Stat 577 "Statistical Learning Theory"

T3: Designing a penalty and Ridge estimator

Xiongzhi Chen

Washington State University

Regularization and shrinkage

Secrets behind shrinkage

- Stein's phenomenon:
  - Given a single observation $\mathbf{z}_0$ from $\mathbf{z} \sim \text{Normal}(\boldsymbol{\theta}, \mathbf{I})$ with $\mathbf{z} \in \mathbb{R}^p$. If $p \geq 3$, then the LSE $\hat{\boldsymbol{\theta}} = \mathbf{z}_0$ of $\boldsymbol{\theta}$ is inadmissible under the loss $E\left[\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2\right]$ compared to the estimator James-Stein estimator

$$\hat{\boldsymbol{\theta}}_{\text{JS}} = \mathbf{z}_0 \left[1 - (p-2)\left\|\mathbf{z}_0\right\|^{-2}\right]$$

- Hodges's super-efficiency phenomenon:
  - Given n i.i.d. observations $\{z_i\}_{i=1}^n$ from $z \sim \text{Normal}(\theta, 1)$ with $z \in \mathbb{R}$. Hodges estimator

$$\hat{\theta}_{n,\text{H}} = \begin{cases} \bar{z}_n & \text{if} \quad |\bar{z}_n| \geq n^{-1/4} \\ 0 & \text{if} \quad |\bar{z}_n| < n^{-1/4} \end{cases} \Rightarrow \begin{cases} n^{\alpha}\left(\hat{\theta}_{n,\text{H}} - \theta\right) \rightsquigarrow 0, \forall \alpha \in \mathbb{R} \\ \sqrt{n}\left(\hat{\theta}_{n,\text{H}} - \theta\right) \sim \sqrt{n}\left(\bar{z}_n - \theta\right) \end{cases}$$

Norms and sparsity

Some norms on Euclidean space

For a vector $\boldsymbol{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n$, we have

l_0-"norm":	$\|\boldsymbol{x}\|_0 = \sum_{i=1}^n 1_{\{x \neq 0\}}(x_i)$
l_p-norm:	$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n \|x_i\|^p\right)^{1/p}, 0 < p < \infty$
l_∞-norm:	$\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} \|x_i\|$

- $\|\boldsymbol{x}\|_0$ is the number of nonzero entries of $\boldsymbol{x}$, Euclidean norm of $\boldsymbol{x}$ is $\|\boldsymbol{x}\|_2$, and l_0-"norm" is not a norm
- Define $0^0 = 0$. Then

$$\lim_{p \to 0+} \|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_0 \quad \text{and} \quad \lim_{p \to \infty} \|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_\infty$$

- Different l_p-norms usually induce different l_p-balls

Norms: axioms and properties

- Definition of *norm* $\|\cdot\|_* : \mathcal{V} \to \mathbb{R}$:
  - Positive-definiteness: $\|\boldsymbol{x}\|_* \geq 0, \forall \boldsymbol{x}$ and $\|\boldsymbol{x}\|_* = 0$ iff $\boldsymbol{x} = 0$
  - Triangular inequality: $\|\boldsymbol{x} + \boldsymbol{y}\|_* \leq \|\boldsymbol{x}\|_* + \|\boldsymbol{y}\|_*$
  - Absolute homogeneity: $\|\lambda \boldsymbol{x}\|_* = |\lambda| \|\boldsymbol{x}\|_*, \forall \lambda \in \mathbb{R}$ and $\forall \boldsymbol{x}$
- $\|\cdot\|_*$ must be convex and Lipschitz, i.e.,

$$\begin{cases} \|\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|_* \leq |\lambda| \|\boldsymbol{x}\|_* + |1-\lambda| \|\boldsymbol{y}\|_* \\ |\|\boldsymbol{x}\|_* - \|\boldsymbol{y}\|_*| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_* \end{cases}$$

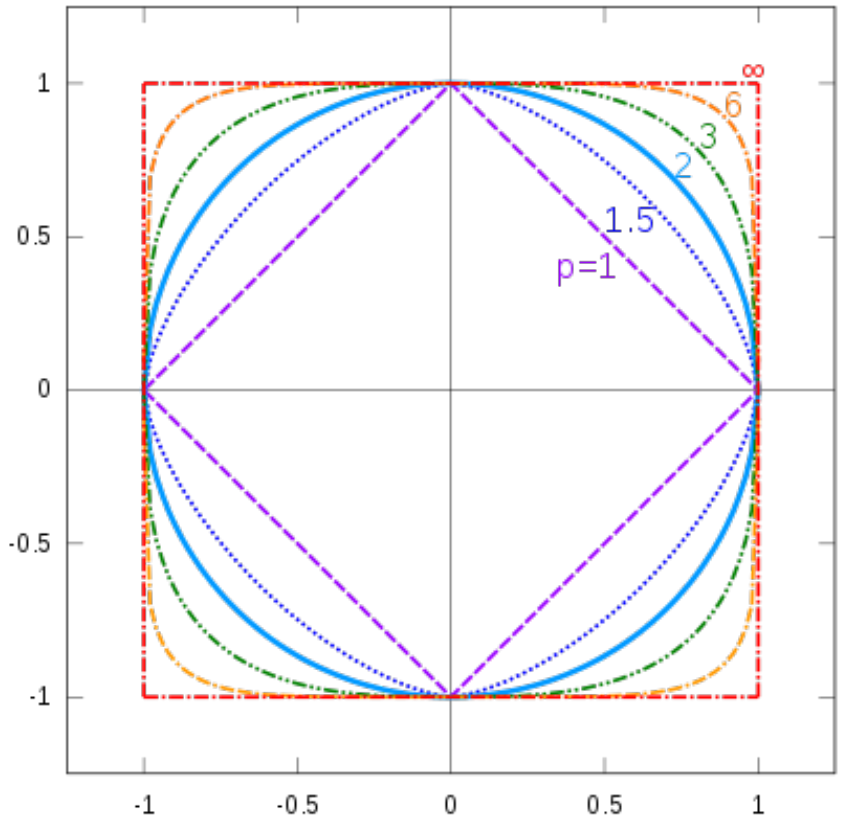
- l_p-norm, $0 < p \leq \infty$, are norms due to *Minkowski inequality*

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$$

  (which can be proved by *Hölder inequality*)
- l_0-"norm" is *not a norm* since it is NOT homogeneous

Geometry: visualization

- Unit l_p-ball $B_p = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_p \leq 1 \right\}$ in $\mathbb{R}^2$ (Credit: Wiki)



Linear model with scalar response

- Predictor $X = \left(X_1, ..., X_p\right)^T \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$
- Coefficient vector $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ and population model

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \tag{1}$$

  where $\varepsilon \in \mathbb{R}$ is the random error term $E\left(\varepsilon\right) = 0$ and $\text{var}\left(\varepsilon\right) = \sigma^2$
- Data version: set $\boldsymbol{\varepsilon} = \left(\varepsilon_1, ..., \varepsilon_N\right)^T \in \mathbb{R}^N$ and each $\varepsilon_i \sim \varepsilon$, then

$$\mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Variable/model selection for (1) is the same as retaining X_j whose $\beta_j \neq 0$

Searching through model space

Always keep β_0 in $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$

- Continuous in model size (and *controlling model size*):
  - Exhaustive: Best subset selection (BSS)
  - Non-exhaustive: Backward-stepwise selection; forward-stepwise selection
- Discontinuous in model size (and *controlling coefficient size*):
  - Non-exhaustive: regularization such as LASSO and "least angle regression (LAR)"
  - Non-exhaustive: implicit regularization via use of priors
- Hybrid:
  - Warm start without regularization and then search model space via regularization (such as "two-stage methods" including adaptive LASSO of Hui ZOU)

Concepts on sparsity

For an unknown coefficient vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$, a general concept of "sparsity" ϑ is

"the proportion of entries of $\boldsymbol{\beta}$ whose magnitude exceed a specific threshold"

- ϑ for variable selection (in linear model): $\|\boldsymbol{\beta}\|_0 \leq s$
- ϑ originated from quantum physics: $\|\boldsymbol{\beta}\|_q \leq s$ for $0 < q \leq \infty$

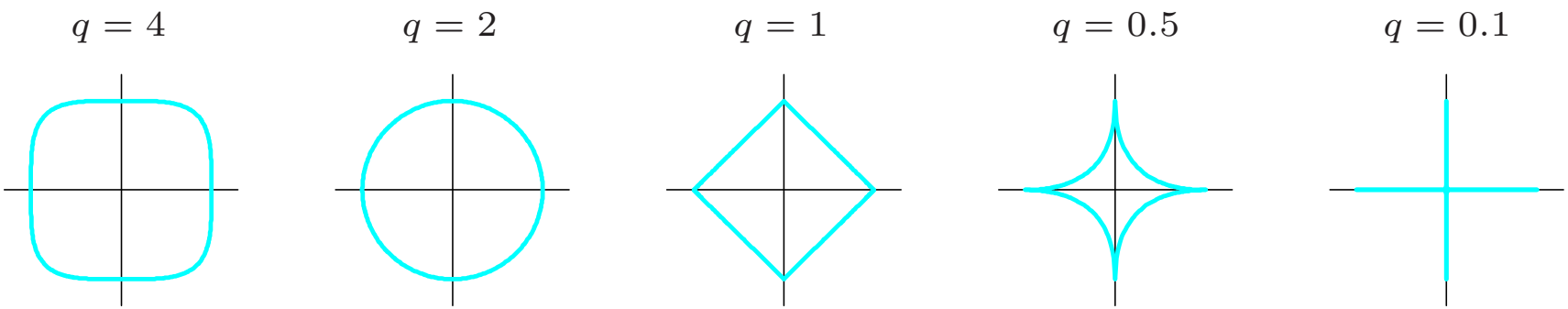


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

False discovery rate

Variable selection and multiple testing

- Model: $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ and

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \Leftrightarrow \mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Sparsity: $\|\boldsymbol{\beta}\|_0 = s$; e.g., $\min_{1 \le i \le s} \left|\beta_j\right| > 0$ and $\max_{s < j \le p} \left|\beta_j\right| = 0$
- Hypotheses: $H_{j0} : \beta_j = 0$ versus $H_{j1} : \beta_j \neq 0$ for $1 \le j \le p$
- Classification table for H_{j0}:

	Truth	
	$\beta_j = 0$	$\beta_j \neq 0$
Selected (reject H_{j0})	False positive	True positive
Excluded (retain H_{j0})	True negative	False negative

Variable selection and false discovery rate

- Classification table for all $\{H_{j0}\}_{j=1}^{p}$ upon variable selection:

	Null is **true**	Null is **false**	Total
Reject null	V	S	R
Retain null	U	T	$p-R$
Total	$p-s$	s	p

- The false discovery rate (FDR) α of the selection procedure $\mathcal{R}$:

$$\alpha(\mathcal{R}) = \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right]$$

- The proportion of true nulls π_0 and the proportion of false nulls π_1 are defined respectively as

$$\pi_0 = \frac{p-s}{p} \quad \text{and} \quad \pi_1 = \frac{s}{p}$$

Power of a variable selection procedure

- Classification table of selection procedure $\mathcal{R}$:

	Null is **true**	Null is **false**	Total
Reject null	V	S	R
Retain null	U	T	$p-R$
Total	$p-s$	s	p

- FDR of the selection procedure $\mathcal{R}$:

$$\alpha = \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right]$$

- Power of $\mathcal{R}$ is measured by "true discovery rate (TDR)"

$$\omega = \mathbb{E}\left[S/s\right]$$

- Other measures available on selection performance

The Benjamini-Hochberg procedure

- Let p_i be the p-value associated with H_{i0}
- Let $\left\{p_{(i)}\right\}_{i=1}^{m}$ be the order statistics of $\{p_i\}_{i=1}^{m}$ such that $p_{(j)} \leq p_{(j+1)}$
- Set

$$r = \max\left\{1 \leq k \leq m : p_{(k)} \leq \alpha k m^{-1}\right\}$$

if r is well-defined, then reject $H_{(i)}$ such that $p_{(i)} \leq p_{(r)}$; otherwise, no rejections are made

Theorem (Benjamini and Yekutieli (2001))

If the joint distribution of p-values satisfy PRDS on the subset I_0 of true null hypotheses and null p-values are super-uniform, then $\hat{\alpha} \leq \pi_0 \alpha$, with equality for independent p-values with continuous, uniform null distribution.

Target of variable selection

- Model $Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$
- Truth $\mathcal{A} = \{j : \beta_j \neq 0\}$ with $|\mathcal{A}| = s$
- Hypotheses $H_{j0} : \beta_j = 0$ versus $H_{j1} : \beta_j \neq 0$ for $1 \leq j \leq p$
- Selection $\mathcal{I} = \{j : H_{j0} \text{ is rejected}\}$
- Performance criteria:
  - "Strong selection consistency": $\Pr(\mathcal{I} = \mathcal{A}) \to 1$
  - "Weak selection consistency": $\Pr(\mathcal{I} \supseteq \mathcal{A}) \to 1$
  - "Estimation consistency": $\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|\right] \to 0$
  - "Oracle property": $\mathcal{I} = \mathcal{A}$ and

$$\mathbf{A}_n \left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{I}} - \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{b}_n\right)\right] \rightsquigarrow \text{Normal}(0, \mathbf{I})$$

Some relationships

- FDR and TDR of the selection procedure $\mathcal{R}$:

$$\alpha = \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right] \text{ and } \omega = \mathbb{E}\left[S/s\right]$$

- Strong or week selection consistency:

$$\Pr\left(\mathcal{I} = \mathcal{A}\right) \to 1 \text{ or } \Pr\left(\mathcal{I} \supseteq \mathcal{A}\right) \to 1$$

- Oracle property and estimation consistency:

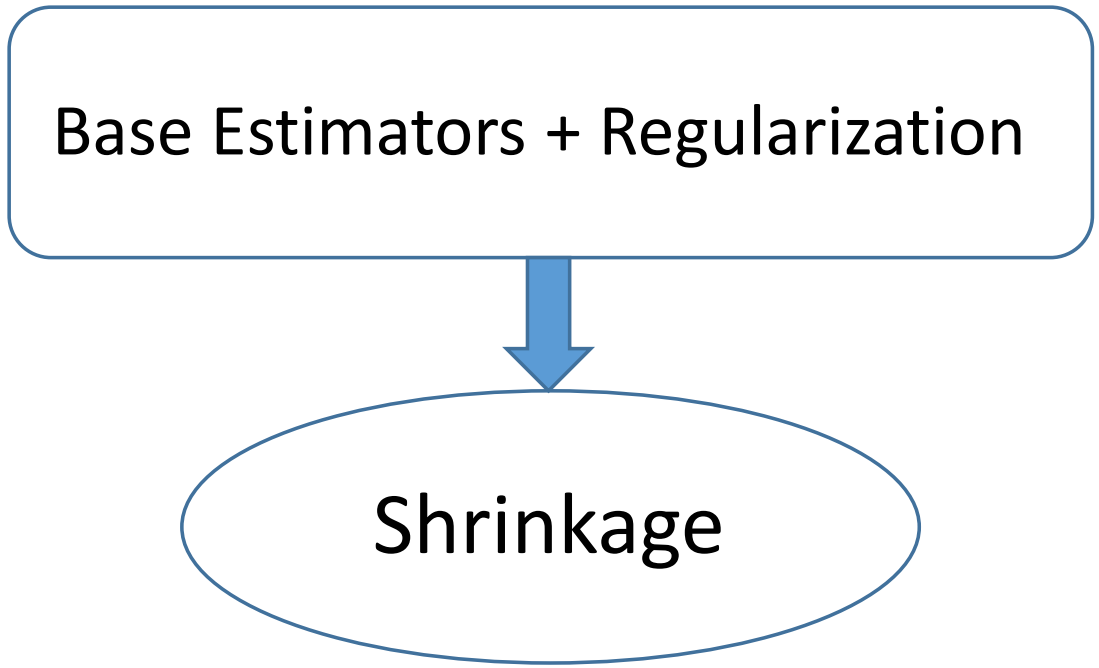
$$\begin{cases} \mathcal{I} = \mathcal{A} \text{ and } \mathbf{A}_n\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{I}} - \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{b}_n\right)\right] \rightsquigarrow \text{Normal}\left(0, \mathbf{I}\right) \\ \mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|\right] \to 0 \end{cases}$$

- Do you know their relationships?

Shrinkage-inducing penalty

General principle to induce shrinkage

- Base estimators: least squares estimate, maximum likelihood estimator, moment estimator, or kernel estimator
- Regularization: implicit/explicit and/or deterministic/random regularization



Penalized RSS

- RSS: e.g., $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ or

$$\sum_{i=1}^{N} \left(y_i - f\left(\mathbf{x}_i\right)\right)^2 = \|\mathbf{y} - f\left(\mathbf{X}\right)\|^2$$

- Penalized RSS:

$$\text{Penalized RSS} = \text{RSS} + \text{Penalty}$$

- Example models: additive model

$$E\left(Y|X\right) = f\left(X\right) + \varepsilon$$

Penalized likelihood

- Likelihood: $h\left(y; \text{other parameters}\right)$ and $l\left(\mathbf{y}; \text{other parameters}\right)$

$$l\left(\mathbf{y}; *\right) = \prod_{i=1}^{n} h\left(y_i; *\right)$$

- Penalized likelihood:

$$\text{Penalized likelihood} = \text{Likelihood} + \text{Penalty}$$

- Example models:

$$g\left[E\left(Y|X\right)\right] = f\left(X\right) \text{ and } g \text{ is a link function}$$

Designing a penalty

A *shrinkage-inducing* penalty should produce an estimator that possesses

- **Continuity**: estimator is continuous in data input to avoid instability in **model prediction**
- for purpose of *variable selection*
  1. **Unbiasedness**: estimator is nearly unbiased when true unknown parameter is large to avoid unnecessary **modeling bias**
  2. **Sparsity**: estimator incorporates a thresholding rule, which automatically sets small estimated coefficients to zero to reduce **model complexity**

Various penalties

Recall $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ and penalty $\varrho : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$.

- Separable penalty: $\varrho\left(\boldsymbol{\beta}\right) = \sum_{i=1}^p \varrho_i\left(\beta_i\right)$ with $\varrho_i : \mathbb{R} \to \mathbb{R}_{\geq 0}$, such as
  - the l_q-penalty

$$\varrho\left(\boldsymbol{\beta}\right) = \|\boldsymbol{\beta}\|_q^q = \sum_{i=1}^p |\beta_i|^q, 0 \leq q < \infty,$$

    including BSS, LASSO and Ridge
- Non-separable penalty: $\varrho\left(\boldsymbol{\beta}\right) \neq \sum_{i=1}^p \varrho_i\left(\beta_i\right)$, such as
  - Benjamini-Hochberg (BH) procedure; see Abramovich et al. (2006)
- Note: BH procedure induces a random penalty

Orthogonal design and penalized RSS

- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Then LSE $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ and

$$\mathrm{RSS}\left(\boldsymbol{\beta}\right) = \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 = \left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2 + \left\|\mathbf{z} - \boldsymbol{\beta}\right\|^2,$$

  where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ since $\mathbf{y} - \hat{\mathbf{y}} \perp \langle \mathbf{X} \rangle$ and $\mathbf{X}^T\mathbf{X} = \mathbf{I}$
- Further, RSS with *separable, piecewise smooth penalty* is

$$\begin{aligned} \mathrm{pRSS}\left(\boldsymbol{\beta}\right) &:= \frac{1}{2}\mathrm{RSS}\left(\boldsymbol{\beta}\right) + \sum_{i=1}^{p} \varrho_\lambda\left(\left|\beta_i\right|\right) \\ &= \frac{1}{2}\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2 + \frac{1}{2}\left\|\mathbf{z} - \boldsymbol{\beta}\right\|^2 + \sum_{i=1}^{p} \varrho_\lambda\left(\left|\beta_i\right|\right) \end{aligned} \tag{2}$$

Componentwise optimization

- Let $\boldsymbol{\theta} = \left(\theta_1, ..., \theta_p\right)^T$ and $\mathbf{z} = \left(z_1, ..., z_p\right)^T$. We have equivalent objective function to pRSS $(\boldsymbol{\beta})$ as

$$u\left(\boldsymbol{\theta}\right) = \frac{1}{2}\left\|\mathbf{z} - \boldsymbol{\theta}\right\|^2 + \sum_{i=1}^{p} \varrho_\lambda\left(\left|\theta_i\right|\right) = \frac{1}{2}\sum_{i=1}^{p}\left(z_i - \theta_i\right)^2 + \sum_{i=1}^{p} \varrho_\lambda\left(\left|\theta_i\right|\right)$$

- First order condition for minimizer $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, ..., \hat{\theta}_p\right)^T$ from

$$a\left(\theta\right) = \frac{1}{2}\left(z - \theta\right)^2 + \varrho_\lambda\left(\left|\theta\right|\right) \quad \text{and} \quad u\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{p} a_i\left(\theta\right)$$

gives

$$\frac{\partial a\left(\theta\right)}{\partial \theta} = \begin{cases} \operatorname{sgn}\left(\theta\right)\left[\left|\theta\right| + \varrho'_\lambda\left(\left|\theta\right|\right)\right] - z & \text{if} \quad \theta \neq 0 \\ \text{subdifferential } \partial a\left(\theta\right) & \text{if} \quad \theta = 0 \end{cases}$$

Shrinkage without hard thresholding

- Recall $a(\theta) = \frac{1}{2}(z-\theta)^2 + \varrho_\lambda(|\theta|)$ and

$$\frac{\partial a(\theta)}{\partial \theta} = \begin{cases} \operatorname{sgn}(\theta)\left[|\theta| + \varrho'_\lambda(|\theta|)\right] - z & \text{if} \quad \theta \neq 0 \\ \text{subdifferential } \partial a(\theta) & \text{if} \quad \theta = 0 \end{cases}$$

- Let $\varrho_\lambda(|\theta|) = 2^{-1}\lambda\theta^2$ for $\lambda > 0$. Then $\frac{\partial a(\theta)}{\partial \theta}$ exits at 0 and the optimal solution

$$\hat{\theta} = \frac{z}{1+\lambda} \tag{3}$$

- Note LSE $\hat{\theta} = z$. Namely, (3) is a "shrinkage" estimator of z (since $|\hat{\theta}| < |z|$ whenever $\lambda > 0$)
- (3) is the *ridge estimator* of z (when $\mathbf{X}^T\mathbf{X} = \mathbf{I}$)

Shrinkage via hard thresholding

- Recall $a(\theta) = \frac{1}{2}(z-\theta)^2 + \varrho_\lambda(|\theta|)$ and

$$\frac{\partial a(\theta)}{\partial \theta} = \begin{cases} \operatorname{sgn}(\theta)\left[|\theta| + \varrho'_\lambda(|\theta|)\right] - z & \text{if} \quad \theta \neq 0 \\ \text{subdifferential } \partial a(\theta) & \text{if} \quad \theta = 0 \end{cases}$$

- Since $\partial a(\theta) = \theta - z + \lambda v$ for $|v| \leq 1$ at $\theta = 0$ for penalty

$$\varrho_\lambda(|\theta|) = 2^{-1}\lambda^2 - 2^{-1}(|\theta| - \lambda)^2 1_{\{|\theta|<\lambda\}}(\theta),$$

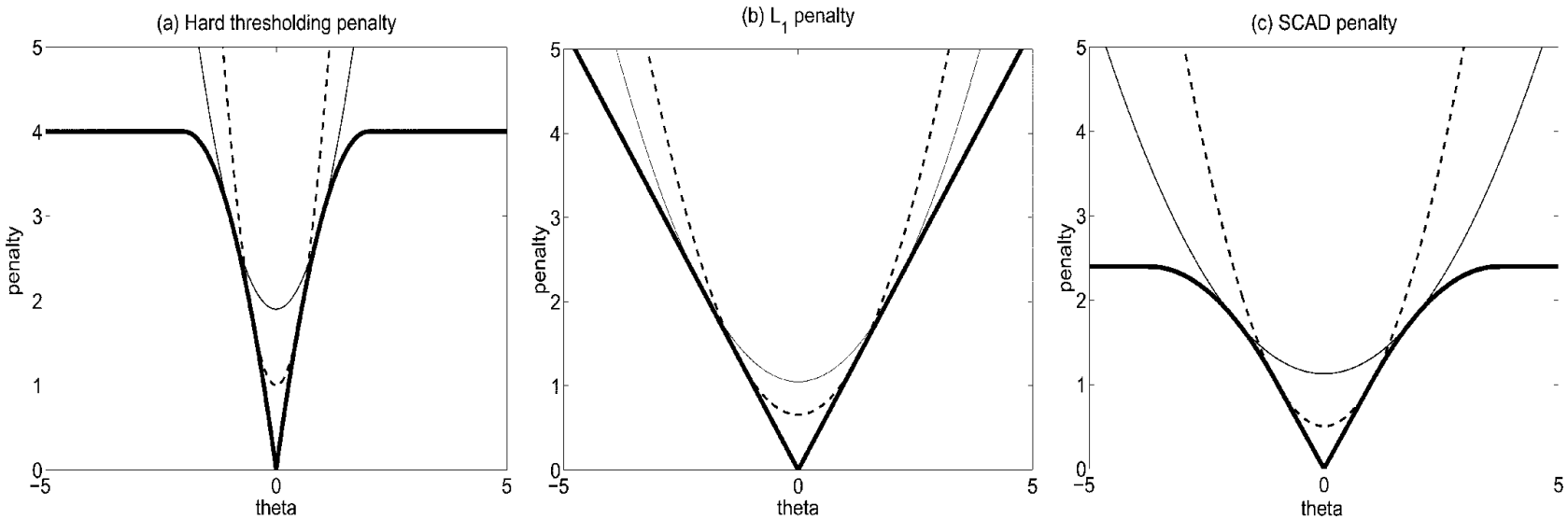
then $\hat{\theta} = 0$ iff $|z| \leq \lambda$, yielding the "hard thresholding" estimator

$$\hat{\theta} = z 1_{\{|z|>\lambda\}}(z) \tag{4}$$

- (4) is equivalent to best subset selection when $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Why?

Three penalties for variable selection

- Hard-thresholding penalty; l_1-penalty for LASSO, and "smoothly clipped absolute deviation" as SCAD
- We want unbiasedness, sparsity and continuity of estimate. Which penalty is preferred?



Nearly unbiasedness and thresholding

- Recall $a(\theta) = \frac{1}{2}(z-\theta)^2 + \varrho_\lambda(|\theta|)$ and

$$\frac{\partial a(\theta)}{\partial \theta} = \begin{cases} \operatorname{sgn}(\theta)\left[|\theta| + \varrho'_\lambda(|\theta|)\right] - z & \text{if} \quad \theta \neq 0 \\ \text{subdifferential } \partial a(\theta) & \text{if} \quad \theta = 0 \end{cases}$$

- *Nearly unbiasedness*: $\hat{\theta} = z$ for $|z|$ large if $\varrho'_\lambda(|\theta|) = 0$ for large $|\theta|$
- Thresholding rule: $\hat{\theta} = 0$ for small $|z|$, i.e.,

$$\hat{\theta} = 0 \text{ when } |z| < \tau_\lambda = \inf_{\theta \neq 0}\left\{|\theta| + \rho'_\lambda(|\theta|)\right\}$$

$$\text{iff} \quad \tau_\lambda > 0 \ \& \ \operatorname*{arginf}_{\theta \in \mathbb{R}}\left\{|\theta| + \rho'_\lambda(|\theta|)\right\} = 0.$$

- Namely, $\rho_\lambda(|\theta|)$ has to be non-differentiable at $\theta = 0$ to induce thresholding rule

Thresholding rule: visualization

- $\underset{\theta\in\mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2}(z-\theta)^2 + \varrho_\lambda(|\theta|) \right\}$ when $\tau_\lambda = \inf_{\theta\neq 0} \left\{ |\theta| + \rho'_\lambda(|\theta|) \right\} > 0$
- $\hat{\theta}(z_1) = 0$ but $\hat{\theta}(z_2) = \sup\left\{ \theta : \frac{\partial a(\theta)}{\partial \theta} = 0 \right\}$ to retain continuity

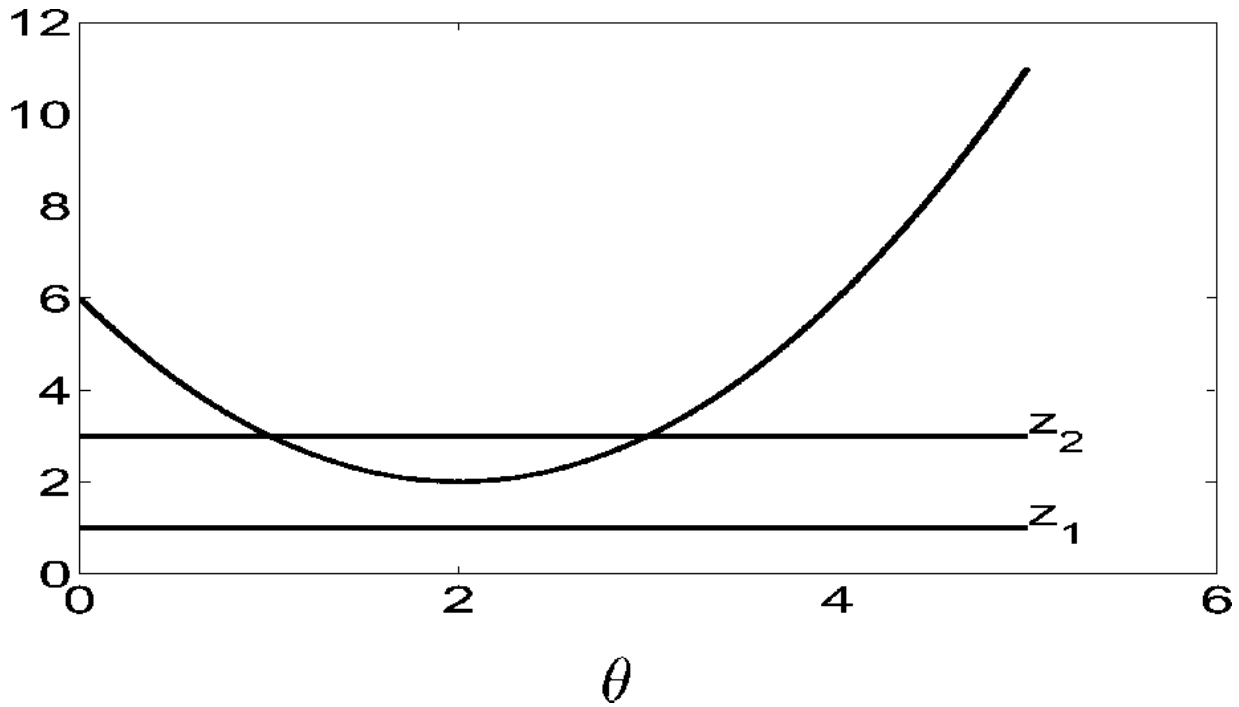


Figure 3. A Plot of $\theta + p'_\lambda(\theta)$ Against $\theta(\theta > 0)$.

Derivatives of three penalties

- Hard-thresholding penalty: $\varrho_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 1_{\{|\theta|<\lambda\}}(\theta)$
- SCAD penalty: $\varrho'_\lambda(\theta) = \lambda\left[1_{\{\theta\le\lambda\}}(\theta) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} 1_{\{\theta>\lambda\}}(\theta)\right], \theta > 0$
- l_q-penalty: $\varrho_\lambda(|\theta|) = \lambda|\theta|^q$

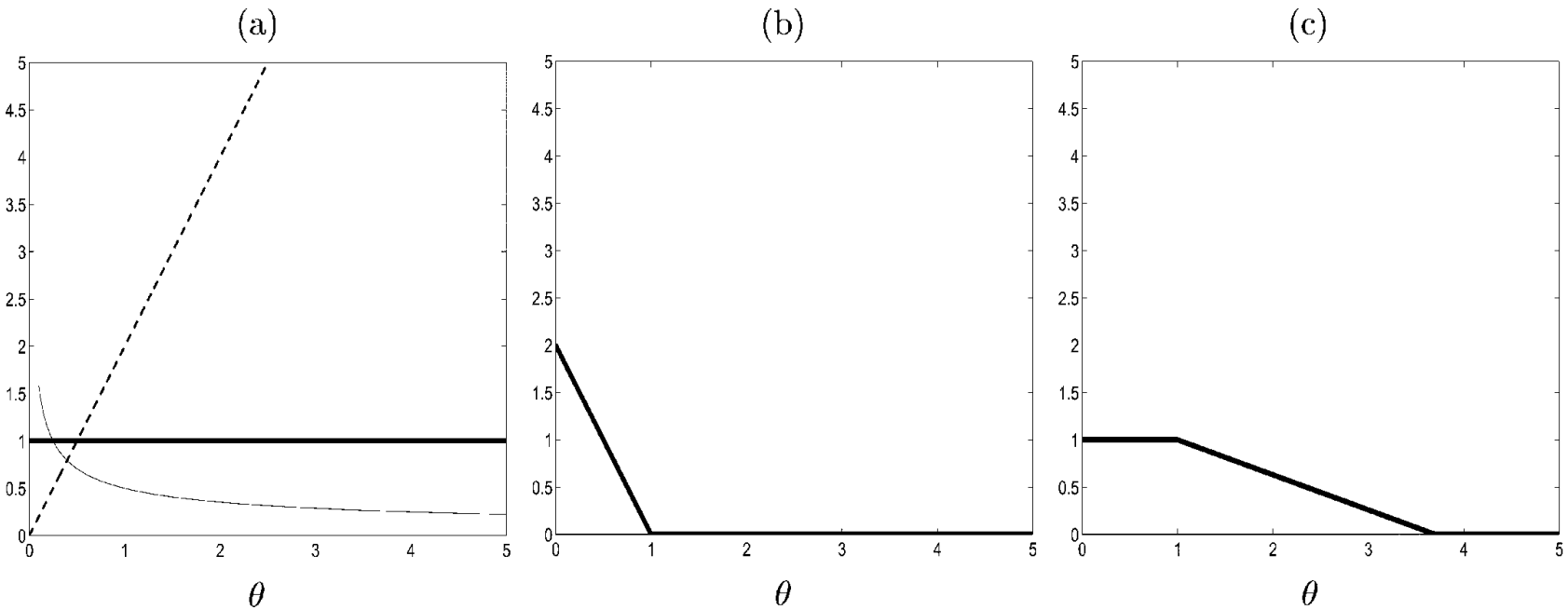
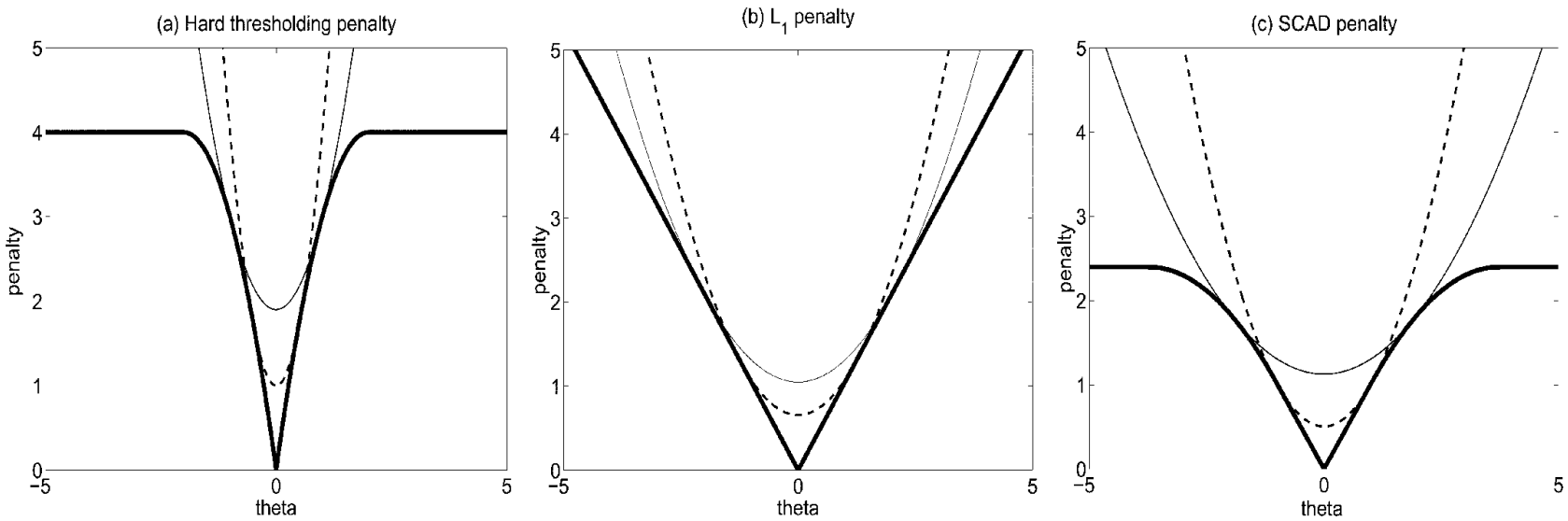


Figure 4. Plot of $p'_\lambda(\theta)$ Functions Over $\theta > 0$ (a) for L_q Penalties, (b) the Hard Thresholding Penalty, and (c) the SCAD Penalty. In (a), the heavy line corresponds to L_1, the dash-dot line corresponds to $L_{.5}$, and the thin line corresponds to L_2 penalties.

Recap on three penalties for variable selection

- Pay attention to the "rounding" parts of a hard thresholding penalty and SCAD penalty
- None of these penalties are differentiable at 0
- All are even functions on $\mathbb{R}$ and non-decreasing on $\mathbb{R}_{>0}$



Estimates given by three penalties

When tuning parameter $\lambda > 0$:

- Hard thresholding estimate is discontinuous in data
- LASSO estimate is intrinsically biased
- Both LASSO and SCAD incorporate hard thresholding

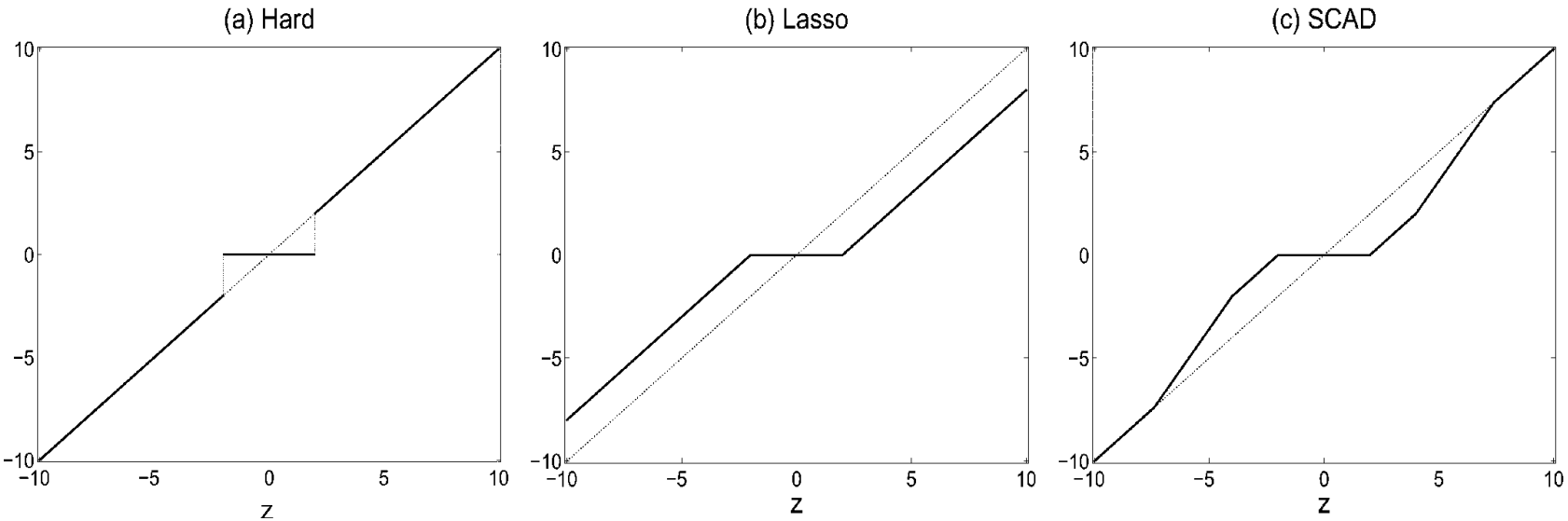


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

Recap on penalties for variable selection: I

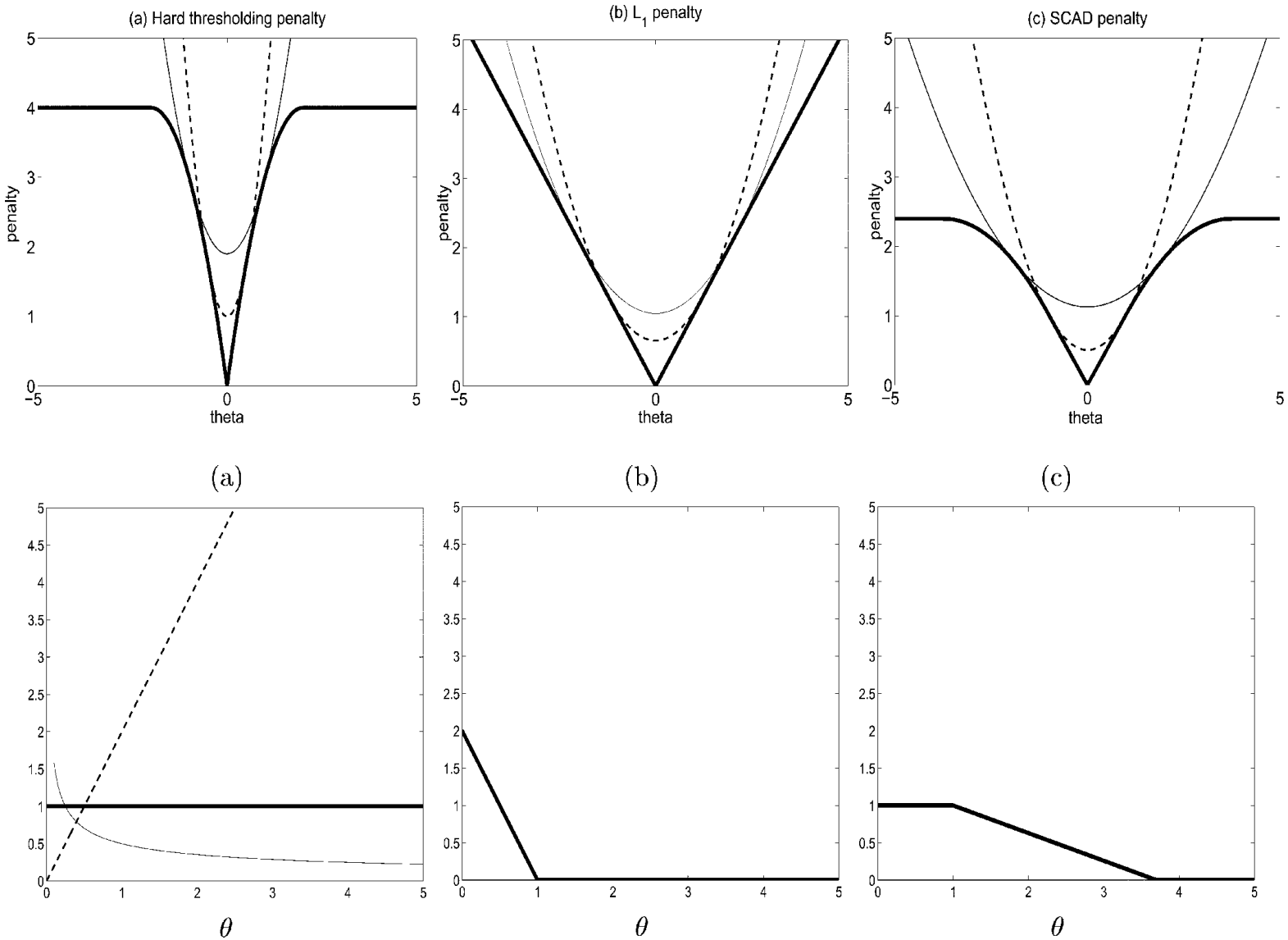


Figure 4. Plot of $p'_\lambda(\theta)$ Functions Over $\theta > 0$ (a) for L_q Penalties, (b) the Hard Thresholding Penalty, and (c) the SCAD Penalty. In (a), the heavy line corresponds to L_1, the dash-dot line corresponds to $L_{.5}$, and the thin line corresponds to L_2 penalties.

Recap on penalties for variable selection: II

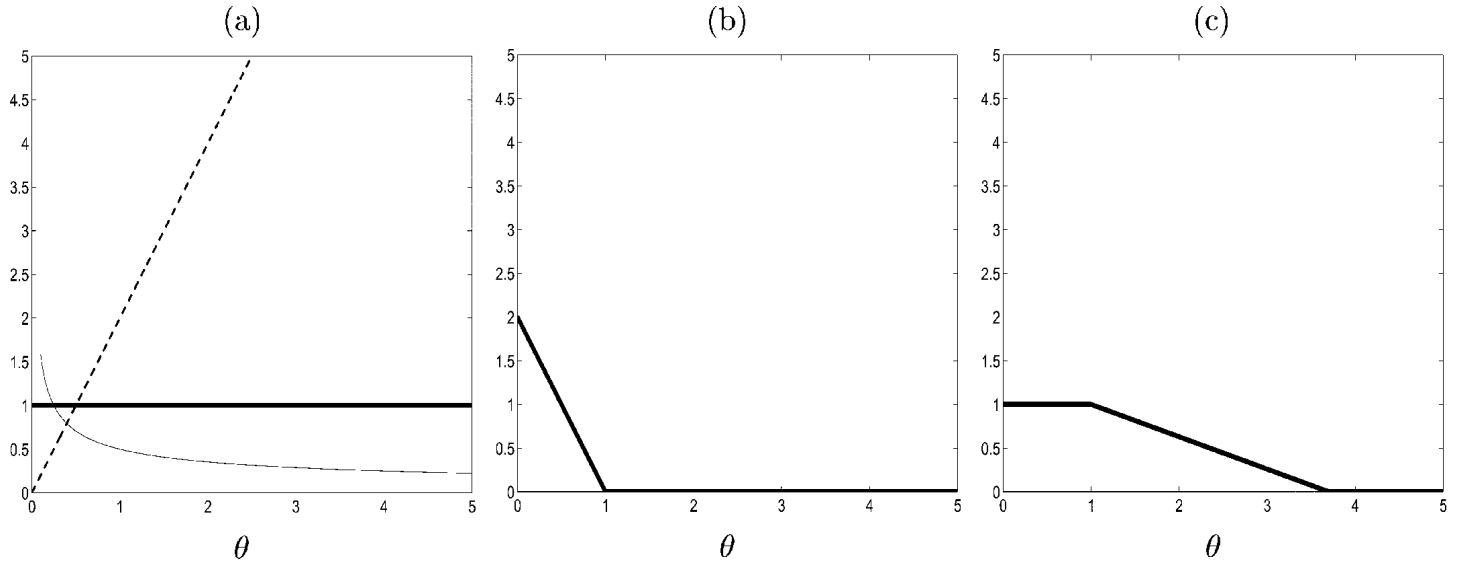


Figure 4. Plot of $p'_\lambda(\theta)$ Functions Over $\theta > 0$ (a) for L_q Penalties, (b) the Hard Thresholding Penalty, and (c) the SCAD Penalty. In (a), the heavy line corresponds to L_1, the dash-dot line corresponds to $L_{.5}$, and the thin line corresponds to L_2 penalties.

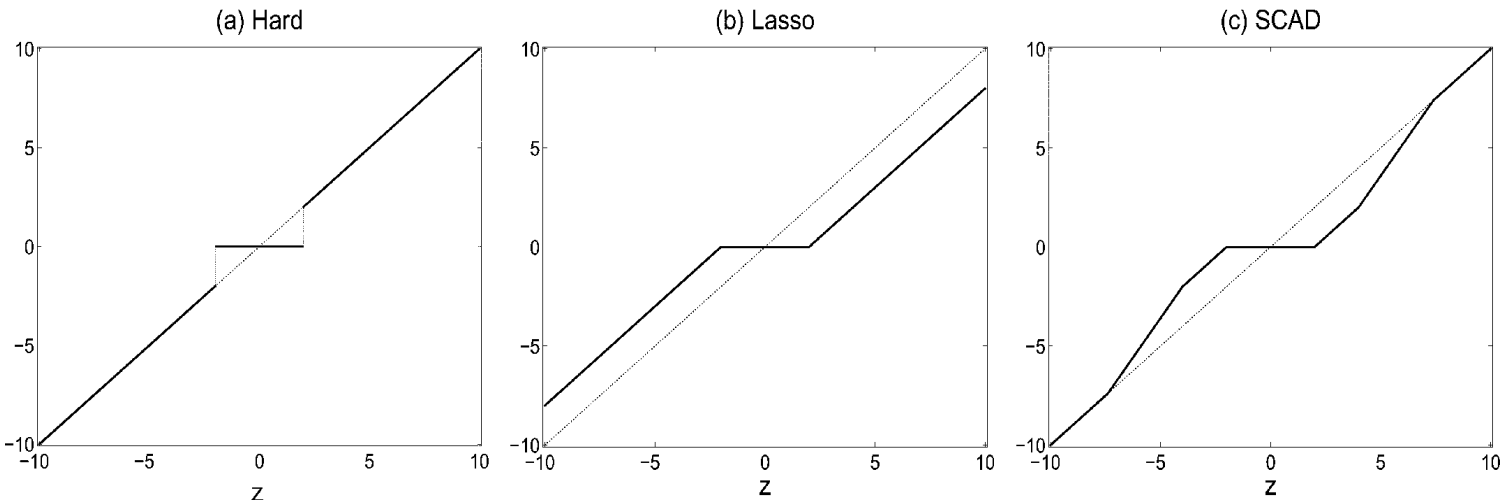


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

Geometry of linear transform

Singular value decomposition

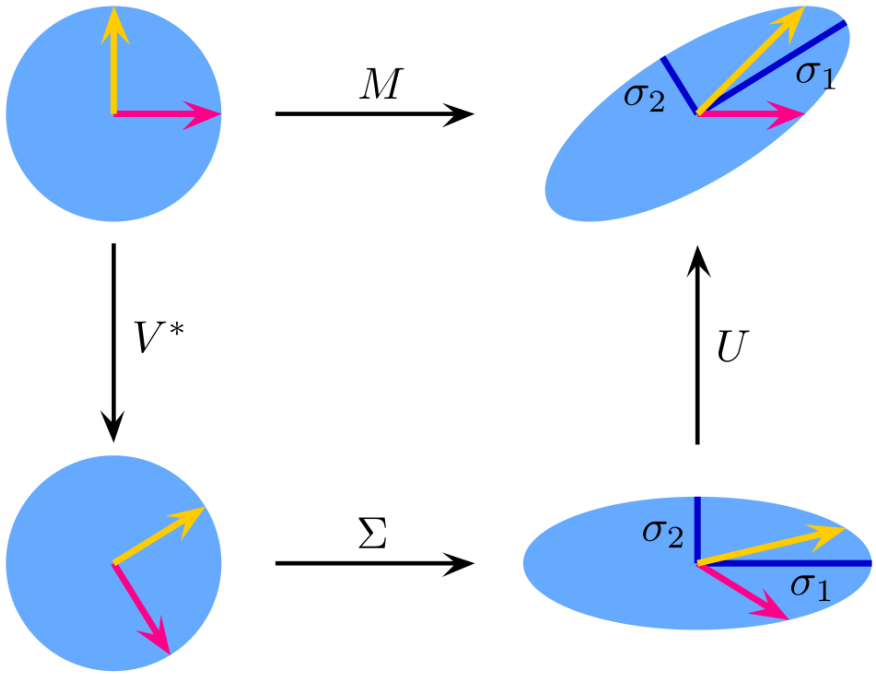
Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $r = \text{rank}(\mathbf{A})$. The "singular value decomposition (SVD)" of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^T$$

- $\mathbf{D} = \text{diag}\{d_1, \ldots, d_r\}$ and $d_1 \geq \cdots \geq d_r > 0$. Each d_i is called a "singular value" of $\mathbf{A}$
- $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$ such that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_r$. Further, $\langle \mathbf{U} \rangle = \langle \mathbf{A} \rangle$ and $\langle \mathbf{V} \rangle = \langle \mathbf{A}^T \rangle$
- $\mathbf{u}_i$ is the ith column of $\mathbf{U}$ and is called a "left singular vector", and $\mathbf{v}_i$ the ith column of $\mathbf{V}$ and is called a "right singular vector"
- $(\mathbf{U}, \mathbf{D}, \mathbf{V})$ above can be extended into $\mathbf{U} \in \mathbb{R}^{m \times n}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ and $\mathbf{D} = \text{diag}\{d_1, \ldots, d_n\}$ such that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ and $d_{r+1} = \cdots = d_n = 0$

Geometric meaning of SVD: illustration

- SVD $M = U\Sigma V^*$ where Σ is diagonal; note the rotations induced by U and V^* and scaling induced by Σ (Credit: Wiki)



$$M = U \cdot \Sigma \cdot V^*$$

Spectral decomposition

Let $\mathbf{A} \in \mathbb{R}^{m \times m}$ with $\mathbf{A} = \mathbf{A}^T$. The "spectral decomposition" of $\mathbf{A}$ is

$$\mathbf{A} = \tilde{\mathbf{V}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T = \sum_{i=1}^{m} \tilde{d}_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i^T$$

- $\tilde{\mathbf{D}} = \text{diag}\left\{\tilde{d}_1, \ldots, \tilde{d}_m\right\}$ with $\left|\tilde{d}_1\right| \geq \cdots \geq \left|\tilde{d}_m\right|$, and each $\tilde{d}_i$ is called an "eigenvalue" of $\mathbf{A}$
- $\tilde{\mathbf{V}} \in \mathbb{R}^{m \times m}$ such that $\tilde{\mathbf{V}}^T \tilde{\mathbf{V}} = \mathbf{I}_m$, and the ith column $\tilde{\mathbf{v}}_i$ of $\tilde{\mathbf{V}}$ is an eigenvector of $\mathbf{A}$ associated with $\tilde{d}_i$
- $\left(\tilde{d}_i, \tilde{\mathbf{v}}_i\right)$ is called the "ith eigenpair"

Remark: Spectral decomposition $\Longleftrightarrow$ SVD; decompositions hold for Hermitian matrix $\mathbf{A} = \mathbf{A}^*$ with operation "conjugate transpose *"

Maximal scaling

- $\mathbf{A} \in \mathbb{R}^{m \times n}$ induces a linear map $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ via *right multiplication* $\mathbf{w} = \mathbf{A}\mathbf{v}$
- Set $\mathbf{S} = \mathbf{A}^T\mathbf{A}$. Then $\mathbf{S} \succeq 0$, i.e., $\mathbf{S}$ is positive semi-definite, and

$$\|\mathbf{w}\|^2 = \|\mathbf{A}\mathbf{v}\|^2 = \mathbf{v}^T \left(\mathbf{A}^T\mathbf{A}\right) \mathbf{v} = \mathbf{v}^T\mathbf{S}\mathbf{v}$$

- SVD $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ gives $\mathbf{S} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$ with $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_n\}$ where

$$d_1 \geq \cdots \geq d_n \geq 0$$

- Set $\|\mathbf{A}\|_{\mathrm{op}} = \max_{1 \leq i \leq n} d_i = d_1$. **Rayleigh's inequality**:

$$\|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{v}\|,$$

i.e., $\mathbf{v} \in \mathbb{R}^n$ can be scaled at most $\|\mathbf{A}\|_{\mathrm{op}}$ by $\mathbf{A}$ when mapped to $\mathbf{A}\mathbf{v}$

Rayleigh's inequality

- Recall $\mathbf{S} = \mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{D} = \text{diag}\{d_1, \ldots, d_n\}$
- Recall $\mathbf{w} = \mathbf{A}\mathbf{v}$. So, $\|\mathbf{w}\|^2 = \mathbf{v}^T\mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{v}$. Let $\tilde{\mathbf{v}} = \mathbf{V}^T\mathbf{v}$. Then

$$\|\tilde{\mathbf{v}}\| = \|\mathbf{v}\|, \text{ i.e., orthogonal transform preserves length}$$

Recall $\|\mathbf{A}\|_{\text{op}} = \max_{1\le i\le n} d_i$. Then

$$\|\mathbf{A}\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{v} = \tilde{\mathbf{v}}^T\mathbf{D}^2\tilde{\mathbf{v}} \le \|\mathbf{A}\|_{\text{op}}^2 \|\tilde{\mathbf{v}}\|^2 = \|\mathbf{A}\|_{\text{op}}^2 \|\mathbf{v}\|^2$$

with equality iff $\mathbf{v} = c\mathbf{v}_1$ with $(d_1, \mathbf{v}_1)$ and $c \neq 0$

- So, $\|\mathbf{A}\mathbf{v}\| \le \|\mathbf{A}\|_{\text{op}} \|\mathbf{v}\|$ and

$$\left(\max_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|, \text{argmax}_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|\right) = \left(\|\mathbf{A}\|_{\text{op}}, \mathbf{v}_1\right)$$

Successive maximal scaling

- Recall $\mathbf{S} = \mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$ with $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and

$$\mathbf{D} = \mathsf{diag}\left\{d_1, \ldots, d_n\right\} \quad \text{and} \quad d_1 \geq \cdots \geq d_n \geq 0$$

- Rayleigh's inequality: $\max_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\|_{\text{op}} = d_1$
- Recall $\mathbf{v}_i$ as ith column of $\mathbf{V}$ and eigenpair $(d_i, \mathbf{v}_i)$. Same arguments (plus induction) can show, for $1 \leq k \leq n-1$,

$$\left(\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle} \|\mathbf{A}\mathbf{v}\|, \operatorname*{argmax}_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle} \|\mathbf{A}\mathbf{v}\| \right) = (d_{k+1}, \mathbf{v}_{k+1})$$

  by noting that $\{\mathbf{v}_i\}_{i=1}^n$ are orthonormal and constraint $\|\mathbf{v}\| = 1$
- The above is called "sucessive maximal scaling (SMS)" lemma

Successive maximal scaling: sketch of proof

- Let $\left(\max_{\|\mathbf{v}\|=1} \|\mathbf{Av}\|, \operatorname{argmax}_{\|\mathbf{v}\|=1} \|\mathbf{Av}\|\right) = (d_1, \mathbf{v}_1)$
- Recall $\mathcal{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. Since $\{\mathbf{v}_i\}_{i=1}^n$ are orthonormal, then $\|\mathbf{v}\| = 1$ and $\mathbf{v} \perp \mathbf{v}_1$ iff $\mathbf{v} \in \langle \{\mathbf{v}_i\}_{i=2}^n \rangle$ and

$$\exists \mathbf{a} = (a_1, ..., a_{n-1})^T \in \mathcal{S}^{n-1} \text{ such that } \mathbf{v} = \sum_{i=2}^{n} a_{i-1}\mathbf{v}_i$$

- So, $\mathbf{Av} = \sum_{i=2}^n a_{i-1}\mathbf{Av}_i$. But $\{\mathbf{v}_i\}_{i=1}^n$ are orthonormal eigenvectors of $\mathbf{S} = \mathbf{A}^T\mathbf{A}$, So,

$$\|\mathbf{Av}\|^2 = \sum_{i=2}^{n} a_{i-1}^2 \|\mathbf{Av}_i\|^2 = \sum_{i=2}^{n} a_{i-1}^2 d_i^2 \leq d_2^2 \|\mathbf{a}\|^2 = d_2^2,$$

  i.e., $\left(\max_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_1} \|\mathbf{Av}\|, \operatorname{argmax}_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_1} \|\mathbf{Av}\|\right) = (d_2, \mathbf{v}_2)$
- Remaining part done via induction

Orthogonal projections and PCA

Directions of successive maximal variability

- Fact: if $\mathbf{v} \in \mathcal{S}^{n-1}$ and $\boldsymbol{x} \in \mathbb{R}^n$, then their inner product $\langle \boldsymbol{x}, \mathbf{v} \rangle$ is the orthogonal projection of $\boldsymbol{x}$ onto $\mathbf{v}$
- Let $X \in \mathbb{R}^n$ have covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ with eigenpairs $(\sigma_i, \mathbf{v}_i)$. Given any orthonormal $\{\mathbf{u}_i\}_{i=1}^n$, define $y_i = \langle X, \mathbf{u}_i \rangle$. Then $\text{var}\left(\langle X, \mathbf{u}_i \rangle\right) = \mathbf{u}_i^T \boldsymbol{\Sigma} \mathbf{u}_i$
- SMS Lemma (applied to $\boldsymbol{\Sigma}$) implies that $\{\mathbf{v}_i\}_{i=1}^n$ are the *directions* for which $\{y_i\}_{i=1}^n$ *successively achieve maximal variances*, i.e.,

$$\begin{cases} \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle} \text{var}\left(\langle X, \mathbf{v} \rangle\right) = d_{k+1} \\ \text{argmax}_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle} \text{var}\left(\langle X, \mathbf{v} \rangle\right) = \mathbf{v}_{k+1} \end{cases}$$

- $\mathbf{v}_i$ is called the "ith principal component direction". Further, $\text{cov}\left(\langle X, \mathbf{v}_i \rangle, \langle X, \mathbf{v}_j \rangle\right) = 0$ for $i \neq j$ when $E(X) = 0$

Principal components

- Let $X \in \mathbb{R}^n$ and $\mathbf{X} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)^T$ be the *column-centered* data matrix (i.e., sample mean for entries of each column of $\mathbf{X}$ is 0)
- SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ gives $\mathbf{S} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, and $\mathbf{S} = m^{-1}\widetilde{\text{cov}}\,(X)$
- Let $\left\{\left(d_i^2, \mathbf{v}_i\right)\right\}_{i=1}^n$ be the eigenpairs of $\mathbf{S}$. Given any orthonormal $\{\tilde{\mathbf{u}}_i\}_{i=1}^n$, define $\mathbf{z}_i = \mathbf{X}\tilde{\mathbf{u}}_i$. Then

$$\mathbf{z}_i = \left(z_{1i}, ..., z_{mi}\right)^T = \left(\left\langle \boldsymbol{x}_1, \tilde{\mathbf{u}}_i \right\rangle, \cdots, \left\langle \boldsymbol{x}_m, \tilde{\mathbf{u}}_i \right\rangle\right)^T$$

  and $\widetilde{\text{cov}}\,(\mathbf{z}_i) = \tilde{\mathbf{u}}_i^T\mathbf{S}\tilde{\mathbf{u}}_i$. (What is the interpretation of $\mathbf{z}_i$?)
- SMS Lemma (applied to $\mathbf{S}$) implies that $\{\mathbf{v}_i\}_{i=1}^n$ are the *directions* for which $\{\mathbf{z}_i\}_{i=1}^n$ *successively* achieve *maximal sample variances*, i.e.,

$$\begin{cases} \max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, ..., \mathbf{v}_k \rangle} \widetilde{\text{var}}\left(\left\langle \mathbf{X}, \mathbf{v} \right\rangle\right) = m^{-1} d_{k+1}^2 \\ \text{argmax}_{\|\mathbf{v}\|=1, \mathbf{v} \perp \langle \mathbf{v}_1, ..., \mathbf{v}_k \rangle} \widetilde{\text{var}}\left(\left\langle \mathbf{X}, \mathbf{v} \right\rangle\right) = \mathbf{v}_{k+1} \end{cases}$$

- Thus, optimal $\mathbf{z}_i = d_i \mathbf{v}_i$ (why?)

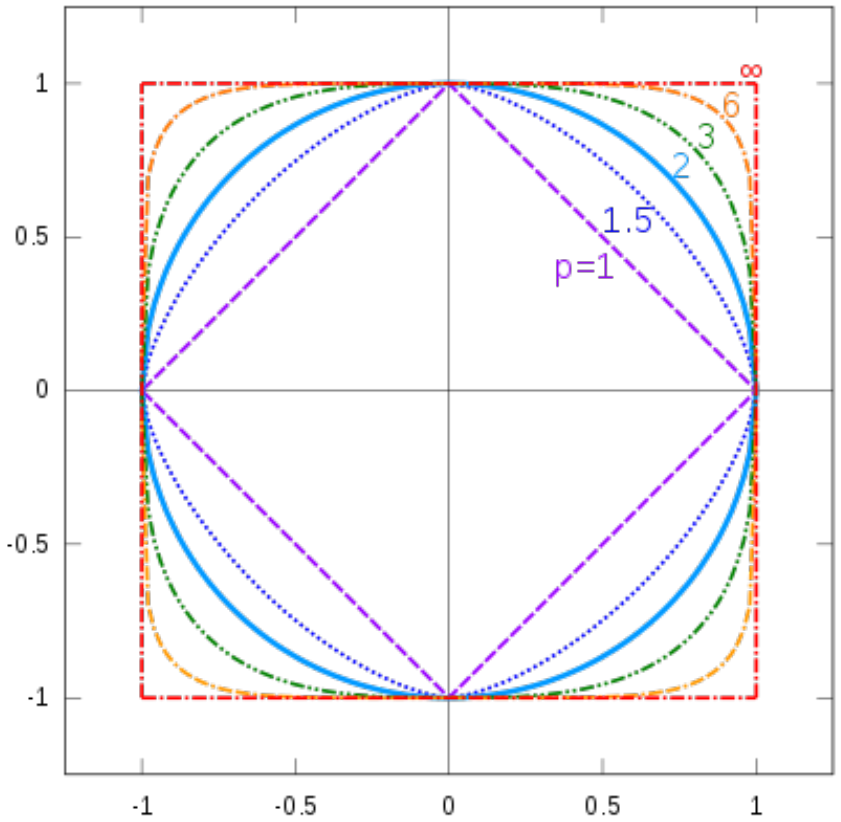
Recap on classic PCA

- Population version of optimal projections:
  - Let $X \in \mathbb{R}^n$ have covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ with eigenpairs $(\sigma_i, \mathbf{v}_i)$
  - Optimal linear combinations of entries of X are $\{y_i = \langle X, \mathbf{v}_i \rangle\}_{i=1}^n$ when $\{\mathbf{v}_i\}_{i=1}^n$ have to be orthonormal
  - $\{y_i\}_{i=1}^n$ *successively* achieve *maximal variances and are mutually uncorrelated*
- Sample version of optimal projections:
  - Let $X \in \mathbb{R}^n$ and $\mathbf{X} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)^T$ be the *column-centered* data matrix
  - Let $\left\{\left(d_i^2, \mathbf{v}_i\right)\right\}_{i=1}^n$ be the eigenpairs of $\mathbf{S} = \mathbf{X}^T\mathbf{X}$
  - Optimal linear combinations of columns of $\mathbf{X}$ are $\{\mathbf{z}_i = \mathbf{X}\mathbf{v}_i\}_{i=1}^n$ when $\{\mathbf{v}_i\}_{i=1}^n$ have to be orthonormal
  - $\{\mathbf{z}_i\}_{i=1}^n$ *successively* achieve *maximal sample variances*
- Both versions involve orthogonal projections

Solution for ridge regression

Norms and geometry

- Unit l_p-ball: $B_p = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_p \leq 1 \right\}$ (Credit: Wikipedia l_p-space)



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

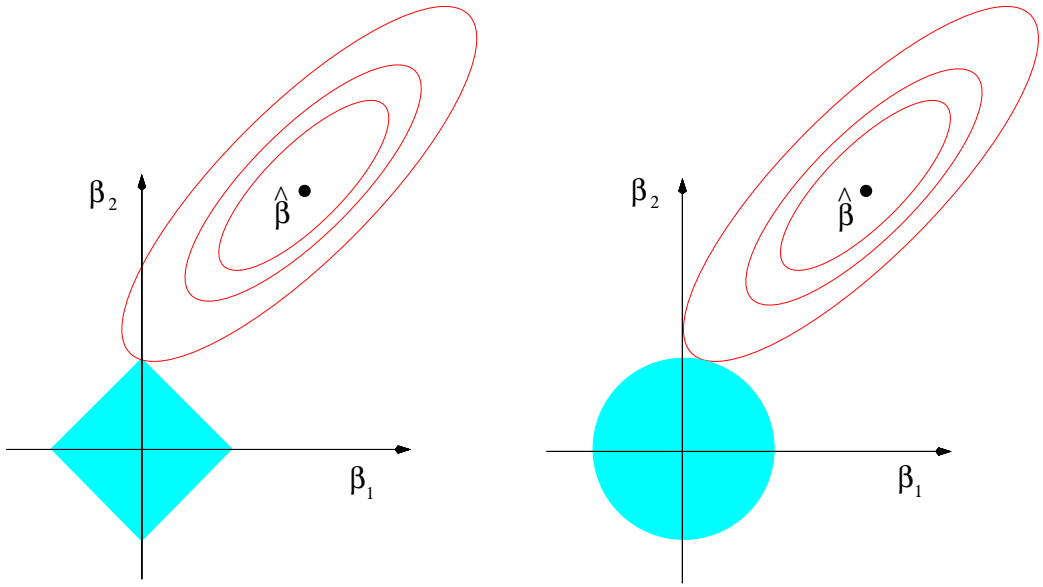


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are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

Linear model with scalar response

- Predictor $X = \left(X_1, ..., X_p\right)^T \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$
- Coefficient vector $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ and population model

$$Y = \sum_{j=1}^{p} X_j \beta_j + \varepsilon,$$

  where $\varepsilon \in \mathbb{R}$ is the random error term $E\left(\varepsilon\right) = 0$ and $\text{var}\left(\varepsilon\right) = \sigma^2$
- Data version: set $\boldsymbol{\varepsilon} = \left(\varepsilon_1, ..., \varepsilon_N\right)^T \in \mathbb{R}^N$ and each $\varepsilon_i \sim \varepsilon$, then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Ridge regression

- l_2-penalized RSS:

$$h_2(\boldsymbol{\beta};\lambda) = \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$
$$= \|\mathbf{y}\|^2 - (\mathbf{X}\boldsymbol{\beta})^T\mathbf{y} + \boldsymbol{\beta}^T\left(\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}$$

- If $\left|\mathbf{X}^T\mathbf{X}\right| \neq 0$, then unique optimizer

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda) = \left(\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \text{ for } \lambda \geq 0$$

  - If $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda) = (1+\lambda)^{-1}\mathbf{X}^T\mathbf{y} = (1+\lambda)^{-1}\hat{\boldsymbol{\beta}}$

- $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(0) = \hat{\boldsymbol{\beta}}$, i.e., LSE and unbiased when $E(\boldsymbol{\varepsilon}) = 0$
- $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\infty) = 0$, i.e., largest based but smallest "variance"
- For some $\lambda \in (0,\infty)$, there is a balance between bias and variance when $\left|\mathbf{X}^T\mathbf{X}\right| = 0$

Ridge regression: solution path

- $\mathrm{df}(\lambda) = \mathrm{trace}\left(\mathbf{X}\left(\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) = \mathrm{trace}\left(\mathbf{H}_\lambda\right) \downarrow \lambda$

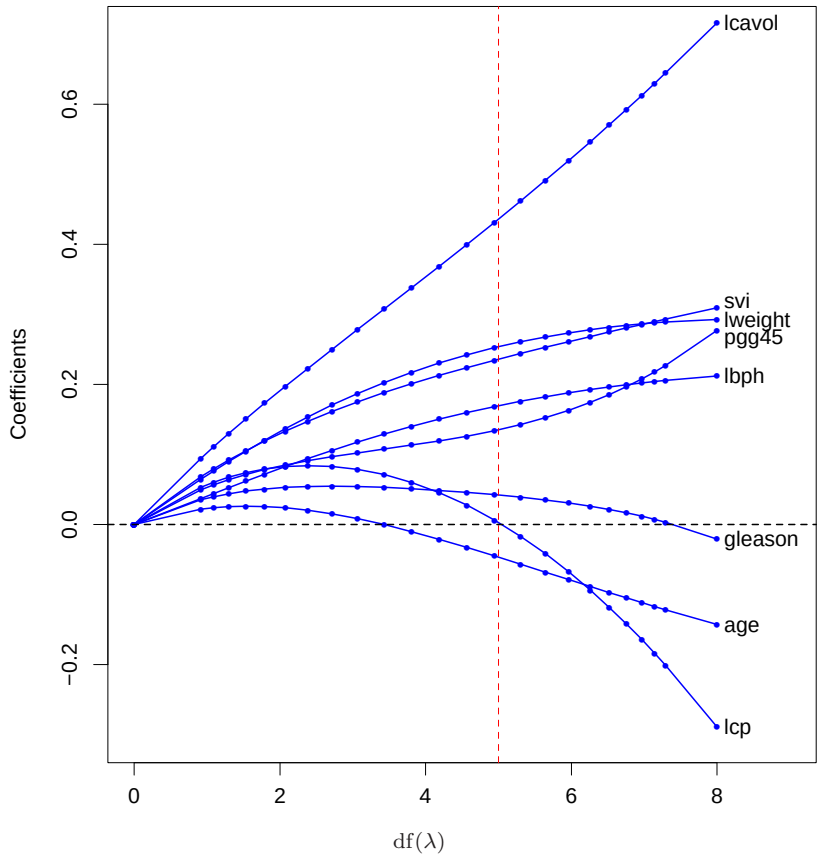
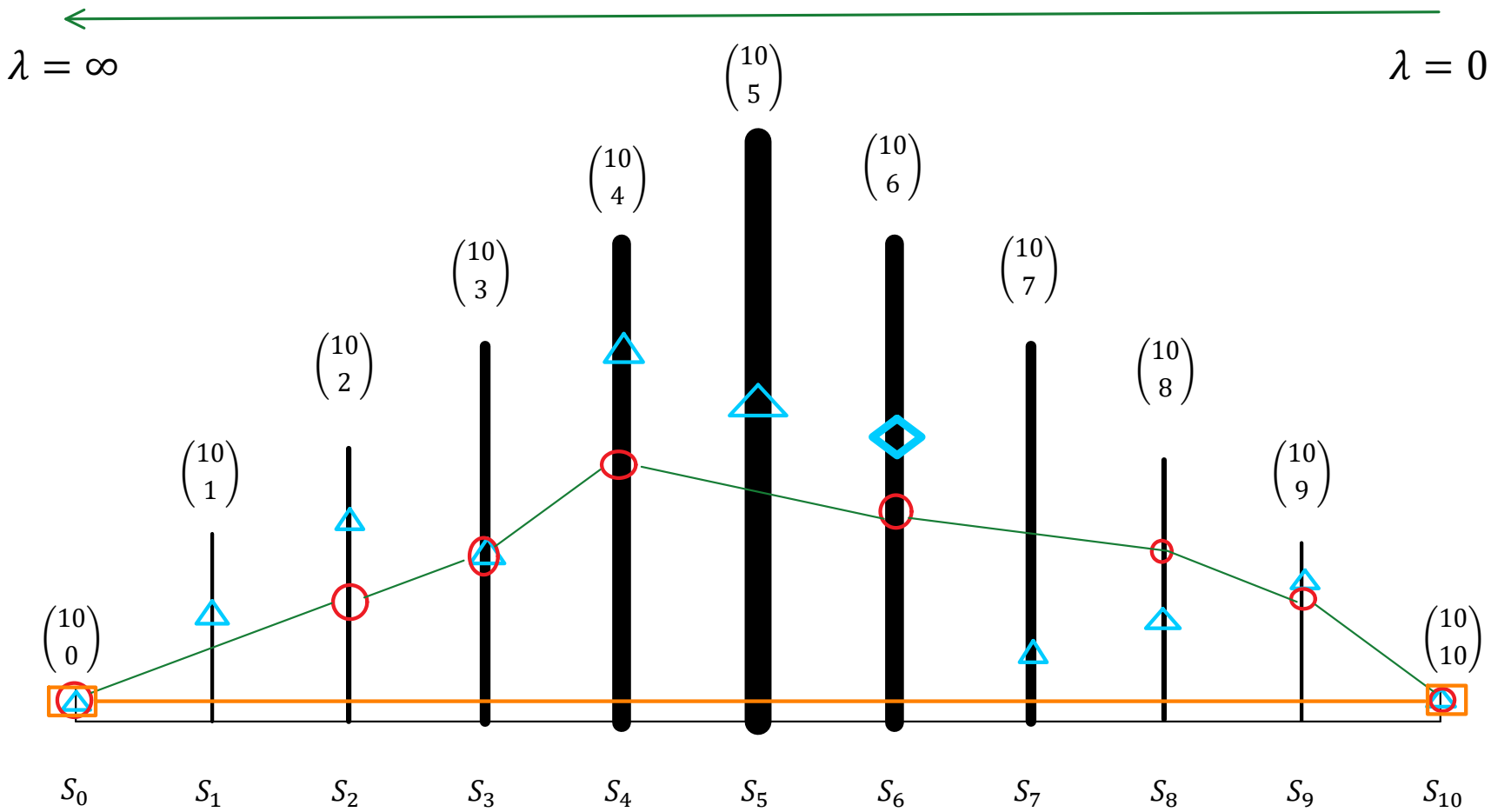


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

Trajectory of selected models

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



Adaptive shrinkage for non-orthogonal design

- Recall ridge estimate $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda) = \left(\lambda \mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ for $\lambda \geq 0$
- Recall $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \in \mathbb{R}^{n\times p}$ and $\mathbf{S} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{u}_i$ is the ith column of $\mathbf{U}$, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and

$$\mathbf{D} = \mathsf{diag}\left\{d_1, \ldots, d_n\right\} \ \& \ d_1 \geq \cdots \geq d_n \geq 0$$

- Adaptive shrinkage on LSE $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{R}}(0)$, i.e., $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda) \neq c_\lambda \hat{\boldsymbol{\beta}}$ for some $c_\lambda \neq 0$ (except when $\mathbf{X}^T\mathbf{X} = \mathbf{I}$)
- Adaptive shrinkage in prediction:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda) = \mathbf{U}\mathbf{D}\left(\lambda\mathbf{I} + \mathbf{D}^2\right)^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y},$$

i.e., $\hat{\mathbf{y}}$ is a sum of projections of $\mathbf{y}$ onto each $\mathbf{u}_j$ with shrinkage proportional to $d_j^2\left(d_j^2+\lambda\right)^{-1}$

Adaptive shrinkage for non-orthogonal design

- Recall $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and $\mathbf{S} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, where $\mathbf{u}_i$ is the ith column of $\mathbf{U}$ and

$$\mathbf{D} = \mathsf{diag}\left\{d_1, \ldots, d_n\right\} \ \& \ d_1 \geq \cdots \geq d_n \geq 0$$

- Recall $\mathbf{z}_i = \mathbf{X}\mathbf{v}_i = d_i\mathbf{u}_i$, $1 \leq i \leq n$ *successively* achieve *maximal sample variances*, i.e., $\widetilde{\mathrm{var}}\left(\mathbf{z}_i\right) = m^{-1}d_i^2$. Recall

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{R}}\left(\lambda\right) = \sum_{j=1}^{p} \mathbf{u}_j d_j^2 \left(d_j^2 + \lambda\right)^{-1} \mathbf{u}_j^T \mathbf{y}$$

- So, $\hat{\mathbf{y}}$ is a sum of shrunken $\mathbf{y}$'s in the directions $\left\{\mathbf{u}_i\right\}_{i=1}^n$ of successive maximal sample variances where a larger such variance leads to more shrinkage

Adaptive shrinkage in prediction: illustration

- Recall $\hat{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{u}_j d_j^2 \left(d_j^2 + \lambda\right)^{-1} \mathbf{u}_j^T \mathbf{y}$

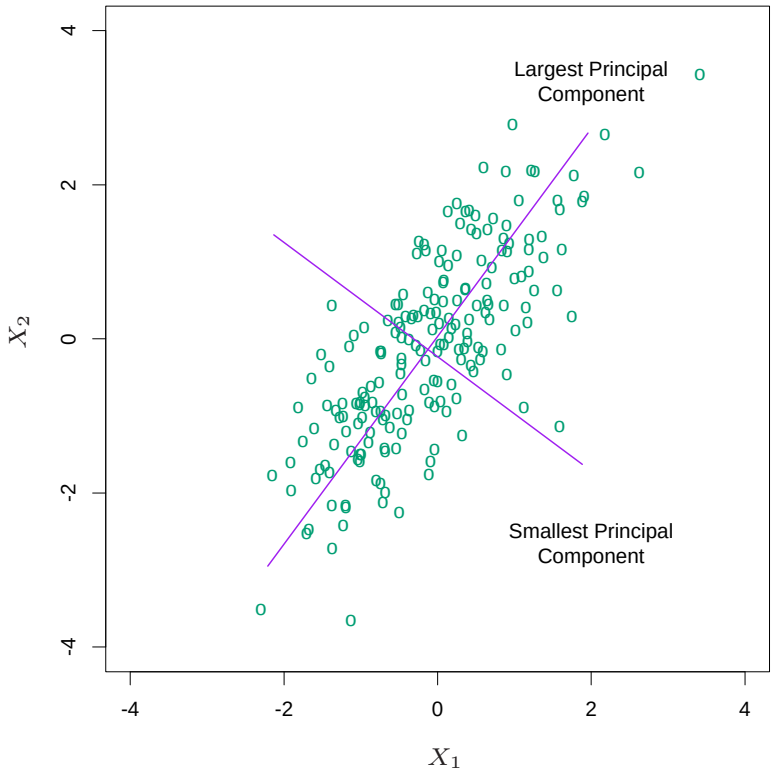


FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in mutliple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.