Inference for Linear Functionals in High-dimensional Linear Models

Zijian Guo

Rutgers University

Tony T. Cai

Tianxi Cai

High-dimensional linear regression

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

- Number of covariates $p \gg$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.

High-dimensional linear regression

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

- Number of covariates $p \gg$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.

Estimation of β: Basis Pursuit (Chen & Donoho, '94); Lasso (Tibshirani, '96); SCAD (Fan & Li, '01); LARS(Efron, Hastie, Johnstone & Tibshirani, '04) Elastic Net (Zou & Hastie, '05); Adaptive Lasso (Zou, '05); Dantzig Selector (Candès & Tao, '07); Lasso and Dantzig (Bickel, Ritov & Tsybakov, '09); MCP (Zhang '10); scaled Lasso (Sun & Zhang, '10); square-root Lasso (Belloni, Chernozhukov & Wang, '11); $\cdots$

High-dimensional linear regression

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

- Number of covariates $p \gg$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.

Estimation of β: Basis Pursuit (Chen & Donoho, '94); Lasso (Tibshirani, '96); SCAD (Fan & Li, '01); LARS(Efron, Hastie, Johnstone & Tibshirani, '04) Elastic Net (Zou & Hastie, '05); Adaptive Lasso (Zou, '05); Dantzig Selector (Candès & Tao, '07); Lasso and Dantzig (Bickel, Ritov & Tsybakov, '09); MCP (Zhang '10); scaled Lasso (Sun & Zhang, '10); square-root Lasso (Belloni, Chernozhukov & Wang, '11); $\cdots$

Inference for Functionals

Inference for Functionals

1. **Linear Functionals** $\eta^{\top}\beta$
   - β_1
   - $\beta_1 - \beta_2$
   - $x_{\text{new}}^{\top}\beta$

Inference for Functionals

1. **Linear Functionals** $\eta^\top \beta$
   - β_1
   - $\beta_1 - \beta_2$
   - $x_{\mathrm{new}}^\top \beta$

2. Quadratic Functionals
   - $\|\beta\|_2^2$
   - $\beta^\intercal \Sigma \beta = \mathrm{Var}(X_{i\cdot}^\intercal \beta)$
   - $\beta_G^\intercal \Sigma_{G,G} \beta_G = \mathrm{Var}(X_{i,G}^\intercal \beta_G)$

Inference for Functionals

1. **Linear Functionals** $\eta^\top \beta$
   - β_1
   - $\beta_1 - \beta_2$
   - $x_{\text{new}}^\top \beta$

2. Quadratic Functionals
   - $\|\beta\|_2^2$
   - $\beta^\intercal \Sigma \beta = \mathrm{Var}(X_{i\cdot}^\intercal \beta)$
   - $\beta_G^\intercal \Sigma_{G,G} \beta_G = \mathrm{Var}(X_{i,G}^\intercal \beta_G)$

3. ℓ_q Accuracy Functionals
   - $\|\widehat{\beta} - \beta\|_2^2$ (Accuracy assessment of $\widehat{\beta}$)
   - $\|\widehat{\beta} - \beta\|_q^q$ for $1 \le q < 2$.

Overview of talk

1. Inference for β_i: Review of De-biasing
2. Minimaxity and Adaptivity
3. Uniform Procedure for All loadings
4. Further Discussion on Optimality

CI for β_i

- **Statistics:** Zhang & Zhang '14; van de Geer, Bühlmann, Ritov & Dezeure '14; Javanmard & Montanari '14;
- **Econometrics:** Chernozhukov, Belloni & Hansen '13; Chernozhukov, Hansen & Spindler '15;

CI for β_i

- **Statistics:** Zhang & Zhang '14; van de Geer, Bühlmann, Ritov & Dezeure '14; Javanmard & Montanari '14;
- **Econometrics:** Chernozhukov, Belloni & Hansen '13; Chernozhukov, Hansen & Spindler '15;
- Main idea: **Bias correction.**

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \text{ with } \lambda \asymp \sqrt{\log p / n}\sigma$$

- De-biased Estimator:

$$\widetilde{\beta}_i = \widehat{\beta}_i + \underbrace{\widehat{u}^\top \frac{1}{n} X^\top \left(y - X\widehat{\beta}\right)}_{\text{Correction term}} \text{ with } \left(\frac{1}{n} X^\top X\right) \widehat{u} \approx e_i.$$

Construction of Projection Direction

Estimation error of $\widehat{\beta}_i$: $\widehat{\beta}_i - \beta_i = e_i^\intercal(\widehat{\beta} - \beta)$

$$\widehat{u}^\intercal \frac{1}{n} X^\intercal \left(Y - X\widehat{\beta}\right) = \widehat{u}^\intercal \widehat{\Sigma}(\beta - \widehat{\beta}) + \widehat{u}^\intercal \frac{1}{n} X^\intercal \epsilon$$

Construction of Projection Direction

Estimation error of $\widehat{\beta}_i$: $\widehat{\beta}_i - \beta_i = e_i^\intercal(\widehat{\beta} - \beta)$

$$\widehat{u}^\intercal \frac{1}{n} X^\intercal \left(Y - X\widehat{\beta}\right) = \widehat{u}^\intercal \widehat{\Sigma}(\beta - \widehat{\beta}) + \widehat{u}^\intercal \frac{1}{n} X^\intercal \epsilon$$

$$= -e_i^\intercal(\widehat{\beta} - \beta) + \underbrace{(\widehat{u}^\intercal \widehat{\Sigma} - e_i^\intercal)(\beta - \widehat{\beta})}_{\text{Remaining Bias}} + \underbrace{\widehat{u}^\intercal \frac{1}{n} X^\intercal \epsilon}_{\textit{Variance}}$$

Construction of Projection Direction

Estimation error of $\widehat{\beta}_i$: $\widehat{\beta}_i - \beta_i = e_i^\intercal(\widehat{\beta} - \beta)$

$$\widehat{u}^\intercal \frac{1}{n} X^\intercal \left(Y - X\widehat{\beta}\right) = \widehat{u}^\intercal \widehat{\Sigma}(\beta - \widehat{\beta}) + \widehat{u}^\intercal \frac{1}{n} X^\intercal \epsilon$$

$$= -e_i^\intercal(\widehat{\beta} - \beta) + \underbrace{(\widehat{u}^\intercal \widehat{\Sigma} - e_i^\intercal)(\beta - \widehat{\beta})}_{\text{Remaining Bias}} + \underbrace{\widehat{u}^\intercal \frac{1}{n} X^\intercal \epsilon}_{\textit{Variance}}$$

De-biased estimator

$$\widetilde{\beta}_i = e_i^\intercal \widehat{\beta} + \widehat{u}^\intercal \frac{1}{n} X^\intercal \left(Y - X\widehat{\beta}\right).$$

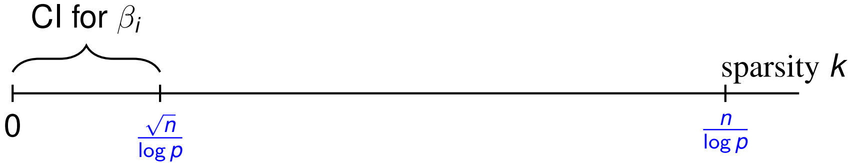
$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} \left\{ \underbrace{u^\intercal \widehat{\Sigma} u}_{\text{Variance}} : \underbrace{\left\| \widehat{\Sigma} u - e_i \right\|_\infty \leq \|e_i\|_2 \lambda_1}_{\text{Constrained Bias}} \right\}$$

Construction of CI for β_1

$$\widetilde{\beta}_i - \beta_i = \underbrace{(\widehat{u}^{\intercal}\widehat{\Sigma} - e_i^{\intercal})(\beta - \widehat{\beta})}_{\text{Remaining Bias}} + \underbrace{\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon}_{\textit{Variance}}$$

1. Variance $\sqrt{n}\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon \mid X \sim N(0, \widehat{u}^{\intercal}\widehat{\Sigma}\widehat{u})$
2. $\sqrt{n}\left|(\widehat{u}^{\intercal}\widehat{\Sigma} - e_i^{\intercal})(\beta - \widehat{\beta})\right| \leq \sqrt{n}\|\widehat{\Sigma}\widehat{u} - e_i\|_{\infty}\|\beta - \widehat{\beta}\|_1 \lesssim \frac{k \log p}{\sqrt{n}}$

Ultra-sparse case $k \ll \frac{\sqrt{n}}{\log p} \Rightarrow$ Variance dominates.



CI over $k \lesssim \frac{n}{\log p}$

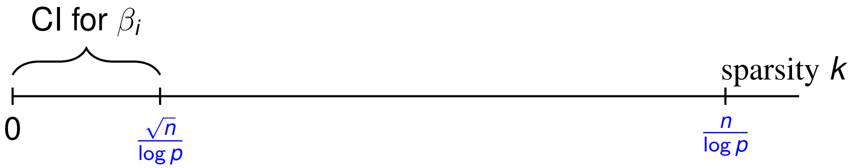
$$\mathrm{CI}_{\beta_1}(k) = \left[\widetilde{\beta}_1 - \rho(k), \quad \widetilde{\beta}_1 + \rho(k)\right],$$

$$\text{with} \;\; \rho(k) = \frac{c_\alpha}{\sqrt{n}}\hat{\sigma} + \underbrace{Ck\frac{\log p}{n}\hat{\sigma}}_{\text{Account for remaining bias}} .$$

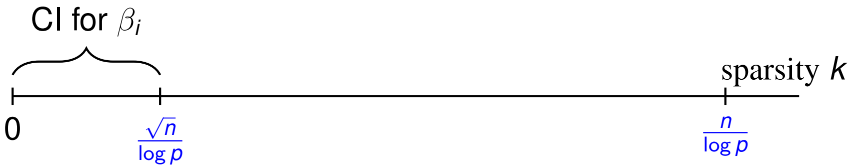
Overview of talk

1. Inference for β_i: Review of De-biasing
2. Minimaxity and Adaptivity
3. Uniform Procedure for All loadings
4. Further Discussion on Optimality

Minimaxity and Adaptivity (Cai and G., '16)



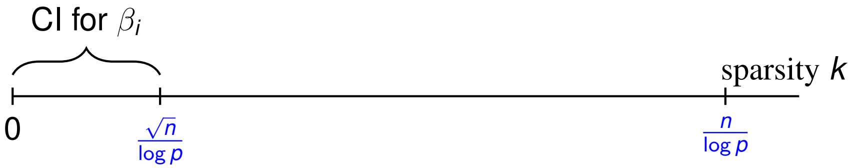
Minimaxity and Adaptivity (Cai and G., '16)



For $k \lesssim \frac{n}{\log p}$,

1. Minimax expected length of CI for β_i.
2. Possible regime to construct adaptive CI for β_i.

Minimaxity and Adaptivity (Cai and G., '16)



For $k \lesssim \frac{n}{\log p}$,

1. Minimax expected length of CI for β_i.
2. Possible regime to construct adaptive CI for β_i.

Adaptivity: without knowing the true sparsity k, construct CI as well as we know k.

Optimal expected length

- Coverage: Guaranteed coverage probability.
- Precision: As short as possible.

$$\Theta(k)=\left\{\theta=(\beta,\Sigma,\sigma): \|\beta\|_0 \leq k, \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1, 0<\sigma \leq M_2\right\}$$

Optimal expected length

- Coverage: Guaranteed coverage probability.
- Precision: As short as possible.

$$\Theta(k)=\left\{\theta=(\beta,\Sigma,\sigma):\|\beta\|_0\leq k,\ \frac{1}{M_1}\leq\lambda_{\min}(\Sigma)\leq\lambda_{\max}(\Sigma)\leq M_1,0<\sigma\leq M_2\right\}$$

- For $0<\alpha<1$, CI has coverage for β_1 over $\Theta(k)$ if

$$\inf_{\theta\in\Theta(k)}\mathbf{P}_\theta(\beta_1\in\mathrm{CI})\geq 1-\alpha.$$

- For given k, the optimal length over $\Theta(k)$,

$$\mathcal{L}^*_\alpha\left(\Theta(k)\right)=\inf_{\substack{\text{CI having coverage}\\ \text{for }\beta_1\text{ over }\Theta(k)}}\underbrace{\sup_{\theta\in\Theta(k)}\mathbf{E}_\theta\mathbf{L}\left(\mathrm{CI}\right)}_{\text{Precision}}.$$

Optimal expected length

Theorem 1(Cai and G., '16)

For $k \leq c \min\{p^{\gamma}, \frac{n}{\log p}\}$ with $0 \leq \gamma < \frac{1}{2}$,

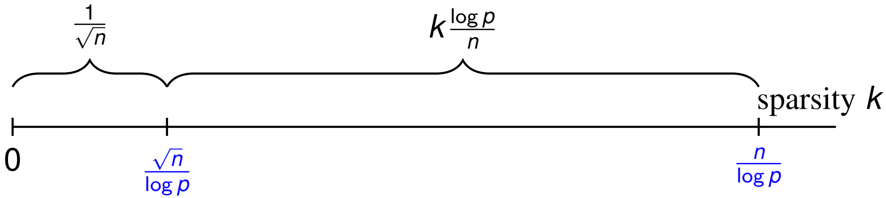
$$L_{\alpha}^{*}\left(\Theta\left(k\right)\right) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}.$$

Optimal expected length

Theorem 1(Cai and G., '16)

For $k \leq c \min\{p^{\gamma}, \frac{n}{\log p}\}$ with $0 \leq \gamma < \frac{1}{2}$,

$$L_{\alpha}^{*}\left(\Theta\left(k\right)\right) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}.$$

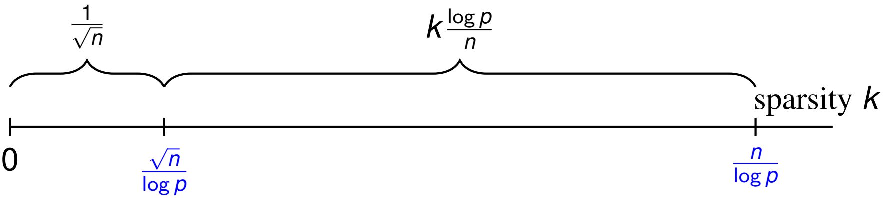


Optimal expected length

Theorem 1 (Cai and G., '16)

For $k \le c \min\{p^{\gamma}, \frac{n}{\log p}\}$ with $0 \le \gamma < \frac{1}{2}$,

$$L^*_{\alpha}(\Theta(k)) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}.$$



CIs of length $\frac{1}{\sqrt{n}}$: NO coverage for $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$.

Adaptive Procedures?

$$\text{Length of CI:} \quad \rho(k) = \frac{c_\alpha}{\sqrt{n}}\hat{\sigma} + Ck\frac{\log p}{n}\hat{\sigma}.$$

Adaptivity $\Longrightarrow$

Without knowing k, possible to construct CIs as well as known k?

Adaptive procedures?

k(unknown true sparsity) $\leq k_u$(known upper bound), $\Theta(k) \subset \Theta(k_u)$

Adaptive procedures?

k(unknown true sparsity) $\leq k_u$(known upper bound), $\Theta(k) \subset \Theta(k_u)$

Is it possible to construct CIs for β_1

1. coverage over $\Theta(k_u)$

Adaptive procedures?

k(unknown true sparsity) $\leq k_u$(known upper bound), $\Theta(k) \subset \Theta(k_u)$

Is it possible to construct CIs for β_1

1. coverage over $\Theta(k_u)$
2. for any $\theta \in \Theta(k)$,

$$\mathbf{E}_\theta \mathbf{L}\left(\mathrm{CI}\right) \lesssim \frac{1}{\sqrt{n}} + k\frac{\log p}{n}?$$

Lack of adaptivity

Theorem 2(Cai and G., '16)

For any $\theta = (\beta, \mathrm{I}, \sigma) \in \Theta(k)$ and $k \leq k_u \leq \sqrt{p}$,

$$\inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \mathbb{E}_\theta L(\mathrm{CI}) \geq c \left(\frac{1}{\sqrt{n}} + k_u \frac{\log p}{n} \right) \sigma.$$

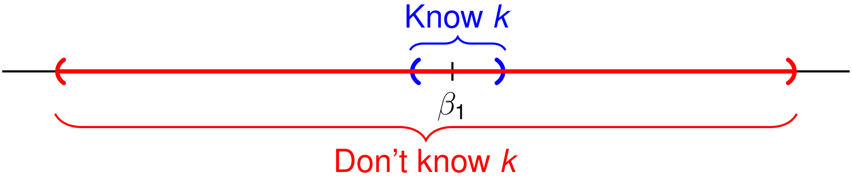
Lack of adaptivity

Theorem 2(Cai and G., '16)

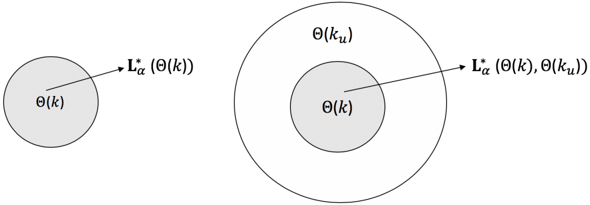
For any $\theta = (\beta, \mathrm{I}, \sigma) \in \Theta(k)$ and $k \leq k_u \leq \sqrt{p}$,

$$\inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \mathbb{E}_\theta L(\mathrm{CI}) \geq c\left(\frac{1}{\sqrt{n}} + k_u \frac{\log p}{n}\right)\sigma.$$

For $\frac{\sqrt{n}}{\log p} \lesssim k_u \lesssim \frac{n}{\log p}$,

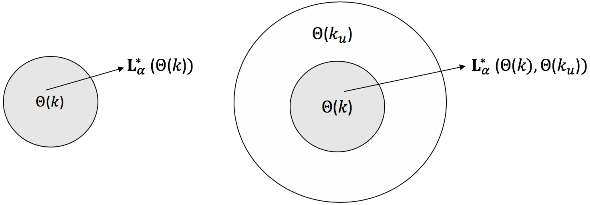


General Adaptation Benchmark



$$L_\alpha^*(\Theta(k), \Theta(k_u)) = \inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L\,(\mathrm{CI})$$

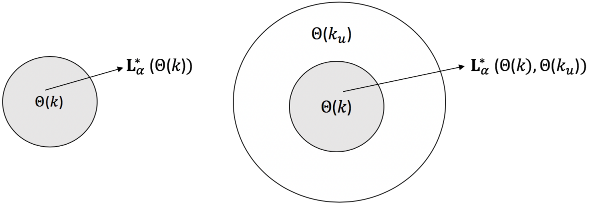
General Adaptation Benchmark



$$L_{\alpha}^{*}(\Theta(k),\Theta(k_u)) = \inf_{\substack{\text{CI having coverage}\\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \sup_{\theta\in\Theta(k)} \mathbb{E}_{\theta} L\,(\mathrm{CI})$$

$$L_{\alpha}^{*}(\Theta(k),\Theta(k_u)) \geq \inf_{\substack{\text{CI having coverage}\\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \mathbb{E}_{\theta} L\,(\mathrm{CI})$$

General Adaptation Benchmark



$$L_\alpha^*(\Theta(k), \Theta(k_u)) = \inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\text{CI})$$

$$L_\alpha^*(\Theta(k), \Theta(k_u)) \geq \inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \mathbb{E}_\theta L(\text{CI})$$

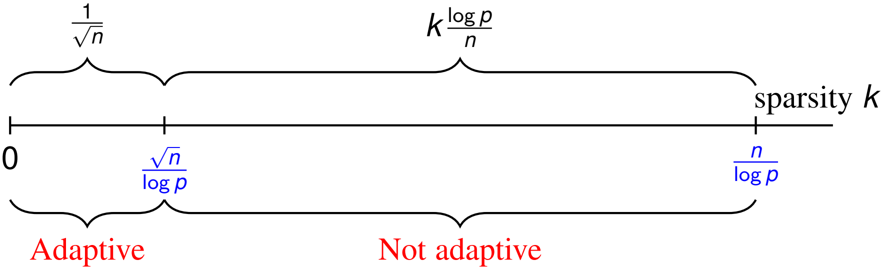
$L_\alpha^*(\Theta(k), \Theta(k_u)) \gg L_\alpha^*(\Theta(k)) \Longrightarrow$ Impossible adaptive CI.

Summary of CI for β_1

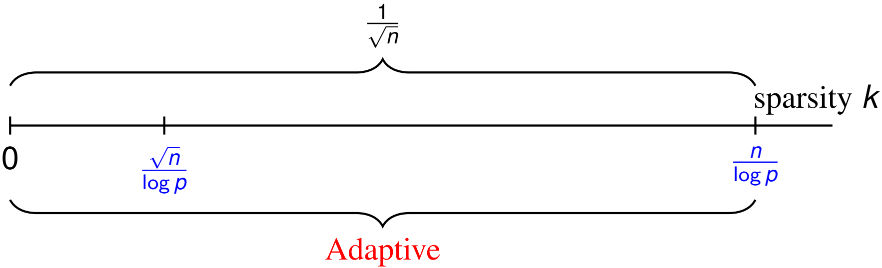
- First constructed CI for β_1 over $k \lesssim \frac{n}{\log p}$.

Summary of CI for β_1

- First constructed CI for β_1 over $k \lesssim \frac{n}{\log p}$.

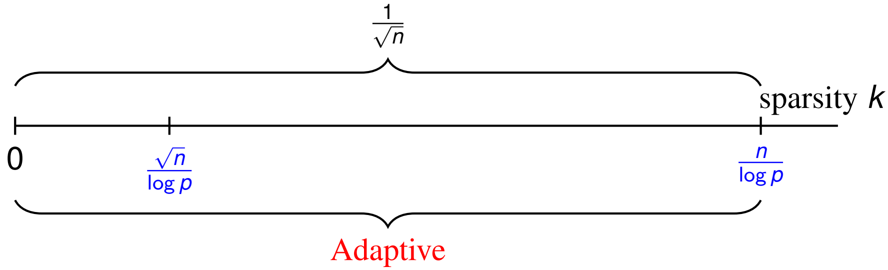


Comparison with known Σ



- CI for β_1 was constructed in Javanmard & Montanari '15.

Comparison with known Σ



- CI for β_1 was constructed in Javanmard & Montanari '15.
- Technical difference: unknown covariance structure between X_{i1} and $X_{i2}, \cdots, X_{ip}$.

Four scenarios

Table: Confidence Intervals for $\eta^\intercal \beta$

	Known Σ	Unknown Σ
Sparse Loading η (e.g., β_1)	✓	✓
Dense Loading η (e.g., $\sum_{i=1}^{p} \beta_i$)	?	?

Exact Loading: Sparse and Dense

We calibrate the sparsity levels as

$$k = p^{\gamma}, \quad k_u = p^{\gamma_u} \quad \text{for} \quad 0 \leq \gamma < \gamma_u \leq \frac{1}{2},$$

We consider exact loadings.

$$\max_{\{i:\eta_i \neq 0\}} |\eta_i| \,/ \min_{\{i:\eta_i \neq 0\}} |\eta_i| \leq C_0,$$

$$\|\eta\|_0 = p^{\gamma_\eta} \quad \text{for} \quad 0 \leq \gamma_\eta \leq 1.$$

Exact Loading: Sparse and Dense

We calibrate the sparsity levels as

$$k = p^{\gamma}, \quad k_u = p^{\gamma_u} \quad \text{for} \quad 0 \leq \gamma < \gamma_u \leq \frac{1}{2},$$

We consider exact loadings.

$$\max_{\{i:\eta_i \neq 0\}} |\eta_i| \,/\, \min_{\{i:\eta_i \neq 0\}} |\eta_i| \leq C_0,$$

$$\|\eta\|_0 = p^{\gamma_\eta} \quad \text{for} \quad 0 \leq \gamma_\eta \leq 1.$$

(E1) $x_{\rm new}$ is called *exact sparse* if $\gamma_\eta \leq \gamma$;

(E2) $x_{\rm new}$ is called *exact dense* if $\gamma_\eta > 2\gamma$;

CI for $\sum_{i=1}^{p} \beta_i$ (Cai and G., '16)

1. Centering at Lasso estimator

$$\mathrm{CI}_{\sum \beta_i}(k) = \left[\sum_{i=1}^{p} \widehat{\beta}_i - Ck\sqrt{\frac{\log p}{n}}\hat{\sigma}, \quad \sum_{i=1}^{p} \widehat{\beta}_i + Ck\sqrt{\frac{\log p}{n}}\hat{\sigma}\right],$$

- NOT using de-biased estimator: Inflation of variance!

CI for $\sum_{i=1}^{p} \beta_i$ (Cai and G., '16)

1. Centering at Lasso estimator

$$\mathrm{CI}_{\sum \beta_i}(k) = \left[\sum_{i=1}^{p} \widehat{\beta}_i - Ck\sqrt{\frac{\log p}{n}}\hat{\sigma}, \quad \sum_{i=1}^{p} \widehat{\beta}_i + Ck\sqrt{\frac{\log p}{n}}\hat{\sigma} \right],$$

   - NOT using de-biased estimator: Inflation of variance!

2. $\mathrm{CI}_{\sum \beta_i}(k)$ achieves optimal expected length $k\sqrt{\frac{\log p}{n}}$.
3. NOT possible to construct adaptive CI.
   - Without knowing k, CI must be longer than $k\sqrt{\frac{\log p}{n}}$.

CI for $\sum_{i=1}^{p} \beta_i$ (Cai and G., '16)

1. Centering at Lasso estimator

$$\mathrm{CI}_{\sum \beta_i}(k) = \left[\sum_{i=1}^{p} \widehat{\beta}_i - Ck\sqrt{\frac{\log p}{n}}\hat{\sigma}, \quad \sum_{i=1}^{p} \widehat{\beta}_i + Ck\sqrt{\frac{\log p}{n}}\hat{\sigma} \right],$$

   - NOT using de-biased estimator: Inflation of variance!

2. $\mathrm{CI}_{\sum \beta_i}(k)$ achieves optimal expected length $k\sqrt{\frac{\log p}{n}}$.

3. NOT possible to construct adaptive CI.
   - Without knowing k, CI must be longer than $k\sqrt{\frac{\log p}{n}}$.

4. The information Σ is NOT useful.

Confidence intervals for $\eta^{\intercal}\beta$

	Known Σ	Unknown Σ
Sparse Loading η	$\frac{\|\eta\|_2}{\sqrt{n}}$	$\|\eta\|_2(\frac{1}{\sqrt{n}} + \frac{k \log p}{n})$
Dense Loading η	$\|\eta\|_\infty k \sqrt{\frac{\log p}{n}}$	

Confidence intervals for $\eta^\intercal \beta$

	Known Σ	Unknown Σ
Sparse Loading η	$\frac{\|\eta\|_2}{\sqrt{n}}$	$\|\eta\|_2(\frac{1}{\sqrt{n}} + \frac{k \log p}{n})$
Dense Loading η	$\|\eta\|_\infty k \sqrt{\frac{\log p}{n}}$	

	Known Σ	Unknown Σ
Sparse Loading η	$k \lesssim \frac{n}{\log p}$	$k \ll \frac{\sqrt{n}}{\log p}$
Dense Loading η	Impossible	

Tony Cai and Zijian Guo. *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity.* AOS, 2017.

Overview of talk

1. Inference for β_i: Review of De-biasing
2. Minimaxity and Adaptivity
3. Uniform Procedure for All loadings
4. Further Discussion on Optimality

Beyond Minimaxity

The minimax results for dense η are pessimistic.

Let's put the minimaxity aside first.

A practical question: Inference procedure for $\eta^{\intercal}\beta$?

1. Works for all η.
2. Requires no knowledge of sparsity.

Literature for $\eta^\intercal \beta$

Cai and Guo (2017)	η is sparse
Athey, Imbens, Wager (2018)	$\|\eta\|_2$ is bounded
Zhu and Bradic (2018)	Certain sparse η

Susan Athey, Guido W Imbens, and Stefan Wager. *Approximate residual balancing: debiased inference of average treatment effects in high dimensions.* JRSSB, 2018.

Yinchu Zhu and Jelena Bradic. *Linear hypothesis testing in dense high-dimensional linear models.* JASA, 2018.

A uniform procedure for all $x_{\rm new} \in \mathbb{R}^p$

Revisit β_i

$$\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\left(Y-X\widehat{\beta}\right)=\widehat{u}^{\intercal}\widehat{\Sigma}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon$$
$$=-e_i^{\intercal}(\widehat{\beta}-\beta)+(\widehat{\Sigma}\widehat{u}-e_i)^{\intercal}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon$$

Bias-corrected estimator

$$\widetilde{\beta}_{1,i}=e_i^{\intercal}\widehat{\beta}+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\left(Y-X\widehat{\beta}\right).$$

$$\widehat{u}=\arg\min_{u\in\mathbb{R}^p}\left\{\underbrace{u^{\intercal}\widehat{\Sigma}u}_{\text{Variance}}:\ \underbrace{\left\|\widehat{\Sigma}u-e_i\right\|_{\infty}\leq\|e_i\|_2\lambda_1}_{\text{Constrained Bias}}\right\}$$

Cai and G. (2017); Athey et.al. (2018)

$$\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\left(Y-X\widehat{\beta}\right)=\widehat{u}^{\intercal}\widehat{\Sigma}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon$$
$$=-\eta^{\intercal}(\widehat{\beta}-\beta)+(\widehat{\Sigma}\widehat{u}-\eta)^{\intercal}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\epsilon$$

Bias-corrected estimator

$$\widetilde{x_{\rm new}^{\intercal}\beta}=\eta^{\intercal}\widehat{\beta}+\widehat{u}^{\intercal}\frac{1}{n}X^{\intercal}\left(Y-X\widehat{\beta}\right).$$

$$\widehat{u}=\arg\min_{u\in\mathbb{R}^{p}}\left\{\underbrace{u^{\intercal}\widehat{\Sigma}u}_{\text{Variance}}:\ \underbrace{\left\|\widehat{\Sigma}u-\eta\right\|_{\infty}\leq\|\eta\|_{2}\lambda_{1}}_{\text{Constrained Bias}}\right\}$$

Challenges for Dense Loadings

Dense η:

$$\text{Feasible Set: } \left\|\widehat{\Sigma}u - \eta\right\|_\infty \leq \|\eta\|_2 \lambda_1$$

$$\|\eta\|_2 \lambda_1 \geq \|\eta\|_\infty \Rightarrow \widehat{u} = 0!$$

Example: If η is decaying as $\eta_j \asymp j^{-\delta}$, then $\|\eta\|_2 \asymp p^{\frac{1}{2}-\delta}$.

Challenges for Dense Loadings

Dense η:

$$\text{Feasible Set: } \left\|\widehat{\Sigma} u-\eta\right\|_{\infty} \leq\|\eta\|_{2} \lambda_{1}$$

$$\|\eta\|_{2} \lambda_{1} \geq\|\eta\|_{\infty} \Rightarrow \widehat{u}=0!$$

Example: If η is decaying as $\eta_j \asymp j^{-\delta}$, then $\|\eta\|_2 \asymp p^{\frac{1}{2}-\delta}$.

Bias-corrected estimator=plug-in estimator,

$$\widetilde{\eta^{\intercal} \beta}=\eta^{\intercal} \widehat{\beta}+\widehat{u}^{\intercal} \frac{1}{n} X^{\intercal}\left(Y-X \widehat{\beta}\right)=\eta^{\intercal} \widehat{\beta}.$$

Curse of dimensionality from dense η.

New Projection Direction

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^{\intercal} \widehat{\Sigma} u$$

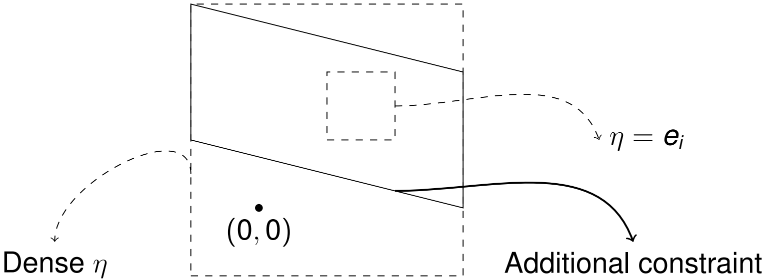
$$\text{subject to } \left\| \widehat{\Sigma} u - \eta \right\|_{\infty} \leq \|\eta\|_2 \lambda_1$$

$$\left| \eta^{\intercal} \widehat{\Sigma} u - \|\eta\|_2^2 \right| \leq \|\eta\|_2^2 \lambda_1$$

The proposed estimator for $\eta^{\intercal}\beta$ is

$$\widehat{\eta^{\intercal}\beta} = \eta^{\intercal}\widehat{\beta} + \widehat{u}^{\intercal} \frac{1}{n} X^{\intercal} \left(Y - X\widehat{\beta} \right) \tag{1}$$

Additional Constraint and Feasible Set



- Small dashed: $\eta = e_i$.
- Large dashed: dense η without additional constraint.
- Solid parallelogram: dense η with additional constraint.

$$\left|\eta^{\intercal}\widehat{\Sigma}u - \|\eta\|_2^2\right| \leq \|\eta\|_2^2\lambda_1$$

Bias-Variance Tradeoff

Bias and Variance Tradeoff.

- Minimizing variance with bias constrained.

$$\left|(\widehat{\Sigma}\widehat{u}-\eta)^{\intercal}(\beta-\widehat{\beta})\right| \leq \|\widehat{\Sigma}\widehat{u}-\eta\|_{\infty}\|\beta-\widehat{\beta}\|_{1}$$

- Minimizing variance with bias and variance constrained.

$$\widehat{u} = \arg\min_{u\in\mathbb{R}^{p}} u^{\intercal}\widehat{\Sigma}u$$

$$\text{subject to } \left\|\widehat{\Sigma}u-\eta\right\|_{\infty} \leq \|\eta\|_{2}\lambda_{1}$$

$$\left|\eta^{\intercal}\widehat{\Sigma}u-\|\eta\|_{2}^{2}\right| \leq \|\eta\|_{2}^{2}\lambda_{1}$$

Enhancing Variance Lemma

Lemma 1 (Cai, Cai, G. (2018)).

Under regularity conditions, we have

$$c_0 \frac{\|\eta\|_2}{\sqrt{n}} \leq \sqrt{\frac{1}{n} \widehat{u}^\intercal \widehat{\Sigma} \widehat{u}} \leq C_0 \frac{\|\eta\|_2}{\sqrt{n}}$$

- Lower bound does not hold without the additional constraint
- Additional constraint leads to a dominating variance

Theory

Theorem 2 (Cai, Cai, G. (2018)).

Under regularity conditions and $\|\beta\|_0 \le c\sqrt{n}/\log p$, *then*

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\eta^{\intercal}\beta} - \eta^{\intercal}\beta\right) \xrightarrow{d} N(0,1) \tag{2}$$

Theory

Theorem 2 (Cai, Cai, G. (2018)).

Under regularity conditions and $\|\beta\|_0 \leq c\sqrt{n}/\log p$*, then*

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\eta^\intercal\beta} - \eta^\intercal\beta\right) \xrightarrow{d} N(0,1) \tag{2}$$

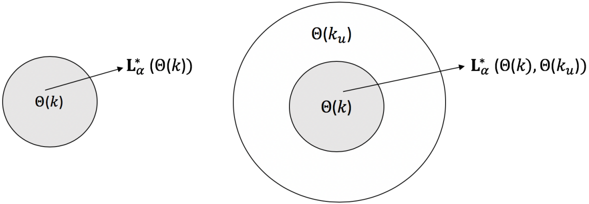
$\mathrm{V} \asymp \frac{\|\eta\|_2}{\sqrt{\mathrm{n}}}$ depends on η.

Works if $\|\beta\|_0 \leq c\sqrt{n}/\log p$.

Overview of talk

1. Inference for β_i: Review of De-biasing
2. Minimaxity and Adaptivity
3. Uniform Procedure for All loadings
4. Further Discussion on Optimality

Adaptive Optimal



$$L_\alpha^*(\Theta(k), \Theta(k_u)) = \inf_{\substack{\text{CI having coverage} \\ \text{for } \beta_1 \text{ over } \Theta(k_u)}} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI})$$

Adaptive optimal: a procedure achieving $L_\alpha^*(\Theta(k), \Theta(k_u))$.

Review of Exact Loading

We calibrate the sparsity levels as

$$k = p^{\gamma}, \quad k_u = p^{\gamma_u} \quad \text{for} \quad 0 \leq \gamma < \gamma_u \leq \mathbf{1},$$

$$c_0 \leq \max_{\{i:\eta_i \neq 0\}} |\eta_i| \,/ \min_{\{i:\eta_i \neq 0\}} |\eta_i| \leq C_0,$$

$$\|\eta\|_0 = p^{\gamma_\eta} \quad \text{for} \quad 0 \leq \gamma_\eta \leq 1.$$

(E1) $x_{\rm new}$ is called *exact sparse* if $\gamma_\eta \leq \mathbf{2}\gamma$;
(E2) $x_{\rm new}$ is called *exact dense* if $\gamma_\eta > 2\gamma$;

Possibility of Adaptive Testing

Suppose that $k \leq k_u \lesssim \frac{\sqrt{n}}{\log p}$,

	$\gamma, \gamma_u, \gamma_\eta$	$L_\alpha^*(\Theta(k))$	Rel	$L_\alpha^*(\Theta(k), \Theta(k_u))$	Adpt
(E1)	$\gamma_\eta \leq 2\gamma$	$\frac{\|\eta\|_2}{\sqrt{n}}$	$\asymp$	$\frac{\|\eta\|_2}{\sqrt{n}}$	Yes
(E2-a)	$\gamma < \gamma_u < \frac{1}{2}\gamma_\eta$	$\|\eta\|_\infty k \sqrt{\frac{\log p}{n}}$	$\ll$	$\|\eta\|_\infty k_u \sqrt{\frac{\log p}{n}}$	No
(E2-b)	$\gamma < \frac{1}{2}\gamma_\eta \leq \gamma_u$	$\|\eta\|_\infty k \sqrt{\frac{\log p}{n}}$	$\ll$	$\frac{\|\eta\|_2}{\sqrt{n}}$	No

- Cut-off for “dense" and “sparse" occurs at $\gamma_\eta = 2\gamma$.

Possibility of Adaptive Testing

Suppose that $k \leq k_u \lesssim \frac{\sqrt{n}}{\log p}$,

	$\gamma, \gamma_u, \gamma_\eta$	$L^*_\alpha(\Theta(k))$	Rel	$L^*_\alpha(\Theta(k), \Theta(k_u))$	Adpt
(E1)	$\gamma_\eta \leq 2\gamma$	$\frac{\|\eta\|_2}{\sqrt{n}}$	$\asymp$	$\frac{\|\eta\|_2}{\sqrt{n}}$	Yes
(E2-a)	$\gamma < \gamma_u < \frac{1}{2}\gamma_\eta$	$\|\eta\|_\infty k\sqrt{\frac{\log p}{n}}$	$\ll$	$\|\eta\|_\infty k_u\sqrt{\frac{\log p}{n}}$	No
(E2-b)	$\gamma < \frac{1}{2}\gamma_\eta \leq \gamma_u$	$\|\eta\|_\infty k\sqrt{\frac{\log p}{n}}$	$\ll$	$\frac{\|\eta\|_2}{\sqrt{n}}$	No

- Cut-off for "dense" and "sparse" occurs at $\gamma_\eta = 2\gamma$.
- If $\gamma_u \geq \frac{1}{2}\gamma_\eta$, then the optimal test is of order $\frac{\|\eta\|_2}{\sqrt{n}}$
- In absence of accurate sparsity information, the proposed inference procedure $\eta^\intercal\beta$ is **adaptive optimal** for **all** exact loadings η.

Take Home Message

- The best we can aim for: $L^*_\alpha(\Theta(k), \Theta(k_u))$
- Dense linear functionals are harder than sparse ones.
- Uniform Procedure over all loadings.

Reference and Acknowledgement

Tony Cai and Zijian Guo. *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity.* AOS, 2017.

Cai, T., Cai, T.T., Guo, Z. (2018). Individualized Treatment Selection: An Optimal Hypothesis Testing Approach In High-dimensional Models. Submitted.

Acknowledgement to NSF and NIH for fundings.

Thank you!

CI for $\eta^\top \beta$ (Cai and G., '16)

Fundamental difference in terms of minimaxity and adaptivity,

1. Sparse loading η : β_i
2. Dense loading η : $\sum_{i=1}^{p} \beta_i$

CI for $\eta^\top \beta$ (Cai and G., '16)

Fundamental difference in terms of minimaxity and adaptivity,

1. Sparse loading η : β_i
2. Dense loading η : $\sum_{i=1}^{p} \beta_i$

Plug-in Lasso Estimators

$$\begin{aligned} \beta_1: &\quad \widehat{\beta}_1 - \beta_1 = \langle e_1, \widehat{\beta} - \beta \rangle \\ \eta^\top \beta: &\quad \eta^\top \widehat{\beta} - \eta^\top \beta = \langle \eta, \widehat{\beta} - \beta \rangle \end{aligned}$$

CI for $\eta^\top \beta$ (Cai and G., '16)

Fundamental difference in terms of minimaxity and adaptivity,

1. Sparse loading η : β_i
2. Dense loading η : $\sum_{i=1}^{p} \beta_i$

Plug-in Lasso Estimators

$$\begin{aligned} \beta_1 : &\quad \widehat{\beta}_1 - \beta_1 = \langle e_1, \widehat{\beta} - \beta \rangle \\ \eta^\top \beta : &\quad \eta^\top \widehat{\beta} - \eta^\top \beta = \langle \eta, \widehat{\beta} - \beta \rangle \end{aligned}$$

- Sparse η: Correct the bias $\Rightarrow$ Similar to β_1.
- Dense η: NOT correct the bias $\Rightarrow$ Inflated variance.

Balance bias and variance.

Simulation Setting

Simulation Setting with $\eta^{\intercal}\beta = 1.08$

- $p = 501$, $n = n_2 = n$
- $\beta_{1,0} = -0.1$, $\beta_{1,j} = 0.4(j-1)$ for $1 \leq j \leq 10$
- $\beta_{2,0} = -0.5$, $\beta_{2,j} = 0.2(j-1)$ for $1 \leq j \leq 5$
- $x_{new,j} \sim N(0,1)$ for $1 \leq i \leq 10$ and $x_{new,j} \sim 0.2 * N(0,1)$ for $i \geq 11$

- **Adaptive optimality**: If the sparsity is unknown, what is the optimal length of CI?

Size

The parameter space

$$\Theta(s) = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \beta, \Sigma_1, \sigma_1 \\ \boldsymbol{\beta}_2, \Sigma_2, \sigma_2 \end{pmatrix} : \|\beta\|_0 \leq s,\ 0 < \sigma_k \leq M_0,\ \lambda_{\min}(\Sigma_k) \geq c_0,\ \text{for } k = 1, 2 \right\},$$

For a test ϕ, its size is

$$\boldsymbol{\alpha}(s, \phi) = \sup_{\theta \in \mathcal{H}_0(s)} \mathbb{E}_{\theta} \phi. \tag{3}$$

with

$$\mathcal{H}_0(s) = \{\boldsymbol{\theta} \in \Theta(s) : \eta^{\intercal}(\beta - \boldsymbol{\beta}_2) \leq 0\}$$

Power

The local alternative parameter space

$$\mathcal{H}_1(s,\tau) = \left\{\boldsymbol{\theta} \in \Theta(s) : x_{\text{new}}^{\intercal}(\beta - \boldsymbol{\beta}_2) = \tau > 0\right\}.$$

The power of ϕ over $\mathcal{H}_1(s,\tau)$ is defined as

$$\boldsymbol{\omega}(s,\tau,\phi) = \inf_{\boldsymbol{\theta} \in \mathcal{H}_1(s,\tau)} \mathbb{E}_{\theta}\phi. \tag{4}$$

Optimality: identify the smallest τ

- The size is controlled over $\mathcal{H}_0(s)$;
- The corresponding power over $\mathcal{H}_1(s,\tau)$ is large

Minimax Detection Boundary

Minimax detection boundary is defined as

$$\tau_{\rm mini}(k, x_{\rm new}) = \arg\min_{\tau} \left\{ \tau : \sup_{\phi:\boldsymbol{\alpha}(s,\phi)\leq\alpha} \boldsymbol{\omega}(s,\tau,\phi) \geq 1-\eta \right\}.$$

A test ϕ is minimax optimal if

$\boldsymbol{\alpha}(s,\phi) \leq \alpha$ and $\boldsymbol{\omega}(s,\phi,\tau) \geq 1-\eta$ for $\tau \asymp \tau_{\rm mini}(k, x_{\rm new})$

Minimax assumes s is known.

Unknown Sparsity Level

Capture the optimality for unknown sparsity level?

We consider two sparsity levels, $k \leq k_u$.

- k denotes the true sparsity level;
- k_u denotes an upper bound for the sparsity level.

The size is uniformly controlled over $\mathcal{H}_0(k_u)$,

$$\boldsymbol{\alpha}(k_u, \phi) = \sup_{\theta \in \mathcal{H}_0(k_u)} \mathbb{E}_\theta \phi \leq \alpha. \quad (5)$$

Adaptive Detection Boundary

The adaptive detection boundary $\tau_{\rm adap}(k_u, k, x_{\rm new})$

$$\tau_{\rm adap}(k_u, k, x_{\rm new}) = \arg\min_{\tau} \left\{ \tau : \sup_{\phi: \boldsymbol{\alpha}(k_u, \phi) \leq \alpha} \boldsymbol{\omega}(k, \tau, \phi) \geq 1 - \eta \right\}.$$

A test ϕ is adaptive optimal if

$$\boldsymbol{\alpha}(k_u, \phi) \leq \alpha \quad \text{and} \quad \boldsymbol{\omega}(k, \tau, \phi) \geq 1 - \eta \quad \text{for} \ \ \tau \asymp \tau_{\rm adap}(k_u, k, x_{\rm new})$$

An adaptive optimal test would be the best that we can aim for if there is lack of accurate information on sparsity.

Adaptive Hypothesis Testing

- If $\tau_{\rm mini}(k, x_{\rm new}) \asymp \tau_{\rm adap}(k_u, k, x_{\rm new})$, the testing problem is adaptive.
- If $\tau_{\rm mini}(k, x_{\rm new}) \ll \tau_{\rm adap}(k_u, k, x_{\rm new})$, the testing problem is NOT adaptive.

Numerical Comparison

Other methods

1. HITS
2. Plug-in scaled Lasso: $x_{\rm new}^{\intercal}(\widehat{\beta} - \widehat{\boldsymbol{\beta}}_2)$
3. Plug-in debiased Lasso: $x_{\rm new}^{\intercal}(\widetilde{\beta} - \widetilde{\boldsymbol{\beta}}_2)$

Numerical Comparison

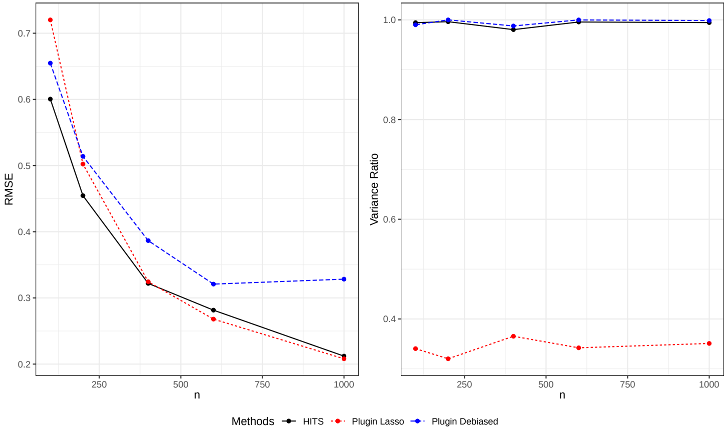
Other methods

1. HITS
2. Plug-in scaled Lasso: $x_{\rm new}^{\intercal}(\widehat{\beta} - \widehat{\boldsymbol{\beta}}_2)$
3. Plug-in debiased Lasso: $x_{\rm new}^{\intercal}(\widetilde{\beta} - \widetilde{\boldsymbol{\beta}}_2)$

Computation comparison

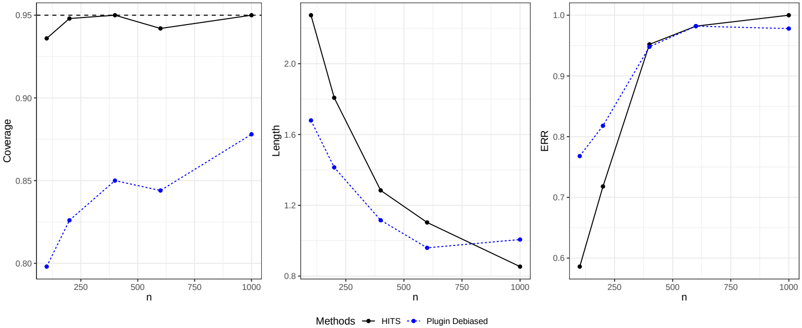
1. HITS: 4 Lasso
2. Plug-in scaled Lasso: 2 Lasso
3. Plug-in debiased Lasso: 1,004 Lasso $(2p+2)$

RMSE



- Plug-in Lasso: hard to do inference
- HITS has smaller RMSE than Plug-in Debiased

ITE and CI



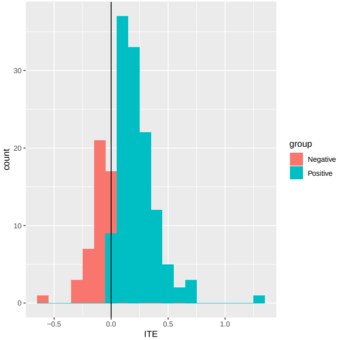
- Better coverage
- Computationally more efficient
- Comparable length and ERR

Real Data Analysis

Rheumatoid Arthritis (RA)

- Treatment 1: methotraxate+ anti-TNF (92 patients)
- Treatment 2: methotraxate (91 patients)
- Outcome $-\log(\mathrm{CRP})$
  Higher value of $Y \rightarrow$ Better treatment response.
- 171 Predictors, including Clinical measurement, EHR and SNP

Real Data Analysis

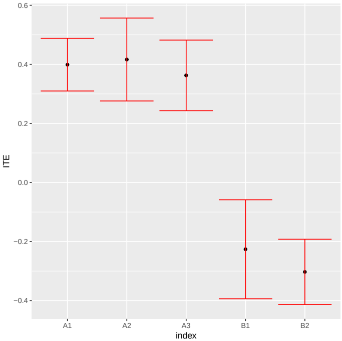


- About 72% benefit from the combination therapy.

Real Data Analysis

Patients	rs12506688	SLE mention	rs2843401	rs8043085	⋯
A	=0	≥ 1	= 0	> 0	⋯
B	>0	No	>0	=0	⋯

(SLE= Systemic Lupus Erythematosus)



The treatment effect is heterogeneous across patients.