Post-selection Inference for all

using NBK Inequalities

Arun Kumar Kuchibhotla

The Wharton School,
University of Pennsylvania.

19 August, 2019

Outline

1. Introduction: Bahadur Representation

2. NBK Inequalities: Linear Regression
   - Application 1: Berry–Esseen Bounds
   - Application 2: Transformations of Response
   - Application 3: Variable Selection
   - Implication 1: Post-selection Inference

3. Summary and Conclusions

Introduction: Bahadur Representation

Let's Remember Cramér

- Suppose $Z_1, \ldots, Z_n$ are observations and we consider estimtor $\hat{\theta}$ satisfying

$$\textstyle\sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) = 0.$$

- MLE, OLS, GLMs and many more estimators are all obtained this way.
- How to do inference using $\hat{\theta}_n$?
- The classical proof of Cramér (1946) proves the Bahadur representation:

$$\sqrt{n}(\hat{\theta} - \theta) \; = \; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbb{E}[\dot{\psi}(Z_1, \theta)])^{-1} \psi(Z_i, \theta) + o_p(1),$$

under some conditions including $Z_1, \ldots, Z_n$ are iid and smoothness of ψ.

- The proof is based on Taylor series expansion (a deterministic tool):

$$0 = \sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) \; \approx \; \sum_{i=1}^{n} \psi(Z_i, \theta) + \sum_{i=1}^{n} \dot{\psi}(Z_i, \theta)(\hat{\theta} - \theta).$$

Let's Remember Cramér

- Suppose $Z_1, \ldots, Z_n$ are observations and we consider estimtor $\hat{\theta}$ satisfying

$$\textstyle\sum_{i=1}^n \psi(Z_i, \hat{\theta}_n) = 0.$$

- MLE, OLS, GLMs and many more estimators are all obtained this way.
- How to do inference using $\hat{\theta}_n$?
- The classical proof of Cramér (1946) proves the Bahadur representation:

$$\sqrt{n}(\hat{\theta} - \theta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbb{E}[\dot{\psi}(Z_1, \theta)])^{-1} \psi(Z_i, \theta) + o_p(1),$$

under some conditions including $Z_1, \ldots, Z_n$ are iid and smoothness of ψ.

- The proof is based on Taylor series expansion (a deterministic tool):

$$0 = \sum_{i=1}^n \psi(Z_i, \hat{\theta}_n) \;\approx\; \sum_{i=1}^n \psi(Z_i, \theta) + \sum_{i=1}^n \dot{\psi}(Z_i, \theta)(\hat{\theta} - \theta).$$

Do we need Z_i independent or even random? What is θ?

Importance of Bahadur Representation

- If $\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2} \sum_{i=1}^{n} W_i + o_p(1)$, for mean zero random variables $W_1, \ldots, W_n$, then by CLT (independent/dependent versions)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z, \quad \text{and} \quad \mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \le t) \to \mathbb{P}(Z \le t),$$

where $Z \sim N(0, \mathrm{Var}(W_1))$. (*Implies Inference.*)

Importance of Bahadur Representation

- If $\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2}\sum_{i=1}^{n} W_i + o_p(1)$, for mean zero random variables $W_1, \ldots, W_n$, then by CLT (independent/dependent versions)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z, \quad \text{and} \quad \mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \le t) \to \mathbb{P}(Z \le t),$$

  where $Z \sim N(0, \mathrm{Var}(W_1))$. (*Implies Inference.*)

- Suppose $\hat{\theta}_1, \hat{\theta}_2$ both satisfy the representation (together):

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}W_{1,i} \\ W_{2,i}\end{pmatrix} + o_p(1).$$

  Then for any t_1, t_2,

$$\mathbb{P}(\sqrt{n}(\hat{\theta}_1 - \theta_1) \le t_1, \sqrt{n}(\hat{\theta}_2 - \theta_2) \le t_2) \;\to\; \mathbb{P}(Z_1 \le t_1, Z_2 \le t_2),$$

  where $(Z_1, Z_2) \sim N(0, \mathrm{Var}(W_{1,1}, W_{2,1}))$. (*Implies simultaneous inference.*)

Importance of Bahadur Representation

- If $\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2}\sum_{i=1}^{n} W_i + o_p(1)$, for mean zero random variables $W_1, \ldots, W_n$, then by CLT (independent/dependent versions)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z, \quad \text{and} \quad \mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \le t) \to \mathbb{P}(Z \le t),$$

  where $Z \sim N(0, \text{Var}(W_1))$. (*Implies Inference.*)

- Suppose $\hat{\theta}_1, \hat{\theta}_2$ both satisfy the representation (together):

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}W_{1,i} \\ W_{2,i}\end{pmatrix} + o_p(1).$$

  Then for any t_1, t_2,

$$\mathbb{P}(\sqrt{n}(\hat{\theta}_1 - \theta_1) \le t_1, \sqrt{n}(\hat{\theta}_2 - \theta_2) \le t_2) \ \to \ \mathbb{P}(Z_1 \le t_1, Z_2 \le t_2),$$

  where $(Z_1, Z_2) \sim N(0, \text{Var}(W_{1,1}, W_{2,1}))$. (*Implies simultaneous inference.*)

- **Bahadur Representation $\Rightarrow$ (Simultaneous) Inference**

NBK Inequalities: Linear Regression[1]

[1] K. (2018), Deterministic Inequalities for Smooth M-estimators. arXiv:1809.05172
Thanks to Mateo Wirth, Bikram Karmakar.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \; = \; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- This holds for any set of observations (*with* $\hat{\Sigma}$ *invertible*).

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- This holds for any set of observations (*with $\hat{\Sigma}$ invertible).*
- Requires neither independence nor a (true linear) model.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- If Z_i satisfy a version of LLN: $\hat{\Sigma} \approx \Sigma$ for some Σ, then for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; (1 + o_p(1)) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma^{-1} X_i (Y_i - X_i^\top \beta),$$

Note: Σ does not have to be $\mathbb{E}\hat{\Sigma}$. Error is multiplicative not additive!!

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- Note $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- Note $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.
- Requires NO model assumptions, NO randomness assumptions, NO assumptions on d/n, NO independence/dependence assumptions.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- Note $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.
- Requires NO model assumptions, NO randomness assumptions, NO assumptions on d/n, NO independence/dependence assumptions.
- What are reasonable choices for Σ and β?

Canonical Choice of β and Σ

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1-\mathcal{D}^\Sigma)_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma.$$

- A natural choice for Σ satisfying $\mathcal{D}^\Sigma \approx 0$ is $\Sigma = \mathbb{E}[\hat{\Sigma}]$.

Canonical Choice of β and Σ

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1-\mathcal{D}^\Sigma)_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma.$$

- A natural choice for Σ satisfying $\mathcal{D}^\Sigma \approx 0$ is $\Sigma = \mathbb{E}[\hat{\Sigma}]$.
- If X_i's are fixed, then $\Sigma = \hat{\Sigma}$ and hence $\mathcal{D}^\Sigma = 0$.

Canonical Choice of β and Σ

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1-\mathcal{D}^\Sigma)_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma.$$

- A natural choice for Σ satisfying $\mathcal{D}^\Sigma \approx 0$ is $\Sigma = \mathbb{E}[\hat{\Sigma}]$.
- If X_i's are fixed, then $\Sigma = \hat{\Sigma}$ and hence $\mathcal{D}^\Sigma = 0$.
- Want $\hat{\beta} - \beta \approx 0$ or "equivalently" $n^{-1}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta) \approx 0$.

Canonical Choice of β and Σ

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1-\mathcal{D}^\Sigma)_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right\|_\Sigma .$$

- A natural choice for Σ satisfying $\mathcal{D}^\Sigma \approx 0$ is $\Sigma = \mathbb{E}[\hat{\Sigma}]$.
- If X_i's are fixed, then $\Sigma = \hat{\Sigma}$ and hence $\mathcal{D}^\Sigma = 0$.
- Want $\hat{\beta} - \beta \approx 0$ or "equivalently" $n^{-1}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^\top\beta) \approx 0$.
- At least require its expectation to be zero. Hence OLS target is

$$\beta := \mathrm{argmin}_{\theta\in\mathbb{R}^d}\ \sum_{i=1}^{n}\mathbb{E}[(Y_i - X_i^\top\theta)^2] \quad\Leftrightarrow\quad \sum_{i=1}^{n}\mathbb{E}[X_i(Y_i - X_i^\top\beta)] = 0.$$

Canonical Choice of β and Σ

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- A natural choice for Σ satisfying $\mathcal{D}^{\Sigma} \approx 0$ is $\Sigma = \mathbb{E}[\hat{\Sigma}]$.
- If X_i's are fixed, then $\Sigma = \hat{\Sigma}$ and hence $\mathcal{D}^{\Sigma} = 0$.
- Want $\hat{\beta} - \beta \approx 0$ or "equivalently" $n^{-1}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta) \approx 0$.
- At least require its expectation to be zero. Hence OLS target is

$$\beta := \mathrm{argmin}_{\theta\in\mathbb{R}^d}\sum_{i=1}^{n}\mathbb{E}[(Y_i - X_i^{\top}\theta)^2] \quad \Leftrightarrow \quad \sum_{i=1}^{n}\mathbb{E}[X_i(Y_i - X_i^{\top}\beta)] = 0.$$

- Under weak dependence and tail assumptions,

$$\|\hat{\beta}-\beta\|_{\Sigma} = O_p\left(\sqrt{\frac{d}{n}}\right), \quad \left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} = O_p\left(\frac{d}{n}\right).$$

Application 1: Berry–Esseen Bounds

Application 1: Berry–Esseen Bounds

Let $\mathcal{C}_d$ be the set of all convex sets in $\mathbb{R}^d$. Set $\mathcal{D}^\Sigma = \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_{op}$ and

$$\Sigma^{-1}K\Sigma^{-1} = \mathrm{Var}\left(n^{-1/2}\textstyle\sum_{i=1}^n \Sigma^{-1}X_i(Y_i - X_i^\top\beta)\right).$$

Theorem (Berry–Esseen bound for OLS)

For all $n \ge 1$ and any $A \in \mathcal{C}_d$,

$$\begin{aligned}
&\left|\mathbb{P}(n^{1/2}(\hat{\beta} - \beta) \in A) - \mathbb{P}\left(N(0, \Sigma^{-1}K\Sigma^{-1}) \in A\right)\right| \\
&\quad \le 5\left|\mathbb{P}\left(n^{-1/2}\sum_{i=1}^n \Sigma^{-1}X_i(Y_i - X_i^\top\beta) \in A\right) - \mathbb{P}(N(0, \Sigma^{-1}K\Sigma^{-1}) \in A)\right| \\
&\quad\quad + C\|\Sigma^{1/2}K^{-1}\Sigma^{1/2}\|_*^{1/4}\left[\frac{d^{1/4}\|K^{1/2}\|_{op}}{n^{1/2}} + \frac{d^{1/4}\|K^{1/2}\|_{HS}}{n^{3/4}}\right] \\
&\quad\quad + \mathbb{P}\left(\mathcal{D}^\Sigma \ge d^{1/4}/(n^{1/4}\sqrt{\log n})\right).
\end{aligned}$$

No model/randomness assumptions. Deterministic!!

Application 1: Berry–Esseen Bounds Contd.

- If $\mathcal{D}^{\Sigma} = O_p(\sqrt{d/n})$, then for any $A \in \mathcal{C}_d$,

$$\begin{aligned}&\left|\mathbb{P}(n^{1/2}(\hat{\beta} - \beta) \in A) - \mathbb{P}\left(N(0, \Sigma^{-1}K\Sigma^{-1}) \in A\right)\right| \\ &\quad \le C\left|\mathbb{P}\left(n^{-1/2}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta) \in A\right) - \mathbb{P}(N(0, \Sigma^{-1}K\Sigma^{-1}) \in A)\right|.\end{aligned}$$

- If X_i's are fixed then $\mathcal{D}^{\Sigma} = 0$ and inequality above holds with $C = 1$.
- If average converges to a normal, then $n^{1/2}(\hat{\beta} - \beta)$ converges to a normal. The above inequality makes this quantitative.
- Implies confidence regions, hypothesis tests.
- Can simultaneously infer about all coordinates of β.
- No model assumptions.

Application 2: Transformations of Response

Application 2: Transformations of Response

- In modeling, it is sometimes of interest to transform the response to match the assumptions like Gaussianity or homoscedasticity. Eg. Box–Cox family.
- Finding such "good" transformation involves some data snooping. Once again the inequality can be used to get a result for final estimator.
- Suppose $\mathcal{G}$ is a class of transformations under consideration and for each $g \in \mathcal{G}$, we have the OLS estimator

$$\hat{\beta}_g := \mathrm{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (g(Y_i) - X_i^\top \theta)^2.$$

For any $g \in \mathcal{G}$, define $\mathtt{Inf}_g(\theta) := n^{-1} \sum_{i=1}^n \Sigma^{-1} X_i (g(Y_i) - X_i^\top \theta)$.

Corollary (Bahadur Representation with Transformed Response)

For any set of observations $Z_i = (X_i, Y_i)$, any Σ, any $g \in \mathcal{G}$ and any $\beta_g \in \mathbb{R}^d$,

$$\left\| \hat{\beta}_g - \beta_g - \mathcal{Inf}_g(\beta_g) \right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1 - \mathcal{D}^\Sigma)_+} \| \mathcal{Inf}_g(\beta_g) \|_\Sigma.$$

In particular this holds for any random $\hat{g} \in \mathcal{G}$ chosen based on the data.

Application 3: Variable Selection

Application 3: Variable Selection

- More often than not, the set of covariates in a reported model is not the same as the set of covariates the analyst started with.
- Finding such "good" set of covariates involves some data snooping.
- Suppose $\mathcal{M}$ is a collection of models (*set of covariates*) and for each $M \in \mathcal{M}$, we have the OLS estimator

$$\hat{\beta}_M := \mathrm{argmin}_{\theta \in \mathbb{R}^{|M|}} \sum_{i=1}^n (Y_i - X_{i,M}^\top \theta)^2.$$

Set for any $M \in \mathcal{M}$, $\mathtt{Inf}_M(\theta) := n^{-1} \sum_{i=1}^n \Sigma_M^{-1} X_{i,M}(Y_i - X_{i,M}^\top \theta)$.

Corollary (Bahadur Representation with Variable Selection)

For any $M \in \mathcal{M}$, any Σ_M, and any $\beta_M \in \mathbb{R}^{|M|}$, we have

$$\left\| \hat{\beta}_M - \beta_M - \mathtt{Inf}_M(\beta_M) \right\|_{\Sigma_M} \le \frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)_+} \|\mathtt{Inf}_M(\beta_M)\|_{\Sigma_M},$$

where $\mathcal{D}_M^\Sigma := \|\Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2} - I_{|M|}\|_{op}$. In particular M can be randomly chosen based on the data.

Rates in a Special Case

- Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and satisfy

$$\mathbb{P}(|\Sigma_M^{-1/2} X_{i,M}^\top \theta| \ge t) \le 2\exp\left(-\frac{t^2}{C^2}\right) \quad \text{for all} \quad \theta, 1 \le i \le n,$$

  and for some constant C,

$$\mathbb{E}[Y_i^4] \le C^2 < \infty \quad \text{for all} \quad 1 \le i \le n.$$

- Then **uniformly** over $1 \le s \le d$,

$$\max_{|M|=s} \max\{\mathcal{D}_M^\Sigma,\ \|\mathtt{Inf}_M(\beta_M)\|_{\Sigma_M}\} = O_p\left(\sqrt{\frac{s\log(ed/s)}{n}}\right).$$

- Hence **uniformly** over $1 \le s \le d$,

$$\max_{|M|=s} \|\hat{\beta}_M - \beta_M\|_{\Sigma_M} = O_p\left(\sqrt{\frac{s\log(ed/s)}{n}}\right),$$

  and

$$\max_{|M|=s} \left\|\hat{\beta}_M - \beta_M - \mathtt{Inf}_M(\beta_M)\right\|_{\Sigma_M} = O_p\left(\frac{s\log(ed/s)}{n}\right).$$

Implication: Post-selection Inference

- **Uniform linear representation** result allows us to claim

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \approx \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty,$$

for some vector functions ψ_M.

- **High-dimensional CLT** implies

$$\max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty \overset{\mathcal{L}}{\approx} \max_{M\in\mathcal{M}} \|G_M\|_\infty,$$

for some Gaussian process $(G_M)_{M\in\mathcal{M}}$.

- **Corresponding multiplier bootstrap** implies

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \overset{\mathcal{L}}{\approx} \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n} g_i\hat{\psi}_M(X_i, Y_i)\right\|_\infty \quad \text{Cond. on } (X_i, Y_i),$$

for $g_1, \ldots, g_n \sim N(0,1)$ (iid).

Summary and Conclusions

Summary and Conclusions

- We have introduced the idea of studying estimators in a deterministic way.
- NBK inequalities solve almost all problems about an estimator in one shot:
  - They imply Berry–Esseen type bounds and hence (finite sample) normal approximation results can follow.
  - They allow for understanding the effects of increasing dependence between observations, increasing dimension.
- Importantly in the context of reproducibility, NBK inequalities allow study of estimators obtained after data snooping.
- In particular, it solves the problem of post-selection inference in a unified way and in the most general framework available till date.
- Application of a (proximal) variant of Newton's method for penalized or constrained estimators leads to first order expansion results.

Summary and Conclusions

- We have introduced the idea of studying estimators in a deterministic way.
- NBK inequalities solve almost all problems about an estimator in one shot:
  - They imply Berry–Esseen type bounds and hence (finite sample) normal approximation results can follow.
  - They allow for understanding the effects of increasing dependence between observations, increasing dimension.
- Importantly in the context of reproducibility, NBK inequalities allow study of estimators obtained after data snooping.
- In particular, it solves the problem of post-selection inference in a unified way and in the most general framework available till date.
- Application of a (proximal) variant of Newton's method for penalized or constrained estimators leads to first order expansion results.

Thanks!

NBK Inequalities: Logistic/Poisson Regression[2]

[2] K. (2018), Deterministic Inequalities for Smooth M-estimators. arXiv:1809.05172
Thanks to Mateo Wirth, Bikram Karmakar.

Logistic/Poisson Regression

- For either $\psi(u) = \log(1+\exp(u))$, Logistic or $\psi(u) = \exp(u)$ Poisson, let

$$\hat{\beta} := \mathrm{argmin}_{\theta\in\mathbb{R}^d}\ L_n(\theta), \quad \text{where} \quad L_n(\theta) := n^{-1}\sum_{i=1}^n \left[\psi(X_i^\top\theta) - Y_iX_i^\top\theta\right],$$

- Define for any $\theta \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$, $\mathcal{D}^\Sigma(\theta) := \|\Sigma^{-1/2}\ddot{L}_n(\theta)\Sigma^{-1/2} - I_d\|_{op}$.

Theorem

For any $\beta \in \mathbb{R}^d$ and any $\Sigma \in \mathbb{R}^{d\times d}$, if

$$\max_{1\le i\le n}\|\Sigma^{-1/2}X_i\| \times \|\Sigma^{-1}\dot{L}_n(\beta)\|_\Sigma \le 0.19(1-\mathcal{D}^\Sigma(\beta))_+, \tag{1}$$

then

$$\frac{\|\hat{\beta}_n - \beta + \Sigma^{-1}\dot{L}_n(\beta)\|_\Sigma}{\|\Sigma^{-1}\dot{L}_n(\beta)\|_\Sigma} \le \frac{\mathcal{D}^\Sigma(\beta)}{(1-\mathcal{D}^\Sigma(\beta))_+} + \frac{10\max_i\|\Sigma^{-1/2}X_i\|\|\Sigma^{-1}\dot{L}_n(\beta)\|_\Sigma}{(1-\mathcal{D}^\Sigma(\beta))_+^2}.$$

Assumption (1) arises becasue of non-linearity of estimating function $\dot{L}_n(\theta)$.

Some Comments

- For $\hat{\beta}$ defined as a minimizer of $L_n(\cdot)$, a canonical choice of Σ, β is given by

$$\beta := \underset{\theta \in \mathbb{R}^d}{\text{argmin}}\ \mathbb{E}[L_n(\theta)] \quad \text{and} \quad \Sigma := \mathbb{E}[\ddot{L}_n(\beta)].$$

Some Comments

- For $\hat{\beta}$ defined as a minimizer of $L_n(\cdot)$, a canonical choice of Σ, β is given by

$$\beta := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}[L_n(\theta)] \quad \text{and} \quad \Sigma := \mathbb{E}[\ddot{L}_n(\beta)].$$

- For independent as well as a weakly dependent sub-Gaussian observations,

$$\max\{\mathcal{D}^{\Sigma}(\beta),\ \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}\} = O_p(\sqrt{d/n}),$$

which implies

$$\|\hat{\beta}_n - \beta + \Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma} = O_p\left(\sqrt{\frac{d}{n}}\right)\|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}.$$

Some Comments

- For $\hat{\beta}$ defined as a minimizer of $L_n(\cdot)$, a canonical choice of Σ, β is given by

$$\beta := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\ \mathbb{E}[L_n(\theta)] \quad \text{and} \quad \Sigma := \mathbb{E}[\ddot{L}_n(\beta)].$$

- For independent as well as a weakly dependent sub-Gaussian observations,

$$\max\{\mathcal{D}^{\Sigma}(\beta),\ \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}\} = O_p(\sqrt{d/n}),$$

  which implies

$$\|\hat{\beta}_n - \beta + \Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma} = O_p\left(\sqrt{\frac{d}{n}}\right)\|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}.$$

- Following the result for logistic and Poisson regression, applications like transformations, variable selection can be carried out easily.

Some Comments

- For $\hat{\beta}$ defined as a minimizer of $L_n(\cdot)$, a canonical choice of Σ, β is given by

$$\beta := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}[L_n(\theta)] \quad \text{and} \quad \Sigma := \mathbb{E}[\ddot{L}_n(\beta)].$$

- For independent as well as a weakly dependent sub-Gaussian observations,

$$\max\{\mathcal{D}^{\Sigma}(\beta),\ \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}\} = O_p(\sqrt{d/n}),$$

which implies

$$\|\hat{\beta}_n - \beta + \Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma} = O_p\left(\sqrt{\frac{d}{n}}\right)\|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}.$$

- Following the result for logistic and Poisson regression, applications like transformations, variable selection can be carried out easily.
- These inequalities are also proved for Cox proportional hazards model, Non-linear least squares, Equality constrained M-estimators among others.

Application: Post-selection Inference

- **Uniform linear representation** result allows us to claim

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \;\approx\; \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty,$$

  for some vector functions ψ_M.

- **High-dimensional CLT** implies

$$\max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty \;\overset{\mathcal{L}}{\approx}\; \max_{M\in\mathcal{M}} \|G_M\|_\infty,$$

  for some Gaussian process $(G_M)_{M\in\mathcal{M}}$.

- **Corresponding multiplier bootstrap** implies

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \;\overset{\mathcal{L}}{\approx}\; \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n} g_i\hat{\psi}_M(X_i, Y_i)\right\|_\infty \quad \text{Cond. on } (X_i, Y_i),$$

  for $g_1, \ldots, g_n \sim N(0, 1)$ (iid).

PoSI Contd.

- To finish inference, need to compute
$$\max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n} g_i\hat{\psi}_M(X_i, Y_i)\right\|_{\infty},$$
for a given set of models $\mathcal{M}$.
- Number the models in $\mathcal{M}$ as $1, 2, \ldots, N$. We have
$$x_j \;:=\; \left\|\frac{1}{n}\sum_{i=1}^{n} g_i\hat{\psi}_j(X_i, Y_i)\right\|_{\infty}.$$
- Need to compute (at least approximately)
$$\|x\|_\infty = \max_{1\le j\le N} |x_j|,$$
for the vector $x = (x_1, \ldots, x_N)$.

Maximum Computation[3]

- Observe that

$$\left(\frac{1}{N}\sum_{j=1}^{N} x_j^q\right)^{1/q} \le \|x\|_\infty \le N^{1/q}\left(\frac{1}{N}\sum_{j=1}^{N} x_j^q\right)^{1/q}.$$

- If W is a random variable drawn uniformly from $\{x_1, \ldots, x_N\}$, then

$$(\mathbb{E}[W^q])^{1/q} \le \|x\|_\infty \le N^{1/q}(\mathbb{E}[W^q])^{1/q}.$$

- Hence (multiplicatively) approximating the maximum is same as approximating the **expectation** of a random variable given access to independent draws.

How many draws required to find $\mathbb{E}[W^q]$ upto a factor of $(1 \pm \varepsilon)$?

[3] Joint work (in progress) with Junhui Cai

Summary

- We have shown how the **analysis of Newton's method** can be used to derive **finite sample results for M-estimators**.
- This idea allow "easier" study of constrained/penalized M-estimators.
- Connections to AMP??
- These results imply post-selection inference for various estimation procedures including **GLMs**, **Cox Model**, **NonLinear Least Squares**, **Equality Constrained MLE**.
- Realizing PoSI in practice requires solving a maximum problem.
- $$\text{PoSI} \;\rightarrow\; \text{Maximum Estimation} \;\rightarrow\; \text{Mean Estimation.}$$
- achievable sample complexity bounds for maximum??

Maximum Computation (Contd.)

- An estimator $\hat{E}_W$ of $\mathbb{E}[W] > 0$ is an (ε, δ) approximate if

$$\mathbb{P}\left(\left|\frac{\hat{E}_W}{\mathbb{E}[W]} - 1\right| \le \varepsilon\right) \;\ge\; 1 - \delta.$$

- If a random variable $W \ge 0$ is known to satisfy

$$\mathrm{Var}(W) \;\le\; L^2(\mathbb{E}[W])^2$$

  then

$$n_{\varepsilon,\delta} \;\asymp\; \frac{2L^2}{\varepsilon^2}\log\left(\frac{1}{\sqrt{2\pi}\delta}\right).$$

- If a random variable $W \in [0, B]$ for some known B, then

$$n_{\varepsilon,\delta} \;\asymp\; C\max\left\{\frac{\mathrm{Var}(W)}{\varepsilon^2(\mathbb{E}[W])^2}, \frac{B}{\varepsilon\mathbb{E}[W]}\right\}\log\left(\frac{1}{\delta}\right),$$

  for some universal constant $C > 0$.