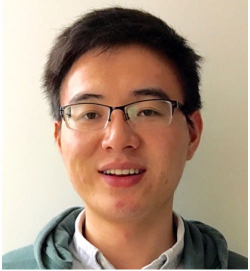
Beyond Gaussian Approximation: Bootstrap in Large Scale Simultaneous Inference

Cun-Hui Zhang, Rutgers University

January 27, 2017

Van Dantzig Seminar, University of Amsterdam
Thanks for the invitation!



Hang Deng

The problem

- Let $X_i = (X_{i,1}, \ldots, X_{i,p})^T$ be independent vectors in $\mathbb{R}^p$,

$$T_n = \max_{j \le p} \sum_{i=1}^n (X_{i,j} - \mathbb{E}X_{i,j})/\sqrt{n}$$

- Let X_i^* be bootstrapped X_i,

$$T_n^* = \max_{j \le p} \sum_{i=1}^n (X_{i,j}^* - \mathbb{E}^* X_{i,j}^*)/\sqrt{n},$$

$$t_\alpha^* : \mathbb{P}^*\{T_n^* \ge t_\alpha^*\} = \alpha$$

The problem

- Let $X_i = (X_{i,1}, \dots, X_{i,p})^T$ be independent vectors in $\mathbb{R}^p$,

$$T_n = \max_{j \le p} \sum_{i=1}^n (X_{i,j} - \mathbb{E}X_{i,j})/\sqrt{n}$$

- Let X_i^* be bootstrapped X_i,

$$T_n^* = \max_{j \le p} \sum_{i=1}^n (X_{i,j}^* - \mathbb{E}^* X_{i,j}^*)/\sqrt{n},$$

$$t_\alpha^* : \mathbb{P}^*\{T_n^* \ge t_\alpha^*\} = \alpha$$

- Under what conditions is the bootstrap consistent,

$$\left|\mathbb{P}\{T_n \ge t_\alpha^*\} - \alpha\right| = o_P(1)?$$

- This consistency in confidence level is a consequence of

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

  i.e. consistency in the Kolmogorov-Smirnov distance

Motivation, some examples

- The non-Gaussian many means problem, $\mu_j = \mathbb{E}\sum_{i=1}^{n} X_{i,j}/n$,

$$\mathbb{P}\Big\{ \max_{1\le j\le p} \big|\widehat{\mu}_j - \mu_j\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha$$

Motivation, some examples

- The non-Gaussian many means problem, $\mu_j = \mathbb{E}\sum_{i=1}^{n} X_{i,j}/n$,

$$\mathbb{P}\Big\{ \max_{1\le j\le p} \big|\widehat{\mu}_j - \mu_j\big| \le t^*_\alpha/\sqrt{n}\Big\} \approx 1-\alpha$$

- Sure screening in regression (Fan & Lv, 08)

$$\mathbb{P}\Big\{ \max_{1\le j\le p} \big|\widehat{\theta}_j - \theta_j\big| \le t^*_\alpha/\sqrt{n}\Big\} \approx 1-\alpha,$$
$$\theta_j = \mathbb{E}\Big[\boldsymbol{x}_j^T \boldsymbol{y}/n \Big| \boldsymbol{X}\Big] \;\text{ or }\; \theta_j = \mathbb{E}\boldsymbol{x}_j^T \boldsymbol{y}/n$$

Motivation, some examples

- The non-Gaussian many means problem, $\mu_j = \mathbb{E}\sum_{i=1}^{n} X_{i,j}/n$,

$$\mathbb{P}\Big\{\max_{1\le j\le p}\big|\widehat{\mu}_j - \mu_j\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha$$

- Sure screening in regression (Fan & Lv, 08)

$$\mathbb{P}\Big\{\max_{1\le j\le p}\big|\widehat{\theta}_j - \theta_j\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha,$$

$$\theta_j = \mathbb{E}\Big[\boldsymbol{x}_j^T\boldsymbol{y}/n\Big|\boldsymbol{X}\Big] \quad \text{or} \quad \theta_j = \mathbb{E}\boldsymbol{x}_j^T\boldsymbol{y}/n$$

- Testing the equality of two matrices (Cai et al 13, Chang et al, 15)

$$\mathbb{P}\Big\{\max_{1\le j,k\le p}\big|\widehat{\theta}_{j,k} - \theta_{j,k}\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha,$$

$$\theta_{j,k} = \mathbb{E}\boldsymbol{x}_j^T\boldsymbol{x}_k/n - \mathbb{E}\boldsymbol{y}_j^T\boldsymbol{y}_k/n$$

Motivation, some examples

- The non-Gaussian many means problem, $\mu_j = \mathbb{E}\sum_{i=1}^n X_{i,j}/n$,

$$\mathbb{P}\left\{ \max_{1\le j\le p} \left|\widehat{\mu}_j - \mu_j\right| \le t_\alpha^*/\sqrt{n}\right\} \approx 1-\alpha$$

- Sure screening in regression (Fan & Lv, 08)

$$\mathbb{P}\left\{ \max_{1\le j\le p} \left|\widehat{\theta}_j - \theta_j\right| \le t_\alpha^*/\sqrt{n}\right\} \approx 1-\alpha,$$
$$\theta_j = \mathbb{E}\left[\boldsymbol{x}_j^T \boldsymbol{y}/n \middle| \boldsymbol{X}\right] \quad \text{or} \quad \theta_j = \mathbb{E}\boldsymbol{x}_j^T\boldsymbol{y}/n$$

- Testing the equality of two matrices (Cai et al 13, Chang et al, 15)

$$\mathbb{P}\left\{ \max_{1\le j,k\le p} \left|\widehat{\theta}_{j,k} - \theta_{j,k}\right| \le t_\alpha^*/\sqrt{n}\right\} \approx 1-\alpha,$$
$$\theta_{j,k} = \mathbb{E}\boldsymbol{x}_j^T\boldsymbol{x}_k/n - \mathbb{E}\boldsymbol{y}_j^T\boldsymbol{y}_k/n$$

- Ridges and density level sets (Chen et al, 15, 16)

Motivation, some examples

- The non-Gaussian many means problem, $\mu_j = \mathbb{E}\sum_{i=1}^n X_{i,j}/n$,

$$\mathbb{P}\Big\{ \max_{1\le j\le p} \big|\widehat{\mu}_j - \mu_j\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha$$

- Sure screening in regression (Fan & Lv, 08)

$$\mathbb{P}\Big\{ \max_{1\le j\le p} \big|\widehat{\theta}_j - \theta_j\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha,$$
$$\theta_j = \mathbb{E}\Big[\boldsymbol{x}_j^T \boldsymbol{y}/n \Big| \boldsymbol{X}\Big] \quad \text{or} \quad \theta_j = \mathbb{E}\boldsymbol{x}_j^T \boldsymbol{y}/n$$

- Testing the equality of two matrices (Cai et al 13, Chang et al, 15)

$$\mathbb{P}\Big\{ \max_{1\le j,k\le p} \big|\widehat{\theta}_{j,k} - \theta_{j,k}\big| \le t_\alpha^*/\sqrt{n}\Big\} \approx 1-\alpha,$$
$$\theta_{j,k} = \mathbb{E}\boldsymbol{x}_j^T \boldsymbol{x}_k/n - \mathbb{E}\boldsymbol{y}_j^T \boldsymbol{y}_k/n$$

- Ridges and density level sets (Chen et al, 15, 16)
- Simultaneous inference about many regression coefficients via de-biasing the Lasso or PLSE (Z-Zhang, 14; Belloni et al, 14, 15; Cheng-Zhang, 16, Dezeure et al, 16)

Bootstrap methods

- Efron's (79) empirical bootstrap,

$$\mathbb{P}^*\left\{X_i^* \leftarrow X_k - \overline{X}\right\} = \frac{1}{n},\ k = 1, \ldots, n, i = 1, \ldots, n$$

- Multiplier/wild bootstrap (Wu, 86; Liu, 88; Liu-Singh, 92; Mammen, 93),

$$X_i^* = W_i(X_i - \overline{X}),\ \mathbb{E}W_i = 0,\ \mathbb{E}W_i^2 = 1$$

- Residual bootstrap in regression (Efron, 79)

Consistency of bootstrap in high-dimension

- Donsker classes: Giné and Zinn (90)

Consistency of bootstrap in high-dimension

- Donsker classes: Giné and Zinn (90)
- $n \gg p^{7/2}$ for all convex sets: Nagaev (76), Senatov (80), Sazonov (81), Götze (91, 93), Bentkus (86, 03)

Consistency of bootstrap in high-dimension

- Donsker classes: Giné and Zinn (90)
- $n \gg p^{7/2}$ for all convex sets: Nagaev (76), Senatov (80), Sazonov (81), Götze (91, 93), Bentkus (86, 03)
- $n \gg (\log p)^7$ for maxima: Chernozhukov et al (13, 14)

Consistency of bootstrap in high-dimension

- Donsker classes: Giné and Zinn (90)
- $n \gg p^{7/2}$ for all convex sets: Nagaev (76), Senatov (80), Sazonov (81), Götze (91, 93), Bentkus (86, 03)
- $n \gg (\log p)^7$ for maxima: Chernozhukov et al (13, 14)
- Gaussian approximation/second moment match
    - Stein (72, 81)
    - Lindeberg (22)

The Stein method: Assume $\mathbb{E}X_i = 0$. Let $f(x_1, \ldots, x_n)$ be a smooth function of the sum $x_1 + \cdots + x_n$ and $Y_i \sim N(0, \mathbb{E}X_i^{\otimes 2})$.

- Slepian's (62) smart interpolation: $Z_i(t) = \cos(t)X_i + \sin(t)Y_i$

$$\mathbb{E}f(\boldsymbol{Y}) - \mathbb{E}f(\boldsymbol{X}) = \int_0^{\pi/2} \sum_{i=1}^{n} \mathbb{E}\Big\langle f^{(1)}(\boldsymbol{Z}(t)), \dot{Z}_i(t)\Big\rangle dt$$

The Stein method: Assume $\mathbb{E}X_i = 0$. Let $f(x_1, \ldots, x_n)$ be a smooth function of the sum $x_1 + \cdots + x_n$ and $Y_i \sim N(0, \mathbb{E}X_i^{\otimes 2})$.

- Slepian's (62) smart interpolation: $Z_i(t) = \cos(t)X_i + \sin(t)Y_i$

$$\mathbb{E}f(\boldsymbol{Y}) - \mathbb{E}f(\boldsymbol{X}) = \int_0^{\pi/2} \sum_{i=1}^{n} \mathbb{E}\Big\langle f^{(1)}(\boldsymbol{Z}(t)), \dot{Z}_i(t)\Big\rangle dt$$

- Stein's (81) leave-one-out method:

$$\mathbb{E}\Big\langle f^{(1)}(\boldsymbol{Z}(t)), \dot{Z}_i(t)\Big\rangle = \int_0^1 \mathbb{E}\Big\langle f^{(3)}(\boldsymbol{Z}_{-i}(t), uZ_i(t)), Z_i^{\otimes 2}(t) \otimes \dot{Z}_i(t)\Big\rangle du$$

due to $\mathbb{E}Z_i(t) \otimes \dot{Z}_i(t) = \sin(t)\cos(t)\mathbb{E}X_i^{\otimes 2} - \sin(t)\cos(t)\mathbb{E}Y_i^{\otimes 2} = 0$

The Stein method: Assume $\mathbb{E}X_i = 0$. Let $f(x_1, \ldots, x_n)$ be a smooth function of the sum $x_1 + \cdots + x_n$ and $Y_i \sim N(0, \mathbb{E}X_i^{\otimes 2})$.

- Slepian's (62) smart interpolation: $Z_i(t) = \cos(t)X_i + \sin(t)Y_i$

$$\mathbb{E}f(\boldsymbol{Y}) - \mathbb{E}f(\boldsymbol{X}) = \int_0^{\pi/2} \sum_{i=1}^{n} \mathbb{E}\Big\langle f^{(1)}(\boldsymbol{Z}(t)), \dot{Z}_i(t)\Big\rangle dt$$

- Stein's (81) leave-one-out method:

$$\mathbb{E}\Big\langle f^{(1)}(\boldsymbol{Z}(t)), \dot{Z}_i(t)\Big\rangle = \int_0^1 \mathbb{E}\Big\langle f^{(3)}(\boldsymbol{Z}_{-i}(t), uZ_i(t)), Z_i^{\otimes 2}(t) \otimes \dot{Z}_i(t)\Big\rangle du$$

due to $\mathbb{E}Z_i(t) \otimes \dot{Z}_i(t) = \sin(t)\cos(t)\mathbb{E}X_i^{\otimes 2} - \sin(t)\cos(t)\mathbb{E}Y_i^{\otimes 2} = 0$

- However,

$$\mathbb{E}Z_i^{\otimes 2}(t) \otimes \dot{Z}_i(t) = \sin(t)\cos^2(t)\mathbb{E}X_i^{\otimes 3} - \sin^2(t)\cos(t)\mathbb{E}Y_i^{\otimes 3} \neq 0$$

even when $\mathbb{E}X_i^{\otimes 3} = \mathbb{E}Y_i^{\otimes 3} \neq 0$

The benefit of third moment match in bootstrap

- Fixed p: Singh (81), Bickel and Freedman (81), Hall (88), Liu (88), Manmen (93)

The benefit of third moment match in bootstrap

- Fixed p: Singh (81), Bickel and Freedman (81), Hall (88), Liu (88), Manmen (93)
- Comparison in the Edgeworth expansion, a refinement of the CLT

The benefit of third moment match in bootstrap

- Fixed p: Singh (81), Bickel and Freedman (81), Hall (88), Liu (88), Manmen (93)
- Comparison in the Edgeworth expansion, a refinement of the CLT
- Large p: We are interested in taking advantage of the third moment match in regimes where the existing theory of Gaussian approximation does not apply

The benefit of third moment match in bootstrap

- Fixed p: Singh (81), Bickel and Freedman (81), Hall (88), Liu (88), Manmen (93)
- Comparison in the Edgeworth expansion, a refinement of the CLT
- Large p: We are interested in taking advantage of the third moment match in regimes where the existing theory of Gaussian approximation does not apply
- The consistency of the bootstrap may depend on the 3rd moment property of X_i

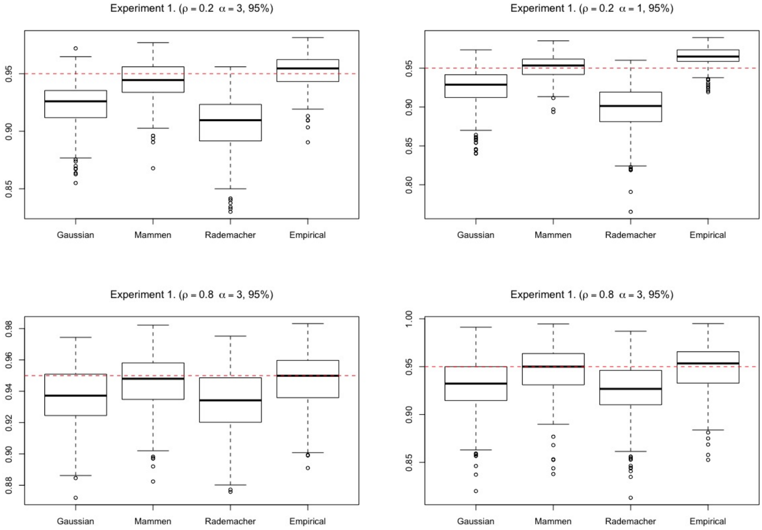
The benefit of third moment match in bootstrap

- Fixed p: Singh (81), Bickel and Freedman (81), Hall (88), Liu (88), Manmen (93)
- Comparison in the Edgeworth expansion, a refinement of the CLT
- Large p: We are interested in taking advantage of the third moment match in regimes where the existing theory of Gaussian approximation does not apply
- The consistency of the bootstrap may depend on the 3rd moment property of X_i

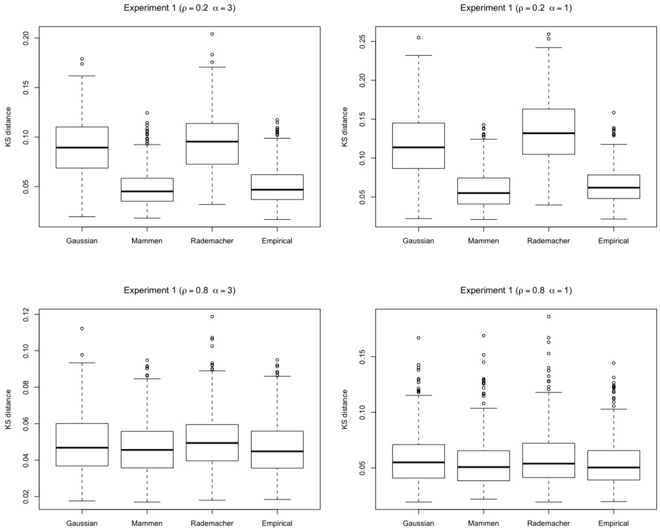
Consistency and second moment properties in the low-dimensional case

- Athreya (1986), Giné and Zinn (1989): For iid $X_i \in \mathbb{R}$, the empirical bootstrap for the mean is consistent if and only if X_1 is in the domain of attraction of the normal law.

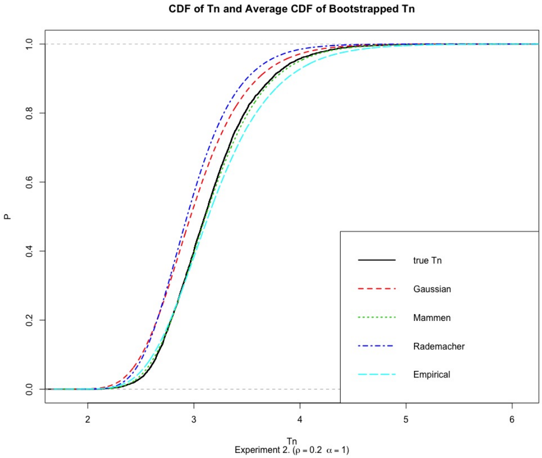
Some simulation results: Coverage probability



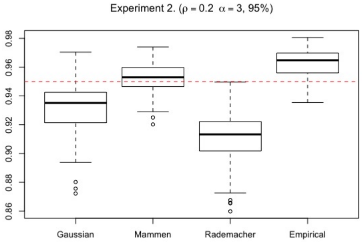
Some simulation results: Kolmogorov-Smirnov distance

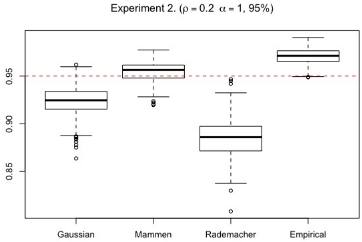


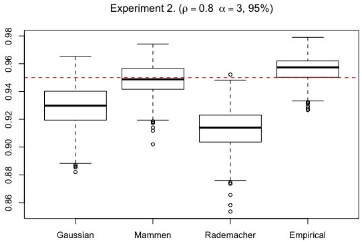
Some simulation results: the bias

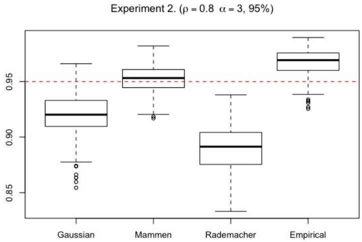


Some more simulation results: Coverage probability

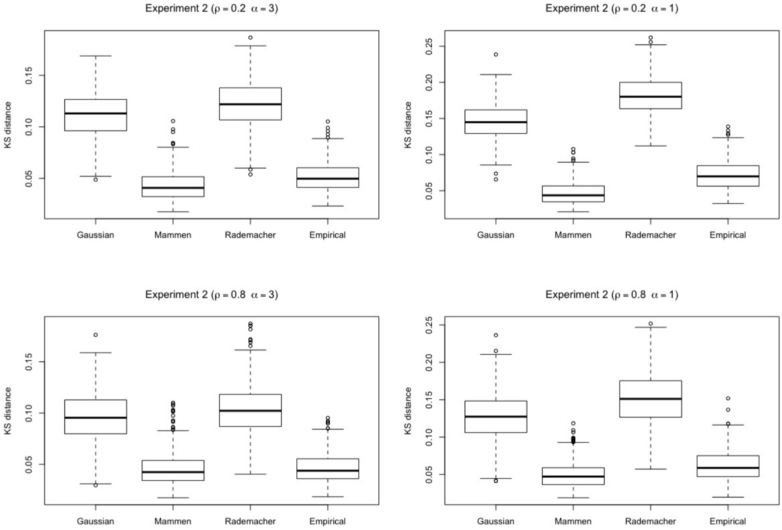








Some more simulation results: Kolmogorov-Smirnov distance



Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \ldots, X_i, Y_{i+1}, \ldots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{ f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1}) \Big\},$$

Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \dots, X_i, Y_{i+1}, \dots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{ f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1}) \Big\},$$

- Leave-one-out: $\boldsymbol{U}_i = (X_1, \dots, X_{i-1}, 0, Y_{i+1}, \dots, Y_n)$
- Taylor expansion

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{1}{m!} \sum_{i=1}^{n} \Big\langle \mathbb{E}f^{(m)}(\boldsymbol{U}_i), \mathbb{E}X_i^{\otimes m} - \mathbb{E}Y^{\otimes m} \Big\rangle + \mathrm{Rem}$$

Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \ldots, X_i, Y_{i+1}, \ldots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{ f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1}) \Big\},$$

- Leave-one-out: $\boldsymbol{U}_i = (X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$
- Taylor expansion

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{1}{m!} \sum_{i=1}^{n} \Big\langle \mathbb{E}f^{(m)}(\boldsymbol{U}_i), \mathbb{E}X_i^{\otimes m} - \mathbb{E}Y^{\otimes m} \Big\rangle + \mathrm{Rem}$$

- This automatically allows comparison of higher moments, with $m^* > 3$

Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \ldots, X_i, Y_{i+1}, \ldots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{ f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1}) \Big\},$$

- Leave-one-out: $\boldsymbol{U}_i = (X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$
- Taylor expansion

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{1}{m!} \sum_{i=1}^{n} \Big\langle \mathbb{E}f^{(m)}(\boldsymbol{U}_i), \mathbb{E}X_i^{\otimes m} - \mathbb{E}Y^{\otimes m} \Big\rangle + \mathrm{Rem}$$

- This automatically allows comparison of higher moments, with $m^* > 3$
- Gaussian approximation (Chatterjee, 06): $m^* = 3$

Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \dots, X_i, Y_{i+1}, \dots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1})\Big\},$$

- Leave-one-out: $\boldsymbol{U}_i = (X_1, \dots, X_{i-1}, 0, Y_{i+1}, \dots, Y_n)$
- Taylor expansion

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{1}{m!} \sum_{i=1}^{n} \Big\langle \mathbb{E}f^{(m)}(\boldsymbol{U}_i), \mathbb{E}X_i^{\otimes m} - \mathbb{E}Y^{\otimes m} \Big\rangle + \mathrm{Rem}$$

- This automatically allows comparison of higher moments, with $m^* > 3$
- Gaussian approximation (Chatterjee, 06): $m^* = 3$
- A problem is the dependence of $\mathbb{E}f^{(m)}(\boldsymbol{U}_i)$ on i

Consistency of the multiplier/wild bootstrap: Suppose that $X_i \in \mathbb{R}^p$ are independent, W_i are iid, and $\{W_i\}$ is independent of $\{X_i\}$. Suppose

$$\mathbb{E}W_i = 0, \quad \mathbb{E}W_i^2 = \mathbb{E}W_i^3 = 1.$$

Let $X_i^* = W_i(X_i - \overline{X})$. Define

$$T_n = \max_{j \le p} \sum_{i=1}^{n} \frac{X_i - \mathbb{E}X_i}{n^{1/2}}, \quad T_n^* = \max_{j \le p} \sum_{i=1}^{n} \frac{X_i^*}{n^{1/2}}.$$

Then, under 4th moment and certain tail probability conditions,

$$\left| \mathbb{P}\Big\{ T_n \le t_\alpha^* \Big\} - \alpha \right| \lesssim \left(\frac{(\log p)^4 \log(1/\epsilon_0)}{n} \right)^{1/6} + \epsilon_0 + \left(\frac{\log^5 p}{n} \right)^{1/5}$$

and

$$n \gg \log^5 p \ \Rightarrow \ \sup_t \left| \mathbb{P}\Big\{ T_n \le t \Big\} - \mathbb{P}^*\Big\{ T_n^* \le t \Big\} \right| = o_P(1)$$

Consistency of bootstrap

- Wild bootstrap with $\mathbb{E}W_i^3 = 1$ or $\mathbb{E}X_i^{\otimes 3} = 0$: Under 4th moment and tail probability conditions,

$$\left|\mathbb{P}\left\{ T_n \leq t_\alpha^* \right\} - \alpha\right| \lesssim \left(\frac{(\log p)^4 \log(1/\epsilon_0)}{n}\right)^{1/6} + \epsilon_0 + \left(\frac{\log^5 p}{n}\right)^{1/5}$$

Consistency of bootstrap

- Wild bootstrap with $\mathbb{E}W_i^3 = 1$ or $\mathbb{E}X_i^{\otimes 3} = 0$: Under 4th moment and tail probability conditions,

$$\left|\mathbb{P}\left\{ T_n \le t_\alpha^* \right\} - \alpha\right| \lesssim \left(\frac{(\log p)^4 \log(1/\epsilon_0)}{n}\right)^{1/6} + \epsilon_0 + \left(\frac{\log^5 p}{n}\right)^{1/5}$$

- Empirical bootstrap: Under 4th moment and tail probability conditions,

$$n \gg \log^5 p \;\Rightarrow\; \sup_t \left|\mathbb{P}\left\{ T_n \le t \right\} - \mathbb{P}^*\left\{ T_n^* \le t \right\}\right| = o_P(1)$$

Consistency of bootstrap

- Wild bootstrap with $\mathbb{E}W_i^3 = 1$ or $\mathbb{E}X_i^{\otimes 3} = 0$: Under 4th moment and tail probability conditions,

$$\left|\mathbb{P}\left\{T_n \le t_\alpha^*\right\} - \alpha\right| \lesssim \left(\frac{(\log p)^4 \log(1/\epsilon_0)}{n}\right)^{1/6} + \epsilon_0 + \left(\frac{\log^5 p}{n}\right)^{1/5}$$

- Empirical bootstrap: Under 4th moment and tail probability conditions,

$$n \gg \log^5 p \;\Rightarrow\; \sup_t \left|\mathbb{P}\left\{T_n \le t\right\} - \mathbb{P}^*\left\{T_n^* \le t\right\}\right| = o_P(1)$$

- Gaussian wild bootstrap: Under 3rd moment and tail probability conditions,

$$\sup_t \left|\mathbb{P}\left\{T_n \le t\right\} - \mathbb{P}^*\left\{T_n^* \le t\right\}\right| \lesssim \left(\frac{\log^7 p}{n}\right)^{1/6}$$

A general comparison theorem: Let

$$\mu^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E} X_i^{\otimes m}, \quad \nu^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E} Y_i^{\otimes m}$$

Under certain smoothness and permutation invariance conditions on f,

$$\mathbb{E} f(\boldsymbol{X}) - \mathbb{E} f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{n}{m!} \left\langle F^{(m)}, \mu^{(m)} - \nu^{(m)} \right\rangle + \mathrm{Rem}$$

with $m^* \geq 2$

$$\left|\mathrm{Rem}\right| \leq C \langle F_{\max}^{(m^*)}, \mu_{\max}^{(m^*)} + \nu_{\max}^{(m^*)} \rangle$$

where $F^{(m)}$ and $F_{\max}^{(m)}$ are respectively weighted averages of $\mathbb{E} f^{(m)}(Z_1, \ldots, Z_n)$ and $\mathbb{E}|f^{(m)}(Z_1, \ldots, Z_n)|$ with $Z_i = X_i$ or Y_i, and for certain $\|\cdot\|$ and u_n

$$\mu_{\max}^{(m)} = \frac{\mathbb{E} \exp(\|X_i\|/u_n)|X_{i,j}|^{\otimes m}}{\mathbb{E} \exp(-\|X_i\|/u_n)}, \quad \nu_{\max}^{(m)} = \cdots$$

Lindeberg's approach

- Interpolation: $\boldsymbol{V}_i = (X_1, \ldots, X_i, Y_{i+1}, \ldots, Y_n)$
- Expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{i=1}^{n} \mathbb{E}\Big\{ f(\boldsymbol{V}_i) - f(\boldsymbol{V}_{i-1}) \Big\},$$

- Leave-one-out: $\boldsymbol{U}_i = (X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$
- Taylor expansion:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{1}{m!} \sum_{i=1}^{n} \Big\langle \mathbb{E}f^{(m)}(\boldsymbol{U}_i), \mathbb{E}X_i^{\otimes m} - \mathbb{E}Y^{\otimes m} \Big\rangle + \mathrm{Rem}$$

- Comparison theory:

$$\mathbb{E}f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{Y}) = \sum_{m=2}^{m^*-1} \frac{n}{m!} \Big\langle F^{(m)}, \mu^{(m)} - \nu^{(m)} \Big\rangle + \mathrm{Rem}$$

A comparison theorem for maxima of sums

- With $F_\beta(x) = \beta^{-1}\log\left(\sum_{j=1}^{p} e^{\beta x_j}\right)$ being a "smooth max function",

$$\|x\|_\infty \le F_\beta(x) \le \|x\|_\infty + \frac{\log p}{\beta}, \qquad \|F_\beta^{(m)}\|_1 \le C_m \beta^{m-1}$$

- For all smooth functions h and constants $b_n > 0$ and $\beta_n \ge b_n \log p$,

$$\begin{aligned} &\left|\mathbb{E}\, h\left(b_n F_{\beta_n}\left(\sum_{i=1}^{n} X_i/\sqrt{n}\right)\right) - \mathbb{E}\, h\left(b_n F_{\beta_n}\left(\sum_{i=1}^{n} Y_i/\sqrt{n}\right)\right)\right| \\ \lesssim\ &\sum_{m=2}^{m^*-1} \frac{b_n \beta_n^{m-1}}{n^{m/2-1}} \left\|\mu^{(m)} - \nu^{(m)}\right\|_\infty + \frac{b_n \beta_n^{m^*-1}}{n^{m^*/2-1}} \left\|\mu_{\max}^{(m^*)} + \nu_{\max}^{(m^*)}\right\|_\infty \end{aligned}$$

where $m^* \ge 2$ and

$$\mu_{\max}^{(m)} = \frac{\mathbb{E}\exp(\|X_i\|_\infty \beta_n/n^{1/2})|X_{i,j}|^{\otimes m}}{\mathbb{E}\exp(-\|X_i\|_\infty \beta_n/n^{1/2})}, \quad \nu_{\max}^{(m)} = \cdots$$

A comparison theorem for maxima of sums

- With $F_\beta(x) = \beta^{-1} \log\left(\sum_{j=1}^p e^{\beta x_j}\right)$ being a "smooth max function",

$$\|x\|_\infty \le F_\beta(x) \le \|x\|_\infty + \frac{\log p}{\beta}, \qquad \|F_\beta^{(m)}\|_1 \le C_m \beta^{m-1}$$

- For all smooth functions h and constants $b_n > 0$ and $\beta_n \ge b_n \log p$,

$$\left| \mathbb{E}\, h\left(b_n F_{\beta_n}\left(\sum_{i=1}^n X_i/\sqrt{n} \right)\right) - \mathbb{E}\, h\left(b_n F_{\beta_n}\left(\sum_{i=1}^n Y_i/\sqrt{n} \right)\right) \right|$$

$$\lesssim \sum_{m=2}^{m^*-1} \frac{b_n \beta_n^{m-1}}{n^{m/2-1}} \left\| \mu^{(m)} - \nu^{(m)} \right\|_\infty + \frac{b_n \beta_n^{m^*-1}}{n^{m^*/2-1}} \left\| \mu_{\max}^{(m^*)} + \nu_{\max}^{(m^*)} \right\|_\infty$$

where $m^* \ge 2$ and

$$\mu_{\max}^{(m)} = \frac{\mathbb{E} \exp(\|X_i\|_\infty \beta_n / n^{1/2}) |X_{i,j}|^{\otimes m}}{\mathbb{E} \exp(-\|X_i\|_\infty \beta_n / n^{1/2})}, \quad \nu_{\max}^{(m)} = \cdots$$

- What is the effect of the approximation by F_β on tail probability? $b_n = ?$

An anti-concentration theorem:

- Recall that

$$T_n = \max_{j \le p} \sum_{i=1}^{n} X_{i,j}/\sqrt{n}.$$

- Under certain moment and tail probability conditions,

$$\max_t \mathbb{P}\Big\{ t \le T_n \le t + \eta \Big\} \lesssim \eta\, \mathbb{E}\, T_n + (\mathbb{E}\, T_n)^4 (\log p)^3 / n,$$

An anti-concentration theorem:

- Recall that

$$T_n = \max_{j \le p} \sum_{i=1}^{n} X_{i,j} / \sqrt{n}.$$

- Under certain moment and tail probability conditions,

$$\max_t \mathbb{P}\Big\{ t \le T_n \le t + \eta \Big\} \lesssim \eta \, \mathbb{E} T_n + (\mathbb{E} T_n)^4 (\log p)^3 / n,$$

- A bound for the modulus of continuity of the distribution function of T_n

An anti-concentration theorem:

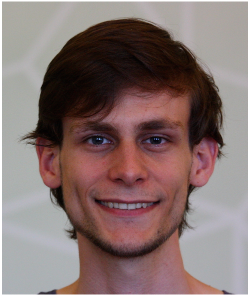
- Recall that

$$T_n = \max_{j \le p} \sum_{i=1}^{n} X_{i,j}/\sqrt{n}.$$

- Under certain moment and tail probability conditions,

$$\max_t \mathbb{P}\Big\{t \le T_n \le t+\eta\Big\} \lesssim \eta\,\mathbb{E}\,T_n + (\mathbb{E}\,T_n)^4(\log p)^3/n,$$

- A bound for the modulus of continuity of the distribution function of T_n
- Chernozhukov et al (13): anti-concentration for Gaussian $\boldsymbol{X}$



Ruben Dezeure

Peter Bühlmann

De-biasing regularized estimators (Dezeure-Bühlmann-Z, 16)

- Linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- De-biasing/LDPE (Z14): e.g. $\widehat{\boldsymbol{\beta}}^{(init)} = \widehat{\boldsymbol{\beta}}^{(lasso)}$:

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

De-biasing regularized estimators (Dezeure-Bühlmann-Z, 16)

- Linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- De-biasing/LDPE (Z14): e.g. $\widehat{\boldsymbol{\beta}}^{(init)} = \widehat{\boldsymbol{\beta}}^{(lasso)}$:

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

- Asymptotic theory:

$$\widehat{\beta}_j - \beta_j = (\boldsymbol{Z}_j^\top \boldsymbol{X}_j)^{-1} \left\{ \boldsymbol{Z}_j^\top \boldsymbol{\varepsilon} - \sum_{k \neq j} \boldsymbol{Z}_j^\top \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta})_k \right\} \approx N\left(0, \frac{\sigma^2}{\|\boldsymbol{Z}_j\|_2^2}\right)$$

  - $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}$, $\widehat{\boldsymbol{\varepsilon}}_{\rm cent} = (\widehat{\boldsymbol{\varepsilon}})_{\rm cent}$
  - $\widehat{\text{s.e.}}_j = (\boldsymbol{Z}_j^\top \boldsymbol{X}_j)^{-1} \|\boldsymbol{Z}_j\|_2 \|\widehat{\boldsymbol{\varepsilon}}_{\rm cent}\|_2 / \sqrt{n}$
  - $T_j = (\widehat{\beta}_j - \beta_j) / \widehat{\text{s.e.}}_j$
  - $\widehat{\text{s.e.}}_{j,\rm robust} = (\boldsymbol{Z}_j^\top \boldsymbol{X}_j)^{-1} \|(\boldsymbol{Z}_j \circ \widehat{\boldsymbol{\varepsilon}})_{\rm cent}\|_2$ for heteroscedastic $\boldsymbol{\varepsilon}$
  - $T_{j,\rm robust} = (\widehat{\beta}_j - \beta_j) / \widehat{\text{s.e.}}_{j,\rm robust}$

Bootstrap methods, a summary

- Residual bootstrap
  - ε^* iid from elements of $\widehat{\varepsilon}_{\text{cent}} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}^{(init)})_{\text{cent}}$
  - $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\beta}^{(init)} + \varepsilon^*$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{Z}_j)$
- Wild bootstrap
  - Draw iid W_i with $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = \mathbb{E}W_i^3 = 1$
  - $\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\beta}^{(init)} + \boldsymbol{W} \circ \widehat{\varepsilon}_{\text{cent}}$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}, \boldsymbol{Y}^*, \boldsymbol{Z}_j)$
- The xyz-paired bootstrap
  - $\widehat{\boldsymbol{X}} \perp \widehat{\varepsilon}_{\text{cent}}$, $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\widehat{\beta}^{(init)} + \widehat{\varepsilon}_{\text{cent}}$, $\widehat{\boldsymbol{Z}} \perp \widehat{\varepsilon}_{\text{cent}}$
  - $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{Z}^*)$: iid sample of rows of $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}})$
  - The plug-in estimates of T_j^* and $T_{j,\text{robust}}^*$ based on $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{Z}_j^*)$
- No re-computation of $\boldsymbol{Z}^*$ in bootstrap replications

Application of the new bootstrap theory to de-biased PLSE

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

- (A1) $\|\boldsymbol{X}\|_{\max} \le C$
- (A2): ε_i independent, $\mathbb{E}\,\varepsilon_i = 0$, $\mathbb{E}\,\varepsilon_i^2 = \sigma_i^2 \ge L$, $\mathbb{E}|\varepsilon_i|^{2+\delta} \le C$
- (A3): $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = o_P(1)/\sqrt{(\log p)\log(1+|G|)}$
- (A4) $\|\widehat{\boldsymbol{\beta}}^{*(init)} - \widehat{\boldsymbol{\beta}}^{(init)}\|_1 = o_{P^*}(1)/\sqrt{(\log p)\log(1+|G|)}$ in probability
- (A5): $\|\boldsymbol{Z}_G^\top \boldsymbol{X}_{-j}/n\|_{\max} \lesssim \sqrt{(\log p)/n}$, $\|\boldsymbol{Z}_j\|_2^2/n \ge L_z$, $\|\boldsymbol{Z}_j\|_{2+\delta}^{2+\delta} \ll \|\boldsymbol{Z}_j\|_2^{2+\delta}$
- (A6) $\|\boldsymbol{Z}_G\|_{\max} \le K$, $\delta = 2$, $\log(|G|) = o(n^{1/5})$

Application of the new bootstrap theory to de-biased PLSE

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

- (A1) $\|\boldsymbol{X}\|_{\max} \le C$
- (A2): ε_i independent, $\mathbb{E}\,\varepsilon_i = 0$, $\mathbb{E}\,\varepsilon_i^2 = \sigma_i^2 \ge L$, $\mathbb{E}|\varepsilon_i|^{2+\delta} \le C$
- (A3): $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = o_P(1)/\sqrt{(\log p)\log(1+|G|)}$
- (A4) $\|\widehat{\boldsymbol{\beta}}^{*(init)} - \widehat{\boldsymbol{\beta}}^{(init)}\|_1 = o_{P^*}(1)/\sqrt{(\log p)\log(1+|G|)}$ in probability
- (A5): $\|\boldsymbol{Z}_G^\top \boldsymbol{X}_{-j}/n\|_{\max} \lesssim \sqrt{(\log p)/n}$, $\|\boldsymbol{Z}_j\|_2^2/n \ge L_z$, $\|\boldsymbol{Z}_j\|_{2+\delta}^{2+\delta} \ll \|\boldsymbol{Z}_j\|_2^{2+\delta}$
- (A6) $\|\boldsymbol{Z}_G\|_{\max} \le K$, $\delta = 2$, $\log(|G|) = o(n^{1/5})$

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

- (A1) and (A2) imply $\|\boldsymbol{X}^T\boldsymbol{\varepsilon}/n\|_\infty = O_P(1)\sqrt{(\log p)/n}$
- (A3) and (A4) hold when $n \gg (s\log p)^2\log(1+|G|)$
- (A5) and (A6,1st) hold if $\boldsymbol{X}$ has iid rows with $\max_{j\in G}\|(\boldsymbol{\Sigma}^{-1})_{j,*}\|_1 = O(1)$

Consistency of the residual bootstrap

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

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

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

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

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

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

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

$$\mathrm{Cov}^*(\boldsymbol{Z}_j^T \boldsymbol{\varepsilon}^*, \boldsymbol{Z}_k^T \boldsymbol{\varepsilon}^*) \not\approx \mathrm{Cov}(\boldsymbol{Z}_j^T \boldsymbol{\varepsilon}, \boldsymbol{Z}_k^T \boldsymbol{\varepsilon})$$

Consistency of the wild bootstrap and xyz-paired bootstrap

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

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

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

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

$$\sup_{t} \left| \mathbb{P}^*\{\max_{j\in G} h(T^*_{j,\mathrm{robust}}) \le t\} - \mathbb{P}\{\max_{j\in G} h(T_{j,\mathrm{robust}}) \le t\} \right| = o_P(1)$$

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

Consistency of the wild bootstrap and xyz-paired bootstrap

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

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

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

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

$$\sup_{t} \left| \mathbb{P}^*\{\max_{j\in G} h(T^*_{j,\text{robust}}) \le t\} - \mathbb{P}\{\max_{j\in G} h(T_{j,\text{robust}}) \le t\} \right| = o_P(1)$$

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

Remark: The theorem is applicable in the heteroscedastic case

$$\mathrm{Cov}^*(\boldsymbol{Z}_j^T \varepsilon^*, \boldsymbol{Z}_k^T \varepsilon^*) \approx \mathrm{Cov}(\boldsymbol{Z}_j^T \varepsilon, \boldsymbol{Z}_k^T \varepsilon)$$

Some simulation results

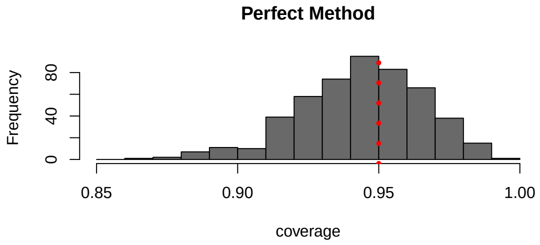


Figure: Histogram of the coverage probabilities of two sided 95% confidence intervals for 500 parameters. It illustrates how the results look like for a perfectly correct method for creating confidence intervals and one uses only 100 realizations to compute the probabilities.

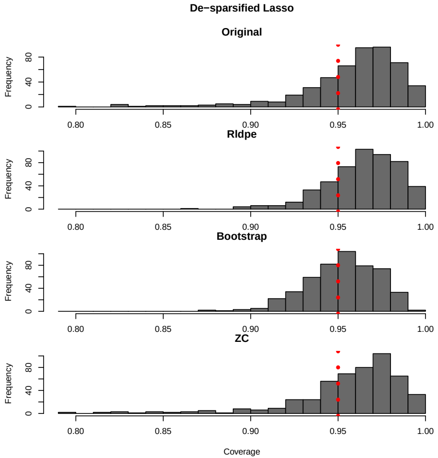


Figure: Histograms of the coverage probabilities of two-sided 95% confidence intervals for all 500 parameters in a linear model ($n = 100$, $p = 500$), computed from 100 independent replications. Perfect performance would look like Figure 1. The fixed design matrix is of Toeplitz type, the single coefficient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

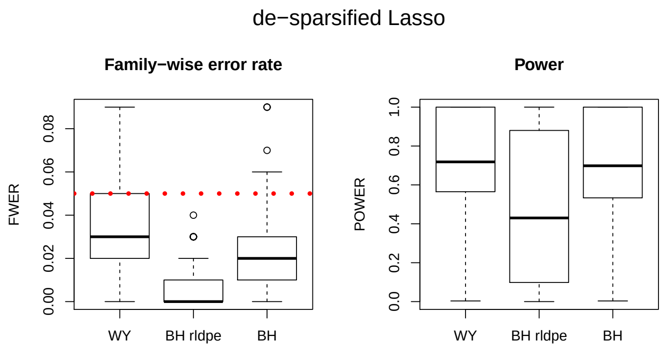


Figure: Boxplot of the familywise error rate and the power for multiple testing for the de-sparsified Lasso. The target is controlling the FWER at level 0.05, highlighted by a red-dotted horizontal line. Two different approaches for multiple testing correction are compared, Westfall-Young (WY) and Bonferroni-Holm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

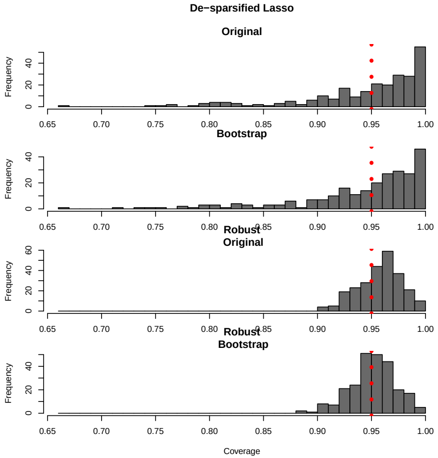


Figure: The same plot as Figure 2 but for **heteroscedastic non-Gaussian errors** and without signal. The robust standard error estimation clearly outperforms the non-robust version. There seems to be hardly any difference between the bootstrap and the original estimator after choosing the standard error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

Thanks!