Model Selection and Regularization

Tianhang Zheng

https://tianzheng4.github.io

Model Selection and Regularization Methods

Subset Selection. We identify a subset of the p predictors that we believe to be related to the response.

Shrinkage. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates (**Regularization**)

Dimension Reduction. We project the p predictors into a M-dimensional subspace, where M < p. This is achieved by computing M different linear combinations, or projections.

Subset Selection

Let M_0 denote the null model, which contains no predictors.

For k = 1,2,...p:

(a) Fit **all** models that contain exactly k predictors

(b) Pick the best among these models and call it M_k. Here the best is defined as having the smallest RSS.

Select a single best model from among $M_0, \ldots, M_k$ using cross-validated prediction error or adjusted R2

Forward Stepwise Selection

```
Let $M_0$ denote the null model, which contains no predictors.

For k = 1,2,...p:
    (a) Consider all p − k models that augment one predictor
    (b) Pick the best among these models and call it $M_k$. Here the best is defined as having the smallest RSS.

Select a single best model from among $M_0, . . . , M_k$ using cross-validated prediction error or adjusted R2
```

Forward Stepwise Selection (Greedy)

Let M_0 denote the null model, which contains no predictors.

```
For k = 1,2,...p:
    (a) Consider all p − k models that augment one predictor
    (b) Pick the best among these models and call it $M_k$. Here the best is defined as having the smallest RSS.
```

Select a single best model from among $M_0, . . . , M_k$ using cross-validated prediction error or adjusted R2

Adjusted R2

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

Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-p-1}$

Shrinkage Methods

Ridge regression and Lasso

Fit a model containing all p predictors using a technique that ***constrains or regularizes the coefficient estimates***, or equivalently, that shrinks the coefficient estimates towards zero. (reduce parameter variance)

Ridge regression

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

Ridge regression objective

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

λ is a hyperparameter

Ridge regression

Before ridge regression, we usually need to standardize the predictors

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

We usually use cross validation to set λ

Lasso

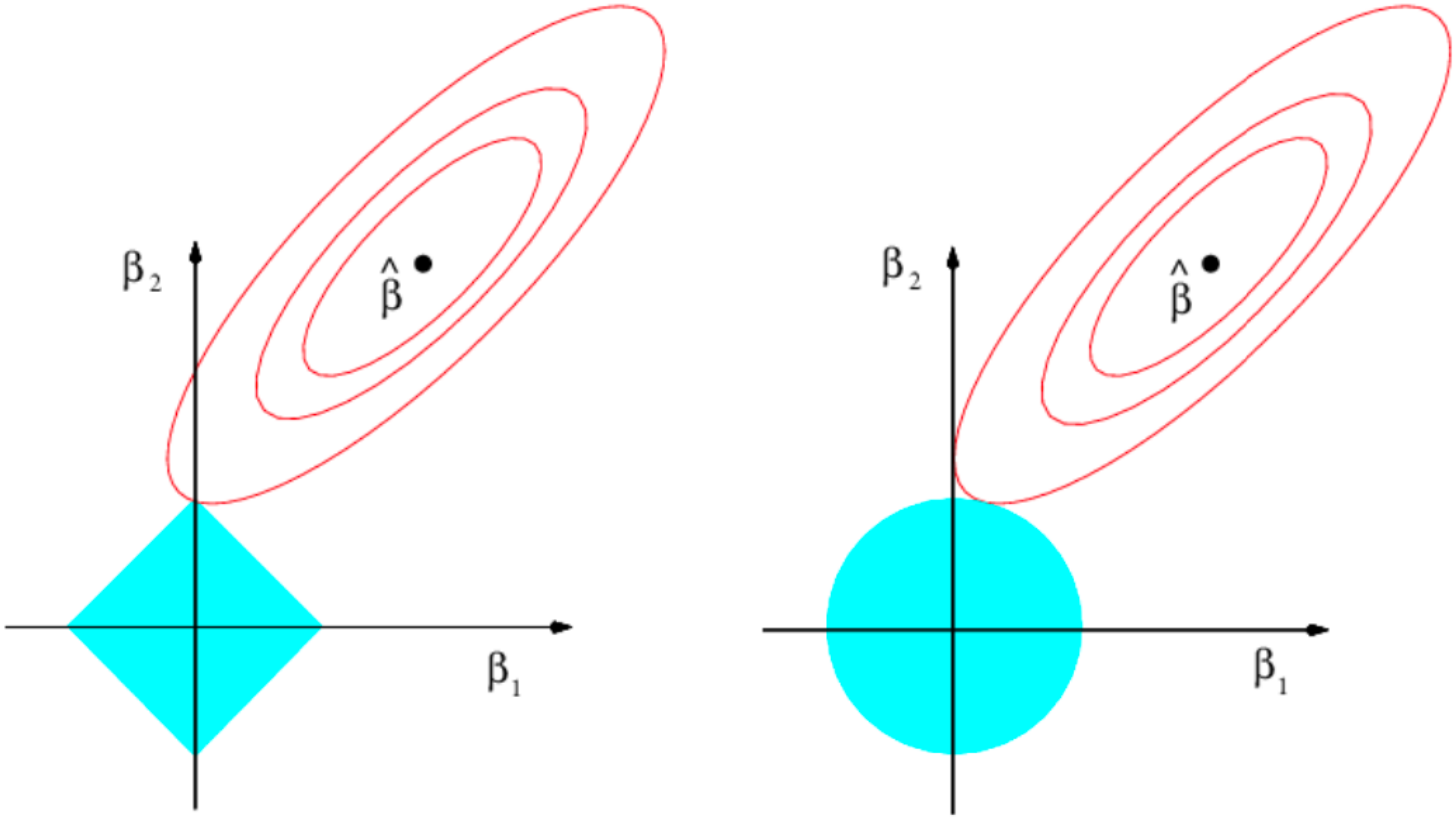
Ridge regression will include all p predictors in the final model (Disadvantage: No predictor selection)

Objective of Lasso

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

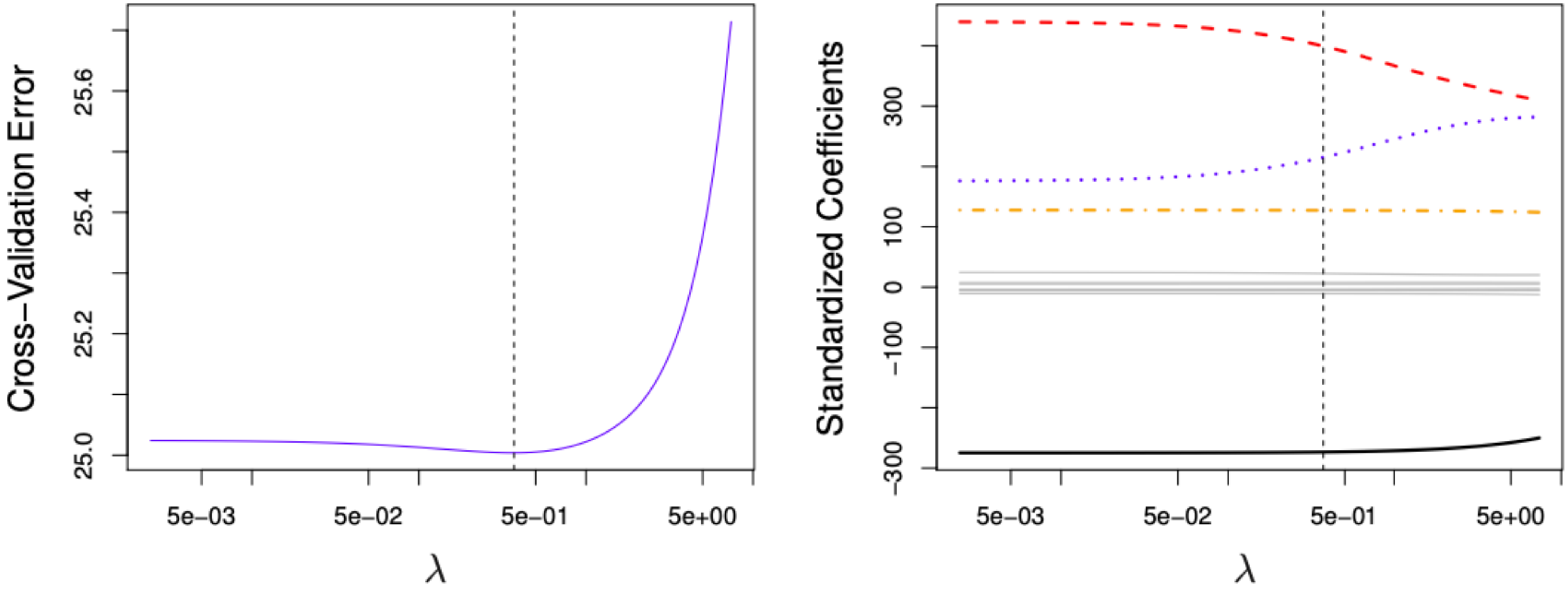
If λ is sufficiently large, then Lasso will force some β to exactly zero (equivalent to predictor selection)

Lasso vs Ridge Regression



Why Lasso can force some β to exactly zero?

An example


Cross−Validation Error
25.0
25.2
25.4
25.6
5e−03
5e−02
5e−01
5e+00
λ
Standardized Coefficients
−300
−100
0
100
300
5e−03
5e−02
5e−01
5e+00
λ

Dimension Reduction

Use linear combinations of the predictors to construct new predictors

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

We can then fit the linear regression model

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

Dimension Reduction

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} \qquad \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Dimension reduction serves to constrain the estimated coefficients as

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

Dimension Reduction

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} \qquad \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Dimension reduction serves to constrain the estimated coefficients as

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

Principal Components Regression

Dimension reduction by Principal Components Analysis (PCA), and conduct linear regression on new predictors

The first principal component is that (normalized) linear combination of the variables with the largest variance.

The second principal component has largest variance, subject to being uncorrelated with the first.

Principal Components Analysis

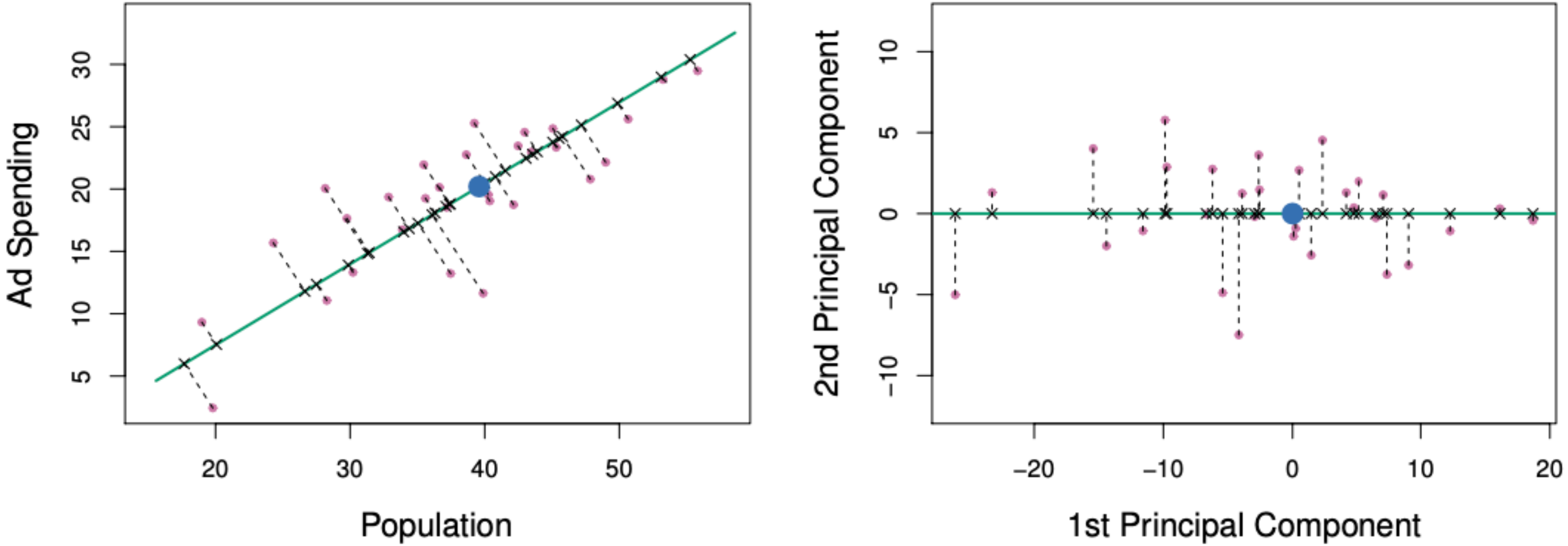
$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} \right\}$$ $\mathbf{X}$: (n, p) not $(n, p+1)$

$w_{(1)}$ is the eigenvector corresponding to the largest eigenvalue

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)} \mathbf{w}_{(s)}^\top$$

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \left\|\hat{\mathbf{X}}_k \mathbf{w}\right\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^\top \hat{\mathbf{X}}_k^\top \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

Principal Components Analysis



Principal Components Regression

PCR identifies linear combinations, or directions, that best represent the predictors

These directions are identified in an unsupervised way, since the response Y is not used to determine the principal component directions

Drawback: no guarantee that the directions that best explain the predictors will be the best directions to use for predicting the response.

Q & A