A. Giovanidis 2019

09. Feature Selection

Data Analysis for Networks - DataNets'19

Anastasios Giovanidis

Sorbonne-LIP6

December 4, 2019

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 .

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, Regularization, and Dimensionality Reduction.

Main idea

A. Giovanidis 2019

☞ 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.

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

We will present three 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.
- Dimension Reduction: This method projects the p predictors to an $M < p$ dimensional space, through M different linear combinations. Then apply LSs fit.

Best subset selection

A. Giovanidis 2019

To find the best set we need to perform LSs for all possible combinations for the p predictors:

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

In total 2^p possibilities.

A. Giovanidis 2019

[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 2019

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 2019

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 2019

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 2019

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 2019

[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 2019

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 2019

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 2019

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 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 2019

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 σ^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 2019

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 2019

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 2019

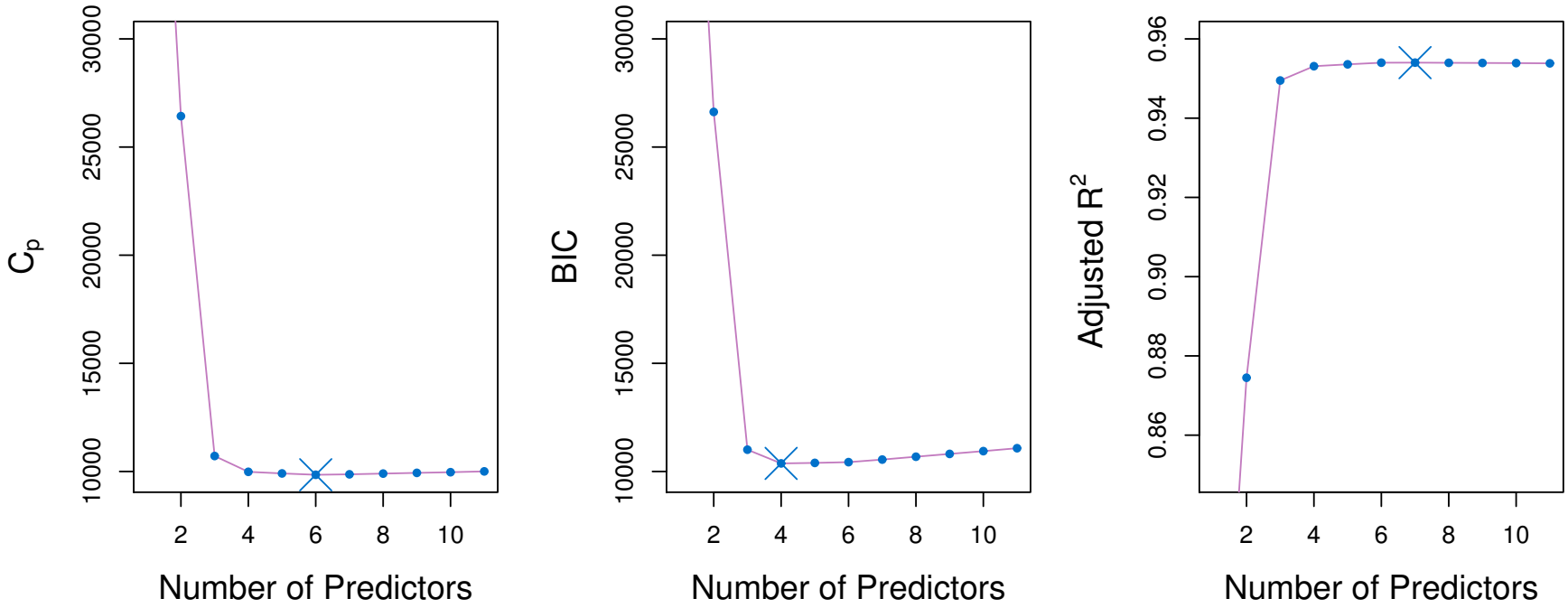


Figure: Feature selection from different metrics.[5]

[5] Source [B.1]

Comparison with Validation and CV tests

A. Giovanidis 2019

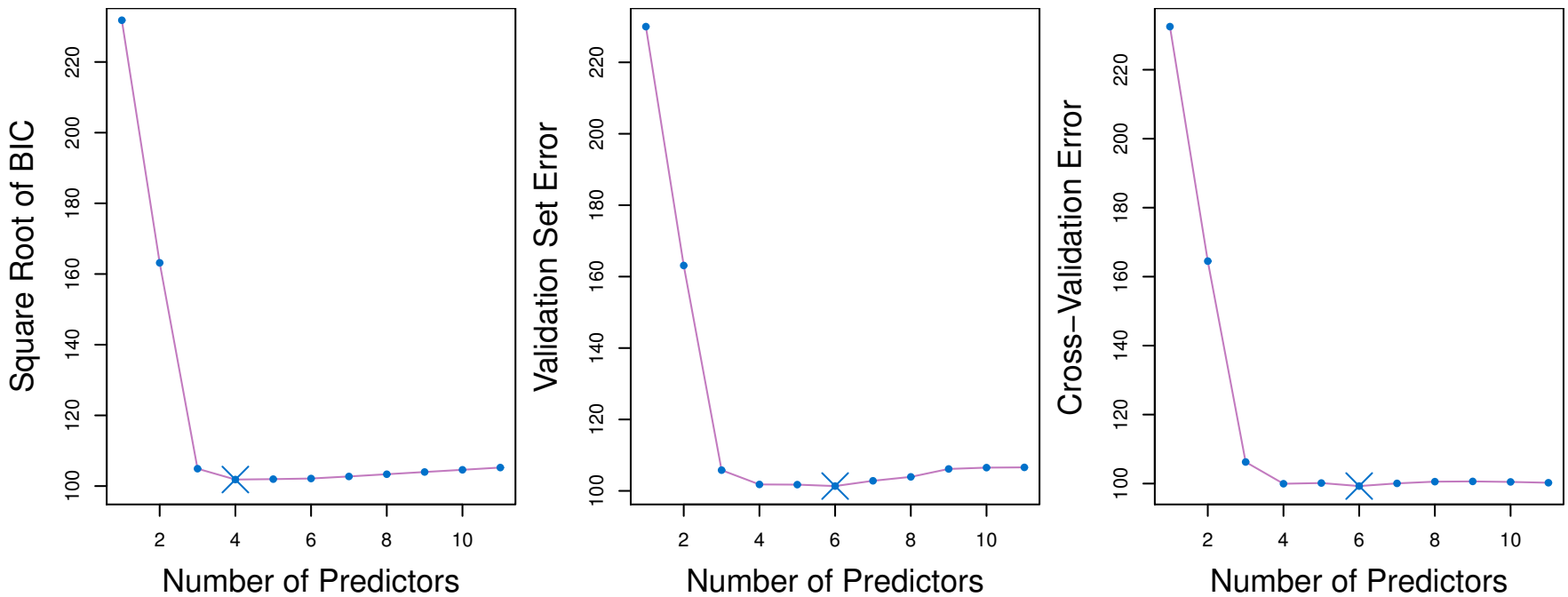


Figure: CV used to be computationally expensive, not any more.[6]

[6] Source [B.1]

Shrinkage

A. Giovanidis 2019

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 2019

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 2019

Properties

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

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

Note 1: The shrinkage penalty is **not** applied to the intercept β_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 2019

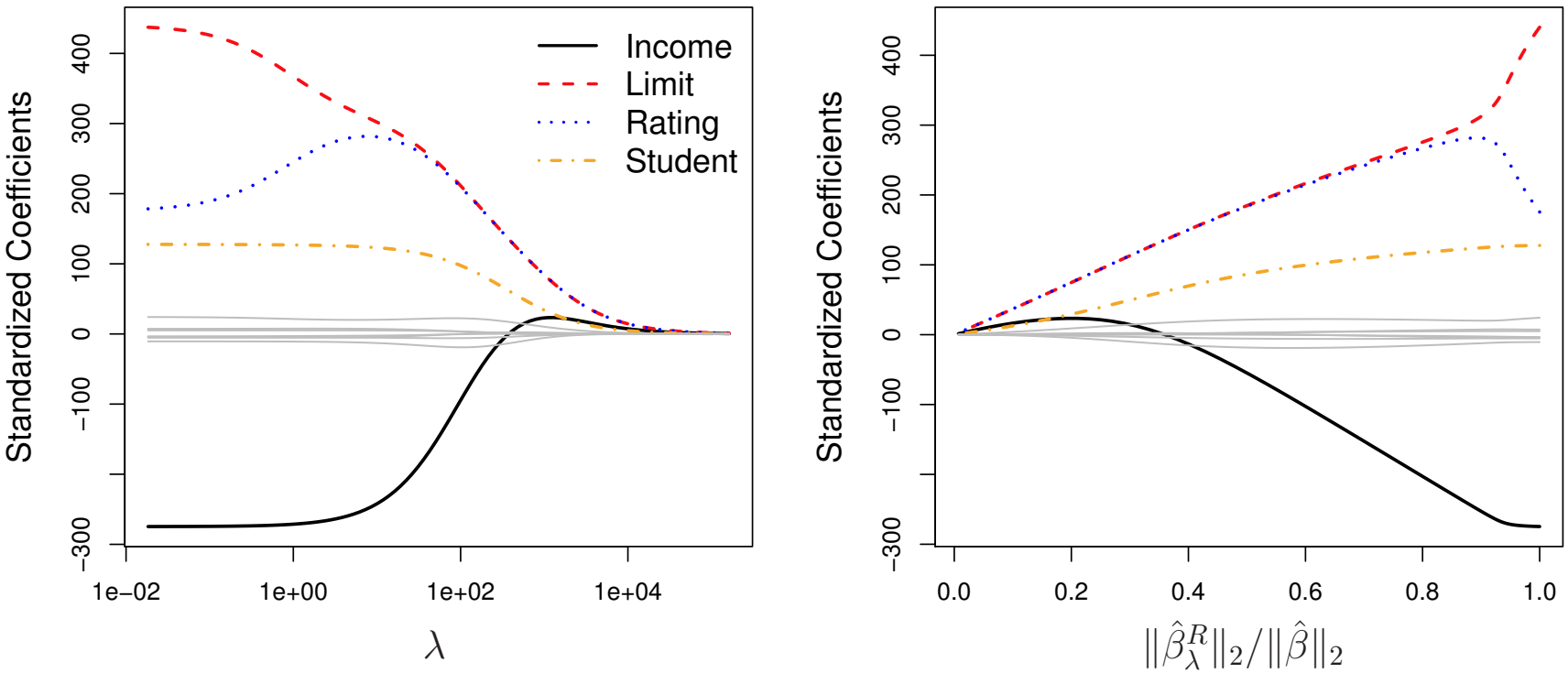


Figure: Change of Ridge Regression coefficients vs λ.[7]

[7] Source [B.1]

A. Giovanidis 2019

Improvement over LSs

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

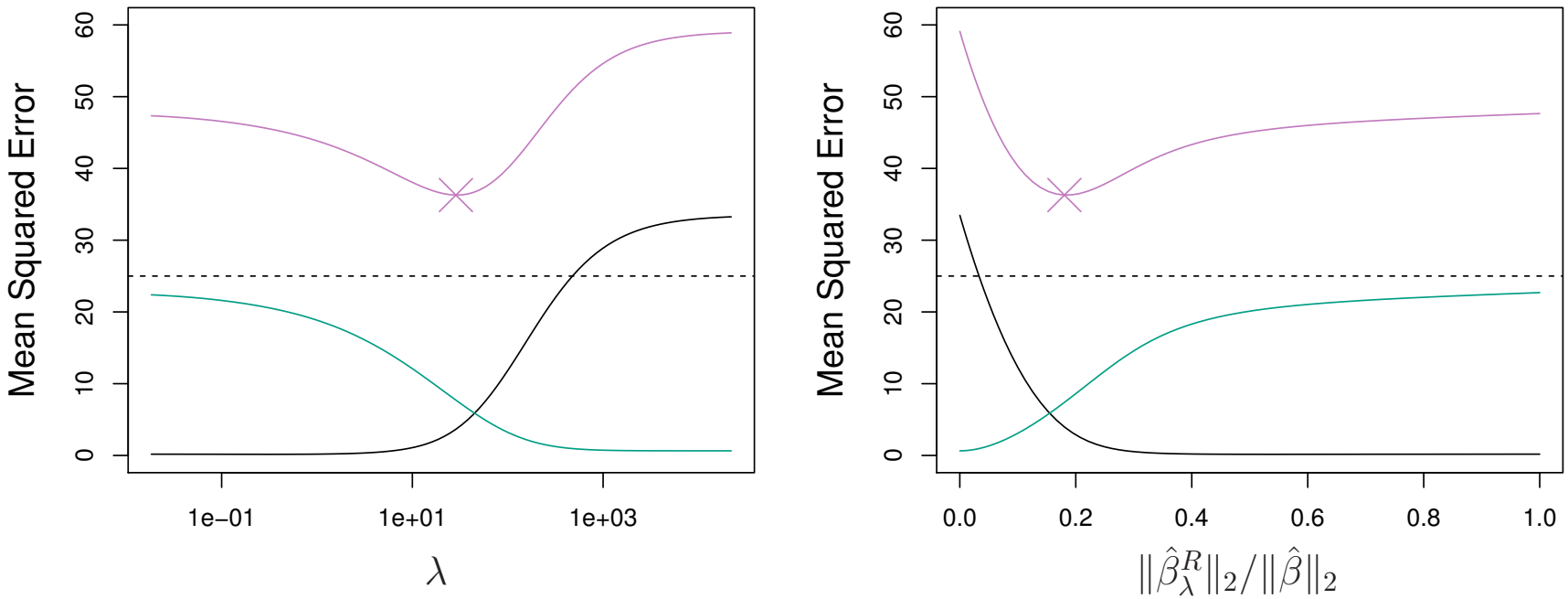


Figure: Bias vs Variance tradeoff and Test MSE.[8]

[8] Source [B.1]

The Lasso

A. Giovanidis 2019

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 ℓ_1-norm instead of ℓ_2).

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 2019

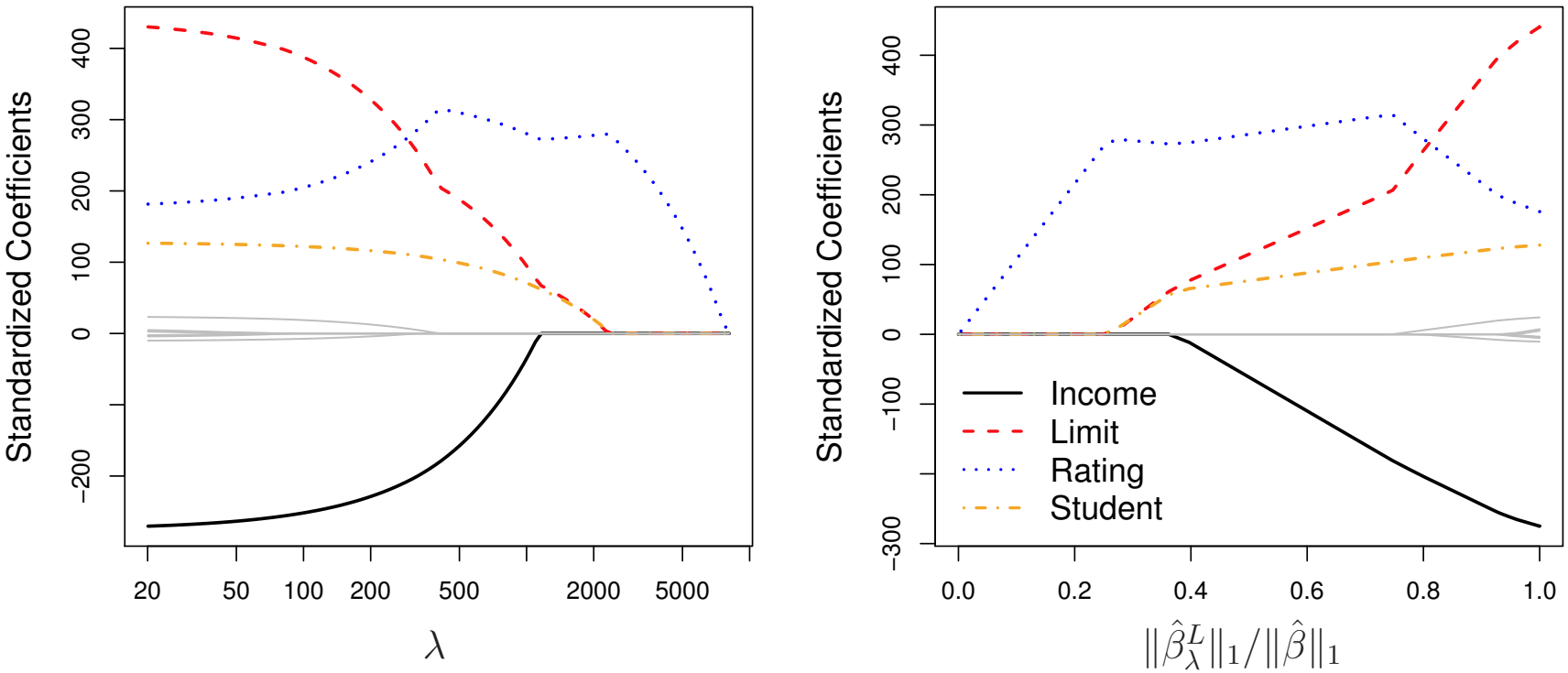


Figure: Change of Lasso coefficients vs λ.[9]

[9] Source [B.1]

A. Giovanidis 2019

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 2019

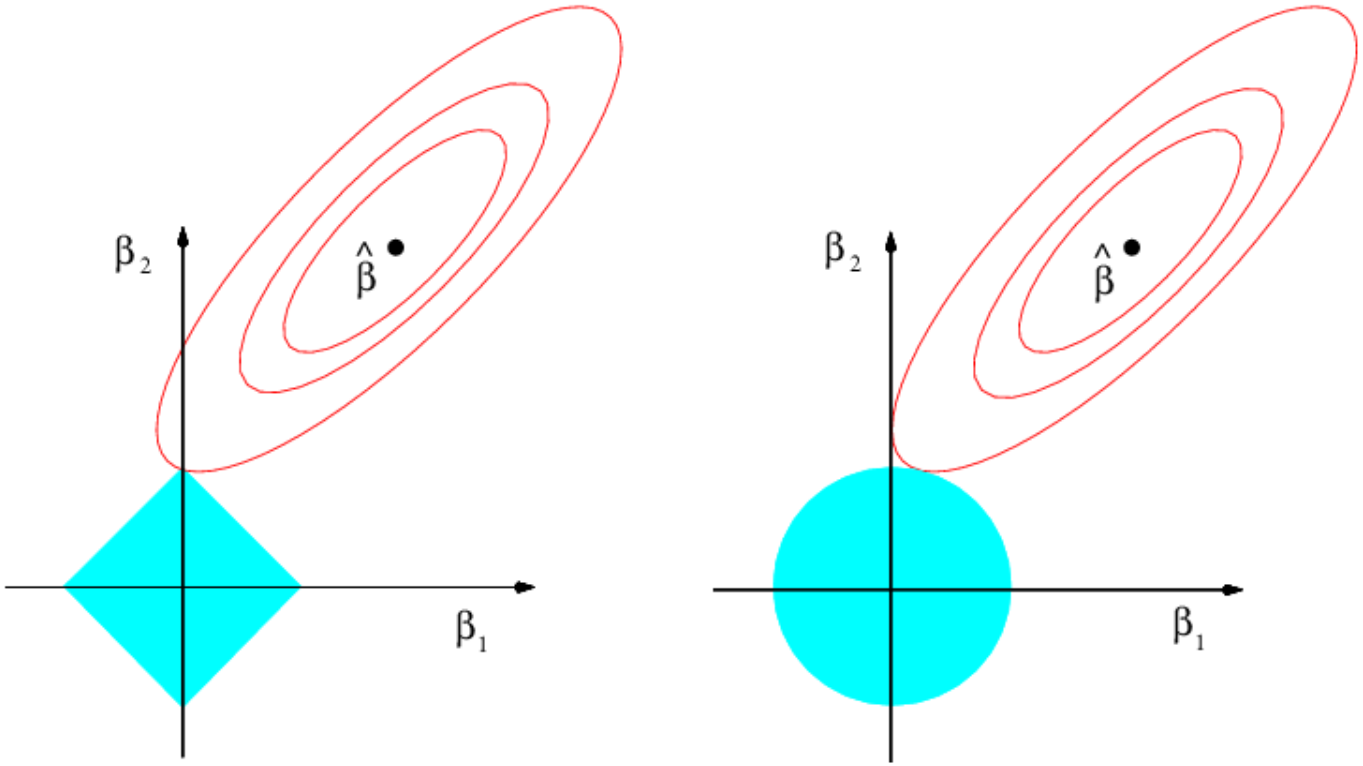


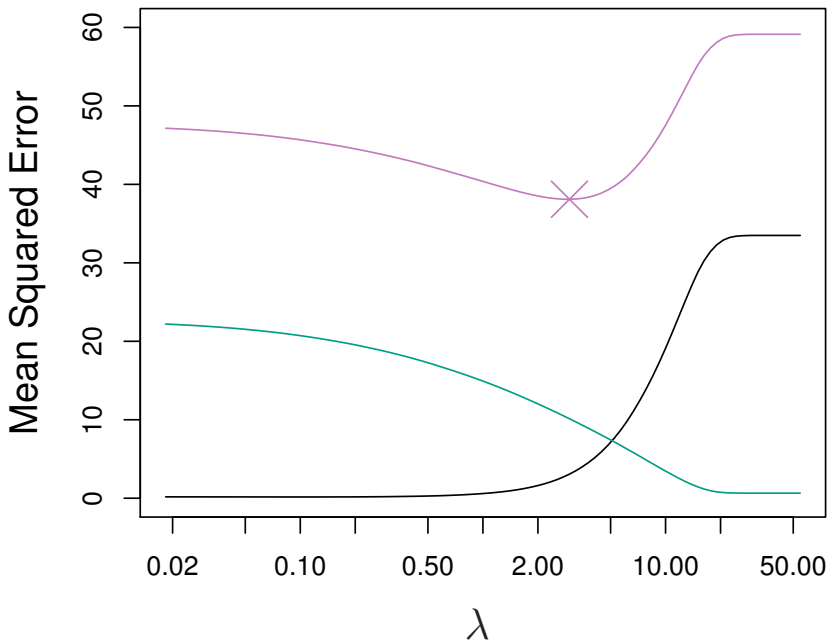
Figure: Why does Lasso lead to estimates equal to 0?[10]

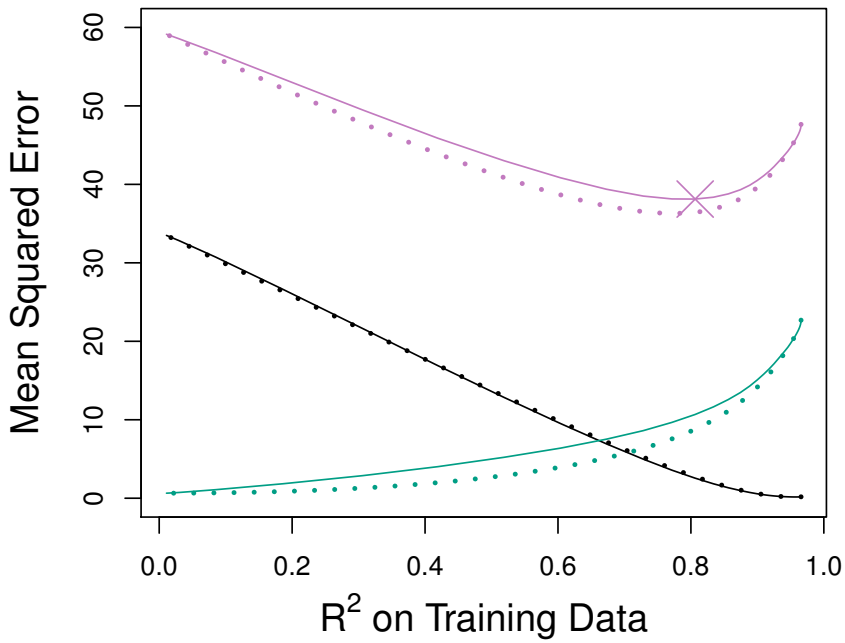
[10] Source [B.1]

Ridge > Lasso

A. Giovanidis 2019

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



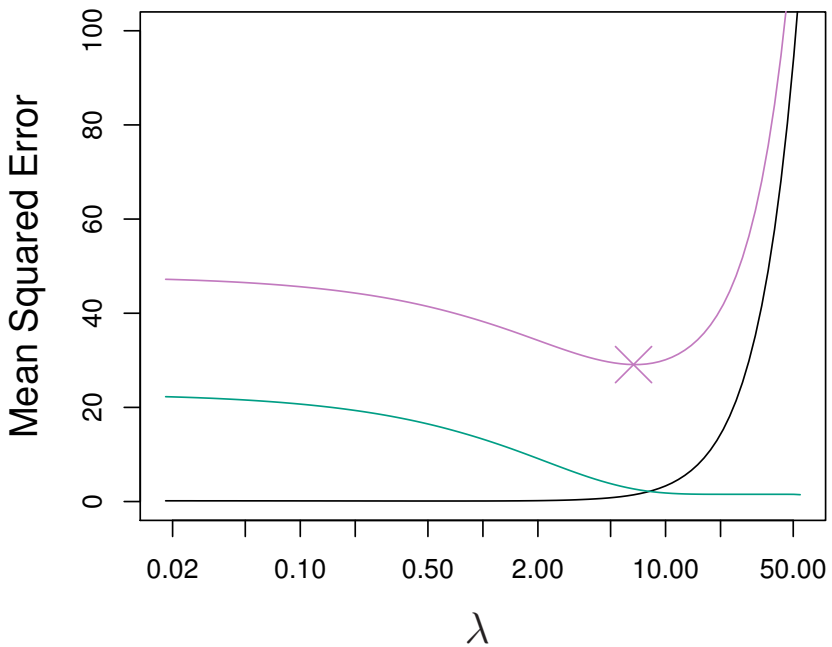


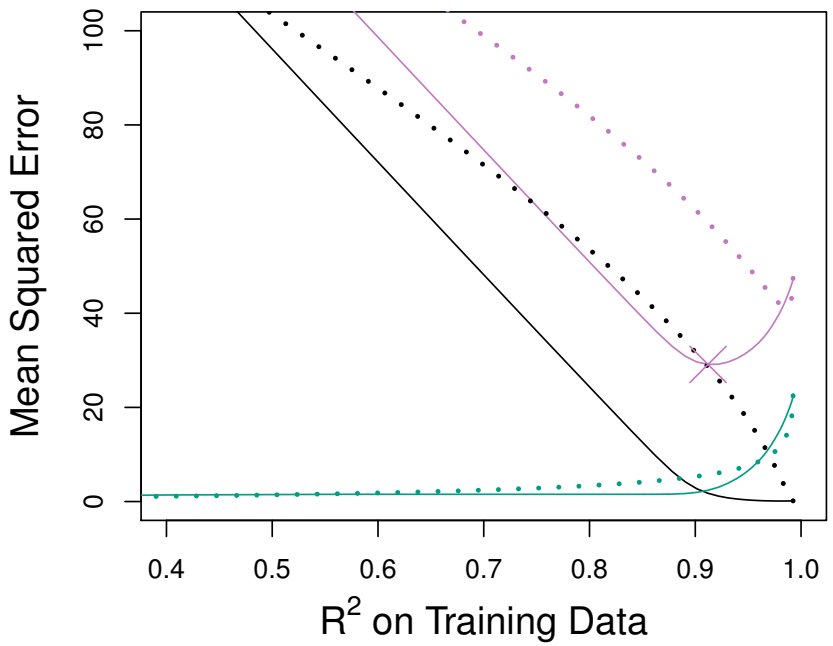
[11] Source [B.1]

Ridge $<$ Lasso

A. Giovanidis 2019

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 2019

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 2019

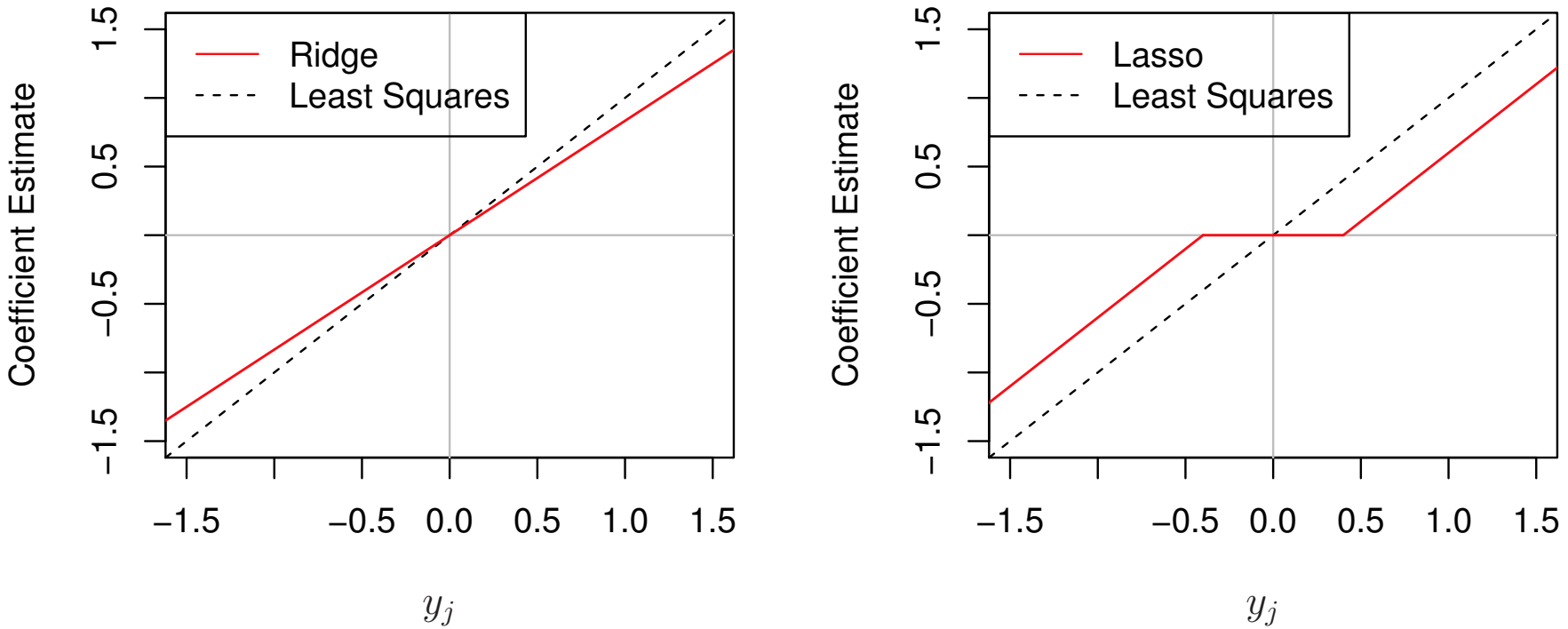


Figure: Ridge and Lasso coefficients over λ, compared to LSs.[13]

[13] Source [B.1]

A. Giovanidis 2019

How to select parameter λ?

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

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

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

A. Giovanidis 2019

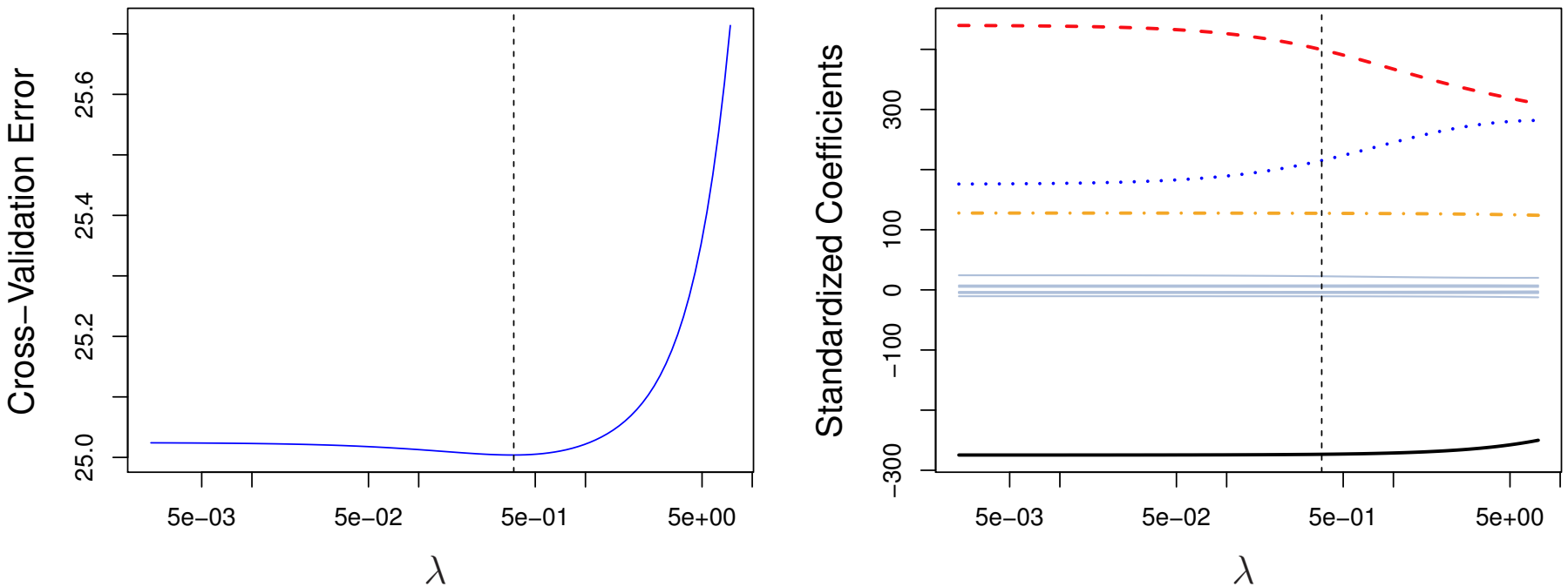


Figure: Ridge parameter tuning and comparison with LSs.[14]

[14] Source [B.1]

A. Giovanidis 2019

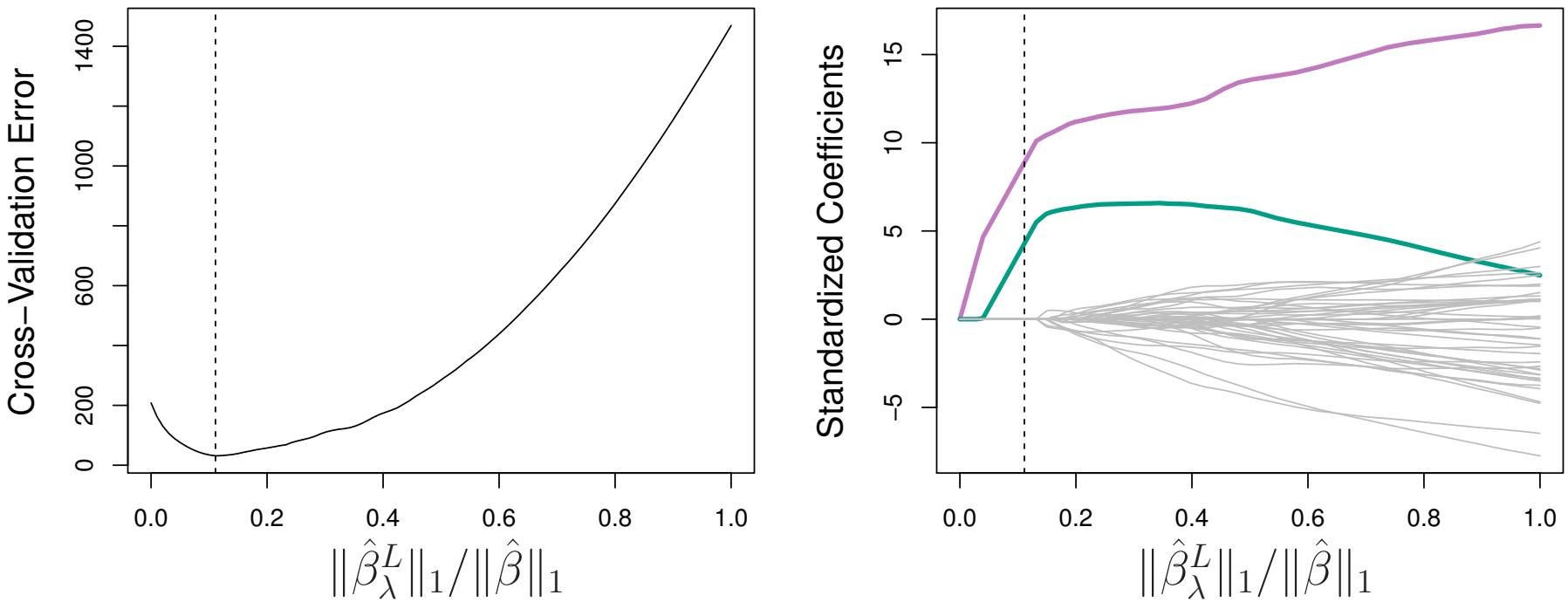


Figure: Lasso parameter tuning and comparison with LSs.[15]

[15] Source [B.1]

Dimension Reduction

A. Giovanidis 2019

Until here, methods control variance in 2 ways. From the original $X_1, \ldots, X_p$ predictors,

- either choose a subset of the original variables, or
- shrink some of their coefficients to zero.

☞ A new method!

1. first, Transform the predictors to a lower dimension $Z_1, \ldots, Z_M$, with $M < p$,
2. then Fit the LS model using the M predictors.

Dimension Reduction cont'd

A. Giovanidis 2019

$Z_1, Z_2, \ldots, Z_M$: linear combinations ($M \leq p$) of our original p predictors.

$$Z_m = \sum_{j=1}^{p} \phi_{j,m} X_j.$$

Unknowns: $\phi_{1,m}, \ldots, \phi_{p,m}$, for $m = 1, \ldots, M$

Then we can fit the linear regression model:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{i,m} + \epsilon_i, \quad i = 1, \ldots, n$$

using the Least Squares. The method can outperform simple LSs!

Dimension Reduction cont'd II

A. Giovanidis 2019

We need not estimate the $p+1$ coefficients: $\beta_0, \beta_1, \ldots, \beta_p$, but rather the less $M+1$ coefficients: $\theta_0, \theta_1, \ldots, \theta_M$.

$$\sum_{m=1}^{M} \theta_m z_{i,m} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{j,m} x_{i,j}$$

$$= \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{j,m} x_{i,j} = \sum_{j=1}^{p} \beta_j x_{i,j}.$$

so that, the β_j's need to take the form:

$$\beta_j := \sum_{m=1}^{M} \theta_m \phi_{j,m}.$$

☞ Works well when p is large compared to n data available.

Partial Least Squares (PLS)

A. Giovanidis 2019

PLS is a dimension reduction method:

1. First, identifies a new set of features $Z_1, \ldots, Z_M$, $M < p$ in a supervised way,
2. then fits a linear model via LSs using the M features.

Supervised way: means uses the response Y to identify new features, related to the response.

PLS cont'd

A. Giovanidis 2019

PLS steps:

- Standardise and centre the p predictors.
- Z_1: Do LSs for all pairs $(X_j, Y) \rightarrow$ use $\phi_{\mathbf{j,1}} := \beta_{\mathbf{j,1}}$, $i = 1, \ldots, p$.

  In $Z_1 = \sum_{j=1}^{p} \phi_{j,1} X_1$, PLS places the highest weight on the variables most strongly related to the response Y.
- Z_2: Do LSs for all pairs (X_j, Z_1) and find new $X_j^{(2)} = X_j - \gamma_{j,1} Z_1$.
  Then do LSs for all pairs $(X_j^{(2)}, Y) \rightarrow$ use $\phi_{\mathbf{j,1}} := \beta_{\mathbf{j,1}}^{\mathbf{(2)}}$, $i = 1, \ldots, p$.
- Z_3: ...

The PLS seeks directions that have high variance and high correlation with the response.

Principal Components Regression (PCR)

A. Giovanidis 2019

Steps of the PCR Method:

1. Standardise $X_j/\sqrt{Var(X_j)}$ and then centre $(X_j - \bar{X}_j)$ all p features.
2. Choose a number $M \leq p$ for the components.
3. Using Principal Component Analysis (PCA) define $Z_1, \ldots, Z_M$ using the coefficients $\phi_{j,m}$, $j = 1, \ldots, p$.
4. Use $Z_1, \ldots, Z_M$ as the predictors in a linear regression model that is fit using LSs.
5. Using different number of components M choose the model which minimises the Test MSE.

☞ The PCR is not a feature selection method!
(uses linear combinations of all p original features)

PCR example

A. Giovanidis 2019

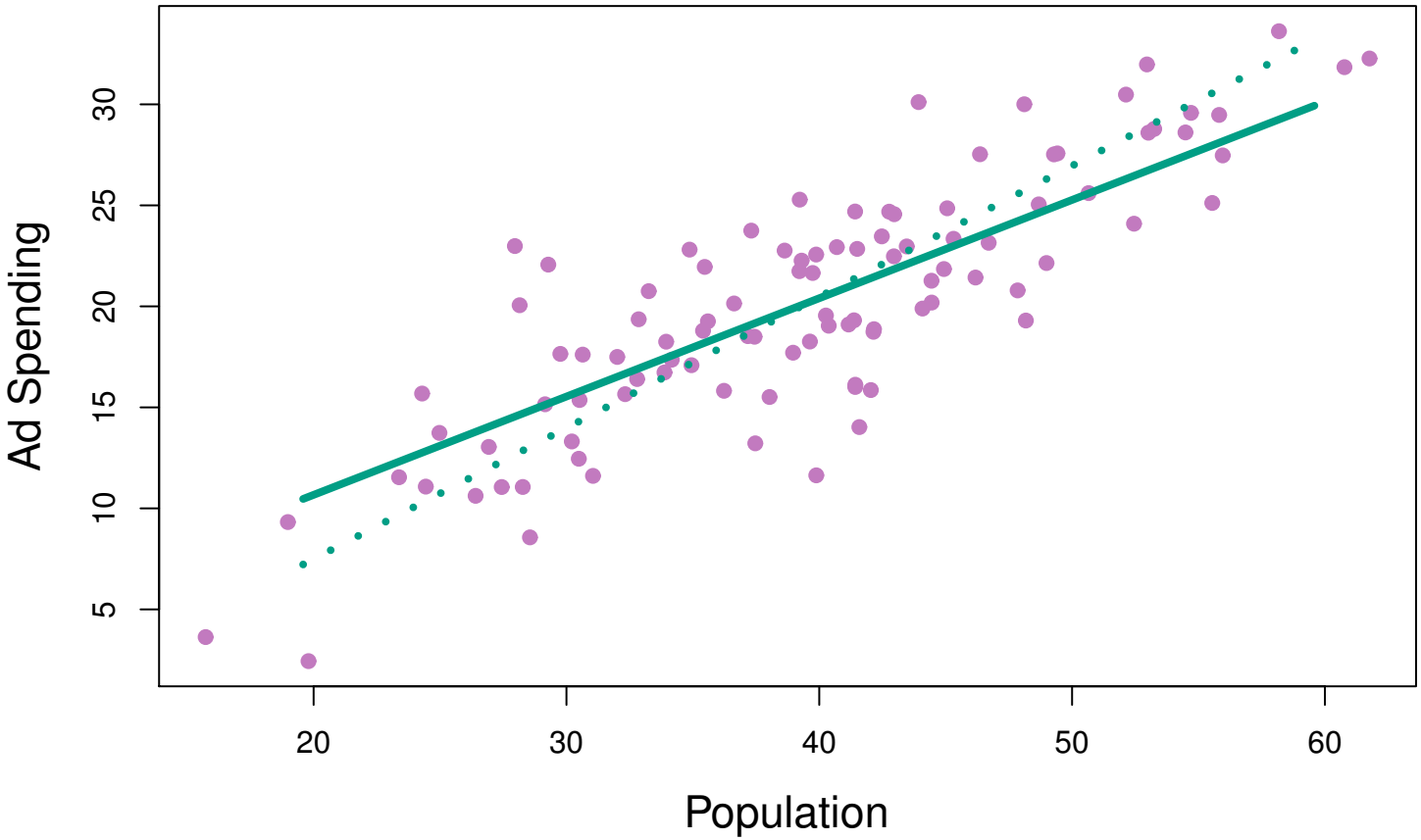


Figure: PCR direction Z_1 for $p = 2$.[16]

[16] Source [B.1]

What is Principal Components Analysis?

A. Giovanidis 2019

PCA is a technique to reduce the dimension of a $(n \times p)$ data matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U}$ is $(n \times p)$, $\mathbf{V}$ is $(p \times p)$ and $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$, called the singular values of $\mathbf{X}$.

☞ Intuition: the right matrix $\mathbf{V}$ and the square of the singular value matrix $\mathbf{D}^2$ also determine the eigen-decomposition of the sample covariance matrix $\mathbf{S} = \mathbf{X}^T\mathbf{X}/n$,

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)^T(\mathbf{U}\mathbf{D}\mathbf{V}^T) \\ &= \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T \\ &\overset{\mathbf{U}^T\mathbf{U}=\mathbf{I}_p}{=} \mathbf{V}\mathbf{D}^2\mathbf{V}^T \end{aligned}$$

PCA cont'd

A. Giovanidis 2019

The principal components directions of $\mathbf{X}$ are the columns of $\mathbf{V}$.

☞ The first principal component direction has the property that $\mathbf{Z_1} = \mathbf{X}v_1$ has the largest sample variance, equal to d_1^2/n.

- This practically means that along the direction v_1, the data vary the most (if projected).
- Also, this direction defines the line which is as close as possible to the data!

PCA Example

A. Giovanidis 2019

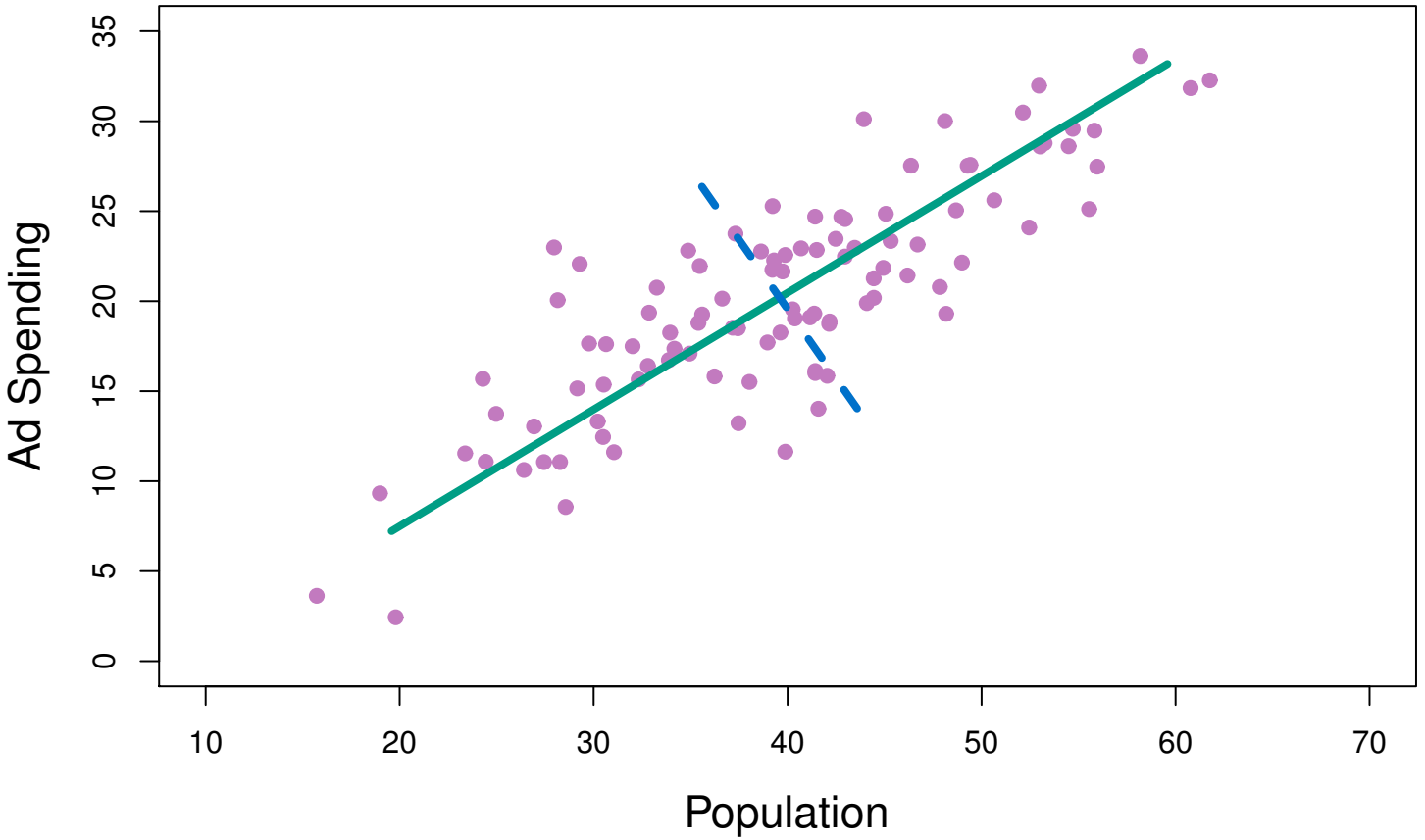


Figure: Principal components directions for $p = 2$.[17]

[17] Source [B.1]

A. Giovanidis 2019

PCA Example cont'd

The two PCA directions are:

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad})$$

$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

- With $p = 2$, we can construct at most $M = 2$ linear combinations.
- The directions $(0.839,\ 0.544)$ and $(0.544,\ -0.839)$ are orthogonal!
- For the Regression using PCA, we need to choose the sets $\{Z_1\}$ and $\{Z_1,\ Z_2\}$ and find which one leads to smallest Test MSE.

PCR vs Ridge and Lasso

A. Giovanidis 2019

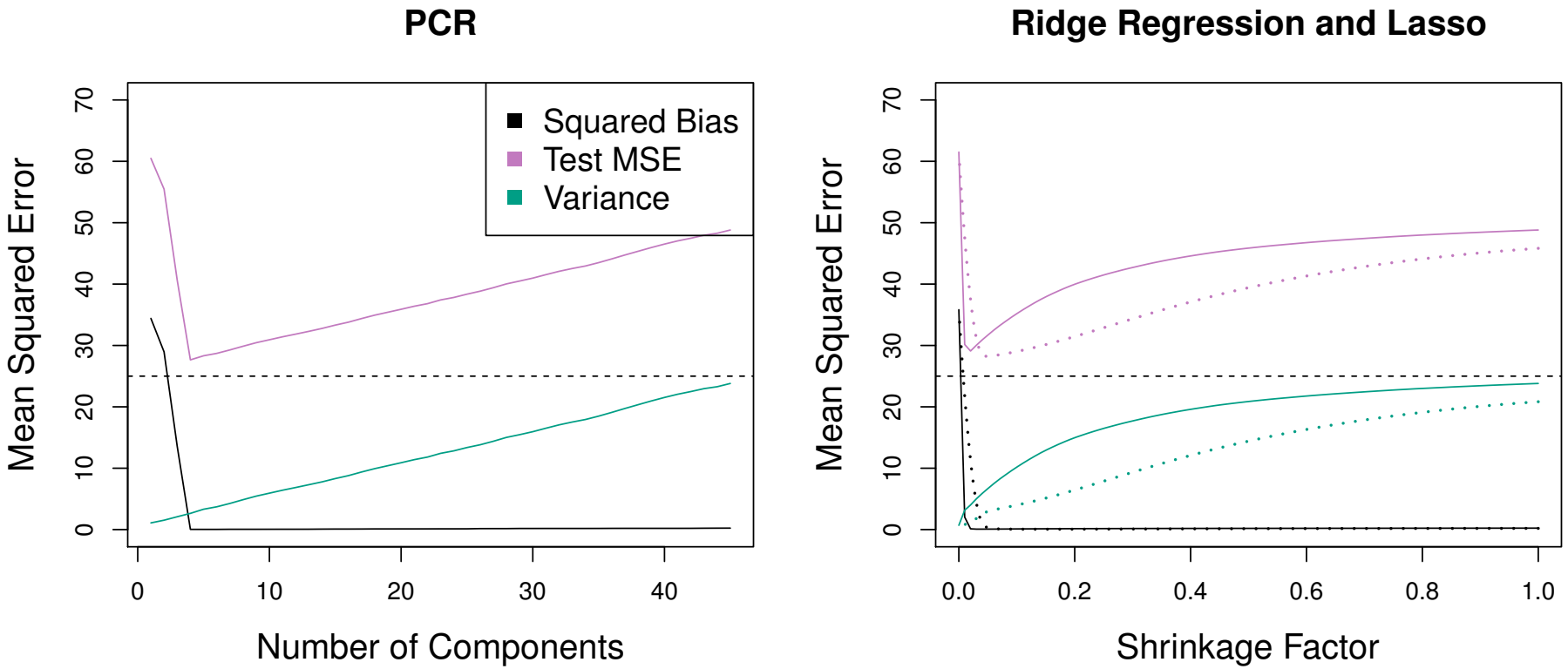


Figure: $p = 45$ features: PCR vs Lasso (solid) and Ridge (dotted).[18]

[18] Source [B.1]

A. Giovanidis 2019

END