5 Central limit theorem in high dimensions

This section is based on Chernozhukov et al. (2017).

5.1 Introduction

Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^p$, where $p \geq 3$ may be large or even much larger than n. Denote X_{ij} the jth coordinate of X_i, so that $X_i = (X_{i1}, \ldots, X_{ip})^\top$. We assume that each X_i is centred, namely $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}^2) < \infty$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Defined the normalised sum

$$S_n^X = (S_{n1}^X, \ldots, S_{np}^X)^\top = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i.$$

We consider Gaussian approximation to S_n^X. Let $Y_1, \ldots, Y_n$ be independent centred Gaussian random vectors in $\mathbb{R}^p$ such that each Y_i have the same covariance matrix as X_i, i.e. $Y_i \sim \mathcal{N}(0, \mathbb{E}[X_i X_i^\top])$. Define the normalised sum

$$S_n^Y = (S_{n1}^Y, \ldots, S_{np}^Y)^\top = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i.$$

We are interested in bounding the quantity

$$\rho_n(\mathcal{A}) = \sup_{A \in \mathcal{A}} |\mathbb{P}(S_n^X \in A) - \mathbb{P}(S_n^Y \in A)|,$$

where $\mathcal{A}$ is a class of Borel sets in $\mathbb{R}^p$.

We are interested in how fast $p = p(n) \to \infty$ is allowed to grow while guaranteeing $\rho(\mathcal{A}) \to 0$.

- When $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}(X_i X_i^\top) = I$,

$$\rho(\mathcal{A}) \leq C_p(\mathcal{A}) \frac{\mathbb{E}(\|X_1\|^3)}{\sqrt{n}},$$

  where $C_p(\mathcal{A})$ is a constant that depends only on p and $\mathcal{A}$.

  - When $\mathcal{A}$ is the class of all Euclidean balls in $\mathbb{R}^p$, $C_p(\mathcal{A})$ is bounded by a universal constant.
  - When $\mathcal{A}$ is the class of Borel measurable convex sets in $\mathbb{R}^p$, $C_p(\mathcal{A}) \leq 400p^{1/4}$. In this case, since $\mathbb{E}(\|X_1\|^3) \geq \{\mathbb{E}(\|X_1\|^2)\}^{3/2} = p^{3/2}$, once we require $\rho(\mathcal{A}) \to 0$, it is required that $p = o(n^{1/3})$.

- When $\mathcal{A}$ is the class of all Borel measurable convex sets, it was shown that $\rho(A) \geq c\mathbb{E}(\|X_1\|^3)/\sqrt{n}$, for some universal constant $c > 0$.

Let $\mathcal{A}$ be the class of all hyperrectangles in the sequel. This allows us to consider Kolmogorov–Smirnov type statistics.

5.2 Main results

Let $\mathcal{A}$ be the collection of all sets A of the form

$$A = \{w \in \mathbb{R}^p : a_j \le w_j \le b_j, \quad \forall j = 1, \dots p\},$$

for some $-\infty \le a_j \le b_j \le \infty$, $j = 1, \dots, p$. To describe the bound on $\rho(\mathcal{A})$, we need some additional notation. Define

$$L_n = \max_{j=1,\dots,p} \sum_{i=1}^{n} \mathbb{E}(|X_{ij}|^3)/n.$$

For $\phi \ge 1$, define

$$M_{n,X}(\phi) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\max_{j=1,\dots,p} |X_{ij}|^3 \mathbb{1}\left\{ \max_{j=1,\dots,p} |X_{ij}| > \sqrt{n}/(4\phi \log(p)) \right\}\right],$$

$$M_{n,Y}(\phi) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\max_{j=1,\dots,p} |Y_{ij}|^3 \mathbb{1}\left\{ \max_{j=1,\dots,p} |Y_{ij}| > \sqrt{n}/(4\phi \log(p)) \right\}\right]$$

and

$$M_n(\phi) = M_{n,X}(\phi) + M_{n,Y}(\phi).$$

Theorem 13. *Suppose that there exists some constant $b > 0$ such that $n^{-1} \sum_{i=1}^{n} \mathbb{E}(X_{ij}^2) \ge b$ for all $j = 1, \dots, p$. Then there exist constants $K_1, K_2 > 0$ depending only on b such that for every constant $\overline{L_n} \ge L_n$, we have*

$$\rho(\mathcal{A}) \le K_1 \left[\left(\frac{\overline{L_n}^2 \log^7(p)}{n} \right)^{1/6} + \frac{M_n(\phi)}{\overline{L_n}} \right], \tag{4}$$

where

$$\phi = K_2 \left(\frac{\overline{L_n}^2 \log^4(p)}{n} \right)^{-1/6}. \tag{5}$$

If $X_1, \dots, X_n$ are such that $\mathbb{E}(X_{ij}^2) = 1$ and for some $B_n \ge 1$, $|X_{ij}| \le B_n$ for all $i = 1, \dots, n$ and $j = 1, \dots, p$, then Theorem 13 shows that

$$\rho(\mathcal{A}) \le K \left\{ n^{-1} B_n^2 \log^7(pn) \right\}^{1/6}.$$

The bound (4) depends on $M_n(\phi)$ whose values are problem specific.

Proposition 14. *Suppose*

- $n^{-1} \sum_{i=1}^{n} \mathbb{E}(X_{ij}^2) \ge b$, *for all $j = 1, \dots, p$ and $b > 0$ some constant;*
- $n^{-1} \sum_{i=1}^{n} \mathbb{E}(|X_{ij}|^{2+k}) \le B_n^k$, *for all $j = 1, \dots, p$, $k = 1, 2$ and $B_n \ge 1$ a sequence of constants;*
- $\mathbb{E}\{\exp(|X_{ij}|/B_n)\} \le 2$, *for all $i = 1, \dots, n$ and $j = 1, \dots, p$.*

Then we have

$$\rho(\mathcal{A}) \lesssim \left(\frac{B_n^2 \log^7(pn)}{n} \right)^{1/6}.$$

Consider the multiplier bootstrap. Let $e_1, \dots, e_n$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables that are independent of $X_1^n = \{X_i\}_{i=1}^n$. Let

$$\bar{X} = \left(\frac{1}{n} \sum_{i=1}^n X_{i1}, \dots, \frac{1}{n} \sum_{i=1}^n X_{ip} \right)^\top$$

and consider the normalised sum

$$S_n^{eX} = \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i (X_i - \bar{X}).$$

We have that, under some mild conditions, for every constant $\bar{\Delta}_n > 0$, on the event $\Delta_{n,r} \leq \bar{\Delta}_n$,

$$\rho^{\mathrm{MB}}(\mathcal{A}) = \sup_{A \in \mathcal{A}} |\mathbb{P}(S_n^{eX} \in A | X_1^n) - \mathbb{P}(S_n^Y \in A)| \lesssim \bar{\Delta}_n^{1/3} \log^{2/3}(p),$$

where

$$\Delta_{n,r} = \max_{1 \leq j,k \leq p} |\widehat{\Sigma}_{jk} - \Sigma/jk|,$$

$$\widehat{\Sigma} = n^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top \quad \text{and} \quad \Sigma = n^{-1} \sum_{i=1}^n \mathbb{E}(X_i X_i^\top).$$

Consider the empirical bootstrap. Let $X_1^*, \dots, X_n^*$ be i.i.d. draws from the empirical distribution of $X_1, \dots, X_n$. Theorem 13 can also lead to an upper bound on

$$\sup_{A \in \mathcal{A}} |\mathbb{P}(S_n^{X^*} \in A | X_1^n) - \mathbb{P}(S_n^Y \in A)|,$$

where $S_n^{X^*} = n^{-1/2} \sum_{i=1}^n (X_i^* - \bar{X})$.

5.3 Proof of Theorem 13

Define

$$\varrho = \sup_{y \in \mathbb{R}^p,\, v \in [0,1]} |\mathbb{P}(\sqrt{v} S_n^X + \sqrt{1-v} S_n^Y \leq y) - \mathbb{P}(S_n^Y \leq y)|,$$

where $Y_1, \dots, Y_n$ are assumed to be independent of the random vectors $X_1, \dots, X_n$.

Lemma 15. *Suppose that there exists some constant $b > 0$ such that $n^{-1} \sum_{i=1}^n \mathbb{E}(X_{ij}^2) \geq b$ for all $j = 1, \dots, p$. Then ϱ satisfies the following inequality for all $\phi \geq 1$,*

$$\varrho \lesssim \frac{\phi^2 \log^2(p)}{n^{1/2}} \{ \phi L_n \varrho + L_n \log^{1/2}(p) + \phi M_n(\phi) \} + \frac{\log^{1/2}(p)}{\phi}$$

up to a constant K that depends only on b.

Define

$$\varrho' = \sup_{A \in \mathcal{A},\, v \in [0,1]} |\mathbb{P}(\sqrt{v} S_n^X + \sqrt{1-v} S_n^Y \in A) - \mathbb{P}(S_n^Y \in A)|.$$

An immediate corollary of Lemma 15 is as follows.

Corollary 16. *Suppose that there exists some constant $b > 0$ such that $n^{-1}\sum_{i=1}^{n}\mathbb{E}(X_{ij}^2) \geq b$ for all $j = 1, \dots, p$. Then ϱ' satisfies the following inequality for all $\phi \geq 1$,*

$$\varrho' \lesssim \frac{\phi^2 \log^2(p)}{n^{1/2}}\{\phi L_n \varrho' + L_n \log^{1/2}(p) + \phi M_n(\phi)\} + \frac{\log^{1/2}(p)}{\phi}$$

up to a constant K that depends only on b.

Proof of Corollary 16. Pick any hyperrectangle

$$A = \{w \in \mathbb{R}^p : \, w_j \in [a_j, b_j] \quad \forall j = 1, \dots, p\}.$$

For $i = 1, \dots, n$, consider the random vectors $\tilde{X}_i$ and $\tilde{Y}_i$ in $\mathbb{R}^{2p}$ defined by $\tilde{X}_{ij} = X_{ij}$ and $\tilde{Y}_{ij} = Y_{ij}$ for $j = 1, \dots, p$, and $\tilde{X}_{ij} = -X_{i,j-p}$ and $\tilde{Y}_{ij} = -Y_{i,j-p}$ for $j = p+1, \dots, 2p$. Then

$$\mathbb{P}(S_n^X \in A) = \mathbb{P}(S_n^{\tilde{X}} \leq y) \quad \text{and} \quad \mathbb{P}(S_n^Y \in A) = \mathbb{P}(S_n^{\tilde{Y}} \leq y),$$

where $y \in \mathbb{R}^{2p}$ is defined by $y_j = b_j$ for $j = 1, \dots, p$ and $y_j = -a_{j-p}$ for $j = p+1, \dots, 2p$. The result then follows from Lemma 15. □

Proof of Theorem 13. The proof relies on Lemma 15 and Corollary 16. Let K' denote a constant from the conclusion of Corollary 16. This constant depends only on b. Set $K_2 = 1/(K' \vee 1)$ in (5), so that

$$\phi = \frac{1}{K' \vee 1}\left(\frac{\overline{L_n}^2 \log^4(p)}{n}\right)^{-1/6}.$$

The result follows from Corollary 16. □

Proof of Lemma 15. We begin with preparing some notation. Let $W_1, \dots, W_n$ be a copy of $Y_1, \dots, Y_n$. Without loss of generality, we may assume that $X_1, \dots, X_n$, $Y_1, \dots, Y_n$ and $W_1, \dots, W_n$ are independent. Consider $S_n^W = n^{-1/2}\sum_{i=1}^{n} W_i$. Then S_n^Y and S_n^W are the same distribution. Then

$$\varrho = \sup_{y \in \mathbb{R}^p, \, v \in [0,1]} |\mathbb{P}(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y \leq y) - \mathbb{P}(S_n^W \leq y)|.$$

Pick any $y \in \mathbb{R}^p$ and $v \in [0, 1]$. Let $\beta = \phi \log(p)$ and define the function

$$F_\beta(w) = \beta^{-1} \log\left(\sum_{j=1}^{p} \exp\{\beta(w_j - y_j)\}\right), \quad w \in \mathbb{R}^p.$$

The function $F_\beta(w)$ has the following property

$$0 \leq F_\beta(w) - \max_{j=1,\dots,p}(w_j - y_j) \leq \beta^{-1}\log(p) = \phi^{-1}, \quad \forall w \in \mathbb{R}^p.$$

Pick a thrice continuously differentiable function $g_0 : \mathbb{R} \to [0, 1]$ whose derivatives up to the third order are all bounded such that $g_0(t) = 1$ for all $t \leq 0$ and $g_0(t) = 0$ for $t \geq 1$. Define $g(t) = g_0(\phi t)$, $t \in \mathbb{R}$, and

$$m(w) = g(F_\beta(w)), \quad w \in \mathbb{R}^p.$$

For brevity of notation, we will use indices to denote partial derivatives of m. For every $j,k,l = 1,\dots,p$, there exists a function $U_{jkl}(w)$ such that

$$|m_{jkl}(w)| \le U_{jkl}(w),$$
$$\sum_{j,k,l=1}^{p} U_{jkl}(w) \lesssim (\phi^3 + \phi\beta + \phi\beta^2) \lesssim \phi\beta^2,$$
$$U_{jkl}(w) \lesssim U_{jkl}(w+\tilde{w}) \lesssim U_{jkl}(w).$$

Define the functions

$$h(w,t) = \mathbb{1}\left\{-\phi^{-1} - t/\beta < \max_{j=1,\dots,p}(w_j - y_j) \le \phi^{-1} + t/\beta\right\}, \quad w \in \mathbb{R}^p,\ t > 0,$$

and

$$w(t) = \frac{1}{\sqrt{t} \wedge \sqrt{1-t}},\ t \in (0,1).$$

The proof consists of two steps. In the first step, we show that

$$|\mathbb{E}(\mathcal{I}_n)| \lesssim \frac{\phi^2 \log^2(p)}{n^{1/2}}(\phi L_n \varrho + L_n \log^{1/2}(p) + \phi M_n(\phi)),$$

where

$$\mathcal{I}_n = m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W).$$

In the second step, we combine this bound with Lemma 17 to complete the proof.

Step 1. Define the Slepian interpolant

$$Z(t) = \sum_{i=1}^{n} Z_i(t), \quad t \in [0,1],$$

where

$$Z_i(t) = \frac{1}{\sqrt{n}}\{\sqrt{t}(\sqrt{v}X_i + \sqrt{1-v}Y_i) + \sqrt{1-t}W_i\}.$$

Note that $Z(1) = \sqrt{v}S_n^X + \sqrt{1-v}S_n^Y$ and $Z(0) = S_n^W$, and so

$$\mathcal{I}_n = m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W) = \int_0^1 \frac{dm(Z(t))}{dt}dt.$$

Denote

$$Z^{(i)} = Z(t) - Z_i(t)$$

and

$$\dot{Z}_i(t) = \frac{1}{\sqrt{n}}\left\{\frac{1}{\sqrt{t}}(\sqrt{v}X_i + \sqrt{1-v}Y_i) - \frac{1}{\sqrt{1-t}}W_i\right\}.$$

For brevity of notation, write $Z = Z(t)$, $Z_i = Z_i(t)$, $Z^{(i)} = Z^{(i)}(t)$ and $\dot{Z}_i = \dot{Z}_i(t)$.

It follows from Taylor's expansion that

$$\mathbb{E}(\mathcal{I}_n) = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathbb{E}[m_j(Z)\dot{Z}_{ij}]dt = \frac{1}{2}(I + II + III),$$

where

$$I = \sum_{j=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathbb{E}[m_j(Z^{(i)})\dot{Z}_{ij}]dt,$$
$$II = \sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathbb{E}[m_{jk}(Z^{(i)})\dot{Z}_{ij}Z_{ik}]dt,$$
$$III = \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \int_0^1 (1-\tau)\mathbb{E}[m_{jkl}(Z^{(i)} + \tau Z_i)\dot{Z}_{ij}Z_{ik}Z_{il}]d\tau dt.$$

By the independence of $Z^{(i)}$ from $\dot{Z}_{ij}$ together with $\mathbb{E}[\dot{Z}_{ij}] = 0$, we have $I = 0$. By independence of $Z^{(i)}$ from $\dot{Z}_{ij}Z_{ik}$ together with

$$\mathbb{E}[\dot{Z}_{ij}Z_{ik}] = \frac{1}{n}\mathbb{E}[vX_{ij}X_{ik} + (1-v)Y_{ij}Y_{ik} - W_{ij}W_{ik}] = 0,$$

we have that $II = 0$. We skip the proof on III here.

Step 2. Let

$$V_n = \sqrt{v}S_n^X + \sqrt{1-v}S_n^Y.$$

Then we have

$$\begin{aligned}
&\mathbb{P}(V_n \leq y - \phi^{-1}) \leq \mathbb{P}(F_\beta(V_n) \leq 0) \leq \mathbb{E}[m(V_n)] \\
\leq&\mathbb{P}(F_\beta(S_n^W) \leq \phi^{-1}) + \mathbb{E}[m(V_n)] - \mathbb{E}[m(S_n^W)] \\
\leq&\mathbb{P}(S_n^W \leq y + \phi^{-1}) + |\mathbb{E}[\mathcal{I}_n]| \\
\leq&\mathbb{E}(S_n^W \leq y - \phi^{-1}) + C\phi^{-1}\log^{1/2}(p) + |\mathbb{E}(\mathcal{I}_n)|.
\end{aligned}$$

The other direction also holds and completes the proof. □

5.4 Auxiliary results

Lemma 17. *Let $Y = (Y_1, \dots, Y_p)^\top$ be a centred Gaussian random vector in $\mathbb{R}^p$ such that $\mathbb{E}(Y_j^2) \geq b$ for all $j = 1, \dots, p$ and some constant $b > 0$. Then for every $y \in \mathbb{R}^p$ and $a > 0$,*

$$\mathbb{P}(Y \leq y + a) - \mathbb{P}(Y \leq y) \leq Ca\sqrt{\log(p)},$$

where $C > 0$ is a constant depending only on b.

Lemma 18. *Let $\psi_i : \mathbb{R} \to [0, \infty)$, $i = 1, 2$ be non-decreasing functions, and let ξ_i, $i = 1, 2$ be independent real-valued random variables. Then*

$$\begin{aligned}
&\mathbb{E}[\psi_1(\xi_)]\mathbb{E}[\psi_2(\xi_1)] \leq \mathbb{E}[\psi_1(\xi_1)\psi_2(\xi_1)], \\
&\mathbb{E}[\psi_1(\xi_)]\mathbb{E}[\psi_2(\xi_2)] \leq \mathbb{E}[\psi_1(\xi_1)\psi_2(\xi_1)] + \mathbb{E}[\psi_1(\xi_2)\psi_2(\xi_2)], \\
&\mathbb{E}[\psi_1(\xi_1)\psi_2(\xi_2)] \leq \mathbb{E}[\psi_1(\xi_1)\psi_2(\xi_1)] + \mathbb{E}[\psi_1(\xi_2)\psi_2(\xi_2)].
\end{aligned}$$