Dual Induction CLT for High-dimensional m-dependent Data

Heejong Bong [1] Arun Kumar Kuchibhotla [2] Alessandro Rinaldo [3]

[1] Department of Statistics, University of Michigan
[2] Department of Statistics & Data Science, Carnegie Mellon University
[3] Department of Statistics & Data Sciences, The University of Texas at Austin

COLLEGE OF LITERATURE, SCIENCE, AND THE ARTS
UNIVERSITY OF MICHIGAN

Central Limit Theorem (CLT)

If $X_1, \dots, X_n$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}[X_i] = \sigma_i^2$, then for large enough n,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i,$$

where Y_i are independent $N(0, \sigma_i^2)$ random variables.

- CLT provides reference distributions for data summaries under minimal assumptions.
- Finite-sample error bounds for the Gaussian approximation is presented by Berry-Esseen bounds.

Berry-Esseen Bound

- The desired rate is $n^{-1/2}$ based on univariate cases:

$$\mu_{\mathscr{A}}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right) \equiv \sup_{A \in \mathscr{A}} \left| \mathbb{P}\left[\sum_{i=1}^n X_i \in A \right] - \mathbb{P}\left[\sum_{i=1}^n Y_i \in A \right] \right|$$
$$\le C \frac{\mathbf{1}}{\sqrt{\mathbf{n}}} \frac{\bar{\nu}_3}{\bar{\sigma}^3},$$

where $\mathscr{A} \equiv \{(-\infty, r] : r \in \mathbb{R}\}$, $\bar{\nu}_3 \equiv \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ X_i\ ^3]$ and $\bar{\sigma}^2 \equiv \frac{1}{n} \sum_{i=1}^n \sigma_i^2$.

- In high-dim. settings, [1]'s work on d-dim. hyper-rectangles ($\mathscr{R}_d$) drew huge attention for its surprisingly efficient polylogarithmic term on d despite suboptimal $n^{-1/8}$ rate:

$$\mu_{\mathscr{R}_d}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right) \le C \left(\frac{\log^7(\mathbf{dn})}{n} \right)^{1/8}.$$

- Recently, [2, 3] established the optimal $n^{-1/2}$ rate up to log terms, assuming non-degenerate covariance and finite third/fourth moments.
- Similar attempts under m-dependence (see Settings) have obtained $n^{-1/6}$ at best [4].
- ***Our contribution: the first $n^{-1/2}$ rate high-dimensional Berry-Esseen bound under m-dependence.***

Settings

- **m-dependence:** for a positive integer m,

$$X_i \perp\!\!\!\perp X_j \text{ for } i \text{ and } j \text{ such that } \ i - j\ \ge m.$$

- **nondegenerate covariance:** for some $\sigma_{\min}, \underline{\sigma} > 0$ and any interval $I \subseteq [n]$,

$$\min_{k \in [p]} \mathrm{Var}\left[\sum_{i \in I} Y_i(k) \right] \ge \sigma_{\min}^2 \cdot \ I \ \text{ and}$$
$$\lambda_{\min}\left(\mathrm{Var}\left[\sum_{i \in I} Y_i \ \ Y_j : j \in I^C \right] \right) \ge \underline{\sigma}^2 \cdot \max\{\ I\ - 2m, 0\}.$$

- **finite q-th moment:**

$$\bar{L}_q \equiv \frac{1}{n} \sum_{i=1}^n L_{q,i} \text{ and } \bar{\nu}_q = \frac{1}{n} \sum_{i=1}^n \nu_{q,i}.$$

where $L_{q,i} \equiv \max_{k=1,\dots,d} \mathbb{E}[\ X_i(k)\ ^q]$ and $\nu_{q,i} \equiv \mathbb{E}[\|X_i\|_\infty^q]$, $i = 1, \dots, n$.

Main Results

Under the aforementioned assumptions, if $q \ge 3$,

$$\mu_{\mathscr{R}_d}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right)$$
$$\le \frac{C \log(n/m)}{\sigma_{\min}} \sqrt{\frac{\log(dn/m)}{n}} \left[(m+1)^2 \frac{\bar{L}_3}{\underline{\sigma}^2} \log^2(d) + \left((m+1)^{q-1} \frac{\bar{\nu}_q}{\underline{\sigma}^2} \right)^{1/(q-2)} \log^{1 \vee 2/(q-2)}(d) \right]$$

for some universal constant $C > 0$. If $q \ge 4$,

$$\mu_{\mathscr{R}_d}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \right)$$
$$\le \frac{C \log(n/m)}{\sigma_{\min}} \sqrt{\frac{\log(dn/m)}{n}}$$
$$\times \left[(m+1)^2 \frac{\bar{L}_3}{\underline{\sigma}^2} \log^{3/2}(d) + (m+1)^{3/2} \frac{\bar{L}_4^{1/2}}{\underline{\sigma}} \log(d) + \left((m+1)^{q-1} \frac{\bar{\nu}_q}{\underline{\sigma}^2} \right)^{1/(q-2)} \log(d) \right]$$

for some universal constant $C > 0$.

- We used the relationship between the Kolmogorov-Smirnov distance and anti-concentration inequalities.
- Newly developed dual-induction argument established the $n^{-1/2}$ rate.

Comparisons to Existing Results

- **high-dim independent cases (i.e., $m = 0$)**

		our result	**[3]**
minimum eigenvalue		of every $\mathrm{Var}[X_i]$	of $\frac{1}{n} \sum_{i=1}^n \mathrm{Var}[X_i]$
moment term		$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|X_i\|_\infty^3]$	$\max_{i=1,\dots,n} \mathbb{E}[\|X_i\|_\infty^4]$
sample complexity		$\frac{1}{\sqrt{n}}$ up to logarithmic factors	
dimension complexity	bounded a.s.	$\log^4(d) = o(n)$	$\log^3(d) = o(n)$
	sub-Gaussian	$\log^4(d) = o(n)$	$\log^4(d) = o(n)$
	sub-exponential	$\log^5(d) = o(n)$	$\log^5(d) = o(n)$
	sub-Weibull(α)	$\log^{3+2/\alpha}(d) = o(n)$	$\log^{3+2/\alpha}(d) = o(n)$

- **1-dim m-dependent cases, vs. [5]:**
  - same optimal dependence on m

Discussions

- Applications to inference on high-dimensional time-series data
- Extension to weaker dependence structure
- Extension to graph dependence structure

References

Bong, H., Kuchibhotla, A. K., & Rinaldo, A. (2022). High-dimensional Berry-Esseen Bound for m-dependent Random Samples. *arXiv preprint arXiv:2212.05355.*

Bong, H., Kuchibhotla, A. K., & Rinaldo, A. (2023). Dual Induction CLT for High-dimensional m-dependent Data. *arXiv preprint arXiv:2306.14299.*

[1] Chernozhukov, V., Chetverikov, D., & Kato, K. (2013). Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-dimensional Random Vectors. *The Annals of Statistics, 41*(6), 2786-2819.

[2] Kuchibhotla, A. K., & Rinaldo, A. (2020). High-dimensional CLT for Sums of Non-degenerate Random Vectors: $n^{-1/2}$-rate. *arXiv preprint arXiv:2009.13673.*

[3] Chernozhukov, V., Chetverikov, D., & Koike, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *The Annals of Applied Probability, 33*(3), 2374-2425.

[4] Chang, J., Chen, X., & Wu, M. (2021). Central limit theorems for high dimensional dependent data. *arXiv preprint arXiv:2104.12929.*

[5] Shergin, V. V. (1980). On the convergence rate in the central limit theorem for m-dependent random variables. *Theory of Probability & Its Applications, 24*(4), 782-796.