High Dimensional Edgeworth Expansion With Applications to Bootstrap and Its Variants

Lihua Lei, Michael Jordan

Department of Statistics, UC Berkeley

Stanford-Berkeley Colloquium, 2016


STATISTICS FOR SPATIAL DATA
STATISTICS FOR SPATIAL DATA
Limit Theorems in Change-Point Analysis
Stochastic Geometry
DISCRETE DISTRIBUTIONS

Francis Ysidro Edgeworth
(1845 - 1926)

Peter Gavin Hall
(1951 - 2016)

Table of Contents

Table of Contents

Setup

- Given the data $\mathcal{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} F$;
- $X_i \in \mathbb{R}^p, \mathbb{E}X_i = 0, \mathbb{E}X_i X_i^T = V$;
- Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$;
- Goal: approximate the distribution of $\bar{X}$, i.e. approx.

$$P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right)$$

for $A \in \mathcal{C}$ where $\mathcal{C}$ denote the collection of all covex sets in $\mathbb{R}^p$.

CLT with Fixed Dimensions

Let Φ be the measure of $N(0, I_{p\times p})$. When the dimension p is fixed:

- Central Limit Theorem (CLT):

$$\sup_{A\in\mathcal{C}} \left| P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right) - \Phi(A)\right| = o(1);$$

CLT with Fixed Dimensions

Let Φ be the measure of $N(0, I_{p\times p})$. When the dimension p is fixed:

- Central Limit Theorem (CLT):

$$\sup_{A\in\mathcal{C}} \left|P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right) - \Phi(A)\right| = o(1);$$

- Berry-Esseen Bound (with third-order moments):

$$\sup_{A\in\mathcal{C}} \left|P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right) - \Phi(A)\right| = O\left(n^{-\frac{1}{2}}\right);$$

Edgeworth Expansion with Fixed Dimensions

- Edgeworth Expansion (with $(\nu+3)$-order moments):

$$\sup_{A\in\mathcal{C}}\left|P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X}\in A\right)-\Phi(A)-\sum_{j=1}^{\nu}n^{-\frac{j}{2}}P_j(A)\right|=O\left(n^{-\frac{\nu+1}{2}}\right);$$

  where $P_j(\cdot)$ are sign measures determined by the cumulants of F.

- When $\nu=1$ and $p=1$,

$$\sup_{A\in\mathcal{C}}\left|P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X}\in A\right)-\Phi(A)-n^{-\frac{1}{2}}P_1(A)\right|=O\left(n^{-1}\right)$$

  where $P_1(A)$ has a density

$$p_1(x)=\frac{1}{6}(x^3-x)\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

- Draw a bootstrap sample $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \hat{F}_n$ where $\hat{F}_n$ is the ecdf of $X_1, \ldots, X_n$;

Edgeworth Expansion and Bootstrap

- Draw a bootstrap sample $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \hat{F}_n$ where $\hat{F}_n$ is the ecdf of $X_1, \ldots, X_n$;
- Heuristically, a first-order edgeworth expansion implies

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n}(V^*)^{-\frac{1}{2}}(\bar{X}^* - \bar{X}) \in A \middle| \mathcal{X}\right) - \Phi(A) - n^{-\frac{1}{2}} P_1^*(A) \right| = O\left(n^{-1}\right);$$

where $V^* = \mathrm{Var}(X_1^*)$ and $P_1^*(\cdot)$ is determined by the cumulants of X_1^*.

Edgeworth Expansion and Bootstrap

- Draw a bootstrap sample $X_1^*, \ldots, X_n^* \stackrel{i.i.d.}{\sim} \hat{F}_n$ where $\hat{F}_n$ is the ecdf of $X_1, \ldots, X_n$;
- Heuristically, a first-order edgeworth expansion implies

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n}(V^*)^{-\frac{1}{2}}(\bar{X}^* - \bar{X}) \in A \middle| \mathcal{X}\right) - \Phi(A) - n^{-\frac{1}{2}} P_1^*(A) \right| = O\left(n^{-1}\right);$$

where $V^* = \mathrm{Var}(X_1^*)$ and $P_1^*(\cdot)$ is determined by the cumulants of X_1^*.

- Recall that

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n} V^{-\frac{1}{2}} \bar{X} \in A\right) - \Phi(A) - n^{-\frac{1}{2}} P_1(A) \right| = O(n^{-1});$$

- The cumulants of F and those of $\hat{F}_n$ are closed and thus

$$\sup_{A\in\mathcal{C}} |P_1(A) - P_1^*(A)| = O\left(n^{-\frac{1}{2}}\right).$$

- The cumulants of F and those of $\hat{F}_n$ are closed and thus

$$\sup_{A\in\mathcal{C}} |P_1(A) - P_1^*(A)| = O\left(n^{-\frac{1}{2}}\right).$$

- As a consequence,

$$\sup_{A\in\mathcal{C}} \left| P\left(\sqrt{n}(V^*)^{-\frac{1}{2}}(\bar{X}^* - \bar{X}) \in A \middle| \mathcal{X}\right) - P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right)\right| = O\left(n^{-1}\right).$$

- The cumulants of F and those of $\hat{F}_n$ are closed and thus

$$\sup_{A \in \mathcal{C}} |P_1(A) - P_1^*(A)| = O\left(n^{-\frac{1}{2}}\right).$$

- As a consequence,

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n}(V^*)^{-\frac{1}{2}}(\bar{X}^* - \bar{X}) \in A \middle| \mathcal{X}\right) - P\left(\sqrt{n}V^{-\frac{1}{2}}\bar{X} \in A\right) \right| = O\left(n^{-1}\right).$$

- This is called *Higher-Order Accuracy* (Hall, 1992).

Table of Contents

CLT in High Dimensions

- The CLT in high dimensions has been investigated since 60's, e.g. Sazonov, 1968; Portnoy, 1986; Gotze, 1991.

CLT in High Dimensions

- The CLT in high dimensions has been investigated since 60's, e.g. Sazonov, 1968; Portnoy, 1986; Gotze, 1991.
- Sharp result is obtained by Bentkus (2003),

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n} V^{-\frac{1}{2}} \bar{X} \in A\right) - \Phi(A) \right| = O\left(\frac{p^{\frac{7}{4}}}{\sqrt{n}}\right);$$

CLT in High Dimensions

- The CLT in high dimensions has been investigated since 60's, e.g. Sazonov, 1968; Portnoy, 1986; Gotze, 1991.
- Sharp result is obtained by Bentkus (2003),

$$\sup_{A \in \mathcal{C}} \left| P\left(\sqrt{n} V^{-\frac{1}{2}} \bar{X} \in A\right) - \Phi(A) \right| = O\left(\frac{p^{\frac{7}{4}}}{\sqrt{n}}\right);$$

- Fundamental limit: $p = o(n^{\frac{2}{7}})$ for CLT to hold;

- In contrast to CLT, very few works on edgeworth expansion in high dimensions; Some results on Banach space but focus on $\mathbb{E}f(\bar{X})$ for smooth f instead of the law of $\bar{X}$ (Gotze, 1981; Bentkus, 1984).

Edgeworth Expansion in High Dimensions

- In contrast to CLT, very few works on edgeworth expansion in high dimensions; Some results on Banach space but focus on $\mathbb{E}f(\bar{X})$ for smooth f instead of the law of $\bar{X}$ (Gotze, 1981; Bentkus, 1984).
- Using existing techniques (Bhattacharya & Rao, 1986):

$$p \preceq \mathrm{PolyLog}(n);$$

Edgeworth Expansion in High Dimensions

- In contrast to CLT, very few works on edgeworth expansion in high dimensions; Some results on Banach space but focus on $\mathbb{E}f(\bar{X})$ for smooth f instead of the law of $\bar{X}$ (Gotze, 1981; Bentkus, 1984).
- Using existing techniques (Bhattacharya & Rao, 1986):

$$p \preceq \mathrm{PolyLog}(n);$$

- Fundamental limit: $n^{-\frac{1}{2}}P_1(A)$ is of order $n^{-\frac{1}{2}}p^3$, without further constraints,

$$p \preceq n^{\frac{1}{6}}.$$

Edgeworth Expansion in High Dimensions

- In contrast to CLT, very few works on edgeworth expansion in high dimensions; Some results on Banach space but focus on $\mathbb{E}f(\bar{X})$ for smooth f instead of the law of $\bar{X}$ (Gotze, 1981; Bentkus, 1984).
- Using existing techniques (Bhattacharya & Rao, 1986):

$$p \preceq \mathrm{PolyLog}(n);$$

- Fundamental limit: $n^{-\frac{1}{2}}P_1(A)$ is of order $n^{-\frac{1}{2}}p^3$, without further constraints,

$$p \preceq n^{\frac{1}{6}}.$$

Question: How fast can the dimension grow with n?

Table of Contents

Theorem 1.

Let $X_1, \ldots, X_n$ be i.i.d. samples with zero mean and covariance matrix V. Assume that

1. $\lambda_{\max}(V) = O(1), \lambda_{\min}(V) = \Omega(1)$;
2. $p = O(n^{\gamma})$ *for some* $\gamma < \frac{1}{14}$;
3. $|X_{ij}| \le B = O(1)$;

Then for any positive integer S,

$$\sup_{A \in \mathcal{C}_n} \left| P(\sqrt{n} V^{-\frac{1}{2}} \bar{X} \in A) - \Phi(A) - \sum_{j=1}^{\nu} n^{-\frac{j}{2}} P_j(A) \right| = O\left(\left(\frac{p^9}{n} \right)^{\frac{\nu+1}{2}} \right).$$

Here $\mathcal{C}_n$ includes all convex sets plus all sets with form

$$\{F^{-1}(A) : A \text{ is convex}\}$$

for arbitrary smooth non-linear functions $F : \mathbb{R}^p \to \mathbb{R}$.

Here $\mathcal{C}_n$ includes all convex sets plus all sets with form

$$\{F^{-1}(A) : A \text{ is convex}\}$$

for arbitrary smooth non-linear functions $F : \mathbb{R}^p \to \mathbb{R}$.

As a corollary, for a smooth F,

$$P(F(\sqrt{n}\bar{X}) \in A) \approx \Phi_{0,V}(F^{-1}(A)) + \sum_{j=1}^{\nu} n^{-\frac{j}{2}} P_j(V^{\frac{1}{2}} F^{-1}(A)).$$

This gives an edgeworth expansion for smooth transform of mean.

Theorem 2.

Let $X_1, \ldots, X_n$ *be i.i.d. samples with zero mean and covariance matrix* V *and* $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \hat{F}_n$. *Assume that*

1. $\lambda_{\max}(V) = O(1), \lambda_{\min}(V) = \Omega(1)$;
2. $p = \Theta(n^\gamma)$ *for some* $0 < \gamma < \frac{1}{17}$;
3. $|X_{ij}| \le B = O(1)$.

Then with probability $1 - \exp\{-\Omega(n^\gamma)\}$,

$$\sup_{A \in \mathcal{C}_n} \left| P\left(V^{*-\frac{1}{2}}(\bar{X}^* - \bar{X}) \in A\right) - P\left(V^{-\frac{1}{2}}\bar{X} \in A\right) \right| \le \frac{Cp^9}{n}.$$

This is strictly better than the bound given by CLT since

$$\frac{p^9}{n} \ll \frac{p^{\frac{7}{4}}}{\sqrt{n}}, \quad \text{when } p = O(n^{\frac{1}{17}}).$$

Thanks!

References

Bentkus, V. (1984). Asymptotic expansions in the central limit theorem in hilbert space. *Lithuanian Mathematical Journal*, *24*(3), 210–225.

Bentkus, V. (2003). On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, *113*(2), 385–402.

Bhattacharya, R. N., & Rao, R. R. (1986). *Normal approximation and asymptotic expansions* (Vol. 64). SIAM.

Gotze, F. (1981). On edgeworth expansions in banach spaces. *The Annals of Probability*, 852–859.

Gotze, F. (1991). On the rate of convergence in the multivariate clt. *The Annals of Probability*, 724–739.

Hall, P. (1992). The bootstrap and edgeworth expansion.

Portnoy, S. (1986). On the central limit theorem in r p when p. *Probability theory and related fields*, *73*(4), 571–583.

Sazonov, V. (1968). On the multi-dimensional central limit theorem. *Sankhyā: The Indian Journal of Statistics, Series A*, 181–204.