Regularised Linear Models

Ke Chen

Reading: Sect. 6.2 [Intro Stat Learn Python]
https://www.statlearning.com/

Lecture Goal

- Understanding the motivation underlying regularisation
- Regularised linear models: ridge regression and the LASSO
- Geometric interpretation of ridge regression and the LASSO
- Variants of regularised linear models for regression

Introduction

- The ultimate goal of statistical learning is towards inductive bias and generalisation.

- The bias-variance trade-off suggests that generalisation of a learning model is determined by not only the properties of a given data set (e.g., sample size and "quality") but also model complexity (flexibility).

- While one can employ model selection methods (empirical or analytical) for effective learning toward the ultimate goal, regularisation turns out to be yet another generic manner to facilitate the generalisation.

- Also, regularisation might tackle other issues encountered in learning.

Motivation

- General form of regularisation: for training data $\mathcal{Z} = \{X, Y\}$, a regularisation penalty is introduced to an original loss function

$$l_R\,(Y, \hat{f}(X, \Theta)) = l(Y, \hat{f}(X, \Theta)) + \lambda R(\Theta)$$

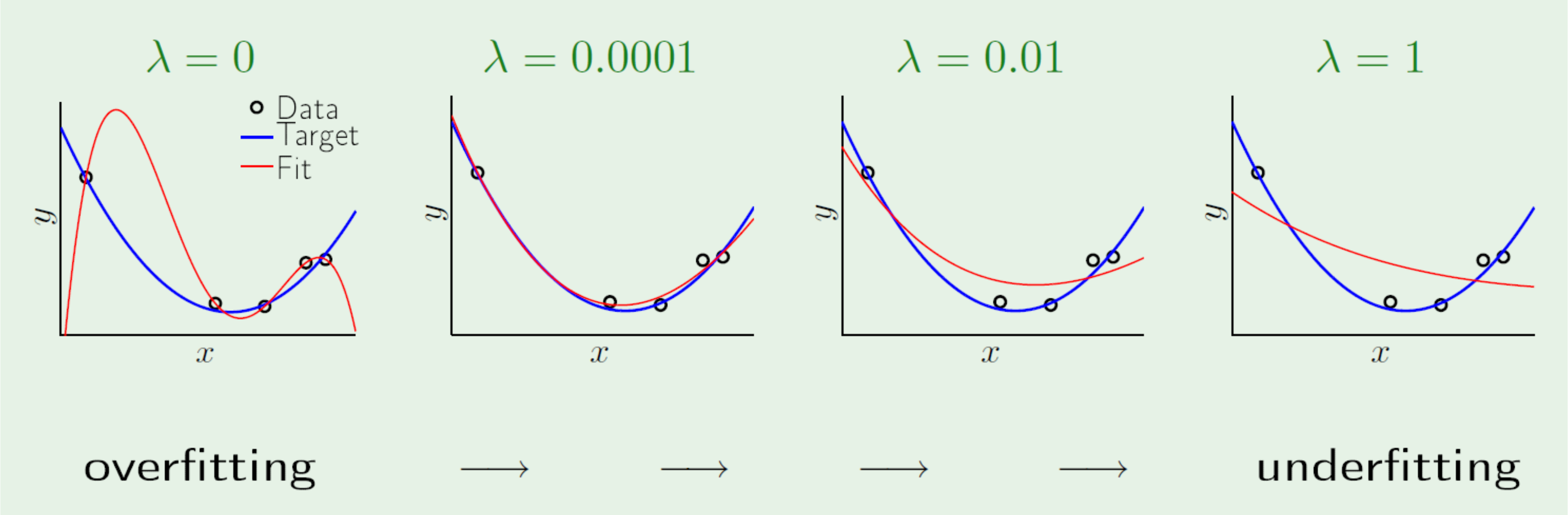
- The general motivation enables both fitting, $l(Y, \hat{f}(X, \Theta))$, and model complexity, $R(\Theta)$, to be considered via a "trade-off factor", λ, during learning.

- Regularisation can also remedy some ill-posed problems, e.g., when the number of training examples (n) is smaller than that of features (p) for $\mathcal{Z} = \{X, \boldsymbol{y}\}$ in linear regression; i.e., when $n < p$,

$$\widehat{\boldsymbol{\beta}} = \boxed{(X^T X)}^{-1} X^T \boldsymbol{y} \longrightarrow \textbf{Singular!}$$

Motivation

- Regularised (linear/polynomial) regression models are able to not only combat the ill-posed problem but also fulfill automatic feature (subset) selection via "shrinkage" of parameters during learning.
- It constrains a learning algorithm to avoid overfitting hence lower test (out-of-sample) error, especially for noisy data.



Ridge Regression

- Ridge regression refers to a regularised linear regression model by applying Tikhonov regularisation to the ordinary least square (OLS) loss.
- The OLS loss is RSS for a linear model of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ on a training data set $\mathcal{Z} = \{X, \boldsymbol{y}\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$,

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

- The ridge regression loss is defined by

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

where λ is a tuneable hyper-parameter (for trade-off) to be determined in advance.

Ridge Regression

- Ridge regression refers to a regularised regression approach by applying Tikhonov regularisation to the ordinary least square (OLS) loss.

- The OLS loss is RSS for a linear model of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ on a training data set $\mathcal{Z} = \{X, \boldsymbol{y}\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$,

$$\mathrm{RSS} = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})$$

- The ridge regression loss is defined with the penalty on $\widetilde{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_p)$ by

$$\mathrm{RSS}(\lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\widetilde{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{\beta}} = \mathrm{RSS} + \lambda\widetilde{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{\beta}}$$

where λ is a tuneable hyper-parameter (for trade-off) to be determined in advance.

Ridge Regression

- The effect of this loss is to add a "shrinkage" or "weight-decay" penalty of the form

$$\lambda \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}} = \lambda \sum_{j=1}^{p} \beta_j^2,$$

  where the tuneable hyper-parameter λ always takes a positive value.

- This has the effect of shrinking the estimated $\tilde{\boldsymbol{\beta}}$ parameters (coefficients) towards zero. It turns out that such a constraint can improve the fit, because shrinking parameters (coefficients) can significantly reduce their model complexity for variance reduction.

- Note that when $\lambda = 0$, the penalty term has no effect, and ridge regression will be the OLS estimate. Thus, selecting a proper (optimal) value for λ is critical.

- Note β_0 has been left out from the shrinkage penalty as it is simply an intercept at the "origin", which can be estimated separately in the ridge regression learning.

Ridge Regression

- For parameter estimate in ridge regression, the training data $\mathcal{Z} = \{X, \boldsymbol{y}\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ must be standardised first for $i = 1, \cdots, n$; $j = 1, \cdots, p$

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

$$\hat{\beta}_0 = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Parameter estimate for ridge regression is done by minimising the ridge regression loss function with respect to $\tilde{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_p)$.
- By setting the 1st derivative of the ridge regression loss function w.r.t. $\tilde{\boldsymbol{\beta}}$ to zero, an analytical solution for a pre-setting λ (λ >0) is obtained:

$$\hat{\boldsymbol{\beta}}_{ridge} = \boxed{(\tilde{X}^T\tilde{X} + \lambda I_p)}^{-1}\tilde{X}^T\boldsymbol{y}$$

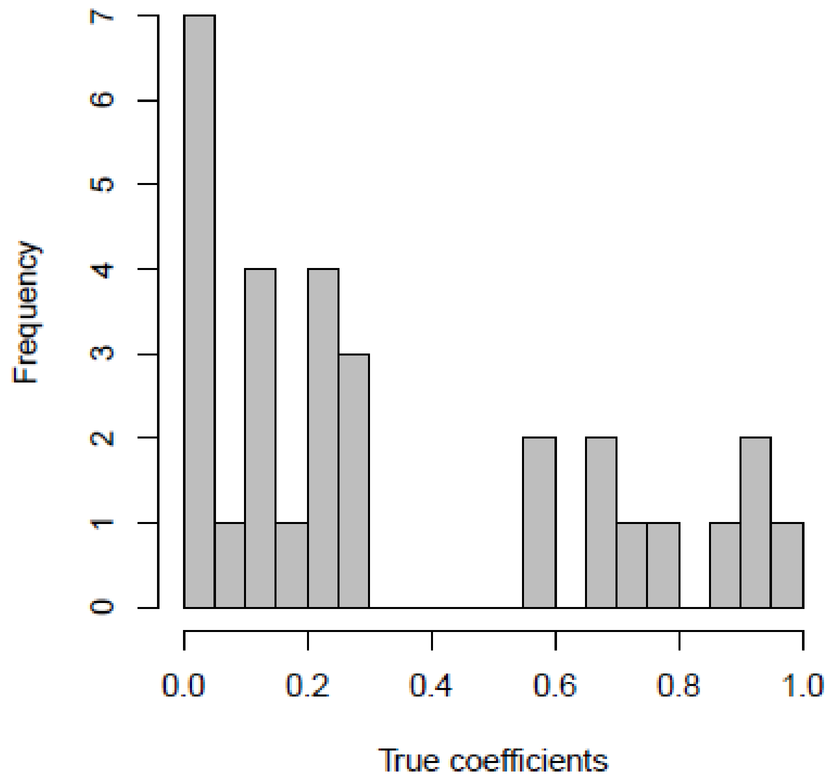
Non-singular!

where I_p is a $p \times p$ identity (unit) matrix.

Ridge Regression

Example: simulation with $n = 50$ and $p = 30$. The entries of the predictor matrix $X \in \mathbb{R}^{50\times 30}$ are all i.i.d. $N(0, 1)$, so overall the variables have low correlation

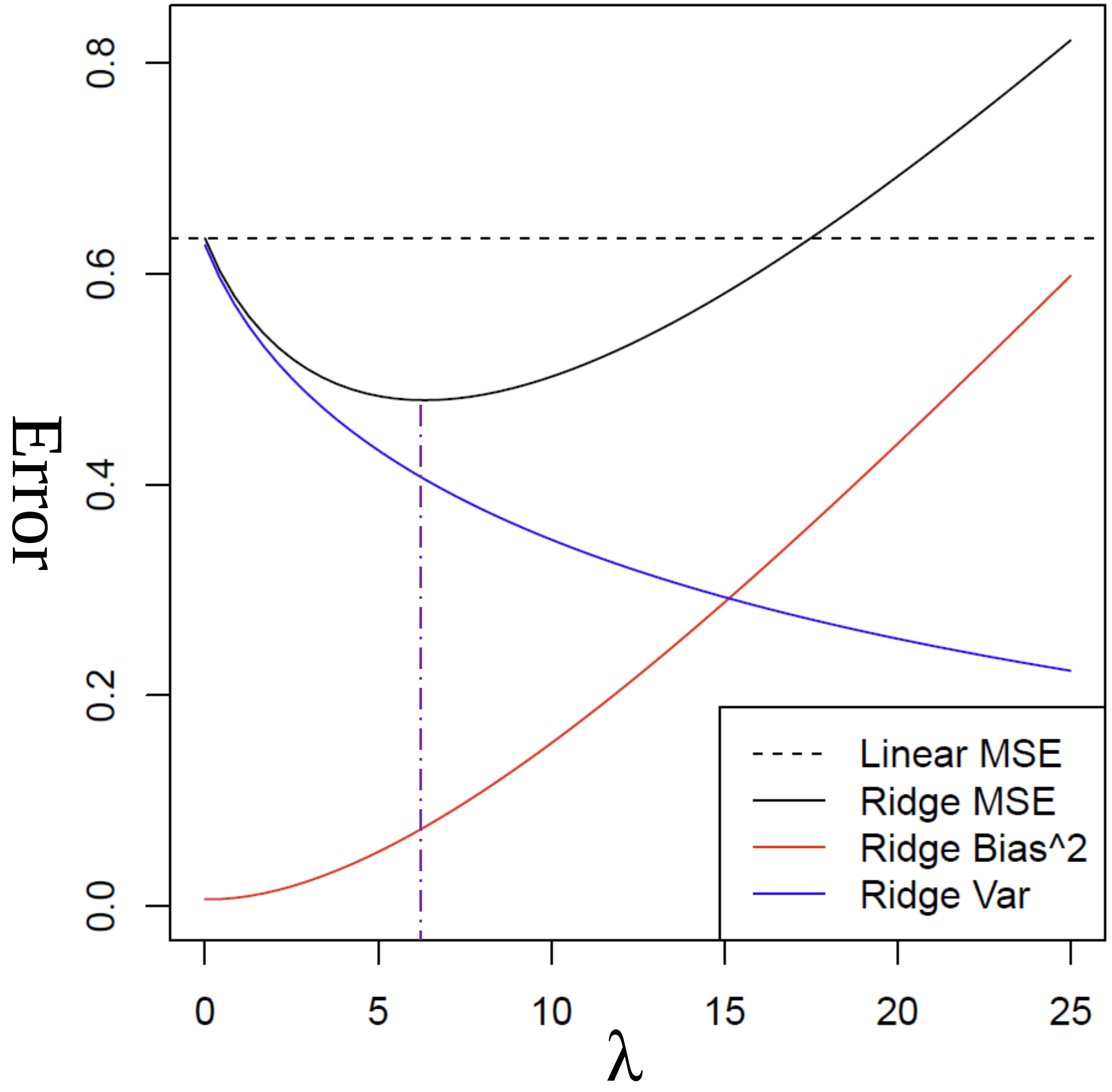
Histogram of the true regression coefficients $\beta^* \in \mathbb{R}^{30}$:



Here 10 coefficients are large (between 0.5 and 1) and 20 coefficients are small (between 0 and 0.3)

Ridge Regression

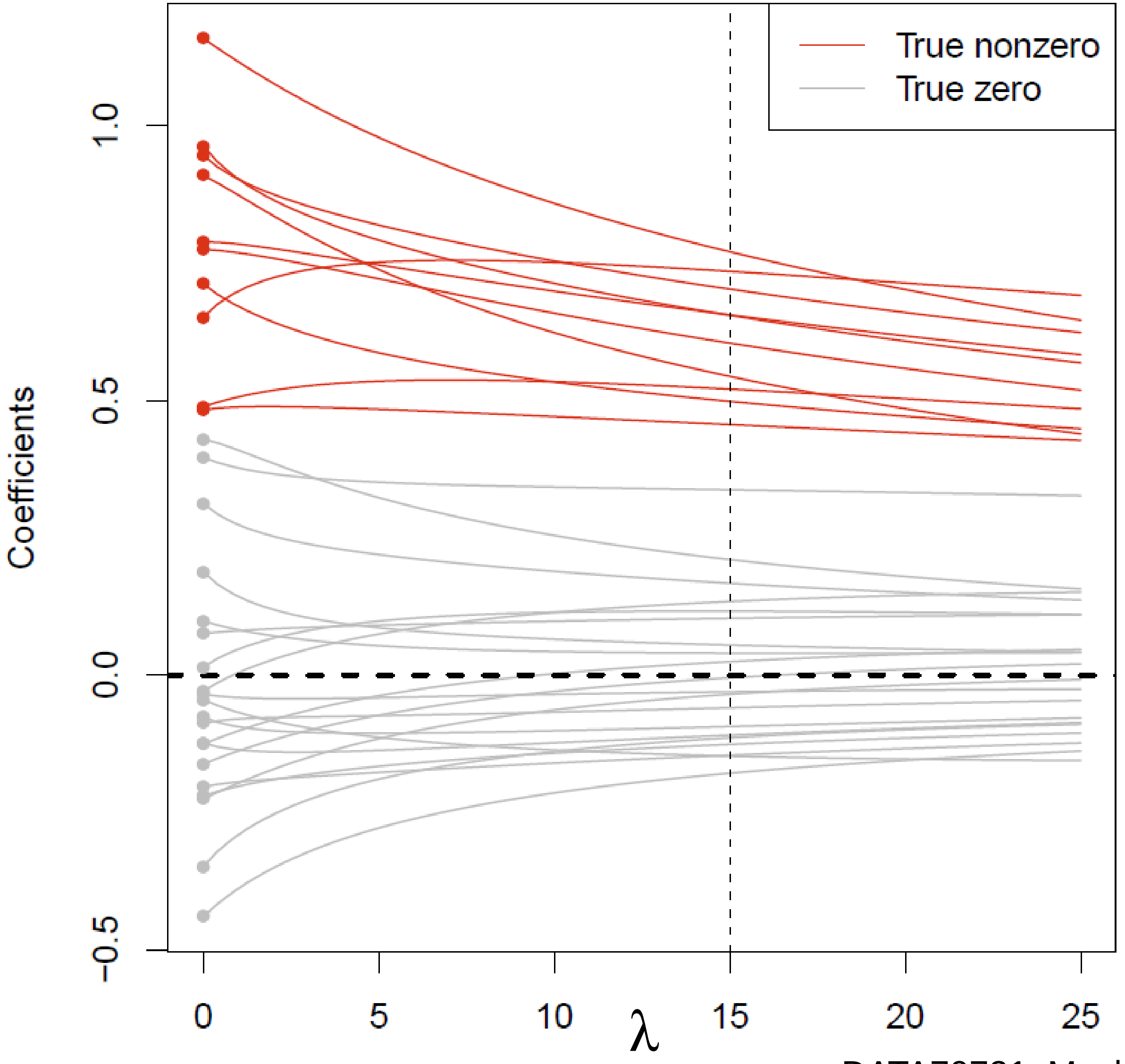
- Example: MSE vs. bias-variance error decomposition at different λ



- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.
- A proper λ used in ridge regression (with regularisation) can lead to a lower MSE than that of linear regression without regularisation.

Ridge Regression

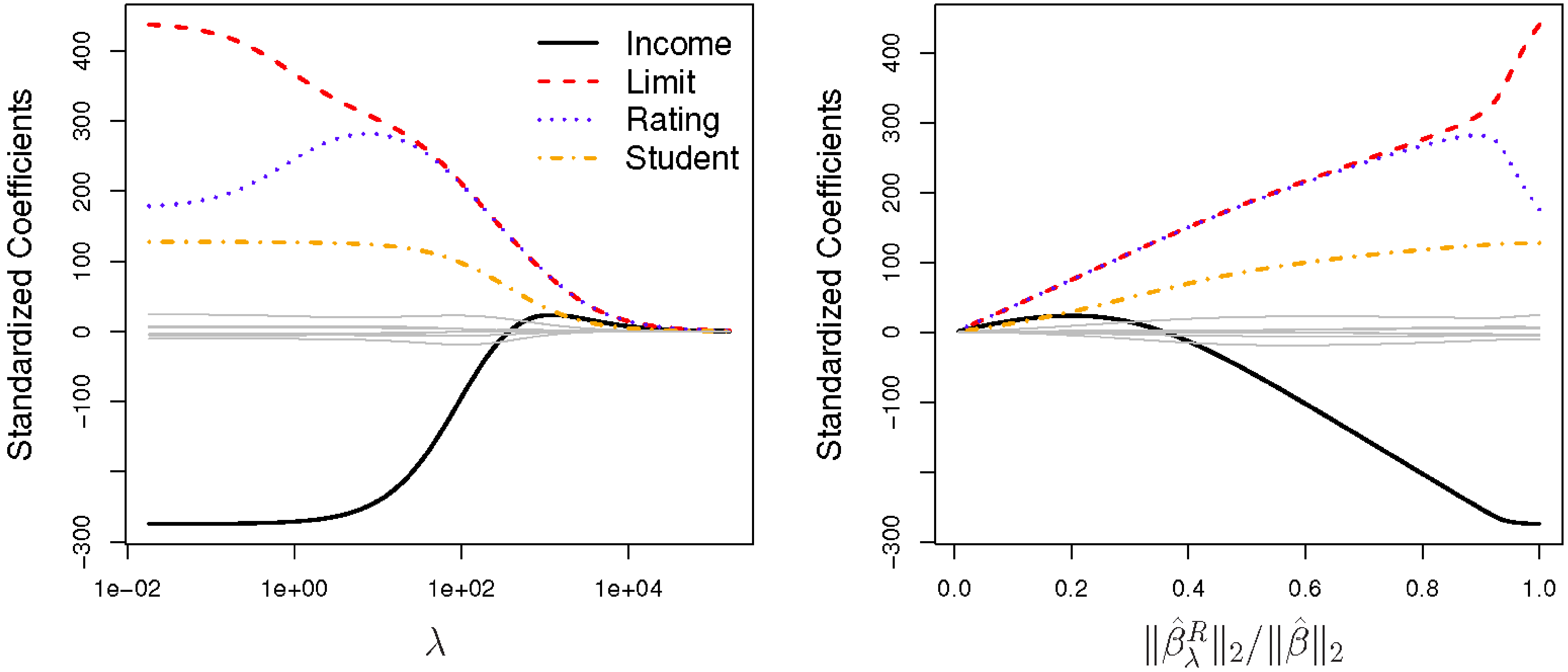
- Example: estimate of parameters on a new synthetic dataset at different λ



- 10 red paths correspond to the true non-zero coefficients; 20 grey paths correspond to true zeros.
- The vertical dashed line at λ=15 marks the point above which the MSE of ridge regression starts losing to that of linear regression.
- An important observation is that the grey coefficient paths are not exactly zero; they are shrunken, but still nonzero.

Ridge Regression

- Example: [predicting credit based on 11 features, Sect. 3.3 in ISLP]



Ridge Regression

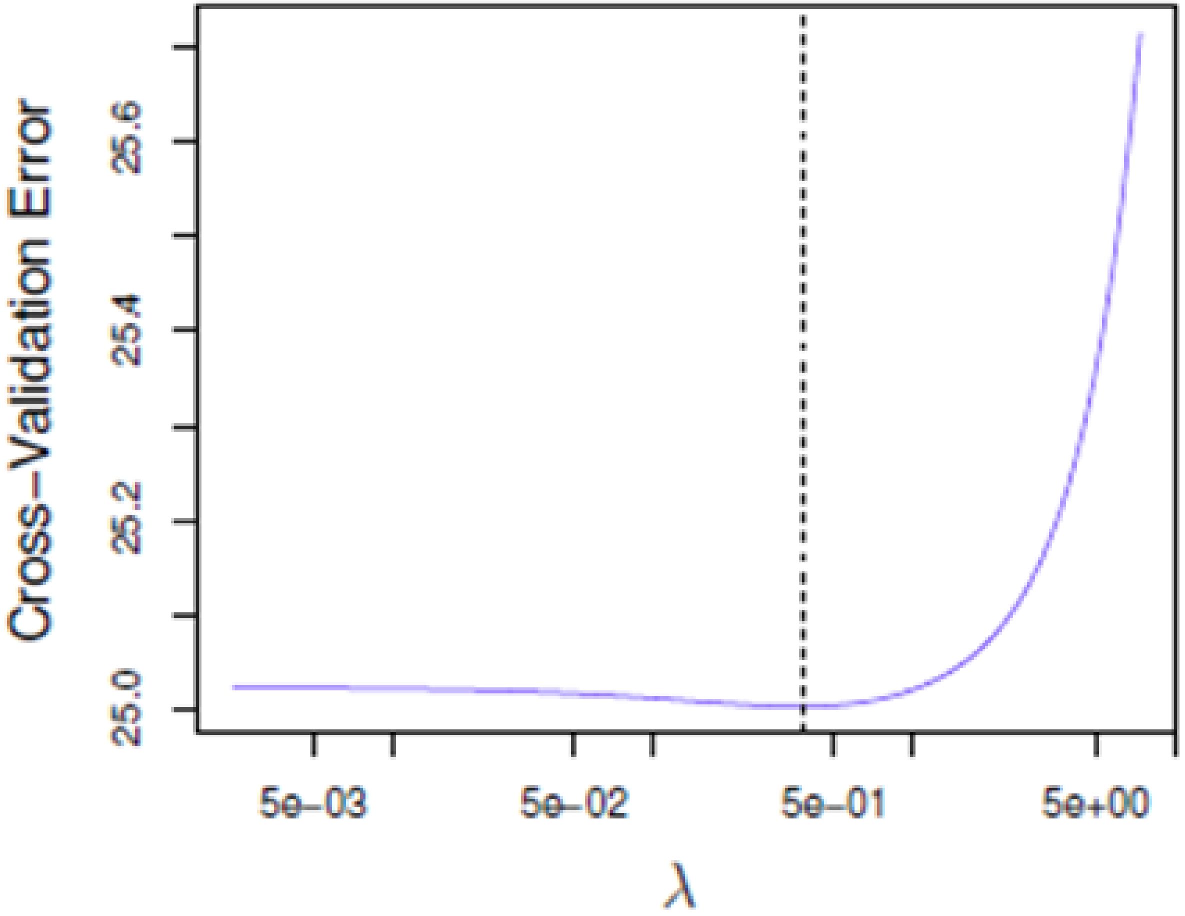
- How to find out a proper (optimal) value of λ?
  - Degree of freedom of a parameter estimate regarding λ is defined by

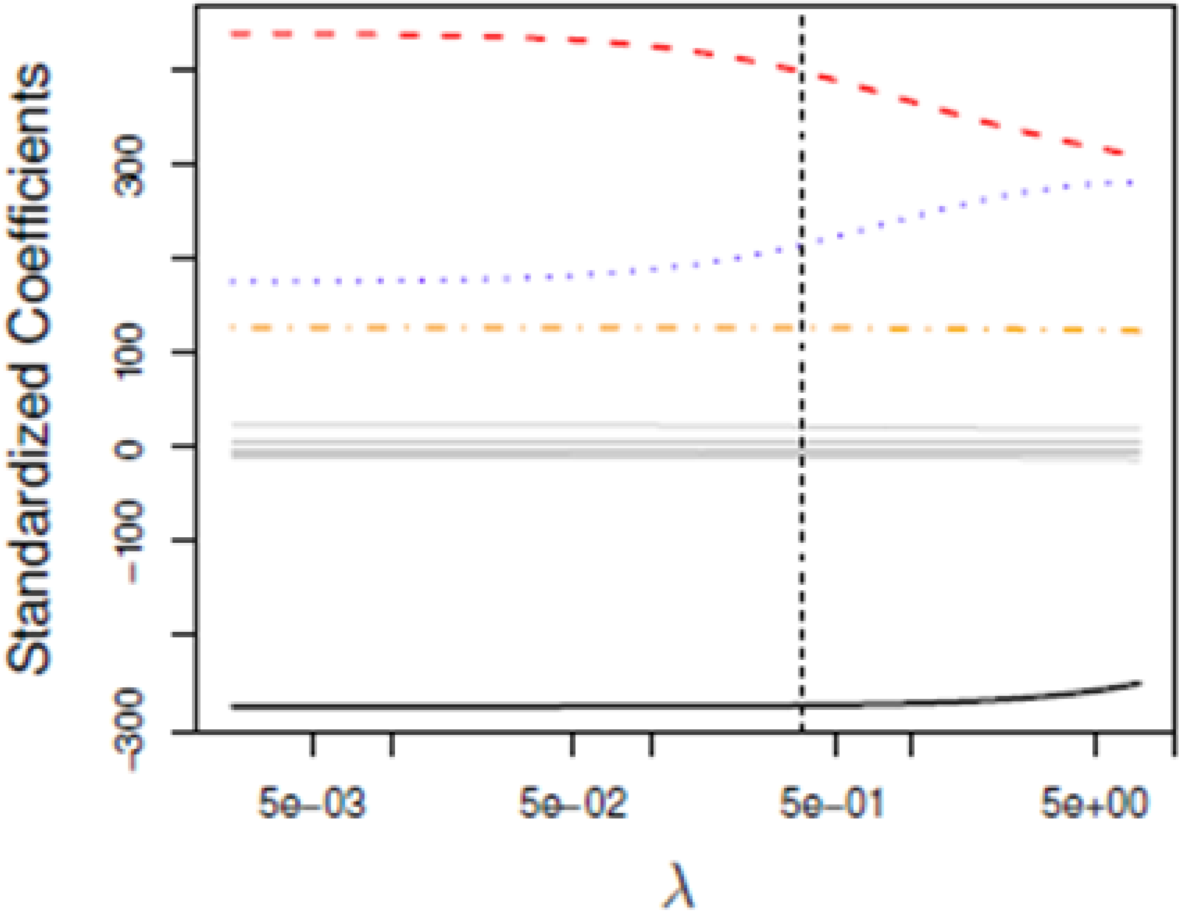
$$df(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

    where d_j is singular values of X obtained with Singular Value Decomposition (SVD) or the eigen values of $X^T X$. Thus, $df(\lambda) = p$ if $\lambda = 0$ and $X^T X$ is non-singular.
  - Simply pick the effective degrees of freedom that one would like associated with the fit, and solve for λ based on singular values of X.
  - As a generic method, cross-validation can always be used to finding a proper λ.
  - Finally, the model is re-fit with all training data and the optimal value of that tuneable hyper-parameter λ.

Ridge Regression

- Example: [apply cross-validation (credit) to find an optimal value of λ]





LASSO

- As observed in the examples, one significant problem of ridge regression is that the l_2 penalty term will never force any of the coefficients to be exactly zero.

- Thus, the final ridge regression model will include all p predictors, which creates a challenge in model interpretation; i.e., what subset of features are non-trivial.

- Least absolute shrinkage and selection operator (LASSO) is a more recent statistical learning alternative to ridge regression to solve the above problem.

- The LASSO works in a similar way to ridge regression, except it uses l_1, a different penalty term, that can shrink some of the parameters exactly to zero.

LASSO

- The LASSO loss is defined on parameters (coefficients) $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ on a training data set $\mathcal{Z} = \{X, \boldsymbol{y}\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ as follows:

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

- The vector version of LASSO loss is defined with the penalty on $\widetilde{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_p)$ by

$$\text{RSS}(\lambda) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda||\widetilde{\boldsymbol{\beta}}||_1 = \text{RSS} + \lambda||\widetilde{\boldsymbol{\beta}}||_1$$

where λ is a tuneable hyper-parameter (for trade-off) to be determined in advance.

- To facilitate parameter estimate, the standardisation procedure (the same as used in ridge regression) must be first applied to the training data on X and deal with β_0 separately.

LASSO

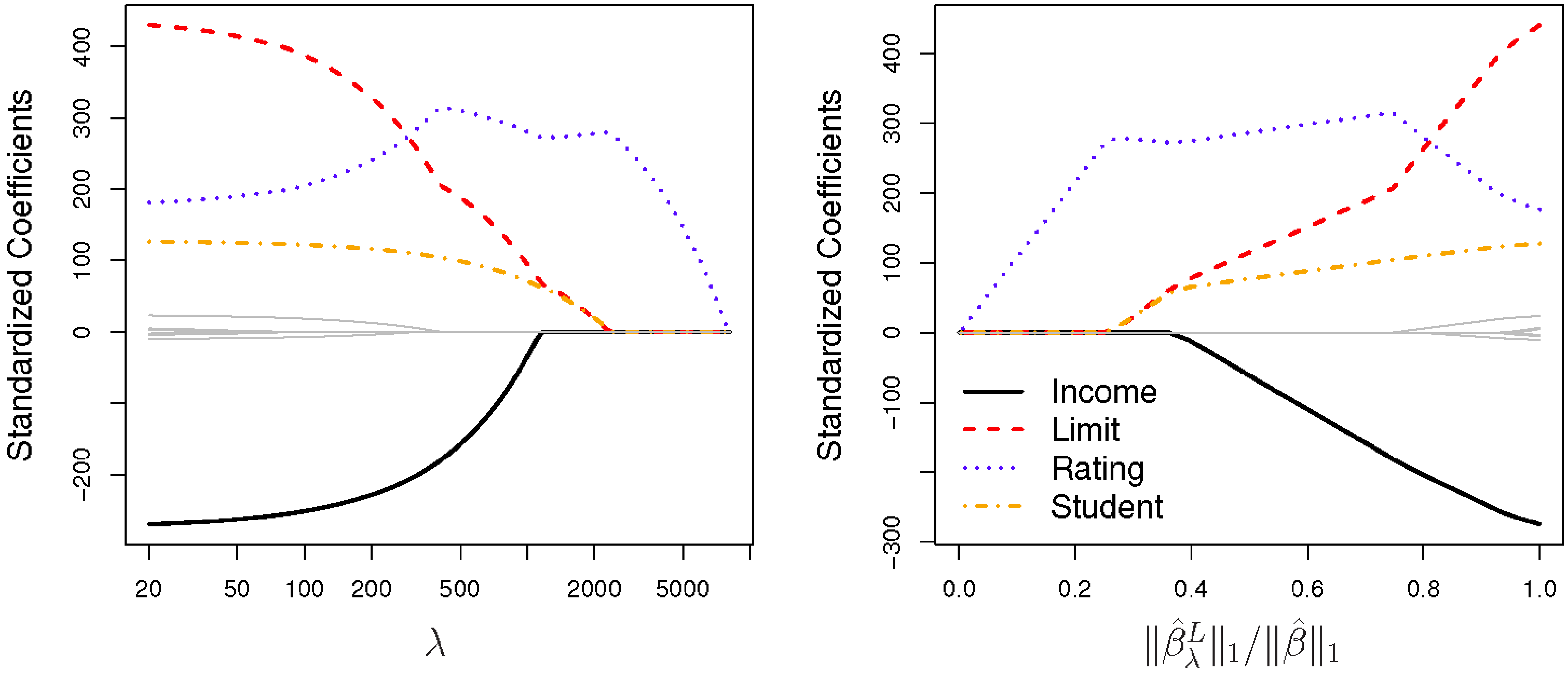
- Unlike ridge regression, parameter estimate in LASSO is quite different as there is not an analytical or close-form solution to minimise the LASSO loss.
- There are several iteration-based LASSO learning algorithms.
- The forward stagewise learning algorithm is the widely used one as follows:

For a given training data set, $\mathcal{Z} = \{X, \boldsymbol{y}\}$, choosing a small ϵ,

1) Initialise the residual $\boldsymbol{r} = \boldsymbol{y}$ and $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
2) Find out the feature $\boldsymbol{x}_j \ (j = 1, \cdots, p)$ most correlated with $\boldsymbol{r}$ (Pearson)
3) Update $\beta_j \leftarrow \beta_j + \delta_j$ where $\delta_j = \epsilon \cdot \mathrm{sign}(\boldsymbol{x}_j^T \boldsymbol{r})$
4) Set $\boldsymbol{r} \leftarrow \boldsymbol{r} - \delta_j \boldsymbol{x}_j$, and repeat Step 2) and 3) until a stopping condition is met.

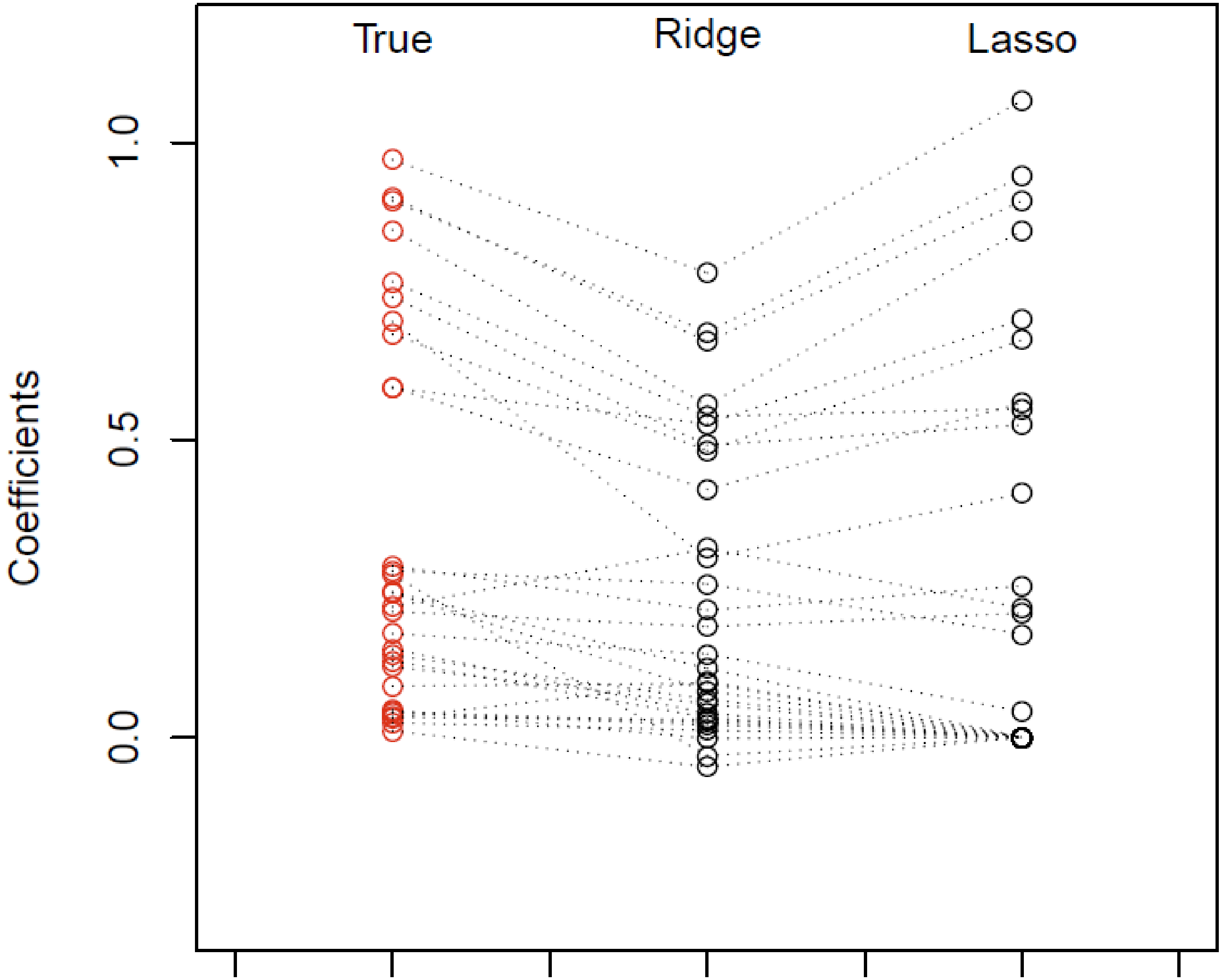
LASSO

- Example: [predicting credit based on 11 features, Sect. 3.3 in ISLR]



LASSO

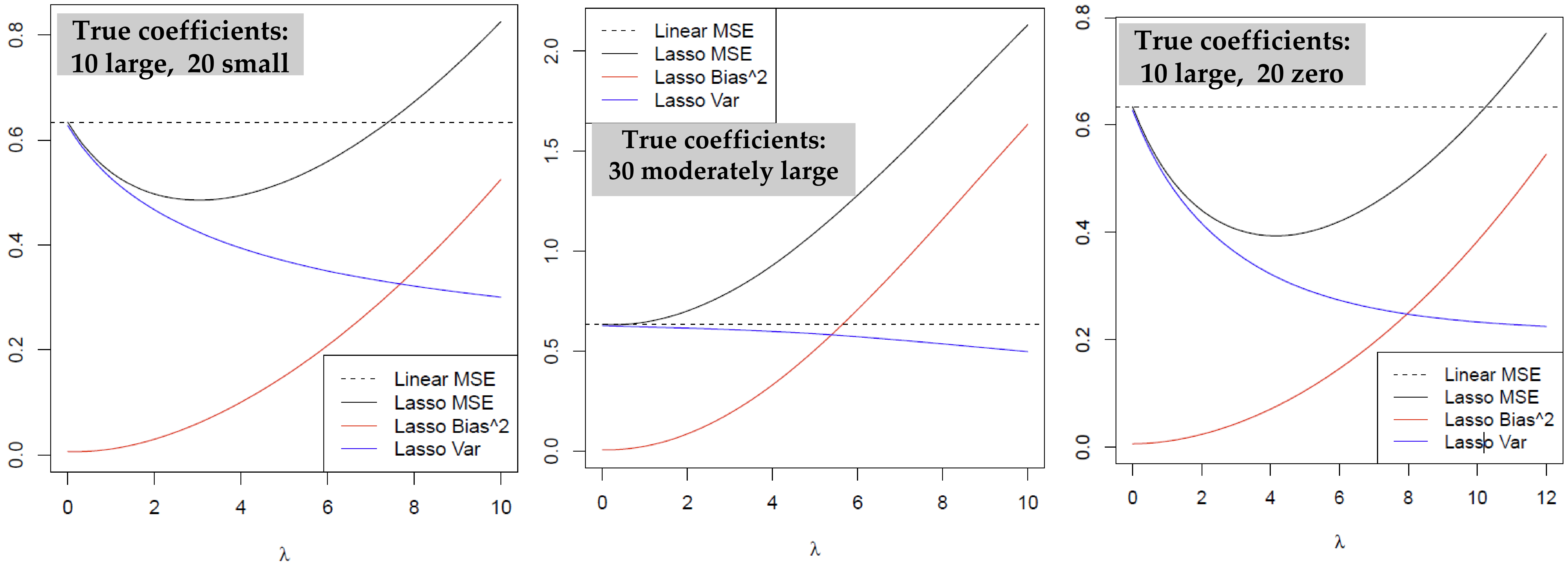
- Example (Same Data in Slide 10): estimate of parameters at the optimal λ



- The red dots correspond to the true coefficients.
- Ridge regression: most of the true coefficients are shrunk but none of true zero-coefficient have been shrunk to zero.
- LASSO: all the true zero-coefficient have been shrunk to zero.

LASSO

- Examples: bias-variance trade-off on different toy data sets ($n=50, p=30$)



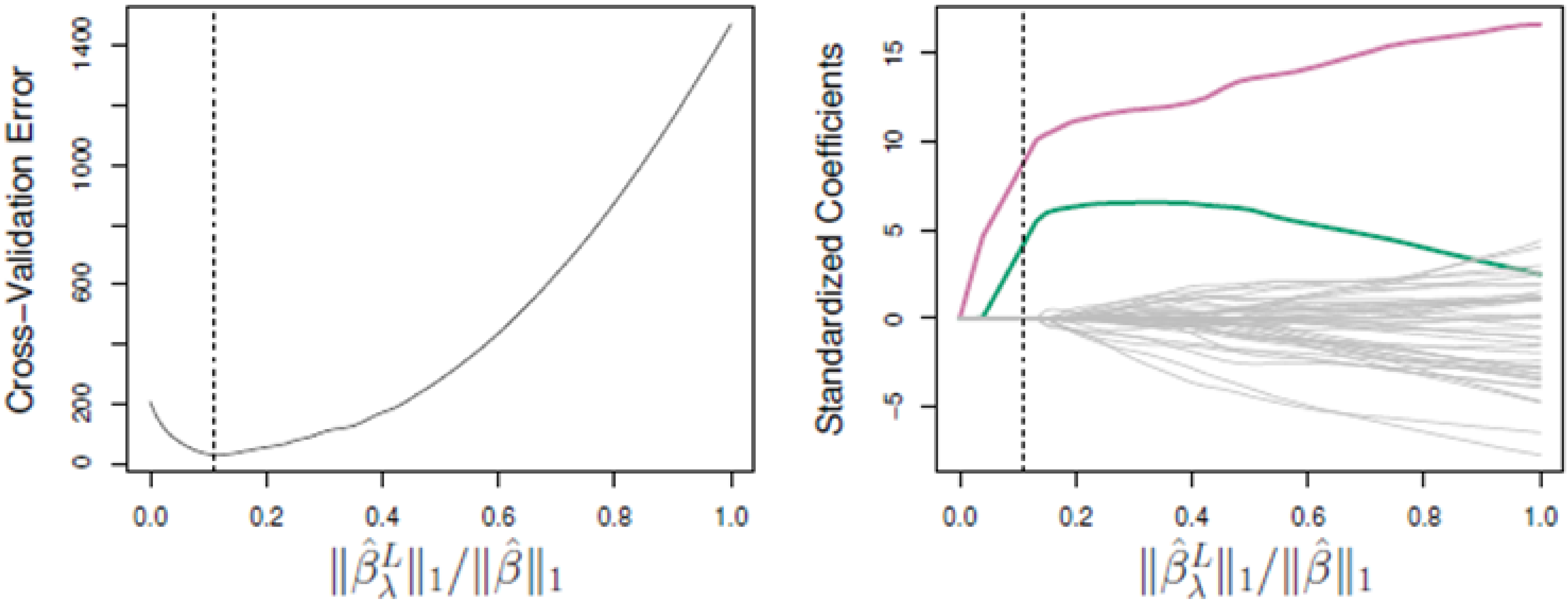
LASSO

- How to find out a proper (optimal) value of λ?
  - Unlike ridge regression, there is **no** analytical solution to an optimal λ.
  - Cross-validation is always used to find out a proper (optimal) λ.
  - Finally, the model is re-fit with all the training data by using all of the variable observations and the selected value of hyper-parameter λ.

LASSO

- Example: apply 10-fold cross-validation to find an optimal value of λ (2 features)

Data generated with 43 zero and 2 non-zero true coefficients (Sect. 6.2, ISLR)



Ridge Regression vs. LASSO

- LASSO has a major advantage over ridge regression; it produces simpler and more interpretable models that involved only a subset of features.

- LASSO leads to qualitatively similar behaviour to ridge regression, in that as λ increases, the variance decreases and the bias increases.

- LASSO often generates more accurate predictions compared to ridge regression but ridge regression may outperform LASSO when all features are nontrivial.

- Once again, cross-validation can be used for model selection in order to determine which one is better on a specific data set in hand.

Ridge Regression vs. LASSO

- To understand why LASSO is better than ridge regression in terms of model interpretation (feature subset selection), the optimisation problems in parameter estimates of LASSO and ridge regression can be reformulated as follows:

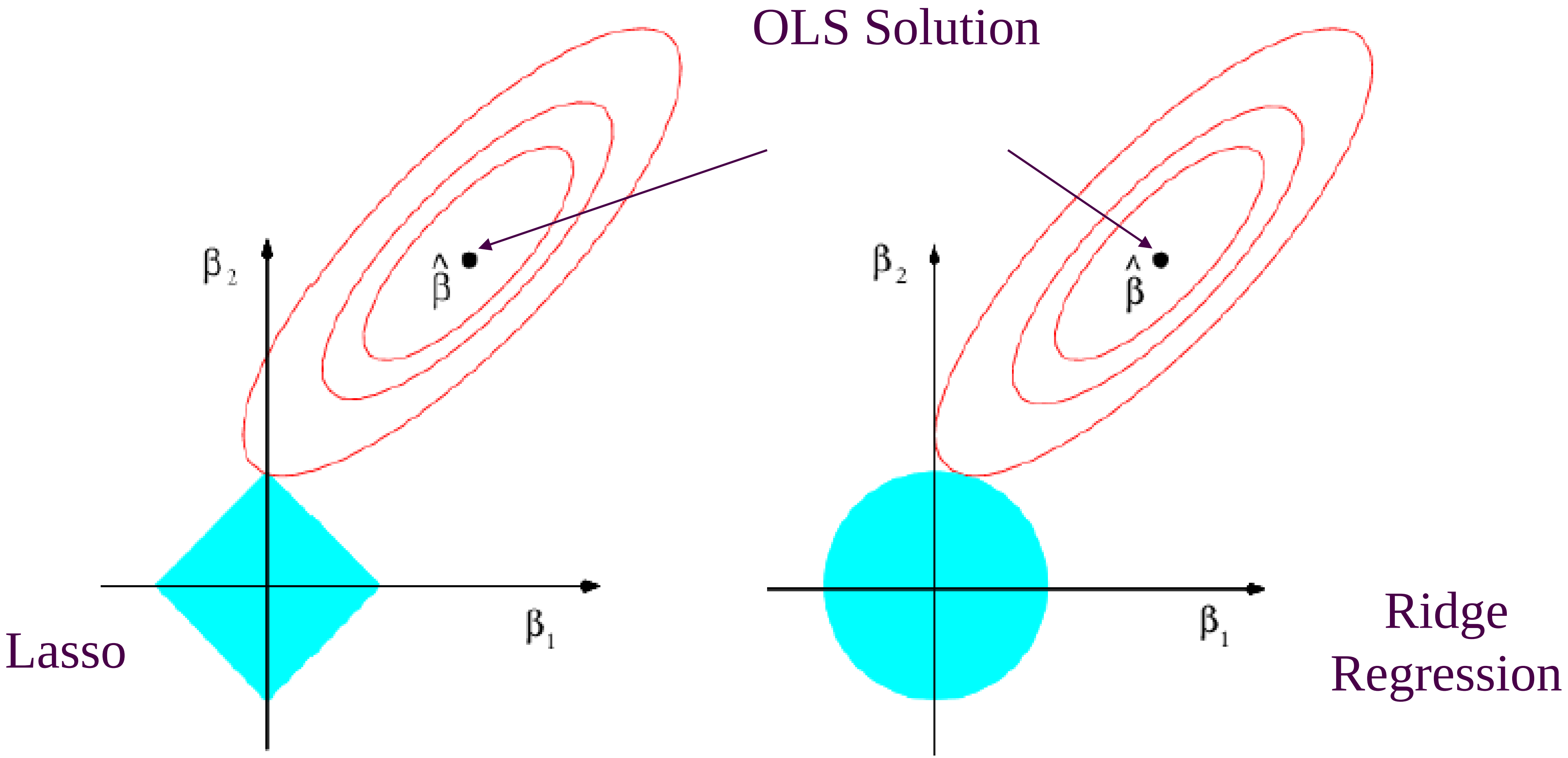
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Ridge Regression vs. LASSO

- LASSO and ridge regression parameter estimates are given by the first point at which an ellipse contacts the constraint region.



Variant of Regularised Linear Model

- Bridge Regression (Frank & Friedman, 1993)
  - apply generic l_q penalty term for regularisation; LASSO if $q = 1$, ridge regression if $q = 2$
  - provide more meaningful results in data analysis both from the theoretical and empirical perspectives
- Elastic Net (Zou & Hastie, 2005)
  - apply both l_1 and l_2 penalty terms for regularisation with two hyperparameters λ_1 and λ_2
  - able to improve the performance of LASSO further and still produce sparse solutions
- Group LASSO (Yuan & Lin, 2006)
  - allow predefined groups of features to be selected into or out of a model together, so that all the members of a particular group are either included or not included

Summary

- Regularisation turns out to be yet another generic idea to facilitate statistical learning toward its ultimate goal and remedy other ill-posed problems.
  - Generic form: empirical loss + λ * regularisation-penalty (structural loss)
  - Two terms in a loss with regularisation are traded off via the "trade-off factor", λ.
- Ridge regression and LASSO are two popular regularised linear models.
  - l_1 and l_2 penalty terms are employed in LASSO and ridge regression.
  - for parameter estimate, an analytical solution exists for ridge regression while an iterated algorithm has to be used in LASSO.
  - turning the hyper-parameter λ is key to success and can always be done via cross-validation
  - In general, LASSO outperforms ridge regression in model interpretation.
  - The performance can be understood with the bias-variance trade-off; as λ increases, the variance decreases and the bias increases.
- Main variants (still an active research area in statistical learning)
  - bridge regression, elastic net, group LASSO, ...