Linear Model Selection and Regularization

Slides on *Introduction to Statistical Learning*, Chapter 6

Edward Thompson

March 2024

See the Typst source: https://typst.app/project/pzEmcVAtZ9sJA-_Y4vrri0

Table of contents

Introduction

Motivation

- **Linear Regression can be too flexible** at the expense of accuracy and interpretability

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

- This was introduced in chapter 2:
  - Particularly true when n is not much greater than p or you have highly colinear predictors
  - Introduced *forward selection*, *backward selection* and *mixed selection* as solutions
  - **This chapter explores these solutions and more**

- Three basic approaches to reducing flexibility
  - **Subset selection**
  - **Shrinkage**
  - **Dimension Reduction**

Subset Selection

Subset Selection

- **Only include some variables in your final model**
- There are different ways to identify which ones to include:
  - Best Subset Selection
  - Forward Stepwise Selection
  - Backward Stepwise Selection
  - Hybrid Approaches

Comparing models with differing numbers of predictors

- Models with more variables will have a lower $\mathrm{RSS} \Rightarrow$ cannot use RSS to compare
- Adjust the RSS (or R^2) to account for # variables:
  - $\boldsymbol{C_p} = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$
  - $\mathbf{AIC} = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$
  - $\mathbf{BIC} = \frac{1}{n}(\mathrm{RSS} + \log(n)d\hat{\sigma}^2)$
  - $\textbf{Adjusted } \boldsymbol{R^2} = 1 - \frac{\frac{\mathrm{RSS}}{n-d-1}}{\frac{\mathrm{TSS}}{n-1}}$
- **Discussion**: why are they different?

Best Subset Selection

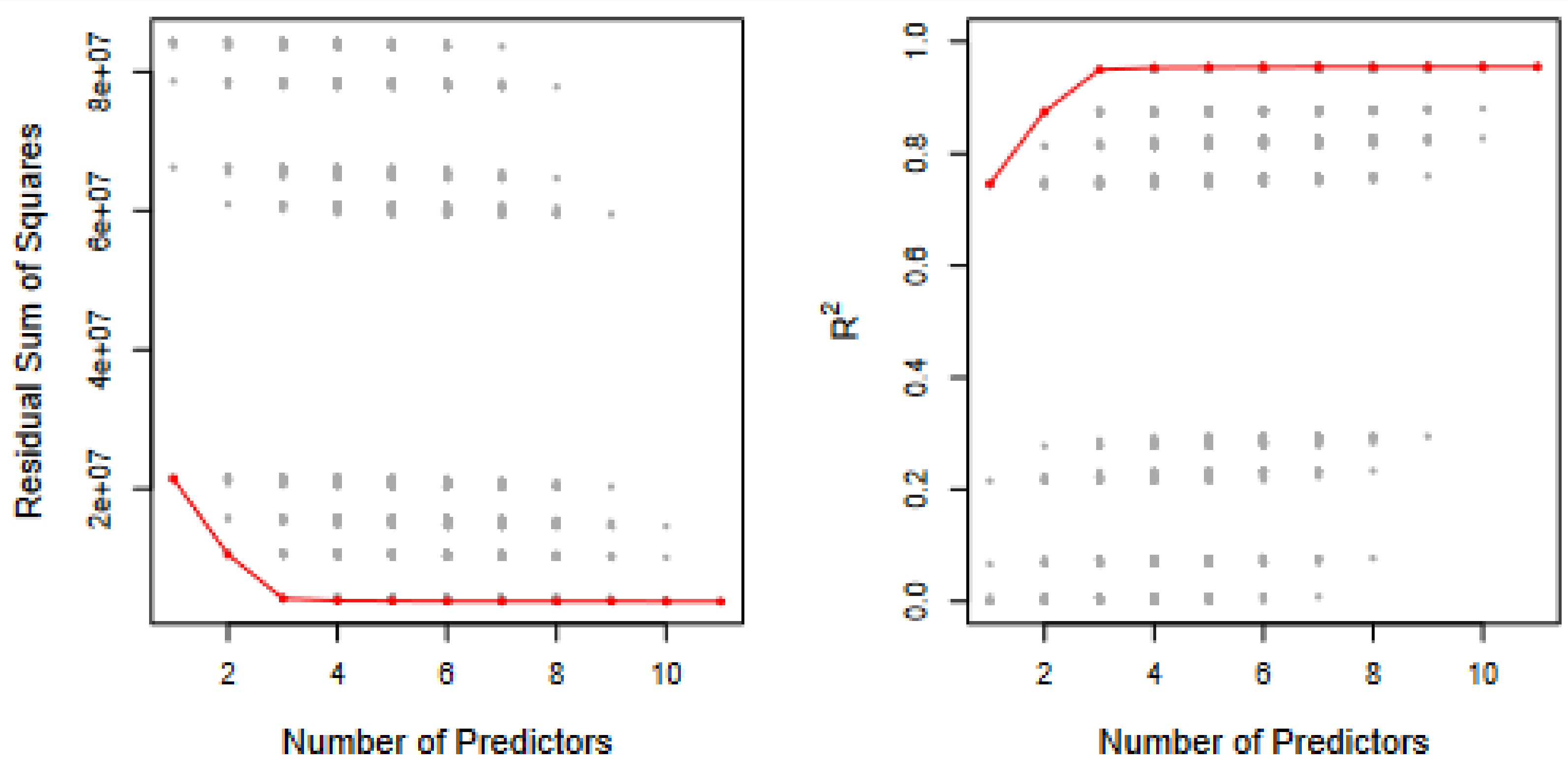
- **Try every model**

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 using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

- If using cross validation repeat step 2 on each training fold and average validation errors in step 3 to choose the best k. Then choose best given k on total training data.

Best Subset Selection Example



- Very slow...

Forward stepwise selection

- **Sequentially add predictors**

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 the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

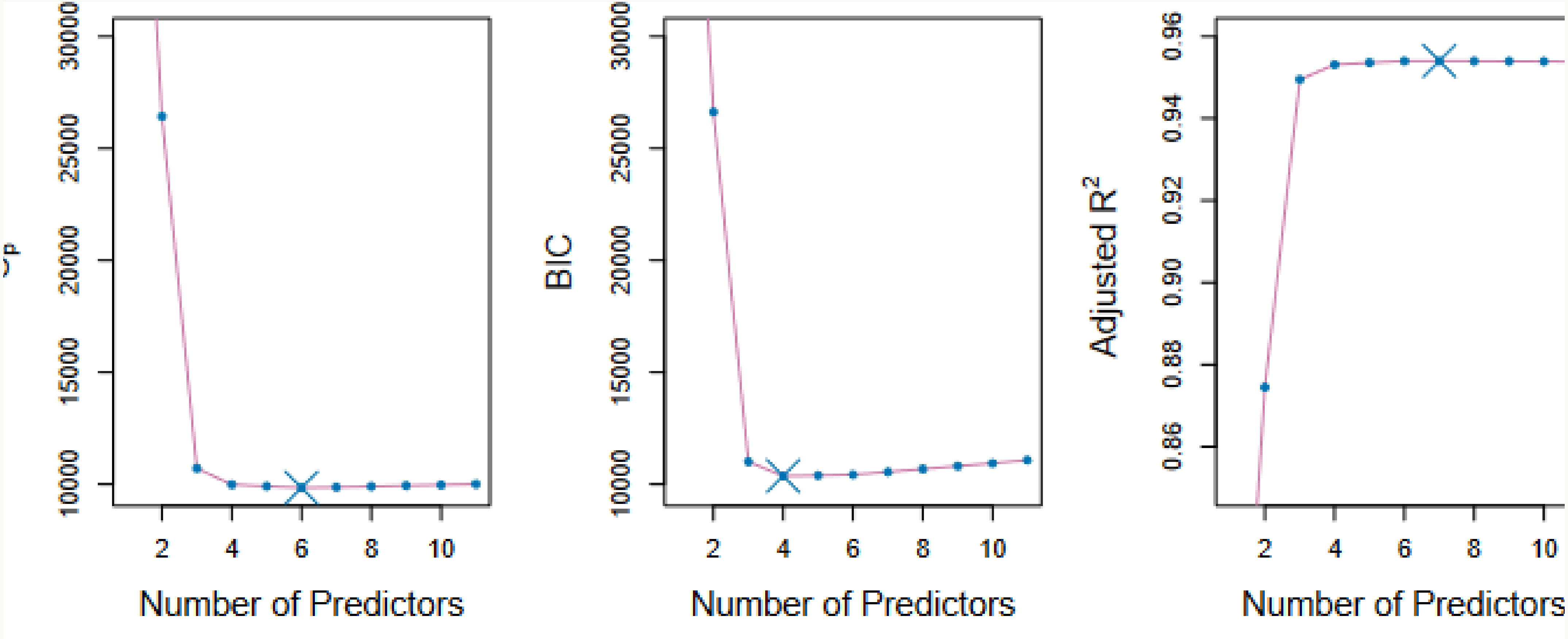
Backward stepwise selection

- **Sequentially remove predictors**

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 the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

Selection example



- Using cross-validation enables using the one-standard-error rule.

Shrinkage

Shrinkage Idea

- In linear regression using least squares we are minimizing:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2$$

- In shrinkage we **add a penalty term** that penalizes the coefficients
- **Ridge regression**

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

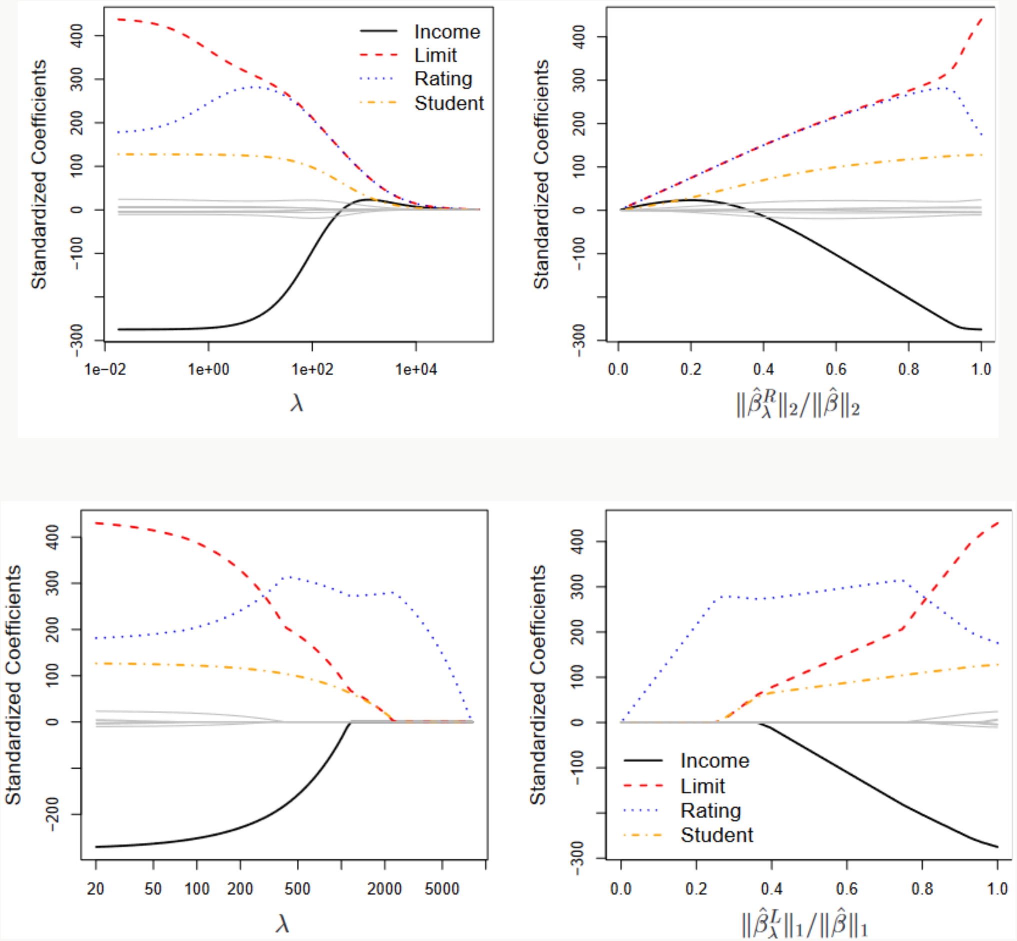
- **LASSO regression**

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

- Need to standardise the predictors:

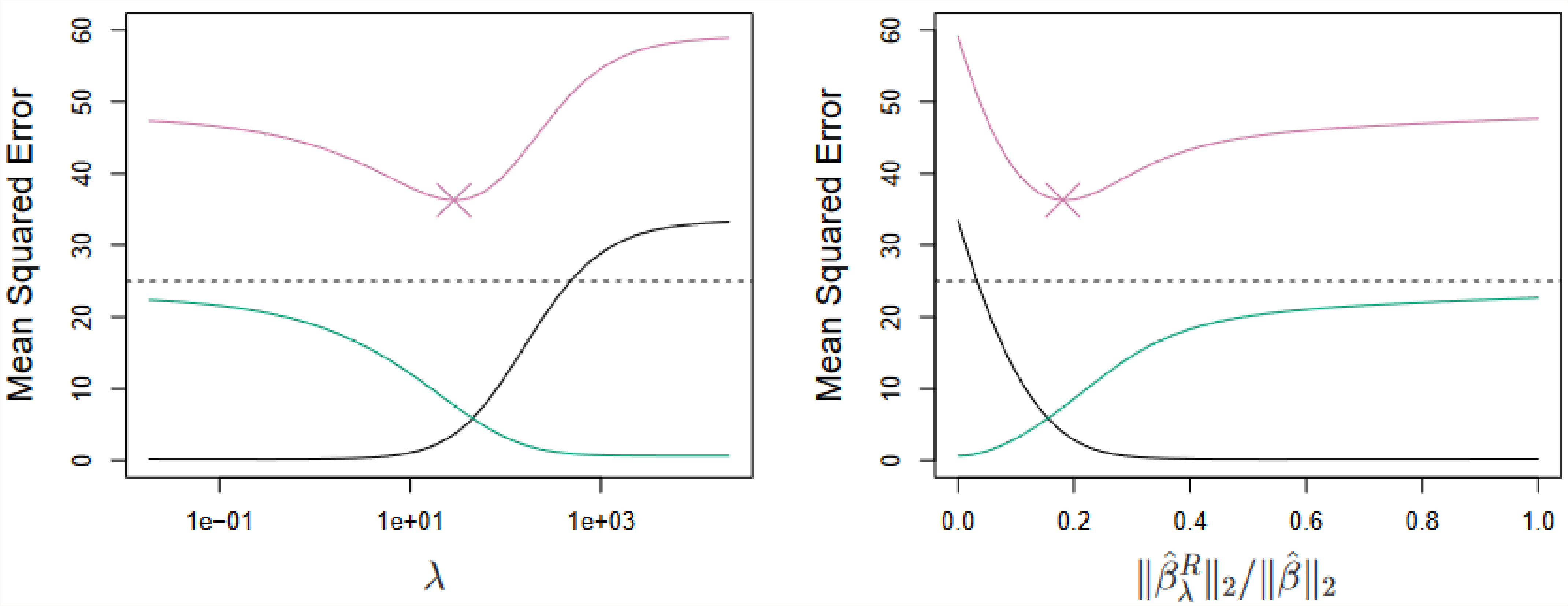
$$\frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2}}$$

Ridge and Lasso Examples



Why does this improve on Least Squares?

- Example on simulated data



Another Perspective

- **Ridge regression**

$$\min_\beta\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_i x_{ij}\right)^2\right\}\text{ subject to}\sum_{j=1}^{n}\beta_j^2\le s$$

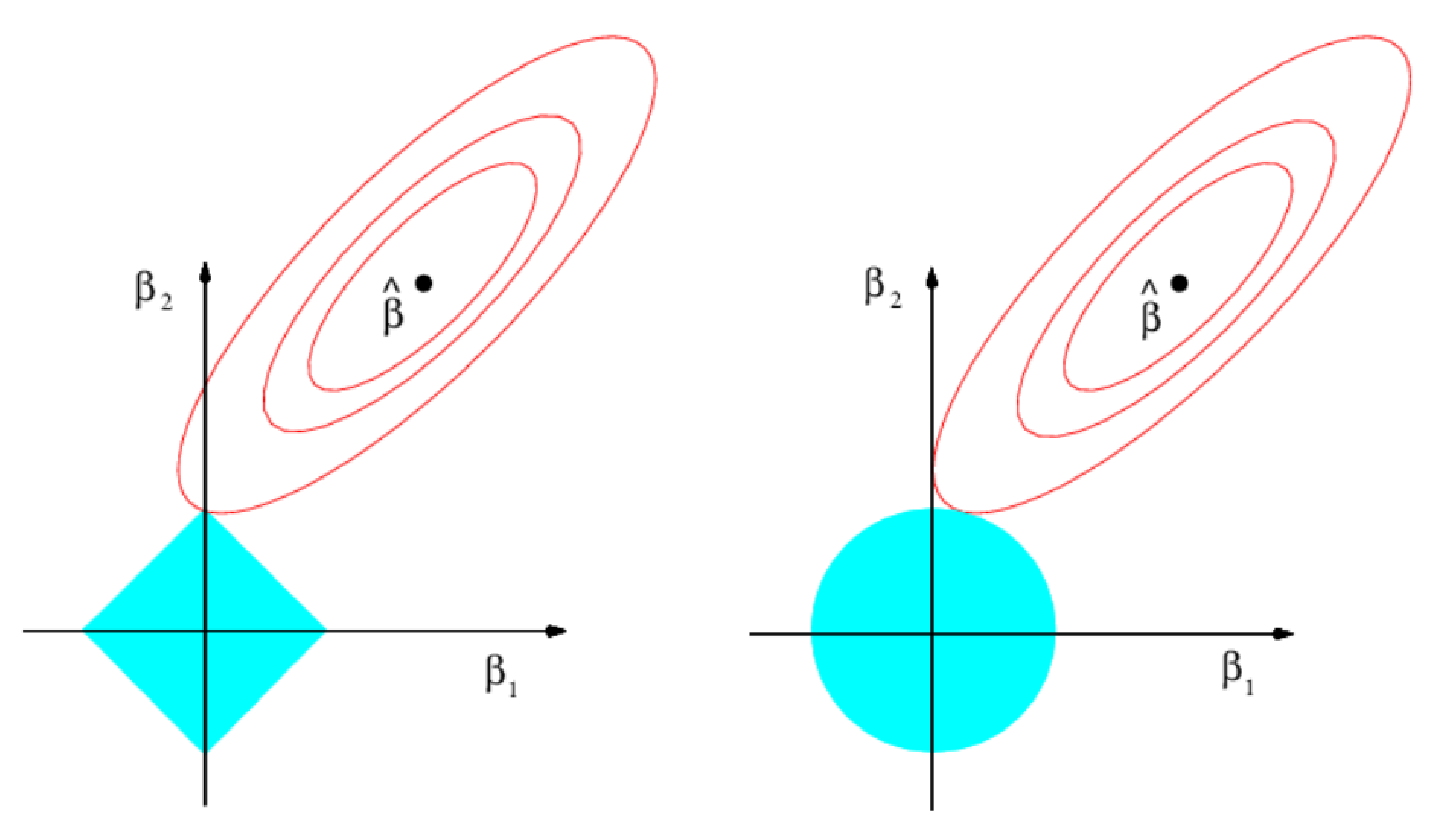
- **LASSO regression**

$$\min_\beta\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_i x_{ij}\right)^2\right\}\text{ subject to}\sum_{j=1}^{n}|\beta_j|\le s$$

- **Best Subset Selection**

$$\min_\beta\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_i x_{ij}\right)^2\right\}\text{ subject to}\sum_{j=1}^{n}I(\beta_j\neq 0)\le s$$

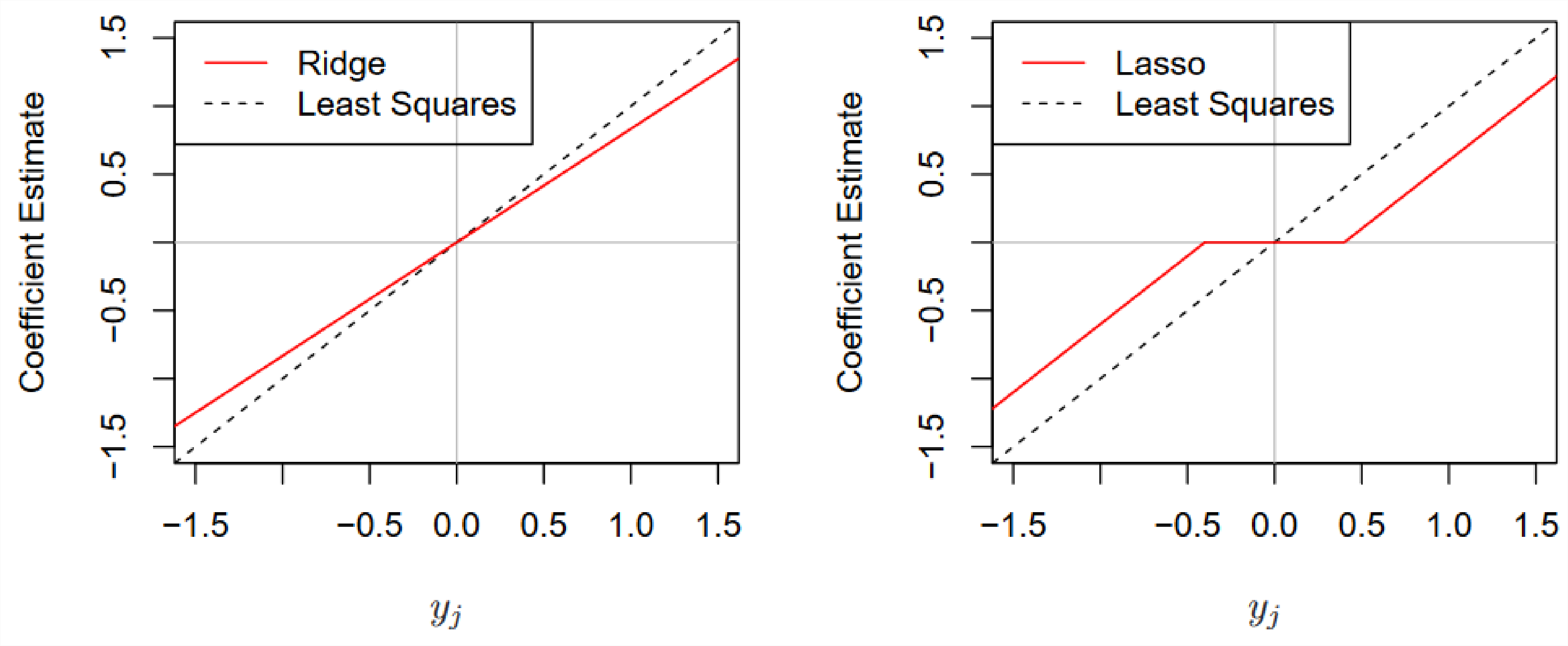
Why does LASSO Lead To Subset Selection?



Why does LASSO Lead To Subset Selection?

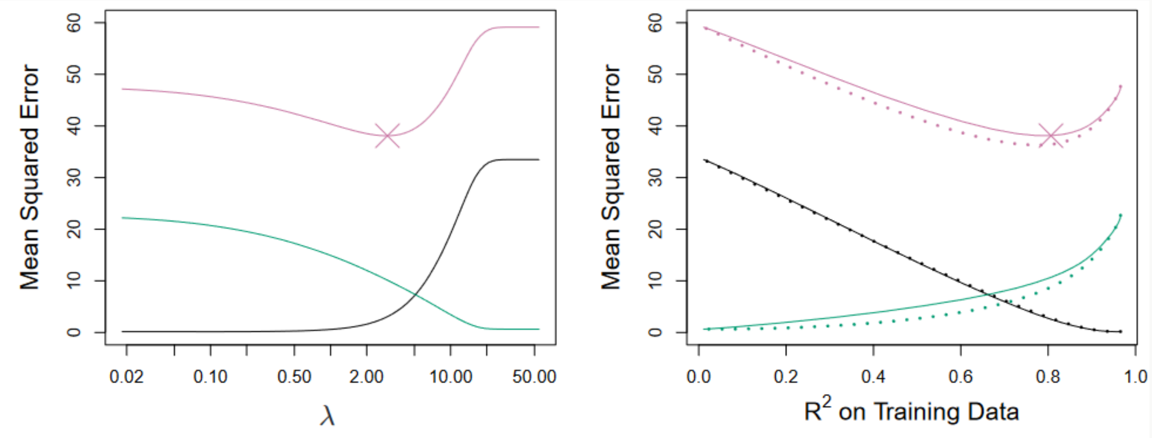
- $X = I$, $n = p$, $\beta_0 = 0$

Ridge: $\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 + \lambda \beta_j^2$ Lasso: $\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 + \lambda \left|\beta_j\right|$

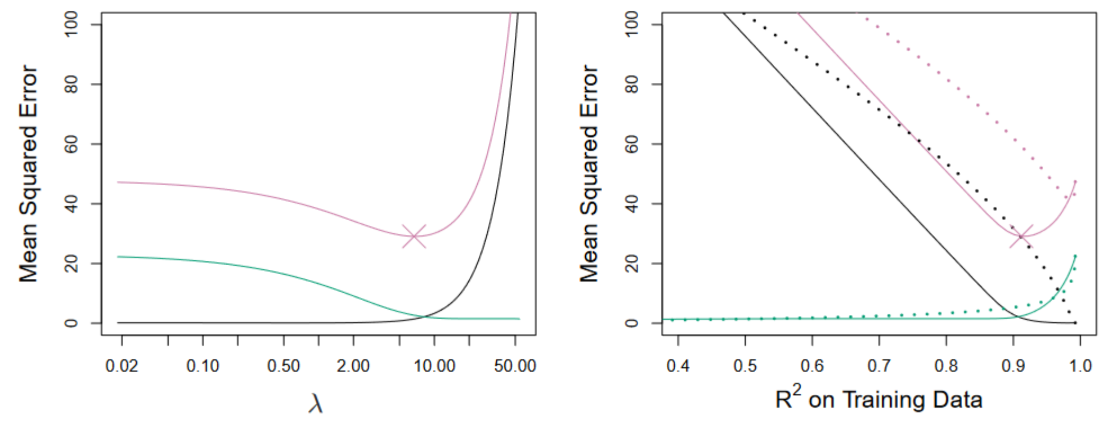


Is LASSO or Ridge better? It depends.

All variables



Few variables

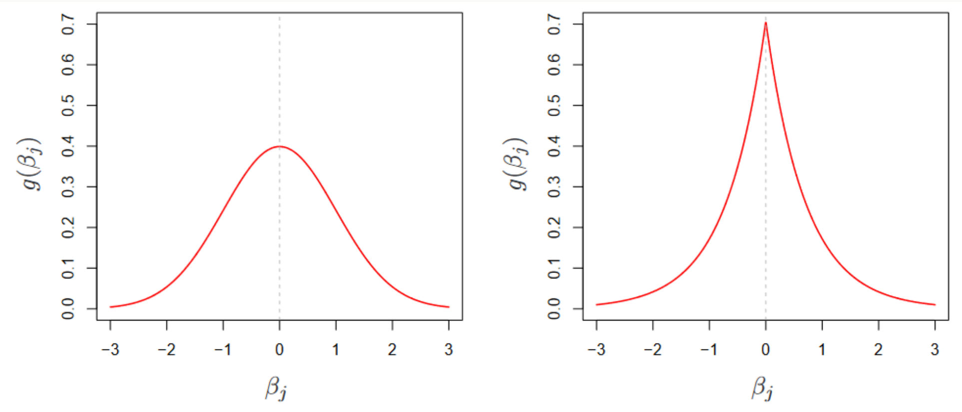


Bayesian Interpretation

- From Bayes' theorem:

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

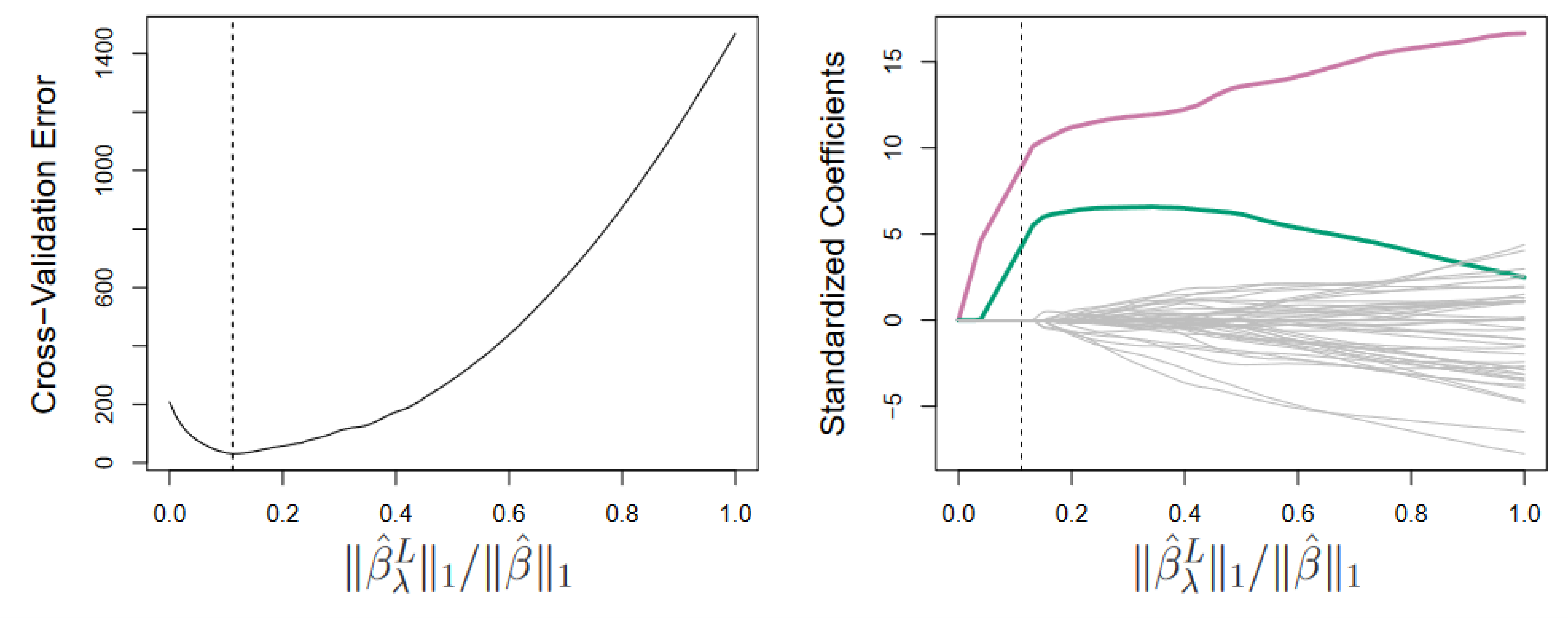
- Assume:
- $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$
- $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ for some density g
- Errors are independent and Gaussian



- Then:
  - g is Gaussian $\Rightarrow$ posterior *mode* of β is Ridge
  - g is Laplacian $\Rightarrow$ posterior *mode* is LASSO

How Should You Pick The Tuning Parameter?

- Cross Validation



Dimension Reduction

Dimension Reduction

- **Combine your predictors to reduce their number**
- Original predictors: $X_1, ..., X_p$
- New predictors $Z_1, ..., Z_M$, where $M < p$:

$$Z_m = f_m(X_1, ..., X_p)$$

- We simplify by restricting to **linear combinations**:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

- Fit regression model:

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

- Equivalent to constraining the values of β:

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

How do you decide what linear combinations to use?

- Heuristics
- Principal Component Regression
- Partial Least Squares

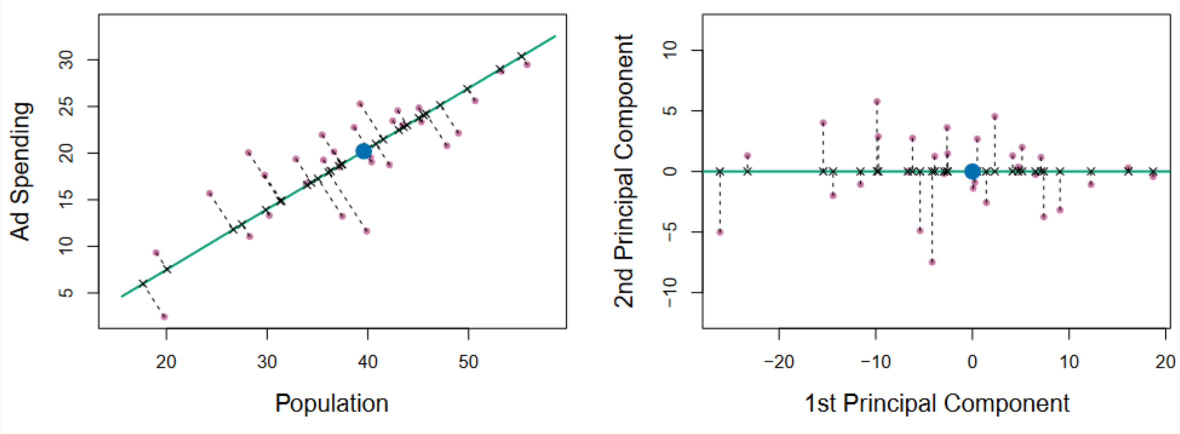
Principal Component Regression

- **Pick first M directions along which your predictor data varies most**
- For a data matrix $\boldsymbol{X}$ of observations $\boldsymbol{x_i} \in \mathbb{R}^p$, choose the first *direction* $\boldsymbol{\phi_1} \in \mathbb{R}^p$:

$$\boldsymbol{\phi_1} = \operatorname{argmax}_{\|\boldsymbol{\phi}\|=1} \left\{ \sum_i \left(\boldsymbol{x_i} \cdot \boldsymbol{\phi} \right)^2 \right\}$$

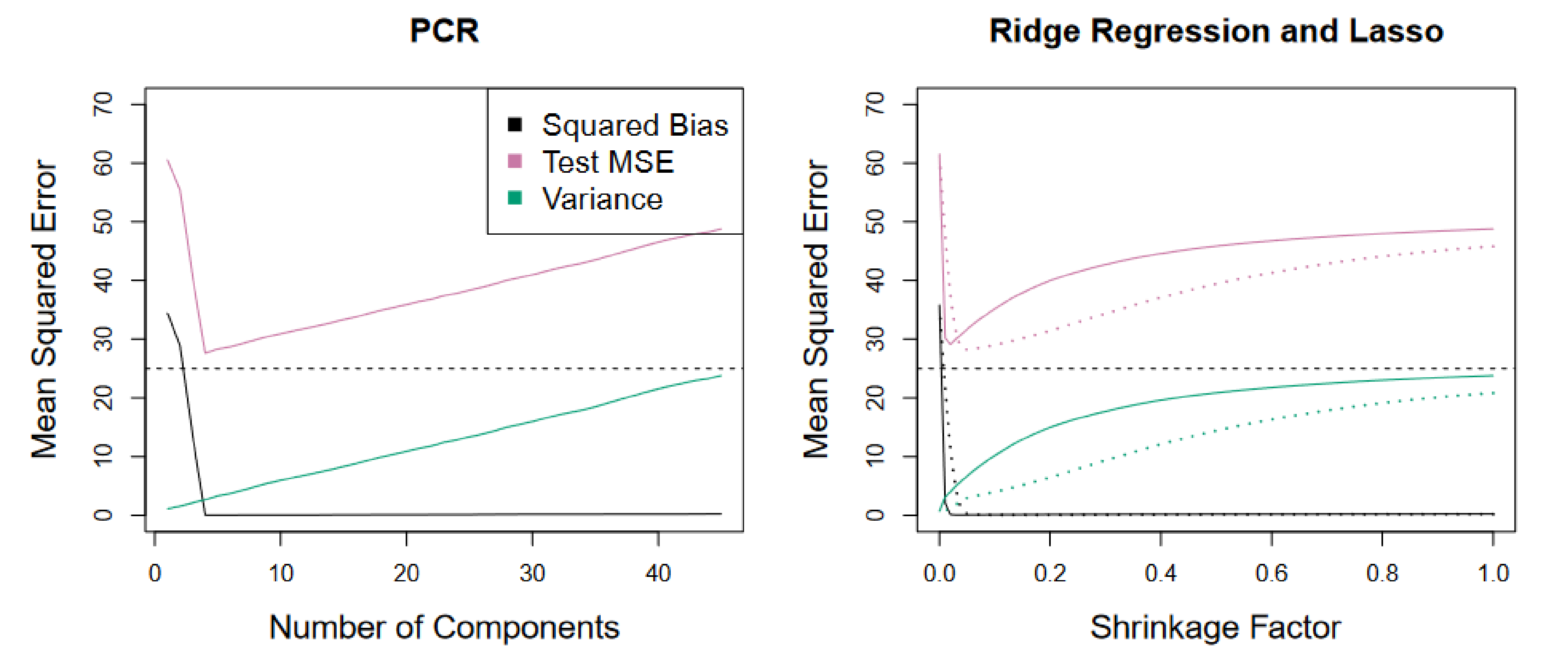
- To find subsequent directions repeat on $\widehat{\boldsymbol{X}}_k$ which has the first $k-1$ directions *removed*:

$$\widehat{\boldsymbol{X}}_k = \boldsymbol{X} - \sum_{s=1}^{k-1} \boldsymbol{X} \boldsymbol{\phi_s} \boldsymbol{\phi}_s^T$$



Principal Component Regression Example

- Example where first 5 components explain Y



Principal Component Regression Notes

- This is *not* feature selection - all predictors are used
- Should standardise predictors - otherwise those will dominate the error
- In practise use eigen-decomposition of the covariance matrix of $\boldsymbol{X}$.
- Implicitly assuming features describing $\boldsymbol{X}$ describe Y - is *unsupervised*

Partial Least Squares

- **Pick first directions that are most related to the response Y**

- Choose the first direction Z_1 by setting ϕ_{j1} equal to the coefficient from a simple linear regression.
- Choose subsequent direction k by regressing predictors on directions $Z_1, \ldots, Z_{k-1}$ and taking the *residual*.

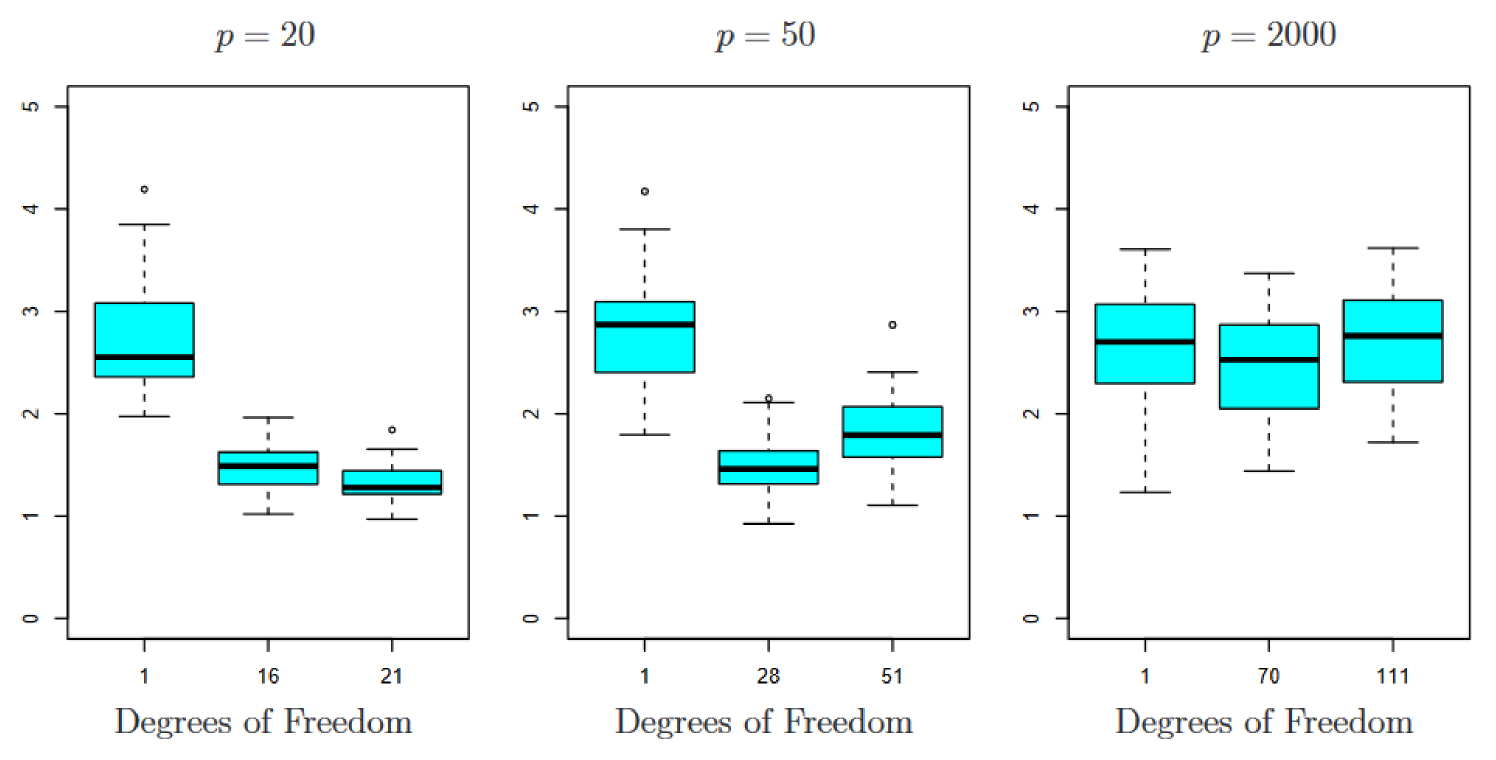
Partial Least Squares

- **Pick first directions that are most related to the response Y**
- Choose the first direction Z_1 by setting ϕ_{j1} equal to the coefficient from a simple linear regression.
- Choose subsequent direction k by regressing predictors on directions $Z_1, \ldots, Z_{k-1}$ and taking the *residual*.

Summary

High dimensions

- Cannot use training R^2, AIC, BIC or C_p.
- Regularisation methods help but are not a panacea. E.g. with $n = 100$ and "correct" $p = 20$:



Summary

Method	Feature Select?	Fast?	Supervised?	Transparent?	Smooth?
Best Subset	☑	✗	☑	☑	☑
Forw'd Stepwise	☑	☑	☑	☑	✗
Backw'd Stepwise	☑	☑	☑	☑	✗
Ridge	✗	☑	☑	☑	☑
Lasso	☑	☑	☑	☑	✗
PCR	✗	☑	✗	✗	✗
Partial Least Sqr	✗	☑	☑	☑	✗

- **Discussion**: What do people recommend?