# Stat 577 "Statistical Learning Theory"

## T4: LASSO and SCAD estimator

Xiongzhi Chen

Washington State University

# Subgradient and subdifferential

# Convex set and convex function

- A set $U \subseteq \mathbb{R}^n$ is convex if

$$\lambda x + (1-\lambda)\, y \in U \text{ for } \forall x, y \in U \text{ and } \forall \lambda \in [0,1]$$

- A function $f : U \to \mathbb{R}$ is convex if

$$f\left(\lambda x + (1-\lambda)\, y\right) \leq \lambda f\left(x\right) + (1-\lambda) f\left(y\right)$$

  for all $x, y \in U$ and $\lambda \in [0,1]$
- A convex set does not have to be open or closed
- A convex function is differentiable almost everywhere
- Always assume $U \subseteq \mathbb{R}^n$ to be non-empty

# Convex function: equivalent definitions

Let $U \subseteq \mathbb{R}^n$ be open and convex. Equivalent definitions of $f : U \to \mathbb{R}$ being convex:

1. $f$ has convex *epigraph*

$$\text{epi}\,(f) = \{(x,z) \in \mathbb{R}^n \times \mathbb{R} : z \geq f(x), x \in U\}$$

2. $f$ is the *upper envelope* of *affine minorant*, i.e., $\forall x_0 \in U$,

$$f(x_0) = \sup\{a(x_0) : a \leq f, a : x \mapsto c_a + \langle \vartheta_a, x \rangle\}$$

   Each $a$ induces a "supporting hyperplane" to epi $(f)$ at $(x, f(x))$

3. For $n = 1$, the function

$$F(x,y) = \frac{f(x) - f(y)}{(x-y)}, x \neq y$$

   is non-decreasing in $x$ (for fixed $y$) and in $y$ (for fixed $x$)

# Univariate convex function

- Let $U = (a, b) \subseteq \mathbb{R}$. Equivalent definition of a convex $f : U \to \mathbb{R}$:
  - $f$ has left derivative $f'_-$ and right derivative $f'_+$ on $(a, b)$ (which implies $f$ is continuous on $(a, b)$), and
  - For all $x_1 < x_2$ and $x_1, x_2 \in (a, b)$,

$$f'_- (x_1) \leq f'_+ (x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f'_- (x_2) \leq f'_+ (x_2)$$

- Corollary:

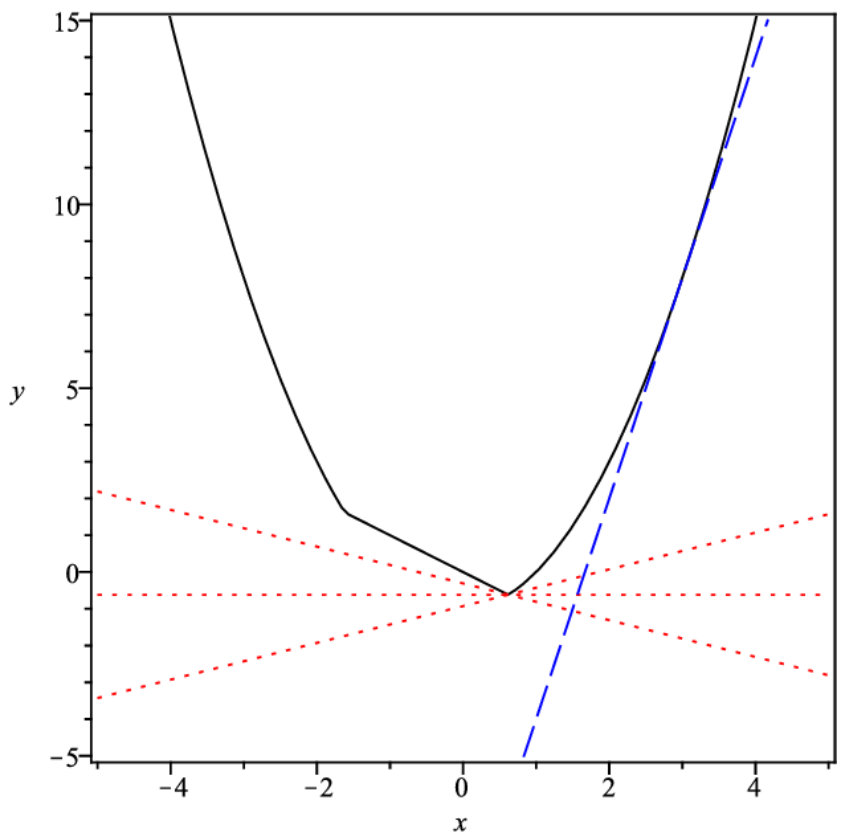
$$f'_- (x) = \sup_{y<x} \frac{f(y) - f(x)}{y - x} \quad \& \quad f'_+ (x) = \inf_{y>x} \frac{f(y) - f(x)}{y - x}$$

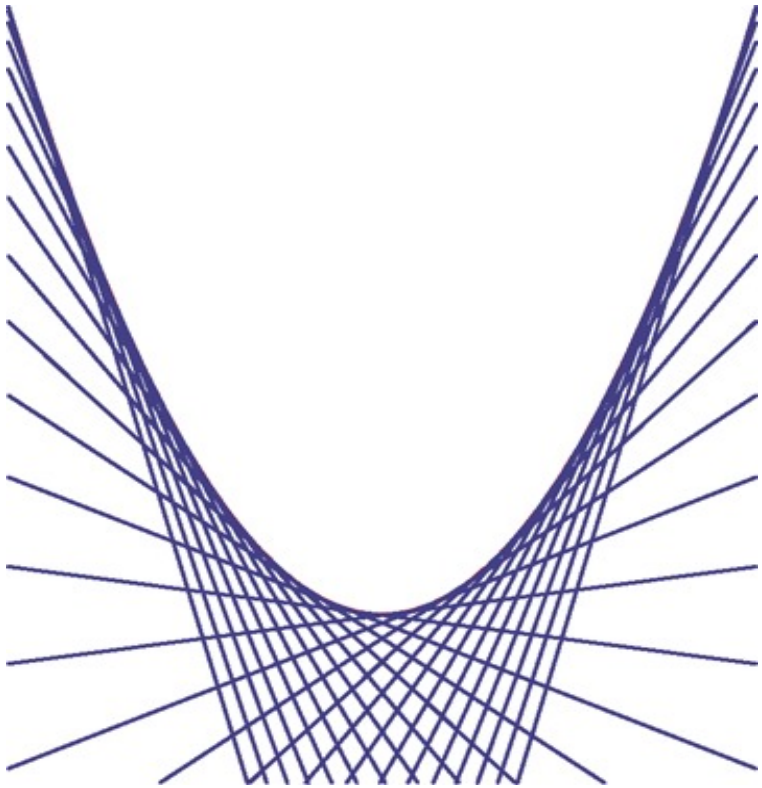
- Subdifferential: for $x_0 \in (a, b)$,

$$\begin{aligned} \partial f(x_0) &:= \{\alpha \in \mathbb{R} : f(x) - f(x_0) \geq \alpha (x - x_0), \forall x \in U\} \\ &= \left[f'_- (x_0), f'_+ (x_0)\right] \end{aligned}$$

# Convex function and envelope: illustration

- *Wiki (Envelope)*



# Subgradient and subdifferential

Let $U \subseteq \mathbb{R}^n$ be open and convex and $f : U \to \mathbb{R}$ be convex.

- Recall "supporting hyperplanes" induced by $a$ in

$$f(x) = \sup \{a(x) : a(x) = c_a + \langle \vartheta_a, x \rangle, a(x) \leq f(x)\} \tag{1}$$

- Subdifferential of $f$ at $x_0 \in U$ is defined as

$$\partial f(x_0) := \{v \in \mathbb{R}^n : f(x) - f(x_0) \geq \langle v, x - x_0 \rangle, \forall x \in U\},$$

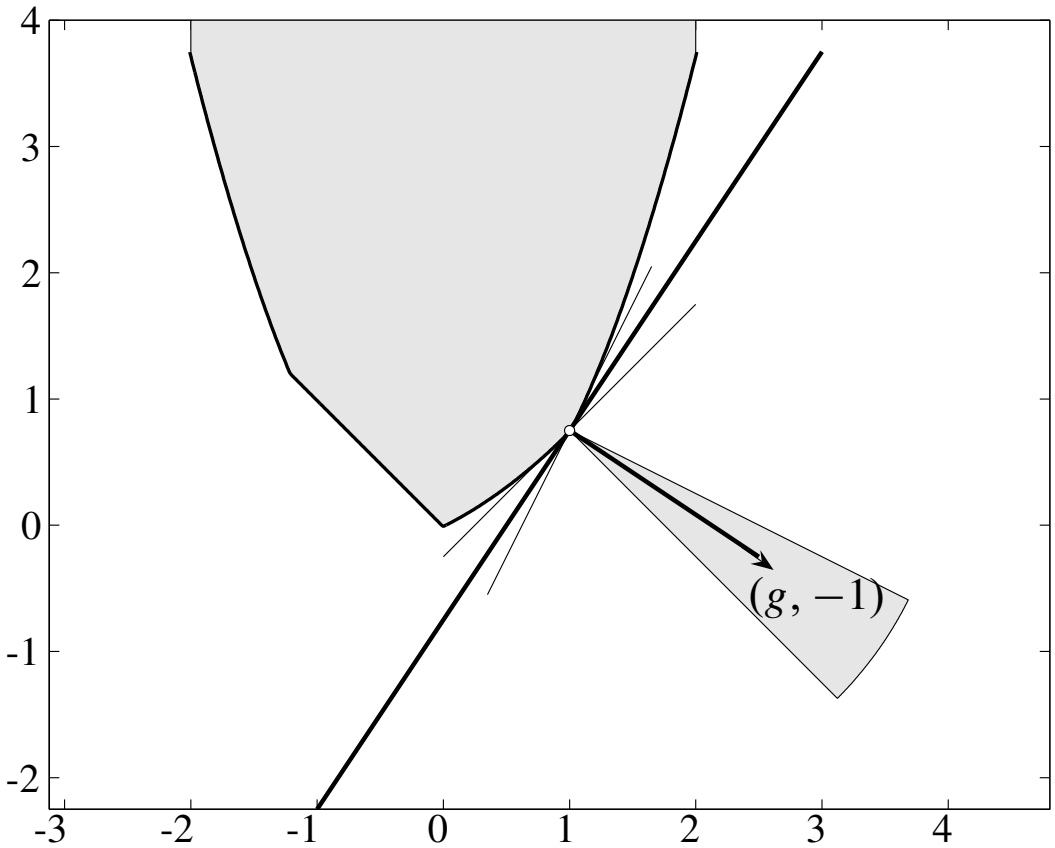
  and $\partial f(x_0) \neq \varnothing$ whenever $f$ is convex and continuous at $x_0$. An element of $\partial f(x_0)$ is a "subgradient" of $f$ at $x_0$
- Subdifferential $\partial f(x_0)$ can be derived from the famous *Theorem on Separation of Convex Sets* and is closely related to supporting hyperplanes

# Subgradient: illustration

- Subgradient $g$ of $f$ at $x_0$:

$$f(x) - f(x_0) \geq \langle g, x - x_0 \rangle, \forall x \in U$$

- Image credit: "Nonlinear optimization" by Andrzej Ruszczyński



# Karush-Kuhn-Tuker (KKT) conditions

- Let $U \subseteq \mathbb{R}^n$ be open and convex. Given $f : U \to \mathbb{R}$, consider

$$\min_{x \in U} f(x) \quad \text{subject to} \quad \begin{cases} g_i(x) \leq 0, 1 \leq \forall i \leq p' \\ h_j(x) = 0, 1 \leq \forall j \leq q' \end{cases} \quad \text{for } x \in U$$

  - If $f$ and $\{g_i\}_{i=1}^{p'}$ are convex on $U$ and $\{h_j\}_{j=1}^{q'}$ are linear on $U$, and
  - If *Slater condition* holds, i.e., there exits $x_0 \in U$ such that $g_i(x_0) < 0$ for $1 \leq i \leq p'$ and $h_j(x_0) = 0$ for $1 \leq j \leq q'$
- Then $\hat{x}$ is a minimizer iff $h_j(\hat{x}) = 0$ for $1 \leq j \leq q'$ and there exist $\hat{\lambda}_i \geq 0$ for $1 \leq i \leq p'$ and $\hat{\mu}_j \in \mathbb{R}$ for $1 \leq j \leq q'$ such that

$$\begin{cases} 0 \in \partial f(\hat{x}) + \sum_{i=1}^{p'} \hat{\lambda}_i \partial g_i(\hat{x}) + \sum_{j=1}^{q'} \hat{\mu}_j \partial h_j(\hat{x}) \\ \hat{\lambda}_i g_i(\hat{x}) = 0 \text{ for } 1 \leq \forall i \leq p' \end{cases}$$

# Solution for LASSO

# LASSO and Ridge estimates

- Contour of RSS can only meet vertex on $l_1$-sphere (due to subdifferential) since $l_1$-norm is singular at 0, leading to variable selection via hard thresholding
- Contour of RSS can meet any point on $l_2$-sphere since $l_2$-norm is nonsingular at 0, leading to shrinkage only

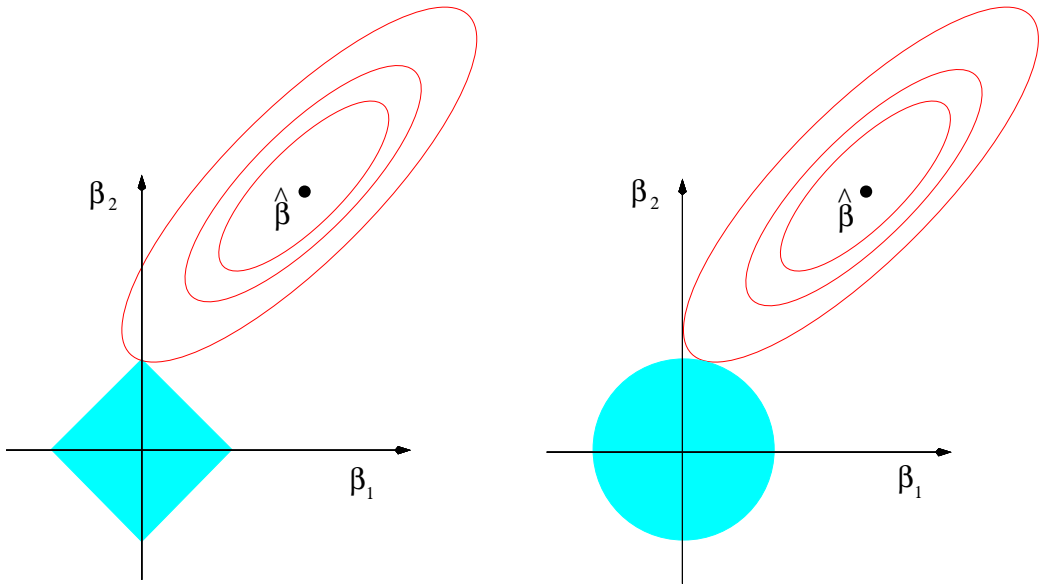
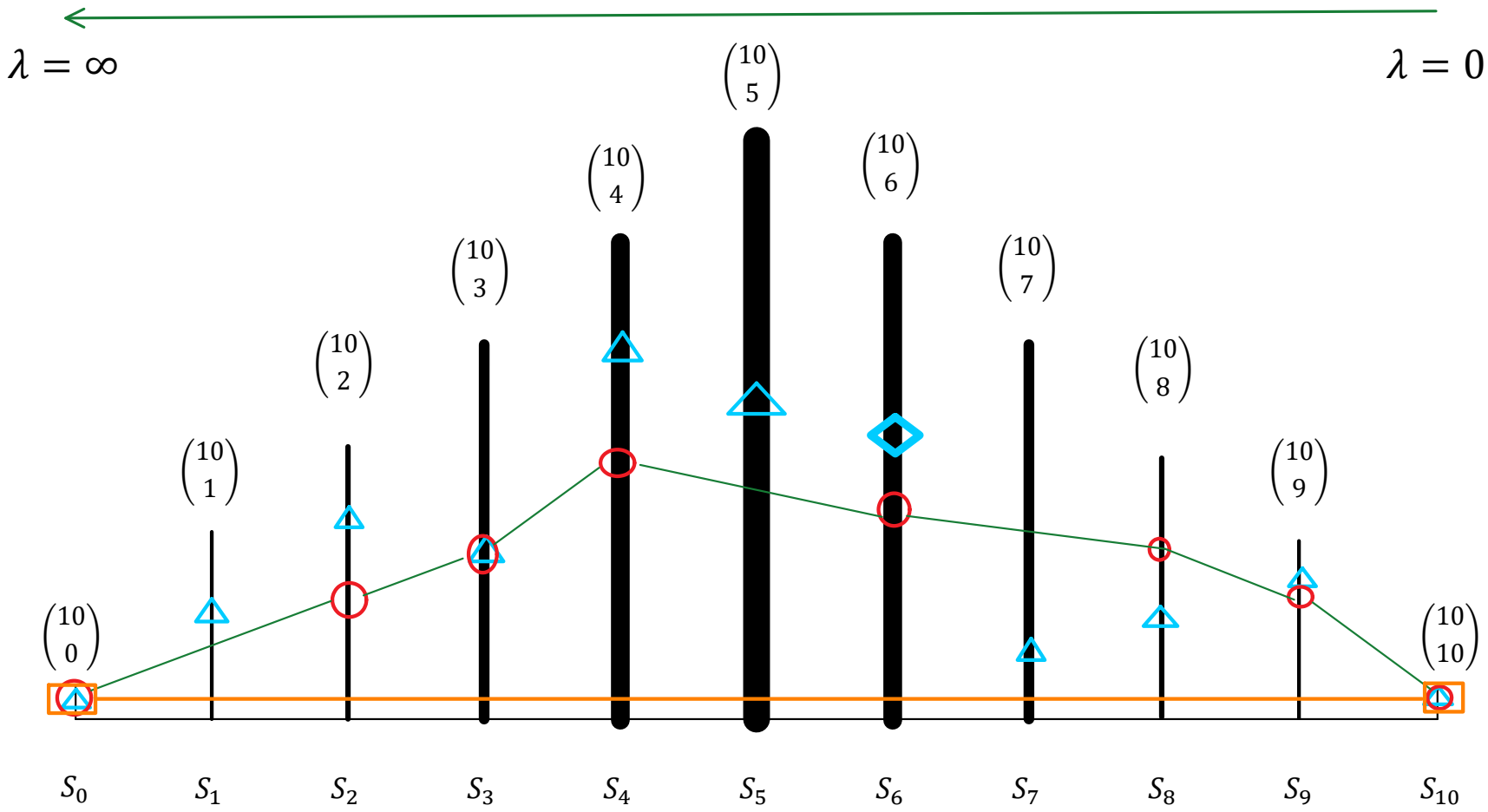



**FIGURE 3.11.** *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses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

# Trajectory of selected models

- "Circle", "Rectangle" and "Triangle": models selected by LASSO, Ridge and BSS; "Diamond": best model by BSS



# LASSO and optimization

- Model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ and LASSO

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 \text{ subject to } \sum_{i=1}^p |\beta_i| \leq t; \\ h_1\left(\boldsymbol{\beta}; \lambda\right) = \frac{1}{2} \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda \sum_{i=1}^p |\beta_i| \end{cases}$$

- Subgradient of $g(x) = |x|$ at $x = 0$ is $\partial g(0) = [-1, 1]$
- By KKT conditions, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)$ is the optimizer iff $0 \in \partial h_1\left(\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda); \lambda\right)$, where

$$\partial h_1\left(\boldsymbol{\beta}; \lambda\right) = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y} + \lambda\mathbf{c},$$

and $\mathbf{c} = \left(c_1, ..., c_p\right)^T$ and

$$\begin{cases} c_i = \operatorname{sgn}\left(\beta_i\right) & \text{if} \quad \beta_i \neq 0 \\ c_i \in [-1, 1] & \text{if} \quad \beta_i = 0 \end{cases}$$

# Orthogonal design and solution

- If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then LSE $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ and $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\hat{\beta}_{\mathrm{L},1}(\lambda), ..., \hat{\beta}_{\mathrm{L},p}(\lambda)\right)^T$ satisfies

$$\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) - \hat{\boldsymbol{\beta}} + \lambda \mathbf{c} = 0 \text{ with } \begin{cases} c_i = \operatorname{sgn}\left(\hat{\beta}_{\mathrm{L},i}(\lambda)\right) & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0 \\ c_i \in [-1, 1] & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \end{cases}$$

- Thresholding: $\hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \Longleftrightarrow \hat{\beta}_i - \lambda c_i = 0, |c_i| \leq 1 \Longleftrightarrow \left|\hat{\beta}_i\right| \leq \lambda$
- Soft thresholding and equivalences: $\hat{\beta}_{L,i}(\lambda) \neq 0 \Longleftrightarrow$
  - $\hat{\beta}_{\mathrm{L},i}(\lambda) - \hat{\beta}_i + \lambda \operatorname{sgn}\left(\hat{\beta}_{\mathrm{L},i}(\lambda)\right) = 0$
  - $\hat{\beta}_{\mathrm{L},i}(\lambda) = \hat{\beta}_i - \lambda$ if $\hat{\beta}_{\mathrm{L},i}(\lambda) > 0$ but $\hat{\beta}_{\mathrm{L},i}(\lambda) = \hat{\beta}_i + \lambda$ if $\hat{\beta}_{\mathrm{L},i}(\lambda) < 0$
  - $\hat{\beta}_i > \lambda$ and $\hat{\beta}_{\mathrm{L},i}(\lambda) = \hat{\beta}_i - \lambda$ or $\hat{\beta}_i < -\lambda$ and $\hat{\beta}_{\mathrm{L},i}(\lambda) = \hat{\beta}_i + \lambda$
  - $\hat{\beta}_{\mathrm{L},i}(\lambda) = \operatorname{sgn}\left(\hat{\beta}_i\right)\left(\left|\hat{\beta}_i\right| - \lambda\right)$
- Summary: $\hat{\beta}_{\mathrm{L},i}(\lambda) = \operatorname{sgn}\left(\hat{\beta}_i\right)\left(\left|\hat{\beta}_i\right| - \lambda\right)_+$

# Visualization of estimate for orthogonal design

- If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ and $\hat{\beta}_{\mathrm{L},i}(\lambda) = \operatorname{sgn}\left(\hat{\beta}_i\right)\left(\left|\hat{\beta}_i\right| - \lambda\right)_+$
  - $\hat{\beta}_{\mathrm{L},i}(\lambda)$ is always biased when $\left|\hat{\beta}_i\right| > \lambda > 0$, i.e., *intrinsic bias*
  - For Gaussian error, explicit distribution $F_{i,\lambda}$ for $\hat{\beta}_{L,i}(\lambda)$ for each *deterministic* $\lambda$ can be obtained. However, once optimal $\lambda$ is chosen from data, such is often impossible to obtain

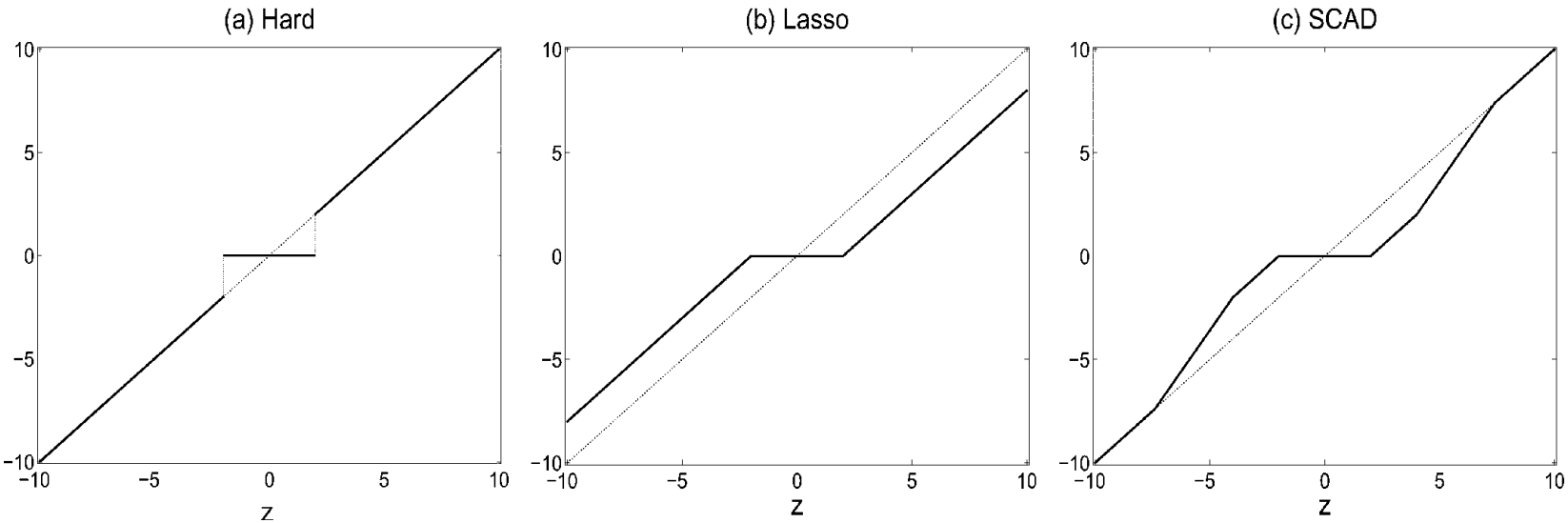



*Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.*

# Consistency of LASSO

# Intrinsic bias of LASSO estimate

If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then LSE $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ and $\hat{\beta}_{L,i}(\lambda) = \operatorname{sgn}\left(\hat{\beta}_i\right)\left(\left|\hat{\beta}_i\right| - \lambda\right)_+$ and

- $\hat{\beta}_{\mathrm{L},i}(\lambda)$ is biased when $\left|\hat{\beta}_i\right| > \lambda > 0$
- $\hat{\beta}_{\mathrm{L},i}(\lambda)$ is biased when $0 < \left|\hat{\beta}_i\right| < \lambda$ for $\lambda > 0$
- Bias caused by $\frac{d}{dx}\varrho_\lambda(|x|) > 0$ for large $|x|$ and non-differentiability of $\varrho_\lambda(|x|)$ at $x = 0$

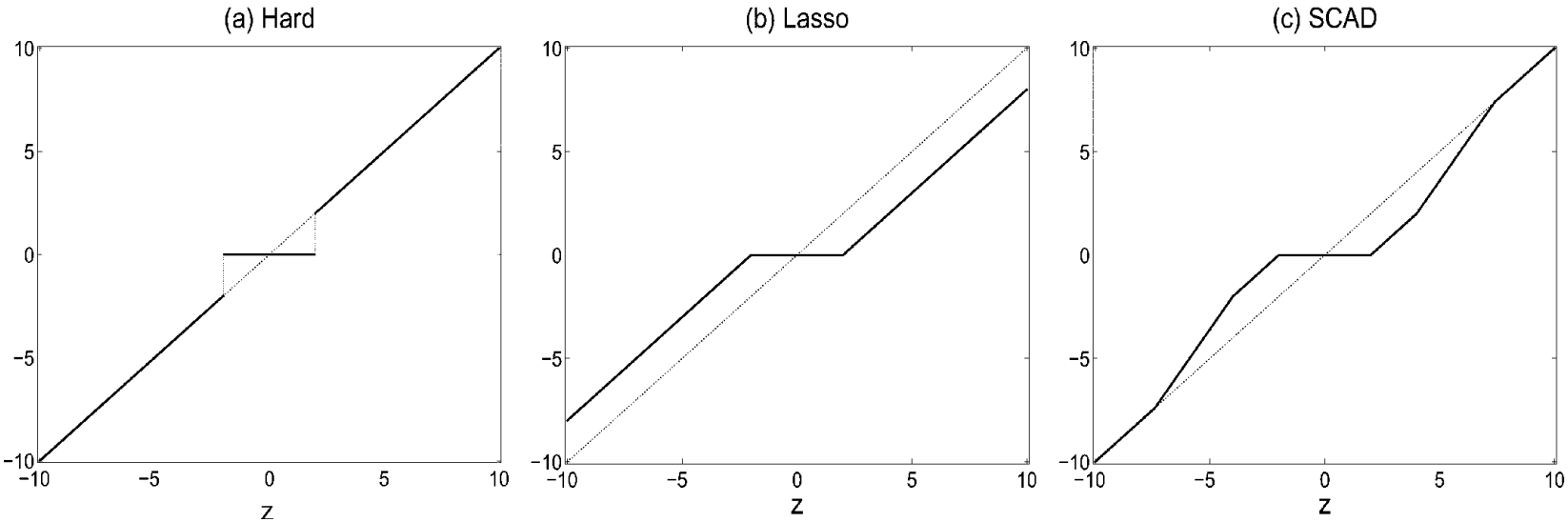



*Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.*

# Bias of LASSO estimate

If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ and $\hat{\beta}_{\mathrm{L},i}(\lambda) = \operatorname{sgn}(\hat{\beta}_i)\left(|\hat{\beta}_i| - \lambda\right)_+$

- Assume $|\hat{\beta}_i| > \lambda > 0$. How can the bias of $\hat{\beta}_{\mathrm{L},i}(\lambda)$ be 0?
- Assume $0 \leq |\hat{\beta}_i| < \lambda$ for $\lambda > 0$. How can the bias of $\hat{\beta}_{\mathrm{L},i}(\lambda)$ be 0?
- Assume $\lambda = 0$. Which $\hat{\beta}_{\mathrm{L},i}(\lambda)$ among $1 \leq i \leq p$ has 0 bias?
- Under what additional conditions will $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)$ be consistent?

# LASSO estimate for general design

- Let $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\hat{\beta}_{\mathrm{L},1}(\lambda), ..., \hat{\beta}_{\mathrm{L},p}(\lambda)\right)^T$ and

$$\begin{cases} c_i = \operatorname{sgn}\left(\hat{\beta}_{L,i}(\lambda)\right) & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0 \\ c_i \in [-1,1] & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \end{cases}$$

- Let $\widetilde{x}_j$ be the $j$th column of $\mathbf{X}$. The following are equivalent:
  - $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) - \mathbf{X}^T\mathbf{y} + \lambda\mathbf{c} = 0 \Longleftrightarrow$
  - $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \hat{\boldsymbol{\beta}} - \lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{c} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} - \lambda\mathbf{c}\right) \Longleftrightarrow$
  - $$\begin{cases} \widetilde{x}_i^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)\right) = \lambda \operatorname{sgn}\left(\hat{\beta}_{L,i}(\lambda)\right) & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0 \\ \left|\widetilde{x}_i^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)\right)\right| \leq \lambda & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \end{cases}$$
- LASSO estimate in general *shrinkage LSE* but is *very complicated*

# Framework for theoretical investigation

- Sparsity $\|\boldsymbol{\beta}\|_0 = s$: $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_s)^T$; $\boldsymbol{\beta}_0 = 0^T$; $\mathcal{A} = \{1, ..., s\}$
- Split: $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_0)$ with $\mathbf{X}_1 = (\widetilde{\boldsymbol{x}}_1, ..., \widetilde{\boldsymbol{x}}_s)$ and $\mathbf{X}_0 = \left(\widetilde{\boldsymbol{x}}_{s+1}, ..., \widetilde{\boldsymbol{x}}_p\right)$, i.e.,

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_0 \end{pmatrix} \mapsto \mathbf{X} = (\mathbf{X}_1, \mathbf{X}_0)$$

- Recall $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} - \lambda\mathbf{c}\right)$ or

$$\begin{cases} \widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)\right) = \lambda \operatorname{sgn}\left(\hat{\beta}_{L,i}(\lambda)\right) & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0 \\ \left|\widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)\right)\right| \leq \lambda & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \end{cases} \tag{2}$$

- Estimated sparsity $||\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)||_0 = \hat{s}$ and $\hat{\mathcal{A}} = \left\{i : \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0\right\}$
- Strategy on theory: investigate (2)

# Intuition on consistency of LASSO

- Recall $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\hat{\beta}_{\mathrm{L},1}(\lambda), ..., \hat{\beta}_{\mathrm{L},p}(\lambda)\right)^T$ satisfies

$$\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} - \lambda\mathbf{c}\right), \tag{3}$$

$$\text{where} \begin{cases} c_i = \operatorname{sgn}\left(\hat{\beta}_{L,i}(\lambda)\right) & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0 \\ c_i \in [-1,1] & \text{if} \quad \hat{\beta}_{\mathrm{L},i}(\lambda) = 0 \end{cases}$$

- Recall $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_s)^T$, $\boldsymbol{\beta}_0 = 0^T$, $\mathcal{A} = \{1, ..., s\}$ and splitting

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_0 \end{pmatrix} \mapsto \mathbf{X} = (\mathbf{X}_1, \mathbf{X}_0) \mapsto \mathbf{C} = \mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_0 \\ \mathbf{X}_0^T\mathbf{X}_1 & \mathbf{X}_0^T\mathbf{X}_0 \end{pmatrix}$$

- Note $\mathbf{c}$ is a function of $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)$. So, $\mathbf{X}_1$ and $\mathbf{X}_0$ have to satisfy some conditions for $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda)$ to achieve selection consistency and/or estimation consistency

# LASSO consistency: iff condition

- Recall $\hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} - \lambda\mathbf{c}\right)$, where $c_i = \operatorname{sgn}\left(\hat{\beta}_{\mathrm{L},i}(\lambda)\right)$ if $\hat{\beta}_{\mathrm{L},i}(\lambda) \neq 0$ but $|c_i| \leq 1$ if $\hat{\beta}_{\mathrm{L},i}(\lambda) = 0$
- Recall $\mathbf{C} = \mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_0 \\ \mathbf{X}_0^T\mathbf{X}_1 & \mathbf{X}_0^T\mathbf{X}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$
- *Irrepresentability* is about $\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}$, where $\mathbf{s} = \operatorname{sgn}(\boldsymbol{\beta}_1)$ and about $\|\boldsymbol{\beta}_1\|_1$ versus $\left\|\lambda\mathbf{C}_{11}^{-1}\mathbf{s}\right\|_1$
- "Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions" by Soumendra N. Lahiri, Annals of Statistics, 2021
  - Namely, lower irrepresentable condition + upper irrepresentable condition == variable selection consistency
  - *Surprise*: LASSO cannot simultaneously achieve variable selection consistency and $\sqrt{n}$-consistency

# LASSO solution parth

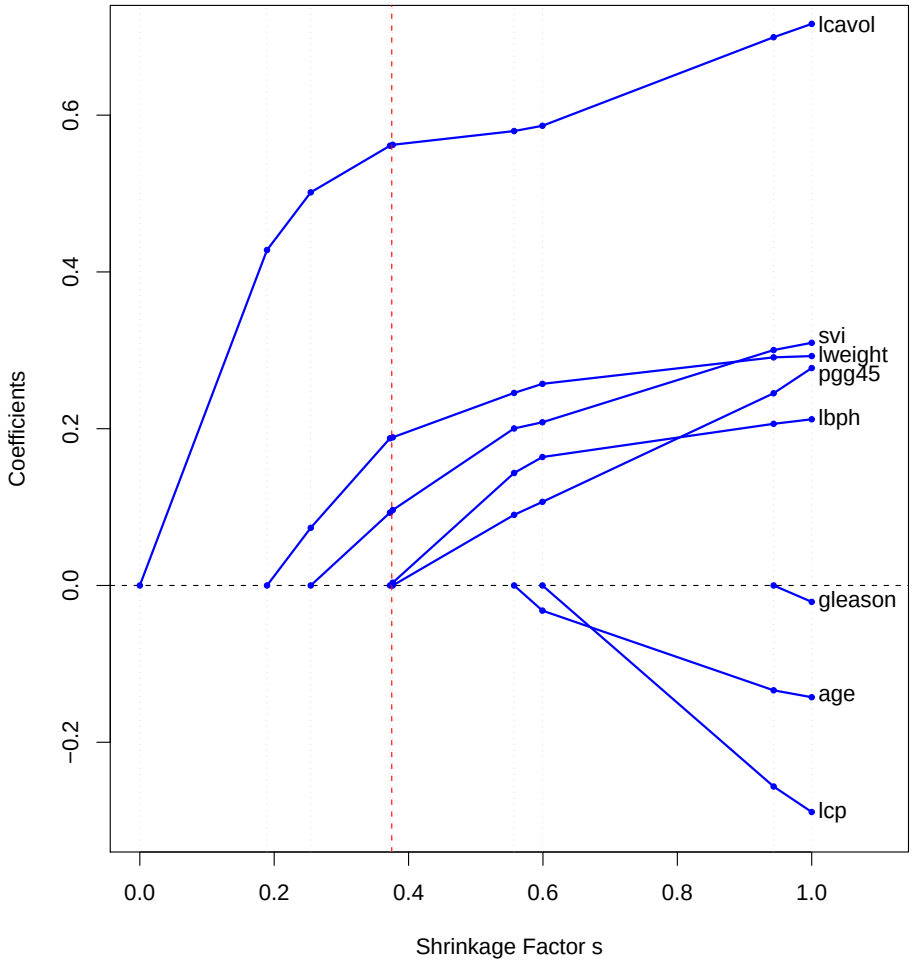



**FIGURE 3.10.** *Profiles of lasso coefficients, as the tuning parameter $t$ is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

- LSE $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, ..., \hat{\beta}_p\right)^T$
- $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ LASSO estimate
- $t = \left\| \hat{\boldsymbol{\beta}}_{\mathrm{L}}(\lambda) \right\|_1$
- $\hat{s} = 0.36$ chosen by 10-fold cross-validation
- Caution: magnitudes of $\hat{\beta}_{\mathrm{L},i}$'s not necessarily monotone decreasing in $s$, even though they are for this example

# SCAD solution

# SCAD penalty and estimate for orthogonal design

- SCAD penalty is a *quadratic spline* and is *not convex*
- Derivative of SCAD penalty for $a > 2$ is

$$\varrho'_\lambda(\theta) = \lambda\left[1_{\{\theta\le\lambda\}}(\theta) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda}1_{\{\theta>\lambda\}}(\theta)\right], \theta > 0$$

- If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then LSE $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ and SCAD estimate

$$\hat{\beta}_{S,i}(\lambda) = \begin{cases} \operatorname{sgn}(\hat{\beta}_i)(|\hat{\beta}_i| - \lambda)_+ & \text{if} \quad |\hat{\beta}_i| \le 2\lambda \\ \dfrac{(a-1)\hat{\beta}_i - \operatorname{sgn}(\hat{\beta}_i)a\lambda}{a-2} & \text{if} \quad 2\lambda < |\hat{\beta}_i| \le a\lambda \\ \hat{\beta}_i & \text{if} \quad |\hat{\beta}_i| > a\lambda \end{cases}$$

and LASSO estimate

$$\hat{\beta}_{L,i}(\lambda) = \operatorname{sgn}(\hat{\beta}_i)(|\hat{\beta}_i| - \lambda)_+$$

- Note the differences between "soft shrinkage" parts

# Visualizing estimates for orthogonal design

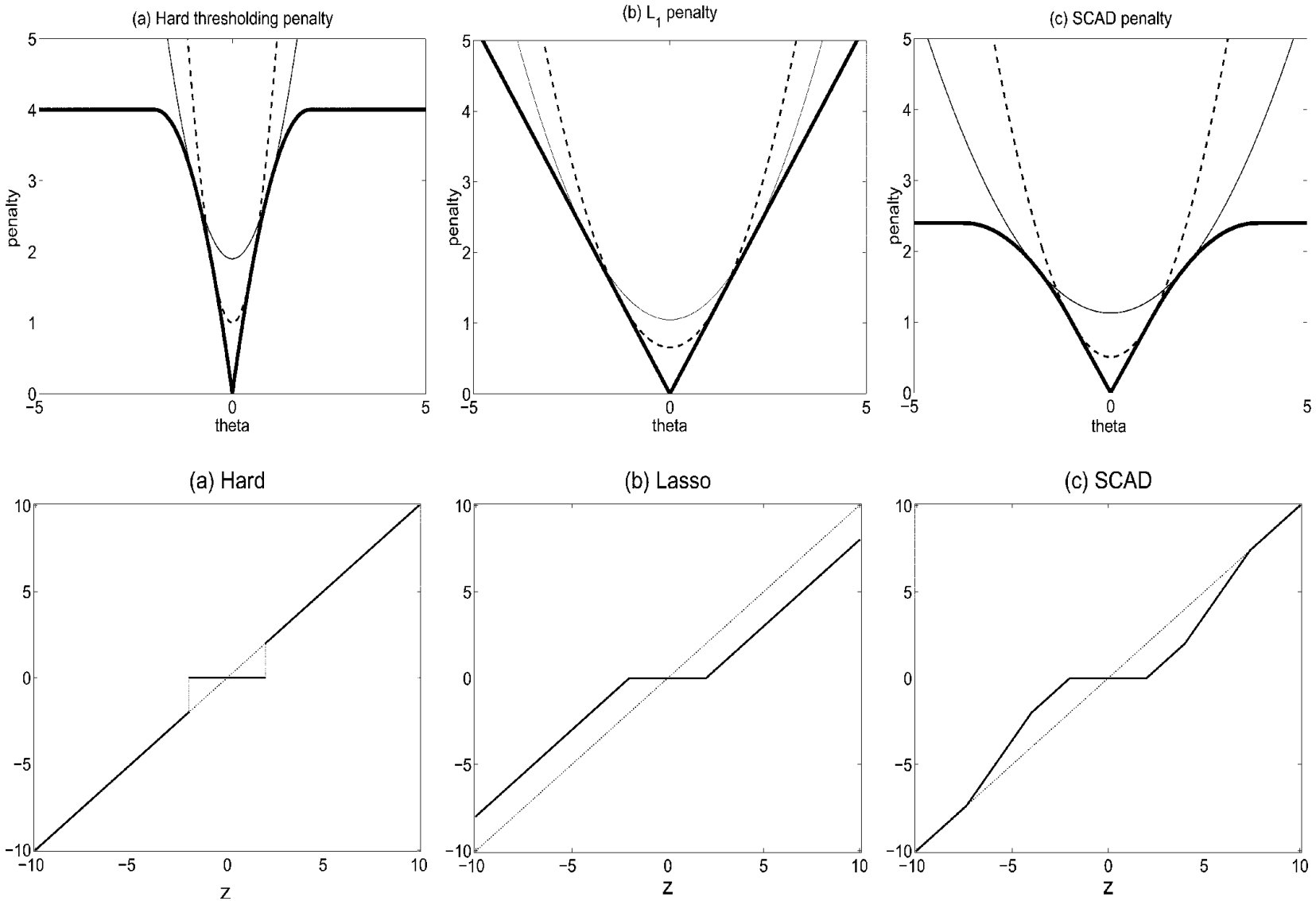



*Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.*

# SCAD estimator in general

- Since SCAD penalty $\varrho_\lambda(|x|)$ is *not convex, KKT conditions are usually only necessary*
- Penalized RSS as objective function

$$h_S(\boldsymbol{\beta};\lambda) = \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{i=1}^{p}\varrho_\lambda(|\beta_i|)$$

and optimizer $\hat{\boldsymbol{\beta}}_S(\lambda) = \left(\hat{\beta}_{S,1}(\lambda), ..., \hat{\beta}_{S,p}(\lambda)\right)^T$

- Let $\widetilde{\boldsymbol{x}}_j$ be the $j$th column of $\mathbf{X}$. Necessary condition on $\hat{\boldsymbol{\beta}}_S(\lambda)$:

$$\begin{cases} \widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_S(\lambda)\right) = \operatorname{sgn}\left(\hat{\beta}_{S,i}(\lambda)\right)\varrho'_\lambda\left(\left|\hat{\beta}_{S,i}(\lambda)\right|\right) \text{ if } \hat{\beta}_{S,i}(\lambda)\neq 0 \\ \left|\widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_S(\lambda)\right)\right| \leq \lambda \text{ if } \hat{\beta}_{S,i}(\lambda)=0 \end{cases},$$

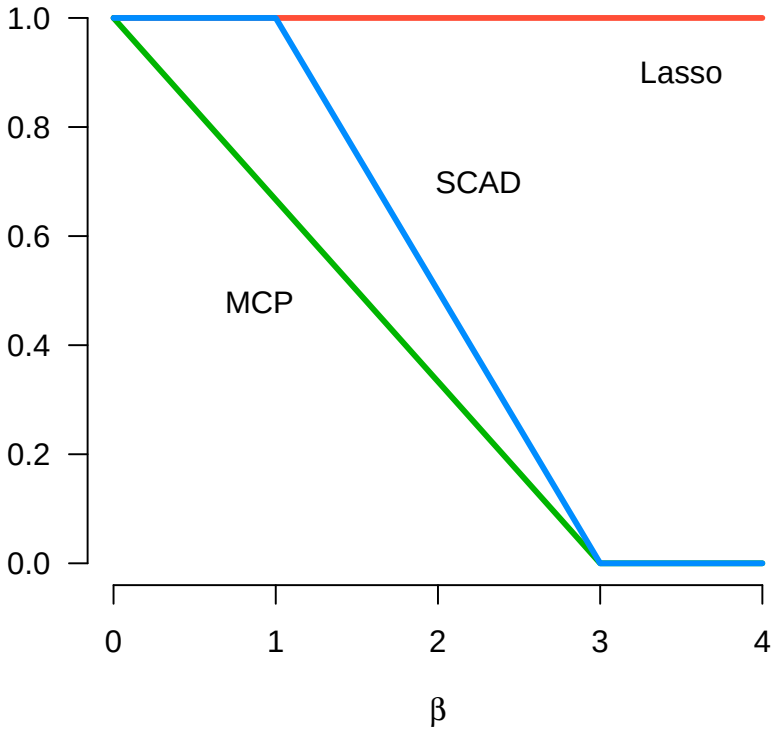
since $\varrho_\lambda(|x|) = \lambda|x|$ for $x$ close to 0

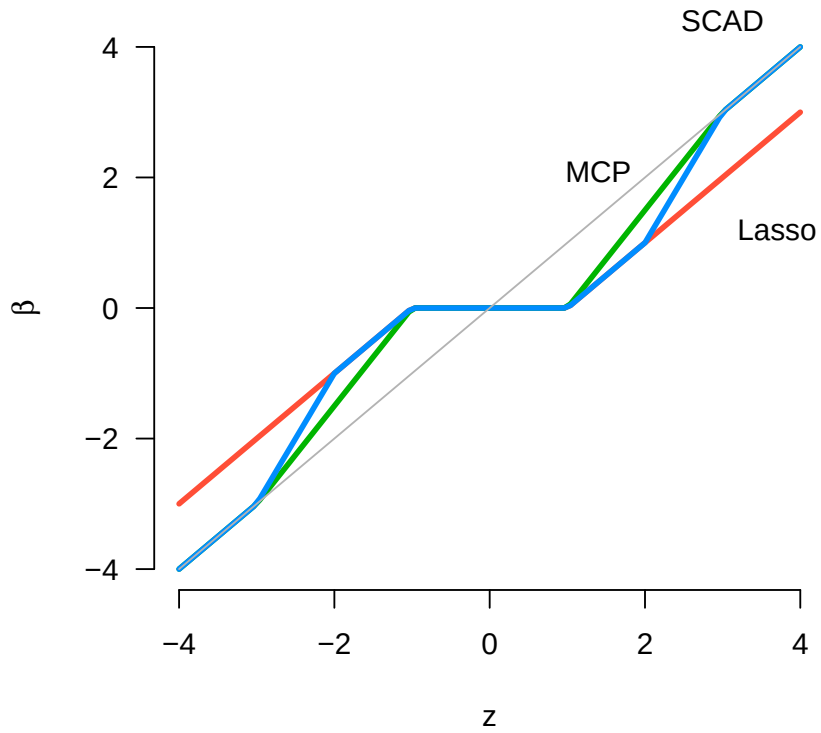
# Non-uniqueness of SCAD optimizer

- Since SCAD penalty $\varrho_\lambda(|x|)$ is not convex, optimizer $\hat{\boldsymbol{\beta}}_S(\lambda)$ is not unique in general
- Despite nearly unbiasedness of $\hat{\boldsymbol{\beta}}_S(\lambda)$, SCAD estimate is not as population as LASSO estimate since the former is harder to compute than the latter
- Through years of research, Prof. Jianqing Fan and his collaborators found that, in some situations, any local optimizer $\hat{\boldsymbol{\beta}}_S(\lambda)$ is good enough, thus avoiding global search of optimizers of $h_S(\boldsymbol{\beta}; \lambda)$
- SCAD estimate, if it is good, often outperforms LASSO estimate in terms of accuracy
- *We know little about performances of high-dimensional models under non-convex penalties*

# Minimax concave penalty (MCP)

- MCP further improves SCAD; see Zhang (2010)
- MCP is not convex, and it reduces bias in estimating moderate sized effects, compared to SCAD and LASSO



# Consistency of SCAD estimator

# Framework on theoretical investigation

- Recall SCAD estimate $\hat{\boldsymbol{\beta}}_{\mathrm{S}}(\lambda)$ as

$$\begin{cases} \widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{S}}(\lambda)\right)=\operatorname{sgn}\left(\hat{\beta}_{\mathrm{S},i}(\lambda)\right)\varrho_\lambda'\left(\left|\hat{\beta}_{\mathrm{S},i}(\lambda)\right|\right) \text{ if } \hat{\beta}_{\mathrm{S},i}(\lambda)\neq 0 \\ \left|\widetilde{\boldsymbol{x}}_i^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{S}}(\lambda)\right)\right|\leq\lambda \text{ if } \hat{\beta}_{\mathrm{S},i}(\lambda)=0 \end{cases}$$

- Selection consistency is similar to that of LASSO
- Estimation consistency should be better than that of LASSO
- The above representation of $\hat{\boldsymbol{\beta}}_{\mathrm{S}}(\lambda)$ can be the starting point of theoretical investigation on $\hat{\boldsymbol{\beta}}_{\mathrm{S}}(\lambda)$
- An alternative framework has been provided by, e.g., Fan and Li (2001) and Fan and Peng (2004) and Fan and Lv (2011)

# Penalized log-likelihood

- Recall penalized RSS as objective function

$$h_S(\boldsymbol{\beta};\lambda) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{i=1}^{p} \varrho_\lambda(|\beta_i|)$$

- To accommodate model

$$\begin{cases} \Pr(Y \in A|X) = \int_A f\left(g^{-1}\left(X^T\boldsymbol{\beta}\right), y\right) dy \\ g(E(Y|X)) = X^T\boldsymbol{\beta} \end{cases}$$

and asymptotic study, consider *penalized log-likelihood*

$$h(\boldsymbol{\beta};\lambda) = -n^{-1}\sum_{i=1}^{n} \ln f\left(g^{-1}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right), y_i\right) + \sum_{i=1}^{p} \varrho_\lambda(|\beta_i|)$$

for i.i.d. observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $(X, Y)$

# Oracle property of optimizer

- Sparsity $\|\boldsymbol{\beta}\|_0 = s$: $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_s)^T$; $\boldsymbol{\beta}_0 = 0^T$; $\mathcal{A} = \{1, ..., s\}$
- Optimizer $\hat{\boldsymbol{\beta}}(\lambda) = \left(\hat{\beta}_1(\lambda), ..., \hat{\beta}_p(\lambda)\right)^T$ such that

$$\hat{\boldsymbol{\beta}}(\lambda) \in \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} h(\boldsymbol{\beta}; \lambda)$$

- Oracle property of $\hat{\boldsymbol{\beta}}(\lambda) = \left(\hat{\boldsymbol{\beta}}_1(\lambda)^T, \hat{\boldsymbol{\beta}}_2(\lambda)^T\right)^T$ as

$$\begin{cases} \text{Matching sparsity: } \hat{\boldsymbol{\beta}}_2(\lambda) = \boldsymbol{\beta}_0 = 0 \\ \text{Asymptotic Normality: } \sqrt{n} A_n \left(\hat{\boldsymbol{\beta}}_1(\lambda) - \boldsymbol{\beta}_1 + \mathbf{b}_n\right) \\ \qquad \rightsquigarrow \text{Normal}(0, \mathbf{I}) \end{cases}$$

- Oracle property is stronger than

$$\sqrt{n} A_n \left(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta} + \mathbf{b}_n\right) \rightsquigarrow \text{Normal}(0, \mathbf{I})$$

# Alternative framework on consistency

- *Existence of $\sqrt{n}$-consistent estimate*: Pick suitable $\alpha_n = O\left(n^{-1/2}\right)$. For any $\varepsilon > 0$, there exists constant $C > 0$ such that

$$\Pr\left(\sup_{\|u\|=C} h\left(\boldsymbol{\beta} + \alpha_n \mathbf{u}; \lambda\right) < h\left(\boldsymbol{\beta}; \lambda\right)\right) > 1 - \varepsilon$$

*iff* with probability tending to one, $h\left(\cdot; \lambda\right)$ has a local maximum $\hat{\boldsymbol{\beta}}\left(\lambda\right)$ in the ball

$$B_C\left(\boldsymbol{\beta}\right) = \left\{\boldsymbol{\beta} + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\right\} \text{ and } \left\|\hat{\boldsymbol{\beta}}\left(\lambda\right) - \boldsymbol{\beta}\right\| = O_{\Pr}\left(\alpha_n\right)$$

- "*Matching sparsity lemma*": for any given $\tilde{\boldsymbol{\beta}} = \left(\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T\right)^T$ such that $\left\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right\| = O_{\Pr}\left(n^{-1/2}\right)$ and any constant $C > 0$,

$$h\left(\left(\tilde{\boldsymbol{\beta}}_1^T, 0^T\right)^T; \lambda\right) = \max_{\left\|\tilde{\boldsymbol{\beta}}_2\right\| \leq C n^{-1/2}} h\left(\left(\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T\right)^T; \lambda\right)$$

# Summary on alternative framework on consistency

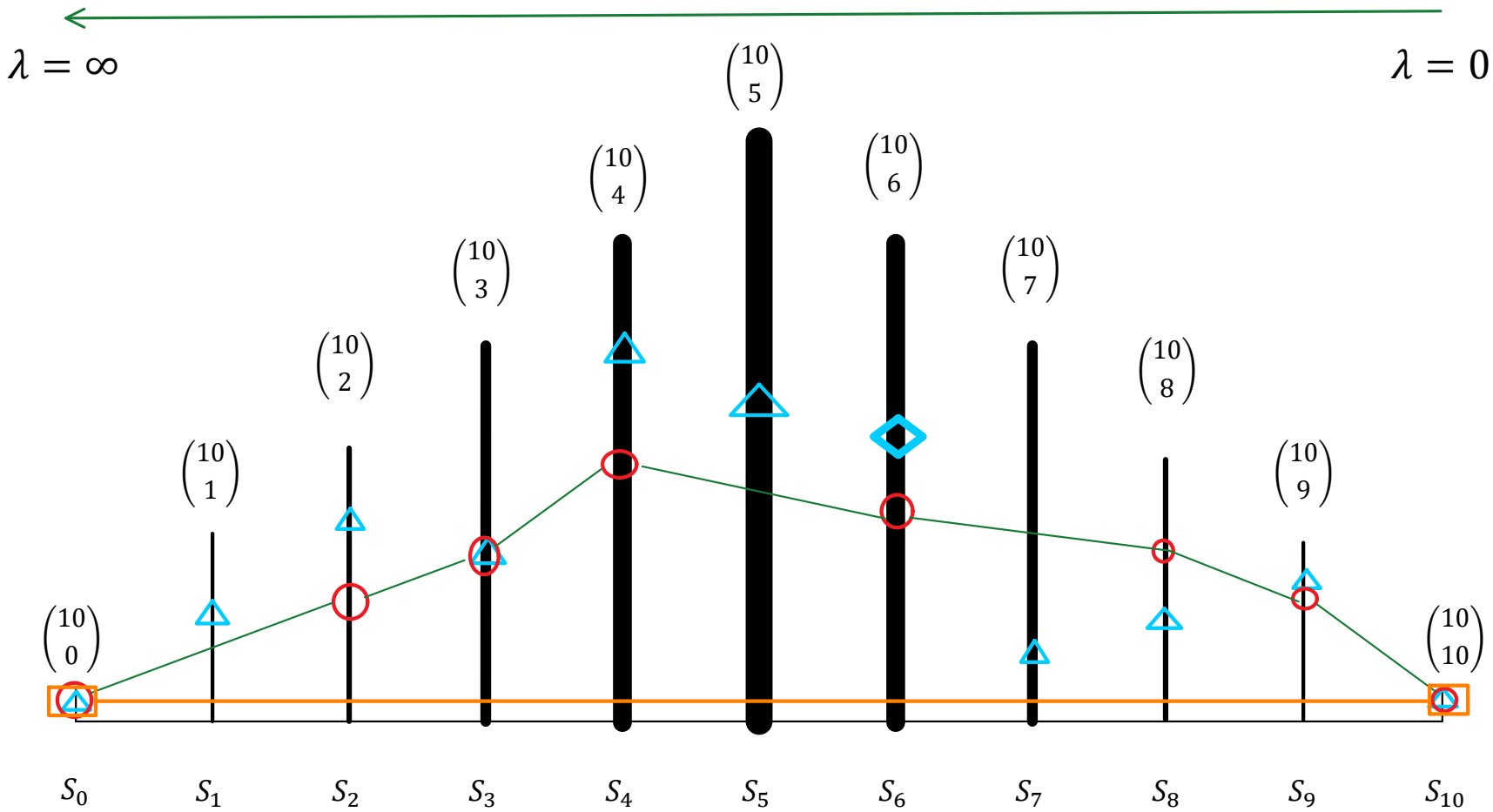
- Existence of $\sqrt{n}$-consistent estimate and matching sparsity lemma
  - form the backbone of this framework
  - will eventually lead to oracle property
  - rely heavily on properties of *sample size, number of coefficients, sparsity, magnitudes of non-zero coefficients, likelihood function, penalty, and tuning parameter* (which is true for other methods)
  - rely heavily on how the above *seven elements interact with each other* (which is true for other methods)
- The general principles of Fan and Li (2001) to achieve oracle property work for the regime of a diverging number of parameters (see, e.g., Fan and Peng (2004) and Fan and Lv (2011))
- Note that properties of design matrix has been incorporated into conditions that govern how likelihood function, penalty and tuning parameter should interact with each other

# Traditional versus modern

- Traditional variable/model selection *penalizes/controls model size*
  - Search algorithm: Best subset selection or forward/backward stepwise selection, all driven by, e.g., RSS
  - Model selection criteria: AIC; BIC; Mallow's $C_p$; adjusted R-squared; *cross-validation*
- Mordern variable/model selection *penalizes/controls effect size*
  - Search algorithm: *11homotopy" via tuning parameter*, i.e., the family of models searched is $\{\mathcal{M}_\lambda : \lambda \geq 0\}$
  - Model selection criterion: maximizing penalized/regularized (log)-likelihood function
  - Tuning parameter selection via *cross-validation and estimated prediction risk*
  - Implementation: pathwise coordinate optimization

# Search algorithm for model space

- "Circle", "Rectangle" and "Triangle": models selected by LASSO, Ridge and BSS; "Diamond": best model by BSS



# Visualizing estimates for orthogonal design

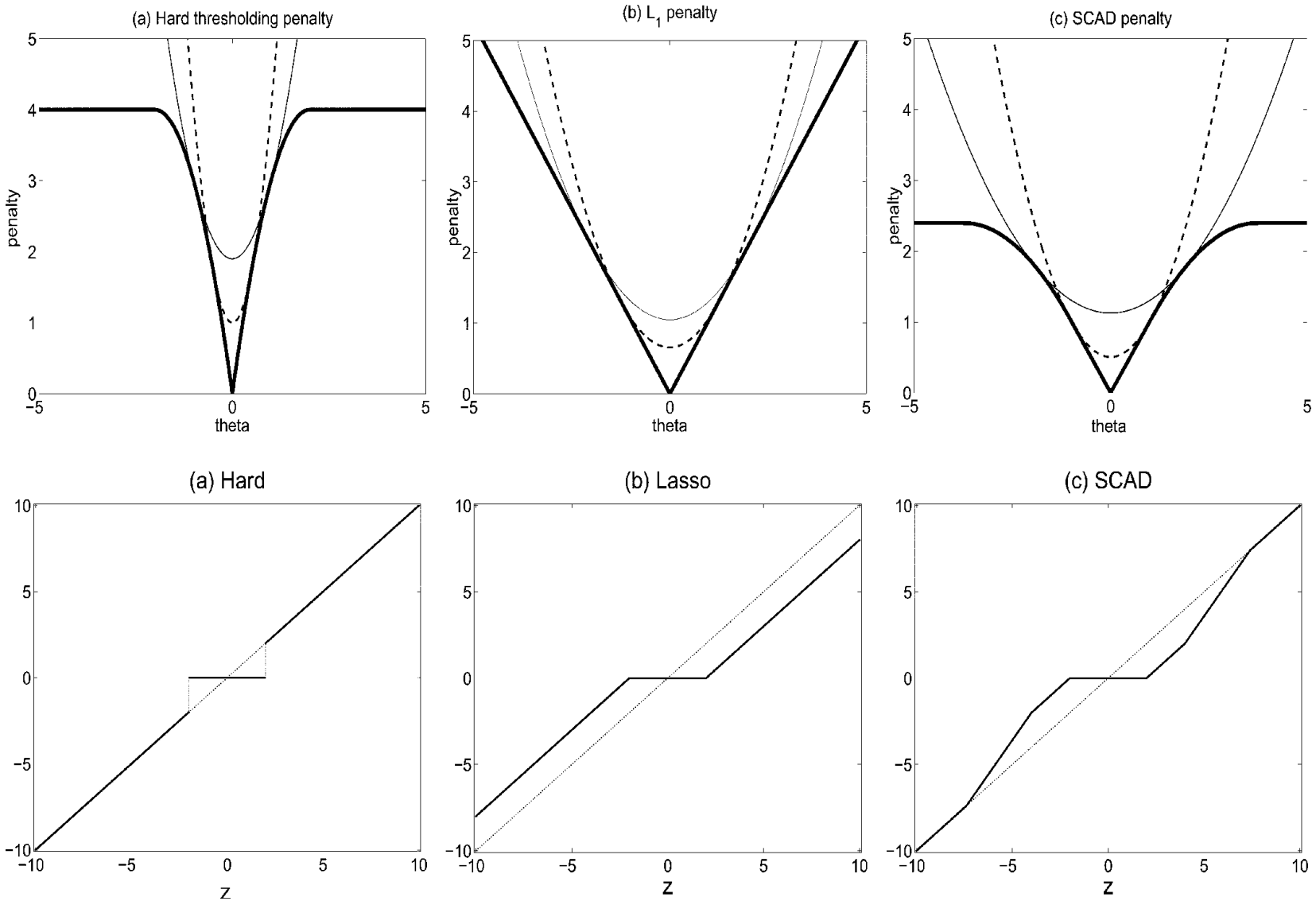



*Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.*

# References


Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Trans Inf Theory*, 57(8):5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of paramters. *Ann. Statist.*, 32(2):928–961.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.