Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications

Zijian Guo

Rutgers University

Harvard University

Collaborator

Tony T. Cai

Overview of talk

Research Problem

$$y_i = X_{i\cdot}^{\intercal} \beta_{p\times 1} + \epsilon_i \quad \text{for } 1 \leq i \leq n$$

- Number of covariates $p \geq$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.
- $\Sigma = \mathrm{Cov}\left(X_{i\cdot}\right)$ and $\sigma^2 = \mathrm{Var}(\epsilon_i)$

Research Problem

$$y_i = X_{i\cdot}^{\intercal} \beta_{p \times 1} + \epsilon_i \quad \text{for } 1 \leq i \leq n$$

- Number of covariates $p \geq$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.
- $\Sigma = \mathrm{Cov}\left(X_{i\cdot}\right)$ and $\sigma^2 = \mathrm{Var}(\epsilon_i)$

Confidence Interval for β_i: Zhang & Zhang '14; van de Geer, Bühlmann, Ritov & Dezeure '14; Javanmard & Montanari '14; Chernozhukov, Belloni & Hansen '13; Chernozhukov, Hansen & Spindler '15;

Research Problem

$$y_i = X_{i\cdot}^{\intercal} \beta_{p\times 1} + \epsilon_i \quad \text{for } 1 \le i \le n$$

- Number of covariates $p \geq$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.
- $\Sigma = \mathrm{Cov}\left(X_{i\cdot}\right)$ and $\sigma^2 = \mathrm{Var}(\epsilon_i)$

Confidence Interval for β_i: Zhang & Zhang '14; van de Geer, Bühlmann, Ritov & Dezeure '14; Javanmard & Montanari '14; Chernozhukov, Belloni & Hansen '13; Chernozhukov, Hansen & Spindler '15;

$$\mathrm{Var}(y_i) = \underbrace{\beta^{\intercal}\Sigma\beta}_{\text{Explained Variance}} + \sigma^2 \tag{1}$$

Research Problem

$$y_i = X_{i\cdot}^{\intercal}\beta_{p\times 1} + \epsilon_i \quad \text{for } 1 \leq i \leq n$$

- Number of covariates $p \geq$ sample size n.
- When $p > n$, $\|\beta\|_0 \leq k$.
- $\Sigma = \mathrm{Cov}\left(X_{i\cdot}\right)$ and $\sigma^2 = \mathrm{Var}(\epsilon_i)$

Confidence Interval for β_i: Zhang & Zhang '14; van de Geer, Bühlmann, Ritov & Dezeure '14; Javanmard & Montanari '14; Chernozhukov, Belloni & Hansen '13; Chernozhukov, Hansen & Spindler '15;

$$\mathrm{Var}(y_i) = \underbrace{\beta^{\intercal}\Sigma\beta}_{\text{Explained Variance}} + \sigma^2 \tag{1}$$

Semi-supervised Inference for $\mathrm{Q} = \beta^{\intercal}\Sigma\beta$

Semi-supervised Data

Semi-supervised data is a mixture of

- Labelled/Supervised data with sample size n
- Unlabelled/Unsupervised data with sample size N

$$\begin{bmatrix} X_{1,\cdot} & y_1 \\ X_{2,\cdot} & y_2 \\ \vdots & \vdots \\ X_{n,\cdot} & y_n \\ \hline X_{n+1,\cdot} & \text{NA} \\ X_{n+2,\cdot} & \text{NA} \\ \vdots & \vdots \\ X_{n+N,\cdot} & \text{NA} \end{bmatrix}$$

Semi-supervised Data

Semi-supervised data is a mixture of

- Labelled/Supervised data with sample size n
- Unlabelled/Unsupervised data with sample size N

$$\begin{bmatrix} X_{1,\cdot} & y_1 \\ X_{2,\cdot} & y_2 \\ \vdots & \vdots \\ X_{n,\cdot} & y_n \\ \hline X_{n+1,\cdot} & \text{NA} \\ X_{n+2,\cdot} & \text{NA} \\ \vdots & \vdots \\ X_{n+N,\cdot} & \text{NA} \end{bmatrix}$$

Efficient integration of labelled and unlabelled data

Semi-supervised Data

1. Electronic Health Records (EHR).
   - Covariates: extracted by natural language processing.
   - Outcomes: labelling is costly and time-consuming

Semi-supervised Data

1. Electronic Health Records (EHR).
   - Covariates: extracted by natural language processing.
   - Outcomes: labelling is costly and time-consuming
2. Integrative Genetics
   - Integrative analysis of multiple GWAS
   - Covariates: same set of genetic variants
   - Outcomes: vary from study to study
3. Missing Outcomes

Semi-supervised Data

1. Electronic Health Records (EHR).
   - Covariates: extracted by natural language processing.
   - Outcomes: labelling is costly and time-consuming
2. Integrative Genetics
   - Integrative analysis of multiple GWAS
   - Covariates: same set of genetic variants
   - Outcomes: vary from study to study
3. Missing Outcomes

Why to study $\mathrm{Q} = \beta^{\intercal}\Sigma\beta$?

Genetic Application: Heritability

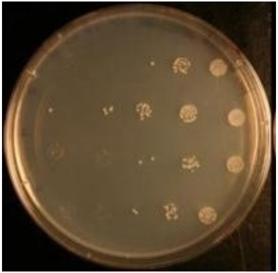


Figure: Yeast Colony YNB

1. Heritability: variance explained by genetic variants (e.g. SNPs)
2. For normalized outcome, represented by $\beta^{\intercal}\Sigma\beta$

Genetic Application: Heritability

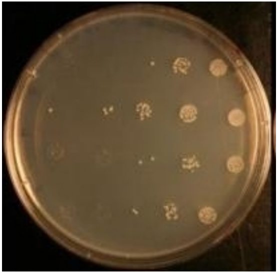


Figure: Yeast Colony YNB

1. Heritability: variance explained by genetic variants (e.g. SNPs)
2. For normalized outcome, represented by $\beta^\intercal \Sigma \beta$
3. Yeast study: $p = 4,410$ SNPs and $n = 1,008$ samples.
4. **Missing heritability.**

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436), 234-237.

Signal Detection

$$H_0 : \beta^{\intercal}\Sigma\beta = 0 \text{ v.s. } H_1 : \beta^{\intercal}\Sigma\beta > 0. \tag{2}$$

$\Sigma \approx \mathrm{I}$: Ingster, Tsybakov & Verzelen(2010); Arias-Castro, Candès, & Plan (2011).

Global testing

$$H_0 : \left(\beta - \beta^{\text{null}}\right)^{\intercal} \Sigma \left(\beta - \beta^{\text{null}}\right) = 0 \text{ v.s. } H_1 : \left(\beta - \beta^{\text{null}}\right)^{\intercal} \Sigma \left(\beta - \beta^{\text{null}}\right) > 0.$$

Applications: Accuracy and Confidence Ball

Prediction Accuracy Assessment of $\widehat{\beta}$

$$\mathbb{E}\left[x_{\text{new}}^{\intercal}(\widehat{\beta}-\beta)\right]^2=\left(\widehat{\beta}-\beta\right)^{\intercal}\Sigma\left(\widehat{\beta}-\beta\right)$$

Inference for $\|\widehat{\beta}-\beta\|_q^q$: Cai and Guo (2017).

Confidence Ball for β

$$\left\{\beta\in\mathbb{R}^p:\left(\widehat{\beta}-\beta\right)^{\intercal}\Sigma\left(\widehat{\beta}-\beta\right)\leq U\right\}$$

Knowledge of σ: Nickl and van de Geer (2013).

Overview of talk

Idea of Calibration/Correction

- $\widehat{\beta}$ and $\widehat{\Sigma}$ denote certain “good" estimators of β and Σ
- A natural estimator is the plug-in estimator $\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}$

Idea of Calibration/Correction

- $\widehat{\beta}$ and $\widehat{\Sigma}$ denote certain "good" estimators of β and Σ
- A natural estimator is the plug-in estimator $\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}$

Error Decomposition

$$\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}-\beta^{\intercal}\Sigma\beta = 2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)-\underbrace{(\widehat{\beta}-\beta)^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)}_{\text{Error of estimating } \beta}+\underbrace{\beta^{\intercal}(\widehat{\Sigma}-\Sigma)\beta}_{\text{Error of estimating } \Sigma}.$$

Idea of Calibration/Correction

- $\widehat{\beta}$ and $\widehat{\Sigma}$ denote certain “good" estimators of β and Σ
- A natural estimator is the plug-in estimator $\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}$

Error Decomposition

$$\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}-\beta^{\intercal}\Sigma\beta = 2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta) - \underbrace{(\widehat{\beta}-\beta)^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)}_{\text{Error of estimating } \beta} + \underbrace{\beta^{\intercal}(\widehat{\Sigma}-\Sigma)\beta}_{\text{Error of estimating } \Sigma}.$$

Idea: Calibrate plug-in estimator by estimating $2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)$.

Calibration/Correction term

$$\left(\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta} - 2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)\right) - \beta^{\intercal}\Sigma\beta = -\underbrace{(\widehat{\beta}-\beta)^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)}_{\text{Error of estimating } \beta} + \underbrace{\beta^{\intercal}(\widehat{\Sigma}-\Sigma)\beta}_{\text{Error of estimating } \Sigma} .$$

Calibration/Correction term

$$\left(\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}-2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)\right)-\beta^{\intercal}\Sigma\beta=-\underbrace{(\widehat{\beta}-\beta)^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)}_{\text{Error of estimating }\beta}+\underbrace{\beta^{\intercal}(\widehat{\Sigma}-\Sigma)\beta}_{\text{Error of estimating }\Sigma}.$$

Estimation of $2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)$

$$\begin{aligned}-2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}(y_i-X_{i\cdot}\widehat{\beta}) &= 2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\intercal}\left(\widehat{\beta}-\beta\right)-2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\epsilon_i. \\ &\approx 2\widehat{\beta}^{\intercal}\widehat{\Sigma}(\widehat{\beta}-\beta)-2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\epsilon_i \end{aligned} \tag{3}$$

CHIVE

Propose the following calibrated/corrected estimator

$$\widehat{\mathrm{Q}}(\widehat{\beta},\widehat{\Sigma},Z)=\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}+\underbrace{2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}(y_i-X_{i\cdot}\widehat{\beta})}_{\text{Calibration Term}}. \tag{4}$$

<u>Calibrated High-dimensional Inference for Variance Explained (**CHIVE**)</u>

CHIVE

Propose the following calibrated/corrected estimator

$$\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}, Z) = \widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta} + \underbrace{2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta})}_{\text{Calibration Term}}. \quad (4)$$

Calibrated High-dimensional Inference for Variance Explained (**CHIVE**)

Required Inputs:

- $\widehat{\beta}$: estimator of β
- $\widehat{\Sigma}$: estimator of Σ
- Labelled data $Z = (X, y)$

Algorithm Inputs

$$\{\widehat{\beta}, \widehat{\sigma}\} = \arg \min_{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sqrt{\frac{2.01 \log p}{n}} \sum_{j=1}^{p} \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j|.$$

$\widehat{\Sigma} = \frac{1}{n+N} \sum_{i=1}^{n+N} X_{i\cdot} X_{i\cdot}^{\intercal}$

- Unlabelled data is only used here.

Algorithm Inputs

$$\{\widehat{\beta}, \widehat{\sigma}\} = \arg \min_{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sqrt{\frac{2.01 \log p}{n}} \sum_{j=1}^{p} \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j|.$$

$\widehat{\Sigma} = \frac{1}{n+N} \sum_{i=1}^{n+N} X_{i\cdot} X_{i\cdot}^{\intercal}$

- Unlabelled data is only used here.

Assumptions on $\widehat{\beta}$ and $\widehat{\sigma}$

- With high probability, the estimator $\widehat{\beta}$ satisfies

$$\max \left\{ \frac{1}{\sqrt{n}} \|X(\widehat{\beta} - \beta)\|_2, \|\widehat{\beta} - \beta\|_2 \right\} \lesssim \sqrt{\frac{k \log p}{n}}, \quad \|\widehat{\beta} - \beta\|_1 \lesssim k \sqrt{\frac{\log p}{n}}.$$

- $\widehat{\sigma}^2$ is a consistent estimator of σ^2.

Convergence Rate

Theorem 1(Cai. & G., 2018)

Suppose that $k \leq cn/\log p$ for some constant $c > 0$, the estimator $\widehat{\mathrm{Q}}$ satisfies

$$\left|\widehat{\mathrm{Q}} - \mathrm{Q}\right| \lesssim \frac{\|\beta\|_2}{\sqrt{n}} + \frac{\|\beta\|_2^2}{\sqrt{N+n}} + \frac{k \log p}{n}. \quad (5)$$

- N: sample size of unlabelled data;
- n: sample size of labelled data;
- k: number of non-zeros in β;
- Depends on $\|\beta\|_2$.

Convergence Rate

Theorem 1(Cai. & G., 2018)

Suppose that $k \leq cn/\log p$ for some constant $c > 0$, the estimator $\widehat{\mathrm{Q}}$ satisfies

$$\left|\widehat{\mathrm{Q}} - \mathrm{Q}\right| \lesssim \frac{\|\beta\|_2}{\sqrt{n}} + \frac{\|\beta\|_2^2}{\sqrt{N+n}} + \frac{k \log p}{n}. \quad (5)$$

- N: sample size of unlabelled data;
- n: sample size of labelled data;
- k: number of non-zeros in β;
- Depends on $\|\beta\|_2$.
- Unlabelled data is helpful.

Optimal Convergence Rate

$$\Theta(k, M) = \left\{(\beta, \Sigma, \sigma) : \|\beta\|_0 \leq k,\ \frac{M}{2} \leq \|\beta\|_2 \leq M,\ c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0\right\}$$

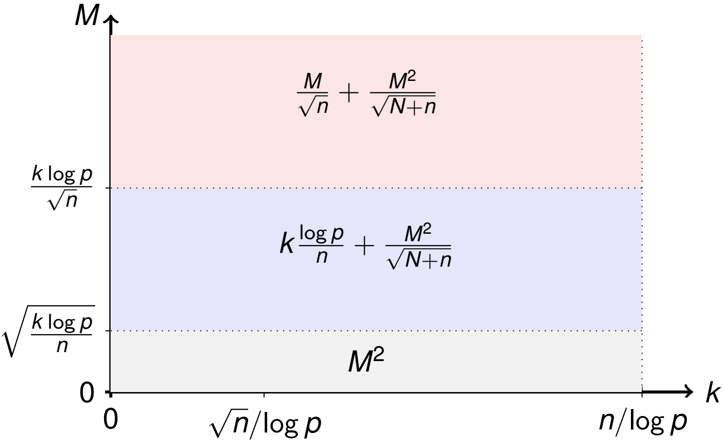
- k: sparsity level;
- M: the signal strength of β in its ℓ_2 norm.

Theorem 2(Cai. & G., 2018)

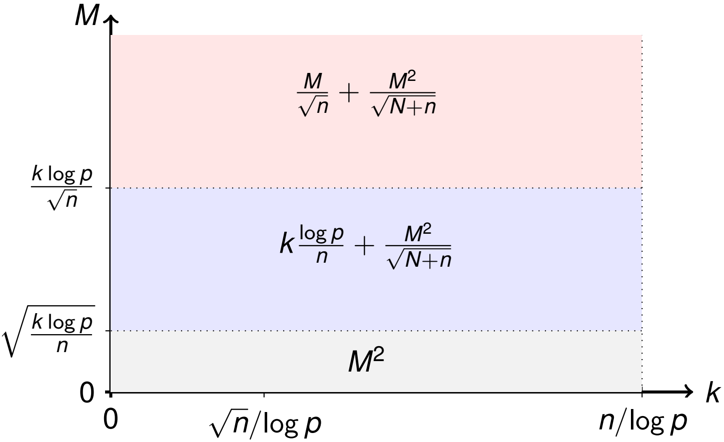
Suppose $k \leq c \min\{n/\log p, p^{\nu}\}$ for some constants $c > 0$ and $0 \leq \nu < \frac{1}{2}$. Then

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\theta \in \Theta(k,M)} \mathbb{P}\left(\left|\widetilde{\mathrm{Q}} - \mathrm{Q}\right| \gtrsim \frac{M^2}{\sqrt{N+n}} + \min\left\{\frac{M}{\sqrt{n}} + \frac{k \log p}{n}, M^2\right\}\right) \geq \frac{1}{4}.$$

Optimal Convergence Rate



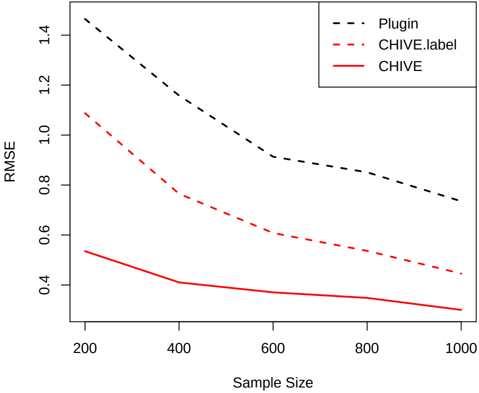
Optimal Convergence Rate



For $M \gtrsim \sqrt{\frac{k \log p}{n}}$, the optimal rate is achieved by CHIVE.

Numerical illustration: RMSE

- $p = 800$, $n \in \{200, 400, 600, 800, 1,000\}$ and $N = 2,000$
- $k = 10$ and $\beta = (0.1, 0.2, 0.3, \cdots, 1, 0, 0, \cdots, 0)$
- True value $\beta^{\intercal}\Sigma\beta = 9.42$



Special Case: Supervised Setting

Supervised Setting

Estimate Σ by $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}$ and then

$$\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}, Z) = \widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta} + 2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta}). \tag{6}$$

Corollary 1(Cai. & G., 2018)

Suppose $k \le c\min\{n/\log p, p^{\nu}\}$ for some constants $c > 0$ and $0 \le \nu < \frac{1}{2}$, the CHIVE estimator achieves the optimal convergence rate

$$\frac{M}{\sqrt{n}} + \frac{M^2}{\sqrt{n}} + \frac{k\log p}{n} \tag{7}$$

over $\Theta(k, M)$ for $M \gtrsim \sqrt{k\log p/n}$.

- Special case of semi-supervised setting with $N = 0$.

Connection to Literature

- $\mathrm{Q} = \mathbb{E}(y_i^2) - \sigma^2$
- Sun and Zhang [2012] and Verzelen and Gassiat [2016]

$$\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}+2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}(y_i-X_{i\cdot}\widehat{\beta}) = \frac{1}{n}\left(\|y\|_2^2 - \|y - X\widehat{\beta}\|_2^2\right) = \frac{1}{n}\|y\|_2^2-\widehat{\sigma}^2.$$

Connection to Literature

- $\mathrm{Q} = \mathbb{E}(y_i^2) - \sigma^2$
- Sun and Zhang [2012] and Verzelen and Gassiat [2016]

$$\widehat{\beta}^\intercal \widehat{\Sigma} \widehat{\beta} + 2\widehat{\beta}^\intercal \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta}) = \frac{1}{n}\left(\|y\|_2^2 - \|y - X\widehat{\beta}\|_2^2\right) = \frac{1}{n}\|y\|_2^2 - \widehat{\sigma}^2.$$

- New perspective: estimate $\beta^\intercal \Sigma \beta$ directly by calibrating the plug-in estimator

Connection to Literature

- $\mathrm{Q} = \mathbb{E}(y_i^2) - \sigma^2$
- Sun and Zhang [2012] and Verzelen and Gassiat [2016]

$$\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta} + 2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta}) = \frac{1}{n}\left(\|y\|_2^2 - \|y - X\widehat{\beta}\|_2^2\right) = \frac{1}{n}\|y\|_2^2 - \widehat{\sigma}^2.$$

- New perspective: estimate $\beta^{\intercal}\Sigma\beta$ directly by calibrating the plug-in estimator
- This new perspective is useful for semi-supervised setting.
- Study of uncertainty quantification.

Statistical Fundamental Limit Comparison

- Consider $k \leq c \min\{n/\log p, p^{\nu}\}$ for $0 \leq \nu < 1/2$
- Sequence model: $y_i = \beta_i + \frac{1}{\sqrt{n}}\epsilon_i$ for $1 \leq i \leq p$.

Model	Target	Optimal Rate over $\Theta(k, M)$
Sequence model	$\|\beta\|_2^2$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{k\log p}{n}, M^2\right\}$
HD regression	$\|\beta\|_2^2$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{k\log p}{n} + M\frac{k\log p}{n}, M^2\right\}$
HD regression	$\beta^{\intercal}\Sigma\beta$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{k\log p}{n} + M^2\frac{1}{\sqrt{n}}, M^2\right\}$

Collier, O., Comminges, L., & Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *AOS*, 45(3), 923-958.

Guo, Z., Wang, W., Cai, T.T., & Li, H.(2017). Optimal estimation of Genetic Relatedness in high-dimensional linear models. *JASA*, to appear.

Overview of talk

Limiting Distribution

Theorem 3(Cai. & G., 2018)

Suppose that $k \ll \sqrt{n}/\log p$ and $\|\beta\|_2 \gg k \log p/\sqrt{n}$,

$$\frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \mathrm{Q}\right)}{\sqrt{4\sigma^2 \beta^\intercal \Sigma \beta + \rho \mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2}} \to N(0,1) \quad (8)$$

where $\rho = \lim_{n\to\infty} \frac{n}{N+n}$.

- Stronger conditions than estimation

Limiting Distribution

Theorem 3(Cai. & G., 2018)

Suppose that $k \ll \sqrt{n}/\log p$ and $\|\beta\|_2 \gg k \log p/\sqrt{n}$,

$$\frac{\sqrt{n}\left(\widehat{\mathrm{Q}}-\mathrm{Q}\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta+\rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2}} \to N(0,1) \quad (8)$$

where $\rho = \lim_{n\to\infty}\frac{n}{N+n}$.

- Stronger conditions than estimation
- $\underbrace{4\sigma^2\beta^\intercal\Sigma\beta}_{\text{Uncertainty for } \beta} + \underbrace{\rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2}_{\text{Uncertainty for } \Sigma}$
- If $\rho = 0$, then $\frac{\sqrt{n}(\widehat{\mathrm{Q}}-\mathrm{Q})}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta}} \to N(0,1)$

Confidence Interval Construction

Estimate $\sqrt{4\sigma^2\beta^\intercal\Sigma\beta+\rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2}/\sqrt{n}$.

- Estimate $4\sigma^2\beta^\intercal\Sigma\beta$ by $\widehat{\phi}_1=\widehat{\sigma}^2\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$,
- Estimate ρ by $\widehat{\rho}=n/(N+n)$,
- Estimate $\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2$ by

$$\widehat{\phi}_2=\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta}-\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}\right)^2.$$

Confidence Interval Construction

Estimate $\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}/\sqrt{n}$.

- Estimate $4\sigma^2\beta^\intercal\Sigma\beta$ by $\widehat{\phi}_1 = \widehat{\sigma}^2\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$,
- Estimate ρ by $\widehat{\rho} = n/(N+n)$,
- Estimate $\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2$ by

$$\widehat{\phi}_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}\right)^2.$$

We propose the following CI,

$$\mathrm{CI}(Z) = \left(\left(\widehat{\mathrm{Q}} - z_{\alpha/2}\widehat{\phi}\right)_+, \ \widehat{\mathrm{Q}} + z_{\alpha/2}\widehat{\phi}\right), \text{ where } \widehat{\phi} = \sqrt{\frac{4\widehat{\phi}_1 + \widehat{\rho}\widehat{\phi}_2}{n}}.$$

Coverage and Length Precision

Theorem 4(Cai. & G., 2018)

Suppose that $k \ll \sqrt{n}/\log p$ and $\|\beta\|_2 \gg k \log p/\sqrt{n}$, then

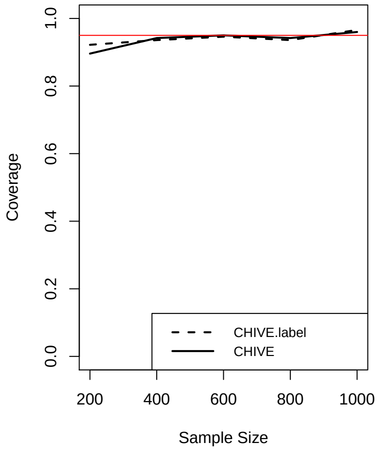
$$\liminf_{n\to\infty} \mathbb{P}\left(\beta^{\intercal}\Sigma\beta \in \mathrm{CI}(Z)\right) \geq 1-\alpha$$

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbf{L}(\mathrm{CI}(Z)) \geq (1+\delta_0)\sqrt{\frac{4\sigma^2\beta^{\intercal}\Sigma\beta}{n} + \frac{\mathbb{E}\left(\beta^{\intercal}X_{1\cdot}X_{1\cdot}^{\intercal}\beta - \beta^{\intercal}\Sigma\beta\right)^2}{N+n}}\right) = 0$$

for any positive constant $\delta_0 > 0$.

Additional unlabelled data leads to shorter confidence intervals.

Numerical illustration: Coverage and Precision



Weak Signals: Super-efficiency

For $k \ll \frac{\sqrt{n}}{\log p}$, coverage property is only for $\|\beta\|_2 \gg \frac{k \log p}{\sqrt{n}}\sigma$.

Weak Signals: Super-efficiency

For $k \ll \frac{\sqrt{n}}{\log p}$, coverage property is only for $\|\beta\|_2 \gg \frac{k \log p}{\sqrt{n}}\sigma$.

1. $\sqrt{\mathrm{variance}}$ level is $\sqrt{\frac{4\sigma^2 \beta^\intercal \Sigma \beta}{n} + \frac{\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2}{N+n}}$
2. Bias level: $k \log p/n$

Weak Signals: Super-efficiency

For $k \ll \frac{\sqrt{n}}{\log p}$, coverage property is only for $\|\beta\|_2 \gg \frac{k \log p}{\sqrt{n}} \sigma$.

1. $\sqrt{\text{variance}}$ level is $\sqrt{\frac{4\sigma^2 \beta^\intercal \Sigma \beta}{n} + \frac{\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2}{N+n}}$
2. Bias level: $k \log p / n$
3. Even for weak signals, CHIVE still shoots at the center.

Weak Signals: Super-efficiency

For $k \ll \frac{\sqrt{n}}{\log p}$, coverage property is only for $\|\beta\|_2 \gg \frac{k \log p}{\sqrt{n}}\sigma$.

1. $\sqrt{\text{variance}}$ level is $\sqrt{\frac{4\sigma^2 \beta^\intercal \Sigma \beta}{n} + \frac{\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2}{N+n}}$
2. Bias level: $k \log p / n$
3. Even for weak signals, CHIVE still shoots at the center.

Randomized Calibration

Generate random variables $u_i \stackrel{iid}{\sim} N(0, \tau_0^2)$ for $1 \leq i \leq n$

Randomized Calibration

Generate random variables $u_i \overset{iid}{\sim} N(0, \tau_0^2)$ for $1 \leq i \leq n$

$$\widehat{\mathrm{Q}}^R\left(\widehat{\beta}, \widehat{\Sigma}, Z, u\right) = \widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta} + 2\frac{1}{n}\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}\widehat{\beta} + u_i\right)(y_i - X_{i\cdot}^{\intercal}\widehat{\beta}). \quad (9)$$

Randomized Calibration

Generate random variables $u_i \overset{iid}{\sim} N(0, \tau_0^2)$ for $1 \leq i \leq n$

$$\widehat{\mathrm{Q}}^R\left(\widehat{\beta}, \widehat{\Sigma}, Z, u\right) = \widehat{\beta}^\intercal \widehat{\Sigma} \widehat{\beta} + 2\frac{1}{n}\sum_{i=1}^{n}\left(X_{i\cdot}^\intercal \widehat{\beta} + u_i\right)(y_i - X_{i\cdot}^\intercal \widehat{\beta}). \quad (9)$$

1. If $u_i = 0$, reduced to non-randomized CHIVE.

Randomized Calibration

Generate random variables $u_i \stackrel{iid}{\sim} N(0, \tau_0^2)$ for $1 \leq i \leq n$

$$\widehat{\mathrm{Q}}^R \left(\widehat{\beta}, \widehat{\Sigma}, Z, u\right) = \widehat{\beta}^{\intercal} \widehat{\Sigma} \widehat{\beta} + 2\frac{1}{n} \sum_{i=1}^{n} \left(X_{i\cdot}^{\intercal} \widehat{\beta} + u_i\right) (y_i - X_{i\cdot}^{\intercal} \widehat{\beta}). \quad (9)$$

1. If $u_i = 0$, reduced to non-randomized CHIVE.
2. For $u_i \stackrel{iid}{\sim} N(0, \tau_0^2)$, then

$$2\frac{1}{n} \sum_{i=1}^{n} u_i (y_i - X_{i\cdot}^{\intercal} \widehat{\beta}) \approx N(0, 4\sigma^2 \tau_0^2 / n).$$

3. The enlarged $\sqrt{\text{variance}}$ level dominates the bias level.

Limiting Distribution

Theorem 5(Cai.& G., 2018)

Suppose $k \ll \sqrt{n}/\log p$ and $\tau_0 > 0$ is a positive constant,

$$\sqrt{n}\frac{\widehat{\mathrm{Q}}^{R}-\mathrm{Q}}{\sqrt{4\sigma^2\left(\beta^{\intercal}\Sigma\beta+\tau_0^2\right)+\rho\mathbb{E}\left(\beta^{\intercal}X_{1\cdot}X_{1\cdot}^{\intercal}\beta-\beta^{\intercal}\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1)$$

where $\rho = \lim \frac{n}{n+N}$.

Limiting Distribution

Theorem 5(Cai.& G., 2018)

Suppose $k \ll \sqrt{n}/\log p$ and $\tau_0 > 0$ is a positive constant,

$$\sqrt{n}\frac{\widehat{\mathrm{Q}}^{R} - \mathrm{Q}}{\sqrt{4\sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right) + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1)$$

where $\rho = \lim \frac{n}{n+N}$.

1. No Free Lunch: variance enlarged by $4\sigma^2\tau_0^2/n$
2. Finite sample: $\tau_0 \geq C\frac{k\log p}{\sqrt{n}}\sigma$

Limiting Distribution

Theorem 5(Cai.& G., 2018)

Suppose $k \ll \sqrt{n}/\log p$ and $\tau_0 > 0$ is a positive constant,

$$\sqrt{n}\frac{\widehat{\mathrm{Q}}^{R}-\mathrm{Q}}{\sqrt{4\sigma^2\left(\beta^{\intercal}\Sigma\beta+\tau_0^2\right)+\rho\mathbb{E}\left(\beta^{\intercal}X_{1\cdot}X_{1\cdot}^{\intercal}\beta-\beta^{\intercal}\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1)$$

where $\rho = \lim \frac{n}{n+N}$.

1. No Free Lunch: variance enlarged by $4\sigma^2\tau_0^2/n$
2. Finite sample: $\tau_0 \geq C\frac{k\log p}{\sqrt{n}}\sigma$
3. Construct CI by estimating the standard error.

Overview of talk

Statistical Application: Signal Detection

Signal Detection

Signal Detection

$$H_0 : \beta^{\intercal}\Sigma\beta = 0 \text{ v.s. } H_1 : \beta^{\intercal}\Sigma\beta > 0.$$

Signal Detection

Signal Detection

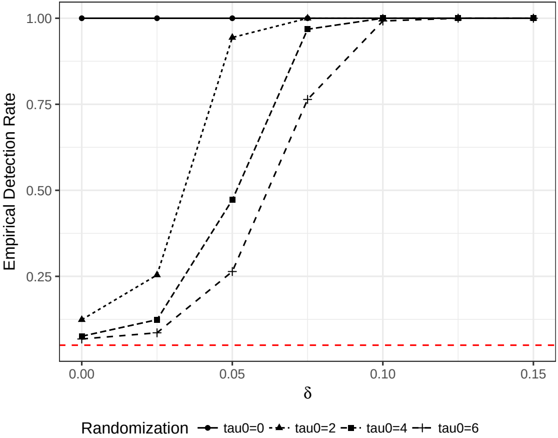
$$H_0 : \beta^\intercal \Sigma \beta = 0 \text{ v.s. } H_1 : \beta^\intercal \Sigma \beta > 0.$$

1. We choose τ_0 and apply randomized calibration,
   - obtain the point estimator $\widehat{\mathrm{Q}}^R(\tau_0)$ for $\beta^\intercal \Sigma \beta$;
   - obtain the SE estimator $\widehat{\phi}^R(\tau_0)$.
2. For $\alpha \in (0, 1)$, propose

$$D(\tau_0) = \mathbf{1}\left(\widehat{\mathrm{Q}}^R(\tau_0) \geq \widehat{\phi}^R(\tau_0) z_\alpha\right).$$

Numerical illustration: Signal Detection

- $n = 600$, $p = 800$, $\beta = (\underbrace{\delta, \cdots, \delta}_{\text{50 repetitions}}, 0, \cdots, 0)$
- $\delta \in \{0, 0.025, 0.05, 0.075, 0.10, 0.125, 0.15\}$



Biological Application: Heritability

Missing Heritability

- Data: $n = 1,008$ yeast, $p = 4,410$ markers, 46 traits;
- Missing heritability (Bloom et al., 2013)
  " the undiscovered factors could have effects that are too small to be detected with current sample sizes".

  Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436), 234-237.

Missing Heritability

- Data: $n = 1,008$ yeast, $p = 4,410$ markers, 46 traits;
- Missing heritability (Bloom et al., 2013)
  " the undiscovered factors could have effects that are too small to be detected with current sample sizes".

  Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436), 234-237.
- $\widehat{\beta}^{\intercal}\widehat{\Sigma}\widehat{\beta}$ tends to lower estimate $\beta^{\intercal}\Sigma\beta$

Confidence Interval for Heritability

	Supervised			Semi-Supervised			
Media	Plug	CHIVE	CI	Plug	CHIVE	CI	Missing
Raffinose	0.3168	0.5105 (0.0410)	[0.4300, 0.5909]	0.3105	0.5041 (0.0399)	[0.4259, 0.5824]	34.33%
Sorbitol	0.2968	0.4893 (0.0431)	[0.4049, 0.5737]	0.2864	0.4789 (0.0417)	[0.3972, 0.5606]	40.58%
YNB	0.3654	0.5927 (0.0347)	[0.5248, 0.6607]	0.3652	0.5926 (0.0347)	[0.5247, 0.6605]	0.20%

1. CHIVE adds back missing heritability due to small effects.

Confidence Interval for Heritability

	Supervised			Semi-Supervised			
Media	Plug	CHIVE	CI	Plug	CHIVE	CI	Missing
Raffinose	0.3168	0.5105 (0.0410)	[0.4300, 0.5909]	0.3105	0.5041 (0.0399)	[0.4259, 0.5824]	34.33%
Sorbitol	0.2968	0.4893 (0.0431)	[0.4049, 0.5737]	0.2864	0.4789 (0.0417)	[0.3972, 0.5606]	40.58%
YNB	0.3654	0.5927 (0.0347)	[0.5248, 0.6607]	0.3652	0.5926 (0.0347)	[0.5247, 0.6605]	0.20%

1. CHIVE adds back missing heritability due to small effects.
2. Shorter CI with unlabelled data
   - around 3% for Sorbitol (with 40.58% outcome missing)
   - around 2% for Raffinose (with 34.33% outcome missing)
3. Colony sizes are genetically heritable (Bloom et al., 2013)

Overview of talk

Summary

1. Inference for $\beta^\intercal \Sigma \beta$
   - Calibration/Correction
   - Randomization (weak signals)

Summary

1. Inference for $\beta^\intercal \Sigma \beta$
   - Calibration/Correction
   - Randomization (weak signals)
2. Semi-supervised: efficient integration of unlabelled data

Summary

1. Inference for $\beta^\intercal \Sigma \beta$
   - Calibration/Correction
   - Randomization (weak signals)
2. Semi-supervised: efficient integration of unlabelled data
3. Statistical and biological applications
   - Heritability
   - Signal to noise ratio
   - Signal detection
   - Prediction accuracy assessment
   - Confidence set construction
   - $\cdots$

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 $1 \leq q < 2$.

References

Cai, T.T., & Guo, Z.(2018). Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications. *Submitted.*

Acknowledgement to NSF and NIH for fundings.

Thank you!

Bias Correction

Error decomposition of $\|\widehat{\beta}\|_2^2$:

$$\|\widehat{\beta}\|_2^2 - \|\beta\|_2^2 = -\underbrace{2\langle \widehat{\beta}, \beta - \widehat{\beta} \rangle}_{\text{Main Error}} - \|\widehat{\beta} - \beta\|_2^2 \qquad (10)$$

Bias Correction

Error decomposition of $\|\widehat{\beta}\|_2^2$:

$$\|\widehat{\beta}\|_2^2 - \|\beta\|_2^2 = -\underbrace{2\langle\widehat{\beta}, \beta - \widehat{\beta}\rangle}_{\text{Main Error}} - \|\widehat{\beta} - \beta\|_2^2 \tag{10}$$

Bias correction idea:

$$\left(\|\widehat{\beta}\|_2^2 + \underbrace{2\langle\widehat{\beta}, \beta - \widehat{\beta}\rangle}_{\text{Main Error}}\right) - \|\beta\|_2^2 = -\|\widehat{\beta} - \beta\|_2^2. \tag{11}$$

Bias Correction

Error decomposition of $\|\widehat{\beta}\|_2^2$:

$$\|\widehat{\beta}\|_2^2 - \|\beta\|_2^2 = -\underbrace{2\langle\widehat{\beta}, \beta - \widehat{\beta}\rangle}_{\text{Main Error}} - \|\widehat{\beta} - \beta\|_2^2 \tag{10}$$

Bias correction idea:

$$\left(\|\widehat{\beta}\|_2^2 + \underbrace{2\langle\widehat{\beta}, \beta - \widehat{\beta}\rangle}_{\text{Main Error}}\right) - \|\beta\|_2^2 = -\|\widehat{\beta} - \beta\|_2^2. \tag{11}$$

Intuition of estimating $\langle\widehat{\beta}, \beta - \widehat{\beta}\rangle$

$$\frac{1}{n}X^\top X(\beta - \widehat{\beta}) = -\frac{1}{n}X^\top \epsilon + \lambda \text{sign}(\widehat{\beta}).$$

Multiply both sides by $\widehat{\beta}^\top \left(\frac{1}{n}X^\top X\right)^{-1}$.

Projection Direction

- (y, X) is split into two subsamples $\left(y^{(1)}, X^{(1)}\right)$ with sample size $n/2$ and $\left(y^{(2)}, X^{(2)}\right)$ with sample size $n/2$.

Projection Direction

- (y, X) is split into two subsamples $\left(y^{(1)}, X^{(1)}\right)$ with sample size $n/2$ and $\left(y^{(2)}, X^{(2)}\right)$ with sample size $n/2$.
- Let $\widehat{\beta}$ denote the scaled Lasso estimator based on $\left(y^{(1)}, X^{(1)}\right)$.

Projection Direction

- (y, X) is split into two subsamples $\left(y^{(1)}, X^{(1)}\right)$ with sample size $n/2$ and $\left(y^{(2)}, X^{(2)}\right)$ with sample size $n/2$.
- Let $\widehat{\beta}$ denote the scaled Lasso estimator based on $\left(y^{(1)}, X^{(1)}\right)$.
- For $u \in \mathbb{R}^p$,

$$\begin{aligned}&\frac{1}{n/2}u^\top\left(X^{(2)}\right)^\top\left(y^{(2)}-X^{(2)}\widehat{\beta}\right)-\langle\widehat{\beta},\beta-\widehat{\beta}\rangle\\=&\underbrace{\frac{1}{n/2}u^\top\left(X^{(2)}\right)^\top\epsilon^{(2)}}_{\text{Variance}}+\underbrace{\left(u^\top\widehat{\Sigma}-\widehat{\beta}^\top\right)\left(\beta-\widehat{\beta}\right)}_{\text{Bias}},\end{aligned} \tag{12}$$

with $\widehat{\Sigma}=\left(X^{(2)}\right)^\top X^{(2)}/(n/2)$.

Projection Direction

1. $\frac{1}{\sqrt{n/2}} u^{\top} \left(X^{(2)}\right)^{\top} \epsilon^{(2)} \mid X^{(2)} \sim N(0, u^{\top}\widehat{\Sigma}u).$
2. $\left|\sqrt{\frac{n}{2}} \left(u^{\top}\widehat{\Sigma} - \widehat{\beta}^{\top}\right) \left(\beta - \widehat{\beta}\right)\right| \leq \sqrt{\frac{n}{2}} \|\widehat{\Sigma}u - \widehat{\beta}\|_{\infty} \|\beta - \widehat{\beta}\|_{1}$

- Define the projection direction $\widehat{u}$ as

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} \left\{ u^{\top}\widehat{\Sigma}u : \|\widehat{\Sigma}u - \widehat{\beta}\|_{\infty} \leq \|\widehat{\beta}\|_2 \frac{\lambda_1}{\sqrt{n/2}} \right\}, \quad (13)$$

where $\lambda_1 \asymp \sqrt{\log p}$.

Functional Debiased Estimator (FDE)

- Estimate $\langle \widehat{\beta}, \beta - \widehat{\beta} \rangle$ by

$$\widehat{u}^{\top} \frac{1}{n/2} \left(X^{(2)} \right)^{\top} \left(y^{(2)} - X^{(2)} \widehat{\beta} \right).$$

- Propose Functional Debiased Estimator (FDE) of $\|\beta\|_2^2$ as

$$\widehat{\|\beta\|_2^2} = \left(\|\widehat{\beta}\|_2^2 + \underbrace{2\widehat{u}^{\top} \frac{1}{n/2} \left(X^{(2)} \right)^{\top} \left(y^{(2)} - X^{(2)} \widehat{\beta} \right)}_{\text{Correction}} \right)_{+}. \quad (14)$$