SDS 387
Linear Models

Fall 2025

Lecture 11 - Thu, Oct 7, 2025

Instructor: Prof. Ale Rinaldo

- Last time CLT in $\mathbb{R}^d$:

Consider a triangular array of random vectors in $\mathbb{R}^d$

$\{ X_{n,j}, j=1,\dots,n \}_{n=1,2,\dots}$ s.t. $\mathbb{E}[X_{n,j}] = 0 \in \mathbb{R}^d$

and that $\underbrace{\text{Var}[X_{n,j}]}_{d \times d \text{ matrix}}$ exists and is invertible $\forall n$ and j

Let $$Y_{n,j} = \left(\sum_{i=1}^{n} \text{Var}[X_{n,i}] \right)^{-1/2} X_{n,j}$$

If

$$(*) \quad \lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}\left[\| Y_{n,j} \|^2 \mathbb{1}\{ \| Y_{n,j} \| > \varepsilon \} \right] = 0 \quad \forall \varepsilon > 0$$

Then $$\sum_{j=1}^{n} Y_{n,j} \xrightarrow{d} N_d(0, I_d)$$ as $n \to$

I_d ↳ identity matrix

Pf/ Use the Cramer Wold device. We need to show that $\forall t \in \mathbb{R}^d$

$$t^T \sum_{j=1}^{n} Y_{n,j} \xrightarrow{d} t^T Z$$

$\hookrightarrow Z \sim N_d(0, I_d)$

It is easy to see that

Exercise or HW $\leftarrow$

$$t^T \sum_{j=1}^{n} Y_{n,j} \sim \underbrace{0, \|t\|^2}_{\text{mean is 0 and variance is } \|t\|^2}$$

To show that $\dfrac{\left(t^T \sum_{j=1}^{n} Y_{n,j}\right)}{\|t\|} \xrightarrow{d} N(0,1)$ we will check that the LF condition holds. So $\forall \varepsilon > 0$

$$\frac{1}{\|t\|^2} \sum_{j=1}^{n} \mathbb{E}\left[\left(t^T Y_{n,j}\right)^2 \mathbb{1}\{|t^T Y_{n,j}| > \varepsilon \|t\|\}\right]$$

By Cauchy Schwartz $|t^T Y_{n,j}| \le \|t\| \, \|Y_{n,j}\|$

$$\le \cancel{\frac{1}{\|t\|^2}} \sum_{j=1}^{n} \mathbb{E}\left[\cancel{\|t\|^2} \|Y_{n,j}\|^2 \mathbb{1}\{\cancel{\|t\|} \|Y_{n,j}\| > \varepsilon \cancel{\|t\|}\}\right]$$

$$= \sum_{j=1}^{n} \mathbb{E}\left[\|Y_{n,j}\|^2 \mathbb{1}\{\|Y_{n,j}\| > \varepsilon\}\right] \to 0$$

as $n \to \infty$ by assumption. (*)

∎

∎ BERRY-ESSEEN BOUNDS $\rightarrow$ See Petrov Chapter 5

Let $X_1, X_2, \ldots, X_n$ be independent univariate r.v.'s s.t. $\mathbb{E}[X_n] = 0$, $\sigma_n^2 = \text{Var}[X_n]$ and $\mathbb{E}|X_n|^3 = \mu_{i,3} < \infty$

Then

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i}{B_n} \le x\right) - \Phi(x) \right| \le C \, \frac{\sum_{i=1}^{n} \mu_{i,3}}{B_n^3}$$

where $B_n^2 = \sum_{i=1}^{n} \sigma_i^2$ and Φ is the cdf of $N(0,1)$

C is an universal constant $< 1/2$

- Assume $\sigma_i^2 = \sigma^2$ and $E[|X_i|^3] = \mu_3$ all i. Then

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{\sqrt{n}\, \bar{X}_n}{\sigma} \le x \right) - \underbrace{\Phi(x)}_{P\left(\sqrt{n}\,\frac{\bar{Z}_n}{\sigma} \le x\right)} \right| \le c \frac{\cancel{n}\, \mu_3}{\sigma^3\, n^{3/\cancel{2}}} = C \frac{\mu_3}{\sigma^3} \frac{1}{\sqrt{n}}$$

$Z_1, \dots, Z_n \overset{iid}{\sim} N(0, \sigma^2)$

- This requires a 3^{rd} moment!

<u>Example</u> $X_1, X_2, \dots, X_n$ independent with $X_i \sim \text{Bernoulli}(p_i)$

$p_i \in (0,1)$ all i

Then the Berry – Esseen bound is as follows:

$$E\left[|X_i - p_i|^3\right] = p_i(1-p_i) \underbrace{\left[(1-p_i)^2 + p_i^2\right]}_{\le 1}$$

$$\le p_i(1-p_i)$$

So, assuming that $p_i \in [\varepsilon, 1-\varepsilon]$ $0 < \varepsilon < 1/2$ all i

The RHS of Berry – Esseen bound is

$$\le C \frac{\sum_{i=1}^{n} p_i(1-p_i)}{\left(\sum_{i=1}^{n} p_i(1-p_i)\right)^{3/2}} = C \frac{1}{\sqrt{\sum_{i=1}^{n} p_i(1-p_i)}}$$

Next, for all i, $\frac{1}{p_i(1-p_i)} \leq \frac{1}{\varepsilon(1-\varepsilon)}$. To see this, e.g. look at the graph of the function $x \in [\varepsilon, 1-\varepsilon] \longmapsto x(1-x)$

(formally, use concavity of the function)

Then,

$$\frac{1}{\sqrt{\sum_{i=1}^{n} p_i(1-p_i)}} \leq \frac{1}{\sqrt{n \min_i p_i(1-p_i)}} \leq \frac{1}{\sqrt{n\,\varepsilon(1-\varepsilon)}}$$

If we let $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$ the we have a CLT as long as $\frac{1}{\sqrt{n}} = o\left(\sqrt{\varepsilon_n(1-\varepsilon_n)}\right)$

Equivalently ε_n can go to zero but slower than $\frac{1}{n}$

For example, if $\varepsilon_n = n^{-\alpha}$ for $\alpha \in (0,1)$ the Berry - Esseen bound is of order $n^{(\alpha-1)/2}$.

HIGH-DIM BERRY ESSEEN BOUNDS

Let $X_1, \dots, X_n$ are independent centered r.v.'s in $\mathbb{R}^d$ s.t. $\mathrm{Cov}[X_i] = \Sigma_i$. Let $Z_1, \dots, Z_n$ be independent centered Gaussians s.t. $\mathrm{Var}[Z_i] = \Sigma_i$
$= \mathrm{var}[X_i]$

Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^d$. Examples:

- set of all convex set
- set of all balls or ellipsoids
- set of all hyper-rectangles

We want to establish the bound:

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}\left(\frac{\sum_i X_i}{\sqrt{n}} \in A \right) - \mathbb{P}\left(\frac{\sum_i Z_i}{\sqrt{n}} \in A \right) \right|$$

$$\leq C(d, \mathcal{A}) \frac{1}{\sqrt{n}} \text{ "Third moment term"}$$