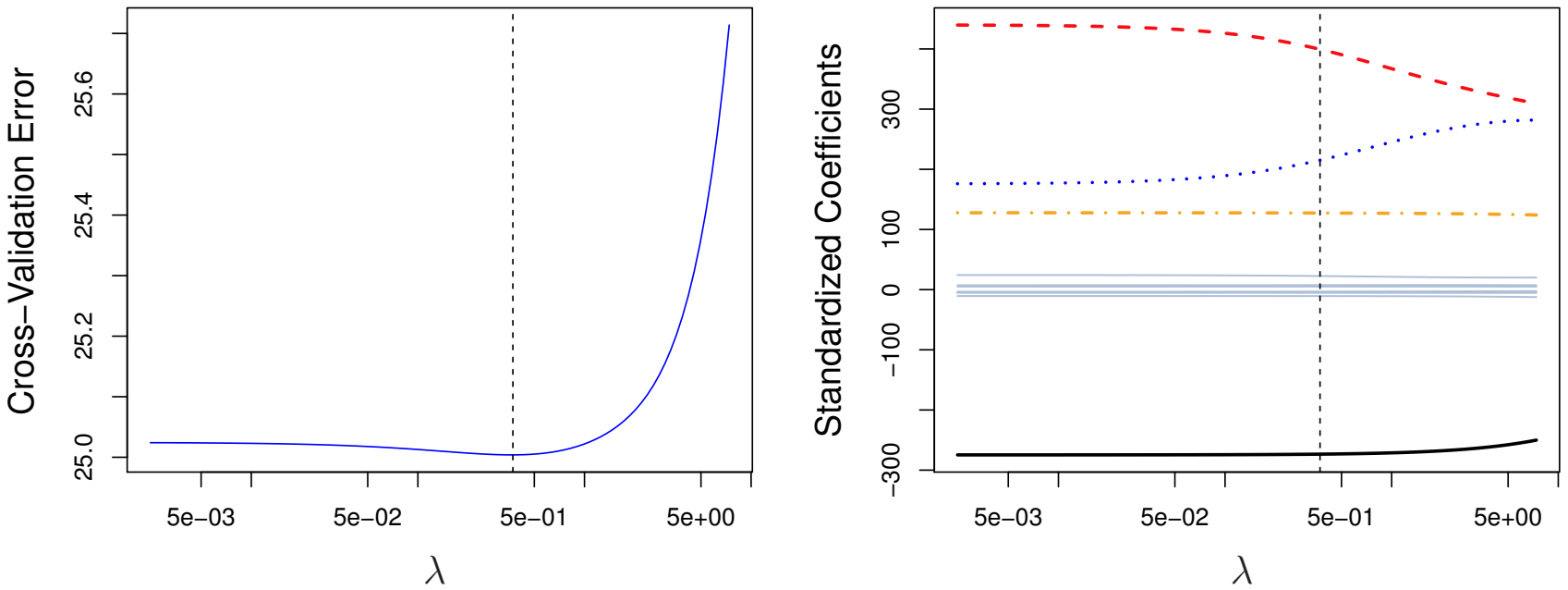

Figure: Credit data: Cross-validation and Lasso regression

3.4. Comparing Ridge and Lasso

Bayesian interpretation: Consider the (Gaussian) linear model

- ✓ $f(Y \mid X, \beta)$: Likelihood function (data information);
- ✓ $g(\beta \mid X) \equiv g(\beta)$: Prior distribution
- ✓ Posterior distribution is given by:

$$p(\beta \mid X, Y) \propto f(Y \mid X, \beta) g(\beta)$$

- ✓ If g is a Gaussian distribution, the posterior mode (or the posterior mean) of β is the Ridge regression solution;
- ✓ If g is a double-exponential (Laplace) distribution with mean zero and scale parameter a function of λ, the posterior mode of β is the Lasso regression solution;
- ✓ The Gaussian prior distribution is flatter and flatter at zero whereas the Laplace prior distribution picks at zero $\Rightarrow$ one expects that some coefficients are (exactly) zero by using the Laplace distribution (i.e., the Lasso regression) whereas the coefficients are expected to be randomly distributed around zero by using the Gaussian distribution (i.e., the Ridge regression).

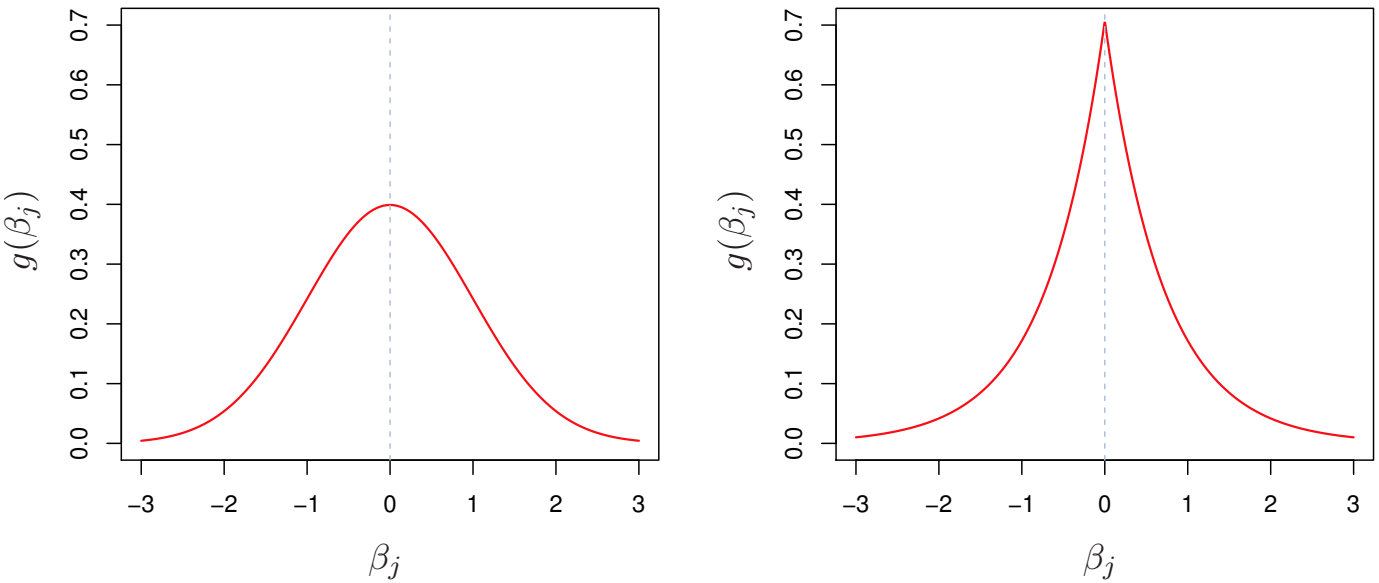

Figure: Prior distributions

Pros and cons of Lasso regressions

Generally,

- ✓ Lasso offers interpretable, stable models, and efficient prediction at a reasonable cost, and it allows for variable selection and solutions when $p > n$.

- ✓ Lasso is a nice compromise regarding the specification of the norm. Indeed, the L_q constrained problem can be written as

$$\widehat{\beta}^{Q-norm} = \underset{\beta}{\texttt{argmin}} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right\}$$

where $\|\beta\|_q^q = |\beta_1|^q + \cdots + |\beta_p|^q$.

  - When $q < 1$: less bias than Lasso but non-convexity of the minimization problem. Variable selection can still be implemented.
  - When $q > 1$: problem is convex but no variable selection
  - When $q = 1$: Lasso is a compromise!

Some rules of thumbs...

- ✓ Lasso better if the "true model" is sparse because it shrinks the remaining coefficients less than Ridge
- ✓ Ridge is better when the predictors are highly correlated: Ridge keeps the redundant variables, but shrinks them whereas Lasso discards all but one.
- ✓ Trade-off (correlation) $\Rightarrow$ Elastic Net regression!

A simple comparison

Consider a linear regression model in which $n = p$, X is the $n \times n$ identity matrix (orthonormal columns of X), and there is no intercept:

$$y_i = x_i\beta_i + u_i = \beta_i + u_i \quad i = 1, \cdots, n.$$

Method	Objective function	Estimates
Least squares	$\sum_{j=1}^{p}(y_j - \beta_j)^2$	$\widehat{\beta}_j^{\text{ls}} = y_j$
Ridge	$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$	$\widehat{\beta}_j^{\text{R}} = \frac{y_j}{1+\lambda}$
Lasso	$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$	$\widehat{\beta}_j^{\text{L}} = y_j - \lambda/2$ if $y_j > \lambda/2$
		$= y_j + \lambda/2$ if $y_j < \lambda/2$
		$= 0$ otherwise

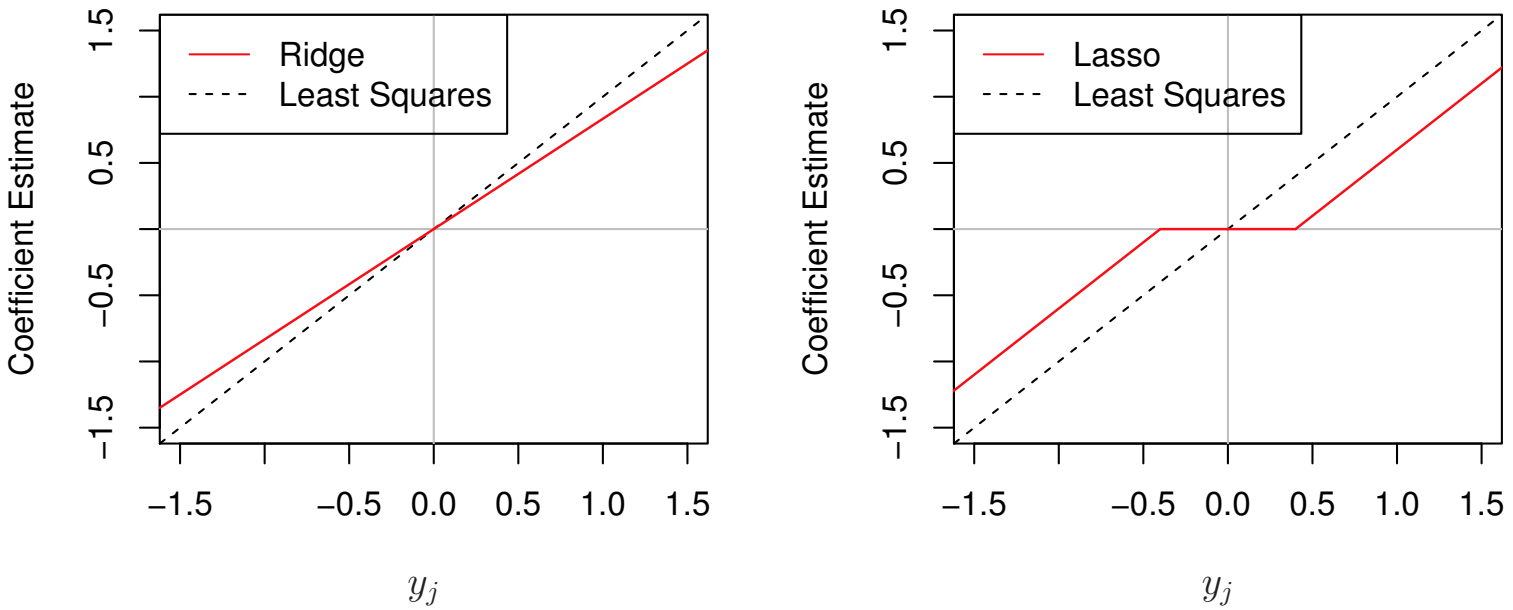

Figure: LS, Ridge, and Lasso regressions

Source: The Elements of statistical learning, Hastie et al. (2001).

3.5. Improving the Lasso regression

- **Problem 1**: Right-hand side variables are identified but their coefficient estimates are biased
  - ✓ Relaxed Lasso
  - ✓ Variable Inclusion and Shrinkage Algorithm (VISA)
  - ✓ Adaptive Lasso
  - ✓ Others: Dantzig selector, LAD-Lasso
- **Problem 2**: Strong correlations $\Rightarrow$ make use of Elastic Net rgression (main advantages: 1. if $p > n$ can select more than n variables; 2. Can select group of redundant variables)
- **Problem 3**: Group of variables—Group Lasso, Composite Absolute Penalties, Fused Lasso, etc...
- **Problem 4**: model extensions—generalized linear model, logistic regression, nonlinear models.
- **Problem 5**: Mixture of data (e.g., quantitative and qualitative data).

Elastic net regression

Definition

Elastic net regression coefficient estimates, denoted $\widehat{\beta}^L$, solve

$$\widehat{\beta}^L = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

or

$$\widehat{\beta}^{EN} = \underset{\beta}{\texttt{argmin}} \left\{ \|y - e_n\beta_0 - X\beta\|_2^2 + \lambda \left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2 \right) \right\}$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$.

Remarks:

- ✓ Elastic net mixes the Lasso and Ridge penalties
  - When $\alpha = 1$: Lasso regression;
  - When $\alpha = 0$: Ridge regression;
- ✓ Elastic net performs generally better than Lasso in the presence of correlated features (even when α gets closer to one);
- ✓ In contrast to the Ridge regression, when $p > n$, elastic net can consider more than n variables.
- ✓ In the presence of a group of relevant and redundant variables, Lasso generally tends to discard all but one variable from this group whereas elastic net will tend to select the entire group of features.

Adaptive Lasso

Definition

The (two-step) adaptive Lasso estimates, denoted $\widehat{\beta}^{AL}$, solve

$$\widehat{\beta}^{L} = \underset{\beta_0,\cdots,\beta_p}{\texttt{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{L}|} \right\}$$

where $\widehat{\beta}^{L}$ is the Lasso estimate of β.

Remarks:

- One makes use of the Lasso estimate of β from a first stage in a prediction optimal way (i.e., the tuning parameter is determined by cross-validation)
- In a second step,one makes use again of cross-validation to select the hyperparameter in the adaptive Lasso, i.e. regularization parameters are determined in a sequential way.
- This can be extended to a k-step adaptive Lasso.
- $\widehat{\beta}_j^L = 0 \Rightarrow \widehat{\beta}_j^{AL} = 0.$
- Adaptive Lasso provides a sparse solution and generally reduces the number of **false positives** (i.e. the selection of irrelevant variables).

Technical appendix

A1. Singular Value Decomposition (SVD)

Definition

Let X denote an $n \times p$ matrix. The singular value decomposition (SVD) takes the form:

$$X = UDV^\top$$

where

- ✓ U is a $n \times p$ *orthogonal* matrix whose columns, u_j $j = 1, \cdots, p$, span the column space of X;
- ✓ V is a $p \times p$ *orthogonal* matrix whose columns, u_j $j = 1, \cdots, p$, span the row space of X;
- ✓ D is a $p \times p$ *diagonal* matrix with diagonal entries or singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

A2. Least squares and Ridge

Using the SVD,

- The least squares fitted vector, denoted $\widehat{y}^{ls}$, is

$$\begin{aligned}\widehat{y}^{ls} = X\widehat{\beta}^{LS} &= X\left(X^\top X\right)^{-1}X^\top y \\ &= UU^\top y.\end{aligned}$$

- The Ridge fitted vector, denoted $\widehat{y}^{r}$, is

$$\begin{aligned}\widehat{y}^{r} = X\widehat{\beta}^{R} &= X\left(X^\top X+\lambda I\right)^{-1}X^\top y \\ &= UD\left(D^2+\lambda I\right)^{-1}DU^\top y \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^\top y\end{aligned}$$

where $0 \leq \frac{d_j^2}{d_j^2+\lambda} \leq 1$.

Interpretation:

- The least squares regression compute the coordinates of y w.r.t. the orthonormal basis U, $U^\top y$.
- The Ridge regression computes also these coordinates but *shrinks* them by the factors $\frac{d_j^2}{d_j^2+\lambda}$:
  - More shrinkage is applied to the columns vectors u_j (basis vectors) with smaller d_j^2;
  - Less shrinkage is applied to the columns vectors u_j (basis vectors) with larger d_j^2;
  - Why? Singular values d_j correspond to "directions" in the column space of X (the vector space engendered by the column vectors, i.e. the explanatory variables). Notably, small singular values corresponds to directions having small variance.

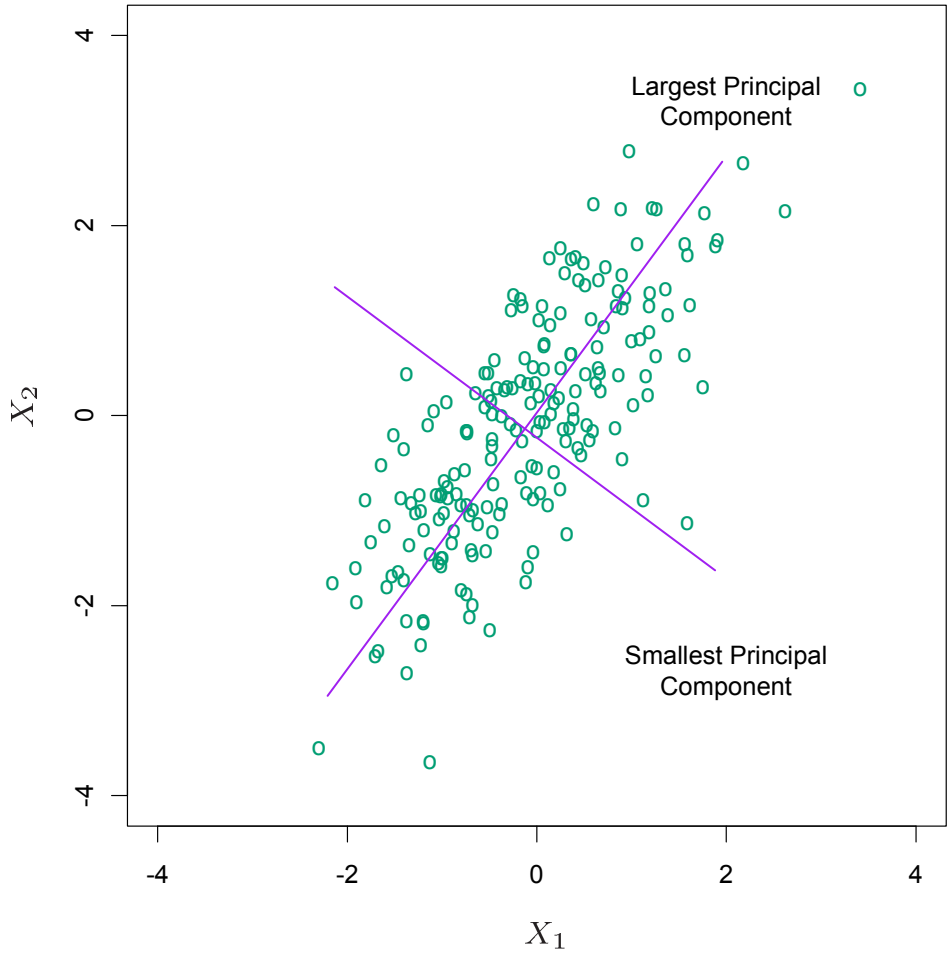

Figure: Shrinkage of the Ridge regression

Ridge regression projects y onto the two components and then shrinks more the coefficients of low-variance component relative to the high-variance one.
Source: The Elements of statistical learning, Hastie et al. (2001).