# Shrinkage

Instructor: Haoran LEI

Hunan University

# Shrinkage Methods

- JS Estimators imply that we can fit a model containing all $p$ predictors using a technique that *shrinks the coefficient estimates towards zero*.
- It turns out that shrinking the coefficient estimates can significantly *reduce their variance,* and thus can improve the fit.
- Two popular shrinkage methods in the context of linear model: **Ridge Regression** and **lasso**.

# From OLS to Ridge regression

- Linear model: $f_L(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.
- Recall that the **LS fitting** procedure estimates $\beta_0$, $\beta_1$, ..., $\beta_p$ using the values that minimize RSS over the training data:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- where $\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$.

# From OLS to Ridge regression

- Linear model: $y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$.
- Ridge regression uses the $\hat{\beta}^R$ that minimize

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

- where $\lambda \geq 0$ is a **tuning parameter**, to be determined separately.

# Ridge regression

- As with LS, **ridge regression** seeks coefficient estimates that fit the data well, by *making the RSS small.*
- However, the second term $\lambda \sum_{j=1}^{p} \beta_j^2$ is small when the $\beta$'s are close to zero, and so it has the effect of *shrinking* the estimates of $\beta_j$ towards zero.
- The tuning parameter $\lambda$ is *determined by cross validation*, and serves to control the relative impact of these two terms on the regression coefficient estimates.

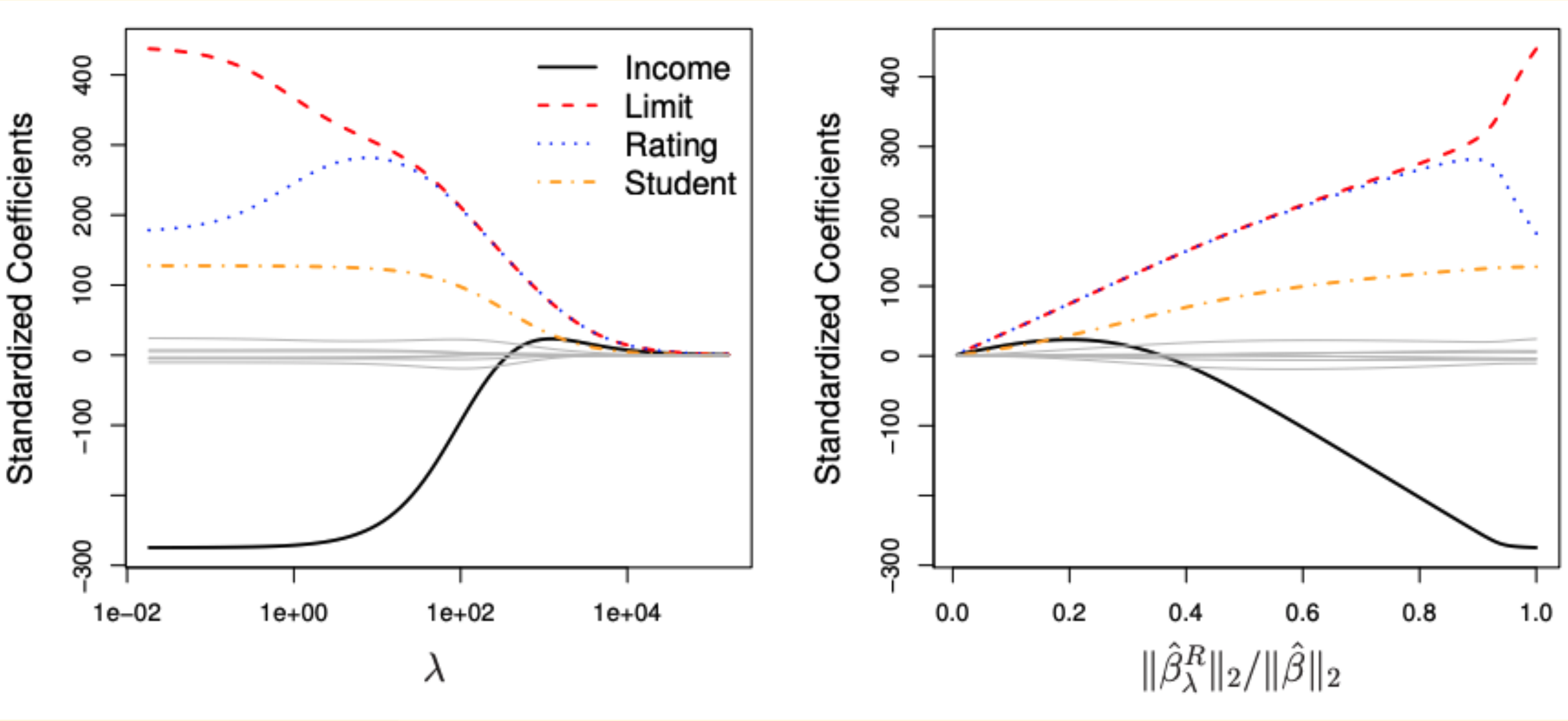



*Figure.* the credit data example: predicting **balance** from other 10 predictors (**age**, **cards**, **gender**, **student**, **limit**, ...)

## Some *"stupid questions"* in case anyone is confused...

- **Q:** How many "models" are there in the left panel?
  **A:** Infinite. Each $\lambda \in (0.01, 10000)$ leads to some $\beta_\lambda^R$.
- **Q:** What are "1e-02" and "1e+04" in the left panel?
  **A:** That's how R represents $10^{-2}$ and $10^4$ (scientific e notation).
- **Q:** What's $||\hat{\beta}||_2$ in the right panel?
  **A:** This is called **the $\ell_2$ norm:** $||x||_2 = \sqrt{x_1^2 + \cdots + x_p^2}$.
  Specifically, $\hat{\beta}$ is OLS estimator and $\hat{\beta}_\lambda^R$ is the Ridge estimator with the tuning parameter $\lambda$.
- **Q:** Why the range of x-axis is [0,1] in the right panel?

# Ridge regression: scaling of predictors

- The OLS estimates are *scale equivariant:* multiplying $X_j$ by a constant $c$ simply leads to a scaling of the LS estimates by a factor of $1/c$.
  - In other words, regardless of how the $j$-th predictor is scaled, $\hat{\beta}_j X_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can *change substantially* when multiplying a given predictor by a constant, due to the penalty term in the ridge regression objective function.

# Ridge regression: scaling of predictors

- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula:

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

- Like OLS, Ridge regression allows an exact formula:

$$\hat{\beta}^R = (X^T X + \lambda I_p)^{-1} X^T Y$$

- You can see that the ridge estimates are *not* scale equivariant.

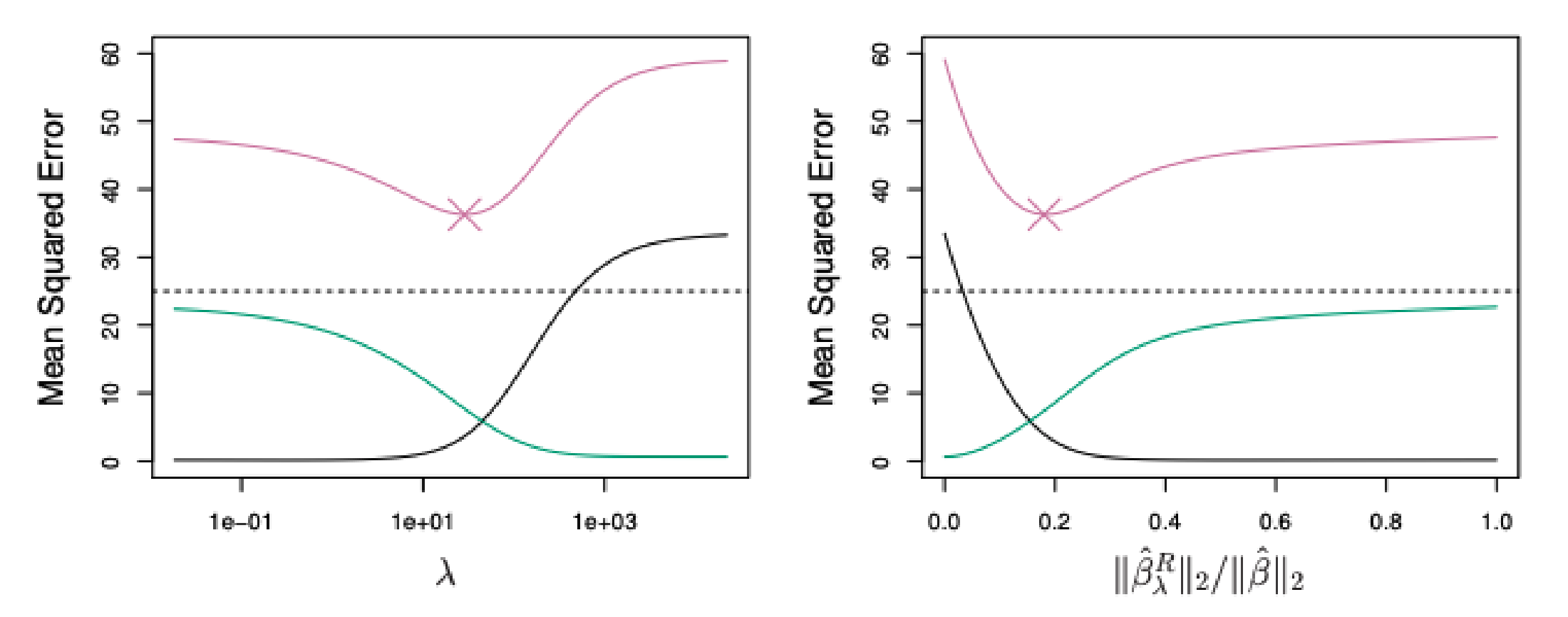



*Fig: Ridge regression and bias-variance tradeoff.* **Squared bias** (*black*), **variance** (*green*), and **test mean squared error** (*purple*) for the ridge regression predictions on a simulated data set.

# Least Absolute Shrinkage and Selection Operator (LASSO)

- **Lasso**, invented by Rob Tibshirani in 1996, is a relatively recent alternative to ridge regression.
- Ridge regression has one obvious disadvantage:
  - unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all $p$ predictors in the final model.
  - In other words, Ridge does *not select features*.
- Lasso is mainly proposed to overcome that disadvantage.

## Least Absolute Shrinkage and Selection Operator (LASSO)

- The lasso coefficients, $\beta_\lambda^L$, minimize the quantity

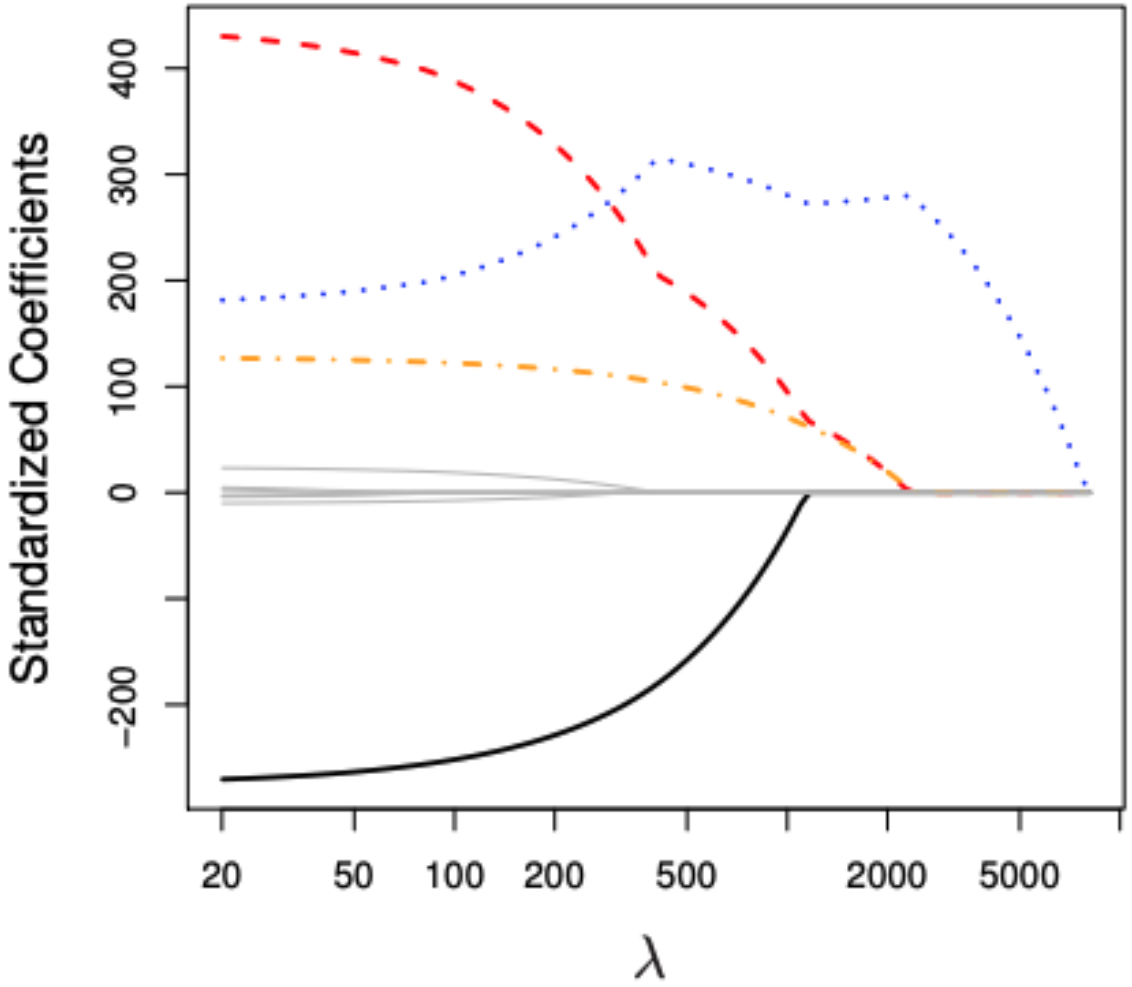
$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

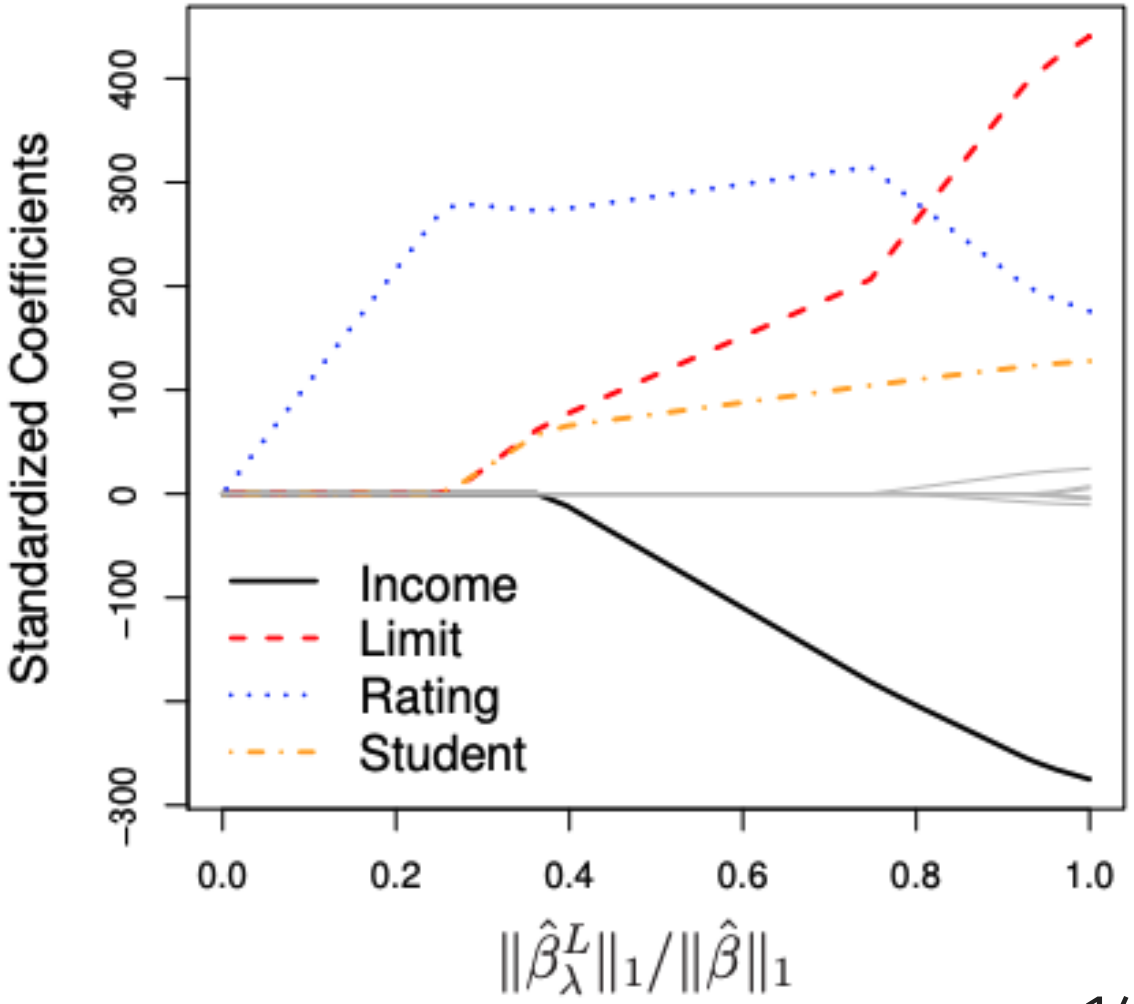
- In other words, the lasso uses an $\ell_1$ **(pronounced "ell 1") penalty** instead of an $\ell_2$ penalty.
  - The $\ell_1$ norm of a coefficient vector $\beta$ is given by $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$.

# Lasso v.s. Ridge Regression

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the $\ell_1$ penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter $\lambda$ is sufficiently large.
- Hence, much like best subset selection, the lasso *performs variable selection*.
- We say that the lasso yields **sparse models** — that is, models that involve only a subset of the variables.

# Example: Credit dataset and lasso



## Review: Lagrangian of an optimization problem

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\min \sum_{i=1}^{N} \Big( y_i - f_L(x_i) \Big)^2 \text{ s.t. } \sum \beta_j^2 \leq s \text{ (Ridge)}$$

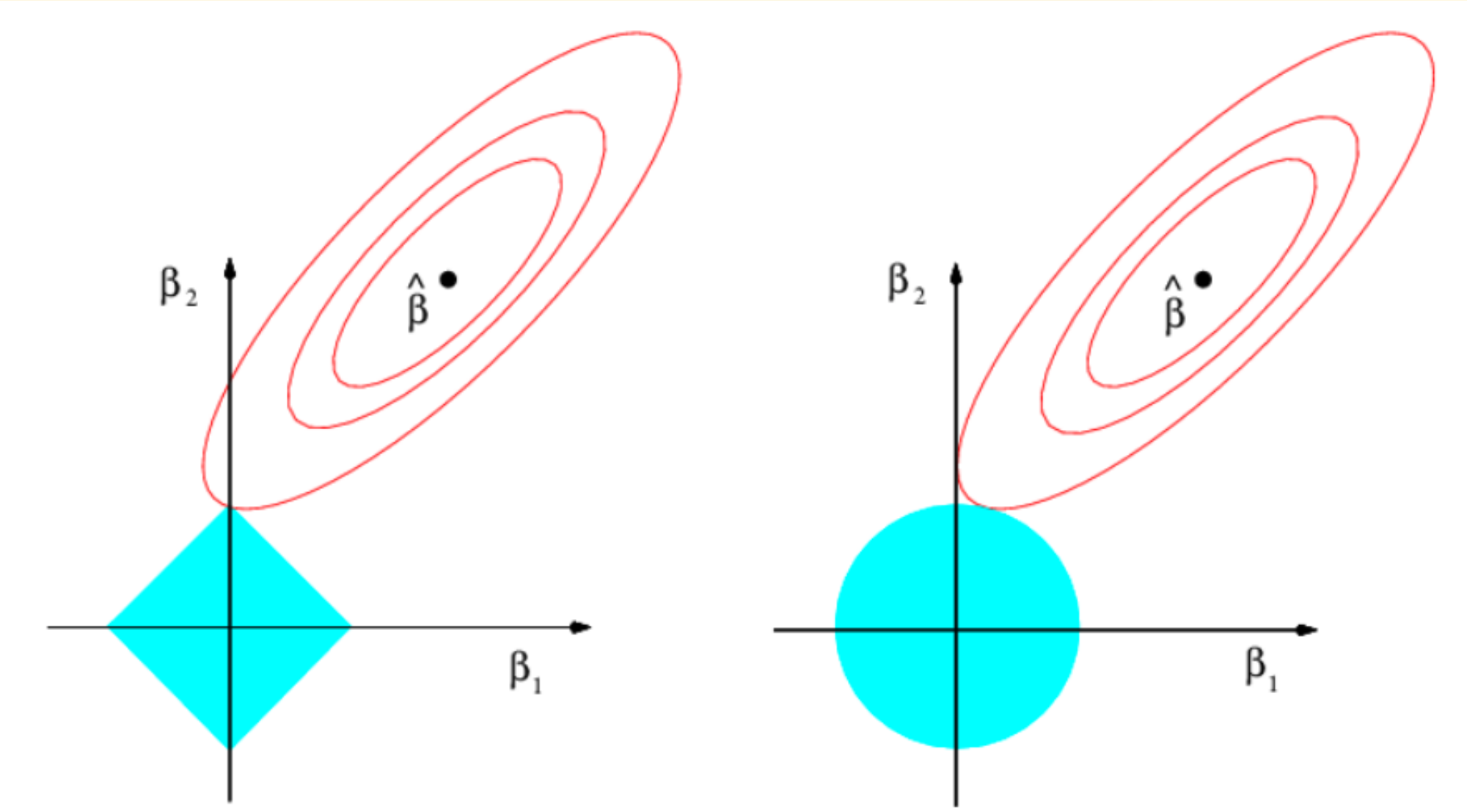
$$\min \sum_{i=1}^{N} \Big( y_i - f_L(x_i) \Big)^2 \text{ s.t. } \sum |\beta_j| \leq s \text{ (Lasso)}$$

respectively, where $f_L(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.

# Why LASSO performs feature selection while Ridge Regression does not?

- Ridge regression uses $\ell_2$ penalty. So in the optimization, the search area is a circle, which leads to an *interior solution.*
- Lasso uses $\ell_1$ penalty. So in the optimization, the search area is a square, which leads to an *corner solution.*

# Lasso v.s. Ridge Regression: illustration



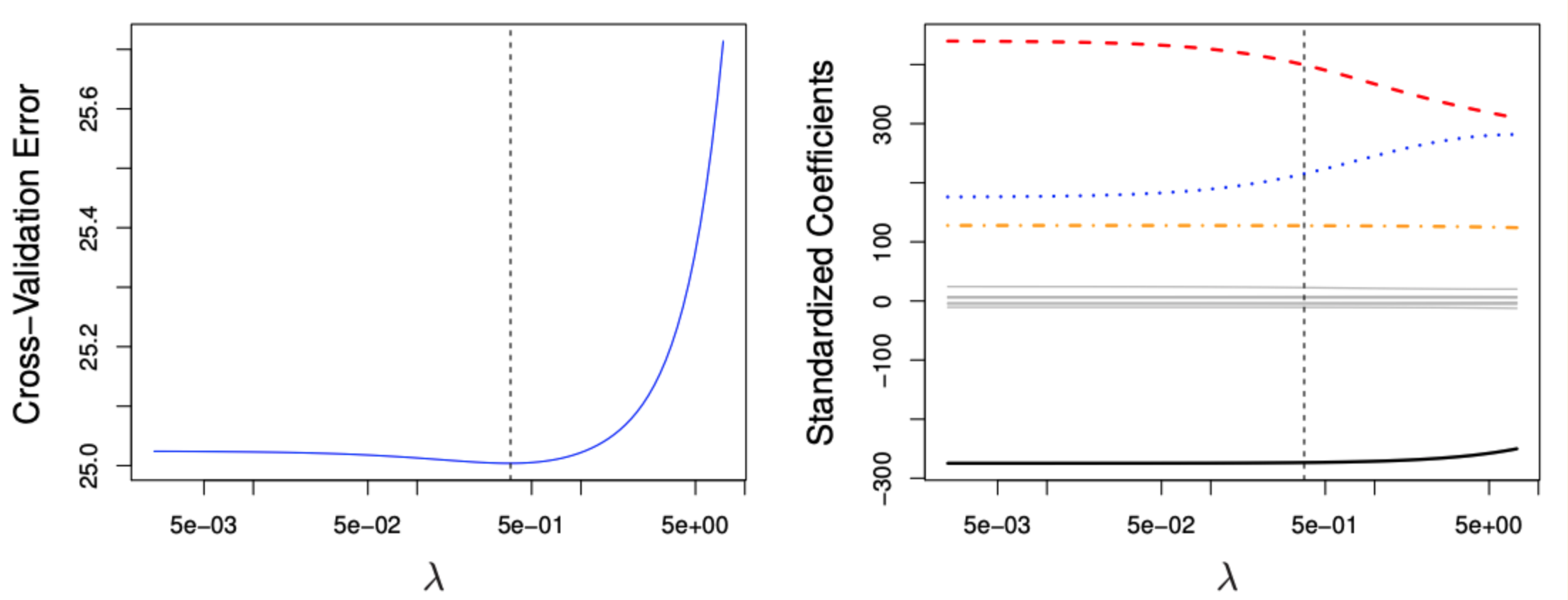
# Lasso v.s. Ridge regression on predicting power

- In general, one might expect the lasso to perform better when the response is a function of *only a relatively small* number of predictors.
- However, the number of predictors that is related to the response is never known *a priori* for real data sets.
- One needs to use cross-validation to determine which approach is better on a particular problem.

## Selecting the Tuning Parameter

- For both Ridge Regression and Lasso, we need to do cross-validation to select a value for the tuning parameter $\lambda$ (or equivalently, the value of the constraint $s$):
  1. We choose a grid of $\lambda$ values, and compute the cross-validation error rate for each value of $\lambda$.
  2. We then select the tuning parameter value for which the cross-validation error is smallest.
  3. Finally, the model is re-fit using all of the available observations and the selected value of $\lambda$.

*Fig: Credit data example.* Cross-validation errors that result from applying **ridge regression** to the Credit data set.



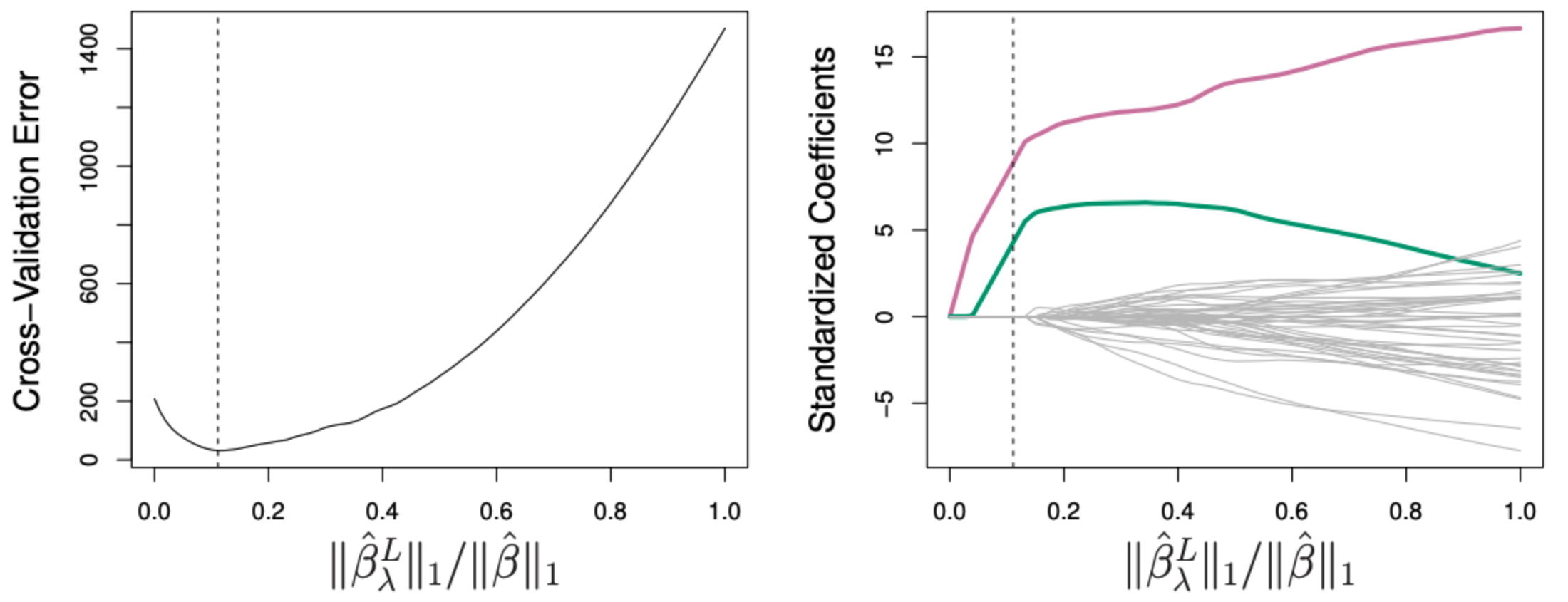



*Fig: CV on lasso.* Ten-fold cross-validation MSE for the **lasso**, applied to the sparse simulated data set.