P9120 - Statistical Learning and Data Mining

Lecture 3 - Model Selection / Variable Selection

Min Qian

Department of Biostatistics, Columbia University

September 19, 2024

Outline

Model Selection

- Training Sample: $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} = (\mathbf{X}, \mathbf{y})$.
- Models: $\{\mathcal{F}_m : m = 1, \ldots, M\}$.
  - # variables in best-subset selection
  - Tuning parameter λ in Ridge, Lasso, etc.
  - # of PCs in PCR.
  - ...
- Fit each model using training data yielding $\{\hat{f}_m : m = 1, \ldots, M\}$
- Choose the model m with smallest $PE(\hat{f}_m)$ or $EPE(\hat{f}_m)$.

How to estimate (expected) prediction error of $\hat{f}_m$?

Holdout (Simple Validation)

Split dataset into two parts:

- Training set: fit each model
- Validation set: estimated prediction error of each fitted model and select the one with minimal error.

Drawbacks:

- In a sparse dataset setting, we may not be able to afford the "luxury" of setting aside a proportion of the dataset for testing.
- Since it is a single train-and-validate experiment, the holdout estimate of error will be misleading if we happen to get an "unfortunate" split.

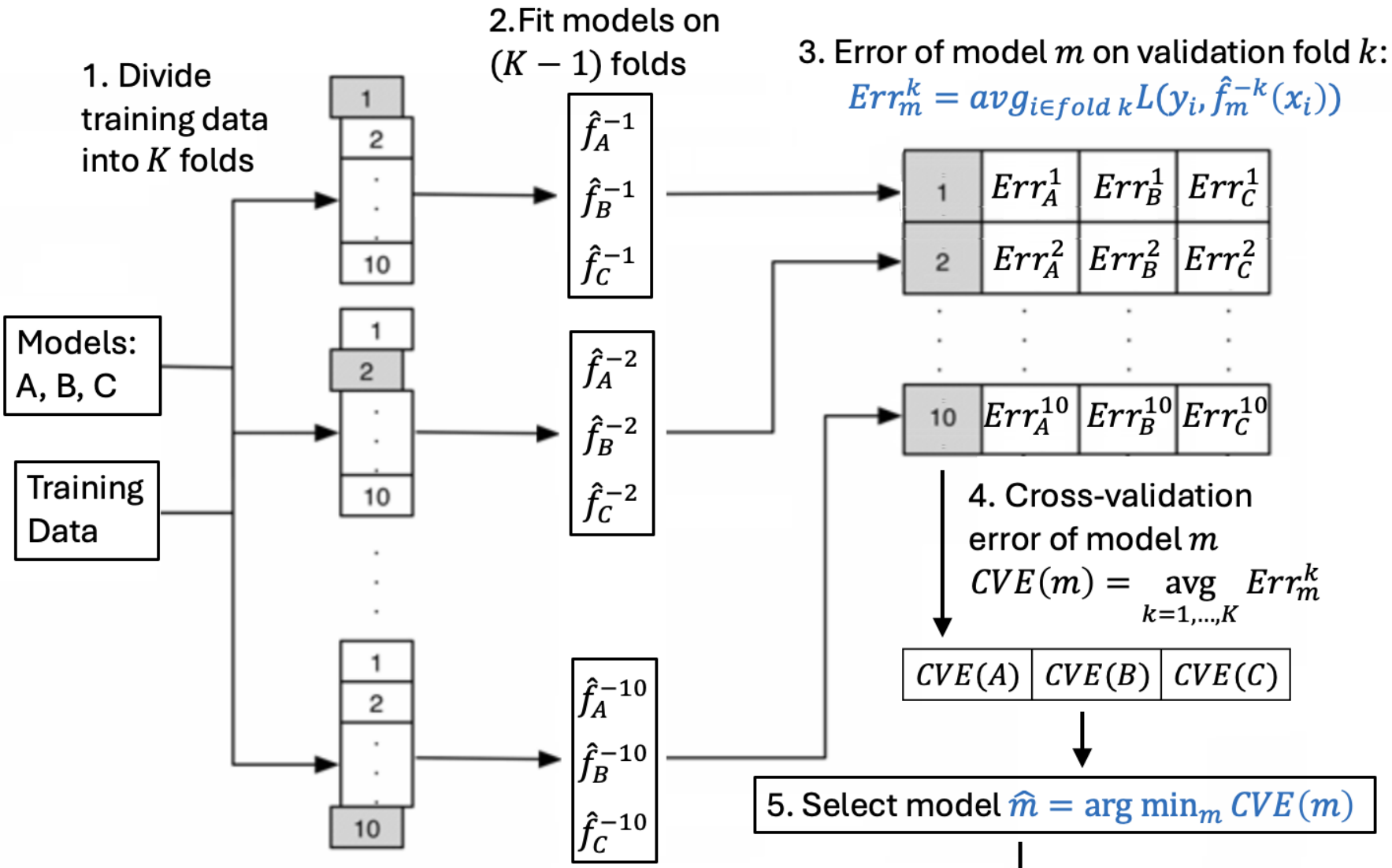
Repeated Random Sub-sampling Validation

- Randomly splits the dataset into training and validation data.
- For each such split, fit the model to the training data, and predictive accuracy is assessed using the validation data.
- The results are then averaged over the splits.

Remarks:

- Some observations may never be used for training or validation.
- Exhibits Monte Carlo variation, meaning that the results will vary if the analysis is repeated with different random splits.

K-fold Cross-Validation (usually $K = 5$ or 10)



Sometimes, CV with One Standard Error Rule

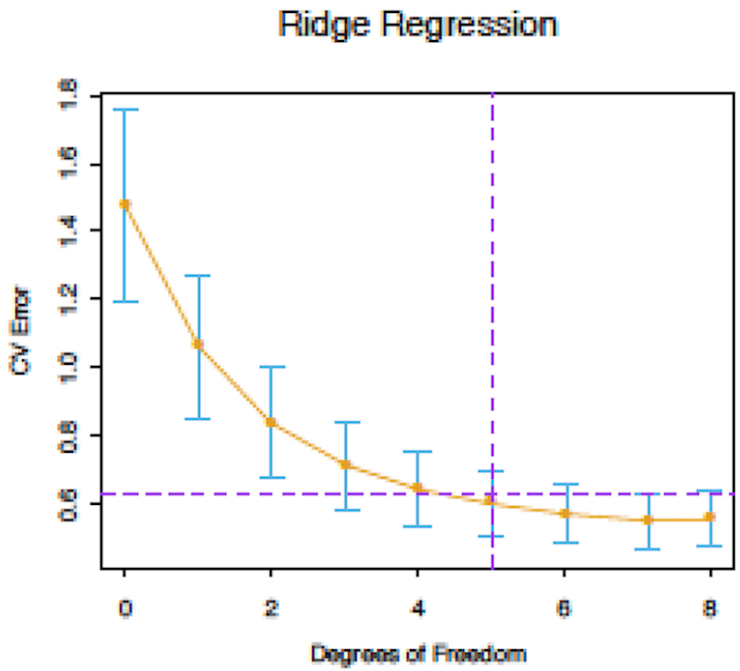
CV error of model m

$$CVE(m) = \frac{1}{K}\sum_{k=1}^{K} Err_m^k,$$

where $Err_m^k = \text{avg}_{i\in\text{fold }k} L(y_i, \hat{f}_m^{-k}(\mathbf{x}_i))$.

Model with minimal CV error:

$$\widehat{m} = \arg\min_m CVE(m).$$



One S.E. Rule finds the simplest (most regularized) model $\widetilde{m}$ whose CV error is within one standard error of the minimal CV error

$$CVE(\widetilde{m}) \leq CVE(\widehat{m}) + SE(\widehat{m}),$$

where $SE(m) = \frac{\text{SD}(Err_m^1,\ldots,Err_m^K)}{\sqrt{K}}$, i.e., the standard error of $CVE(m)$.

Leave-one-out Cross-Validation

Leave-one-out is the degenerate case of K-Fold CV with $K = n$.

$$LOOCV(m) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_m^{-i}(\mathbf{x}_i)]^2.$$

Bias-Variance trade off for $\widehat{EPE}$ (Discusssion@stackexchange)

- Common conception: LOOCV has smaller bias but larger variance than K-fold CV.
- For stable algorithms, LOOCV often performs the best.
- For unstable algorithms, have to think about bias and variance (in general k-fold may be better, but not the case for some biased estimators).

Generalized Cross-Validation

For many linear estimate (i.e. $\hat{f}_m(\mathbf{X}) = \mathbf{S}\mathbf{y}$) under squared-error loss,

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_m^{-i}(\mathbf{x}_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_m(\mathbf{x}_i)}{1 - S_{ii}}\right]^2,$$

where S_{ii} is the i-th diagonal element of $\mathbf{S}$ (see the book by Hastie and Tibshirani, 1990).

Approximating S_{ii} by $\text{trace}(\mathbf{S})/n$ yields

$$GCV(\hat{f}_m) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_m(\mathbf{x}_i)}{1 - \text{trace}(\mathbf{S})/n}\right]^2.$$

What's the advantage of GCV as compared to leave-one-out CV?

Relating GCV to C_p

$$GCV(\hat{f}_m) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_m(\mathbf{x}_i)}{1 - \text{trace}(\mathbf{S})/n}\right]^2.$$

Using the fact $1/(1-x)^2 \approx 1 + 2x$, we have

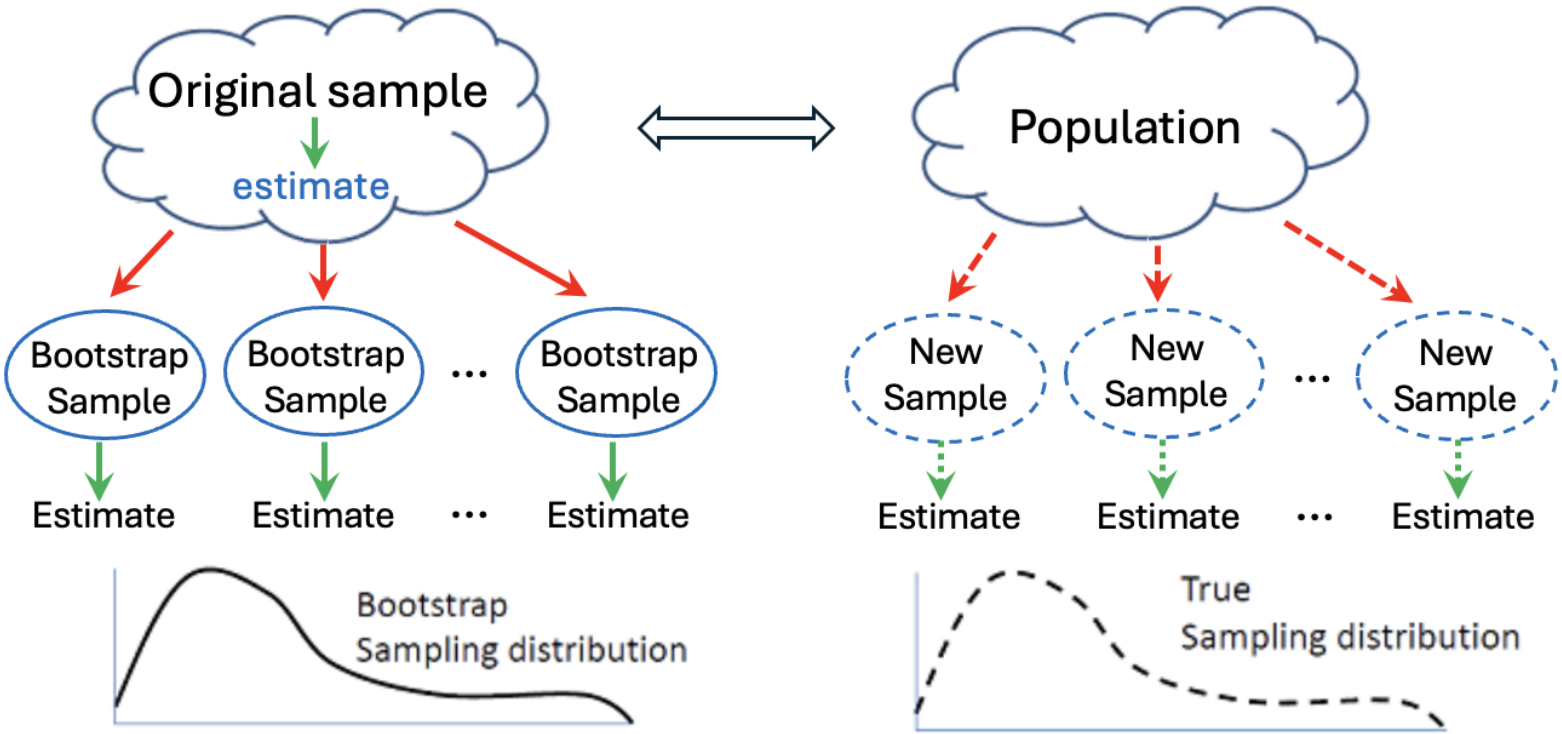
$$\begin{aligned} GCV(\hat{f}_m) \approx & \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_m(\mathbf{x}_i)]^2\left(1 + \frac{2\text{trace}(\mathbf{S})}{n}\right) \\ = & \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_m(\mathbf{x}_i)]^2 + \frac{2\text{trace}(\mathbf{S})}{n}\left(\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_m(\mathbf{x}_i)]^2\right) \end{aligned}$$

If $\hat{f}_m$ is OLS, then $\text{trace}(\mathbf{S}) = p$.

GCV is approximately equivalent to C_p.

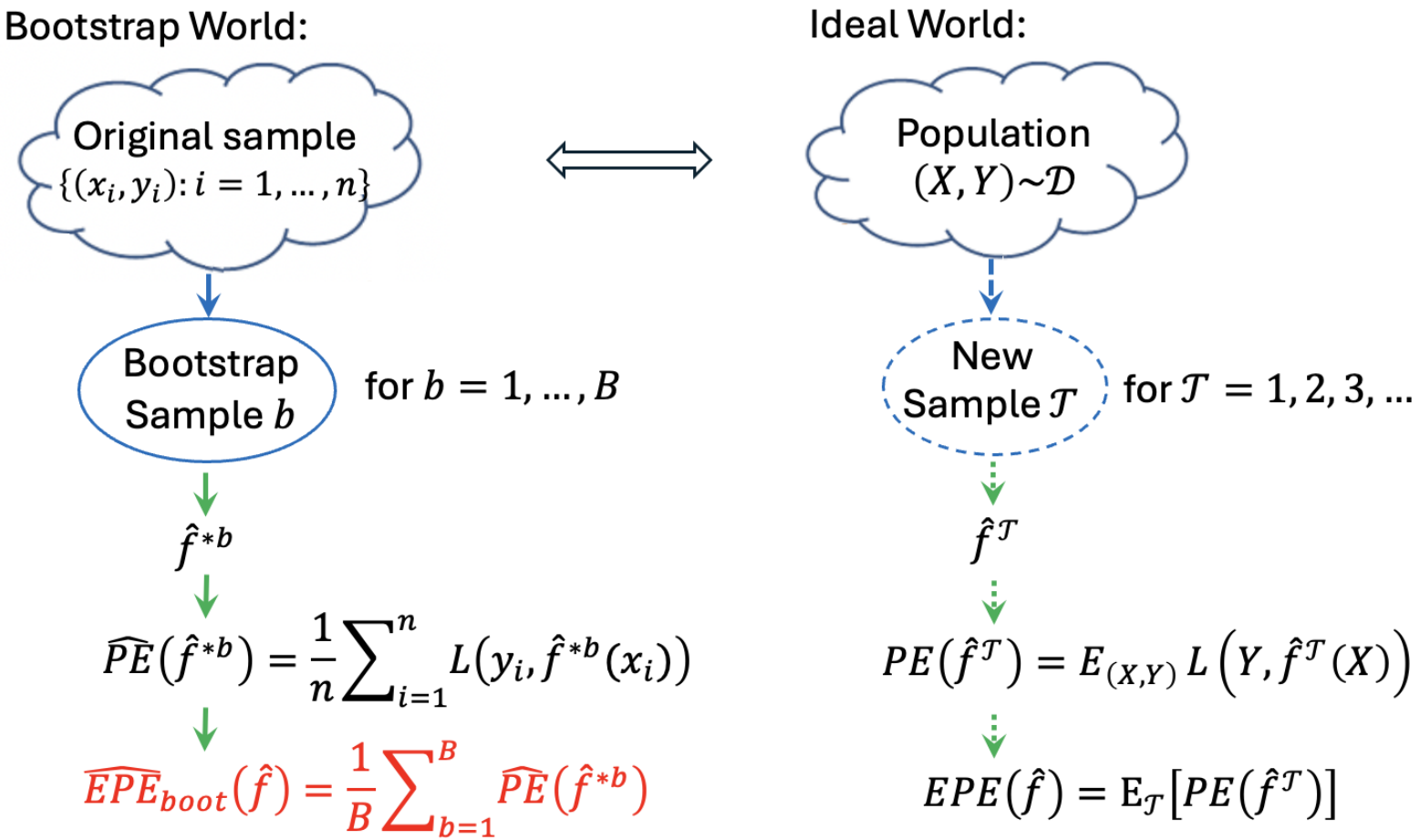
Bootstrap Schematic

A procedure for quantify the uncertainty of an estimator by resampling (often with replacement) one's data or a model estimated from the data.



Bootstrap estimate of prediction error

Original sample $\rightarrow \hat{f} \rightarrow$ How to estimate $\text{EPE}(\hat{f})$?



$\widehat{\text{PE}}^{*b}, b = 1, \ldots, B$ provides an estimate for the sampling dist. of $\text{PE}(\hat{f})$.

Issues with $\widehat{EPE}_{boot}$

1. Large overlap between original and bootstrap samples. Likely to under-estimate the prediction error!

2. Within each boot. sample, expected distinct obs. is about $0.632n$.

$$Pr(\text{obs. } i \in \text{boot. sample b}) = 1 - \left(1 - \frac{1}{n}\right)^n \approx 1 - e^{-1} = 0.632$$

The EPE of $\hat{f}^{*b}$ tends to be larger than EPE of $\hat{f}$ due to smaller sample size (upward bias).

Improved Bootstrap estimates of EPE

- For overlap issue, evaluate $\hat{f}^{*b}$ using data not in the boot. sample.

$$\widehat{\text{EPE}}^{(1)} = \frac{1}{B}\sum_{b=1}^{B}\left[avg_{i\notin b}L(y_i, \hat{f}^{*b}(\mathbf{x}_i))\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}L(y_i, \hat{f}^{*b}(\mathbf{x}_i)).$$

- For upward bias issue,

$$\widehat{\text{EPE}}^{(0.632)} = 0.368 \times \text{TE}(\hat{f}) + 0.632 \times \widehat{\text{EPE}}^{(1)}.$$

based on the fact that each boot. sample only used $0.632n$.

- An improved version:

$$\widehat{\text{EPE}}^{(0.632+)} = (1-\hat{w}) \times \text{TE}(\hat{f}) + \hat{w} \times \widehat{\text{EPE}}^{(1)},$$

where $\hat{w}$ is a scaled weight measures the amount of overfitting $(\widehat{\text{EPE}}^{(1)} - \text{TE}(\hat{f}))$.

Three Forms of Bootstrap Sampling

- Non-parametric bootstrap
  - A bootstrap sample is formed by sampling with replacement from the original data.
- Parametric bootstrap
  - Fit a parametric model with the original data;
  - A bootstrap sample is formed by generating data from the fitted model.
- Residual bootstrap (used in regression)
  - Fit a regression model with the original training set, $\hat{f}$.
  - Compute residuals: $e_i = y_i - \hat{f}(\mathbf{x}_i)$, $i = 1, \ldots, n$.
  - Sample with replacement from $e_1, \ldots, e_n$, denoted by $e_1^*, \ldots, e_n^*$.
  - A bootstrap sample is formed by $(\mathbf{X}, \mathbf{y}^*)$, where $y_i^* = \hat{f}(\mathbf{x}_i) + e_i^*$.

Bootstrap for Statistical Inference

- Parameter of interest θ (e.g., population mean).
- Training sample estimate $\hat{\theta}$ (e.g., sample mean).
- Bootstrap estimate $\hat{\theta}^{*b}$ (e.g., bootstrap sample mean), $b = 1, \ldots, B$.

$(1 - \alpha) \times 100\%$ confidence intervals for θ:

- Standard normal: $\hat{\theta} \pm Z_{\alpha/2}\hat{\sigma}$, where

$$\hat{\sigma}^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^{*b} - avg(\hat{\theta}^{*b})\right]^2$$

- Percentile bootstrap: $(\hat{\theta}^{*b}_{\alpha/2}, \hat{\theta}^{*b}_{1-\alpha/2})$, where $\hat{\theta}^{*b}_{\alpha/2}$ is the lower $\alpha/2$-percentile of the empirical cumulative distribution of $\hat{\theta}^{*b}$.

- Hybrid bootstrap (centered percentile bootstrap): $(2\hat{\theta} - \hat{\theta}^{*b}_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^{*b}_{\alpha/2})$.

Bootstrap: Strength and Weakness

Strength:

- It is a straightforward way to derive estimates of standard errors, confidence intervals and other quantities when the theoretical distribution of a statistic of interest is complicated or unknown.

Potential risks/weakness:

- The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (e.g. independence of samples) where these would be more formally stated in other approaches.
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees.
- Computationally intensive (not a problem in most cases now!)

Model Selection Summary

Key: Usually involves Bias-variance trade off.

- Test-based methods: traditional, but still widely used.
- Criterion based methods (C_p, AIC, BIC): theory derived under fixed design matrix, but are widely used in general settings.
- Cross-validation: most common model selection method.
- Bootstrap: can be used to estimate EPE, but is more commonly applied for statistical inference.

Variable Selection

Variable selection methods learned so far:

- Forward-, Backward-, Stepwise selection
- Best subset selection: C_p, AIC, BIC, CV, Bootstrap

From now on,

- focus on linear model with square error loss;
- use $\hat{\boldsymbol{\beta}}$ to denote the OLS estimate;
- simultaneous parameter estimation and variable selection.

Ridge vs. Best-subset Selection

Ridge regression: least squares with an l_2-penalty.

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Best-subset selection:

$$\hat{\boldsymbol{\beta}}^{\text{subset}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \right\}.$$

Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996)

Least squares with an l_1-penalty:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Shrinkage
- Sparsity: some fitted coefficients are exactly zero

Continuous variable selection

Mean Zero Version

Assume $\sum_{i=1}^n y_i = 0$ and $\sum_{i=1}^n x_{ij} = 0$ for $j = 1, \ldots, p$.

$$\hat{\boldsymbol{\beta}}^{\text{subset}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p 1_{\beta_j \neq 0} \right\}.$$

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}.$$

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

(We can always center $\mathbf{y}$ and each column of $\mathbf{X}$ first.)

Orthonormal $\mathbf{X}$

In general, the Lasso solution doesn't have a closed form.

Under orthonormal design ($\mathbf{X}^T\mathbf{X} = I$),

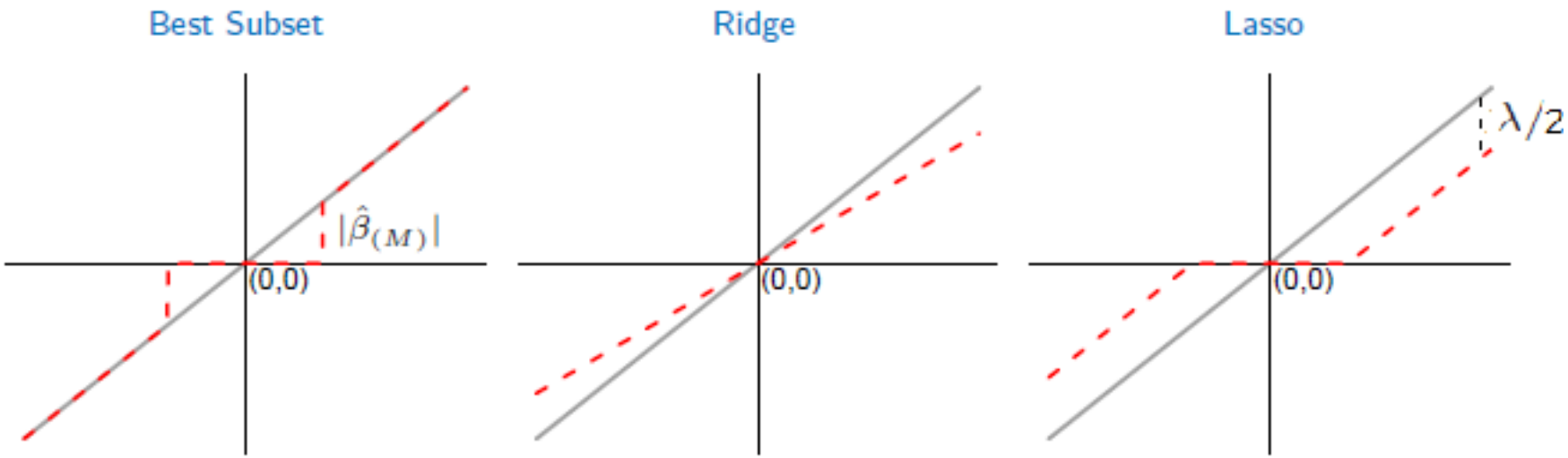
$$\begin{aligned}\hat{\beta}_j^{\text{lasso}}(\lambda) &= \begin{cases} \hat{\beta}_j - \lambda/2, & \text{if} \quad \hat{\beta}_j > \lambda/2 \\ 0, & \text{if} \quad |\hat{\beta}_j| \leq \lambda/2 \\ \hat{\beta}_j + \lambda/2 & \text{if} \quad \hat{\beta}_j < -\lambda/2 \end{cases} \\ &= \text{sign}(\hat{\beta}_j)\left(\hat{\beta}_j - \frac{\lambda}{2}\right)_+ .\end{aligned}$$

- Lasso shrinks large coefficients by a constant.
- Lasso truncates small coefficients to zero.

($\hat{\beta}_j$ is the OLS estimate.)

Orthonormal $\mathbf{X}$

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda/2)_+$



($\hat{\beta}_{(M)}$ is the M-th largest OLS coefficient in magnitude.)

General $\mathbf{X}$

Ridge:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(t) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ s.t. } \sum_{j=1}^{p} \beta_j^2 \leq t$$

Lasso:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}(t) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ s.t. } \sum_{j=1}^{p} |\beta_j| \leq t$$

Note that

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)^T \mathbf{X}^T \mathbf{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) + \text{constant}$$

General $\mathbf{X}$ $(p = 2)$

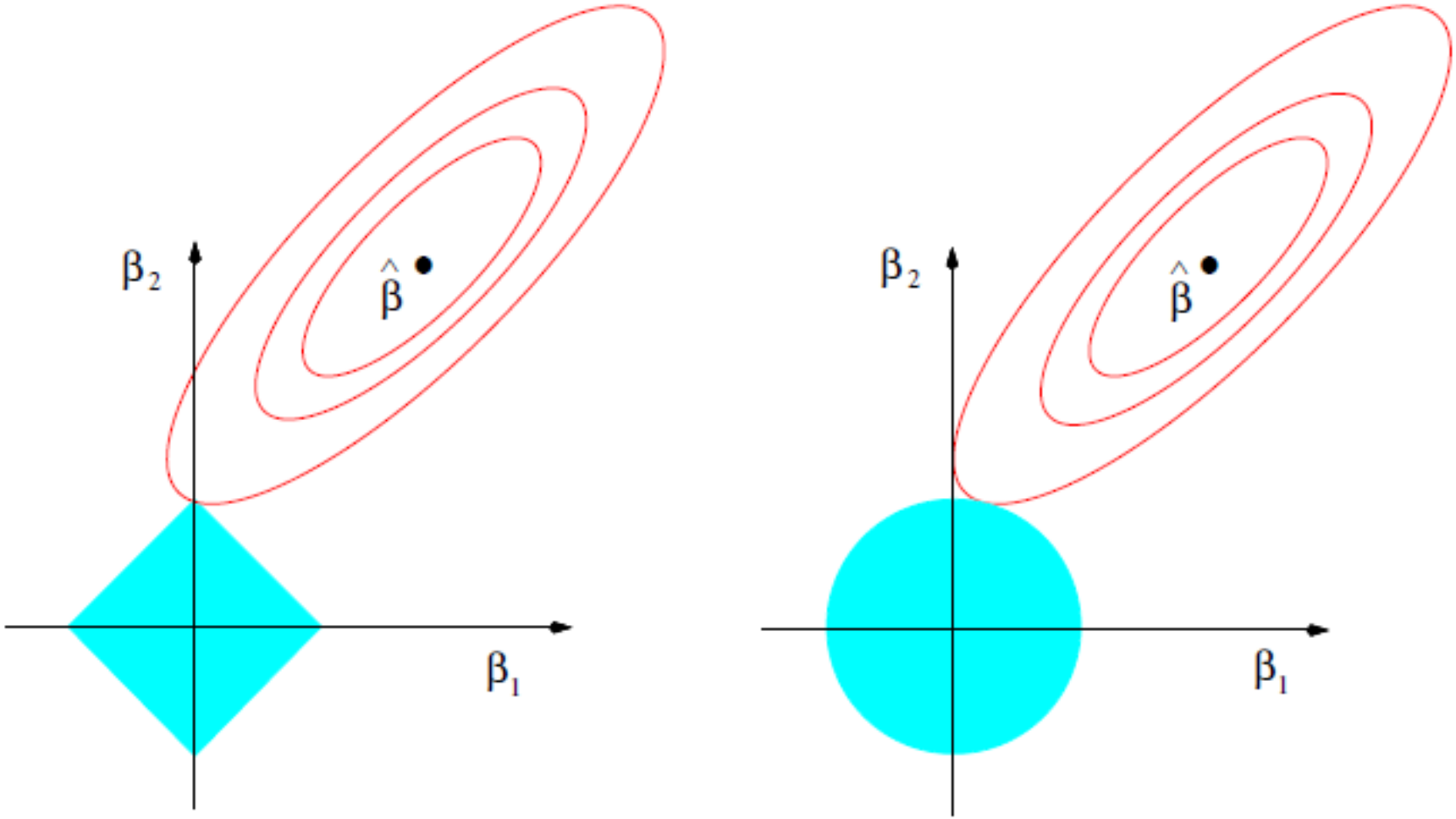
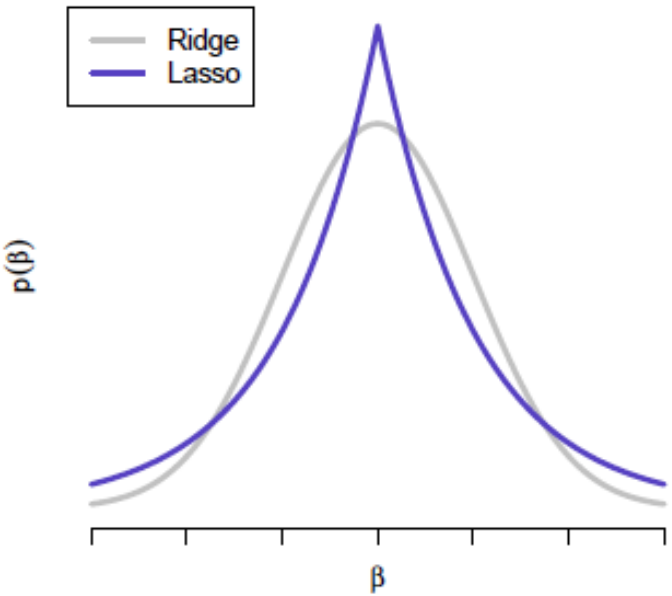


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Bayesian Perspective

Gaussian model: $\mathbf{y} \sim N(\beta_0 + \mathbf{X}^T\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

	Prior of $\boldsymbol{\beta}$	-2log(posterior dist. of $\boldsymbol{\beta}$) $\propto$
OLS	Uniform	$\frac{1}{\sigma^2}\sum_{i=1}^n \left(y_i - \beta_0 - x_i\boldsymbol{\beta}\right)^2$
Ridge	$N(0, \tau^2\mathbf{I}_p)$	$\frac{1}{\sigma^2}\sum_{i=1}^n \left(y_i - \beta_0 - x_i\boldsymbol{\beta}\right)^2 + \frac{1}{\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}$
Lasso	$\prod_{j=1}^p \left[\frac{1}{2\tau}\exp\left(-\frac{\lvert\beta_j\rvert}{\tau}\right)\right]$	$\frac{1}{\sigma^2}\sum_{i=1}^n \left(y_i - \beta_0 - x_i\boldsymbol{\beta}\right)^2 + \frac{2}{\tau}\lVert\boldsymbol{\beta}\rVert_1$



Ridge prior (normal) is centered around 0, so the posterior mode is likely to shrink towards zero.

Lasso prior (double exponential) is "pointy" at 0, so there is a chance that the posterior mode will be identically zero.

Extension of Lasso

- Group Lasso: $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G)$

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{\mathbf{g=1}}^{\mathbf{G}} \|\boldsymbol{\beta}_{\mathbf{g}}\|_{\mathbf{2}}$$

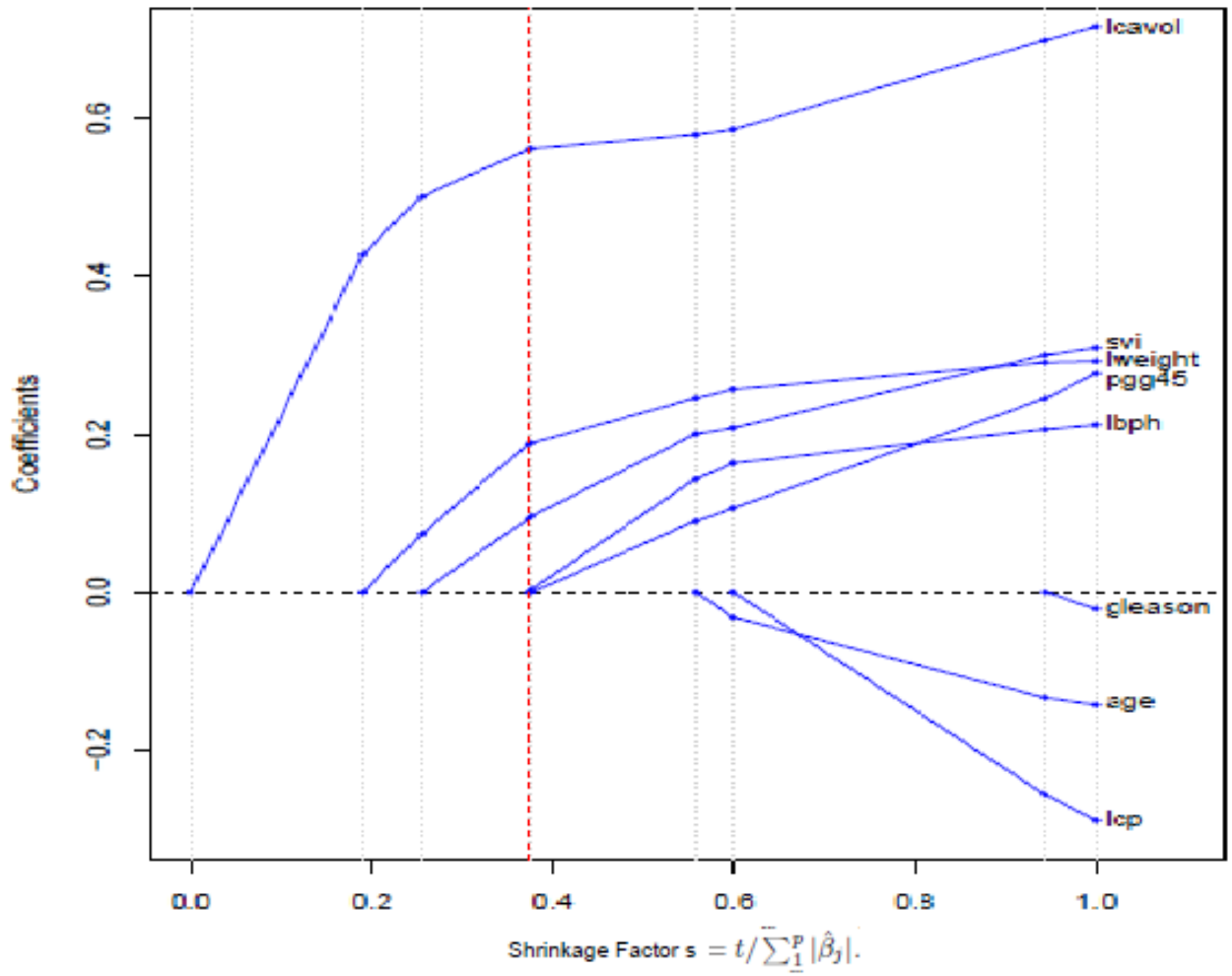
- Generalized Lasso:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\mathbf{D}\boldsymbol{\beta}\|_{\mathbf{1}}$$

example: fused lasso

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{\mathbf{j=2}}^{\mathbf{p}} |\beta_{\mathbf{j}} - \beta_{\mathbf{j-1}}|$$

Lasso Solution Path: Piecewise Linear



The tuning parameter t (or λ) can be selected using CV, GCV, ...

Effective Degrees of Freedom

- For a linear model with OLS, the degrees of freedom is the number of free parameters to vary.

Diff between Expected PE and TE:

$$EPE_{in}(\hat{f}) - \mathrm{E}_{\mathcal{T}} TE(\hat{f}) = \frac{2}{n}\sum_{i=1}^{n} Cov\left(y_i, \hat{f}(x_i)\right) = \frac{2}{n}\mathrm{trace}(\boldsymbol{S})\sigma^2 = \frac{2}{n}p\sigma^2$$

1. 1
2. 2

1. 1
2. 2
3. 3

1. 1
2. 2
3. 3
4. 4

Effective Degrees of Freedom:

$$df(\hat{f}) \triangleq \frac{\sum_{i=1}^{n} Cov\left(y_i, \hat{f}(x_i)\right)}{\sigma^2} = \mathrm{trace}(\boldsymbol{S}) = p$$

1. Square error loss
2. Additive model $Y = f(X) + \epsilon$, with $E(\epsilon) = 0, var(\epsilon) = \sigma^2$
3. Linear estimate $\hat{f}(\mathbf{X}) = \boldsymbol{S}\boldsymbol{y}$
4. OLS

Ridge regression:

$$df(\lambda) = trace(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T)$$
$$= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

How to Choose λ using C_p, AIC or BIC

$$C_p(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)\right)^2 + \frac{2}{n}df\left(\hat{\mathbf{y}}^{\text{lasso}}(\lambda)\right)\sigma^2.$$

Zou et al. (2007): when $\mathbf{X}$ has full column rank,

$$df\left(\hat{\mathbf{y}}^{\text{lasso}}(\lambda)\right) \approx \ \#\text{ of nonzero elements in } \hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$$

Similarly,

$$AIC(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)\right)^2\right] + \frac{2}{n}df\left(\hat{\mathbf{y}}^{\text{lasso}}(\lambda)\right)$$

$$BIC(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)\right)^2\right] + \frac{\log(n)}{n}df\left(\hat{\mathbf{y}}^{\text{lasso}}(\lambda)\right).$$

General l_q Penalty (Bridge Regression)

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j|^q \leq t$$

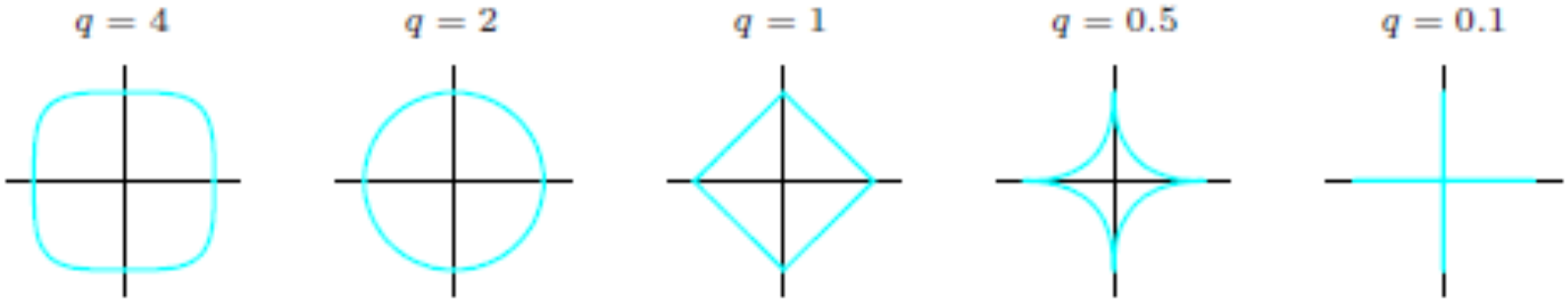


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

- $q = 2$: ridge regression
- $q = 1$: lasso, the smallest q such that the constraint region is convex.
- $q = 0$: best-subest (define $0^q \to 0$ as $q \to 0$)
- For $q \leq 1$, more weights are imposed on the coordinate directions.

Penalized Least Squares

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \text{pen}_\lambda(|\beta_j|) \right]$$

Fan and Li (2001) advocated penalty functions with three properties:

- Unbiasedness: The resulting estimator should have low bias, especially when the true coefficient β_j is large.
- Sparsity: The resulting estimator should automatically set small estimated coefficients to zero to accomplish variable selection.
- Continuity: The resulting estimator should be continuous in data to reduce instability in model prediction.

Penalized Least Squares under Orthonormal $\mathbf{X}$

$$\min_{\beta_j} \left[\frac{1}{2}(\beta_j - \hat{\beta}_j)^2 + \text{pen}_\lambda(|\beta_j|) \right]$$

Note that the solution satisfies

$$\frac{\partial \left[\frac{1}{2}(\beta_j - \hat{\beta}_j)^2 + \text{pen}_\lambda(|\beta_j|) \right]}{\partial \beta_j} = \text{sign}(\beta_j) \left(|\beta_j| + \text{pen}'_\lambda(|\beta_j|) \right) - \hat{\beta}_j = 0$$

- Unbiasedness
- Sparsity/Selection
- Continuity

l_q Penalty

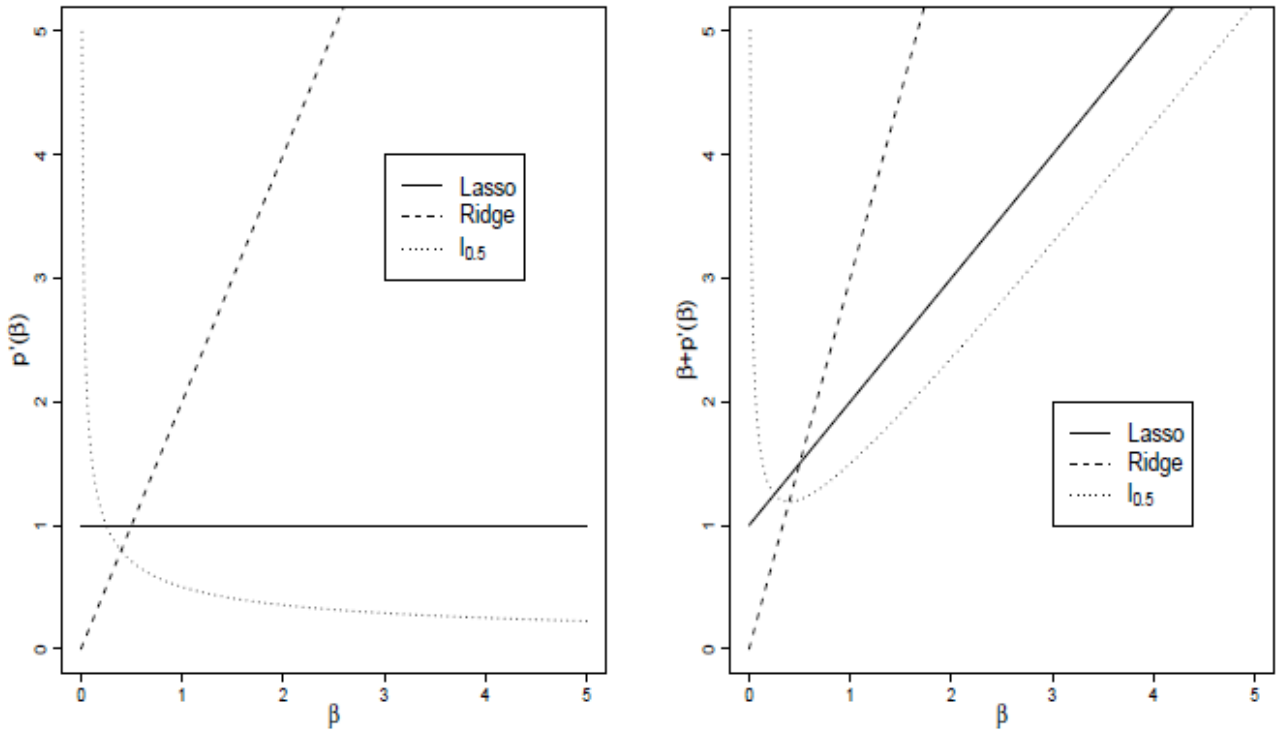


Figure: Plot of $\text{pen}'_\lambda(\beta)$ (left) and $\beta + \text{pen}'_\lambda(\beta)$ (right) for $\lambda = 1$.

- Unbiasedness if $\text{pen}'_\lambda(\beta) = 0$ for large positive β.
- Sparsity if $\min_{\beta:\beta\geq 0}[\beta + \text{pen}'_\lambda(\beta)] > 0$.
- Continuity if and only if $\arg\min_{\beta:\beta\geq 0}[\beta + \text{pen}'_\lambda(\beta)] = 0$.

SCAD (Fan and Li, 2001)

Smoothly clipped absolute deviation (SCAD) penalty:

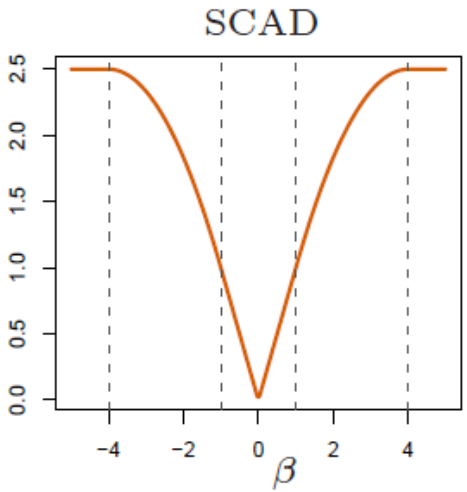
$$\min_{\boldsymbol{\beta}} \frac{1}{2}\left[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \sum_{j=1}^{p}\text{pen}_\lambda(|\beta_j|)\right],$$

where

$$\text{pen}_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| < 2\lambda \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & \text{if } 2\lambda \le |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| \ge a\lambda \end{cases}$$

for some $a > 2$.

- Can choose λ and a using CV or GCV. etc.
- $a = 3.7$ is suggested in the paper.



SCAD Penalty

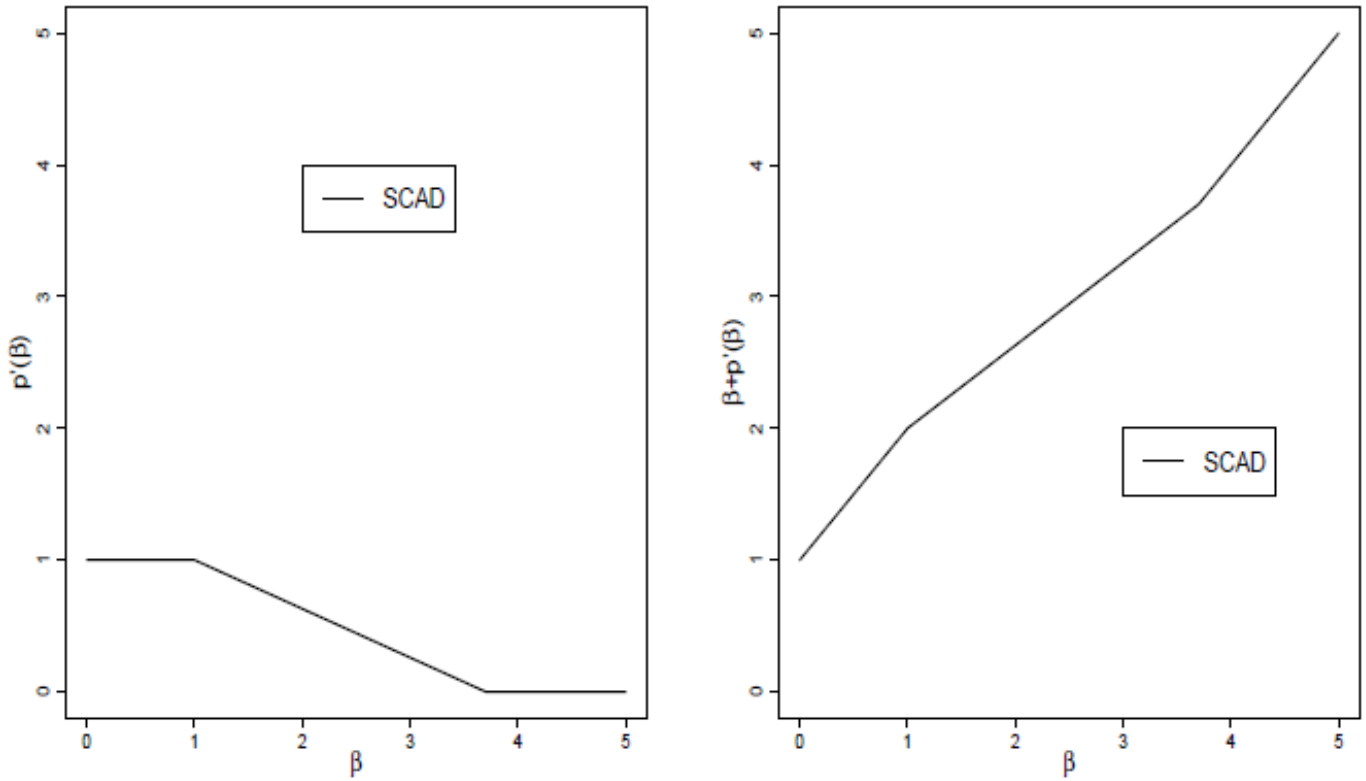
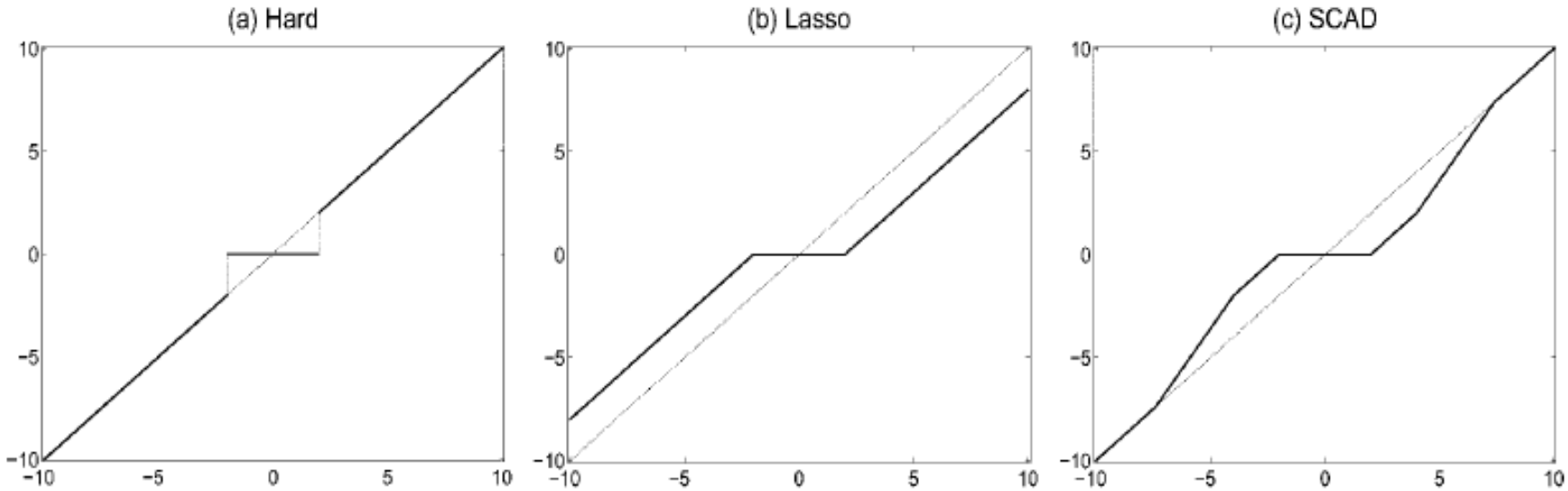


Figure: Plot of $\text{pen}'_\lambda(\beta)$ (left) and $\beta + \text{pen}'_\lambda(\beta)$ (right) for $\lambda = 1$.

SCAD satisfies all three conditions!

SCAD Solution under Orthonormal $\mathbf{X}$



$$\hat{\beta}_j^{\text{scad}} = \begin{cases} \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda), & \text{if } |\beta_j| < 2\lambda \\ \frac{(a-1)|\hat{\beta}_j| - \text{sign}(\hat{\beta}_j)a\lambda}{a-2}, & \text{if } 2\lambda \leq |\beta_j| < a\lambda \\ \hat{\beta}_j & \text{if } |\beta_j| \geq a\lambda \end{cases}$$

Oracle Property (Fan and Li, 2001)

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,$$

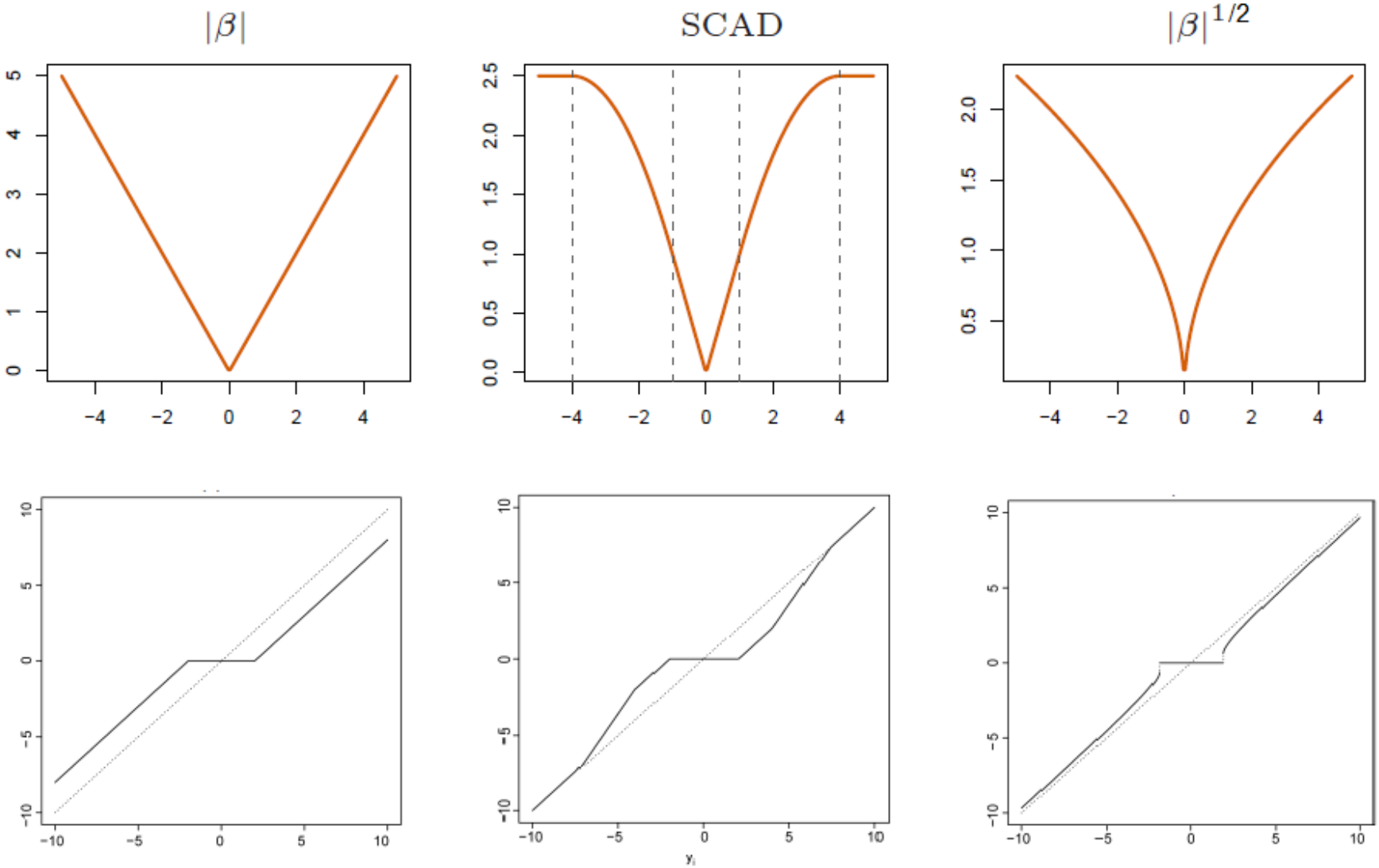
$\boldsymbol{\beta}_1$ contains all nonzero elements, and $\boldsymbol{\beta}_2 = \mathbf{0}$.

- Variable selection consistency: $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}_2^{\text{scad}} = 0) \to 1$.
- Asymptotic normality for true non-zero coefficients: $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{scad}} - \boldsymbol{\beta}_1) \xrightarrow{d} N(0, C^{-1}\sigma^2)$, where $C = \lim_{n\to\infty}(\mathbf{X}_1^T\mathbf{X}_1/n)$.

To deal the non-convex penalty,

- The paper proposed a local quadratic approximation.
- Zou and Li (2008) proposed a local linear approximation, so that LARS algorithm applies.
- Breheny and Huang (2011) extends coordinate descent algorithms to SCAD. (R package: 'ncvreg').

SCAD vs. l_q $(0 \leq q \leq 1)$



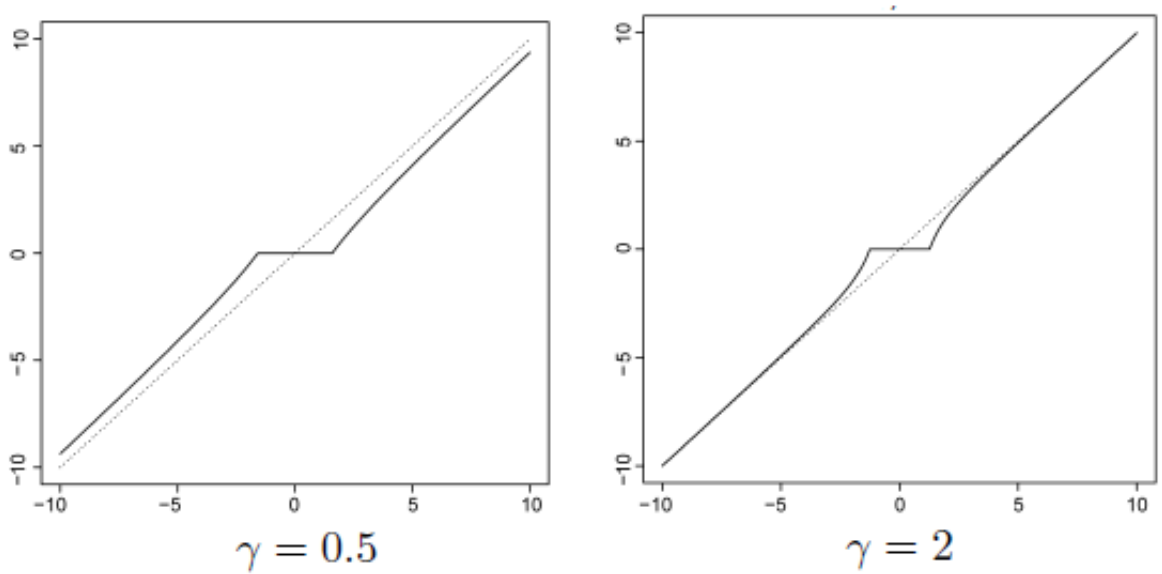
Can we do something similar with an l_q type penalty?

Adaptive Lasso (Zou, 2006)

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right],$$

where $\hat{w}_j = |\hat{\beta}_j|^{-\gamma}$ for some $\gamma > 0$. (Can replace $\hat{\beta}_j$ with any other $\sqrt{n}$-consistent estimate of true β.)

Adaptive lasso solution under orthonormal $\mathbf{X}$:



It has oracle property while retaining the attractive convexity property.

In the Presence of Strong Correlations among X

- Ridge shrinks the coefficients of correlated variables toward each other.
- Lasso is somewhat indifferent to the choice among a set of strong but correlated variables.

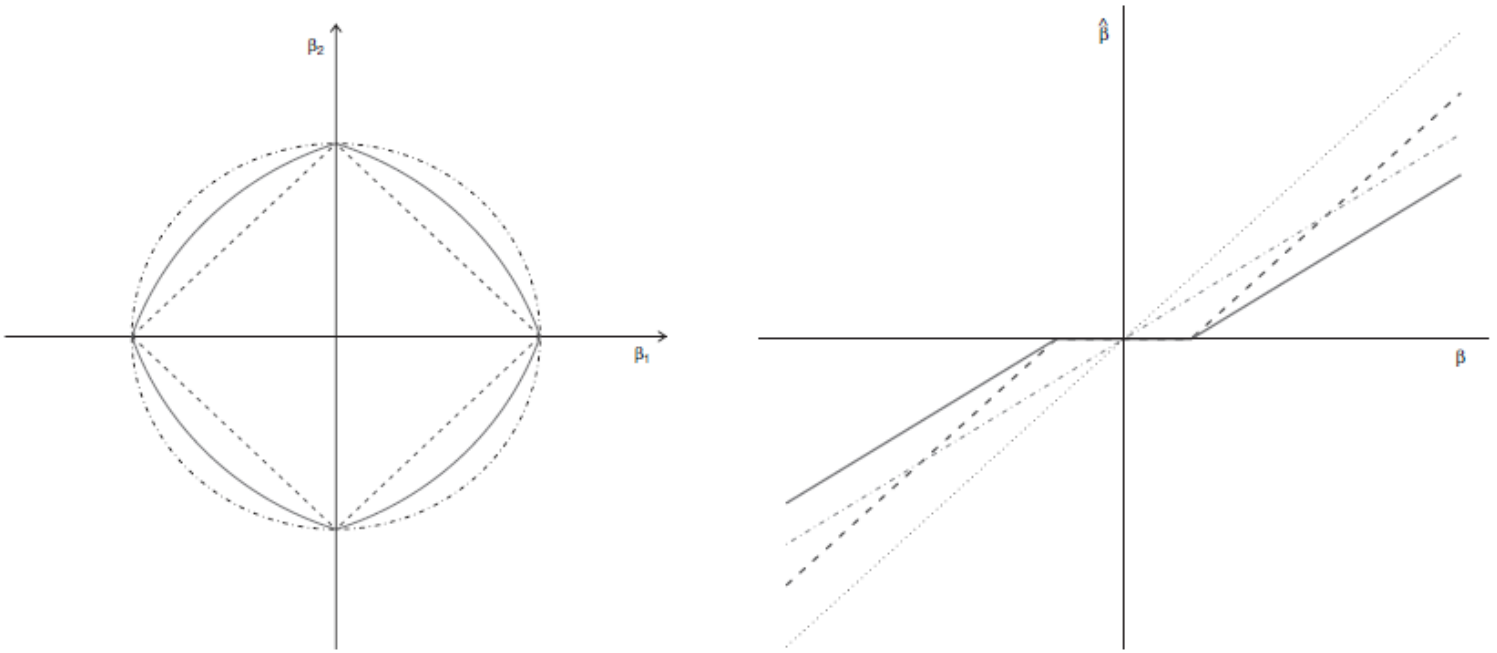
Example: $p = 2$, X_1 and X_2 are identical and important.

Elastic Net (Zou and Hastie, 2006)

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right) \right],$$

for $\alpha \in [0, 1]$.

- l_1 penalty encourages a sparse solution.
- l_2 penalty encourages highly correlated variables to be averaged. Enable the selection of more than n variables.



MCP (minimax concave penalty, Zhang 2010)

$$\text{pen}_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2a} & \text{if } |\beta_j| < a\lambda \\ \frac{a\lambda^2}{2} & \text{if } |\beta_j| \geq a\lambda \end{cases} \qquad (a > 1)$$

Out of all penalty functions continuously differentiable on $(0, \infty)$ that satisfy $pen'_\lambda(0^+) = \lambda$ (selection) and $pen'_\lambda(\beta) = 0$ for all $\beta > a\lambda$ (unbiasedness), the MCP minimizes the maximum concavity.

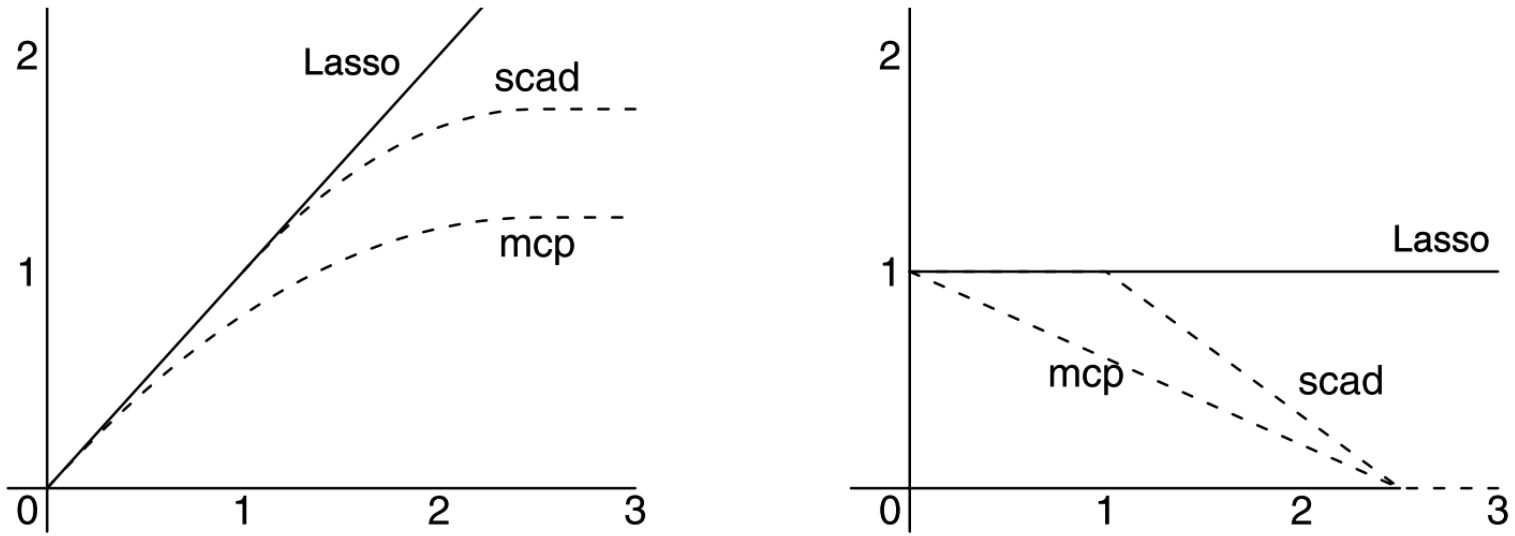


Figure: Plot of $\text{pen}_\lambda(\beta)$ (left) and $\text{pen}'_\lambda(\beta)$ (right) for $\lambda = 1$.

Summary

Computation:

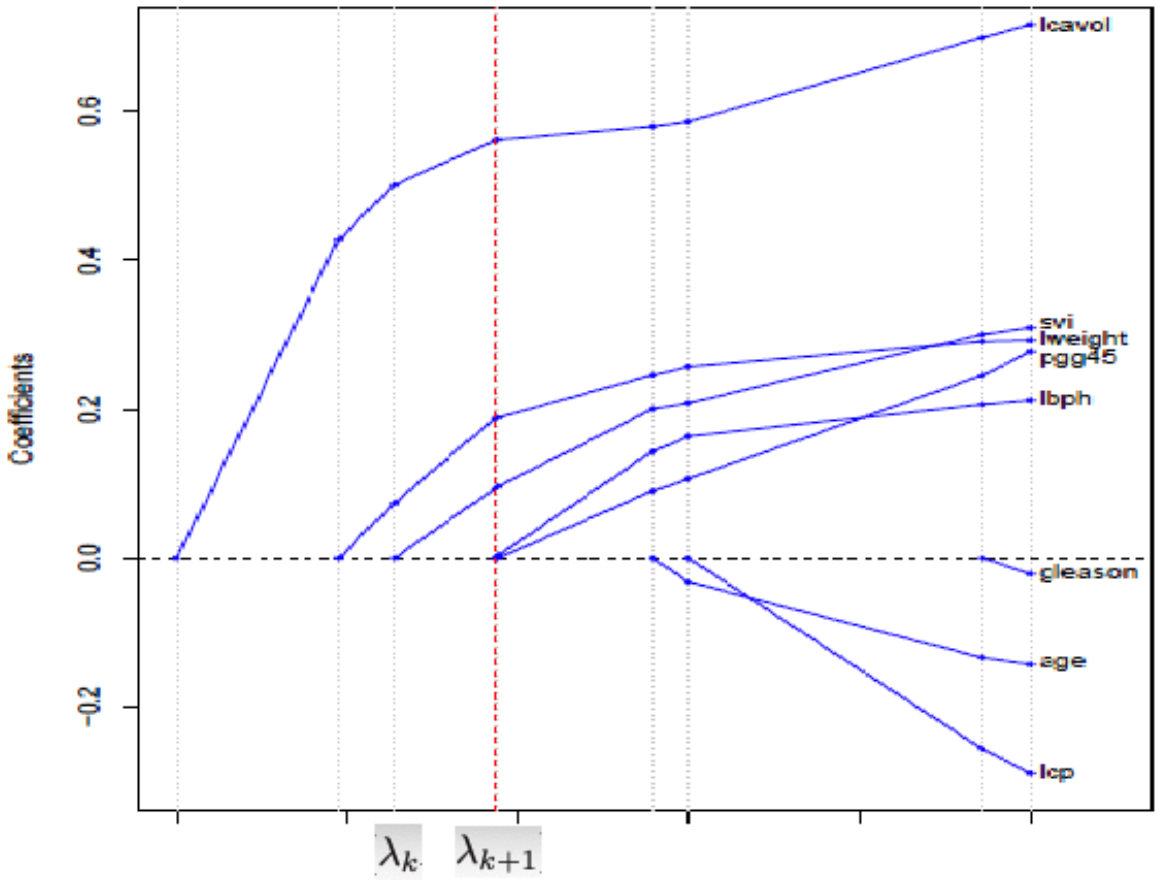
- Convex programming.
- Piecewise linear solution path (Rosset and Zhu, 2007)
- Coordinate descent.

Oracle property (Fan and Li, 2001)

- Select the right model with probability tending to 1.
- Coefficient estimate has the same asymptotic distribution as if the true model was known in advance.

	Lasso	AdaLasso	SCAD/MCP	Enet
Oracle Property	No	Yes	Yes	No
Computation (old)	Path	Path	Non-convex	Convex
Computation (new) (coordinate descent)	glmnet	glmnet	ncvreg	glmnet

Post-Selection Inference for Lasso: Covariance Test



$\hat{\boldsymbol{\beta}}^{\text{lasso}}_{\mathcal{A}_{k-1}}(\lambda_{k+1})$: Lasso estimate restricted to active predictors just before λ_k, with $\lambda = \lambda_{k+1}$.

Under H_0 that the current lasso model contains all truly active variables,

$$T_k = \frac{1}{\sigma^2}\left[\langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda_{k+1})\rangle - \langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\rangle\right] \xrightarrow{d} \text{Exp}(1)$$

Covariance Test (Lockhart et al. 2014)

$$F_k = \frac{1}{\hat{\sigma}^2}\left[\langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda_{k+1})\rangle - \langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\rangle\right]$$

- When $n > p$, $\hat{\sigma}^2 = \text{RSS of } \boldsymbol{\beta}^{\text{ols}}/(n-p)$.

$$F_k \xrightarrow{d} F_{2,n-p}.$$

- When $p \geq n$, the authors suggested $\hat{\sigma}^2 = \text{RSS of } \boldsymbol{\beta}^{\text{ols}}_{cv}/(n-p)$ and use distribution $F_{2,n-r}$, where $\boldsymbol{\beta}^{\text{ols}}_{cv}$ is the OLS estimate from model selected using cross-validation, and r is the number of parameters in the selected model.

Debiased Lasso

$$\widetilde{\boldsymbol{\beta}}^{\text{DL}} = \hat{\boldsymbol{\beta}}^{\text{lasso}} + \frac{1}{n}\mathbf{M}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}\right),$$

where $M \approx \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}$.

Under appropriate conditions with fixed design matrix $\mathbf{X}$,

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}^{\text{DL}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, M\hat{\Sigma}M^T\sigma_\epsilon^2)$$

- Zhang and Zhang (2014) "Confidence intervals for low dimensional parameters in high dimensional linear models." JRSSB.
- Van de Geer et al (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." AOS.
- Javanmard and Montanari (2014) "Confidence intervals and hypothesis testing for highdimensional regression." JMLR.

R Package: hdi

Review paper:

Dezeure, Bühlmann, Meier, and Meinshausen (2015)
"High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi." Statist. Sci. 30 (4) 533 - 558.

- Multi sample splitting
- Debiased lasso in Zhang and Zhang (2014)
- Ridge projection and bias correction
- Control for multiple testing

Reminders

- Homework #1 has been posted. It is due at 9pm EST on October 5th.
- Quiz is due at 9pm EST on September 23rd.