INTRODUCTION

To Statistical Learning

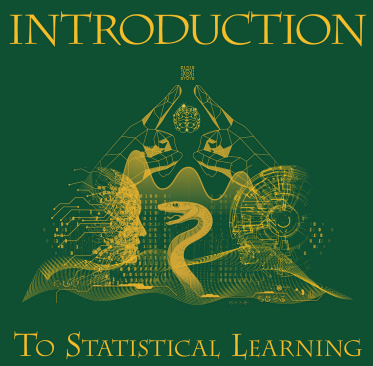
Introduction to Statistical Learning

Feat. Selection and Regularization - Class 8

Giora Simchoni

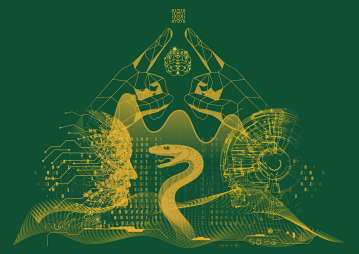
gsimchoni@gmail.com and add #intro2sl in subject

Stat. and OR Department, TAU



Goals of Selection and Regularization

INTRODUCTION



TO STATISTICAL LEARNING

Why select features? Why regularize?

E.g., did we not see the Gauss-Markov Theorem?

- Improve prediction accuracy
  - Recall: the Bias-Variance Tradeoff $\Rightarrow$ allowing some bias might decrease variance!
  - Recall: $op \approx \mathcal{O}\left(\frac{p\sigma^2}{n}\right)$
  - If $p > n$: $X^T X$ has no inverse, infinite solutions
- Improve interpretability
  - Discarding features with small "unlikely" coefficients

    $\Rightarrow$ lowering model complexity

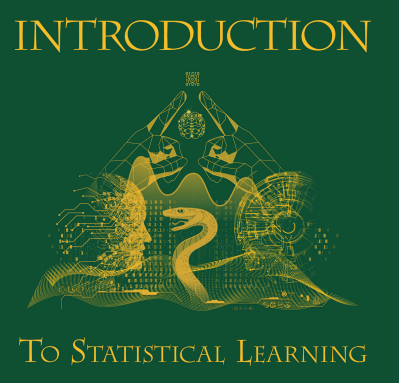
    $\Rightarrow$ parsimony!
  - Often coincides with improving prediction accuracy
- In general: "don't believe *everything* the data says"

Feature Selection and Regularization

We will focus on:

- Subset selection (best, stepwise, stagewise)
- Regularized regression (Ridge, Lasso)
- Dimensionality reduction (PCR)

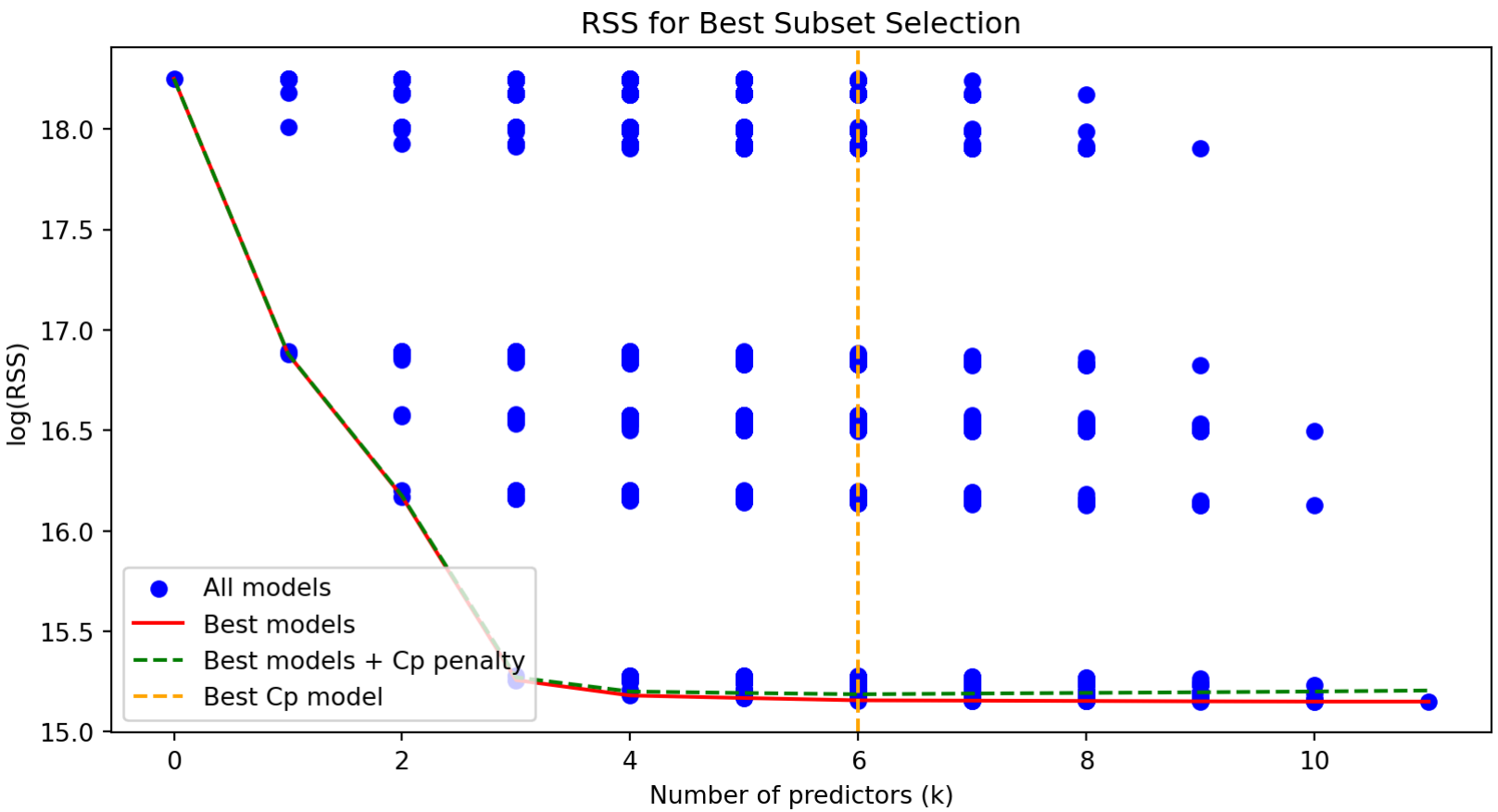
Subset Selection



Best subset selection

1. M_0 model: predict $\hat{y} = \hat{\beta}_0 = \bar{y}$
2. For $k = 1, \ldots, p$:
    i. Fit all $\binom{p}{k}$ models containing k features
    ii. Pick the best M_k with $\min RSS$
3. Select the best model from $M_0, \ldots, M_p$ with the C_p/AIC criterion or CV

Best subset selection

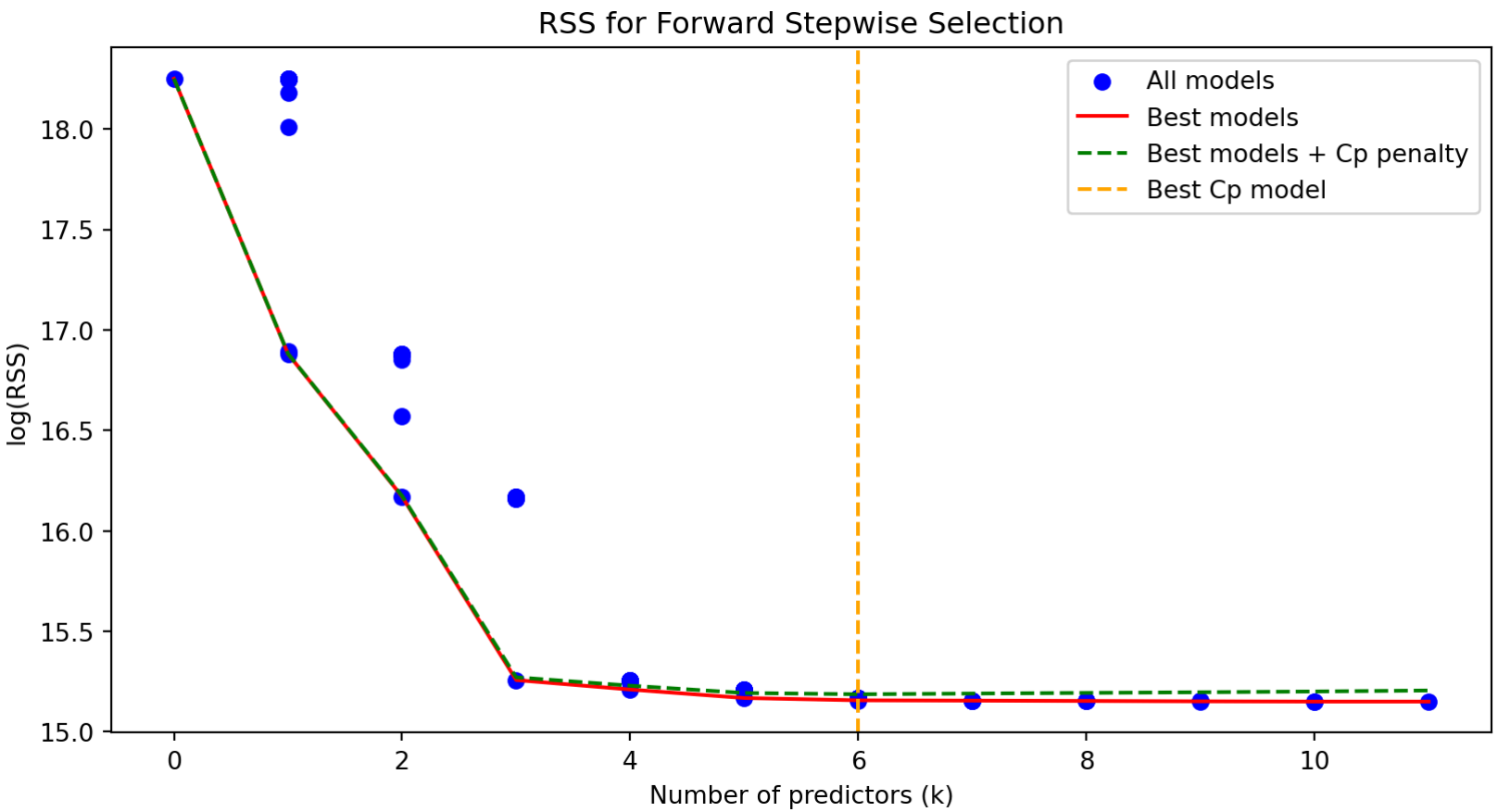


How many models are run?

Forward stepwise selection

1. M_0 model: predict $\hat{y} = \hat{\beta}_0 = \bar{y}$
2. For $k = 0, \ldots, p-1$:
    i. Fit all $p - k$ models adding 1 additional feature
    ii. Pick the best M_{k+1} with $\min RSS$
3. Select the best model from $M_0, \ldots, M_p$ with the C_p/AIC criterion or CV

Forward stepwise selection



How many models are run?

Stepwise regression main disadvantage

k	Best subset	Forward stepwise
1	{Rating}	{Rating}
2	{Rating, Income}	{Rating, Income}
3	{Rating, Income, Student}	{Rating, Income, Student}
4	{Income, Student, Limit, Cards}	{Rating, Income, Student, Limit}

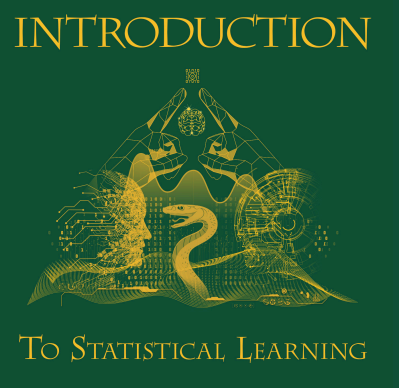
What happens if $p > n$?

Forward stagewise selection

0. Standardize all features, input some $\tau_{thresh} \in (0, 1)$ and $\varepsilon > 0$ step size
1. Residual $\mathbf{r} = \mathbf{y} - \bar{y}, \beta_1, \ldots, \beta_p = 0$
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$, and let $\tau = Corr(\mathbf{r}, \mathbf{x}_j)$
3. While $|\tau| > \tau_{thresh}$:
    i. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \varepsilon \cdot \text{sign}(\tau)$
    ii. Update $\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j$
    iii. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$, and let $\tau = Corr(\mathbf{r}, \mathbf{x}_j)$

Why would we want to “slow-learn”?

Regularized Regression: Ridge



Ridge regression

- Instead of reducing the number of parameters $\Rightarrow$ constrain them, penalize their norm
- With ℓ_2 norm we get the penalized RSS criterion for some regularization/penalty parameter $\lambda > 0$:

$$PRSS(\lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \|\beta^*\|_2^2$$

- Standardize the features in X, then:
  - $\hat{\beta}_0 = \bar{y}$
  - λ punishes features of different scale comparably

Ridge solution

$$PRSS(\lambda) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\frac{\partial PRSS}{\partial \beta} = -2X^T y + 2X^T X\beta + 2\lambda\beta$$

$$2X^T y - 2X^T X\beta - 2\lambda\beta = 0$$

$$(X^T X + \lambda I_p)\beta = X^T y$$

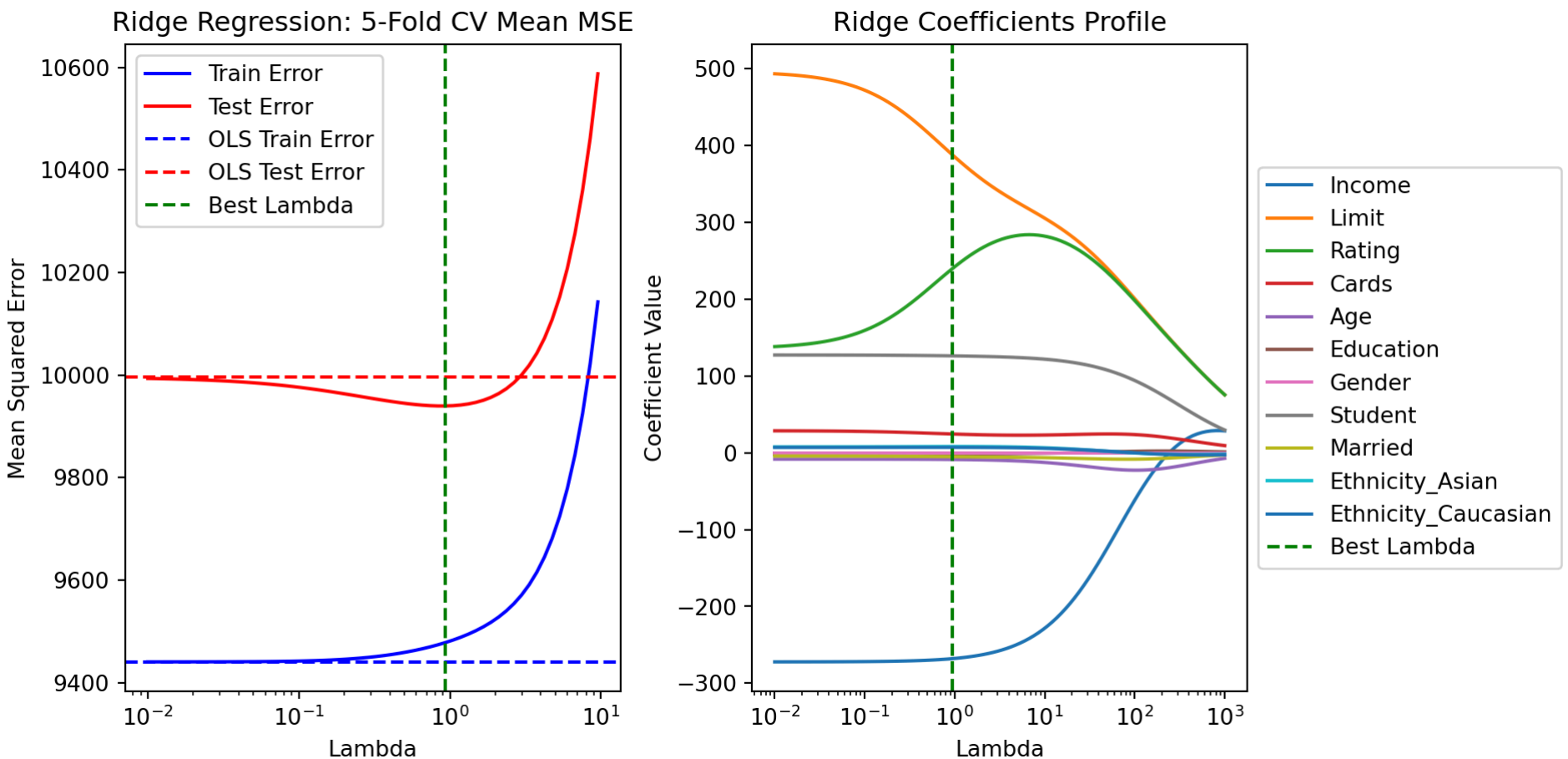
$$\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T y$$

Ridge (original) justification

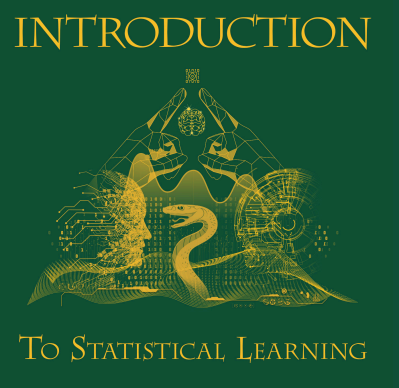
$$\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T y$$

- Numerical stability:
  - If X's columns are highly correlated, $X^T X$ is ill-conditioned and its inverse is unstable, variance of $\hat{\beta}$ is high – but $X^T X + \lambda I_p$ improves on this
- Feasibility:
  - If X's columns are linearly dependent $X^T X$ is not invertible but $X^T X + \lambda I_p$ is!
  - If $p > n$: same!
- Guaranteed prediction error reduction for some λ:
  - Bias increases but variance decreases *more*

Choosing the λ hyperparameter



Regularized Regression: Lasso

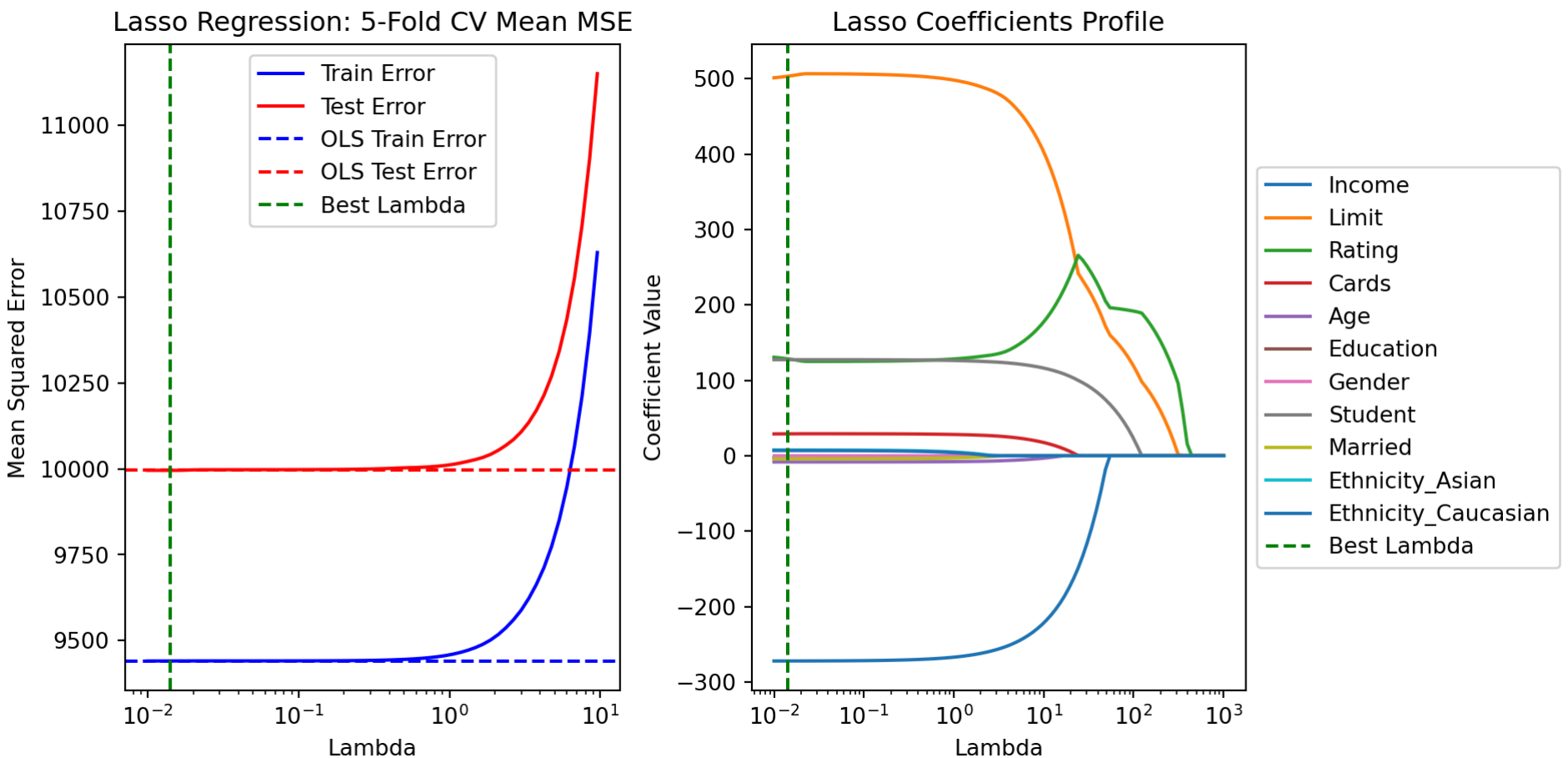


Lasso regression

$$PRSS(\lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \|\beta^*\|_1$$

- As with Ridge, standardize the features in X, then:
  - $\hat{\beta}_0 = \bar{y}$
  - λ punishes features of different scale comparably

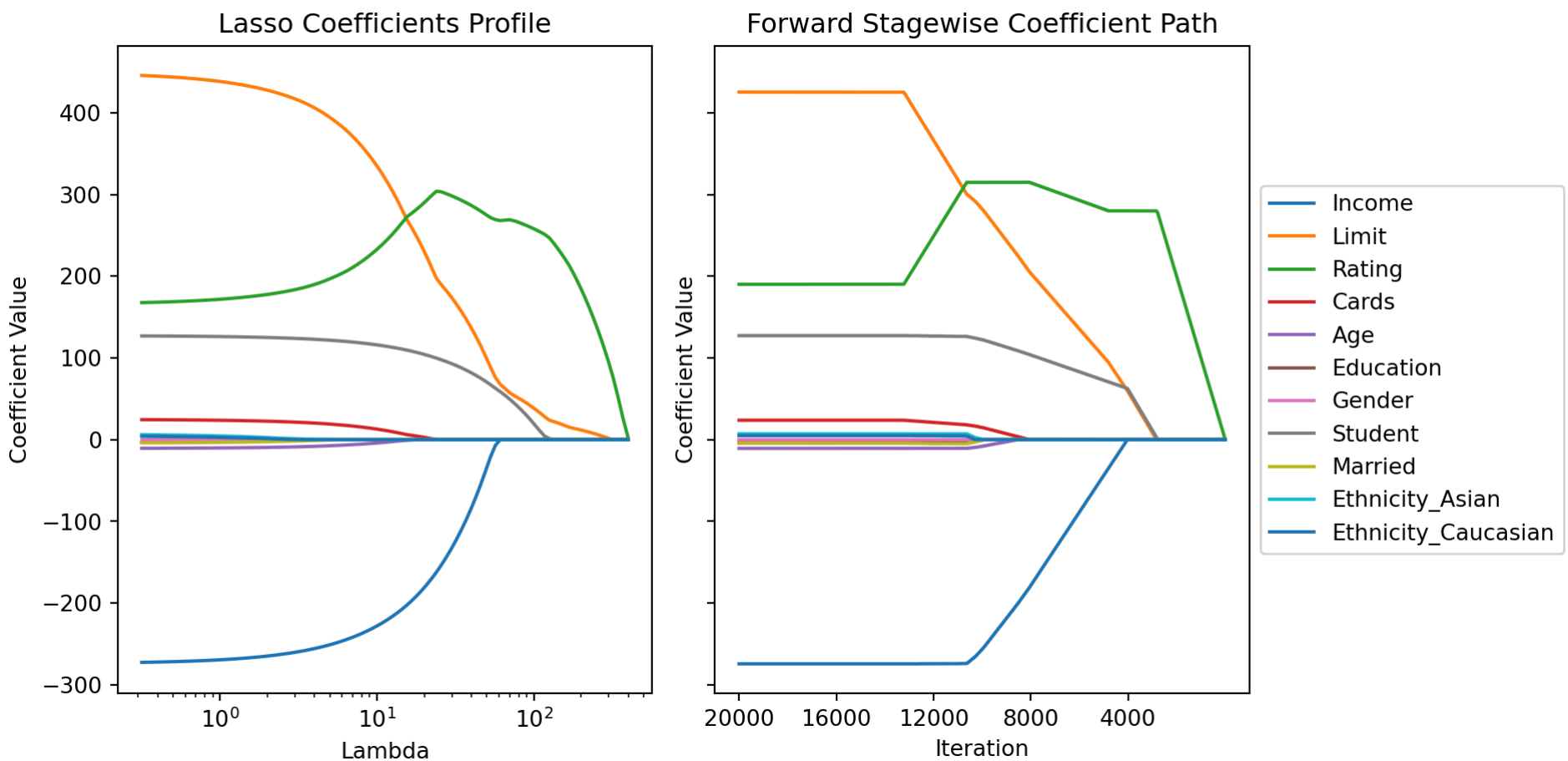
The Lasso Path



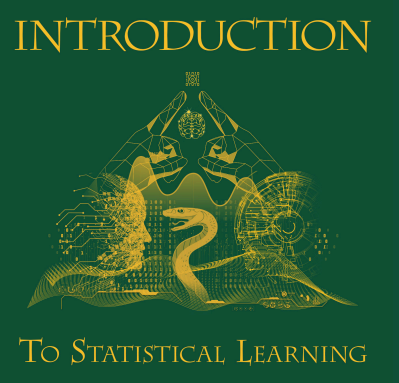
- Lasso's selling point: *sparsity*!
- Highly useful, especially for $p > n$ datasets

Lasso solution

- Quadratic Programming, LARS
- But also, surprisingly:



On Shrinkage and Sparsity



Dual criteria

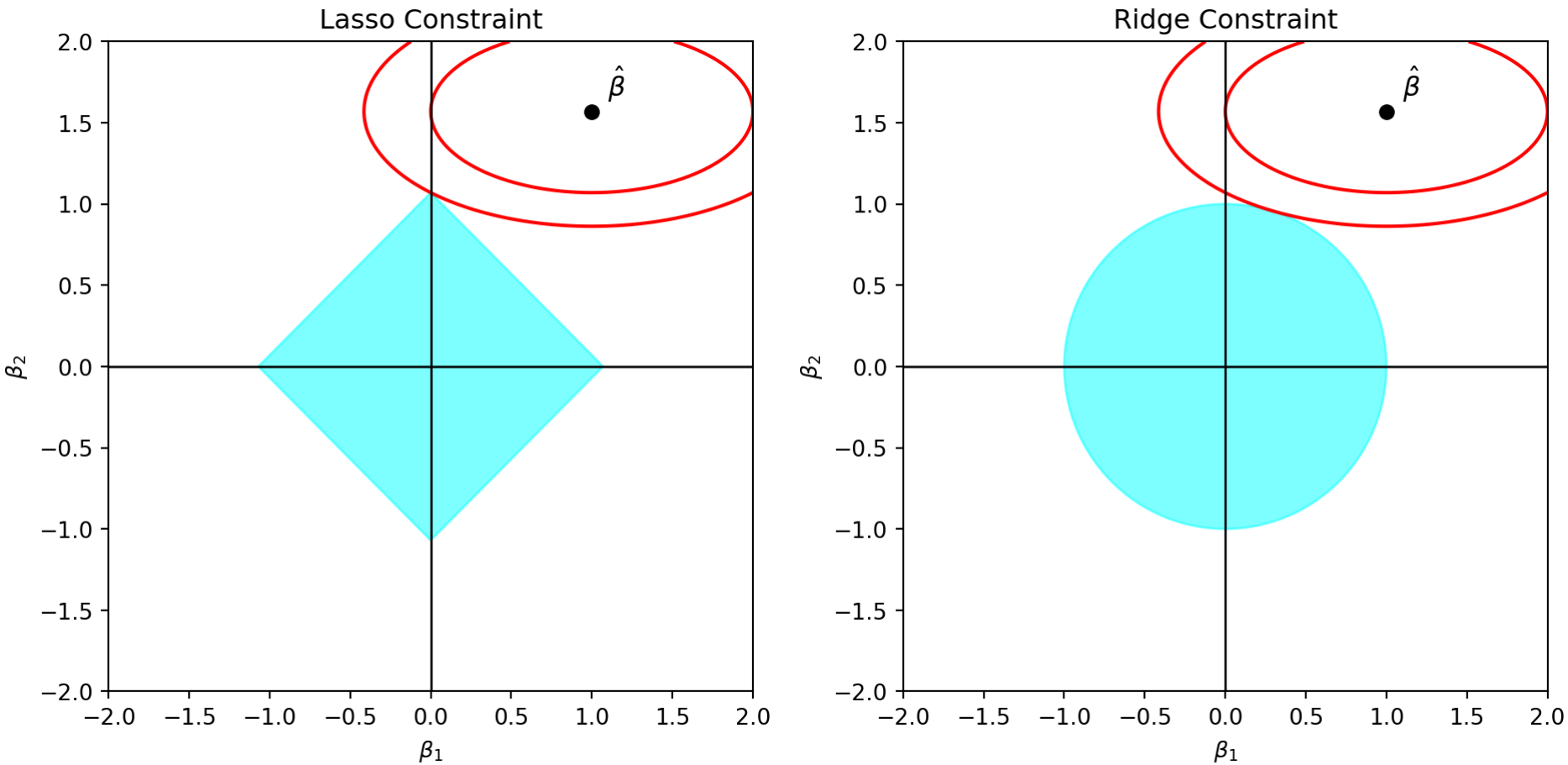
There is 1-to-1 correspondence between Ridge and Lasso's previous criteria and:

- Ridge: $\min_\beta RSS$ s.t. $\sum_{j=1}^{p} \beta_j^2 \leq s$
- Lasso: $\min_\beta RSS$ s.t. $\sum_{j=1}^{p} |\beta_j| \leq s$

In this context we can also write for Best Subset regression:

- $\min_\beta RSS$ s.t. $\sum_{j=1}^{p} \mathbb{I}[\beta_j \neq 0] \leq s$

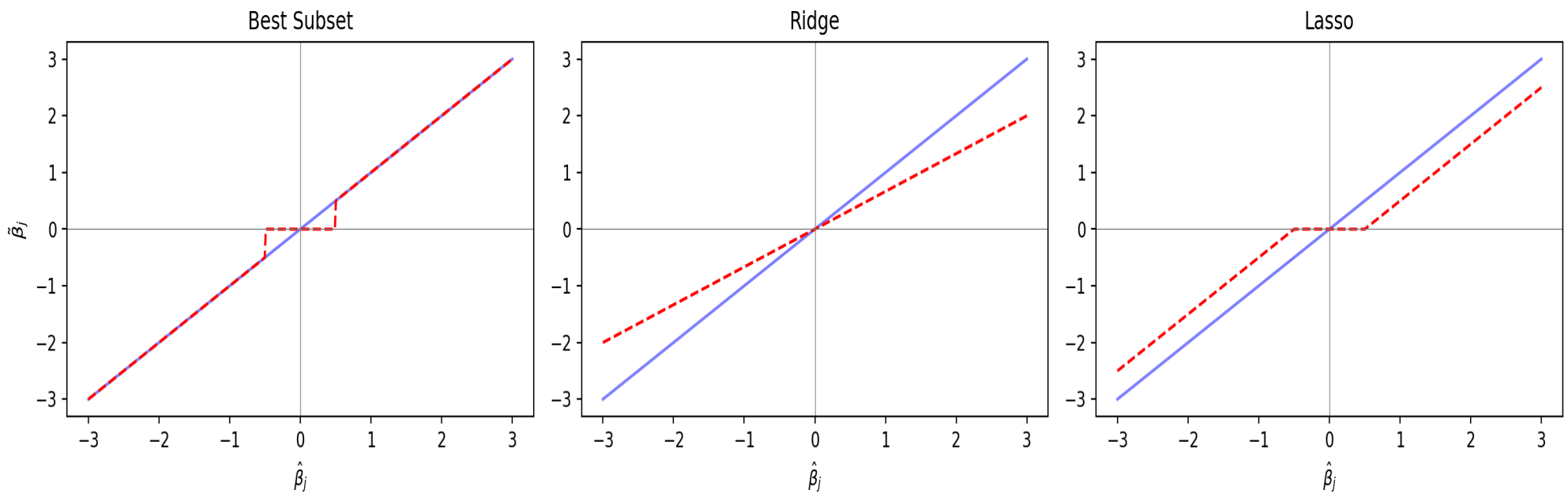
Seeing Shrinkage and Sparsity (I)



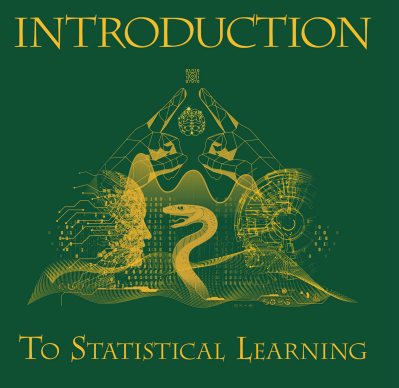
Seeing Shrinkage and Sparsity (II)

If X has orthonormal columns ($X^T X = I_p$) and $\hat{\beta}_j$ is the OLS estimator:

Estimator	$\tilde{\beta}_j$
Best M subset	$\hat{\beta}_j \cdot \mathbb{I}\left[\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert\right]$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathbf{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Bayesian Viewpoint



Conditional Distribution

- Recall Bayes Rule:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \text{ or posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

- For continuous distributions:

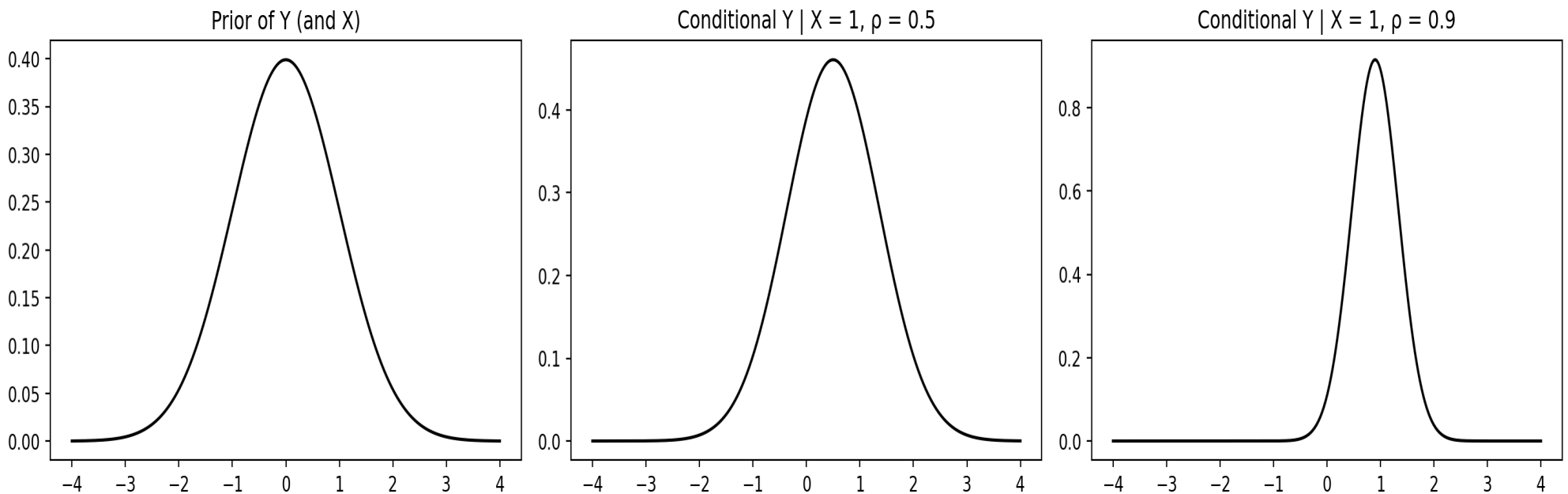
$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)} \propto f_{X|Y}(x|y) f_Y(y)$$

- $f_{Y|X}$ may not always have closed form, but sometimes…

Conditional Gaussian

For Gaussian distribution, if $X \sim \mathcal{N}(\mu_X, \sigma_X^2), Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \rho = Corr(X, Y)$:

$$Y|X = x \sim \mathcal{N}\left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$$



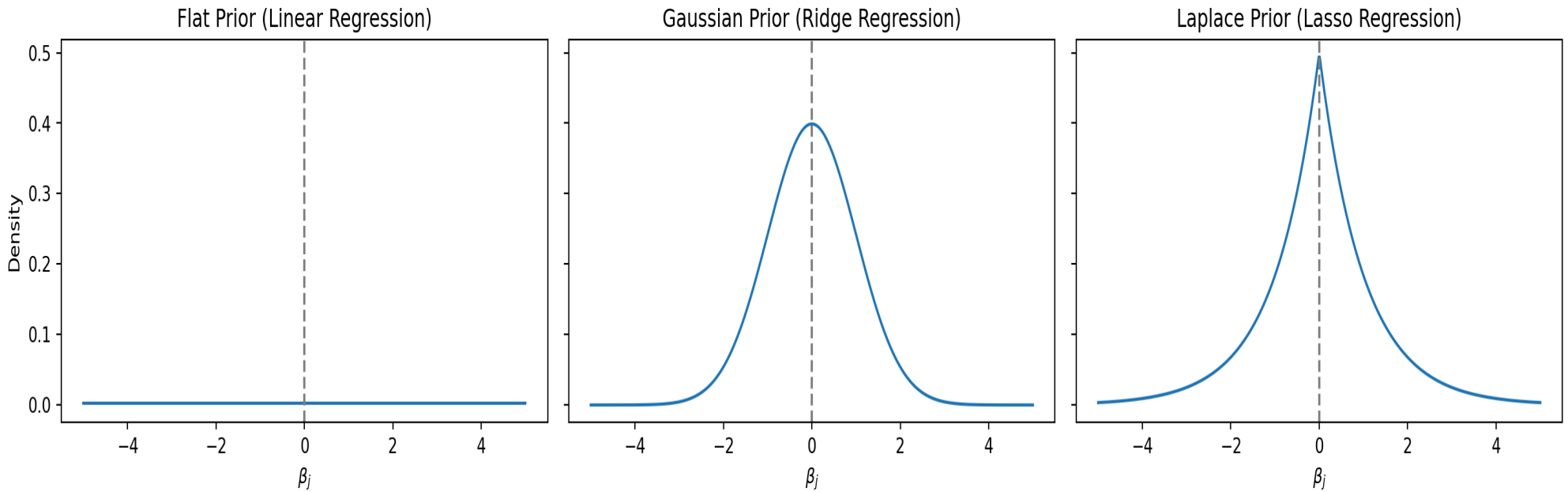
Bayesian Statistics

- In Bayesian statistics a parameter θ isn't a *fixed* number but a RV having a prior distribution $P(\theta)$
- In comes data sample $Y = y_1, \ldots, y_n$ with distribution $P(Y|\theta)$ (likelihood)
- We calculate the posterior distribution given the data $P(\theta|Y) \propto P(Y|\theta)P(\theta)$
- θ's final estimate is the mean or mode of $P(\theta|Y)$

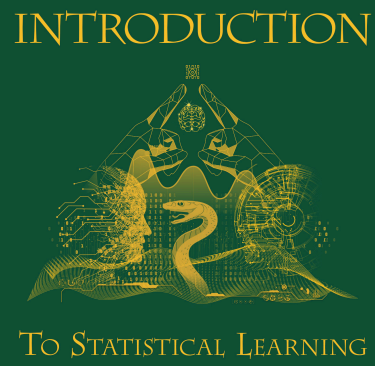
> Why would we take this approach?

- For linear regression and normal prior:
  - Prior: $\beta \sim \mathcal{N}(0, \tau^2 I_p)$
  - Data: $Y|\beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$
  - Posterior: $\beta|Y \sim \mathcal{N}\left((X^T X + \frac{\sigma^2}{\tau^2} I_p)^{-1} X^T y, \Sigma\right)$

Ridge and Lasso priors

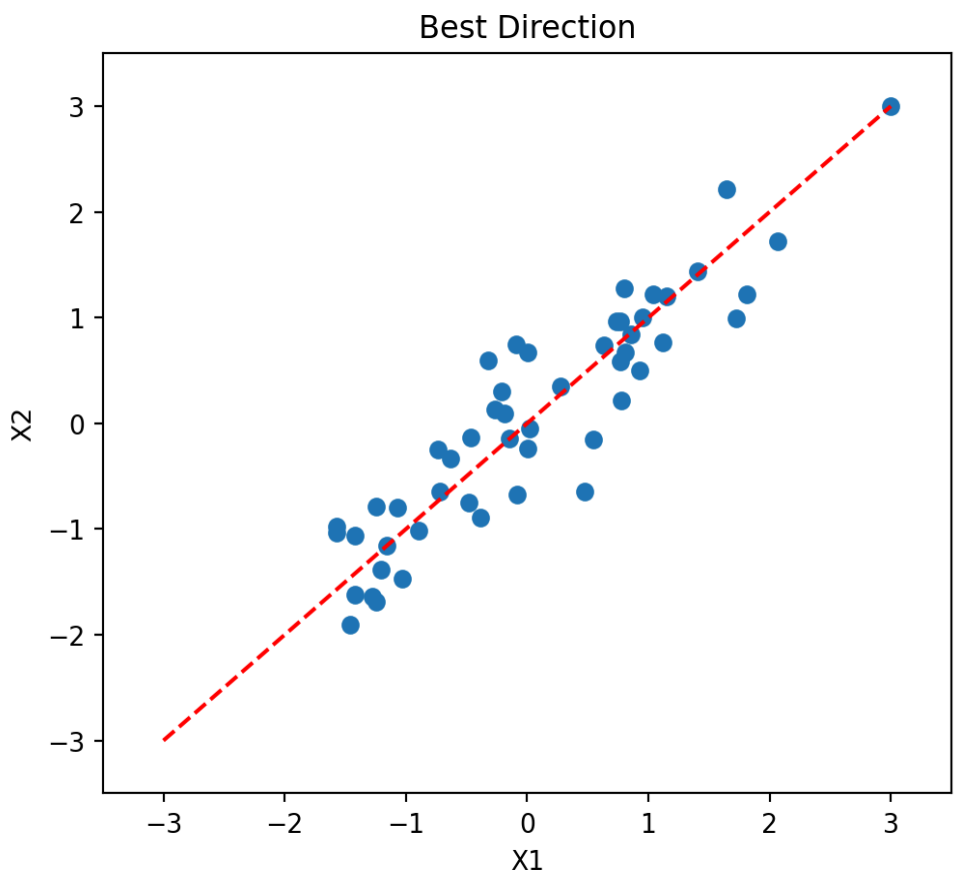


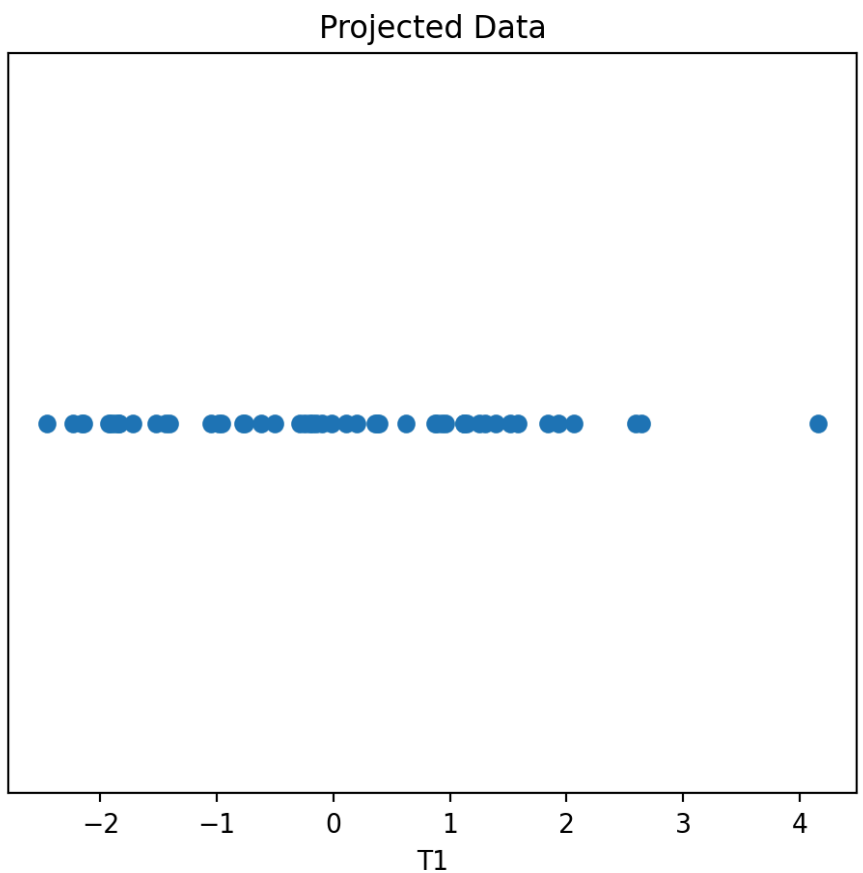
Dimensionality Reduction Methods: PCR



Recall: PCA

PCA will find the best direction to project the data on, while preserving the maximum "information":





The PCA Problem

- Goal: Find the q direction(s) with the most dispersion
- Projection is direction $\mathbf{v}$: $X\mathbf{v} \in \mathbb{R}^n$. Examples:
  - $\mathbf{v} = (1, 0, \ldots, 0)^T$: pick first coordinate from each observation
  - $\mathbf{v} = (1/\sqrt{p}, 1/\sqrt{p}, \ldots, 1/\sqrt{p})^T$: project on diagonal (average all coordinates)
- Dispersion in direction $\mathbf{v}$: $||X\mathbf{v}||^2 = \mathbf{v}^T(X^TX)\mathbf{v}$.
- Finding the best direction which maximizes dispersion: $\mathbf{v}_1 = \arg\max_{\mathbf{v}:\|\mathbf{v}\|^2=1} \left\|X\mathbf{v}\right\|^2$
- $\mathbf{v}_1$ is the first Principal Component direction: the best direction to project on!
- Similarly find the next PC directions $\mathbf{v}_2, \ldots, \mathbf{v}_q$ and stack them to matrix $W_{p\times q}$
- Data with reduced dimensionality:
  - $T_{n\times q} = X_{n\times p}W_{p\times q}$ taking only the first q principal directions

PCA regression (PCR)

- Standardize the features in X
- Find $W_{p \times q}$ (usually via SVD decomposition)
- Perform (regular) linear regression on $T_{n \times q} = XW$

$$y_i = \theta_0 + \theta_1 \cdot t_{i1} + \cdots + + \theta_q \cdot t_{iq} + \varepsilon_i \quad \text{or} \quad y = T\theta + \varepsilon$$

- q becomes a hyperparameter
- $y = T\theta + \varepsilon = X(W\theta) + \varepsilon \rightarrow \beta = W\theta$, hence β is still constrained
- SVD decomposition also shows similarities to Ridge regression

PCA regression (PCR)