Tight concentration inequality for sub-Weibull random variables with variance constraints

CFE-CMStatistics 2023

December 17th, 2023

Heejong Bong

Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Arun Kumar Kuchibhotla

Department of Statistics & Data Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

Concentration inequalities have been essential tools in high-dimensional statistical theory

For each session $l \in [m]$, we observe independent random vectors:

$$X_1^{(l)}, \dots, X_n^{(l)} \in \mathbb{R}^p.$$

Objective: shared covariance structure across m sessions.

error

CLT terms: $\sum_l \left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right] \right)$

high-dimensional CLT (Chernozhukov, Chetverikov & Kato, 2013)

non-CLT bias terms: $\sum_l \left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right] \right)^2$

concentration inequalities (Kuchibhotla & Chakrabortty, 2018)

Bong, H. (2022). Discovery of Functional Predictivity across Brain Regions from Local Field Potentials, PhD thesis, Carnegie Mellon University.

Concentration inequality is imperial in modern statistical theory

For each session $l \in [m]$, we observe independent random vectors:

$$X_1^{(l)}, \dots, X_n^{(l)} \in \mathbb{R}^p.$$

Objective: shared covariance structure across m sessions.

error

- **CLT terms:** $\sum_l \left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right]\right)$

  high-dimensional CLT (Chernozhukov, Chetverikov & Kato, 2013)

- **non-CLT bias terms:** $\sum_l \left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right]\right)^2$

  concentration inequalities (Kuchibhotla & Chakrabortty, 2018)

  ⇒ tight concentration inequalities for sub-Weibull random variables (BH. & Kuchibhotla, A., 2023)

Concentration inequalities are available for sub-Gaussian and sub-exponential random variables

Sub-Gaussian random variables:

$$\mathbb{P}[|\mathrm{X}| \geq \mathrm{t}] \leq 2\exp\left(-\frac{\mathrm{t}^2}{\nu^2}\right), \text{ for all } t > 0.$$

Sub-exponential random variables:

$$\mathbb{P}[|\mathrm{X}| \geq \mathrm{t}] \leq 2\exp\left(-\min\left\{\frac{t^2}{\nu^2}, \frac{t}{L\nu}\right\}\right), \text{ for all } t > 0.$$

However, $\left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right]\right)^2$ is not sub-Gaussian or sub-exponential.

Sub-Weibull random variables can model heavier tails

X is sub-Weibull(order α) with parameters ν and L if

$$\mathbb{P}[|\mathrm{X}| \geq \mathrm{t}] \leq 2\exp\left(-\min\left\{\frac{t^2}{\nu^2}, \frac{t^\alpha}{(L\nu)^\alpha}\right\}\right), \quad \text{for all } t > 0.$$

- $\alpha = 2 \;\Rightarrow\; X$ is sub-Gaussian
- $\alpha = 1 \;\Rightarrow\; X$ is sub-exponential
- $\alpha < 1 \;\Rightarrow\; X$ has heavier tails; e.g., $\left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2}\right]\right)^2$ is sub-Weibull(½).

Theorem

For independent sub-Weibull X_i with parameters $\nu_i = 1$ and $L_i > 0$,

$$\mathbb{P}[|\sum_i a_i X_i| \geq t] \leq 2\exp\left(-\frac{1}{C_\alpha}\min\left\{\frac{t^2}{\sum_i a_i^2 (1 \vee L_i)^2}, \frac{t^\alpha}{\max_i a_i^\alpha L_i^\alpha}\right\}\right), \quad t > 0.$$

Bong, H., & Kuchibhotla, A. K. (2023). Tight Concentration Inequality for sub-Weibull Random Variables with Generalized Bernstien Orlicz norm. *arXiv preprint arXiv:2302.03850*.

Theorem

For independent sub-Weibull X_i with parameters $\nu_i = 1$ and $L_i > 0$,

$$\mathbb{P}[|\sum_i a_i X_i| \geq t] \leq 2\exp\left(-\frac{1}{C_\alpha}\min\left\{\frac{t^2}{\sum_i a_i^2(1 \vee L_i)^2}, \frac{t^\alpha}{\max_i a_i^\alpha L_i^\alpha}\right\}\right),\quad t > 0.$$

vs. Bernstein's inequality

For independent sub-exponential X_i with parameters $\nu_i = 1$ and $L_i > 0$,

$$\mathbb{P}[|\sum_i a_i X_i| \geq t] \leq 2\exp\left(-min\left\{\frac{t^2}{\sum_i a_i^2(1 \vee L_i)^2}, \frac{t}{max_i\, a_i L_i}\right\}\right),\quad t > 0.$$

Bong, H., & Kuchibhotla, A. K. (2023). Tight Concentration Inequality for sub-Weibull Random Variables with Generalized Bernstien Orlicz norm. *arXiv preprint arXiv:2302.03850*.

Theorem

For independent sub-Weibull X_i with parameters $\nu_i = 1$ and $L_i > 0$,

$$\mathbb{P}[|\sum_i a_i X_i| \geq t] \leq 2\exp\left(-\frac{1}{C_\alpha}\min\left\{\frac{t^2}{\sum_i a_i^2 (1 \vee L_i)^2}, \frac{t^\alpha}{\max_i a_i^\alpha L_i^\alpha}\right\}\right), \quad t > 0.$$

Furthermore, the upperbound is tight:

$$\sup_{\|X_i\|_{\phi_{\alpha,L_i}}=1} \mathbb{P}[|\sum_i a_i X_i| \geq t] \geq \frac{1}{C_\alpha}\exp\left(-C_\alpha \min\left\{\frac{t^2}{\sum_i a_i^2 (1 \vee L_i)^2}, \frac{t^\alpha}{\max_i a_i^\alpha L_i^\alpha}\right\}\right), \quad t > 0,$$

where C_α is a constant depending on α.

Bong, H., & Kuchibhotla, A. K. (2023). Tight Concentration Inequality for sub-Weibull Random Variables with Generalized Bernstien Orlicz norm. *arXiv preprint arXiv:2302.03850*.

Applying the tight concentration inequality, we obtain a reduced sample complexity

$$\sum_l \left(\sum_i X_i^{(l)2} - \mathbb{E}\left[\sum_i X_i^{(l)2} \right] \right)^2$$

- sample complexity by the new theorem:

$$m + \log(qmn_0p) = o\left(\frac{\sqrt{mn_0p}}{d} \right)$$

- sample complexity of the previous result:

$$m + \log^2(qmn_0p) = o\left(\frac{\sqrt{mn_0p}}{d} \right)$$

Proof technique

- Tight moment bound using Latała's method:

$$\|\textstyle\sum_i a_i X_i\|_p \le C(\alpha) \max\left\{\sqrt{p \sum_i a_i^2 (1 \vee L_i)^2}\,, p^{\frac{1}{\alpha}} \max_i a_i^\alpha L_i^\alpha\right\}.$$

- Tight *Orlicz norm* bound following Kuchibhotla & Chakrabortty (2022):

$$\mathbb{E}\left[\exp\left(\min\left\{\left(\frac{\sum_i a_i X_i}{\sqrt{\sum_i a_i^2 (1 \vee L_i)^2}}\right)^2, \left(\frac{\sum_i a_i X_i}{\max_i a_i^\alpha L_i^\alpha}\right)^\alpha\right\}\right)\right] \le C(\alpha).$$

- Tight tail probability bound using Cramér-Chernoff technique and Paley-Zygmund inequality.

Bentkus' approach offers improved concentration inequality

Assumption:

$$\mathbb{E}[X_i] = 0,\ \mathrm{Var}[X_i] \leq A_i^2 \text{ and } \mathbb{P}[X_i > B] = 0.$$

Working inequality:

$$\mathbf{1}\{\nu \geq 0\} \leq \left(1 + \frac{\nu}{\alpha}\right)_+^{\alpha},$$

where $(x)_+ \coloneqq \max\{x, 0\}$.

Bentkus, V. (2004). On Hoeffding's inequalities. *Annals of probability*, *32*(2), 1650-1673.

Bentkus' approach offers improved concentration inequality

Application of Markov's inequality:

$$\mathbb{P}[\sum_i X_i \geq u] \leq \inf_{\lambda \geq 0} \mathbb{E}\left[1 + \frac{\lambda(\sum_i X_i - u)_+^\alpha}{\alpha}\right] = \inf_{x \leq u} \frac{\mathbb{E}\left[(\sum_i X_i - x)_+^\alpha\right]}{(u-x)_+^\alpha}.$$

Resulting tail probability inequality:

$$\mathbb{P}[\sum_i X_i \geq u] \leq \inf_{x \leq u} \sup_{X_i \sim \text{Assumption}} \frac{\mathbb{E}\left[(\sum_i X_i - x)_+^\alpha\right]}{(u-x)_+^\alpha} = \inf_{x \leq u} \frac{\mathbb{E}\left[(\sum_i G_i - x)_+^\alpha\right]}{(u-x)_+^\alpha},$$

where $G_i := \begin{cases} -\frac{A_i^2}{B} & \text{w.p. } \frac{B^2}{A_i^2 + B^2}, \\ B & \text{w.p. } \frac{A_i^2}{A_i^2 + B^2}. \end{cases}$

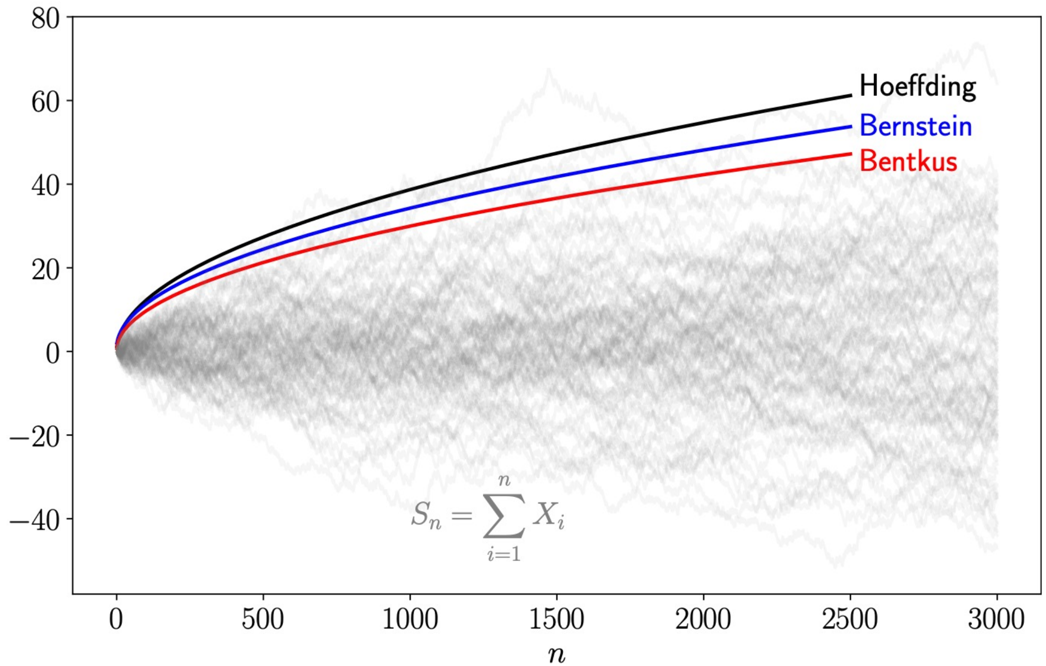
Bentkus, V. (2004). On Hoeffding's inequalities. *Annals of probability*, *32*(2), 1650-1673.

Bentkus' approach offers improved concentration inequality

Working inequality:

$$\mathbf{1}\{\nu \geq 0\} \leq \left(1 + \frac{\nu}{\alpha}\right)_+^{\alpha} \leq e^{\nu},$$

where $(x)_+ \coloneqq \max\{x, 0\}$.



Kuchibhotla, A. K., & Zheng, Q. (2020). Near-optimal confidence sequences for bounded random variables. *arXiv preprint arXiv:2006.05022*.

Bentkus' approach offers improved concentration inequality

$$G_i := \begin{cases} -\frac{A_i^2}{B} & \text{w.p. } \frac{B^2}{A_i^2 + B^2}, \\ B & \text{w.p. } \frac{A_i^2}{A_i^2 + B^2}. \end{cases}$$

Cramér-Chernoff technique:

$$1 \le \lim_{n\to\infty} \sup_{u\in\mathbb{R}} \frac{1}{\mathbb{P}\left[\sum_i G_i \ge u\right]} \inf_{\lambda\ge 0} \frac{\mathbb{E}[\exp(\lambda \sum_i G_i)]}{\exp(\lambda u)} = \infty$$

Bentkus' technique: for all $n \in \mathbb{N}$ and $u > 0$,

$$1 \le \frac{1}{\mathbb{P}\left[\sum_i G_i \ge u\right]} \inf_{x\le u} \frac{\mathbb{E}[(\sum_i G_i - x)_+^2]}{(u-x)_+^2} \le \frac{e^2}{2}.$$

⇒ sharp confidence sequence (Kuchibhotla & Zheng, 2021).

Kuchibhotla, A. K., & Zheng, Q. (2020). Near-optimal confidence sequences for bounded random variables. *arXiv preprint arXiv:2006.05022*.

We seek to refine concentration inequalities for unbounded X_i's using a similar approach

Assumption:

$$\mathbb{E}[X_i] = 0,\ \mathrm{Var}[X_i] \le A_i^2 \text{ and } \mathbb{P}[|X_i| \ge t] \le \min\left\{1, 2\exp\left(-\left(\frac{t}{K_i}\right)^\alpha\right)\right\} \text{ for } t > 0.$$

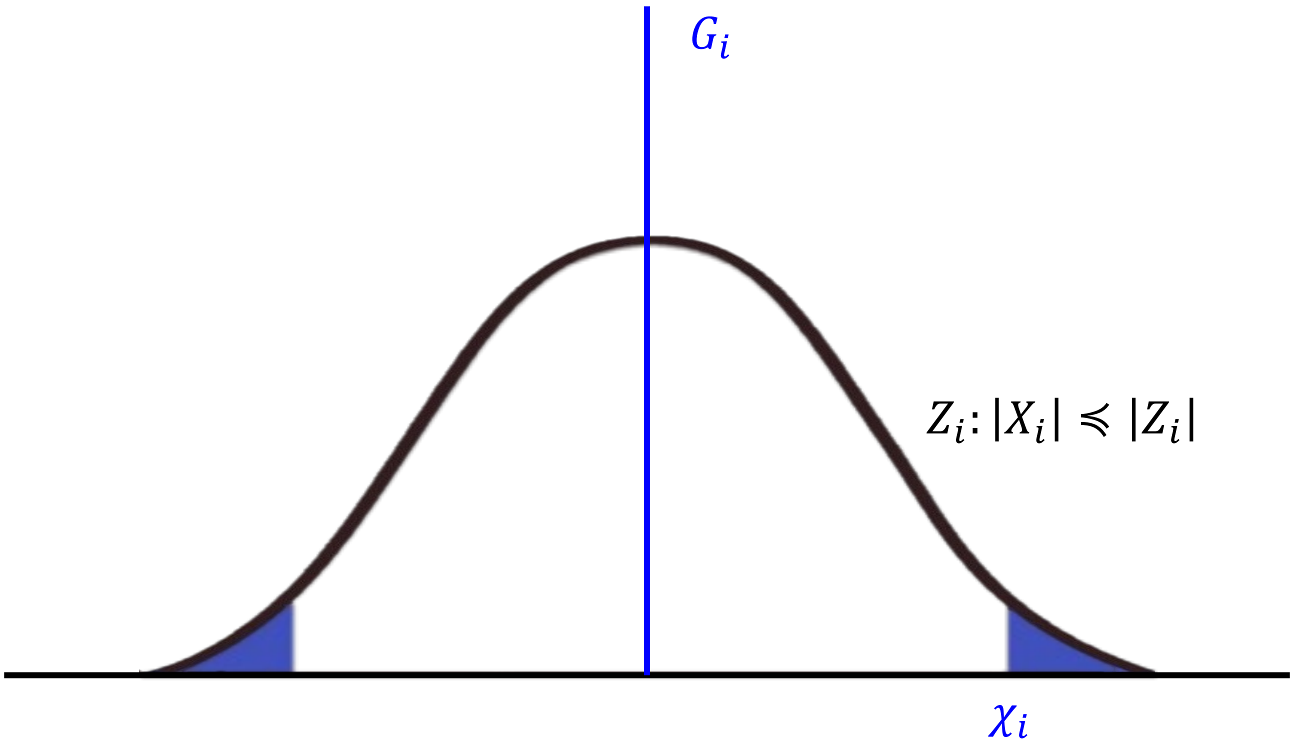
Theorem

For $\alpha \ge 3$,

$$\mathbb{P}\left[\sum_i X_i \ge u\right] \le \inf_{x \le u} \sup_{X_i \sim \mathrm{Assumption}} \frac{\mathbb{E}\left[\left(\sum_i X_i - x\right)_+^\alpha\right]}{(u-x)_+^\alpha} = \inf_{x \le u} \frac{\mathbb{E}\left[\left(\sum_i G_i - x\right)_+^\alpha\right]}{(u-x)_+^\alpha},$$

where $G_i \coloneqq Z_i 1\{|Z_i| \ge \chi_i\}$, $\chi_i = \inf\left\{t > 0: Var[Z_i 1\{|Z_i| \ge t\}] \le A_i^2\right\}$ and Z_i satisfies $\mathbb{P}[|Z_i| \ge t] = \min\left\{1, 2\exp\left(-\left(\frac{t}{K_i}\right)^\alpha\right)\right\}$.

Bong, H., & Kuchibhotla, A. K. (in progress). Tight concentration inequality for sub-Weibull random variables with variance constraints.



Bong, H., & Kuchibhotla, A. K. (in progress). Tight concentration inequality for sub-Weibull random variables with variance constraints.

Construction of empirical confidence interval

Suppose that we observe i.i.d sub-Weibull random variables

$$X_1, \dots, X_n \in \mathbb{R}$$

with known sub-Weibull parameters but unknown finite variance.

1. Obtain an upperbound of A_i using concentration inequalities for $\sum_i X_i^2$.

2. Obtain χ_i and G_i numerically.

3. Obtain confidence interval of $\sum_i X_i$ using

$$\mathbb{P}[\sum_i X_i \geq u] \leq \inf_{x \leq u} \frac{\mathbb{E}\left[\left(\sum_i G_i - x\right)_+^\alpha\right]}{(u-x)_+^\alpha}.$$

Future directions

- Theoretical and simulation study of the empirical confidence interval.
- Tightness of the resulting concentration inequality such as

$$1 \leq \frac{1}{\mathbb{P}\left[\sum_i G_i \geq u\right]} \inf_{x \leq u} \frac{\mathbb{E}[(\sum_i G_i - x)_+^2]}{(u - x)_+^2} \leq \frac{e^2}{2}.$$

- Comparison to the results of the Cramér-Chernoff technique.
- Construction of adaptive empirical confidence sequence.

References

1. Bong, H., & Kuchibhotla, A. K. (2023). Tight Concentration Inequality for sub-Weibull Random Variables with Generalized Bernstien Orlicz norm. *arXiv preprint arXiv:2302.03850*.

2. Bong, H., & Kuchibhotla, A. K. (in progress). Tight concentration inequality for sub-Weibull random variables with variance constraints.

3. Bong, H. (2022). Discovery of Functional Predictivity across Brain Regions from Local Field Potentials, PhD thesis, Carnegie Mellon University.

4. Kuchibhotla, A. K., & Chakrabortty, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, *11*(4), 1389-1456.

5. Bentkus, V. (2004). On Hoeffding's inequalities. *Annals of probability*, *32*(2), 1650-1673.

6. Kuchibhotla, A. K., & Zheng, Q. (2020). Near-optimal confidence sequences for bounded random variables. *arXiv preprint arXiv:2006.05022*.