High-Dimensional Linear Regression

Cun-Hui Zhang

June, 2016

Peking University

Problem: High-dimensional (HD) linear regression,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad p \gg n$$

Goals

- Prediction
- Estimation of the coefficient vector
- Variable selection
- Statistical inference: confidence intervals and p-values (time permits)
- Closely related problems (time permits)

Basic assumptions

- Standardized design: $\|\boldsymbol{X}_j\|_2^2 = n$ where $\boldsymbol{X}_j$ are columns of $\boldsymbol{X}$
- Noise level σ, e.g. sufficiently large or $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$ or sub-Gaussian
- Universal threshold level: $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$
- Condition on noise: For a certain target $\boldsymbol{\beta}^*$, e.g. $\boldsymbol{\beta}$ or oracle estimator,

$$\|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty \leq \lambda_{univ}$$

- Approximate sparsity for prediction: For a certain $\overline{\boldsymbol{\beta}}$

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/(n\lambda_{univ}^2) + \|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1/\lambda_{univ} + |\mathcal{S}| \leq s^*$$

- Capped-ℓ_1 sparsity for coefficient estimation:

$$\|\boldsymbol{\beta}_{\mathcal{S}^c}\|_1/\lambda_{univ} + |\mathcal{S}| \leq s^*$$

- Hard sparsity for variable selection:

$$\|\boldsymbol{\beta}\|_0 = s \leq s^*$$

- Complexity relative to sample size: $s^* \log p \leq a_0 n$, with (small) fixed a_0
- Replacement of $\log p$ by $\log(p/s^*)$ and adjust $\|\cdot\|_\infty$ in noise condition

Minimax rates for the estimation of HD objects

- Universal penalty level: $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$
- Prediction loss: When $\log(p/s^*) \asymp \log p$,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/n \leq C_{\text{pred}} s^* \lambda_{univ}^2$$

- Estimation loss: When $\log(p/s^*) \asymp \log p$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \leq C_{\text{est,q}} s^* \lambda_{univ}^q$$

- Variable selection: Signal strength for selection/sign consistency,

$$\beta_{\min} = \min_{\beta_j \neq 0} |\beta_j| \geq C_{\text{select}} \lambda_{univ}$$

$$\Rightarrow \quad \mathbb{P}\Big\{\text{sgn}(\widehat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})\Big\} \to 1$$

- Proper conditions on $\boldsymbol{X}$ in terms of C_{pred}, $C_{\text{est,q}}$ and C_{select}

Noise inflation factor due to model uncertainty:

$$\sqrt{2\log(p/s^*)}$$

Lower bounds. We assume

- $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_{n\times n})$
- $\|\boldsymbol{X}_j\|_2^2 = n \ \forall j$

Theorem 1. *Let $\mathbb{P}_G$ correspond to a prior G under which β_j are iid random variables and $Z_j|\beta_j \sim N(\beta_j, \sigma^2/n)$. Then, for any nonnegative loss function L_0,*

$$\inf_{\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})} \mathbb{E}_G \sum_{j=1}^{p} L_0(\widehat{\beta}_j, \beta_j) \geq p \ \inf_{t(\cdot)} \mathbb{E}_G L_0(t(Z_1), \beta_1).$$

Proof: The l.h.s. equals to

$$\begin{aligned}
& \sum_{j=1}^{p} \mathbb{E}_G \Big\{ \inf_{\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})} \mathbb{E}_G \Big[L_0(\widehat{\beta}_j, \beta_j) \Big| \boldsymbol{X}, \boldsymbol{Y} \Big] \Big\} \\
\geq & \sum_{j=1}^{p} \mathbb{E}_G \Big\{ \inf_{\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})} \mathbb{E}_G \Big[L_0(\widehat{\beta}_j, \beta_j) \Big| \boldsymbol{X}, \boldsymbol{Y}, \beta_k, k \neq j \Big] \Big\} \\
= & \sum_{j=1}^{p} \mathbb{E}_G \Big\{ \inf_{t(\cdot)} \mathbb{E}_G \Big[L_0(t(Z_j), \beta_j) \Big| Z_j \Big] \Big\}
\end{aligned}$$

Lemma 1 (cf. Donoho-Johnstone, 94). *Let* $\sigma_n = \sigma/\sqrt{n}$. *Suppose* $\mathbb{P}_G\{\beta_1 \neq 0\} = \mathbb{P}_G\{\beta_1 = \mu\} = \pi_0$ *with* $\pi_0 < \epsilon_0 \leq 1 - \pi_0$, $\mu = \mu_0\sigma_n$ *and* $\mu_0 = \sqrt{2\log(1/\pi_0)} - \sqrt{2\log(1/\epsilon_0)}$. *Then,*

$$\inf_{t(\cdot)} \mathbb{E}_G|t(Z_1) - \beta_1|^q \quad \geq \quad (1 - \epsilon_0/2)(1 - \epsilon_1)\pi_0(\sigma_n\mu_0)^q$$

where $\epsilon_1 = 1 - 1/[\{\epsilon_0/(1-\pi_0)\}^{1/(q-1)} + 1]^{q-1}$ *for* $q > 1$ *and* $\epsilon_1 = 0$ *for* $q = 1$.

Sketch of proof: Let $\xi = Z_1/\sigma_n$ and $\delta_1 = I\{\beta_1 \neq 0\}$. Then, $\xi|\delta_1 \sim N(\mu_0\delta_1, 1)$ and $\delta_1 \sim$Bernoulli(π_0) under $\mathbb{E}_G$, so that

$$\inf_{t(\cdot)} \frac{\mathbb{E}_G|t(Z_1) - \beta_1|^q}{\sigma_n^q\mu_0^q} = \inf_{t(\cdot)} \mathbb{E}_G|t(\xi) - \delta_1|^q = \int \frac{\pi_0\varphi(\xi - \mu_0)}{\{\omega_1^{1/(q-1)}(\xi) + 1\}^{q-1}} d\xi$$

with $\omega_1(\xi) = \mathbb{P}_G\{\delta_1 = 1|\xi\}/\mathbb{P}_G\{\delta_1 = 0|\xi\} = \pi_0\varphi(\xi - \mu_0)/\{(1-\pi_0)\varphi(\xi)\}$. Let $a_0 = \sqrt{2\log(1/\epsilon_0)} = \sqrt{2\log(1/\pi_0)} - \mu_0$. For $\xi \leq \mu_0 + a_0$, we have $\omega_1(\xi) \leq \pi_0\varphi(a_0)/\{(1-\pi_0)\varphi(\mu_0 + a_0)\} = \epsilon_0/(1-\pi_0)$. Thus,

$$\begin{aligned} \min_{t(\cdot)} \mathbb{E}_G|t(\xi) - \delta_1|^q \quad &\geq \quad (1 - \epsilon_1)\int_{-\infty}^{\mu_0 + a_0} \pi_0\varphi(\xi - \mu_0)d\xi \\ &\geq \quad (1 - \epsilon_1)\pi_0(1 - \epsilon_0/2). \end{aligned}$$

Let $p \gg s^* \to \infty$, $\lambda_{mm} = \sigma\sqrt{(2/n)\log(p/s^*)}$ and

$$\Theta_{n,p,s^*} = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_0 \le s^*, \|\boldsymbol{v}\|_\infty \le \lambda_{mm}\}.$$

Theorem 2. Estimation lower bounds: *Let $\epsilon < 1 \le q < \infty$. Then,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{n,p,s^*}} \mathbb{E}_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \ge (1+o(1))s^*\lambda_{mm}^q,$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{n,p,s^*}} \mathbb{P}_{\boldsymbol{\beta}}\Big\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \ge (1-\epsilon)s^*\lambda_{mm}^q\Big\} \ge (\epsilon + o(1))/3^q.$$

Proof: For the $\mathbb{P}_G$ in Theorem 1 and Lemma 1 and $s^* = (1+\epsilon_0)\pi_0 p$,

$$\inf_{\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})} \mathbb{E}_G\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \ge (1+o(1))p\pi_0\mu^q \approx s^*\mu^q \approx s^*\lambda_{mm}^q.$$

Let $N = \|\boldsymbol{\beta}\|_0$ and $\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{b}} \mathbb{E}_G\big[\|\boldsymbol{\beta} - \boldsymbol{b}\|_q^q \big| \boldsymbol{X}, \boldsymbol{Y}, N \le s^*\big]$. We have

$$(1+o(1))s^*\mu^q \le \mathbb{E}_G\big[\|\boldsymbol{\delta}^* - \boldsymbol{\beta}\|_q^q \big| N \le s^*\big] + \mathbb{E}_G\|\boldsymbol{\delta}^* - \boldsymbol{\beta}\|_q^q I\{N > s^*\}.$$

As $\|\boldsymbol{\beta}\|_q^q \le s^*\mu^q$ when $N \le s^*$, $\|\boldsymbol{\delta}^*\|_q^q \le s^*\mu^q$ by the convexity of $\|\cdot\|_q^q$. Thus,

$$\mathbb{E}_G\|\boldsymbol{\delta}^* - \boldsymbol{\beta}\|_q^q I\{N > s^*\} \le 2^{q-1}\mu^q\mathbb{E}_G\big(s^* + N\big)I\{N > (1+\epsilon_0)\mathbb{E}_G N\}.$$

As $N \sim$binomial(p, π_0), the r.h.s. is $o(p\pi_0\mu^q) = o(s^*\mu^q)$. Thus,

$$(\text{minimax } \ell_q^q \text{ risk})(\Theta_{n,p,s^*}) \ge \mathbb{E}_G\big[\|\boldsymbol{\delta}^* - \boldsymbol{\beta}\|_q^q \big| N \le s^*\big] \ge (1+o(1))s^*\lambda_{mm}^q.$$

Proof of Theorem 2 (continuation): Let $c = (1-\epsilon)^{1/q}$,

$$L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = I\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \geq c(s^*)^{1/q}\lambda_{mm}\},$$
$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} I\Big\{\|\widehat{\boldsymbol{\beta}}\|_q \leq (1+c)(s^*)^{1/q}\lambda_{mm}\Big\}.$$

We have $\mu \leq \lambda_{mm}$ and $\|\boldsymbol{\beta}\|_q^q \leq s^*\mu^q$ for $N \leq s^*$. It follows that

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \leq c^q s^* \lambda_{mm}^q \{1 - L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\} + (2+c)^q s^* \lambda_{mm}^q L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}).$$

As $\|\boldsymbol{\beta}\|_q^q = N\mu^q \leq N\lambda_{mm}^q$,

$$\begin{aligned}\mathbb{E}_G\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \quad \leq \quad & c^q s^* \lambda_{mm}^q + \{(2+c)^q - c^q\} s^* \lambda_{mm}^q \max_{\boldsymbol{\beta}\in\Theta} \mathbb{E}_{\boldsymbol{\beta}} L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \\ & + 2^{q-1} s^* \lambda_{mm}^q \mathbb{E}_G\Big(N/s^* + (1+c)^q\Big) I\{N > (1+\epsilon_0)k\}.\end{aligned}$$

As $\mathbb{E}_G\big(N/s^* + (1+c)^q\big) I\{N > (1+\epsilon_0)k\} \to 0$, Lemma 1 yields

$$\begin{aligned}\{(2+c)^q - c^q\} \max_{\boldsymbol{\beta}\in\Theta} \mathbb{E}_{\boldsymbol{\beta}} L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \quad & \geq \quad \mathbb{E}_G\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q/(s^*\lambda_{mm}^q) - c^q + o(1) \\ & \geq \quad 1 - c^q + o(1).\end{aligned}$$

This gives $\max_{\boldsymbol{\beta}\in\Theta} \mathbb{E}_{\boldsymbol{\beta}} L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq (1 - c^q + o(1))/3^q$.
See Ye-Z (2010, JMLR).

Let $\overline{\mathbf{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$ and $\phi_\pm(m;\mathbf{\Sigma})$ be the upper and lower sparse eigenvalues,

$$\begin{aligned}\phi_+(m;\mathbf{\Sigma}) &= \max_{|A|=m,\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^T\mathbf{\Sigma}_{A,A}\boldsymbol{u},\\ \phi_-(m;\mathbf{\Sigma}) &= \min_{|A|=m,\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^T\mathbf{\Sigma}_{A,A}\boldsymbol{u}.\end{aligned}$$

Theorem 3. Prediction lower bounds: *Under the assumptions of Theorem 2,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{n,p,s^*}} \mathbb{E}_{\boldsymbol{\beta}}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/n \geq (1+o(1))s^*\lambda_{mm}^2\phi_-(2s^*;\overline{\mathbf{\Sigma}})/4,$$

and

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{n,p,s^*}} \mathbb{P}_{\boldsymbol{\beta}}\left\{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/n \geq \frac{(1-\epsilon)^2 s^*\lambda_{mm}^2}{4/\phi_-(2s^*;\overline{\mathbf{\Sigma}})}\right\} \geq \epsilon/4 + o(1).$$

Remark: When $\boldsymbol{X}$ has iid $N(0,\mathbf{\Sigma})$ rows,

$$\frac{\phi_-(m;\overline{\mathbf{\Sigma}})}{\phi_-(m;\mathbf{\Sigma})} \geq 1-\epsilon(a), \quad \frac{\phi_+(m;\overline{\mathbf{\Sigma}})}{\phi_+(m;\mathbf{\Sigma})} \leq 1+\epsilon(a),$$

with $\epsilon(0+) = 0$ not dependent on $\mathbf{\Sigma}$, where $a = m(\log p)/n$. Thus, conditions on $\phi_\pm(m,\overline{\mathbf{\Sigma}})$, e.g. the restricted isometry property (RIP) and sparse Reisz condition (SRC), are of ℓ_2-type.

Theorem 4. Selection lower bound: *Let $\mathscr{S}(s, \beta_*)$ be the set of all coefficient vectors $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\|_0 = s$ and $\min_{\beta_j \neq 0} |\beta_j| \geq \beta_*$, with $s < p - 1$. Then,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in \mathscr{S}(s, \beta_*)} \mathbb{P}\big\{\mathrm{supp}(\widehat{\boldsymbol{\beta}}) \neq \mathrm{supp}(\boldsymbol{\beta})\big\} \geq 1 - \frac{2\beta_*^2 + (\sigma^2/n)\log 2}{(\sigma^2/n)\log(p-s)}.$$

Proof: Let $m = p - s + 1$. For $j = 1, \ldots, m$ let $\mathbb{P}_j$ be the probability under which $\beta_j = \beta_{p-s+2} = \cdots = \beta_p = \beta_*$ and $\beta_k = 0$ for all $k \neq j$ and $k \leq p - s + 1$. Clearly $\boldsymbol{\beta} \in \mathscr{S}(s, \beta_*)$ under $\mathbb{P}_j$. Since the Kullback-Leibler information between $N(\boldsymbol{u}, \sigma^2 \boldsymbol{I}_{n\times n})$ and $N(\boldsymbol{v}, \sigma^2 \boldsymbol{I}_{n\times n})$ is $\|\boldsymbol{u} - \boldsymbol{v}\|_2^2/(2\sigma^2)$,

$$K(\mathbb{P}_j, \mathbb{P}_k) = \mathbb{E}\|\boldsymbol{X}_j\beta_* - \boldsymbol{X}_k\beta_*\|_2^2/(2\sigma^2) \leq 2n\beta_*^2/\sigma^2.$$

Let $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ and $\delta(\boldsymbol{Z}) = \min\{j : \widehat{\beta}_j \neq 0\}$. It follows from Fano's lemma that

$$\min_j \mathbb{P}_j\big\{\mathrm{supp}(\widehat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta})\big\} \leq \frac{2\beta_*^2/(\sigma^2/n) + \log 2}{\log(m-1)}.$$

See Wainwright (2007, Tech Rep) and Z (2007, Tech Rep; 2010, AOS)

Lemma 2 (Fano 61). *Let $\mathbb{P}_1, \ldots, \mathbb{P}_m$ be probability measures for data $\boldsymbol{Z}$, $\delta(\boldsymbol{z})$ a $\{1, \ldots, m\}$-valued statistic, and $K(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \log(d\mathbb{P}/d\mathbb{Q})$ the Kullback-Leibler information. Then,*

$$\frac{1}{m} \sum_{j=1}^{m} \mathbb{P}_j \Big\{ \delta(\boldsymbol{Z}) = j \Big\} \leq \frac{1}{m^2} \sum_{1 \leq j,k \leq m} \frac{K(\mathbb{P}_j, \mathbb{P}_k) + \log 2}{\log(m-1)}.$$

Penalized LSE: Global or suitable local minimizers of

$$\mathcal{L}_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \|p_\lambda(\boldsymbol{\beta})\|_1$$

where $\|p_\lambda(\boldsymbol{\beta})\|_1 = \sum_{j=1}^p p_\lambda(\beta_j)$

Basic properties of penalty functions

- Zero baseline penalty: $p_\lambda(0) = 0$
- Symmetry: $p_\lambda(-t) = p_\lambda(t)$
- Monotonicity: $p_\lambda(x) \leq p_\lambda(y)$ for all $0 \leq x < y < \infty$
- Sparsity: $p_\lambda(0+) + p'_\lambda(0+) > 0$
- Subadditivity: $p_\lambda(x+y) \leq p_\lambda(x) + p_\lambda(y)$ for all $0 \leq x \leq y < \infty$
- One-sided differentiability: $p'_\lambda(t\pm)$ are defined for all real t

Convention: $p'_\lambda(t) = x$ means

$$\min\left\{p'_\lambda(t+), p'_\lambda(t-)\right\} \leq x \leq \max\left\{p'_\lambda(t+), p'_\lambda(t-)\right\}$$

Important quantities

- Maximum concavity: $\kappa(p_\lambda) = \sup_{t,\epsilon\neq 0}\{p'_\lambda(t-\epsilon) - p'_\lambda(t)\}/\epsilon$
- Bias threshold: $a_\lambda = \inf\left\{t > 0 : p_\lambda(x) = p_\lambda(t)\ \forall\ |x| \geq t\right\}$
  – For $p = 1$, PLSE = LSE when $|\widehat{\beta}^{(lse)}| \geq a_\lambda$
- Global threshold level: $\lambda_* = \lambda_*(p_\lambda) = \inf_{t>0}\left\{t/2 + p_\lambda(t)/t\right\}$
  – For $p = 1$, $\widehat{\beta} = 0$ attains global minimum iff $|\widehat{\beta}^{(lse)}| \leq \lambda_*$

Standardization of penalty functions (scaling key quantities in λ)

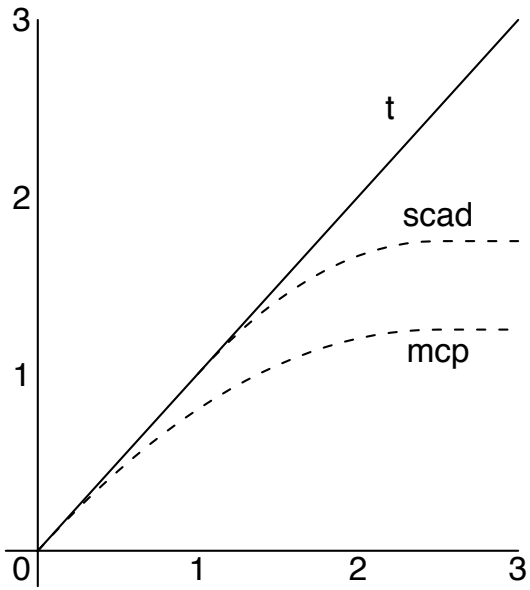
- Standardized local threshold level: $p'_\lambda(0+) = \lambda$ if $\max|p'_\lambda(0)| < \infty$
- Standardized global threshold level: $\lambda_*(p_\lambda) = \lambda$ if $\max|p'_\lambda(0)| = \infty$
- Standardized bias threshold: $a = a_\lambda/\lambda$
- Standardized local threshold level: $a^* = p'_\lambda(0+)/\lambda$
- Standardized global threshold level: $a_* = \lambda_*/\lambda$
- Standardized maximum penalty: $\gamma^* = \|p_\lambda(t)\|_\infty/\lambda^2$

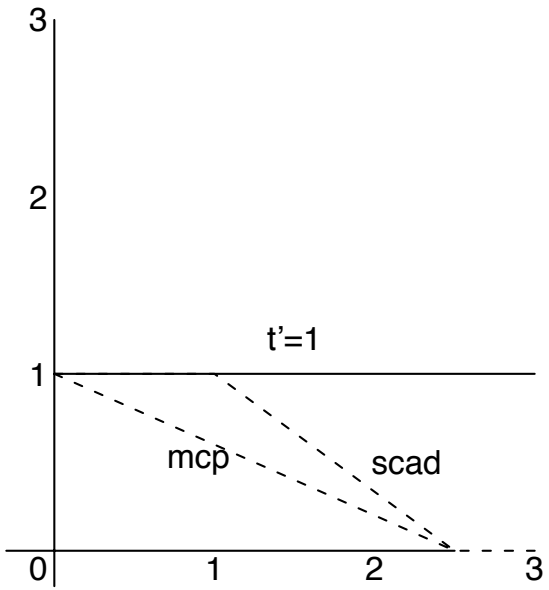
The above quantities are indeed constants when $p_\lambda(t) = \lambda^2 p_1(t/\lambda)$, as in the following examples. See Fan-Li (2001, JASA), Z (2010, AOS), and Z-Zhang (2012, StatSci) for general discussion of penalty functions

Penalty	$p_1(t)$	Bias a	Maximum Concavity $\kappa(p_\lambda)$	Thresh Levels a^*/a_*	Max Penalty γ^*
ℓ_0	$2^{-1}I\{t \neq 0\}$	0	∞	$\infty/1$	$1/2$
Bridge/ℓ_α	$C_\alpha \lvert t\rvert^\alpha$	∞	∞	$\infty/1$	∞
Lasso/ℓ_1	$\lvert t\rvert$	∞	0	$1/1$	∞
SCAD	$\int_0^{\lvert t\rvert}\left(1-\frac{(x-1)_+}{a-1}\right)_+ dx$	$a \geq 2$	$1/(a-1)$	$1/1$	$a/2+1/2$
MCP	$\int_0^{\lvert t\rvert}\left(1-x/a\right)_+ dx$	$a \geq 1$	$1/a$	$1/1$	$a/2$
Capped-ℓ_1	$\min(\lvert t\rvert, a)$	$a \geq 1/2$	∞	$1/1$	a

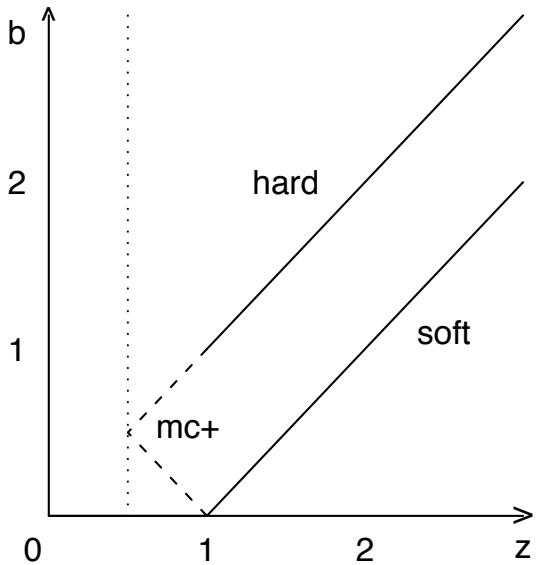
Table: Specific penalty functions: $C_\alpha = \{2(1-\alpha)\}^{1-\alpha}/(2-\alpha)^{2-\alpha}, 0 < \alpha < 1$

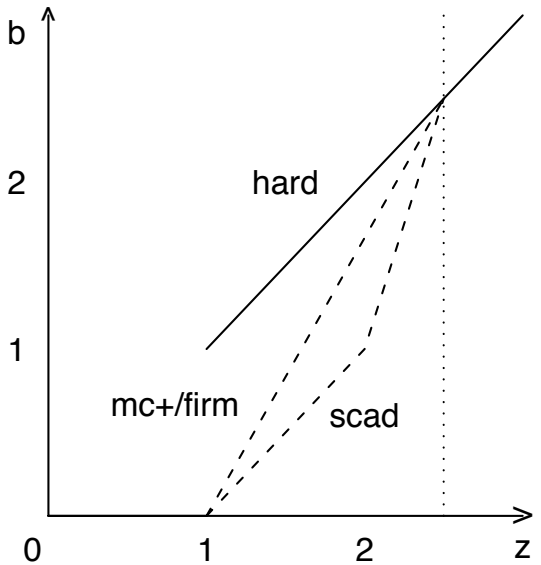
The Lasso, SCAD and MCP: $p_1(t)$ and $p_1'(t)$ for $t \geq 0$





The Lasso, SCAD and MC+ estimators for $p = 1$
(vertical line indicates the bias threshold a)





The Lasso: With the ℓ_1 penalty $p_\lambda(t) = \lambda|t|$,

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(\ell_1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1 \right\}$$

Motivation

- A convex minimization problem
  - Easier to compute than the traditional subset selection (ℓ_0 penalty)
  - Do not have to deal with (potentially bad) local minima
- Rate optimality in prediction, coefficient estimation and variable selection in the rate minimax sense (adaptive in complexity s^*)

Some drawbacks

- Biasedness
  - Does not take advantage of signal strength well
  - Requires more restrictive ℓ_∞-type conditions for selection consistency
- Performance unclear for individual $\widehat{\beta}_j$

Convex minimization

- With convex loss $\mathcal{L}_0$ and convex regularizer R_λ, let

$$\mathcal{L}_\lambda(\boldsymbol{\beta}) = \mathcal{L}_0(\boldsymbol{\beta}) + R_\lambda(\boldsymbol{\beta})$$

- **The KKT condition:** $\widehat{\boldsymbol{\beta}}$ is a minimizer of $\mathcal{L}_\lambda(\boldsymbol{\beta})$ iff

$$\partial\Big\{\mathcal{L}_0(\widehat{\boldsymbol{\beta}}) + R_\lambda(\widehat{\boldsymbol{\beta}})\Big\} \ni \mathbf{0}$$

- Sub-differential: For any convex function f,

$$\partial f(x) = \Big\{v : f(x+h) - f(x) - \langle v, h\rangle \geq 0 \ \forall h\Big\}$$

- Bregman divergence: For convex f and $\nabla f(y) \in \partial f(y)$

$$D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq 0$$

- Symmetric Bregman divergence:

$$D_f^s(x,y) = D_f(x,y) + D_f(y,x) = \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$$

Proposition 1. KKT for the Lasso: *A vector* $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ *is a Lasso solution iff*

$$g_j = \boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = p'_\lambda(\widehat{\beta}_j) \begin{cases} = \lambda \operatorname{sgn}(\widehat{\beta}_j) & \widehat{\beta}_j \neq 0 \\ \in \lambda\,[-1,1] & \forall j. \end{cases}$$

Moreover, $\widehat{\boldsymbol{\beta}}$ *is the unique solution of the Lasso iff* $\|\boldsymbol{X}\boldsymbol{u}\|_2^2 > 0$ *for all* $\boldsymbol{u} \neq 0$ *satisfying the conditions* $\operatorname{supp}(\boldsymbol{u}) \subseteq \{j : |g_j| = \lambda\}$ *and* $g_j u_j \geq 0$ *for all* j *with* $\widehat{\beta}_j = 0$, *e.g.* $\operatorname{rank}(\boldsymbol{X}_{\mathcal{S}_1}) = |\mathcal{S}_1|$ *with* $\mathcal{S}_1 = \{j : |g_j| = \lambda\}$.

Proof:

- Recall that $\mathcal{L}_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1$
- Joint differentiability of $\mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}} + \boldsymbol{u})$ in $\boldsymbol{u}$ for small $\boldsymbol{u}$ with fixed $\operatorname{sgn}(\boldsymbol{u})$
- $\partial p_\lambda(t) = \lambda \operatorname{sgn}(t)$ for $t \neq 0$
- $\partial p_\lambda(t) = \lambda[-1,1]$ for $t = 0$
- Uniqueness: Check $(\partial/\partial t)^2 \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}} + t\boldsymbol{u}) = \|\boldsymbol{X}\boldsymbol{u}\|_2^2/n$ for small $t > 0$

Selection consistency of the Lasso: Define

- Gram matrix: $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$
- Support: $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta})$
- Orthogonal projection to the column space of $\boldsymbol{X}_A$: $\boldsymbol{P}_A$
- Oracle estimator $\widehat{\boldsymbol{\beta}}^o$: $\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} = \boldsymbol{X}_{\mathcal{S}}^{\dagger}\boldsymbol{Y}$, $\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}^c} = \boldsymbol{0}$
- Neighborhood stability/exact recovery/irrepresentability coefficient (Meinshausen-Bühlmann, 06; Tropp, 06; Zhao-Yu, 06):

$$\kappa_{\mathrm{select}} = \|\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}}\,\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty$$

Theorem 5 (i) Sufficient conditions: *Suppose* $\mathrm{rank}(\boldsymbol{X}_{\mathcal{S}}) = |\mathcal{S}|$ *and*

$$\max_{j\notin\mathcal{S}} |\boldsymbol{X}_j^T\boldsymbol{P}^{\perp}_{\mathcal{S}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})/n| \le (1-\kappa_{\mathrm{select}})\lambda,$$
$$\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}}) = \mathrm{sgn}(\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} - \lambda\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}}\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}})).$$

Then, a Lasso solution is sign consistent,

$$\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}).$$

Moreover, the solution is unique if the first condition holds strictly.

Uniform signal strength/$\beta_{\min}$ condition:

$$\min_{j\in\mathcal{S}}\left(|\beta_j| - \mathrm{sgn}(\beta_j)\lambda(\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}}\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}}))_j - \lambda'(\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}})^{1/2}_{j,j}\right) > 0,$$

Theorem 5 (ii) Necessary conditions: *Suppose* $\mathrm{rank}(\boldsymbol{X}_{\mathcal{S}}) = |\mathcal{S}|$ *and* $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ *is a symmetric continuous random vector. If either* $\kappa_{\mathrm{select}} \geq 1$ *or the* $\beta_{\min}$ *condition fails to hold for* $\lambda' = 0$*, then*

$$\mathbb{P}\Big\{\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta})\Big\} \leq 1/2.$$

If $\kappa_{\mathrm{select}} < 1$ *and the* $\beta_{\min}$ *condition fails to hold for a certain* λ'*, then*

$$\begin{aligned} & \mathbb{P}\Big\{\mathrm{supp}(\widehat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta})\Big\} \\ \leq \quad & \max_{j\in\mathcal{S}} \mathbb{P}\Big\{\mathrm{sgn}(\beta_j)(\widehat{\beta}^o_j - \beta_j)/(\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}})^{1/2}_{j,j} > -\lambda'\Big\}. \end{aligned}$$

Corollary: *Suppose* $\mathrm{rank}(\boldsymbol{X}_{\mathcal{S}}) = |\mathcal{S}|$*,* $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ *is sub-Gaussian,* $\max_j \|\boldsymbol{P}^{\perp}_{\mathcal{S}}\boldsymbol{X}_j\|_2^2 \leq n$*,* $\lambda \geq (1-\kappa_{\mathrm{select}})^{-1}_{+}\sigma\sqrt{(2/n)\log((p-|\mathcal{S}|)/\epsilon)}$*, and the* $\beta_{\min}$ *condition holds with* $\lambda' = \sigma\sqrt{(2/n)\log(|\mathcal{S}|/\epsilon)}$*. Then,*

$$P\big\{\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta})\big\} \geq 1 - 3\epsilon.$$

Sufficient and nearly necessary conditions for Lasso selection consistency:

- Strong irrepresentable condition: $\kappa_{\text{select}} = \|\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty < 1$
- Penalty level: $\lambda \geq \lambda_{univ}/(1-\kappa_{\text{select}})$
- Signal strength: $\beta_{\min} \geq \lambda\|\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty + \lambda'\|\operatorname{diag}(\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1})\|_\infty^{1/2}$

The signal strength requirement is of optimal rate, but the ℓ_∞-type conditions do not scale well in $\|\boldsymbol{\beta}\|_0 = \|\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\|_2^2$

Proof of Theorem 5:

- By KKT, $\widehat{\boldsymbol{\beta}}$ is a Lasso solution with $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta})$ iff

$$\begin{aligned}\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}})/n = \lambda\operatorname{sgn}(\beta_j),\ j \in \mathcal{S},\\ |\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}})/n| \leq \lambda \text{ and } \widehat{\beta}_j = 0,\ j \notin \mathcal{S}.\end{aligned}$$

- The first part yields $\widehat{\boldsymbol{\beta}}_{\mathcal{S}} = \widehat{\boldsymbol{\beta}}_{\mathcal{S}}^o - \lambda\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})$
- The KKT (i.e. nasc) for $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta})$ can be written as

$$\begin{aligned}\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^o - \lambda\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})) = \operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}}),\\ \|\boldsymbol{X}_{\mathcal{S}^c}^T\boldsymbol{P}_{\mathcal{S}}^{\perp}\boldsymbol{\varepsilon}/n + \lambda\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty \leq \lambda.\end{aligned}$$

Prediction and coefficient estimation

- Noise condition: $\|\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty \le \eta\lambda$ with $\eta \le 1$, e.g. $\lambda \ge \lambda_{univ}/\eta$
- Let $\boldsymbol{w}_{\mathcal{S}} = \boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)/(\lambda n) - \mathrm{sgn}(\overline{\boldsymbol{\beta}}_{\mathcal{S}})$, with $\|\boldsymbol{w}_{\mathcal{S}}\|_\infty \le 1+\eta$, and

$$\psi(\boldsymbol{u}) = \boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1.$$

Lemma 3. A basic inequality: *For any $\mathcal{S}$ and $\overline{\boldsymbol{\beta}}$,*

$$\begin{aligned} & \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/n \\ \le\ & \|\boldsymbol{X}\overline{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + 4\lambda\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 + 2\lambda\psi(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}). \end{aligned}$$

Proof: Let $\boldsymbol{h}^* = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $\overline{\boldsymbol{h}} = \widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}$. We have

$$\begin{aligned} &\text{KKT: } \lambda\partial\|\widehat{\boldsymbol{\beta}}\|_1 = -\partial\mathcal{L}_0(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/n, \\ &\overline{\boldsymbol{h}}^T\partial\|\widehat{\boldsymbol{\beta}}\|_1 = \overline{\boldsymbol{h}}^T\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)/(\lambda n) - (\boldsymbol{X}\overline{\boldsymbol{h}})^T(\boldsymbol{X}\boldsymbol{h}^*)/(\lambda n), \end{aligned}$$

$$\begin{aligned} \overline{\boldsymbol{h}}^T\partial\|\widehat{\boldsymbol{\beta}}\|_1 &\ge \|\overline{\boldsymbol{h}}_{\mathcal{S}^c}\|_1 - 2\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 + \overline{\boldsymbol{h}}_{\mathcal{S}}^T\partial\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1, \\ \overline{\boldsymbol{h}}_{\mathcal{S}}^T\partial\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1 &\ge \overline{\boldsymbol{h}}_{\mathcal{S}}^T\mathrm{sgn}(\overline{\boldsymbol{\beta}}_{\mathcal{S}}) = \overline{\boldsymbol{h}}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)/(\lambda n) - \boldsymbol{w}_{\mathcal{S}}^T\overline{\boldsymbol{h}}_{\mathcal{S}}, \\ \overline{\boldsymbol{h}}^T\partial\|\widehat{\boldsymbol{\beta}}\|_1 &\ge (1-\eta)\|\overline{\boldsymbol{h}}_{\mathcal{S}^c}\|_1 - 2\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 + \overline{\boldsymbol{h}}^T\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)/(\lambda n) - \boldsymbol{w}_{\mathcal{S}}^T\overline{\boldsymbol{h}}_{\mathcal{S}}, \end{aligned}$$

$$\begin{aligned} &(\boldsymbol{X}\overline{\boldsymbol{h}})^T(\boldsymbol{X}\boldsymbol{h}^*)/n \le -2\lambda\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 + \lambda\psi(\overline{\boldsymbol{h}}), \\ &2ab = a^2 + b^2 - (a-b)^2 \end{aligned}$$

Prediction error bounds

- Let $\psi(\boldsymbol{u}) = \boldsymbol{w}_{\mathcal{S}}^T \boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1$ and define

$$\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, \overline{\Delta}) = \sup\left\{ \frac{\left[\psi(\boldsymbol{u}) + \left\{\psi^2(\boldsymbol{u}) + 2\overline{\Delta}\right\}^{1/2}\right]^2}{4(1 \vee |\mathcal{S}|)} : \boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u} = 1 \right\}$$

- Basic inequality: With $\overline{\Delta} = \|\boldsymbol{X}\overline{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/(\lambda^2 n) + 4\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1/\lambda$,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \lambda^2\overline{\Delta} + 2\lambda\psi(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) - \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/n$$

Theorem 6. *For any $\overline{\boldsymbol{\beta}}$ and $\mathcal{S}$,*

$$\begin{cases} \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/n \le \lambda^2\overline{\Delta}, & \mathcal{S} = \emptyset, \\ \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \lambda^2\overline{\Delta} + \lambda^2|\mathcal{S}|\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, 0), & \mathcal{S} \neq \emptyset, \\ \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \lambda^2(1 \vee |\mathcal{S}|)\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, \overline{\Delta}), & \overline{\boldsymbol{\beta}} = \boldsymbol{\beta}^*. \end{cases}$$

Proof: With $\overline{\boldsymbol{h}}/\lambda = t\boldsymbol{u}$ at the largest possible t and $c = 0$ or $c = \overline{\Delta}$, $(\lambda^2 c + 2\lambda\psi(\overline{\boldsymbol{h}}) - \overline{\boldsymbol{h}}^T\overline{\boldsymbol{\Sigma}}\overline{\boldsymbol{h}})/\lambda^2 = c + 2t\psi(\boldsymbol{u}) - t^2 \le (1 \vee |\mathcal{S}|)\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, c)$.

Remarks

- Approximate sparsity: For certain $\boldsymbol{\beta}^*$ satisfying the noise condition and $\overline{\boldsymbol{\beta}}$

$$\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/(n\lambda_{univ}^2) + \|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1/\lambda_{univ} + |\mathcal{S}| \leq s^*$$

- Prediction error bound of order $\lambda^2 s^*$:

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \leq \|\boldsymbol{X}\overline{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + 4\lambda\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 + \lambda^2|\mathcal{S}|\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, 0)$$

- Regularity condition on $\boldsymbol{X}$: With $\boldsymbol{w}_{\mathcal{S}} = \boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/(\lambda n) - \text{sgn}(\overline{\boldsymbol{\beta}}_{\mathcal{S}})$,

$$\overline{C}_{\text{pred}}^{(\ell_1)}(\mathcal{S}, 0) = \sup_{\boldsymbol{u}\neq 0}\left\{\frac{\left[\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1\right]_+^2}{(1 \vee |\mathcal{S}|)\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}}\right\} = O(1)$$

- Compatibility condition: As $\|\boldsymbol{w}_{\mathcal{S}}\|_\infty \leq 1 + \eta$, a sufficient condition is

$$\sup\left\{\frac{(1+\eta)^2\|\boldsymbol{u}_{\mathcal{S}}\|_1^2}{(1 \vee |\mathcal{S}|)\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}} : (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < (1+\eta)\|\boldsymbol{u}_{\mathcal{S}}\|_1\right\} = O(1)$$

- Restricted eigenvalue (RE) condition: Another sufficient condition is

$$\sup\left\{\frac{(1+\eta)^2\|\boldsymbol{u}_{\mathcal{S}}\|_2^2}{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u}} : (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < (1+\eta)\|\boldsymbol{u}_{\mathcal{S}}\|_1\right\} = O(1)$$

- All the above conditions are of ℓ_2-type

Error bounds for coefficient estimation

- Let $\psi(\boldsymbol{u}) = \boldsymbol{w}_{\mathcal{S}}^T \boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1$ and define

$$\overline{C}^{(\ell_1)}_{\mathrm{est,q}}(\mathcal{S}, \overline{\Delta}) = \sup\left\{ \frac{\|\boldsymbol{u}\|_q \left[\psi(\boldsymbol{u}) + \left\{\psi^2(\boldsymbol{u}) + 2\overline{\Delta}\right\}^{1/2}\right]}{2(1 \vee |\mathcal{S}|)^{1/(q\wedge 2)}} : \boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u} = 1 \right\}$$

- Recall the basic inequality: With $\overline{\Delta} = \|\boldsymbol{X}\overline{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/(\lambda^2 n) + 4\|\overline{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1/\lambda$,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \lambda^2 \overline{\Delta} + 2\lambda\psi(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) - \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\overline{\boldsymbol{\beta}}\|_2^2/n$$

Theorem 7. *For any $\overline{\boldsymbol{\beta}}$ and $\mathcal{S}$,*

$$\begin{cases} \|\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}\|_q \le 2\lambda|\mathcal{S}|^{1/(q\wedge 2)} \overline{C}^{(\ell_1)}_{\mathrm{est,q}}(\mathcal{S}, \overline{\Delta}/2), & \forall\ \overline{\boldsymbol{\beta}} \\ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \le \lambda|\mathcal{S}|^{1/(q\wedge 2)} \overline{C}^{(\ell_1)}_{\mathrm{est,q}}(\mathcal{S}, \overline{\Delta}), & \overline{\boldsymbol{\beta}} = \boldsymbol{\beta}^*. \end{cases}$$

Proof: With $\overline{\boldsymbol{h}}/\lambda = t\boldsymbol{u}$ at the largest t, $\|\overline{\boldsymbol{h}}\|_q = \lambda t \|\boldsymbol{u}\|_q$.

Remarks

- Approximate sparsity: For certain β^* satisfying the noise condition and $\overline{\beta}$

$$\|\boldsymbol{X}\beta^* - \boldsymbol{X}\overline{\beta}\|_2^2/(n\lambda_{univ}^2) + \|\overline{\beta}_{\mathcal{S}^c}\|_1/\lambda_{univ} + |\mathcal{S}| \leq s^*$$

  - The Lasso has prediction error $\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{X}\beta^*\|_2^2/n \lesssim s^*\lambda^2$
  - The Lasso actually estimates $\overline{\beta}$ with $\|\widehat{\beta} - \overline{\beta}\|_q^q \lesssim s^*\lambda^q$, $1 \leq q \leq 2$
  - ℓ_q error bounds for $2 < q \leq \infty$ based on the (sign-restricted) CIF
  - The error bounds match the rate of the minimax lower bounds
- The error bounds in Theorems 6 and 7 are the sharpest possible based on the basic inequality
- Many error bounds for prediction and coefficient estimation have been derived in the literature, mostly based on cruder versions of the basic inequality and under the stronger hard sparsity condition with $\overline{\Delta} = 0$
- Similar prediction and estimation error bounds can be established for the Dantzig selector

$$\widehat{\beta}^{(DS)} = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{b}\|_1 : \|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})/n\|_\infty \leq \lambda \right\}, \ \lambda \geq \lambda_{univ}$$

The false positive and sparsity of the Lasso

- Sparse condition number of the Gram matrix $\overline{\mathbf{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$,

$$\phi_{\text{cond}}(m; \mathcal{S}, \mathbf{\Sigma}) = \max_{|B\setminus\mathcal{S}|\le(1\vee m)} \left\{\phi_{\max}(\overline{\mathbf{\Sigma}}_{B,B})/\phi_{\min}(\overline{\mathbf{\Sigma}}_{B,B})\right\}$$

- Constant factor for controlling the false positive

$$C_{\text{FP}}^{(\ell_1)}(\mathcal{S}, \overline{\mathbf{\Sigma}}) = \inf\left\{t : \frac{\phi_{\text{cond}}(t|\mathcal{S}|; \mathcal{S}, \overline{\mathbf{\Sigma}}) - 1}{2(1-\eta)^2/(1+\eta)^2} + \frac{1}{|\mathcal{S}|} \le t\right\}.$$

Theorem 8. *Let* $\mathcal{S} \supseteq \text{supp}(\boldsymbol{\beta}) \cup \text{supp}(\boldsymbol{\beta}^*)$ *and* $\widehat{S} = \text{supp}(\widehat{\boldsymbol{\beta}})$. *Then,*

$$|\widehat{S} \setminus \mathcal{S}| \le |\mathcal{S}| C_{\text{FP}}^{(\ell_1)}(\mathcal{S}, \overline{\mathbf{\Sigma}}).$$

Heuristics: The KKT implies that for any $B \subseteq \widehat{S} \setminus \mathcal{S}$,

$$\begin{aligned} \lambda^2|B| &= \|\boldsymbol{X}_B(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_2^2 \\ &\le \phi_{\max}\big(\overline{\mathbf{\Sigma}}_{B,B}\big)\|\boldsymbol{P}_B(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_2^2/n. \end{aligned}$$

Done if the r.h.s. is smaller than $\lambda^2|\mathcal{S}|C_{\text{FP}}^{(\ell_1)}(\mathcal{S}, \overline{\mathbf{\Sigma}})$ when $|B| \le |\mathcal{S}|C_{\text{FP}}^{(\ell_1)}(\mathcal{S}, \overline{\mathbf{\Sigma}})$.
See Z-Huang (2008), Z (2010) and Z-Zhang (2012).

LSE after Lasso selection

- LSE in a given model $\mathcal{M} \subset \{1, \ldots, p\}$:

$$\widehat{\boldsymbol{\beta}}^{(\mathcal{M})} = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 : \mathrm{supp}(\boldsymbol{b}) \subset \mathcal{M} \right\}$$

- Relaxed lower sparse eigenvalue: With $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$,

$$\phi_*(m; \mathcal{S}, \overline{\boldsymbol{\Sigma}}) = \min_{|B \setminus \mathcal{S}| \le (1 \vee m)} \phi_{\min}(\overline{\boldsymbol{\Sigma}}_{B,B})$$

Theorem 9. *Let $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta})$ and $\widehat{S} = \mathrm{supp}(\widehat{\boldsymbol{\beta}})$ and $m_1 = \lfloor C_{\mathrm{pred}}^{(\ell_1)}(\mathcal{S})|\mathcal{S}| \rfloor$. Suppose that $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ is sub-Gaussian and $\lambda = \eta^{-1}\sigma\sqrt{(2/n)\log(p/\epsilon)}$. Then,*

$$\begin{aligned} \frac{\|\widehat{\boldsymbol{\beta}}^{(\widehat{S})} - \boldsymbol{\beta}\|_2^2}{\phi_*\left(m_1; \mathcal{S}, \overline{\boldsymbol{\Sigma}}\right)} &\le \left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(\widehat{S})} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2/n \\ &\le |\mathcal{S}|\left\{ C_{\mathrm{pred}}^{(\ell_1)}(\mathcal{S})(\lambda^2 + 7.5\sigma^2/n) + 2.5 C_{\mathrm{FP}}^{(\ell_1)}(\mathcal{S}, \overline{\boldsymbol{\Sigma}})(\lambda_1^2 + 3\sigma^2/n) \right\} \end{aligned}$$

with at least probability $1 - 2\epsilon - \epsilon^{m_1}$, where $\lambda_1 = \sigma\sqrt{(2/n)\log(ep/(\epsilon m_1))}$

Proof: $|\widehat{S} \setminus \mathcal{S}| \le m_1$ and $\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(\widehat{S})} - \boldsymbol{X}\boldsymbol{\beta}^*\right\|_2^2/n \le \left\|\boldsymbol{P}_{\widehat{S}}^{\perp}\boldsymbol{X}\boldsymbol{\beta}^*\right\|_2^2/n + \|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\|_2^2/n$.

Some references:

- The Lasso: Tinbshirani (96), Chen-Donoho-Saunders (01)
- Prediction/persistency: Greenshtein-Ritov (04)
- Dantzig selector/restricted isometry condition: Candes-Tao (05,07)
- Lasso/sparse Riesz condition: Z-Huang (08), Zhang (09)
- RE condition: Bickel-Ritov-Tsybakov (09), Koltchinskii (09)
- Compatibility condition: van de Geer (07), van de Geer-Bühlmann (09)
- Cone invertibility factors (CIF): Ye-Z (10)
- Capped-ℓ_1 sparsity: Z-Huang (08), Z-Zhang (12)
- Approximate sparsity: Sun-Z (12)
- False positive/Post Lasso LSE: Sun-Z (12)

Restricted eigenvalue: Let $\xi = (1+\eta)/(1-\eta)$, $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$ and

$$\mathrm{RE}^2(\mathcal{S}, \xi; \overline{\boldsymbol{\Sigma}}) = \inf\left\{\boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u} : \|\boldsymbol{u}_{\mathcal{S}}\|_2 = 1, \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < \xi\|\boldsymbol{u}_{\mathcal{S}}\|_1\right\}.$$

Lemma 4. *For integer $m > 0$ and $\boldsymbol{u} \in \mathbb{R}^p$, there exists $\boldsymbol{v} \in \mathbb{R}^p$ such that*

$$\boldsymbol{v}_{\mathcal{S}} = \boldsymbol{u}_{\mathcal{S}},\ \mathrm{supp}(\boldsymbol{v}) \subseteq \mathrm{supp}(\boldsymbol{u}),\ \|\boldsymbol{v}_{\mathcal{S}^c}\|_0 \le m,\ \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 = \|\boldsymbol{u}_{\mathcal{S}^c}\|_1,$$

and

$$\boldsymbol{v}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{v} - \boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u} \le \|\boldsymbol{u}_{\mathcal{S}^c}\|_1^2/m \le |\mathcal{S}|\xi^2\|\boldsymbol{u}_{\mathcal{S}}\|_2^2/m.$$

Proof: Assume $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 > 0$. Let $\pi_j = |u_j|/\|\boldsymbol{u}_{\mathcal{S}^c}\|_1$ and $\boldsymbol{Z}^i$ be a iid vectors with

$$\mathbb{P}\left\{\boldsymbol{Z}^i = \boldsymbol{u}_{\mathcal{S}} + \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \mathrm{sgn}(u_j)\boldsymbol{e}_j\right\} = \pi_j,\ j \notin \mathcal{S}.$$

Let $\boldsymbol{V} = \sum_{i=1}^m \boldsymbol{Z}^i/m$. We have $\mathbb{E}\boldsymbol{Z} = \boldsymbol{u}$ and

$$\mathbb{E}\boldsymbol{V}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{V} - \boldsymbol{u}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{u} = \frac{1}{m}\mathbb{E}(\boldsymbol{Z}-\boldsymbol{u})_{\mathcal{S}^c}^T\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}^c}(\boldsymbol{Z}-\boldsymbol{u})_{\mathcal{S}^c} \le \frac{\|\boldsymbol{u}_{\mathcal{S}^c}\|_1^2}{m}\|\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}^c}\|_{\max}.$$

Thus, $\boldsymbol{v}$ can be one of the realizations of $\boldsymbol{V}$.

Restricted eigenvalue:

$$\mathrm{RE}^2(\mathcal{S}, \xi; \overline{\boldsymbol{\Sigma}}) = \inf \left\{ \boldsymbol{u}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{u} : \|\boldsymbol{u}_{\mathcal{S}}\|_2 = 1, \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 < \xi \|\boldsymbol{u}_{\mathcal{S}}\|_1 \right\}.$$

A lower bound of the restricted eigenvalue is

$$\mathrm{RE}^2(\mathcal{S}, \xi; \overline{\boldsymbol{\Sigma}}) \geq \inf \left\{ \boldsymbol{v}^T \overline{\boldsymbol{\Sigma}} \boldsymbol{v} : \|\boldsymbol{v}_{\mathcal{S}}\|_2 = 1, \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 < \xi \|\boldsymbol{v}_{\mathcal{S}}\|_1, \|\boldsymbol{v}_{\mathcal{S}^c}\|_0 \leq m \right\} - \xi^2 |\mathcal{S}|/m.$$

If $\boldsymbol{X}$ has iid $N(0, \boldsymbol{\Sigma})$ (or sub-Gaussian) rows, then

$$\mathrm{RE}^2(\mathcal{S}, \xi; \overline{\boldsymbol{\Sigma}}) \geq \left\{ 1 - \frac{C_0}{n} \binom{p}{|\mathcal{S}| + m} \right\} \mathrm{RE}^2(\mathcal{S}, \xi; \boldsymbol{\Sigma}) - \frac{\xi^2 |\mathcal{S}|}{m}.$$

This strengthens the results in Rudelson-Zhou (2013)

Concave penalties: Global or suitable local minimizers of

$$\mathcal{L}_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \|p_\lambda(\boldsymbol{\beta})\|_1$$

where $\|p_\lambda(\boldsymbol{\beta})\|_1 = \sum_{j=1}^p p_\lambda(\beta_j)$

- Critical point:
  $\liminf_{t\to 0_+} \inf_{0<\|\boldsymbol{u}\|_2\le 1} t^{-1}\left\{\mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}} + t\boldsymbol{u}) - \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}})\right\} \ge 0$
- Local minimizer: There exists $t_0 > 0$ such that
  $\mathcal{L}_\lambda(\boldsymbol{b}) \ge \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}}), \quad \forall\, 0 < \|\boldsymbol{b} - \widehat{\boldsymbol{\beta}}\|_2 \le t_0$
- Strict local minimizer: There exists $t_0 > 0$ such that
  $\mathcal{L}_\lambda(\boldsymbol{b}) > \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}}), \quad \forall\, 0 < \|\boldsymbol{b} - \widehat{\boldsymbol{\beta}}\|_2 \le t_0$
- Global minimizer:
  $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \mathcal{L}_\lambda(\boldsymbol{b})$

Theorem 10. KKT-type conditions for local solutions: Suppose that $p_\lambda(t)$ is twice left- and right-differentiable for all real t.
(i) A vector $\widehat{\boldsymbol{\beta}}$ is a critical point iff

$$\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = p'_\lambda(\widehat{\beta}_j), \quad \forall\, j = 1, \ldots, p.$$

In particular, $|\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n| \le p'_\lambda(0+)$ for all j with $\widehat{\beta}_j = 0$.
(ii) A critical point $\widehat{\boldsymbol{\beta}}$ is a strict local minimizer if

$$\|\boldsymbol{X}\boldsymbol{u}\|_2^2/n + \sum_{u_j>0} p''_\lambda(\widehat{\beta}_j+)u_j^2 + \sum_{u_j<0} p''_\lambda(\widehat{\beta}_j-)u_j^2 > 0$$

for all $\boldsymbol{u}$ with $\{j : u_j > 0\} \subseteq \mathcal{S}_+$ and $\{j : u_j < 0\} \subseteq \mathcal{S}_-$, where
$\mathcal{S}_\pm = \{j : \boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = p'_\lambda(\widehat{\beta}_j\pm)\}$.
(iii) Suppose $p_\lambda(t)$ is a quadratic spline in $[0, \infty)$. Then, a critical point $\widehat{\boldsymbol{\beta}}$ is a strict local minimizer iff the condition in (ii) holds, and it is a local minimizer iff

$$\|\boldsymbol{X}\boldsymbol{u}\|_2^2/n + \sum_{u_j>0} p''_\lambda(\widehat{\beta}_j+)u_j^2 + \sum_{u_j<0} p''_\lambda(\widehat{\beta}_j-)u_j^2 \ge 0$$

for the same set of $\boldsymbol{u}$.

Oracle local solutions: Let $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta})$ and

$$\theta_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta}) = \inf\left\{\theta : \frac{\|\boldsymbol{v}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}\|_\infty}{\theta\lambda + \lambda'} \le 1 \;\Rightarrow\; \|\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}^{-1} p'_\lambda(\boldsymbol{v}_{\mathcal{S}})\|_\infty \le \theta\lambda\right\},$$

$$\kappa_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta}) = \sup\left\{\left\|\overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}^{-1}\left(\frac{p'_\lambda(\boldsymbol{v}_{\mathcal{S}})}{\lambda}\right)\right\|_\infty : \frac{\|\boldsymbol{v}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}\|_\infty}{\theta_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta})\lambda + \lambda'} \le 1\right\}.$$

Theorem 11. *Let $\widehat{\boldsymbol{\beta}}^o$ be the oracle LSE. Suppose $p'_\lambda(t)$ is continuous in $t > 0$, $\mathrm{rank}(\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}) = |\mathcal{S}|$, $\min_{j\in\mathcal{S}} |\beta_j| \ge \theta_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta})\lambda + \lambda'$, and*

$$\|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}\|_\infty \le \lambda', \quad \|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n\|_\infty \le \lambda(1 - \kappa_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta}))_+.$$

Then, there exists a local solution $\widehat{\boldsymbol{\beta}}$ such that

$$\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}), \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty \le \theta_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta})\lambda + \lambda'.$$

Moreover, if in addition $\theta_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta}) = 0$, then $\kappa_{\mathrm{select}}(p_\lambda, \boldsymbol{\beta}) = 0$ and

$$\widehat{\boldsymbol{\beta}}^o \text{ is a local minimizer.}$$

Remarks:

- Folded concave penalties provide oracle local solutions under much weaker condition on $\boldsymbol{X}$ than the Lasso, e.g. when $\widehat{\boldsymbol{\beta}}^o$ is an oracle solution under the signal strength condition $\theta_{\text{select}}(p_\lambda, \boldsymbol{\beta}) = 0$
- Due to the multiplicity of solutions, the challenges of concave PLSE are
  – to find such oracle solutions if exists
  – to cause little harm or even some gain in performance for prediction and coefficient estimation, e.g. in view of the rate minimaxity of the Lasso, when such oracle solution does not exist
  – to compute such solutions with reasonable cost

Outline of the proof of Theorem 11: Apply Brouwer's fixed-point theorem to

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = p'_\lambda(\widehat{\boldsymbol{\beta}}),\ \text{supp}(\widehat{\boldsymbol{\beta}}) = \mathcal{S},$$

on $\mathcal{S}$ to find $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$ and then check the condition on $\mathcal{S}^c$ using $\kappa_{\text{select}}(p_\lambda, \boldsymbol{\beta})$.
See Z-Zhang (2012).

Proof of Theorem 11:

- Recall that $\|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}\|_\infty \le \lambda'$.
- Let $B = \{\boldsymbol{h}_{\mathcal{S}} : \|\boldsymbol{h}_{\mathcal{S}}\|_\infty \le \theta_{\text{select}}(p_\lambda, \boldsymbol{\beta})\lambda\}$.
- For $\boldsymbol{h}_{\mathcal{S}} \in B$, $\boldsymbol{v}_{\mathcal{S}} = \widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} + \boldsymbol{h}_{\mathcal{S}}$ satisfies $\|\boldsymbol{v}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}\|_\infty \le \theta_{\text{select}}(p_\lambda, \boldsymbol{\beta})\lambda + \lambda'$.
- It follows that $\|\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}} p'_\lambda(\boldsymbol{v}_{\mathcal{S}})\|_\infty \le \theta_{\text{select}}(p_\lambda, \boldsymbol{\beta})\lambda$.
- Thus, $\boldsymbol{h}_{\mathcal{S}} = -\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}} p'_\lambda(\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} + \boldsymbol{h}_{\mathcal{S}})$ has a solution (Brouwer).
- Let $\widehat{\boldsymbol{\beta}}_{\mathcal{S}} = \widehat{\boldsymbol{\beta}}^o_{\mathcal{S}} + \boldsymbol{h}_{\mathcal{S}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} = \boldsymbol{0}$. We have

$$\boldsymbol{X}^T_{\mathcal{S}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = \boldsymbol{X}^T_{\mathcal{S}}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^o - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = -\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}\boldsymbol{h}_{\mathcal{S}} = p'_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}).$$

- Moreover, by the definition of $\kappa_{\text{select}}(p_\lambda, \boldsymbol{\beta})$,

$$\begin{aligned} & \left\|\boldsymbol{X}^T_{\mathcal{S}^c}(\boldsymbol{Y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}})/n\right\|_\infty \\ = & \left\|\boldsymbol{X}^T_{\mathcal{S}^c}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n + \overline{\boldsymbol{\Sigma}}_{\mathcal{S}^c,\mathcal{S}}\overline{\boldsymbol{\Sigma}}^{-1}_{\mathcal{S},\mathcal{S}} p'_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{S}})\right\|_\infty \\ \le & \left\|\boldsymbol{X}^T_{\mathcal{S}^c}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n\right\|_\infty + \kappa_{\text{select}}(p_\lambda, \boldsymbol{\beta})\lambda \\ \le & \lambda. \end{aligned}$$

Local solution path: Let $\kappa(p_\lambda) = \sup_{t_1<t_2}\{p'_\lambda(t_1) - p'_\lambda(t_2)\}/(t_2 - t_1)$,

- $\mathscr{P}(\lambda_0, \kappa_0) = \left\{p_\lambda(\cdot) : p'_\lambda(0+) = \lambda \geq \lambda_0,\ \|p'_\lambda(\cdot)\|_\infty \leq \lambda,\ \kappa(p_\lambda) \leq \kappa_0\right\}$
- $\mathscr{B}(\lambda_0, \kappa_0) =$ the set of all local solutions for some $p_\lambda \in \mathscr{P}(\lambda_0, \kappa_0)$
- $\mathscr{B}_0(\lambda_0, \kappa_0) =$ the set of all vectors connected to $\mathbf{0}$ in $\mathscr{B}(\lambda_0, \kappa_0)$
- $\boldsymbol{\beta}^*$ be a target vector satisfying $\|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty < \eta\lambda_0$ with $\eta < 1$
- $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$
- $\mathrm{RE} = \inf_{\boldsymbol{u}\neq 0}\left\{\|\boldsymbol{X}\boldsymbol{u}\|_2/(n^{1/2}\|\boldsymbol{u}\|_2) : \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \|\boldsymbol{u}_{\mathcal{S}}\|_1 + \eta\|\boldsymbol{u}\|_1\right\}$

Theorem 12. (Feng-Z, 2015). *Let $\widehat{\boldsymbol{\beta}}$ be a local minimizer in $\mathscr{B}_0(\lambda_0, \kappa_0)$ with a penalty $p_\lambda \in \mathscr{P}(\lambda_0, \kappa_0)$. Suppose $\mathrm{RE}^2 \geq \kappa_0$. Then, for all seminorms $\|\cdot\|$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \leq \lambda \sup\left\{\frac{(1+\eta)\|\boldsymbol{u}\|}{\|\overline{\boldsymbol{\Sigma}}\boldsymbol{u}\|_\infty} : (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}}\right\}$$

where $\boldsymbol{w}_{\mathcal{S}}^T = \{\boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n - p'_\lambda(\boldsymbol{\beta}^*_{\mathcal{S}})\}/\lambda$. In particular,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \leq \frac{4\lambda^2|S|(1+\eta)^2}{(1-\eta)^2\mathrm{RE}^2}, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{2\lambda|S|(1+\eta)}{(1-\eta)\mathrm{RE}^2},$$

and when $\phi_{\min}(\overline{\boldsymbol{\Sigma}}_{\mathcal{S},\mathcal{S}}) > \kappa_0$ and $\boldsymbol{\beta}^*$ is an oracle solution for p_λ,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*.$$

Proof: Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. For some $\widetilde{\eta} < \eta$, we have

$$\begin{aligned}
h_j \boldsymbol{X}_j^T(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Y})/n &= -h_j p_\lambda'(\widehat{\beta}_j), && \text{“KKT” for } \widehat{\boldsymbol{\beta}} \\
-h_j p_\lambda'(\widehat{\beta}_j) &\le \kappa_0 h_j^2 - h_j p_\lambda'(\beta_j^*), && \text{bound on concavity} \\
\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n &\le \boldsymbol{h}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n + \kappa_0\|\boldsymbol{h}\|_2^2 - \boldsymbol{h}^T p_\lambda'(\boldsymbol{\beta}^*), && \text{summing over } j \\
-\boldsymbol{h}^T p_\lambda'(\boldsymbol{\beta}^*) &= -\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 - \boldsymbol{h}_{\mathcal{S}}^T p_\lambda'(\boldsymbol{\beta}_{\mathcal{S}}^*), && \text{favorable choice of } p_\lambda' \\
\boldsymbol{h}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n &\le \widetilde{\eta}\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 + \boldsymbol{h}_{\mathcal{S}}\boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n, && \text{condition on noise}
\end{aligned}$$

Combining the above three inequalities, we have a basic inequality

$$\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n + (1-\widetilde{\eta})\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \kappa_0\|\boldsymbol{h}\|_2^2 + \lambda\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{h}_{\mathcal{S}}.$$

Key argument: As $\|\boldsymbol{w}_{\mathcal{S}}\|_\infty \le 1 + \widetilde{\eta}$, the above and the RE condition yield

$$\begin{aligned}
& \kappa_0\|\boldsymbol{h}\|_2^2 \le (\eta - \widetilde{\eta})\|\boldsymbol{h}\|_1 \\
\Rightarrow\quad & \|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \|\boldsymbol{h}_{\mathcal{S}}\|_1 + \eta\|\boldsymbol{h}\|_1 \\
\Rightarrow\quad & (1-\widetilde{\eta})\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \lambda\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{h}_{\mathcal{S}} \\
\Rightarrow\quad & \|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \|\boldsymbol{h}_{\mathcal{S}}\|_1 + \widetilde{\eta}\|\boldsymbol{h}\|_1.
\end{aligned}$$

Thus, $\mathscr{B}_0(\lambda_0, \kappa_0) - \boldsymbol{\beta}^* \subseteq \mathscr{C}(\mathcal{S}, \eta) = \{\boldsymbol{u} : (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \le \boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}}, \boldsymbol{w}_{\mathcal{S}} = \boldsymbol{w}_{\mathcal{S}}^{(p_\lambda)}\}$.
As $\|\boldsymbol{h}\| \le \|\overline{\boldsymbol{\Sigma}}\boldsymbol{h}\|_\infty \sup_{\boldsymbol{u}\in\mathscr{C}(\mathcal{S},\eta)} \|\boldsymbol{u}\|/\|\boldsymbol{\Sigma}\boldsymbol{u}\|_\infty$, the error bound follows from

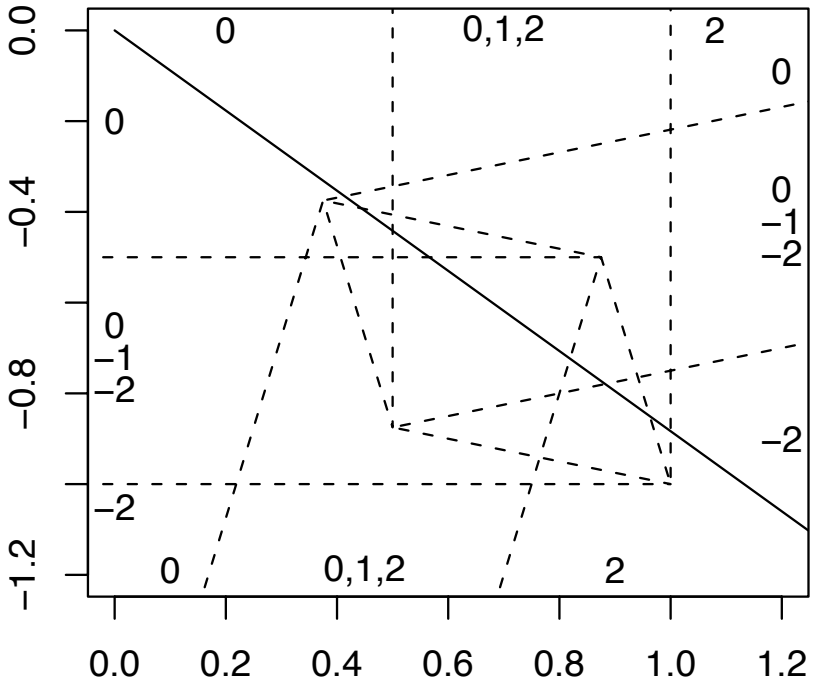
$$\|\overline{\boldsymbol{\Sigma}}\boldsymbol{h}\|_\infty \le \|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty + \|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty \le \lambda + \eta\lambda.$$

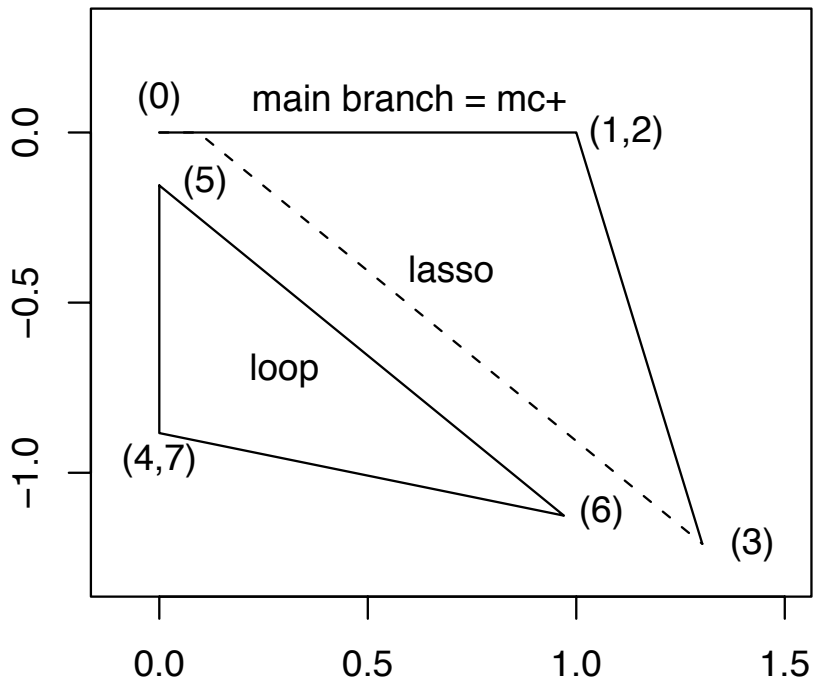
Remarks:

- $\mathscr{B}_0(\lambda_0, \kappa_0)$ is the set of all local solutions computable by path following algorithms stating from the origin, with the constraints $\lambda \geq \lambda_0$ and $\kappa(p_\lambda) \leq \kappa_0$ on the penalty and concavity levels
- The RE condition, which is of the ℓ_2 type, is arguably nearly the weakest proven one on $\boldsymbol{X}$ for the rate optimal prediction and coefficient estimation error bounds for the Lasso, e.g. in models where the sharper bounds in Theorems 6 and 7 are not much sharper
- Under the RE condition, any path following solution achieves the same or sharper prediction and coefficient estimation error rates as the Lasso
- When a large majority of the components of $|\widehat{\boldsymbol{\beta}}^o_{\mathcal{S}}|$ are large, $\boldsymbol{w}_{\mathcal{S}}$ can be small for $\widehat{\boldsymbol{\beta}}^o$. In this case, concave penalty outperforms the Lasso in prediction and coefficient estimation
- Under the same RE condition, the path following solution achieves variable selection consistency, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, completely removing the need for the restrictive ℓ_∞-type conditions such as the irrepresentable condition required for the selection consistency of the Lasso

A path following algorithm:

- $p_\lambda(t) = \lambda^2 p_1(|t|/\lambda)$ with a quadratic spline $p_1(t)$ in $[0, \infty)$
- $\widehat{\boldsymbol{\beta}}(\lambda)$: minimizing $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda^2 \|p_1(\boldsymbol{\beta}/\lambda)\|_1$
- Rescale: $\tau = 1/\lambda$, $\boldsymbol{b} = \boldsymbol{b}(\tau) = \widehat{\boldsymbol{\beta}}(\lambda)/\lambda$, $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{y}/n$
- $\boldsymbol{b}$: minimizing $\boldsymbol{b}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{b}/2 - \tau \boldsymbol{Z}^T\boldsymbol{b} + \|p_1(\boldsymbol{b})\|_1$
- "KKT": $\overline{\boldsymbol{\Sigma}}\boldsymbol{b}(\tau) - \tau\boldsymbol{Z} + p_1'(\boldsymbol{b}(\tau)) = 0$, linear spline path
- Direction of path: $\overline{\boldsymbol{\Sigma}}\boldsymbol{b}'(\tau) - \boldsymbol{Z} + p_1''(\boldsymbol{b}(\tau)) \circ \boldsymbol{b}'(\tau) = 0$
- **Initialization:** $\tau^{(0)} = 0$, $\widehat{S}^{(0)} = \emptyset$, $\boldsymbol{b}(\tau) = 0$, $k = 1$
- **Iteration:** – Find $\tau^{(k)}$ as the beginning point for new $\widehat{S}^{(k)}$ or new $p_\lambda''(\boldsymbol{b})$
  – Find the new $\widehat{S}^{(k)}$ & $p_\lambda''(\boldsymbol{b})$
  – Compute $\partial \boldsymbol{b}_{\widehat{S}}^{(k)} = \left(\overline{\boldsymbol{\Sigma}} + \mathrm{diag}(p_1''(\boldsymbol{b}))\right)^{-1}_{\widehat{S},\widehat{S}} \boldsymbol{Z}_{\widehat{S}}$ with new $\widehat{S}^{(k)}$ & $p_\lambda''(\boldsymbol{b})$
  – $\boldsymbol{b}(\tau) = \boldsymbol{b}(\tau^{(k)}) + (\tau - \tau^{(k)})\partial \boldsymbol{b}^{(k)}$ from $\tau = \tau^{(k)}$ to $\tau = \tau^{(k+1)}$
  – $k = k + 1$
- References
  – Osborne et al (2000), Efron et al (2004), LARS for the Lasso
  – Z (2010), PLUS for the Lasso, SCAD, MCP and ...


0.0
−0.4
−0.8
−1.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
0,1,2
2
0
0
0
−1
−2
0
−1
−2
−2
0
0,1,2
2


(0)
main branch = mc+
(1,2)
(5)
lasso
loop
(4,7)
(6)
(3)
0.0
−0.5
−1.0
0.0
0.5
1.0
1.5

Majorize-Minimization (MM) algorithms:

- $\mathcal{L}_\lambda(\boldsymbol{\beta}) = \mathcal{L}_0(\boldsymbol{\beta}) + R_\lambda(\boldsymbol{\beta})$, $\widehat{\boldsymbol{\beta}}^{(old)}$
- $f(\boldsymbol{\beta}) \geq \mathcal{L}_\lambda(\boldsymbol{\beta})$, $f(\widehat{\boldsymbol{\beta}}^{(old)}) = \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}}^{(old)})$
- $\widehat{\boldsymbol{\beta}}^{(new)} = \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$
- $\mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}}^{(new)}) \leq f(\widehat{\boldsymbol{\beta}}^{(new)}) \leq f(\widehat{\boldsymbol{\beta}}^{(old)}) = \mathcal{L}_\lambda(\widehat{\boldsymbol{\beta}}^{(old)})$

- Majorizing the penalty
  - Local quadratic approximation (LQA; Fan-Li, 2001):

$$p_\lambda(|\beta_j|) \leq p_\lambda(|\widehat{\beta}_j^{(old)}|) + p'_\lambda(|\widehat{\beta}_j^{(old)}|)(\beta_j^2 - |\widehat{\beta}_j^{(old)}|^2)/(2|\widehat{\beta}_j^{(old)}|)$$

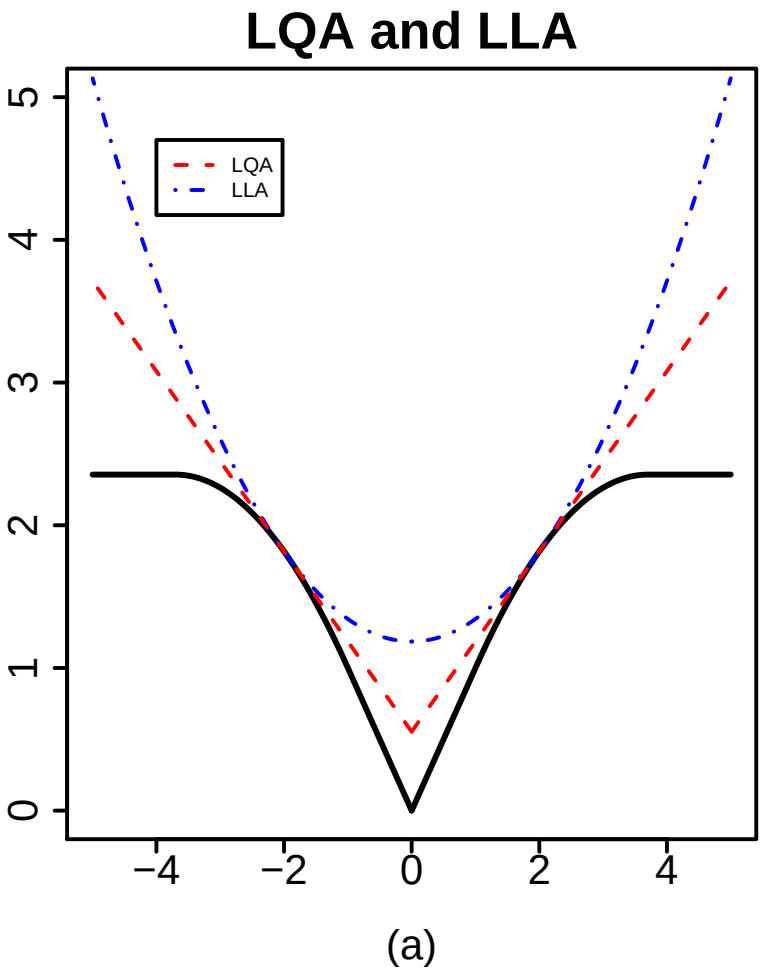
  - Local linear approximation (LLA; Zou-Li, 2008):

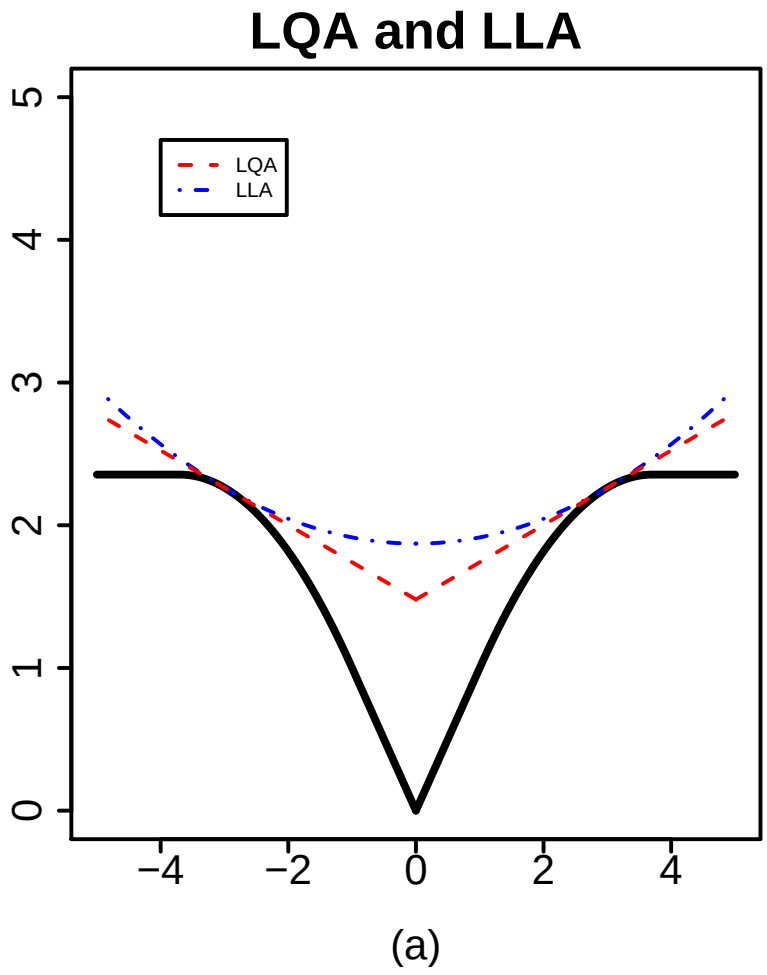
$$p_\lambda(|\beta_j|) \leq p_\lambda(|\widehat{\beta}_j^{(old)}|) + p'_\lambda(|\widehat{\beta}_j^{(old)}|)(|\beta_j| - |\widehat{\beta}_j^{(old)}|) = p'_\lambda(|\widehat{\beta}_j^{(old)}|)|\beta_j| + C_j$$

- Majorizing the loss
  - (Fast) iterative shrinkage-thresholding algorithm (FISTA; Nesterov, 83; Daubechies et al, 2004; Beck-Teboulle, 2009)

$$\mathcal{L}_0(\boldsymbol{\beta}) \leq \mathcal{L}_0(\widehat{\boldsymbol{\beta}}^{(old)}) + \left\langle \nabla \mathcal{L}_0(\widehat{\boldsymbol{\beta}}^{(old)}), \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(old)} \right\rangle + \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(old)}\|_2^2/(2s)$$

with $\sup_{\boldsymbol{\beta}} \|\nabla^{\otimes 2} \mathcal{L}_0(\boldsymbol{\beta})\|_{spectrum} \leq 1/s$





See Zou-Li 08; Thanks to J. Fan

Properties of LLA: Let $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(old)}$ and $\widehat{\boldsymbol{\beta}}$ be an estimator satisfying

$$\left|\boldsymbol{X}_j^T(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Y})/n + p'_\lambda(\widetilde{\beta}_j)\partial|\widehat{\beta}_j|\right| \le \lambda\nu_j$$

Basic inequality: Suppose $|p'_\lambda(t_1) - p'_\lambda(t_2)| \le \kappa(p_\lambda)|t_1 - t_2|\ \forall\ 0 < t_1 < t_2$. Similar to the basic inequality for the path following algorithm, we have

$$\boldsymbol{h}^T\overline{\boldsymbol{\Sigma}}\boldsymbol{h} + (1-\eta)\lambda\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \lambda\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{h}_{\mathcal{S}} + \|\boldsymbol{h}\|_2\big(\kappa(p_\lambda)\|\widetilde{\boldsymbol{h}}\|_2 + \lambda\|\boldsymbol{\nu}\|_2\big)$$

with $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\widetilde{\boldsymbol{h}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, and $\boldsymbol{w}_{\mathcal{S}} = \{\boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n - p'_\lambda(\boldsymbol{\beta}^*_{\mathcal{S}})\}/\lambda$.

Theorem 13: (i) *Suppose p'_λ satisfies the Lipschitz condition. Then,*

$$\|\boldsymbol{h}\| \le \lambda \sup\left\{\|\boldsymbol{u}\|\psi(\boldsymbol{u}) : \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} = 1\right\}$$

with $\psi(\boldsymbol{u}) = \boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 + \|\boldsymbol{u}\|_2\big\{\kappa(p_\lambda)\|\widetilde{\boldsymbol{h}}\|_2/\lambda + \|\boldsymbol{\nu}\|_2\big\}$.
(ii) *Let* $\rho_0 = \sup\{\|\boldsymbol{u}\|_2 : \psi(\boldsymbol{u}) > 0, \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} = 1\}$. *Then,*

$$\|\boldsymbol{h}\|_2 \le \rho_0\kappa(p_\lambda)\|\widetilde{\boldsymbol{h}}\|_2 + \lambda \sup_{\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}=1}\left\{\|\boldsymbol{u}\|_2\big(\boldsymbol{w}_{\mathcal{S}}^T\boldsymbol{u}_{\mathcal{S}} - (1-\eta)\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 + \|\boldsymbol{u}\|_2\|\boldsymbol{\nu}\|_2\big)\right\}$$

– Multistage LLA/approximate solution (Zhang, 2010, 2013; Z-Zhang, 2012)
– RSC (Negahban et al, 2012)

Estimation of HD objects

- Universal penalty level: $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$
- Sparsity condition: $\sum_{j=1}^{p} \min(1, |\beta_j|/\lambda_{univ}) \leq s^*$
- Prediction: With at least probability $1 - p^{-a_0}$,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/n \leq C_{\text{pred}} s^* \lambda_{univ}^2$$

- Estimation: With at least probability $1 - p^{-a_0}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \leq C_{\text{est,q}} s^* \lambda_{univ}^q$$

- Variable selection:

$$\beta_{\min} = \min_{\beta_j \neq 0} |\beta_j| \geq C_{\text{select}} \lambda_{univ}$$

$$\Rightarrow \quad \mathbb{P}\Big\{\text{sgn}(\widehat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})\Big\} \geq 1 - p^{-a_0}$$

- Proper conditions on $\boldsymbol{X}$ yield bounded C_{pred}, $C_{\text{est,q}}$ and C_{select}

How about confidence interval and p-value for β_j?

Statistical inference after model selection

- A model M is selected based on screening data
- Fresh data $(\boldsymbol{X}, \boldsymbol{Y})$ are observed independent of M, $n \gg |M|$
- Confidence interval for $\theta = \boldsymbol{a}^T \boldsymbol{\beta}$ after model selection:

$$\boldsymbol{a}_M^T \widehat{\boldsymbol{\beta}}_M^{(lse)} \pm 1.96\sigma \left(\boldsymbol{a}_M (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{a}_M \right)^{1/2}$$

Statistical inference after model selection

- A model M is selected based on screening data
- Fresh data $(\boldsymbol{X}, \boldsymbol{Y})$ are observed independent of M, $n \gg |M|$
- Confidence interval for $\theta = \boldsymbol{a}^T \boldsymbol{\beta}$ after model selection:

$$\boldsymbol{a}_M^T \widehat{\boldsymbol{\beta}}_M^{(lse)} \pm 1.96\sigma \left(\boldsymbol{a}_M (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{a}_M \right)^{1/2}$$

A superefficiency paradox

- If an oracle model A is available such that

$$A \supseteq M \cup \{j : \beta_j \neq 0\}, \quad |M| \vee \|\boldsymbol{\beta}\|_0 < |A| \ll n,$$

- do we use

$$\boldsymbol{a}_A^T \widehat{\boldsymbol{\beta}}_A^{(lse)} \pm 1.96\sigma \left(\boldsymbol{a}_A (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1} \boldsymbol{a}_A \right)^{1/2} ?$$

- If we do, then the CI with the oracle information is wider:

$$\left(\boldsymbol{a}_A (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1} \boldsymbol{a}_A \right)^{1/2} > \left(\boldsymbol{a}_M (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{a}_M \right)^{1/2}.$$

Conservative approach: Berk et al (2013)

Estimation of the noise level

- Theoretical results for the Lasso suggests

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1 \right\}$$

  with $\lambda = \sigma\lambda_0$, $\lambda_0 = \eta^{-1}\sqrt{(2/n)\log(p/k)}$, $1 \le k \ll n/\log p$

- A "naive estimate" of σ (Sun-Z, 2010): A recursive solution of

$$\widehat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{\sigma}\lambda_0)\|_2^2/n$$

- Convex minimization formulation for the same $\widehat{\sigma}$ (Antoniadis, 2010):

$$\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\} = \arg\min_{\{\boldsymbol{\beta},\sigma\}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2\sigma n) + \sigma/2 + \lambda_0\|\boldsymbol{\beta}\|_1 \right\}$$

- Theory for this scaled Lasso (Sun-Z, 2012, 2013):

$$\left|\widehat{\sigma}/\sigma^* - 1\right| \le O(1) C_{\mathrm{est},1}(s^* + k)\lambda_0^2$$
$$(s^* + k)\log p \ll n^{1/2} \;\Rightarrow\; \sqrt{n}(\widehat{\sigma}/\sigma - 1) \to N(0, 1/2)$$

  where $(\sigma^*)^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/n = \sigma^2\chi_n^2$

- An equivalent "square-root Lasso" formulation of $\widehat{\boldsymbol{\beta}}$ (Belloni et al, 2011)
- An earlier scaled Lasso proposal (Städler et al, 2010) yields $\|\widehat{\sigma}/\sigma^* - 1| \lesssim s^*\lambda_0^2 + \lambda_0\|\boldsymbol{\beta}/\sigma\|_1$

Bias correction/Statistical inference after regularized estimation

- Linear bias correction from an initial estimator, e.g. $\widehat{\boldsymbol{\beta}}^{(init)} = \widehat{\boldsymbol{\beta}}^{(lasso)}$:

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + \boldsymbol{w}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)})$$

- Error decomposition:

$$\widehat{\beta}_j - \beta_j \quad = \quad \boldsymbol{w}_j^\top \boldsymbol{\varepsilon} - (\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j)^\top (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})$$

- The LS property: $\boldsymbol{X}^\top \boldsymbol{w}_j = \boldsymbol{e}_j$, i.e. $\boldsymbol{w}_j \propto \boldsymbol{X}_j^\perp \perp \boldsymbol{X}_k \ \forall\ k \neq j$
- A relaxed LS property is sufficient
  - Noise factor: $\tau_j = \|\boldsymbol{w}_j\|_2$, hopefully $\tau_j \asymp n^{-1/2}$
  - Bias factor: $\eta_j = \|\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j\|_\infty / \|\boldsymbol{w}_j\|_2$, hopefully $\eta_j \leq \sqrt{2 \log p}$
  - Asymptotic theory based on ℓ_∞-ℓ_1 split:

$$\eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \ll 1 \ \Rightarrow\ (\widehat{\beta}_j - \beta_j)/(\widehat{\sigma}\|\boldsymbol{w}_j\|_2) \to N(0,1)$$

  - Sample size requirement: $n \gg (s^* \log p)^2$ under "proper conditions"
  - Bonferroni adjustment: $\eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \ll \sqrt{\log p}$
- Low-dimensional projection estimator (LDPE, Z-Zhang, 14; arXiv 11)

Bias correction (details)

- Error decomposition:

$$\widehat{\beta}_j - \beta_j \quad = \quad \boldsymbol{w}_j^\top \boldsymbol{\varepsilon} - (\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j)^\top (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})$$

- A relaxed LS property is sufficient
  - Noise factor: $\tau_j = \|\boldsymbol{w}_j\|_2$, hopefully $\tau_j \asymp n^{-1/2}$
  - Bias factor: $\eta_j = \|\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j\|_\infty / \|\boldsymbol{w}_j\|_2$, hopefully $\eta_j \leq \sqrt{2 \log p}$
  - Asymptotic theory based on ℓ_∞-ℓ_1 split:

$$\eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \ll 1 \;\Rightarrow\; (\widehat{\beta}_j - \beta_j)/(\widehat{\sigma}\|\boldsymbol{w}_j\|_2) \to N(0,1)$$

  - Sample size requirement: $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$
    - Sparsity assumption: $\sum_j \min(|\beta_j|/\lambda_{univ}, 1) \leq s^*$
    - ℓ_1 error bound: $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \leq C_{\mathrm{est},1} s^* \lambda_{univ}$
    - $\eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \leq C_{\mathrm{est},1} s^* \sigma (\log p)\sqrt{2/n}$
    - Sample size requirement: $n \gg (s^* \log p)^2$
  - Bonferroni adjustment for simultaneous interval estimation of all β_j
    - $\eta_j \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 \ll \sqrt{\log p}$
    - Sample size requirement: $n \gg (s^*)^2 \log p$

Finding a "score vector" for LD projection

- Under the sample size and proper regularity conditions,
  - $(\widehat{\beta}_j - \beta_j)/\tau_j \to N(0, \sigma^2)$ with $\tau_j = \|\boldsymbol{w}_j\|_2$
  - when $\eta_j = \|\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j\|_\infty / \|\boldsymbol{w}_j\|_2 \le C\sqrt{2\log p}$
- When $\mathrm{rank}(\boldsymbol{X}) = p$,
  - $\boldsymbol{w}_j = \boldsymbol{X}_j^\perp / \|\boldsymbol{X}_j^\perp\|_2^2$ yields the LSE
  - $\eta_j = 0$ and $\tau_j = 1/\|\boldsymbol{X}_j^\perp\|_2$
- When $p > n$,
  - $\boldsymbol{X}_j^\perp = 0$ when $\boldsymbol{X}$ is in "general position"
  - $\boldsymbol{z}_j =$ a relaxed version of $\boldsymbol{X}_j^\perp$
  - $\boldsymbol{w}_j = \boldsymbol{z}_j / \boldsymbol{z}_j^T \boldsymbol{X}_j$ to ensure $\boldsymbol{w}_j^T \boldsymbol{X}_j = 1$; Alternatively, $\boldsymbol{w}_j = \boldsymbol{z}_j / \|\boldsymbol{z}_j\|_2^2$
- For random designs with $\mathbb{E}\boldsymbol{X}^T\boldsymbol{X}/n = \boldsymbol{\Sigma}$
  - Regression model

$$\boldsymbol{X}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}_{-j} + \boldsymbol{z}_j^o, \quad \sigma_j^2 = \mathbb{E}\|\boldsymbol{z}_j^o\|_2^2/n = 1/(\boldsymbol{\Sigma}^{-1})_{jj}$$

  - Population version of $\boldsymbol{X}_j^\perp$: $\boldsymbol{z}_j^o = \boldsymbol{X}\boldsymbol{u}^o \propto \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$, $\boldsymbol{u}_j^o = 1$, $\boldsymbol{u}_{-j}^o = -\boldsymbol{\gamma}_{-j}$
  - When $\boldsymbol{z}_j^o$ is unknown: $\boldsymbol{z}_j \approx \boldsymbol{z}_j^o$, or $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{u}$ with $\boldsymbol{u} \approx \boldsymbol{u}^o$
  - Expected: $\tau_j = (1 + o(1))n^{-1/2}/\sigma_j$
- Algorithms:

$$\begin{aligned} \text{Choice 1:} \quad & \boldsymbol{z}_j = \text{residual vector of PLSE}(\boldsymbol{X}_{-j}, \boldsymbol{X}_j) \\ \text{Choice 2:} \quad & \boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \left\{ \|\boldsymbol{z}\|_2^2 : \boldsymbol{z}^T\boldsymbol{X}_j = n,\ \|\boldsymbol{z}^T\boldsymbol{X}_{-j}/n\|_\infty \le \lambda' \right\} \end{aligned}$$

Finding a score vector with Lasso/PLSE

- Under the sample size and proper regularity conditions,
  - $(\widehat{\beta}_j - \beta_j)/\tau_j \to N(0, \sigma^2)$ with $\tau_j = \|\boldsymbol{w}_j\|_2$
  - when $\eta_j = \|\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j\|_\infty / \|\boldsymbol{w}_j\|_2 \le C\sqrt{2\log p}$
  - $\boldsymbol{w}_j = \boldsymbol{z}_j / \boldsymbol{z}_j^T \boldsymbol{X}_j$, $\boldsymbol{X}_j^T \boldsymbol{w}_j = 1$
- Lasso: $\boldsymbol{z}_j = \boldsymbol{X}_j - \boldsymbol{X}_{-j}\widehat{\gamma}_{-j}$

$$\widehat{\gamma}_{-j} = \arg\min_{\gamma} \left\{ \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\gamma\|_2^2/(2n) + \lambda'\|\gamma\|_1 \right\}$$

  - KKT: $\boldsymbol{X}_{-j}^T(\boldsymbol{X}_j - \boldsymbol{X}_{-j}\widehat{\gamma}_{-j})/n = \boldsymbol{X}_{-j}^T\boldsymbol{z}_j/n \in \lambda'\partial\|\widehat{\gamma}_{-j}\|_1$
  - $\boldsymbol{X}_j^T\boldsymbol{z}_j/n = \|\boldsymbol{z}_j\|_2^2/n + (\boldsymbol{X}_{-j}\widehat{\gamma}_{-j})^T\boldsymbol{z}_j/n = \|\boldsymbol{z}_j\|_2^2/n + \lambda'\|\widehat{\gamma}_{-j}\|_1$
  - $\tau_j = \|\boldsymbol{w}_j\|_2 = \|\boldsymbol{z}_j\|_2/|\boldsymbol{z}_j^T\boldsymbol{X}_j| \le 1/\|\boldsymbol{z}_j\|_2$
  - $\eta_j = \|\boldsymbol{X}_{-j}^T\boldsymbol{w}_j\|_\infty/\|\boldsymbol{w}_j\|_2 = \|\boldsymbol{X}_{-j}^T\boldsymbol{z}_j\|_\infty/\|\boldsymbol{z}_j\|_2 = \lambda' n/\|\boldsymbol{z}_j\|_2 \le \sqrt{2\log p}$
  - $\lambda' = \|\boldsymbol{z}_j/n^{1/2}\|_2\sqrt{(2/n)\log p} = \widehat{\sigma}_j\lambda_0$
  - This suggests $\eta_j = \sqrt{2\log p}$ for the scaled Lasso
- $\eta_j = \|\boldsymbol{X}_{-j}^T\boldsymbol{z}_j\|_\infty/\|\boldsymbol{z}_j\|_2 \downarrow \lambda'$
- $\tau_j \le 1/\|\boldsymbol{z}_j\|_2 \uparrow \lambda'$

Finding a score vector by quadratic programming

- Under proper regularity conditions,
  - $(\widehat{\beta}_j - \beta_j)/\tau_j \to N(0, \sigma^2)$ with $\tau_j = \|\boldsymbol{w}_j\|_2$
  - when $\eta_j = \|\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j\|_\infty / \|\boldsymbol{w}_j\|_2 \le \sqrt{2 \log p}$
  - $\boldsymbol{w}_j = \boldsymbol{z}_j / \boldsymbol{z}_j^T \boldsymbol{X}_j$
  - $\eta_j = \|\boldsymbol{X}_{-j}^T \boldsymbol{z}_j\|_\infty / \|\boldsymbol{z}_j\|_2$ as $\boldsymbol{X}_j^T \boldsymbol{w}_j = 1$
  - $\tau_j = \|\boldsymbol{z}_j\|_2 / |\boldsymbol{z}_j^T \boldsymbol{X}_j| = \|\boldsymbol{z}_j\|_2 / n$ when $\boldsymbol{X}_j^T \boldsymbol{z}_j = n$
- Quadratic programming: minimize τ_j given $\eta_j \le \sqrt{2 \log p}$

$$\begin{aligned} \boldsymbol{z}_j &= \arg\min_{\boldsymbol{z}} \left\{ \tau_j : \boldsymbol{z}^T \boldsymbol{X}_j = n,\ \eta_j \le \sqrt{2 \log p} \right\} \\ &= \arg\min_{\boldsymbol{z}} \left\{ \|\boldsymbol{z}\|_2^2 : \boldsymbol{z}^T \boldsymbol{X}_j = n,\ \|\boldsymbol{X}_{-j}^T \boldsymbol{z}\|_\infty \le \|\boldsymbol{z}\|_2 \sqrt{2 \log p} \right\} \end{aligned}$$

- Z-Zhang (14; arXiv 11), Javanmard-Montanari (14)
- Near estimability:
  - If $\eta_j \le \sqrt{2 \log p}$ is attainable, or $C\sqrt{2 \log p}$, β_j is "nearly estimable"
  - Otherwise, β_j is not nearly estimable
  - For Gaussian designs, $\eta_j \le \sqrt{2 \log p}$ is attainable with large probability

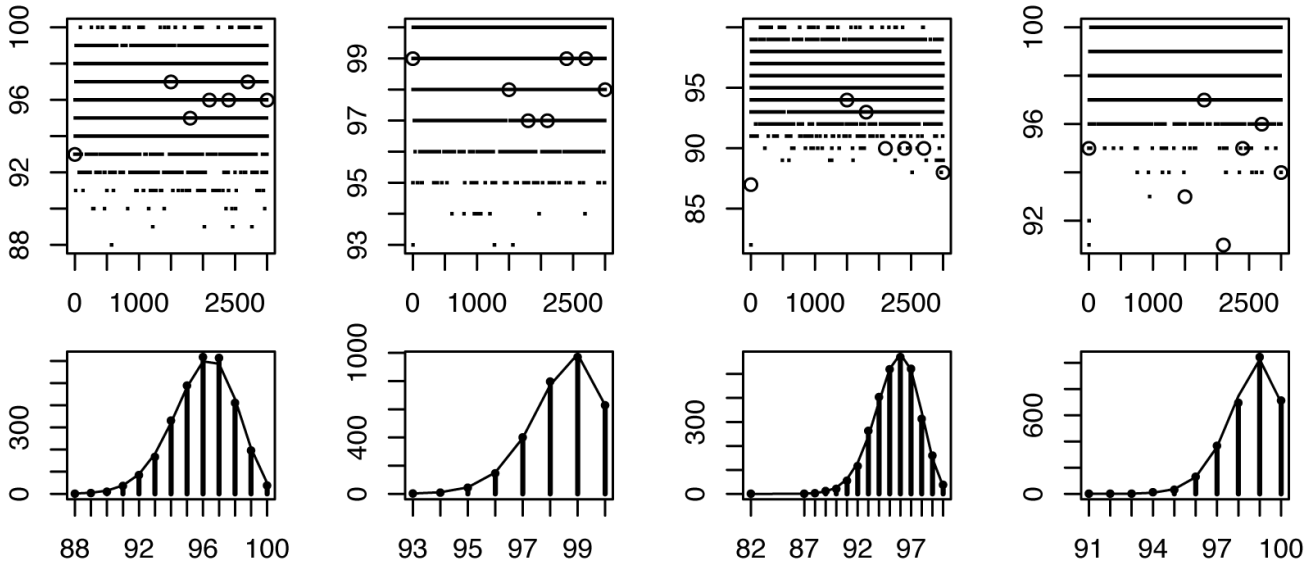
Some simulation results:

- $n = 200$, $p = 3000$, $\sigma = 1$, $\lambda_{univ} = \sqrt{(2/n)\log p} = 0.283$, $\beta_j = 3\lambda_{univ} = 0.849$ for $j = 1500, 1800, \ldots, 3000$, and $\beta_j = 3\lambda_{univ}/j^{\alpha}$ otherwise; $\beta_j \neq 0$ for all j
- $(s, s * (\log p)/\sqrt{n}) = (8.93, 5.05)$ and $(29.24, 16.55)$ respectively for $\alpha = 1$ and 2, while the theory requires $s(\log p)/\sqrt{n} \to 0$, where $s = \sum_j \min(|\beta_j|/\lambda_{univ}, 1)$.
- Generate $(\widetilde{\boldsymbol{X}}, \boldsymbol{X}, \varepsilon)$ in each replication, where $\widetilde{X}$ has iid $N(0, \boldsymbol{\Sigma})$ rows with $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{p\times p}$ and $\boldsymbol{X}$ is the column normalized version of $\widetilde{\boldsymbol{X}}$
- Four settings, labeled (A), (B), (C), and (D), respectively, with $(\alpha, \rho) = (2, 1/5)$, $(1, 1/5)$, $(2, 4/5)$, and $(1, 4/5)$
- Case (D) is most difficult

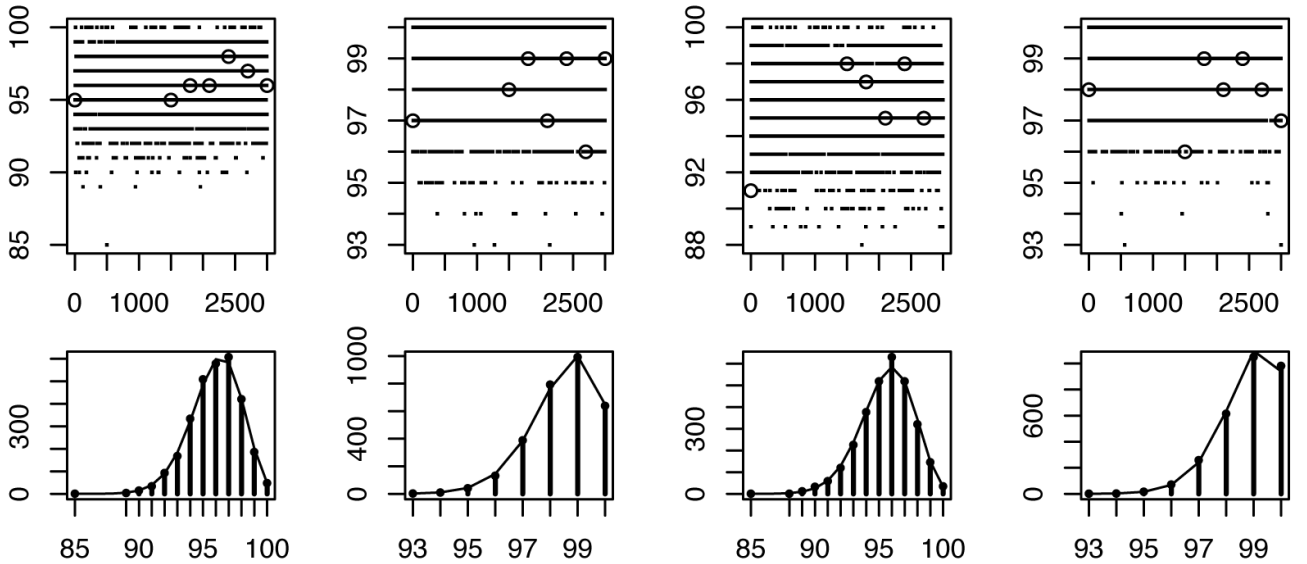
		(A)	(B)	(C)	(D)
all β_j	LDPE	0.9597	0.9845	0.9556	0.9855
	R-LDPE	0.9595	0.9848	0.9557	0.9885
maximal β_j	LDPE	0.9571	0.9814	0.9029	0.9443
	R-LDPE	0.9614	0.9786	0.9414	0.9786

Table: Mean coverage probability of LDPE and R-LDPE.

LDPE



Restricted LDPE



Bias correction (summary)

- Linear bias correction from an initial estimator, e.g. $\widehat{\boldsymbol{\beta}}^{(init)} = \widehat{\boldsymbol{\beta}}^{(lasso)}$:

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + \boldsymbol{w}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)})$$

- Error decomposition:

$$\widehat{\beta}_j - \beta_j \quad = \quad \boldsymbol{w}_j^\top \boldsymbol{\varepsilon} - (\boldsymbol{X}^\top \boldsymbol{w}_j - \boldsymbol{e}_j)^\top (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})$$

- Asymptotic normality $(\widehat{\beta}_j - \beta_j)/(\widehat{\sigma}\|\boldsymbol{w}_j\|_2) \to N(0,1)$ when
  - $\boldsymbol{w}_j = \boldsymbol{z}_j / \boldsymbol{z}_j^T \boldsymbol{X}_j$
  - $\eta_j = \|\boldsymbol{X}_{-j}^T \boldsymbol{z}_j\|_\infty / \|\boldsymbol{z}_j\|_2 \le \sqrt{2 \log p}$
  - $\sum_j \min(|\beta_j|/\lambda_{univ}, 1) \le s^*$
  - $s^* \log p \ll n^{1/2}$
- Algorithms:

  Choice 1: $\boldsymbol{z}_j =$ residual vector of PLSE$(\boldsymbol{X}_{-j}, \boldsymbol{X}_j)$

  Choice 2: $\boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \left\{ \|\boldsymbol{z}\|_2^2 : \boldsymbol{z}^T \boldsymbol{X}_j = n,\ \|\boldsymbol{X}_{-j}^T \boldsymbol{z}\|_\infty \le \|\boldsymbol{z}\|_2 \sqrt{2 \log p} \right\}$

- Approximate confidence interval for β_j is available if either Choice 2 is feasible or Choice 1 yields a feasible solution for Choice 2
- Statistical inference for β_j is "difficult" conditionally on $\boldsymbol{X}$ if the algorithms are infeasible, i.e. failing to be nearly estimable

Connection to semiparametric inference (Z, 11)

- Semiparametric inference:

$$\text{Parametric component} + \text{NP component}$$

- Low-dimensional statistical inference with high-dimensional data, or "semi-LD inference":

$$\text{LD component} + \text{HD component}$$

- General methodology:

$$\begin{gathered}\text{HD estimation} \Rightarrow \text{semi-LD inference} \\ \text{is parallel to} \\ \text{NP estimation} \Rightarrow \text{semiparametric inference}\end{gathered}$$

- We borrow ideas from Engle et al (81), Bickel (82), Wahba (1984), Chen (85,88), Rice (86), Heckman (86), Bickel et al (90), ...

Fisher information for $\theta = \tau(\boldsymbol{\beta})$ in small submodels (Stein, 56)

- Suppose $\boldsymbol{X}$ has iid rows with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}^T\boldsymbol{X}/n)$
- Let

$$\boldsymbol{a} = (\partial/\partial\boldsymbol{\beta})\tau(\boldsymbol{\beta})$$

- The minimum Fisher information in the submodel $\boldsymbol{\beta} = \boldsymbol{u}\theta$:

$$F_\theta = \min_{\boldsymbol{u}^T\boldsymbol{a}=1} \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}/\sigma^2 = (\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a})^{-1}/\sigma^2$$

- The least favorable submodel:

$$\boldsymbol{u}^o = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}\big(\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\big)^{-1}$$

- One-step correction: With $\boldsymbol{u} \approx \boldsymbol{u}^o$ and $\boldsymbol{u}^T\boldsymbol{a} = 1$,

$$\widehat{\theta} = \tau(\boldsymbol{\beta}^{(init)}) + \arg\max_{\phi} \text{ log-lik}(\boldsymbol{\beta}^{(init)} + \boldsymbol{u}\phi)$$

- Le Cam (69), Bickel (82), Schick (86), Klaassen (87), Bickel et al (90); Z(11, Oberwolfach), Z-Zhang (14), van de Geer et al (2014), van de Geer (2014); Sun-Z (12), Ren et al (13)

The LDPE attains the marginal minimum Fisher information:

- Random $\boldsymbol{X}$ with iid sub-Gaussian rows; $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{X}^\top \boldsymbol{X}/n$, $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Sigma})$,

$$s = \sum_{j=1}^{p} \min(1, \boldsymbol{D}_j \beta_j^2 / (\sigma^2 (2/n) \log p))$$

- Assume that $\boldsymbol{X}\widehat{\boldsymbol{D}}^{-1/2}$ is subGaussian and

$$\lambda_{\min}\Big(\boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{-1/2}\Big) \geq c_*, \ s \log p \ll n^{1/2}$$

- Then, the minimum Fisher information is $F_j = 1/\{\sigma^2(\boldsymbol{\Sigma}^{-1})_{jj}\}$, and

$$(n\widehat{F}_j)^{1/2}(\widehat{\beta}_j - \beta_j) \to N(0,1), \quad \widehat{F}_j/F_j \to 1$$

- Interpretation: the conditional inference "works" with high marginal probability
- Z (11), van de Geer-Bühlmann-Ritov (14)

The sample size condition is rate optimal

- Suppose $\boldsymbol{X}$ has iid $N(0, \boldsymbol{\Sigma})$ rows under $\mathbb{P}$. Let

$$\mathscr{P} = \left\{ \mathbb{P} : c_* \leq \text{eigen}\left(\boldsymbol{D}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}^{-1/2} \right) \leq c^*, s \log p \leq c'_* n \right\}$$

- For certain $c''_* > 0$ depending on $\{c_*, c^*, c'_*\}$ only,

$$\inf_{\widehat{\theta}} \sup_{\mathscr{P}} \mathbb{P}\left\{ n^{1/2} |\widehat{\theta} - \theta| \geq c''_* s(\log p)/n^{1/2} \right\} > 0$$

- Le Cam (73), Ren et al (13)
- When $n^{1/2} \ll s \log p \ll n$, $n^{-1/2}$ rate is impossible without some additional assumption
- Javanmard-Montanari (14): $\boldsymbol{\Sigma} = \boldsymbol{I}$ (known) with $n \gg s \log p$

Some additional references

- Leeb-Pötscher (06)
- Laber-Murphy (11)
- Bühlmann (13), Belloni et al (14)
- Meinshausen-Bühlmann (10)
- Lockhart et al (14), Lee et al (14)

De-biasing regularized estimators

- Low-dimensional projection from $\widehat{\boldsymbol{\beta}}^{(init)}$, e.g. $\widehat{\boldsymbol{\beta}}^{(lasso)}$:

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \boldsymbol{z}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}),$$

  i.e. $\boldsymbol{w}_j = \boldsymbol{z}_j / \boldsymbol{z}_j^\top \boldsymbol{X}_j$ with $\boldsymbol{w}_j^\top \boldsymbol{X}_j = 1$

- Asymptotic theory:

$$\widehat{\beta}_j - \beta_j \quad = \quad (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \boldsymbol{z}_j^\top \boldsymbol{\varepsilon} - (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \sum_{k \neq j} \boldsymbol{z}_j^\top \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})_k$$

  Bias after de-biasing: $\|\boldsymbol{z}_j\|_2^{-1} |\sum_{k \neq j} \boldsymbol{z}_j^\top \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})_k| = o_P(1)$

  - Standard error for homoscedastic ε: $\text{s.e.}_j = (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \|\boldsymbol{z}_j\| \sigma$
  - Studentization: $T_j = (\widehat{\beta}_j - \beta_j) / \widehat{\text{s.e.}}_j$
  - Conservative simultaneous interval/inference (Bonferroni):

$$\max_{j \in G} |T_j| \leq \Phi^{-1}(1 - \alpha/(2|G|))$$

- Bootstrap: more accurate simultaneous inference, heteroscedastic ε

Some references

- Joint with Ruben Dezeure and Peter Bühlmannn
- Residual bootstrap: Efron (79)
- Paired bootstrap and heteroscedasticity: Efron (79), Liu-Singh (92)
- Wild bootstrap and heteroscedasticity: Wu (86), Liu-Singh (92), Mammen (93), Chernozhukov et al. (13), Zhang-Cheng (16)
- Chatterjee-Lahiri (11)

Bootstrapping the LDPE: $\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \boldsymbol{z}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)})$

- Notation: $(\boldsymbol{v})_{\text{cent}} = \boldsymbol{v} - (\boldsymbol{1}\boldsymbol{1}^\top/n)\boldsymbol{v}$, the Hadamard product $\boldsymbol{u} \circ \boldsymbol{v}$
- Asymptotic theory
  - $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}$, $\widehat{\boldsymbol{\varepsilon}}_{\text{cent}} = (\widehat{\boldsymbol{\varepsilon}})_{\text{cent}}$
  - $\widehat{\text{s.e.}}_j = (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \|\boldsymbol{z}_j\|_2 \|\widehat{\boldsymbol{\varepsilon}}_{\text{cent}}\|_2/\sqrt{n}$
  - $T_j = (\widehat{\beta}_j - \beta_j)/\widehat{\text{s.e.}}_j$
  - $\widehat{\text{s.e.}}_{j,\text{robust}} = (\boldsymbol{z}_j^\top \boldsymbol{X}_j)^{-1} \|(\boldsymbol{z}_j \circ \widehat{\boldsymbol{\varepsilon}})_{\text{cent}}\|_2$ for heteroscedastic $\boldsymbol{\varepsilon}$
  - $T_{j,\text{robust}} = (\widehat{\beta}_j - \beta_j)/\widehat{\text{s.e.}}_{j,\text{robust}}$

Bootstrapping the LDPE: $\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + (z_j^\top X_j)^{-1} z_j^\top (Y - X\widehat{\beta}^{(init)})$

- Notation: $(v)_{\text{cent}} = v - (11^\top/n)v$, the Hadamard product $u \circ v$
- Asymptotic theory
  - $\widehat{\varepsilon} = Y - X\widehat{\beta}^{(init)}$, $\widehat{\varepsilon}_{\text{cent}} = (\widehat{\varepsilon})_{\text{cent}}$
  - $\widehat{\text{s.e.}}_j = (z_j^\top X_j)^{-1} \|z_j\|_2 \|\widehat{\varepsilon}_{\text{cent}}\|_2/\sqrt{n}$
  - $T_j = (\widehat{\beta}_j - \beta_j)/\widehat{\text{s.e.}}_j$
  - $\widehat{\text{s.e.}}_{j,\text{robust}} = (z_j^\top X_j)^{-1} \|(z_j \circ \widehat{\varepsilon})_{\text{cent}}\|_2$ for heteroscedastic ε
  - $T_{j,\text{robust}} = (\widehat{\beta}_j - \beta_j)/\widehat{\text{s.e.}}_{j,\text{robust}}$
- Residual bootstrap
  - ε^* iid uniform distribution from elements of $\widehat{\varepsilon}_{\text{cent}}$
  - $Y^* = X\widehat{\beta}^{(init)} + \varepsilon^*$
  - $\widehat{\beta}^{*(init)}$ = the Lasso estimator based on (X, Y^*)
  - $\widehat{\beta}_j^* = \widehat{\beta}_j^{*(init)} + (z_j^\top X_j)^{-1} z_j^\top (Y^* - X\widehat{\beta}^{*(init)})$
  - $\widehat{\text{s.e.}}_j^* = (z_j^\top X_j)^{-1} \|z_j\|_2 \|(Y^* - X\widehat{\beta}^{*(init)})_{\text{cent}}\|_2/\sqrt{n}$,
  - $T_j^* = (\widehat{\beta}_j^* - \widehat{\beta}_j^{(lasso)})/\widehat{\text{s.e.}}_j^*$
  - $\widehat{\text{s.e.}}_{j,\text{robust}} = (z_j^\top X_j)^{-1} \|(z_j \circ (Y^* - X\widehat{\beta}^{*(init)}))_{\text{cent}}\|_2$
  - $T_{j,\text{robust}}^* = (\widehat{\beta}_j^* - \widehat{\beta}_j^{(lasso)})/\widehat{\text{s.e.}}_{j,\text{robust}}^*$

Wild bootstrap and xyz-paired bootstrap

- Wild bootstrap
  - Draw iid W_i with $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = 1$
  - Set $\varepsilon^* = \boldsymbol{W} \circ \widehat{\boldsymbol{\varepsilon}}_{\text{cent}}$, $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)} + \boldsymbol{\varepsilon}^*$
  - Proceed as in the residual bootstrap, i.e. plug-in with $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{z}_j)$

Wild bootstrap and xyz-paired bootstrap

- Wild bootstrap
  - Draw iid W_i with $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = 1$
  - Set $\varepsilon^* = \boldsymbol{W} \circ \widehat{\boldsymbol{\varepsilon}}_{\text{cent}}$, $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)} + \boldsymbol{\varepsilon}^*$
  - Proceed as in the residual bootstrap, i.e. plug-in with $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{z}_j)$
- xyz-paired bootstrap
  - Paired bootstrap: sample rows of $(\boldsymbol{X}, \boldsymbol{Y})$ with replacement

Wild bootstrap and xyz-paired bootstrap

- Wild bootstrap
  - Draw iid W_i with $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = 1$
  - Set $\varepsilon^* = \boldsymbol{W} \circ \widehat{\varepsilon}_{\text{cent}}$, $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\beta}^{(init)} + \varepsilon^*$
  - Proceed as in the residual bootstrap, i.e. plug-in with $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{z}_j)$
- xyz-paired bootstrap
  - Paired bootstrap: sample rows of $(\boldsymbol{X}, \boldsymbol{Y})$ with replacement
  - Construction of a proper linear model
    - $\widehat{\varepsilon}_{\text{cent}} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}^{(init)})_{\text{cent}}$,
    - $\widehat{\boldsymbol{X}}_k = \boldsymbol{X}_k - (\widehat{\varepsilon}_{\text{cent}}\widehat{\varepsilon}_{\text{cent}}^\top / \|\widehat{\varepsilon}_{\text{cent}}\|_2^2)\boldsymbol{X}_k$; $\widehat{\boldsymbol{X}}_k^\top \widehat{\varepsilon}_{\text{cent}} = 0$
    - $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\widehat{\beta}^{(init)} + \widehat{\varepsilon}_{\text{cent}}$
    - $\widehat{\boldsymbol{z}}_j = \boldsymbol{z}_j - (\widehat{\varepsilon}_{\text{cent}}\widehat{\varepsilon}_{\text{cent}}^\top / \|\widehat{\varepsilon}_{\text{cent}}\|_2^2)\boldsymbol{z}_j$; $\widehat{\boldsymbol{z}}_j^\top \widehat{\varepsilon}_{\text{cent}} = 0$
  - $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{z}_j^*, j \in G)$: iid sample of rows of $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{z}}_j, j \in G)$
  - T_j^* = the plug-in version based on $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{z}_j^*)$
  - $T_{j,\text{robust}}^*$ = the plug-in version based on $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{z}_j^*)$

Bootstrap methods, a summary

- Residual bootstrap
  - ε^* iid from elements of $\widehat{\varepsilon}_{\text{cent}} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)})_{\text{cent}}$
  - $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)} + \varepsilon^*$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{z}_j)$
- Wild bootstrap
  - Draw iid W_i with $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = 1$
  - $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)} + \boldsymbol{W} \circ \widehat{\varepsilon}_{\text{cent}}$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{z}_j)$
- The xyz-paired bootstrap
  - $\widehat{\boldsymbol{X}} \perp \widehat{\varepsilon}_{\text{cent}}$, $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}^{(init)} + \widehat{\varepsilon}_{\text{cent}}$, $\widehat{\boldsymbol{z}}_j \perp \widehat{\varepsilon}_{\text{cent}}$
  - $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{z}_j^*, j \in G)$: iid sample of rows of $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{z}}_j, j \in G)$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{z}_j^*)$
- No re-computation of $\boldsymbol{z}_j$ in bootstrap replications

Theoretical assumptions for simultaneous inference of $\beta_j, j \in G$:

- (A1): $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = o_P(1)/\sqrt{(\log p)\log(1+|G|)}$
- (A2): $\|\boldsymbol{z}_G^\top \boldsymbol{X}_{-j}/n\|_{\max} \lesssim \sqrt{(\log p)/n}$, $\|z_j\|_2^2/n \geq L_z$, $\|z_j\|_{2+\delta}^{2+\delta} \ll \|\boldsymbol{z}_j\|_2^{2+\delta}$
- (A3): ε_i independent, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma_i^2 \geq L$, $\mathbb{E}|\varepsilon_i|^{2+\delta} \leq C$
- (A4) $\|\widehat{\boldsymbol{\beta}}^{*(init)} - \widehat{\boldsymbol{\beta}}^{(init)}\|_1 = o_{P^*}(1)/\sqrt{(\log p)\log(1+|G|)}$ in probability
- (A5) $\|X\|_{\max} \leq C$
- (A6) $\max_{j\in G} \|\boldsymbol{z}_j\|_\infty \leq K$, $\delta = 2$, $\log(|G|) = o(n^{1/7})$

Theoretical assumptions for simultaneous inference of $\beta_j, j \in G$:

- (A1): $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = o_P(1)/\sqrt{(\log p)\log(1+|G|)}$
- (A2): $\|\boldsymbol{z}_G^\top \boldsymbol{X}_{-j}/n\|_{\max} \lesssim \sqrt{(\log p)/n}$, $\|z_j\|_2^2/n \geq L_z$, $\|z_j\|_{2+\delta}^{2+\delta} \ll \|\boldsymbol{z}_j\|_2^{2+\delta}$
- (A3): ε_i independent, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma_i^2 \geq L$, $\mathbb{E}|\varepsilon_i|^{2+\delta} \leq C$
- (A4) $\|\widehat{\boldsymbol{\beta}}^{*(init)} - \widehat{\boldsymbol{\beta}}^{(init)}\|_1 = o_{P^*}(1)/\sqrt{(\log p)\log(1+|G|)}$ in probability
- (A5) $\|X\|_{\max} \leq C$
- (A6) $\max_{j\in G} \|\boldsymbol{z}_j\|_\infty \leq K$, $\delta = 2$, $\log(|G|) = o(n^{1/7})$

For proper PLSE as $\widehat{\boldsymbol{\beta}}^{(init)}$ and under regularity conditions on $\boldsymbol{X}$ (RE or weaker)

- (A3) and (A5) imply $\|\boldsymbol{X}^T\boldsymbol{\varepsilon}/n\|_\infty \lesssim \lambda_{univ}$
- (A1) and (A4) hold when $n \gg (s\log p)^2 \log(1+|G|)$
- (A2) and (A6) hold if $\boldsymbol{X}$ has iid rows with $\mathrm{Var}(X_{ij}|X_{i,-j}) \geq v_0 > 0$

Consistency of the residual bootstrap

- Homoscedastic case: $\mathbb{E}\varepsilon_i^2 = \sigma^2$ for all $i \le n$
  - Suppose conditions (A1)-(A5) holds. If $|G| = O(1)$, then

$$\sup_{t_j, j \in G} \left| \mathbb{P}^*\{T_j^* \le t_j, j \in G\} - \mathbb{P}\{T_j \le t_j, j \in G\} \right| = o_P(1)$$

    with $T_j \to N(0,1)$ for each $j \in G$
  - If in addition (A6) holds, then

$$\sup_t \left| \mathbb{P}^*\{\max_{j \in G} h(T_j^*) \le t\} - \mathbb{P}\{\max_{j \in G} h(T_j) \le t\} \right| = o_P(1)$$

    for $h(t) = t$, $h(t) = -t$ or $h(t) = |t|$.
- Heteroscedastic case: Suppose (A1)-(A5). Then,

$$\sup_t \left| \mathbb{P}^*\{T_{j,\text{robust}}^* \le t\} - \mathbb{P}\{T_{j,\text{robust}} \le t\} \right| = o_P(1)$$

  with $T_{j,\text{robust}} \to N(0,1)$ for each $j \in G$

Consistency of the wild bootstrap and xyz-paired bootstrap

- Suppose conditions (A1)-(A5) holds. If $|G| = O(1)$, then

$$\sup_{t_j, j \in G} \left| \mathbb{P}^*\{T_j^* \leq t_j, j \in G\} - \mathbb{P}\{T_j \leq t_j, j \in G\} \right| = o_P(1)$$

  with $T_j \to N(0,1)$ for each $j \in G$

- If in addition (A6) holds and $\log p \ll n^{1/2}$, then

$$\sup_t \left| \mathbb{P}^*\{\max_{j \in G} h(T_j^*) \leq t\} - \mathbb{P}\{\max_{j \in G} h(T_j) \leq t\} \right| = o_P(1)$$

  for $h(t) = t$, $h(t) = -t$ or $h(t) = |t|$.

The theorem is applicable in the heteroscedastic case

Chernozukov et al (13)

- $\boldsymbol{\xi} \in \mathbb{R}^{n \times p}$ with independent rows, $\mathbb{E}\boldsymbol{\xi} = 0$, $\mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{\xi}/n = \boldsymbol{\Sigma}^{(1)}$
- $\boldsymbol{\zeta} \in \mathbb{R}^{n \times p}$, Gaussian, with $\mathbb{E}\boldsymbol{\zeta} = 0$, $\mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{\xi}/n = \boldsymbol{\Sigma}^{(2)}$
- CLT for maxima:

$$\max_t \left| \mathbb{P}\Big(\max_{j \le p} \mathbf{1}^\top \boldsymbol{\xi} \le t \Big) - \mathbb{P}\Big(\max_{j \le p} \mathbf{1}^\top \boldsymbol{\zeta} \le t \Big) \right| = o(1)$$

under the following conditions

- $c_1 \le \Sigma^{(k)}_{jj} \le c_2$, $n^{-1} \sum_{i=1}^n \mathbb{E}\xi_{ij}^4 \le M$
- $\log p \ll n^{1/7}$
- $\|\boldsymbol{\xi}\|_{\max} + \|\boldsymbol{\zeta}\|_{\max} = O_P(n^{1/2})/(\log(np))^{3/2}$
- $\|\boldsymbol{\Sigma}^{(1)} - \boldsymbol{\Sigma}^{(2)}\|_{\max} \ll 1/(\log p)^2$

Some simulation results

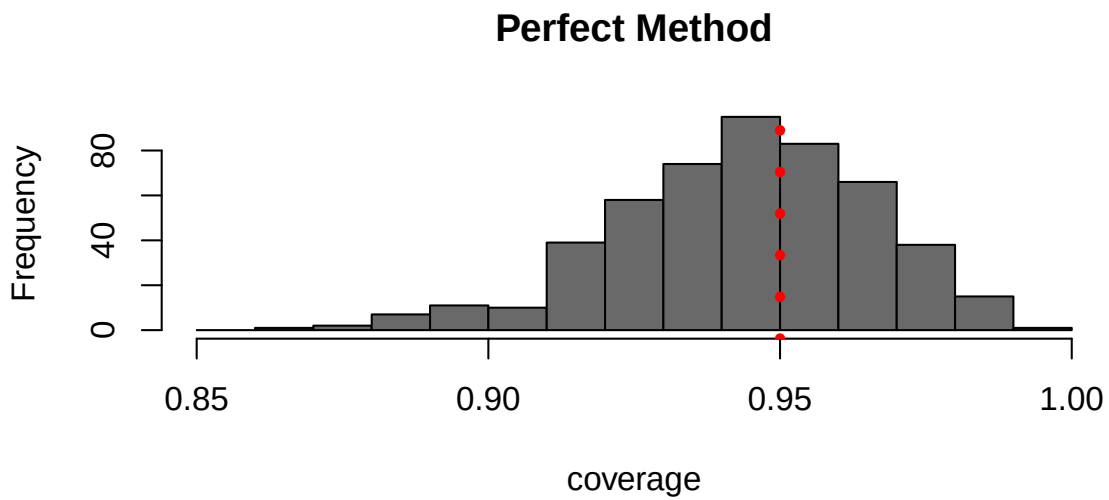


Figure: Histogram of the coverage probabilities of two sided 95% confidence intervals for 500 parameters. It illustrates how the results look like for a perfectly correct method for creating confidence intervals and one uses only 100 realizations to compute the probabilities.

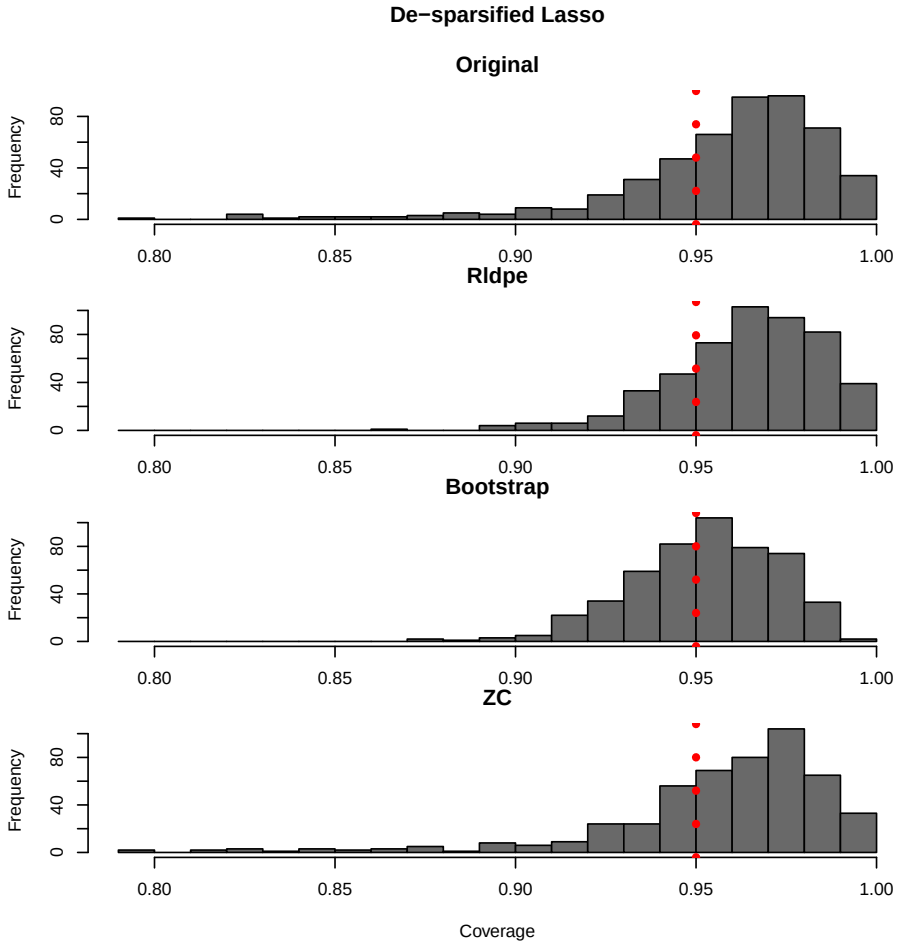


Figure: Histograms of the coverage probabilities of two-sided 95% confidence intervals for all 500 parameters in a linear model ($n = 100, p = 500$), computed from 100 independent replications. Perfect performance would look like Figure 1. The fixed design matrix is of Toeplitz type, the single coefficient vector of type $U(-2, 2)$ and **homoscedastic Gaussian errors**. The original estimator has more over-coverage and under-coverage than the bootstrapped estimator. The RLDPE estimator has little under-coverage, like the bootstrapped estimator, but it has too high coverage probabilities overall. The ZC approach to bootstrapping, which only bootstraps the linearized part of the estimator, doesn't show any improvements over the original de-sparsified Lasso.

de−sparsified Lasso

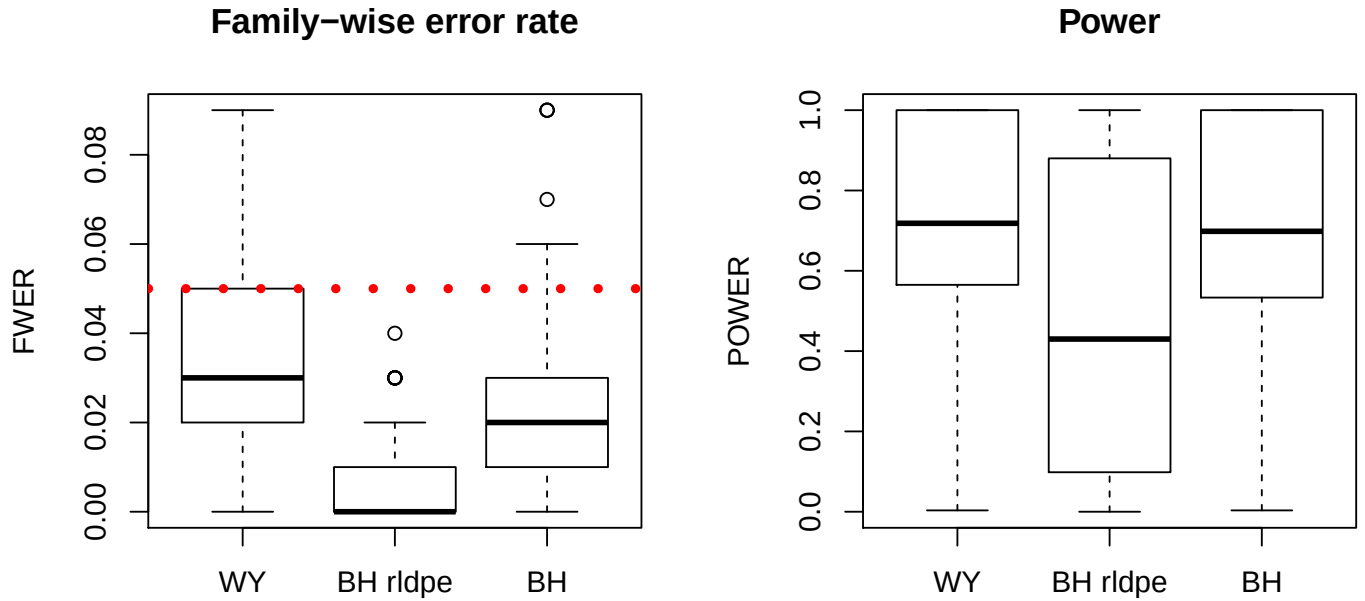


Figure: Boxplot of the familywise error rate and the power for multiple testing for the de-sparsified Lasso. The target is controlling the FWER at level 0.05, highlighted by a red-dotted horizontal line. Two different approaches for multiple testing correction are compared, Westfall-Young (WY) and Bonferroni-Holm (BH). For Bonferroni-Holm, we make the distinction between the original method and the RLDPE approach. 300 linear models are investigated in total, where 50 Toeplitz design matrices are combined with 50 coefficient vectors for each of the 6 types $U(0,2)$, $U(0,4)$, $U(-2,2)$, $\mathrm{fixed}\ 1$, $\mathrm{fixed}\ 2$, $\mathrm{fixed}\ 10$. The variables belonging to the active set are chosen randomly. The errors in the linear model were chosen to be **homoscedastic Gaussian**. Each of the models has a data point for the error rate and the power in the boxplot. The error rate and power probabilities were calculated by averaging over 100 realizations.

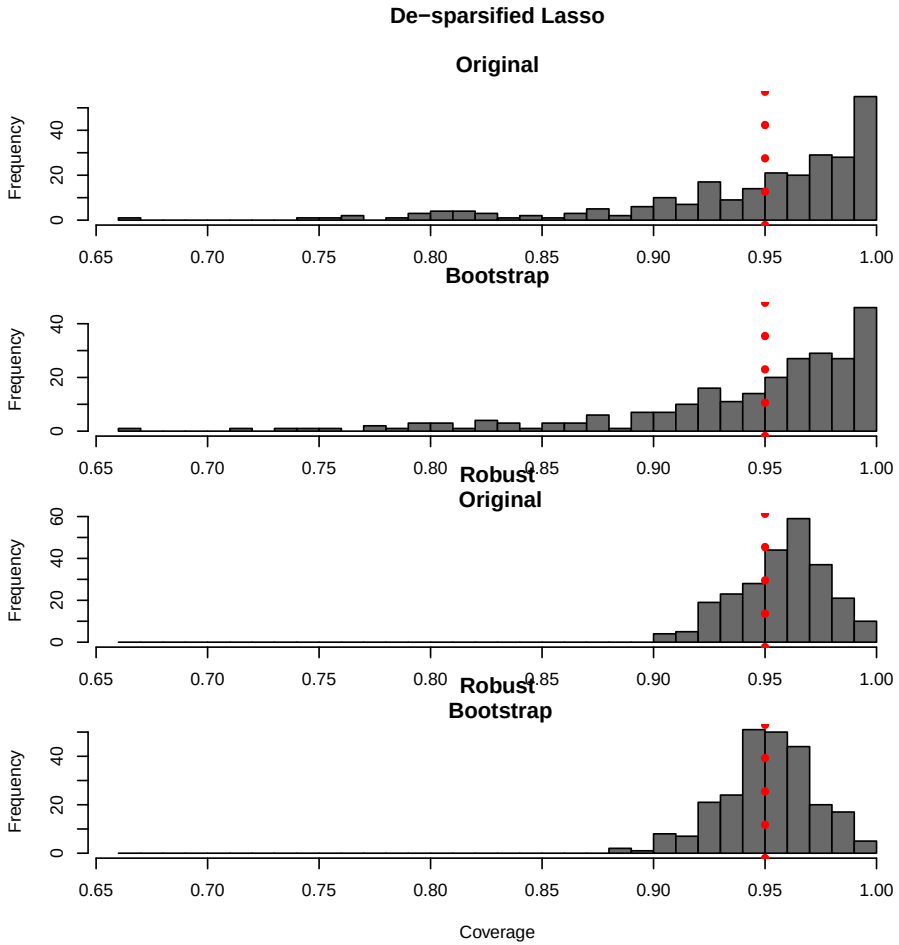


Figure: The same plot as Figure 2 but for **heteroscedastic non-Gaussian errors** and without signal. The robust standard error estimation clearly outperforms the non-robust version. There seems to be hardly any difference between the bootstrap and the original estimator after choosing the standard error estimation.

Summary

- Bootstrapping the de-biased Lasso turns out to improve the coverage of confidence intervals without increasing the confidence interval lengths
- The use of the conservative RLDPE (Zhang and Zhang, 2014) is not necessary: the bootstrap achieves reliable coverage, while for the original de-biased Lasso/LDPE, the RLDPE seems worthwhile to achieve reasonable coverage while paying a price in terms of efficiency
- Bootstrapping only the linearized part of the de-biased estimator as proposed by Zhang and Cheng (2016) is clearly sub-ideal in comparison to bootstrapping the entire estimator and using the plug-in principle as advocated here
- For multiple testing, the bootstrapped estimator had familywise error rates that were closer to the target level while Bonferroni-Holm adjustment is too conservative.
- The robust standard error turned out to be critical when dealing with heteroscedastic errors.

Multivariate inference

- $\boldsymbol{\theta} \in \mathbb{R}^d$
- The (minimum) Fisher information $\boldsymbol{F} = \boldsymbol{F}_{\boldsymbol{\theta}}$ for the estimation of $\boldsymbol{\theta}$
- $\widehat{\boldsymbol{\theta}}_n = n^{-1} \sum_{i=1}^n \boldsymbol{\psi}(\text{data}_i) + n^{-1/2}\text{Rem}$
- $n^{1/2}\left\{ n^{-1} \sum_{i=1}^n \boldsymbol{\psi}(\text{data}_i) - \boldsymbol{\theta} \right\} \approx N(0, \boldsymbol{F}^{-1}) \in \mathbb{R}^d$
- Asymptotic χ^2-confidence region: $\text{Rem} \ll \sqrt{d} \asymp \|N(0, \boldsymbol{F}^{-1})\|_2$
- Efficient inference: $\text{Rem} = o(1)$
  - Testing $H_0 : \boldsymbol{\theta} = 0$
  - $\chi_d \approx \sqrt{d} + N(0, 1/2)$
- Inference is much harder when $p \gg n \gg d \to \infty$

Background: Inference about a complex effect

- Additive regression
- Varying coefficient models with longitudinal data
- Nonparametric graphical models

Fundamental problem: Inference about a large group of coefficients

$$\boldsymbol{\beta}_G, \quad |G| \to \infty,$$

in the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n), \ \boldsymbol{\beta} \in \mathbb{R}^p.$$

Direct extension from univariate bias correction

- We have

$$\widehat{\beta}_j = \beta_j + \boldsymbol{w}_j^T \boldsymbol{\varepsilon} + \mathrm{Rem}_j/\sqrt{n} \quad \mathrm{Rem}_j \asymp (s \log p)/\sqrt{n}$$

- This yields asymptotic optimality in the sense of

$$\widehat{\beta}_G = \beta_G + (\boldsymbol{w}_j, j \in G)^T \boldsymbol{\varepsilon} + \mathrm{Rem}_G/\sqrt{n},$$

- For χ^2-confidence regions, the sample size requirement is still

$$(s \log p) \ll \sqrt{n}$$

- However, for the approximate χ^2 test, direct application of the univariate bound with Cauchy-Schwarz leads to an undesirable requirement

$$\|\mathrm{Rem}_G\|_2 \lesssim (s \log p)\sqrt{|G|/n} \to 0$$

Debiasing under group sparsity

- For approximate χ^2-confidence regions, it is sufficient to have

$$\frac{s + g \log M}{\sqrt{|G|n}} \to 0$$

  where $g = \#\{j : \beta_{G_j} \neq 0\}$, $s = \sum_{\beta_{G_j} \neq 0} |G_j|$, $M = \#$ of groups

- For approximate χ^2-tests, it is sufficient to have

$$\frac{s + g \log M}{\sqrt{n}} \to 0$$

- Larger $|G|$ the better under proper assumptions
- Twin benefits of group sparsity: factor $\sqrt{|G|}$; $(s + g \log M)$ vs. $s \log p$

Working Assumptions

- Group structure:

$$\cup_{j=1}^{M} G_j = \{1, \ldots, p\}, \quad G_j \cap G_k = \emptyset$$

- Group sparsity:

$$\mathrm{supp}(\beta^*) \subset G_{S^*}, \ |S^*| \leq g, \ |G_{S^*}| \leq s$$

- The existence of estimators satisfying

$$\left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \omega_j \left\| \boldsymbol{X}_{G_j} \widehat{\beta}_{G_j}^{(init)} - \boldsymbol{X}_{G_j} \beta_{G_j}^* \right\|_2 \lesssim \left(\frac{s + g \log M}{n} \right)$$

  with $\omega_j \approx \sqrt{(|G_j| + 2 \log M)/n}$

- Group Lasso: Yuan-Lin (06), Bach (08), Koltchinskii and Yuan (08), Obozinski et al (08), Huang-Zhang (10), Lounici et al (11)

Bias correction

- Model:

$$\boldsymbol{Y} = \boldsymbol{X}_G \boldsymbol{\beta}_G + \sum_{G_k \not\subseteq G} \boldsymbol{X}_{G_k \setminus G} \boldsymbol{\beta}_{G_k \setminus G} + \boldsymbol{\varepsilon}$$

- Bias correction with a projection $\boldsymbol{P}_G$:

$$\widehat{\boldsymbol{\beta}}_G = (\boldsymbol{P}_G \boldsymbol{X}_G)^\dagger \boldsymbol{P}_G \Big(\boldsymbol{Y} - \sum_{G_k \not\subseteq G} \boldsymbol{X}_{G_k \setminus G} \widehat{\boldsymbol{\beta}}^{(init)}_{G_k \setminus G} \Big)$$

- Error decomposition for confidence region:

$$\boldsymbol{P}_G(\boldsymbol{X}_G \widehat{\boldsymbol{\beta}}_G - \boldsymbol{X}_G \boldsymbol{\beta}^*_G) = \boldsymbol{P}_G \boldsymbol{\varepsilon} - \sum_{G_k \not\subseteq G} \boldsymbol{P}_G \boldsymbol{X}_{G_k \setminus G} \Big(\widehat{\boldsymbol{\beta}}^{(init)}_{G_k \setminus G} - \boldsymbol{\beta}^*_{G_k \setminus G} \Big)$$

- The LS property $\boldsymbol{P}_G \boldsymbol{X}_{G_k \setminus G} = \boldsymbol{0}$ leads to

$$\| \boldsymbol{P}_G(\boldsymbol{X}_G \widehat{\boldsymbol{\beta}}_G - \boldsymbol{X}_G \boldsymbol{\beta}^*_G) \|_2^2 / \sigma^2 \sim \chi^2_{\mathrm{rank}(\boldsymbol{P}_G)}$$

Relaxed LS property

- Error decomposition for confidence region:

$$\boldsymbol{P}_G(\boldsymbol{X}_G\widehat{\boldsymbol{\beta}}_G - \boldsymbol{X}_G\boldsymbol{\beta}^*_G) = \boldsymbol{P}_G\boldsymbol{\varepsilon} - \mathrm{Rem}_G$$
$$\mathrm{Rem}_G = \sum_{G_k \not\subseteq G} \boldsymbol{P}_G \boldsymbol{X}_{G_k \setminus G}\left(\widehat{\boldsymbol{\beta}}^{(init)}_{G_k \setminus G} - \boldsymbol{\beta}^*_{G_k \setminus G}\right)$$

- Test statistic for $H_0 : \boldsymbol{X}_G\boldsymbol{\beta}^*_G = 0$:

$$T_G = \|\boldsymbol{P}_G\boldsymbol{X}_G\widehat{\boldsymbol{\beta}}_G\|_2/\widehat{\sigma}$$

- Upper bound for the residual bias after bias correction:

$$\|\mathrm{Rem}_G\| \leq \sum_{G_k \not\subseteq G} \|\boldsymbol{P}_G\boldsymbol{Q}_{G_k \setminus G}\|_S \left\|\boldsymbol{X}_{G_k \setminus G}\left(\widehat{\boldsymbol{\beta}}^{(init)}_{G_k \setminus G} - \boldsymbol{\beta}^*_{G_k \setminus G}\right)\right\|_2$$

- Sufficient condition for approximate χ^2-test:

$$\left|T_G - \frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma}\right| \leq \frac{\|\mathrm{Rem}_G\|_2}{\widehat{\sigma}} + \left|\frac{\sigma}{\widehat{\sigma}} - 1\right| \frac{\|\boldsymbol{P}_G\boldsymbol{\varepsilon}\|_2}{\sigma} = o(1)$$

- Sufficient condition for approximate χ^2-test:

$$\left| T_G - \frac{\|\boldsymbol{P}_G \boldsymbol{\varepsilon}\|_2}{\sigma} \right| \leq \frac{\|\mathrm{Rem}_G\|_2}{\widehat{\sigma}} + \left| \frac{\sigma}{\widehat{\sigma}} - 1 \right| \frac{\|\boldsymbol{P}_G \boldsymbol{\varepsilon}\|_2}{\sigma} = o(1)$$

iff

$$\|\mathrm{Rem}_G\|_2 = o(1), \quad \left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| = \frac{o(1)}{\sqrt{|G|}}$$

- Upper bound for the residual bias

$$\|\mathrm{Rem}_G\| \leq \sum_{G_k \not\subseteq G} \|\boldsymbol{P}_G \boldsymbol{Q}_{G_k \setminus G}\|_S \left\| \boldsymbol{X}_{G_k \setminus G} \left(\widehat{\boldsymbol{\beta}}^{(init)}_{G_k \setminus G} - \boldsymbol{\beta}^*_{G_k \setminus G} \right) \right\|_2$$

- Working Assumption:

$$\left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \omega_j \left\| \boldsymbol{X}_{G_j} \widehat{\boldsymbol{\beta}}^{(init)}_{G_j} - \boldsymbol{X}_{G_j} \boldsymbol{\beta}^*_{G_j} \right\|_2 \lesssim \left(\frac{s + g \log M}{n} \right)$$

Suppose

$$\|\boldsymbol{X}_{G_k \setminus G} \boldsymbol{b}_{G_k \setminus G}\|_2 \leq M_k \|\boldsymbol{X}_{G_k} \boldsymbol{b}_{G_k}\|_2$$
$$\|\boldsymbol{P}_G \boldsymbol{Q}_{G_k \setminus G}\|_S \leq \omega_k'$$

for all $\boldsymbol{b}_{G_k}$ and all k. Suppose that the Working Assumptions hold with

$$\frac{|G|}{n} + \frac{s + g \log M}{n^{1/2}} \left(\frac{|G|^{1/2}}{n^{1/2}} + \max_{G_k \not\subseteq G} M_k \frac{\omega_k'}{\omega_k} \right) = o(1)$$

Then,

$$\boldsymbol{P}_G \boldsymbol{X}_G (\widehat{\beta}_G - \beta_G^*) = \boldsymbol{P}\boldsymbol{\varepsilon} + o(1)$$

and

$$\left| T_G - \frac{\boldsymbol{P}_G \boldsymbol{\varepsilon}}{\sigma} \right| = o(1)$$

Recall that $\omega_k \approx \sqrt{(|G_k| + 2 \log M)/n}$

Finding $\boldsymbol{P}_G$:

- An optimization scheme

$$\boldsymbol{P}_G = \arg\min_{\boldsymbol{P}} \left\{ \|\boldsymbol{P}\boldsymbol{Q}_G^{\perp}\|_S : \boldsymbol{P} = \boldsymbol{P}^T = \boldsymbol{P}^2, \|\boldsymbol{P}_G \boldsymbol{Q}_{G_k \setminus G}\|_S \le \omega_k' \right\}$$

- An extension of an optimization scheme discussed in (Z-Zhang, 11)
- The conclusions in the previous slide hold for all feasible solutions
- Feasibility for sub-Gaussian designs with $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{X}^T\boldsymbol{X}/n$:

$$\begin{array}{l} \boldsymbol{X}_G = \boldsymbol{X}_{-G}\boldsymbol{\Gamma}_{-G,G} + \boldsymbol{Z}_G^o \\ \boldsymbol{P}_G^o = \text{projection to the column space of } \boldsymbol{Z}_G^o \\ \omega_k' = \xi\sqrt{|G|/n + |G_k \setminus G|/n + (\log M)/n} \end{array}$$

$$\begin{array}{l} \mathrm{rank}(\boldsymbol{P}_G^o \boldsymbol{Q}_G) = |G|, \ \|\boldsymbol{P}_G \boldsymbol{Q}_G^{\perp}\|_S \le \sqrt{1 - a_n^2} \\ a_n \approx \lambda_{\min}\left(\boldsymbol{\Sigma}_{G,G}^{-1/2} (\boldsymbol{\Sigma}^{-1})_{G,G} \boldsymbol{\Sigma}_{G,G}^{-1/2}\right) \end{array}$$

Finding a feasible solution

- Recall that our theory holds for all feasible solutions of the following:

$$\boldsymbol{P}_G = \arg\min_{\boldsymbol{P}} \left\{ \|\boldsymbol{P}\boldsymbol{Q}_G^{\perp}\|_S : \boldsymbol{P} = \boldsymbol{P}^T = \boldsymbol{P}^2, \|\boldsymbol{P}_G \boldsymbol{Q}_{G_k \setminus G}\|_S \leq \omega_k' \right\}$$

  with $\omega_k' \asymp \sqrt{|G|/n + |G_k \setminus G|/n + (\log M)/n}$

- Finding a feasible solution

$$\widehat{\Gamma}_{-G,G} = \arg\min \left\{ \frac{1}{2n} \|\boldsymbol{X}_G - \boldsymbol{X}_{-G}\boldsymbol{\Gamma}_{-G,G}\|_F + R(\boldsymbol{\Gamma}_{-G,G}) \right\}$$

  - Duality:

$$R(\boldsymbol{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi \omega_k''}{n^{1/2}} \|\boldsymbol{X}_{G_k \setminus G} \boldsymbol{\Gamma}_{G_k \setminus G, G}\|_N$$

  - Group Lasso:

$$R(\boldsymbol{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi \omega_k''}{n^{1/2}} \|\boldsymbol{X}_{G_k \setminus G} \boldsymbol{\Gamma}_{G_k \setminus G, G}\|_F$$

Working Assumptions

- Group structure:

$$\cup_{j=1}^{M} G_j = \{1, \ldots, p\}, \quad G_j \cap G_k = \emptyset$$

- Group sparsity:

$$\text{supp}(\beta^*) \subset G_{S^*}, \ |S^*| \leq g, \ |G_{S^*}| \leq s$$

- Initial estimator

$$\left| \frac{\widehat{\sigma}}{\sigma^*} - 1 \right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \omega_j \left\| \boldsymbol{X}_{G_j} \widehat{\beta}_{G_j}^{(init)} - \boldsymbol{X}_{G_j} \beta_{G_j}^* \right\|_2 \lesssim \left(\frac{s + g \log M}{n} \right)$$

with $\omega_j \approx \sqrt{(|G_j| + 2 \log M)/n}$

- Group Lasso:

$$\widehat{\beta} = \arg\min_{\boldsymbol{b}} \left\{ \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2n} + \sigma \sum_{j=1}^{M} \omega_j \|\boldsymbol{b}_{G_j}\|_2 \right\}$$

- Scaled group Lasso:

$$\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\} = \arg\min_{\boldsymbol{b},\sigma} \left\{ \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sum_{j=1}^{M} \omega_j \|\boldsymbol{b}_{G_j}\|_2 \right\}$$

- Extension from the scaled Lasso: Städler et al (10), Antoniadis (10), Sun-Z (10,12), Belloni et al (11)
- Sign-restricted cone invertibility factor;

$$SCIF = \inf_{\boldsymbol{u} \in \mathscr{C}_-} \frac{\max_j \omega_j^{-1} \|\boldsymbol{X}_{G_j} \boldsymbol{X}\boldsymbol{u}\|_2 \sum_{j \in S^*} \omega_j^2}{n \sum_{j \in S^*} \omega_j \|\boldsymbol{u}_{G_j}\|_2}$$

with

$$\mathscr{C}_- = \{\boldsymbol{u} : \quad \sum_{j \notin S^*} \omega_j \|\boldsymbol{u}_{G_j}\|_2 \leq \xi \sum_{j \in S^*} \omega_j \|\boldsymbol{u}_{G_j}\|_2 \neq 0, \quad \boldsymbol{u}_{G_j}^T \boldsymbol{X}_{G_j}^T \boldsymbol{X}\boldsymbol{u} \leq 0 \quad \forall j \notin S^*\}$$

Suppose $\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S \leq 1$ and $1/SCIF = O(1)$. Suppose

$$\omega_j = A\sqrt{(|G_j| + 2\log M)/n}, \quad A > 1$$

Then,

$$\left|\frac{\widehat{\sigma}}{\sigma^*} - 1\right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \omega_j \left\| \boldsymbol{X}_{G_j} \widehat{\boldsymbol{\beta}}_{G_j}^{(init)} - \boldsymbol{X}_{G_j} \boldsymbol{\beta}_{G_j}^* \right\|_2 \lesssim \left(\frac{s + g\log M}{n} \right)$$

Thanks!