A. Giovanidis 2021

# 07. Feature Selection

Network Data Analysis – NDA'21

Anastasios Giovanidis

Sorbonne-LIP6




October 27, 2021

# Bibliography

B.1 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An introduction to statistical learning: with applications in R". Springer Texts in Statistics. ISBN 978-1-4614-7137-0
Chapter 6
DOI 10.1007/978-1-4614-7138-7

B.2 Giorgos Dimopoulos, Ilias Leontiadis, Pere Barlet-Ros, Konstantina Papagiannaki. "Measuring Video QoE from Encrypted Traffic", IMC '16 Proceedings of the 2016 Internet Measurement Conference Pages 513-526 .

A. Giovanidis 2021

# Intro

In the multiple-regression setting, we assumed that the linear model with additive noise:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

describes the relationship between a response $Y$ and a set of $p \geq 1$ predictor variables $X_1, X_2, \ldots, X_p$.
- The model fit uses least squares (LSs) to estimate the $\hat{\beta}_i$'s.

But, is it always a good fit? Are there any ways to improve this fit?

☞ Feature Selection and Regularization.

# Main idea

A. Giovanidis 2021

☞ Either shrink the coefficients for some feature variables or remove them completely!

Why?

- Prediction Accuracy: If $n >> p$ then LSs do have low variance. But when e.g. $n \leq p$ the model is highly variable!
- Model Interpretability: Some variables used as predictors may not be relevant with the response. Better remove them to reduce model complexity.

A. Giovanidis 2021

# Network example

☞ In [B.2] the authors want to classify Video QoE from encrypted traffic. One of the questions is the quality of stalling.

There are potentially many available features to be used (around $p = 70$)

- Only 4-out-of-70 features are actually important factors that correlate with stalling:
  - BDP mean (related to throughput)
  - packet re-transmission max
  - chunk-size min
  - chunk size standard deviation.

❂ In fact chunk size is a very strong indicator, because at the event of stalling, the size of the chunks decrease so that they are reliably transmitting and start filling-up the buffer.

# Methods

A. Giovanidis 2021

We will present two main methods that modify the LSs:

- Subset Selection: Identify a subset of the original $p$ predictors to be relevant (say $p_s < p$). Then apply LSs fit.
- Shrinkage: Fit the model with all $p$ features, but shrink some coefficients even to zero.
  $\rightarrow$ This method reduces variance.

# Best subset selection

A. Giovanidis 2021

To find the best set we need to perform LSs for all possible combinations for the $p$ predictors: (remember the *F-score* and *p-value*)

- All models with 1 predictor: $p$.
- All models with 2 predictors: $\begin{pmatrix} \mathrm{p} \\ 2 \end{pmatrix} = \frac{p(p-1)}{2}$.
- etc.

In total $2^p$ possibilities.

A. Giovanidis 2021

1

**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 $R^2$.

☞ Pick models with smallest Train RSS, Select with smallest Test RSS.

[1] Source [B.1]

A. Giovanidis 2021

# Many possibilities to test...

The method needs to test too many feature combinations:

- for $p = 10$, approx $1,000$ models,
- for $p = 20$, over $1,000,000$ possibilities!
- etc.

The Best subset selection becomes computationally infeasible for large sets of features.

☞ We need to find other ways to select good subsets stepwise.

# Forward Stepwise Selection

A. Giovanidis 2021

The method:

- Begins with a model without predictors,
- adds predictors to the model one-at-a-time,
- until all predictors are in the model.

At each step the variable that gives the greatest additional improvement to the fit is added.

**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 $R^2$.

☞ Pick models with smallest Train RSS, Select with smallest Test RSS.

[2] Source [B.1]

A. Giovanidis 2021

# Advantages

The method is computationally advantageous compared to Best selection:

- Instead of $2^p$ fitting models, it needs to compute only $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models.
- e.g. for $p = 20$, fit 211 models instead of $1,048,576$ !
- Can be used also when $n < p$ (stops at $n$ features)

It is not guaranteed to find the best possible model out of the $2^p$.

A. Giovanidis 2021

3

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

[3] Source [B.1]

# Backward Stepwise Selection

A. Giovanidis 2021

[4]

**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 $R^2$.

[4] Source [B.1]

# Choosing the Optimal Model

A. Giovanidis 2021

All methods hand-pick a small number of models based on a small value of Train RSS or high value of $R^2$.

☞ The model with all the predictors will have the smallest Train RSS.

✴ To choose exactly one model among these, we need to find the one with smallest Test error. To do so we can:

- Estimate the Test Error, by adjusting the Train Error to account for Bias.
- Directly estimate the Test Error using a validation set or cross-validation.

A. Giovanidis 2021

# Adjustment of the Train Error

☞ The training RSS will decrease as more variables are included in the model, but **not the Test RSS** necessarily.

We cannot use Train error to select among models with different numbers of variables.

Adjust the Train error to select the model with best Test prediction:

- **Mallow's $C_p$-estimate**: $C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$.
- **Akaike Information Criterion**: $AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$.
- **Bayesian Information**: $BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$.
- **Adjusted** $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$, where $TSS = \sum(y_i - \bar{y})^2$.

A. Giovanidis 2021

# Understanding Akaike I

The original definition of Akaike reads

$$AIC = \frac{1}{n}\left(2d - 2\log(\hat{L})\right)$$

where $\hat{L}$ is the log-likelihood and $d \leq p$ is the number of predictors used. Akaike adds a cost which scales linearly with the number of used predictors.
If the model tested is

$$y = f(x) + \epsilon \Rightarrow \mathbb{E}[y] = f(x),$$

then, the error per data is

$$\epsilon_i = y_i - \mathbb{E}[f(x_i)]$$

Suppose the model describes well the data, so that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

A. Giovanidis 2021

# Understanding Akaike II

The log-likelihood of an error sample is

$$\hat{L}(\mathcal{D}_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi)$$

The second term changes slowly over $\sigma^2$. The third term is constant. Altogether

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

❋ Note that $C_p$ and AIC are proportional to each other!

☞ To estimate the variance we will use (with $TSS = \sum_i (y_i - \bar{y})^2$ the total sum of squares for the response)

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}TSS.$$

# Understanding BIC

A. Giovanidis 2021

Similar to Akaike, but now the cost depends on the $\log(n)$ of the samples

$$BIC = \frac{1}{n}\left(\log(n)d - 2\log(\hat{L})\right),$$

where $\hat{L}$ is the log-likelihood and $d$ is the number of predictors used.

As in AIC

$$BIC = \frac{1}{n}\left(\log(n)d + \frac{1}{\sigma^2}RSS\right).$$

Since $\log(n) > 2$ for $n > 7$ the model places a heavier penalty on models with many features.

We choose the $AIC, BIC, C_p$ model with the **lowest** value!

A. Giovanidis 2021

# Understanding Adjusted $R^2$

Remember the usual definition of $R^2$

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{Explained\ \ Variation}{Total\ \ Variation}.$$

☞ The more we add predictors, the more the RSS decreases and the more the Train $R^2$ increases!

For a least squares model with $d$ features the adjusted $R^2$ statistic is

$$Adjusted\ R^2 \quad = \quad 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}.$$

Unlike the other metrics, here a **large value of Adjusted** $R^2$ indicates a model with a small Test error.

Maximising the Adjusted $R^2$ is equivalent to minimizing $\frac{RSS}{n-d-1}$.
This statistic also pays a price for inclusion of unnecessary variables.

# Overview of adjustment metrics

A. Giovanidis 2021

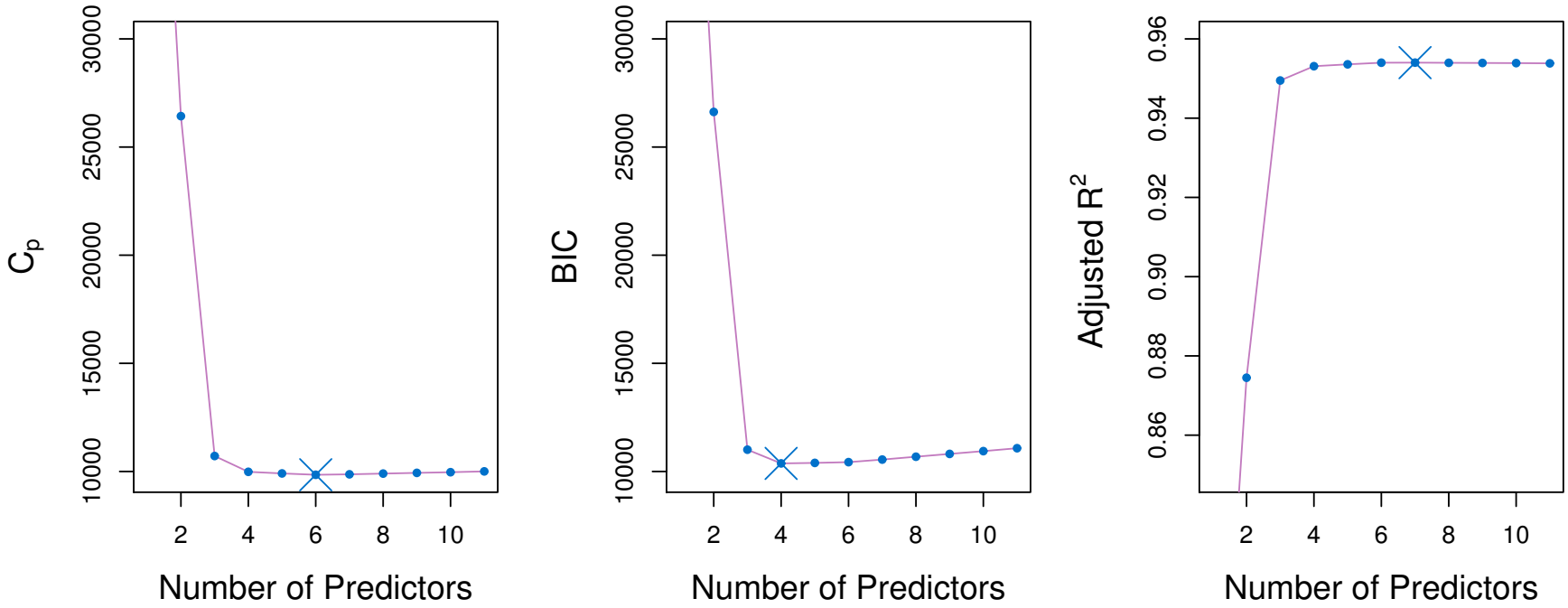



Figure: Feature selection from different metrics.[5]

[5] Source [B.1]

# Comparison with Validation and CV tests

A. Giovanidis 2021

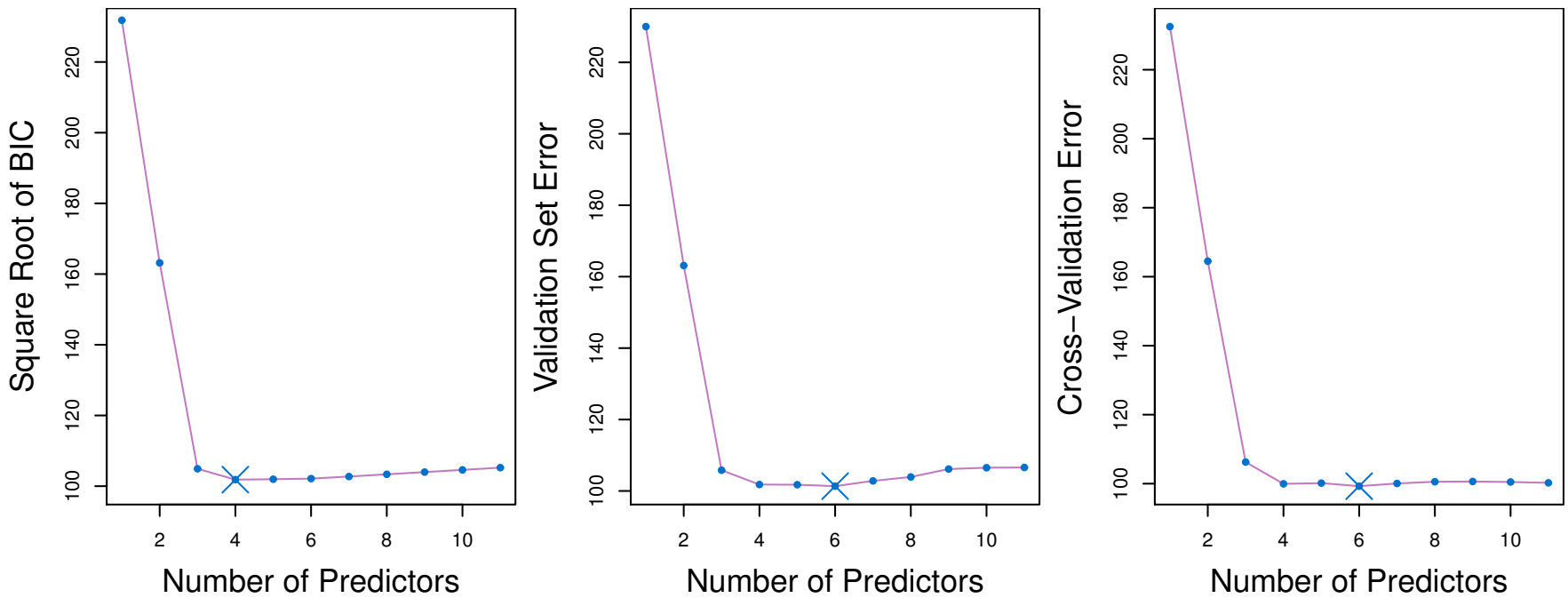



Figure: CV used to be computationally expensive, not any more.[6]

[6] Source [B.1]

# Shrinkage

A. Giovanidis 2021

We have seen methods to optimally select a subset of appropriate features, leaving the rest out.

☞ As an alternative, we can keep all $p$ features, but use a technique that constraints or regularizes the coefficient estimates.

Estimates can be shrunk towards zero! This technique can significantly reduce variance.

Ridge Regression and the Lasso.

# Ridge Regression

A. Giovanidis 2021

Similar to LSs fit, the Ridge Regression solves

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

where the lefthand side is just the RSS, and $\lambda \geq 0$ is a tuning parameter, to be determined.

The second term is called shrinkage penalty.

A. Giovanidis 2021

# Properties

- When $\lambda = 0$: it is just the Least-Squares fit.
- When $\lambda \to \infty$ $\beta$'s will approach zero.
- Find the "best" set of parameters $\beta$.

☞ Each choice of $\lambda$ produces a different set of estimates $\hat{\beta}_\lambda$.

**Note 1:** The shrinkage penalty is **not** applied to the intercept $\beta_0$.
**Note 2:** Best apply ridge-regression after standardizing the predictors (all on the same scale / standard deviation 1)

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

# Ridge Example

A. Giovanidis 2021

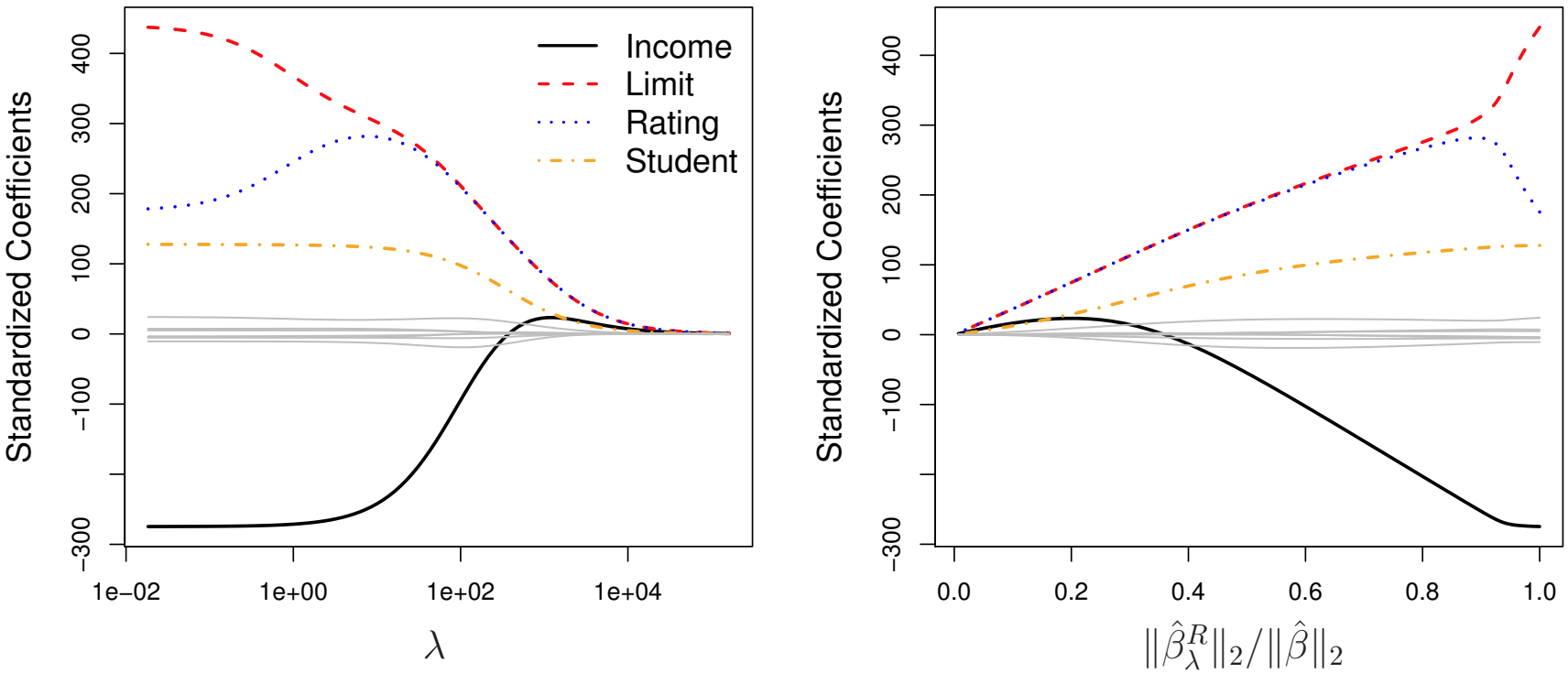



Figure: Change of Ridge Regression coefficients vs $\lambda$.[7]

[7] Source [B.1]

A. Giovanidis 2021

# Improvement over LSs

☞ As $\lambda$ increases, the flexibility of the ridge regression fit decreases: decreased variance but increased Bias.

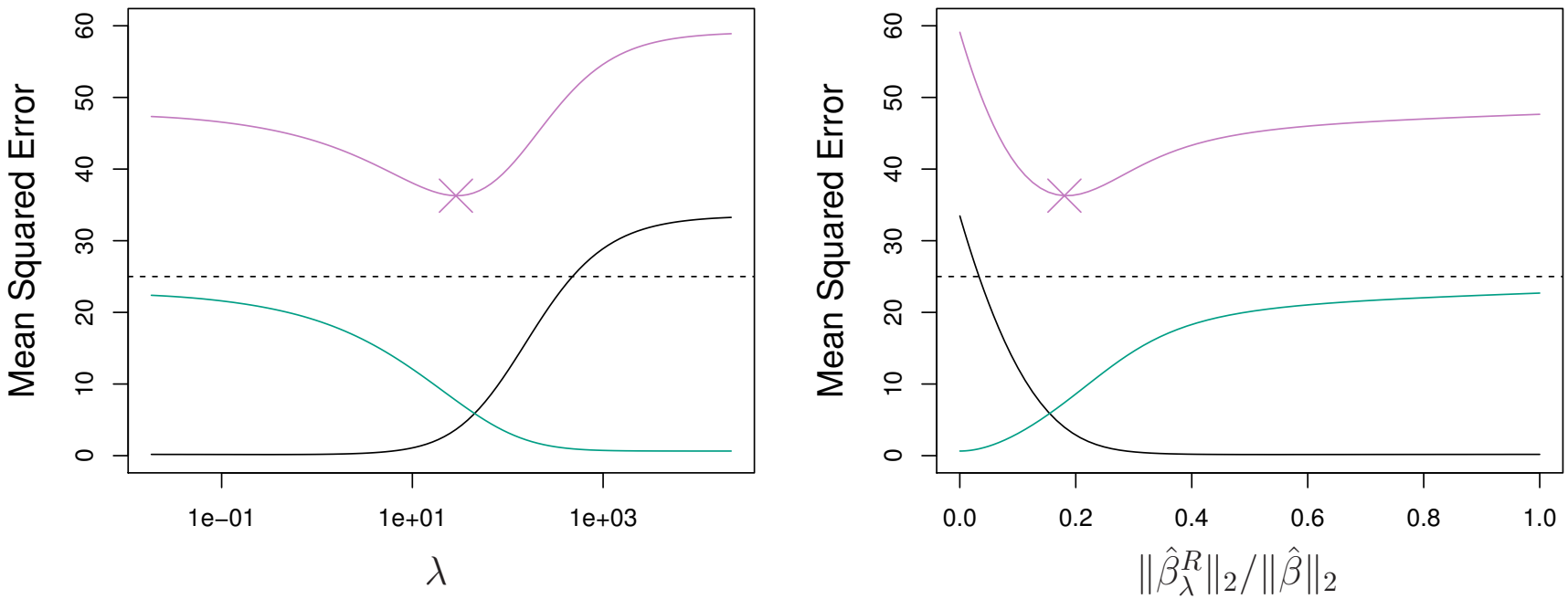



Figure: Bias vs Variance tradeoff and Test MSE.[8]

[8] Source [B.1]

# The Lasso

A. Giovanidis 2021

Similar to Ridge Regression, the Lasso solves a different problem

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

where the lefthand side is just the RSS, and $\lambda \geq 0$ is again a tuning parameter, to be determined.

The second term is the lasso penalty (uses $\ell_1$-norm instead of $\ell_2$).

A. Giovanidis 2021

# Advantages

As formulation, the Lasso is similar to Ridge Regression, with a penalty that uses a different norm.

What is new here?

- The Lasso penalty can force some estimates to be exactly zero $\rightarrow$ performs Variable Selection.
- Lasso's models are sparse involving a subset of variables.
- Simple, more interpretable models.

# Example Lasso

A. Giovanidis 2021

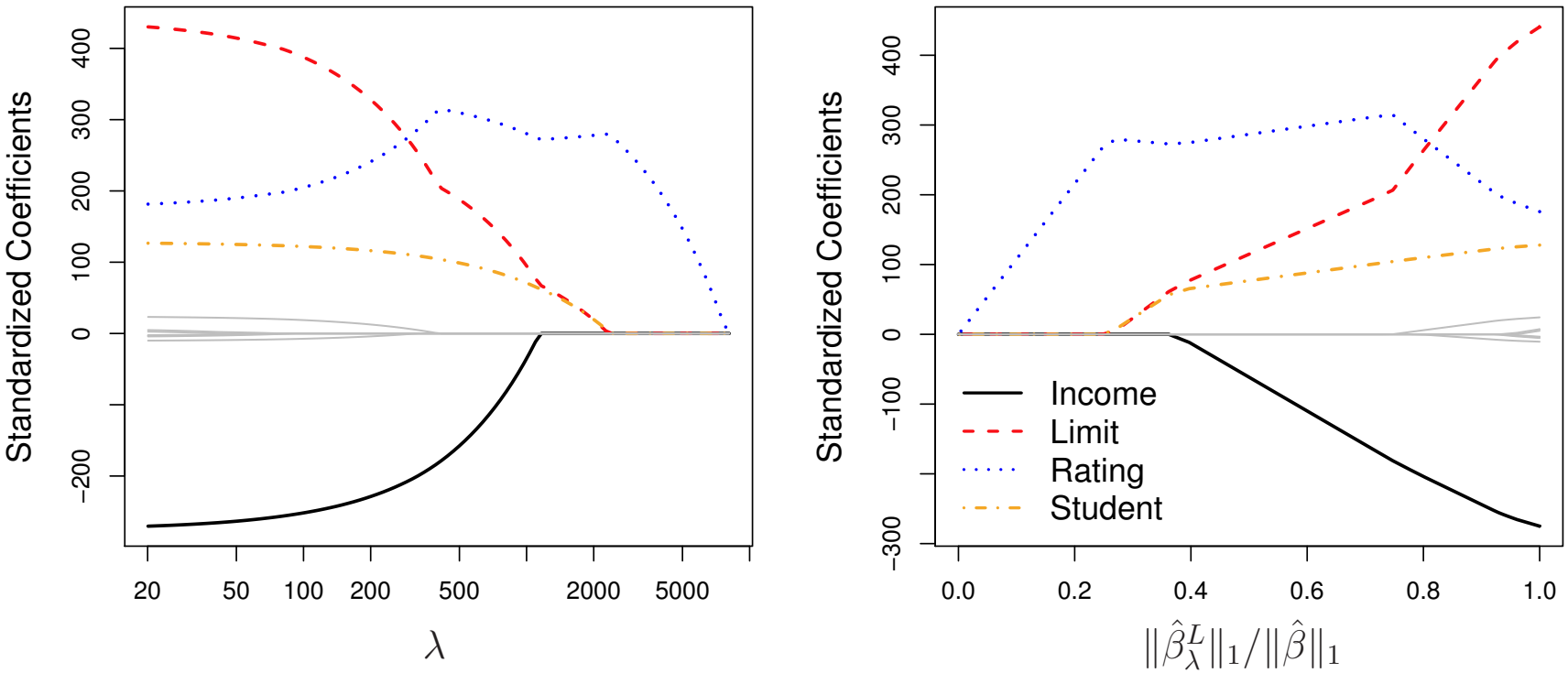



Figure: Change of Lasso coefficients vs $\lambda$.[9]

[9] Source [B.1]

A. Giovanidis 2021

# Equivalent Problems

The Ridge, Lasso, and Best subset selection are each equivalent to:

$$\min_\beta \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad s.t. \quad \sum_{j=1}^{p} \beta_j^2 \leq s \quad (\textit{Ridge})$$

$$\min_\beta \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad s.t. \quad \sum_{j=1}^{p} |\beta_j| \leq s \quad (\textit{Lasso})$$

$$\min_\beta \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad s.t. \quad \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) \leq s \quad (\textit{Best}).$$

# Illustrative Explanation

A. Giovanidis 2021

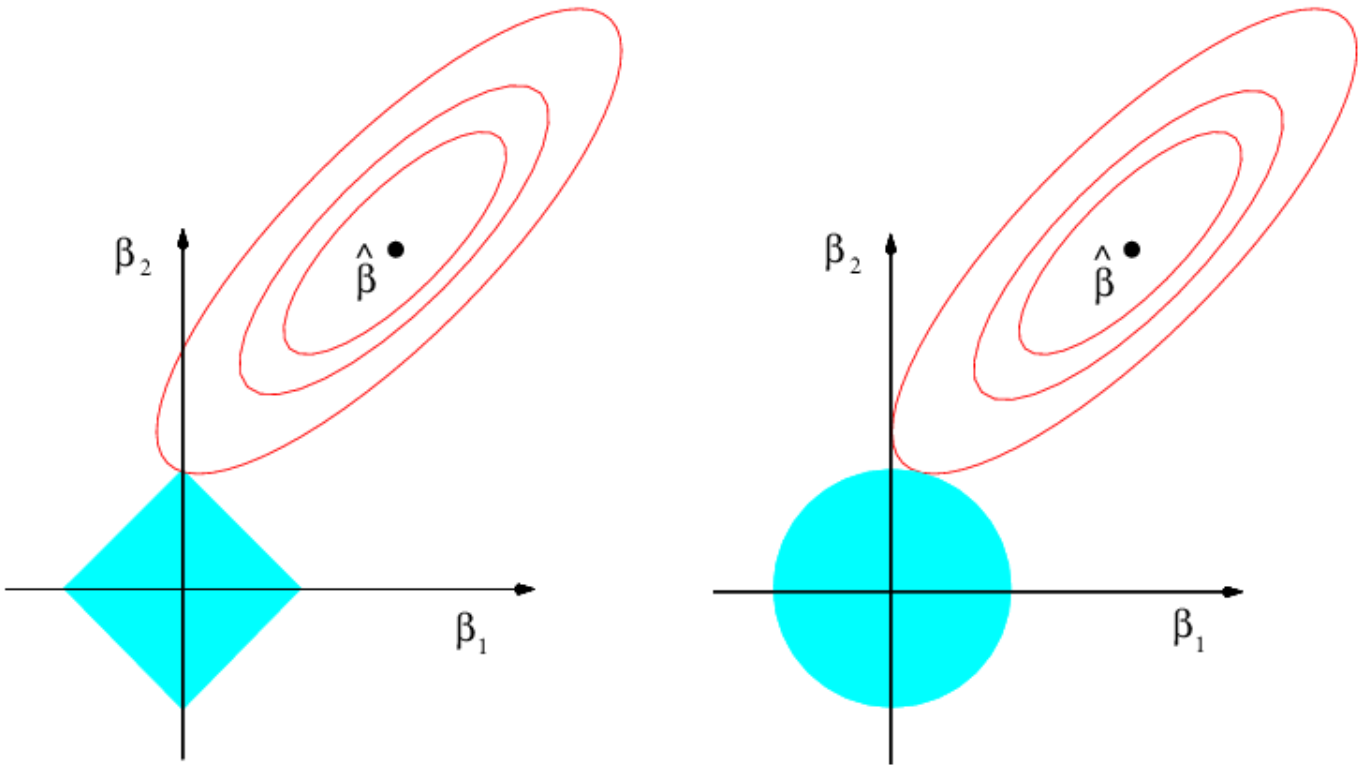


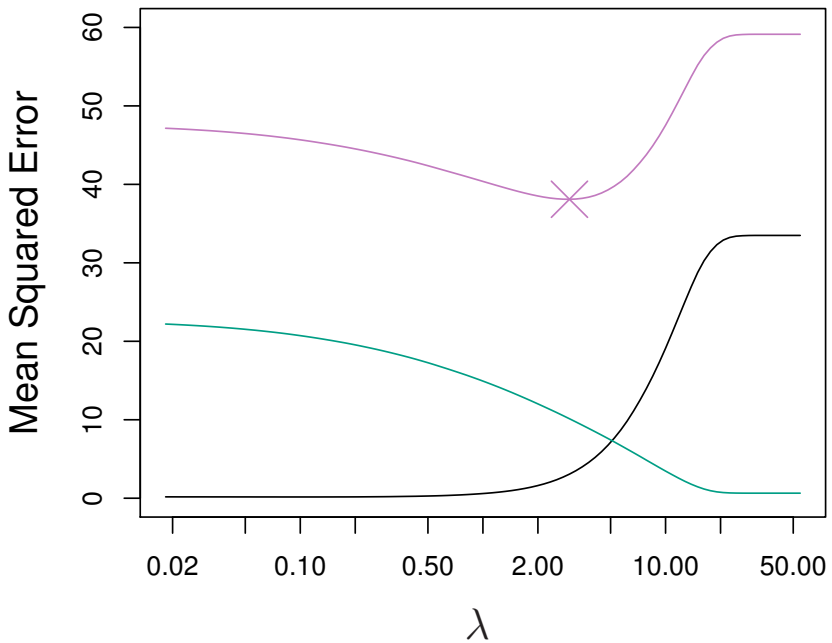

Figure: Why does Lasso lead to estimates equal to 0?[10]

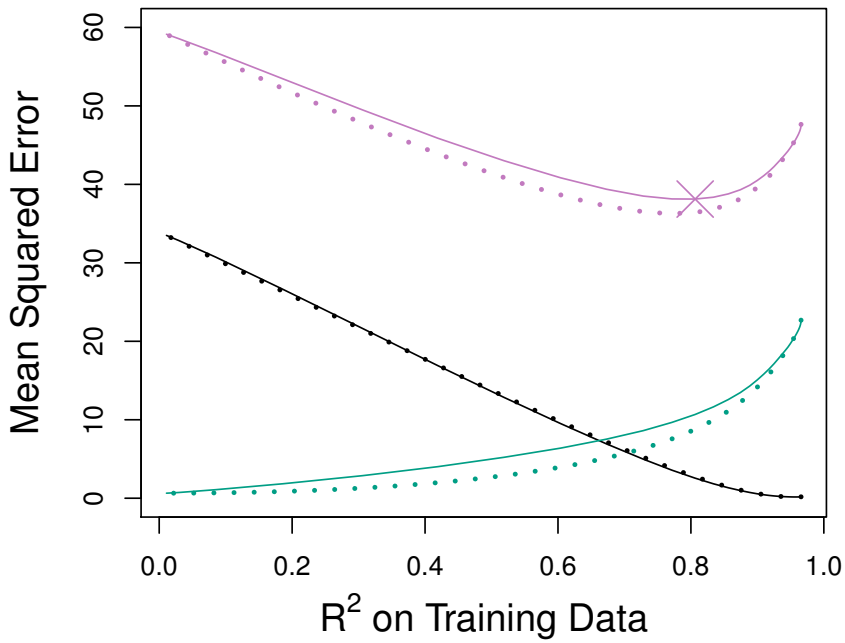
[10] Source [B.1]

A. Giovanidis 2021

# Ridge > Lasso

Here: Ridge needs all 45 coefficients $\neq 0$. Lasso chose 2-out-of-45 features.[11]



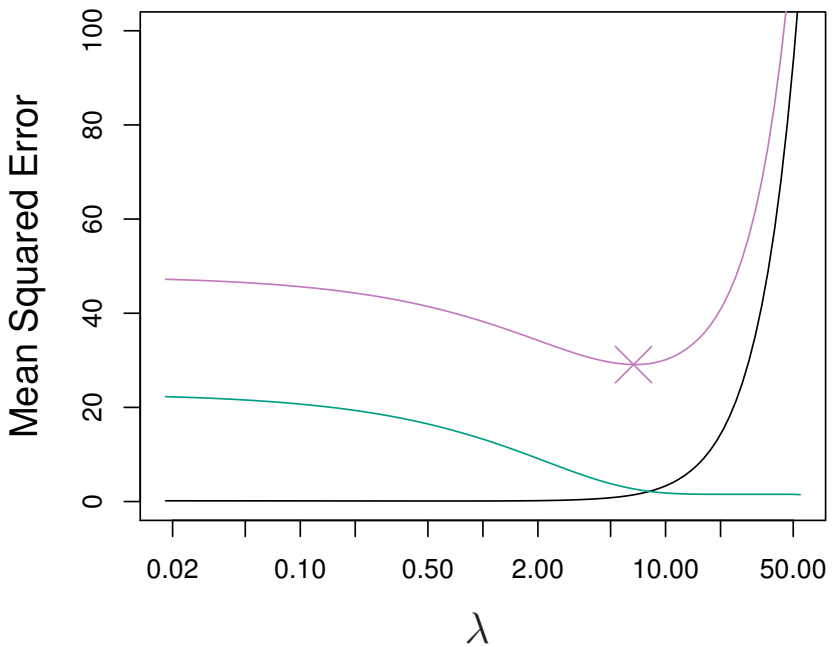



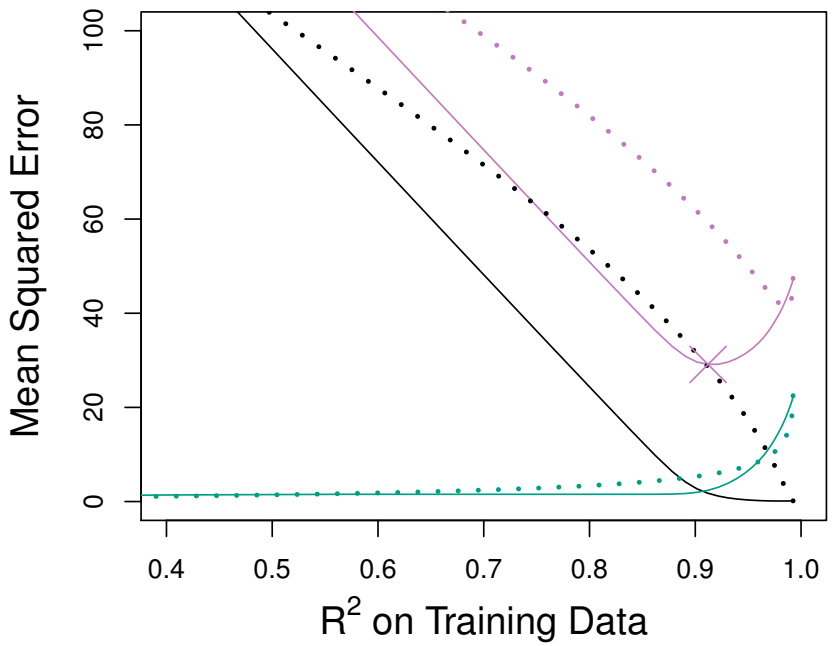

[11] Source [B.1]

A. Giovanidis 2021

# Ridge $<$ Lasso

Here: True response is a function of only 2 predictors and the rest are irrelevant. Ridge needs again all 45 coefficients $\neq 0$. Lasso chose 2-out-of-45 features.[12]




[12] Source [B.1]

A. Giovanidis 2021

# Special Case $n = p$

☞ Data centred around $\bar{x}$, no need for intercept.

- Least Squares: min $\sum_{j=1}^{p}(y_j - \beta_j)^2$. Solution:

$$\hat{\beta}_j \quad = \quad y_j.$$

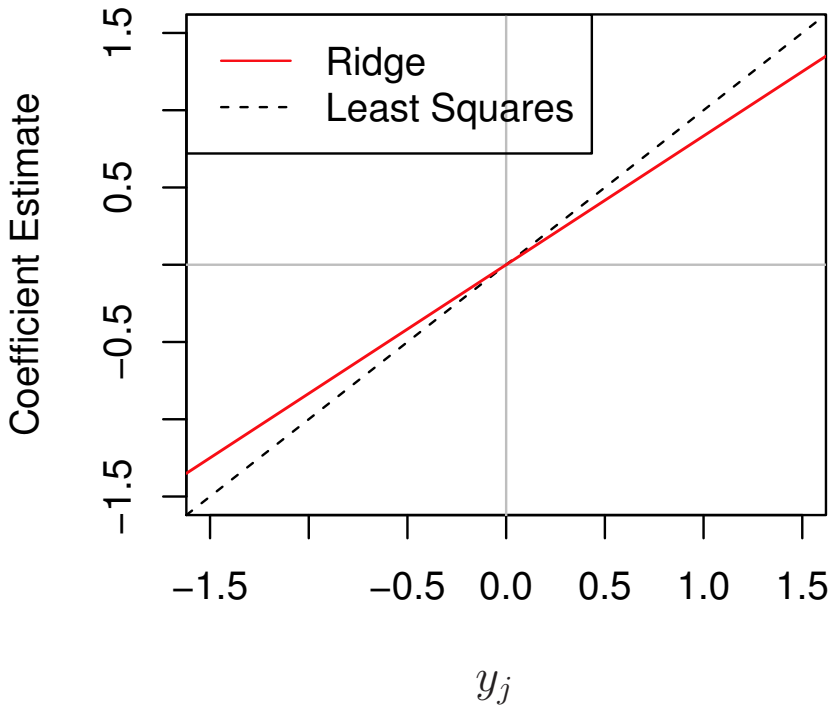
- Ridge Regression: min $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$. Solution:

$$\hat{\beta}_j^{(R)} \quad = \quad y_j/(1+\lambda).$$

- Lasso: min $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$. Solution:

$$\hat{\beta}_j^{(L)} \quad = \quad \begin{cases} y_j - \lambda/2, & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2, & \text{if } y_j < -\lambda/2 \\ 0, & \text{if } |y_j| \leq \lambda/2 \end{cases}$$

A. Giovanidis 2021



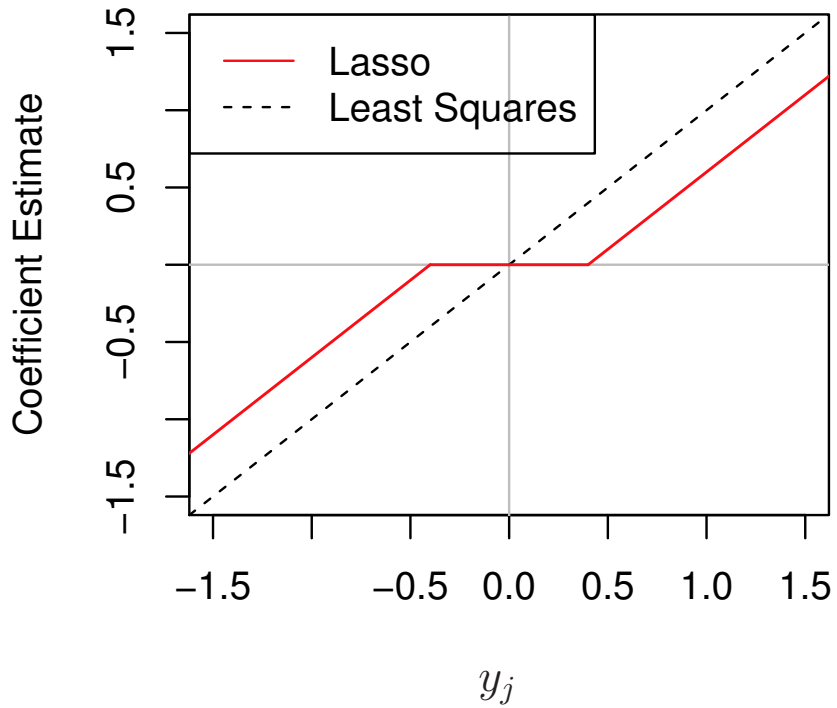




Figure: Ridge and Lasso coefficients over $\lambda$, compared to LSs.[13]

[13] Source [B.1]

A. Giovanidis 2021

# How to select parameter $\lambda$?

For both Ridge and Lasso the tuning parameter $\lambda$ (equivalently $s$) needs to be determined.

☞ Again find the minimum Test MSE using Cross-Validation!

- Choose a grid of $\lambda$ values.
- Compute the cross-validation error for each value of $\lambda$.
- Select the tuning parameter value with minimum CV error.
- Finally, re-fit the model using all observations and the chosen $\lambda$.

A. Giovanidis 2021

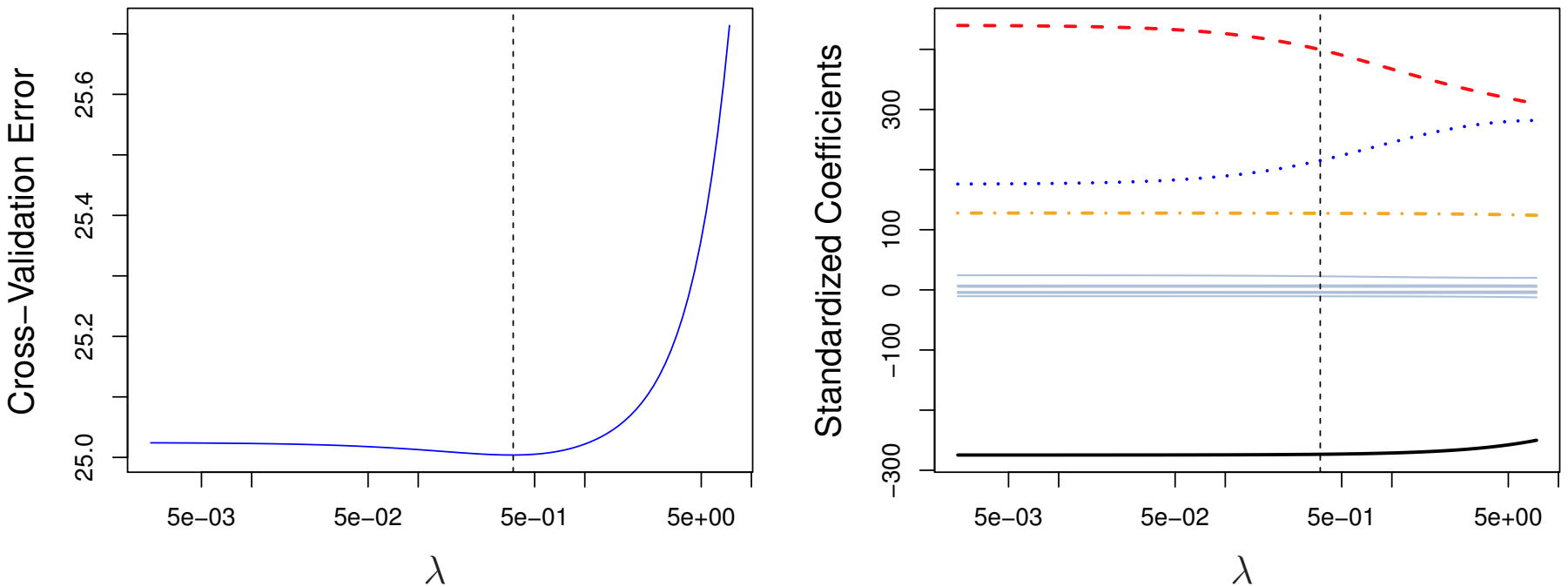



Figure: Ridge parameter tuning and comparison with LSs.[14]

[14] Source [B.1]

A. Giovanidis 2021

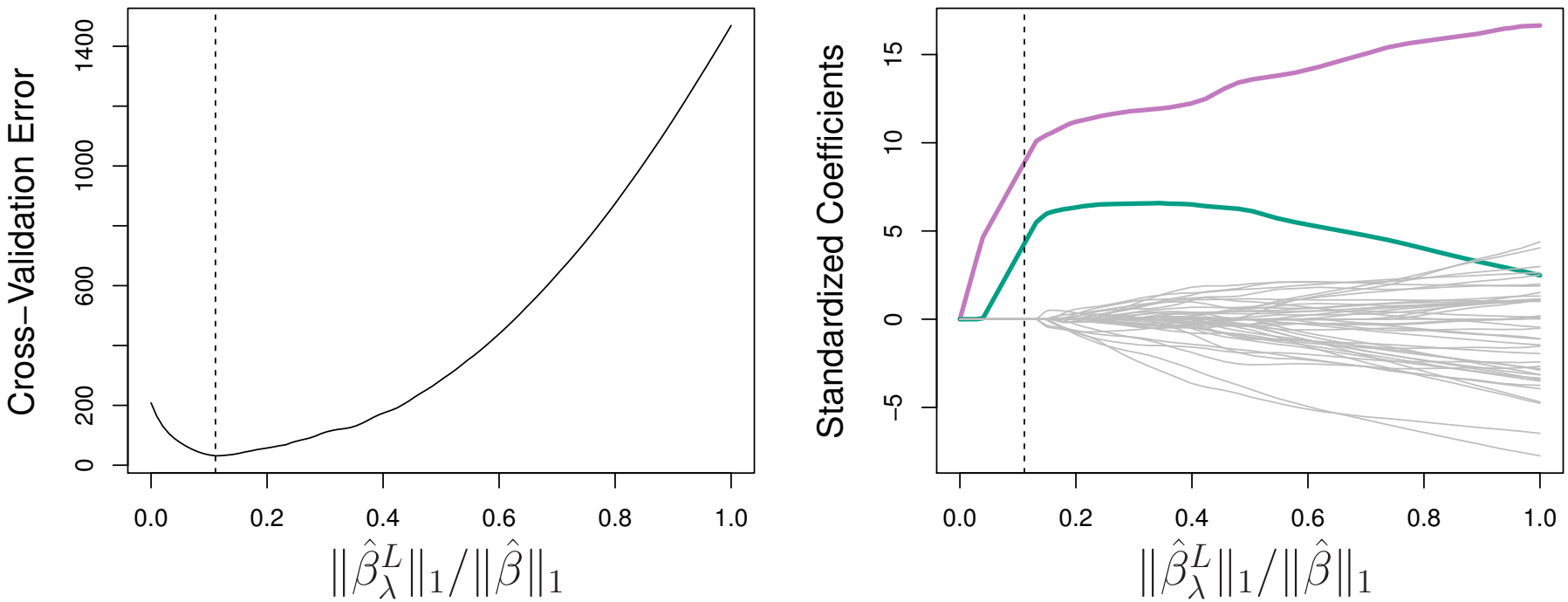



Figure: Lasso parameter tuning and comparison with LSs.[15]

[15] Source [B.1]

# END