Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models

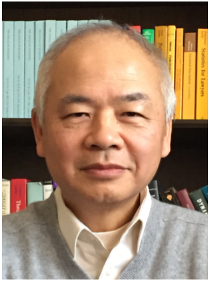
Author
Cun-Hui Zhang and Stephanie S. Zhang

Group 11
Yuanchen Qian, Caiwei Zhang, Lei Liu, Huihang Liu

University of Science and Technology of China

2018.12.12

About author



CunHui Zhang
Distinguished Professor, Ph.D. in Statistics,
Columbia University

High dimentional data
Empirical Bayes
Functional MRI
Network data
Semiparametric and nonparametric methods
Survival analysis, statistical inference
Probability theory

Outline

- Introduction
- Methodology
- Important theoretical results
- Simulations
- Discussion

Section 1

Introduction

Background

In high-dimensional statistics, much work has been made on consistency for prediction, estimation of high-dimensional objects or variable selection.

Regularized linear regression

- ℓ_1 regularized methods
- non-convex penalized methods
- greedy methods
- screening methods ...

Related work

Some related works have concerned with statistics inference:

- Knight and Fu(2000): Lasso type estimators cannot obtain a proper asymptotics distribution of unknown cofficients, even in low-dim situation.
- Leeb and Potscher(2006): Consistent estimation of the distribution of the least squares estimator after model selection is impossible.
- Berk et.al(2010); Laber and Murphy(2011): Conservative statistical inference after model selection may not yield accurate confidence regions or p-values when p is large.

Uniform signal strength condition

Existing variable selection approaches based on selection consistency theory typically requires a uniform signal strength condition:

$$\min_{\beta_j \neq 0} |\beta_j| \geq C\sigma\sqrt{(2/n)\log p},\ C > 1/2, \tag{1}$$

Advantages of OLS in low-dim

Linear model in low dimensions ($p < n$):

$$y = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I).$$

We get the estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y$ with explict form of covariance structure as following

$$cov(c'\hat{\boldsymbol{\beta}}, d'\hat{\boldsymbol{\beta}}) = \sigma^2 c'(X'X)^{-}d.$$

and the confidence interval, page 129.

Section 2

Methodology

Model setting

- Considering the following linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n\times p}$ is a design matrix with columns $\mathbf{x}_j$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \boldsymbol{\beta}_p)^T$ is a vector of unknown regression coefficients.

  - Standardize the design to $\|\mathbf{x}_j\|_2^2 = n$.
  - The design matrix $\mathbf{X}$ is assumed to be deterministic.

Least squares estimator

- In classical theory of linear models, the least squares estimator of an estimable regression coefficient β_j can be written as

$$\hat{\beta}_j^{(lse)} := (\mathrm{x}_j^{\perp})^T \mathrm{y} / (\mathrm{x}_j^{\perp})^T \mathrm{x}_j, \tag{3}$$

  where $\mathrm{x}_j^{\perp}$ is a projection of x_j to the orthogonal complement of column space of $\mathrm{X}_{-j} = (\mathrm{x}_k, k \neq j)$.
- The $\mathrm{x}_j^{\perp}$ is defined by
  - $(\mathrm{x}_j^{\perp})^T \mathrm{x}_k = 0$, for $\forall k \neq j$.

Problems

- In high dimensional situation $p > n$, $rank(\mathrm{X}_{-j}) = n$ for all j.
  - $\mathrm{x}_j^{\perp} = 0$.
  - $\hat{\beta}_j^{(lse)}$ is undefined.
- We want to preserve the properties of the least squares estimator.
  - The covariance structure of the least squares estimator:

$$\mathrm{cov}(\hat{\beta}_j^{(lse)}, \hat{\beta}_k^{(lse)}) = \sigma^2 \frac{(\mathrm{x}_j^{\perp})^T \mathrm{x}_k^{\perp}}{\|\mathrm{x}_j\|_2^2 \|\mathrm{x}_k\|_2^2} \tag{4}$$

- Motivation of LDPE:
  - Replace $\mathrm{x}_j^{\perp}$ with z_j.
  - Relaxing the constraint $\mathrm{z}_j^T \mathrm{x}_k = 0$ for $k \neq j$.

Bias-corrected linear estimators

- For any $\mathbf{z}_j$ that is not orthogonal to $\mathbf{x}_j$, the corresponding univariate linear regression estimator satisfies

$$\hat{\beta}_j^{(lin)} = \frac{\mathbf{z}_j^T \mathbf{y}}{\mathbf{z}_j^T \mathbf{x}_j} = \beta_j + \frac{\mathbf{z}_j^T \varepsilon}{\mathbf{z}_j^T \mathbf{x}_j} + \sum_{k \neq j} \frac{\mathbf{z}_j^T \mathbf{x}_k \beta_k}{\mathbf{z}_j^T \mathbf{x}_j}.$$

  - Here, $\hat{\beta}_j^{(lin)}$ has the same covariance structure with $\hat{\beta}_j^{(lse)}$.

- Note that the bias of $\hat{\beta}_j^{(lin)}$ is linear in β_k, which is unbounded. It is impossible to have $\mathbf{z}_j^T \mathbf{x}_k = 0$ for all $k \neq j \quad (\mathbf{z}_j \neq 0)$.

Low dimensional projection estimator

- Bias correction with a non-linear initial estimator $\hat{\beta}^{(init)}$:

$$\hat{\beta}_j = \hat{\beta}_j^{(lin)} - \sum_{k \neq j} \frac{z_j^T x_k \hat{\beta}_k^{(init)}}{z_j^T x_j} = \frac{z_j^T y}{z_j^T x_j} - \sum_{k \neq j} \frac{z_j^T x_k \hat{\beta}_k^{(init)}}{z_j^T x_j}. \quad (5)$$

  - The estimation error of $\hat{\beta}_j$:

$$\hat{\beta}_j - \beta_j = \frac{z_j^T \varepsilon}{z_j^T x_j} + \frac{\sum_{k \neq j} z_j^T x_k (\beta_k - \hat{\beta}_k^{(init)})}{z_j^T x_j} \triangleq A + B. \quad (6)$$

  - It can be viewed as a sum of noise term and bias term.

Error analysis of LDPE(1)

- The approximation error of the LDPE (Term B) can be controlled:

$$\Big|\sum_{k\neq j} z_j^T x_k(\beta_k - \widehat{\beta}_k^{(init)})\Big| \leq \Big(\max_{k\neq j} |z_j^T x_k|\Big)\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1. \quad (7)$$

- For z_j, define

$$\eta_j = \max_{k\neq j} |z_j^T x_k|/\|z_j\|_2, \qquad \tau_j = \|z_j\|_2/|z_j^T x_j|. \quad (8)$$

  - Bias factor η_j: $\eta_j\|\hat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1$ controls the approximation error.
  - Noise factor τ_j: $\tau_j\sigma$ is the standard deviation of noise term.

Error analysis of LDPE (2)

- Since $z_j^T \varepsilon \sim N(0, \sigma^2 \|z_j\|_2^2)$, equation (5) yields

$$\eta_j \|\hat{\beta}^{(init)} - \beta\|_1/\sigma = o(1) \Rightarrow \tau_j^{-1}(\hat{\beta}_j - \beta_j) \approx N(0, \sigma^2). \quad (9)$$

- Confidence intervals can be constructed by condition (8) and a consistent estimator of σ.
- Need to solve:
  - Choose proper z_j.
  - Choose $\hat{\beta}^{(init)}$.

How can we choose z_j ?

- Choose z_j as the residual of lasso:

$$z_j = x_j - X_{-j}\hat{\gamma}_j,\ \hat{\gamma}_j = \arg\min_{b}\left\{\frac{\|x_j - X_{-j}b\|_2^2}{2n} + \lambda_j\|b\|_1\right\}. \quad (10)$$

- Karush-Kuhn-Tucker conditions for equation (9)

$\Rightarrow \quad |x_k^T z_j/n| \le \lambda_j$ for all $k \neq j$

$\Rightarrow \quad \eta_j \le n\lambda_j/\|z_j\|_2.$

How can we pick initial estimator of β?

- The scaled lasso is a joint convex minimization method

$$\{\hat{\beta}^{(init)}, \sigma\} = \arg\min_{\mathrm{b},\sigma} \left\{ \frac{\|\mathrm{y} - \mathrm{Xb}\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda_0 \|\mathrm{b}\|_1 \right\}. \quad (11)$$

- The scaled lasso is biased, an alternative method scaled lasso-LSE can be applied:

$$\{\hat{\beta}^{(init)}, \sigma\} = \arg\min_{\mathrm{b},\sigma} \{ \frac{\|\mathrm{y} - \mathrm{Xb}\|_2^2}{2\sigma max(n - |\hat{S}^{(scl)}|, 1)} + \frac{\sigma}{2} \} \quad (12)$$

where $\hat{S}^{(scl)}$ is the set of non-zero estimated coefficients produce by scaled lasso.

Procedure of computing $\mathbf{z}_j$

Along the Lasso path for regressing $\mathbf{x}_j$ against $\mathbf{X}_{-j}$, let

$$\begin{aligned} \gamma_j(\lambda) &= \arg\min_{\mathbf{b}} \{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1\}, \\ \mathbf{z}_j(\lambda) &= \mathbf{x}_j - \mathbf{X}_{-j}\gamma_j(\lambda), \\ \eta_j(\lambda) &= \max_{k\neq j}|\mathbf{x}_k^T\mathbf{z}_j(\lambda)|/\|\mathbf{z}_j(\lambda)\|_2, \\ \tau_j(\lambda) &= \|\mathbf{z}_j(\lambda)\|_2/|\mathbf{x}_j^T\mathbf{z}_j(\lambda)|, \end{aligned} \tag{13}$$

be the coefficient estimator γ_j, residual $\mathbf{z}_j$, the bias factor η_j, and the noise factor τ_j, as functions of λ.

Proposition 1

(a) In the Lasso path, $\|\mathrm{z}_j(\lambda)\|_2$, $\eta_j(\lambda)$, and $\sigma_j(\lambda)$ are nondecreasing functions of λ, and $\tau_j(\lambda) \leq 1/\|\mathrm{z}_j(\lambda)\|_2$. Moreover, $\gamma_j(\lambda) \neq 0$ implies $\eta_j(\lambda) = \lambda n/\|\mathrm{z}_j(\lambda)\|_2$.
(b) Let $\lambda_{univ} = \sqrt{(2/n)\log p}$. Then,

$$\sigma_j(C\lambda_{univ}) > 0 \text{ iff } \{\lambda > 0 : \eta_j(\lambda) \leq C\sqrt{2\log p}\} \neq \emptyset, \tag{14}$$

and in this case, the algorithm in Table 2 provides

$$\eta_j \leq \eta_j^* \leq (1 \vee C)\sqrt{2\log p}, \tag{15}$$

$$\tau_j \leq n^{-1/2}(1+\kappa_0)/\hat{\sigma}_j(C\lambda_{univ}). \tag{16}$$

Moreover, when $\mathrm{z}_j(0) = \mathrm{x}_j^{\perp} = 0$, $\eta_j(0+)\inf\{\|\gamma_j\|_1 : \mathrm{X}_{-j}\gamma_j = \mathrm{x}_j\} = \sqrt{n}$.
(c) Let $0 < a_0 < 1 \leq C_0 < \infty$. Suppose that for $s = a_0 n/\log p$

$$\inf_{\delta}\sup_{\boldsymbol{\beta}}\left\{\|\delta(\mathrm{X},\mathrm{y})-\boldsymbol{\beta}\|_2^2 : \mathrm{y} = \mathrm{X}\boldsymbol{\beta}, \sum_{j=1}^{p}\min(|\beta_j|/\lambda_{univ}, 1) \leq s+1\right\} \leq 2C_0 s(\log p)/n.$$

Computation of z_j

Figure 1: Computation of z_j from the Lasso (12)

Input:	$\eta_j^* = \sqrt{2\log p}$
	$\kappa_0 = 0.25$
Step 1:	Compute z_j for $\lambda \geq \lambda_*$,
	Compute η_j and τ_j for $\lambda \geq \lambda_*$ η_j^*
Step 2:	If $\eta_j(\lambda_*) \geq \eta_j^*$, return $z_j \leftarrow z_j(\lambda_*)$;
	otherwise
	$\tau_j^* \leftarrow (1+\kappa_0)\min\{\tau_j(\lambda) : \eta_j(\lambda) \geq \eta_j^*\}$
	$\lambda \leftarrow \arg\min\{\eta_j(\lambda) : \tau_j(\lambda) \geq \tau_j^*\}$
	$z_j \leftarrow z_j(\lambda)$

[1]

[1] λ_* is the smallest non-zero penalty level in lasso path.

Restricted LDPE

- The reason for using restricted lasso relaxation for $\mathbf{z}_j$.
  - The summands with larger absolute correlation $|\mathbf{x}_j^T \mathbf{x}_k / n|$ are likely to have a greater contribution to the bias due to initial estimation error $|\hat{\beta}_k^{(init)} - \beta_k|$.
- How to implement restricted LDPE(RLDPE)?
  - Force smaller $|\mathbf{z}_j^T \mathbf{x}_k / n|$ for large $|\mathbf{x}_j^T \mathbf{x}_k / n|$ with a weighted relaxation:

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{X}_{-j}\gamma_j, \quad \gamma_j = \arg\min_{\mathbf{b}} \left\{ \frac{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n} + \lambda_j \sum_{k \neq j} w_k |b_k| \right\}, \qquad (17)$$

- Simply set $w_k = 0$ for large $|\mathbf{x}_j^T \mathbf{x}_k / n|$ and $w_k = 1$ for other k in the RLDPE.

Confidence interval

- The covariance of the noise component in (5) is proportional to

$$V = (V_{jk})_{p\times p}, \quad V_{jk} = \frac{z_j^T z_k}{|z_j^T x_j||z_k^T x_k|} = \sigma^{-2} cov\left(\frac{z_j^T \varepsilon}{z_j^T x_j}, \frac{z_k^T \varepsilon}{z_k^T x_k}\right). \tag{18}$$

- For sparse vectors a with bounded $||a||_0$, an approximate $(1-\alpha)100\%$ confidence interval is

$$\left|a^T \hat{\boldsymbol{\beta}} - a^T \boldsymbol{\beta}\right| \le \hat{\sigma}\Phi^{-1}(1-\alpha/2)(a^T V a)^{1/2}, \tag{19}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ is the vector of LDPEs $\hat{\beta}_j$ in (4), Φ is the standard normal distribution function.

Section 3

Important theoretical results

Conditions

Let $\lambda_{univ} = \sqrt{(2/n)\log p}$. Suppose the model (1) holds with a vector β satisfying the following capped-ℓ_1 sparsity condition:

$$\sum_{j=1}^{p} \min\left\{|\beta_j|/(\sigma\lambda_{univ}), 1\right\} \le s. \tag{20}$$

This condition holds if β is ℓ_0 sparse with $\|\beta\|_0 \le s$ or ℓ_q sparse with $\|\beta\|_q^q/(\sigma\lambda_{univ})^q \le s$, $0 < q \le 1$. Let $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$. A generic condition we impose on the initial estimator is

$$P\left\{\|\hat{\beta}^{(init)} - \beta\|_1 \ge C_1 s\sigma^* \sqrt{(2/n)\log(p/\epsilon)}\right\} \le \epsilon \tag{21}$$

for a certain fixed constant C_1 and all $\alpha_0/p^2 \le \epsilon \le 1$, where $\alpha_0 \in (0,1)$ is a preassigned constant. We also impose a similar generic condition on an estimator $\hat{\sigma}$ for the noise level:

$$P\left\{|\hat{\sigma}/\sigma^* - 1| \ge C_2 s(2/n)\log(p/\epsilon)\right\} \le \epsilon, \forall \alpha_0/p^2 \le \epsilon \le 1, \tag{22}$$

with a fixed C_2.

Theorem 1

Let $\hat{\beta}_j$ be the LDPE with an initial estimator $\hat{\boldsymbol{\beta}}^{(init)}$. Let η_j and τ_j be the bias and noise factors in (7), $\sigma^* = \|\varepsilon\|_2/\sqrt{n}$, $\max(\epsilon'_n, \epsilon''_n) \to 0$, and $\eta^* > 0$. Suppose (20) holds with $\eta^* C_1 s\sqrt{(2/n)\log(p/\epsilon)} \le \epsilon'_n$. If $\eta_j \le \eta^*$, then

$$P\left\{\left|\tau_j^{-1}(\hat{\beta}_j - \beta_j) - \mathrm{z}_j^T\varepsilon/\|\mathrm{z}_j\|_2\right| > \sigma^*\epsilon'_n\right\} \le \epsilon. \tag{23}$$

If in addition (21) holds with $C_2 s(2/n)\log(p/\epsilon) \le \epsilon''_n$, then for all $t \ge (1+\epsilon'_n)/(1-\epsilon''_n)$,

$$P\left\{|\hat{\beta}_j - \beta_j| \ge \tau_j\hat{\sigma}t\right\} \le 2\Phi_{n-1}(-(1-\epsilon''_n)t + \epsilon'_n) + 2\epsilon, \tag{24}$$

where $\Phi_n(t)$ is the student-t distribution function with n degrees of freedom. Moreover, for the covariance matrix V in and all fixed m,

$$\lim_{n\to\infty} \inf_{\mathrm{a}\in\mathscr{A}_{n,p,m}} P\left\{\left|\mathrm{a}^T\hat{\boldsymbol{\beta}} - \mathrm{a}^T\boldsymbol{\beta}\right| \le \hat{\sigma}\Phi^{-1}(1-\alpha/2)(\mathrm{a}^T\mathrm{V}\mathrm{a})^{1/2}\right\} = 1-\alpha, \tag{25}$$

where $\Phi(t) = P\{N(0,1) \le t\}$ and $\mathscr{A}_{n,p,m} = \{\mathrm{a} : \|\mathrm{a}\|_0 \le m, \max_{j\le p}|a_j|\eta_j \le \eta^*\}$.

Remark 1

- Condition (22) establishes the joint asymptotic normality of the LDPE under condition (20) This allows us to write the LDPE as an approximate Gaussian sequence.

$$\hat{\beta}_j = \beta_j + N(0, \tau_j^2\sigma^2) + o_P(\tau_j\sigma). \tag{26}$$

- Condition (23) and (24) justify the approximate coverage probability of the resulting confidence interval.
- The uniform signal strength condition is not required for condition (20) and (21).

Simultaneous confidence interval

Theorem 2

Suppose (20) holds with $\eta^* C_1 s\sqrt{(2/n)\log(p/\epsilon)} \le \epsilon'_n$. Then,

$$P\left\{\max_{\eta_j \le \eta^*} \left|\tau_j^{-1}(\hat{\beta}_j - \beta_j) - \mathrm{z}_j^T \epsilon / \|\mathrm{z}_j\|_2\right| > \sigma^* \epsilon'_n\right\} \le \epsilon. \qquad (27)$$

If (21) also holds with $C_2 s(2/n)\log(p/\epsilon) \le \epsilon''_n$, then

$$P\left\{\max_{\eta_j \le \eta^*} |\hat{\beta}_j - \beta_j|/(\tau_j \hat{\sigma}) > t\right\} \le 2\Phi_n(-(1-\epsilon''_n)t + \epsilon'_n)\#\{j : \eta_j \le \eta^*\} + 2\epsilon. \qquad (28)$$

If, in addition to (20) and (21), $\max_{j \le p} \eta_j \le \eta^*$ and $\max(\epsilon'_n, \epsilon) \to 0$ as $\min(n,p) \to \infty$, then for fixed $\alpha \in (0,1)$ and $c_0 > 0$,

$$\liminf_{n\to\infty} P\left\{\max_{j \le p}\left|\frac{\hat{\beta}_j - \beta_j}{\tau_j(\hat{\sigma} \wedge \sigma)}\right| \le c_0 + \sqrt{2\log(p/\alpha)}\right\} \ge 1 - \alpha. \qquad (29)$$

Thresholded LDPE

- From (25), the $\hat{\beta}_j$ can be viewed as an approximate Gaussian sequence.
- The approximate Gaussian sequence is not sparse but can be thresholded. Using either the hard or the soft thresholding method:

$$\hat{\beta}_j^{(thr)} = \begin{cases} \hat{\beta}_j I(|\hat{\beta}_j| > \hat{t}_j), \\ sgn(\hat{\beta}_j)(|\hat{\beta}_j| - \hat{t}_j)^+, \end{cases} \tag{30}$$

with

$$\hat{S}^{(thr)} = \{j : |\hat{\beta}_j| > \hat{t}_j\}$$

where $\hat{t}_j \approx \hat{\sigma}\tau_j \Phi^{-1}(1 - \alpha/(2p))$ with $\alpha > 0$.

Theorem 3

Let $L_0 = \Phi^{-1}(1-\alpha/(2p))$, $\tilde{t}_j = \tau_j \sigma L_0$, and $\hat{t}_j = (1+c_n)\hat{\sigma}\tau_j L_0$ with positive constants α and c_n. Suppose condition (queshao) holds with $\eta^* C_1 s/\sqrt{n} \le \epsilon_n'$, $\max_{j\le p}\eta_j \le \eta^*$, and

$$P\left\{ \frac{(\hat{\sigma}/\sigma)\vee(\sigma/\hat{\sigma}) - 1 + \epsilon_n'\sigma^*/(\hat{\sigma}\wedge\sigma)}{1-(\hat{\sigma}/\sigma-1)_+} > c_n \right\} \le 2\epsilon. \tag{31}$$

Let $\boldsymbol{\beta}^{(thr)} = (\beta_1^{(thr)}, \ldots, \beta_p^{(thr)})^T$ be the soft thresholded LDPE with these $\hat{t}_j$. Then, there is an event Ω_n with $P\{\Omega_n^c\} \le 3\epsilon$ such where $L_n = 4/L_0^3 + 4c_n/L_0 + 12c_n^2 L_0$. Moreover, with at least probability $1-\alpha-3\epsilon$,

$$\{j : |\beta_j| > (2+2c_n)\tilde{t}_j\} \subseteq \hat{S}^{(thr)} \subseteq \{j : \beta_j \neq 0\}. \tag{32}$$

Section 4

Simulation

Setting

Set $n = 200$, $p = 3000$, and run several simulation experiments with 100 replications in each setting.

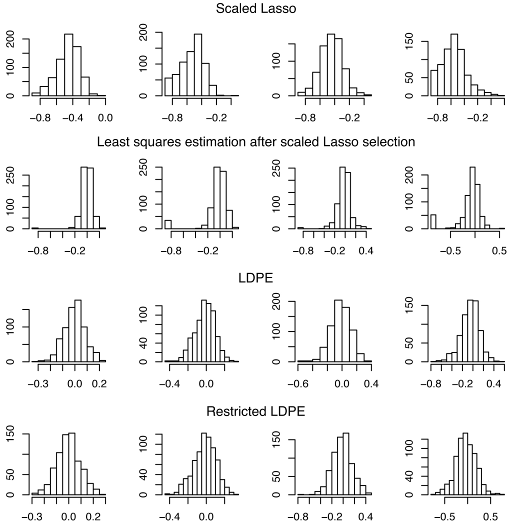
- Generate independent copy of (X, y):
  Given a particular $\rho \in (-1, 1)$, $\tilde{X} = (\tilde{x}_{ij})_{n \times p}$ has iid $N(0, \Sigma)$ rows with $\Sigma = (\rho^{|j-k|})_{p \times p}$, $x_j = \tilde{x}_j \sqrt{n}/|\tilde{x}_j|_2$, and (X, y) is as in (1) with $\sigma = 1$.
- Generate $\boldsymbol{\beta}$:
  Given a particular $\alpha \geq 1$, $\beta_j = 3\lambda_{univ}$ for $j = 1500, 1800, 2100, \ldots, 3000$, and $\beta_j = 3\lambda_{univ}/j^{\alpha}$ for all other j, where $\lambda_{univ} = \sqrt{(2/n) \log p}$.
- Set α and ρ:
  This simulation example includes four cases, labeled (A), (B), (C), and (D), respectively: $(\alpha, \rho) = (2, \frac{1}{5})$, $(1, \frac{1}{5})$, $(2, \frac{4}{5})$, and $(1, \frac{4}{5})$.

Comparison between different methods

Table 3. Summary statistics for various estimates of the maximal $\beta_j = |\beta|_\infty$: the lasso, the scaled lasso, the scaled lasso–LSE method, the oracle estimator, the LDPE and the RLDPE

Setting	*Statistic*	*Lasso*	*Scaled lasso*	*Scaled lasso–LSE*	*Oracle*	*LDPE*	*RLDPE*
A	Bias	−0.2965	−0.4605	−0.0064	−0.0045	−0.0038	−0.0028
	Standard deviation	0.0936	0.1360	0.1004	0.0730	0.0860	0.0960
	Median absolute error	0.2948	0.4519	0.0549	0.0507	0.0531	0.0627
B	Bias	−0.2998	−0.5341	−0.0476	0.0049	−0.0160	−0.0167
	Standard deviation	0.1082	0.1590	0.2032	0.0722	0.1111	0.1213
	Median absolute error	0.2994	0.5150	0.0693	0.0500	0.0705	0.0799
C	Bias	−0.3007	−0.4423	−0.0266	−0.0049	−0.0194	−0.0181
	Standard deviation	0.1207	0.1520	0.1338	0.1485	0.1358	0.1750
	Median absolute error	0.3000	0.4356	0.0657	0.0994	0.0902	0.1150
D	Bias	−0.3258	−0.5548	−0.1074	−0.0007	−0.0510	−0.0405
	Standard deviation	0.1367	0.1844	0.2442	0.1455	0.1768	0.2198
	Median absolute error	0.3319	0.5620	0.0857	0.0955	0.1112	0.1411

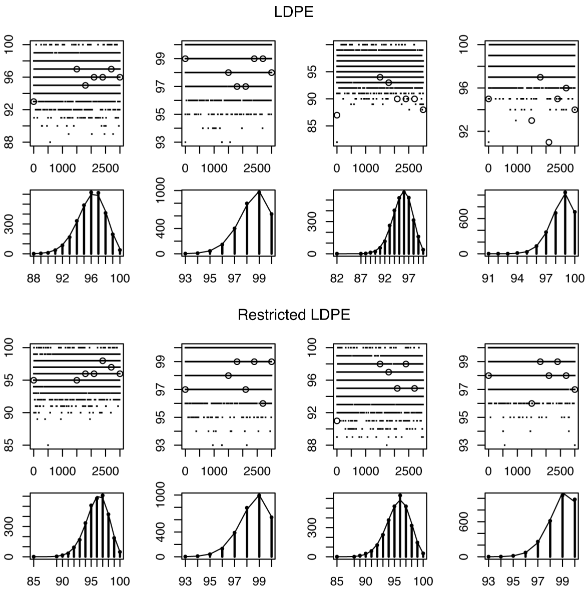
Histogram of error



Result of mean coverage probability

Figure 2: Mean coverage probability of LDPE and R-LDPE.

		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


LDPE
Restricted LDPE

Simulation 1: Summary statistics for different estimators

We set $p = 300$, $n = 50$, replication$= 100$, let Σ, $\alpha = 2$, $\rho = 1/5$ the same as Setting A.

Figure 3: Summary statistics for different estimators of the maximal $\beta_j = \|\beta\|_\infty$

	Lasso	Scaled lasso	LDPE
Bias	2.070	2.110	0.580
Standard deviation	0.046	0.042	0.181

Simulation 2: Mean coverage probability

We set $p = 300$, $n = 50$, replication$= 100$. Generating design matrix X from different Σ to compare the mean coverage probability by LDPE.

- Setting A: $\Sigma = (\rho^{|j-k|})_{p\times p}, \rho = 0.2$
- Setting B: $\Sigma = (\rho^{|j-k|})_{p\times p}, \rho = 0.8$
- Setting C: $\Sigma = (\rho)_{p\times p}, j \neq k, \rho = 0.2$
- Setting D: $\Sigma = (\rho)_{p\times p}, j \neq k, \rho = 0.8$

Figure 4: Mean coverage probability of LDPE.

	A	B	C	D
all β_j	0.919	0.934	0.910	0.913
maximal β_j	0.940	0.980	0.950	0.960

Simulation 3: Choosing initial estimator of β by ℓ_0 and scaled lasso

We set $p = 300$, $n = 50$, replication=100. $\beta_j = 3\lambda_{univ}$ for $j = 15, 18, \ldots, 30$, and $\beta_j = 3\lambda_{univ}/j^\alpha$ for other j.

- Setting A: $\Sigma = (\rho^{|j-k|})_{p\times p}$, $\rho = 0.2$
- Setting B: $\Sigma = (\rho^{|j-k|})_{p\times p}$, $\rho = 0.6$

Figure 5: Mean coverage probability of ℓ_0 and Scaled lasso.

		A	B
all β_j	ℓ_0	0.915	0.942
	Scaled lasso	0.907	0.953
maximal β_j	ℓ_0	0.930	0.980
	Scaled lasso	0.920	0.970

Summary of LDPE

Advantages

- Construct confidence intervals for regression coefficients
- Without requirement of the uniform signal strength condition
- Handle the situation with high-corelated design matrix.

Thank you !

Discussion

- Why lasso to obtain z_j? What about other methods?
- Conditions.
- Limits of Thresholded-LDPE in variable selection.