Large Sample Theory PartI

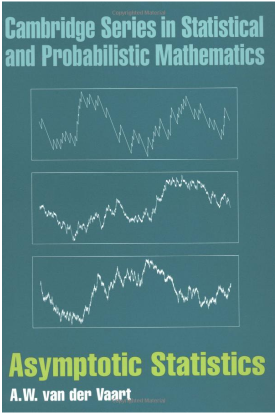
Song Xi Chen, Xiaojun Song

Department of Business Statistics and Econometrics
Center for Statistical Science
Peking University

July 13, 2021

Textbook

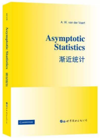
Van der Vaart, A. W. (1998). Asymptotic statistics. Cambridge university press.



新书推荐 |《渐近统计》(一部介绍渐近统计的权威教材)

世界图书出版公司北京公司 1月28日

《渐近统计》是一部介绍渐近统计的权威教材，内容实用而且数学理论论述严谨。



《渐近统计》

(新书即将上市)

作者：[荷兰] 范德瓦特（A. W. van der Vaart）
原名：Asymptotic Statistics
页数：462
定价：129.00元
装帧：平装
ISBN：9787519254025
出版社：世界图书出版公司

书中除了介绍渐近统计的核心理论——似然推断、M估计、渐近效率、U统计和秩过程等内容，书中还涉及该领域的最新研究论题，如半参数模型、自助法、经验过程等其他应用。本书各章有习题。

渐近统计 中文修订版

范德瓦特 著

世界图书出版公司北京公司版权所有

张慧铭译1-5,18-19,24-25章;王鹏飞译11-15章;王然译9,20,2
王宏浩译10,16-17章;刘云啸译23章;沈铂涵译6-7章;魏浩宇译25章;陈

February 16, 2020

摘 要

van der Vaart教授是荷兰皇家艺术与科学院院士，也是当今具有国
的统计学家。在经验过程和半参数统计方面，他做出过重要的基础性贡
著的《弱收敛与经验过程》(Weak Convergence and Empirical Processes
与Jon Wellner 合著) 和《渐近统计》(Asymptotic Statistics, 1998) 已成
领域的经典文献和必读书目。近年来他的兴趣主要集中在非参数贝叶斯方
维统计及统计遗传学、复杂网络等应用领域。由于必须以数据分析和应用
驱动，统计学理论往往呈现出碎片化的倾向。他的工作却能高屋建瓴，从
半参数统计的理论基础，到2000年后贝叶斯方法的频率学派解释，再到最
模型的研究，无不让人自然地领会到统计学理论的统一和谐之美。

Jordan教授访谈[1]：《渐近统计》是加州大学伯克利分校的教材，书中
寄托于经验过程想法的很多统计推断过程。比如M-估计，包含了最大似然
验风险最小化等等。这是一本发人深省的书。Jordan 教授期望学生们花一
至少去读过本书中的一些章节而且他认为这些将要进入机器学习领域的学
需要最终读完这些书籍，更需要将它们都读上至少三遍——**第一遍能够勉**
了，第二遍去尝试做相关的仿真实验和理论推导，第三遍看到之后会觉
都是显而易见的。

Main references

References

- **Serfling, R. (1980) Approximation Theorems in Mathematical Statistics. Wiley.** [Second Textbook]
- Ferguson, T. S. (1997). A course in large sample theory. CRC.
- Proschan, M. A., & Shaw, P. A. (2018). Essentials of probability theory for statisticians. CRC.
- White, H. (2000). Asymptotic Theory for Econometricians: Revised Edition. Emerald.
- Vaart, A. W., & Wellner, J. A. (1996). Weak convergence and empirical processes: with applications to statistics. Springer.
- Bosq, D. (2012). Nonparametric statistics for stochastic processes: estimation and prediction 2ed. Springer.
- Bijma, F., Jonker, M., & Van der Vaart, A. (2017). An introduction to mathematical statistics. Amsterdam University Press.

Let $Z_1, \cdots, Z_n$ be a set of data, and $T_n = T(Z_1, \cdots, Z_n)$ be a Statistic.

Task

- T_n can be an estimator to a parameter θ s.t. $\hat{\theta}_n = T_n$;
- or T_n can be a test statistic for a hypothesis: $H_0 : \theta \in \Omega_0$.
- A key task of Inference is to derive/find the distribution of $\hat{\theta}_n$, say $F_{\hat{\theta}_n}$,

$$F_{\hat{\theta}_n}(x) = P(\hat{\theta}_n \leq x) \quad \text{for } x \in R^d.$$

Why **Asymptotic Statistics**?

Dilemma and Benefits

- Exact fixed sample (non-asymptotical) analysis on statistics is HARD.
- But, letting $n \to \infty$ simplifies things and amazingly quality approximation to $F_{\hat{\theta}_n}(x)$ may be obtained.

The use of asymptotic approximation is two-fold.

Van der Vaart's book:

- It can be used for **asymptotical inference** (find approximate **confidence regions** and **testing**).
- Approximations can be used theoretically to study the quality (efficiency) of statistical inference procedures.

Stochastic Convergence

Let $\{X_n\}$ be a sequence of $\mathbb{R}^p$ random vectors and $d(x, y)$ be a Euclidean distance in $\mathbb{R}^p$, $\{X_n, X\}$ is defined on a common $(\Omega, \mathcal{A}, P)$.

- **Almost-Sure Convergence** ($X_n \stackrel{\text{a.s.}}{\to} X$): The sequence $\{X_n\}$ is said to *converge almost surely* to X

  $$\text{if } d(X_n, X) \to 0 \text{ with probability one}: \ \mathrm{P}(\lim_n d(X_n, X) = 0) = 1.$$

  $X_n \stackrel{\text{a.s.}}{\to} X$= "100% sure + 100% accurate."

- **Convergence in Probability** ($X_n \stackrel{\text{P}}{\to} X$): A sequence of random variables $\{X_n\}$ is said to *converge in probability* to X if for all $\varepsilon > 0$,

  $$\lim_n \mathrm{P}(d(X_n, X) < \varepsilon) = 1.$$

  $X_n \stackrel{\text{P}}{\to} X$= "100% sure+ **not 100%** accurate."

- **Convergence in $\mathbf{r_{th}}$ mean** ($X_n \stackrel{\mathsf{L^r}}{\to} X$): A sequence of random variables $\{X_n\}$ is said to *converge in r_{th} Mean* to X if

  $$\lim_n \mathbb{E}[d(X_n, X)]^r = 0.$$

Convergence in Distribution (Weak Convergence)

Convergence in distribution is a type of weak convergence for r.v.s, it is the most useful stochastic convergence in statistical inferences.

It does not require that the $\{X_n, X\}$ are in the same $(\Omega, \mathcal{A}, P)$.

Definition 2.1 (Convergence in Distribution, $X_n \xrightarrow{d} X$)

Let $\{F_n\}$ be a sequence of distribution functions for a sequence of r.v.s $\{X_n\}$. Then X_n is said to *converge in distribution* to a r.v. X (with distribution F) if

$$\lim_n F_n(x) = F(x), \forall x \in \mathcal{C}_F$$

where $\mathcal{C}_F := \{x | F(x) \text{ is continuous in } x\}$.

Remark 1

1. The discontinuous set $\mathcal{C}_F^c$ is the countable set.

2. The spaces of $\{X_n, X\}$ can different as $\xrightarrow{d}$ focus on the cdfs (free of $(\Omega, \mathcal{A})$).

A Lemma

Known in mathematical analysis: For any continuous function F, F is u.f. continuous on $[-M, M]$.

Lemma 2.2

If F is a continuous distribution function, then F is uniformly continuous in $\mathbb{R}$.

Remark 2

The lemma is valid for F not necessarily a cdf as long as F has limits at $\pm\infty$.

Polya's Theorem

Theorem 2.3

Suppose that $X_n \xrightarrow{d} X$ for a random variable X with F_n and F being the continuous distribution functions of X_n and X. Then $\sup_x |F_n(x) - F(x)| \to 0$ *as* $n \to \infty$.

- The proof uses the Covering Method: divide a diverging interval $[-M, M]$ by a partition $\{[x_{i-1}, x_i]\}_{i=0}^{K+1}$ of equal width δ (except the last one).
- Define $\Delta_n = \max_{i=0}^{K+1}\{|F_n(x_i) - F(x_i)|$. As K is finite for any given M, $\lim_{n\to\infty} \Delta_n = 0$.
- Use the previous lemma.

Asymptotic Normality

Definition 2.4

A sequence of random variables $\{X_n\}$ is said to be asymptotically normal (AN) with "mean" μ_n and "variance" σ_n^2 ($\sigma_n^2 > 0$ when n is sufficiently large) if $\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0, 1)$, denoted as $X_n \sim AN(\mu_n, \sigma_n^2)$.

- μ_n and σ_n^2 are not necessarily to be the mean and variance of X_n. In fact, the mean and variance of X_n may not exist.
- What $X_n \xrightarrow{d}$ is unknown !
- That $X_n \xrightarrow{d} N(\mu_n, \sigma_n^2)$ is obviously wrong !
- Nevertheless,

$$\sup_t \left| P(X_n \leq t) - P(N(\mu_n, \sigma_n^2) \leq t) \right| \to 0$$

as $n \to \infty$.

Multivariate AN

Some facts:

If X_n is AN(μ_n, σ_n^2), then

1. X_n is AN$(\bar{\mu}_n,\ \bar{\sigma}_n^2)$ if and only of (iff) $\frac{\bar{\sigma}_n}{\sigma_n} \to 1$ and $\frac{\mu_n - \bar{\mu}_n}{\sigma_n} \to 0$.
2. $a_n X_n + b_n$ is AN$(\bar{\mu}_n,\ \bar{\sigma}_n^2)$ iff $a_n \to 1$ and $\frac{\mu_n(a_n - 1) + b_n}{\sigma_n} \to 0$.

Definition 2.5 (Multivariate AN)

A seq of random vectors $\{X_n\}$ istb AN with "mean" μ_n and "variance" Σ_n (Σ_n is positive definite when n is sufficiently large) , if

$$a' X_n \text{ is } AN\,(a'\mu_n, a'\Sigma a),\ \forall a \in \mathbb{R}^P.$$

How to establish $X_n \xrightarrow{d} X$?

Lemma 2.6 (Portmanteau Lemma)

For any random vectors X_n and X, the following the following statements are equivalent.

1. $X_n \xrightarrow{d} X$;
2. $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for any* $f \in C_B$;
3. $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for any* $f \in C_{B,Lip}$;
4. $\liminf \mathrm{E} f(X_n) \ge \mathrm{E}f(X)$ *for all nonnegative, continuous* f;
5. $\liminf \mathrm{P}(X_n \in G) \ge \mathrm{P}(X \in G)$ *for any open set* G;
6. $\limsup \mathrm{P}(X_n \in F) \le \mathrm{P}(X \in F)$ *for any* closed set F;
7. $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ *for any Borel set* B *with* $\mathrm{P}(X \in \delta B) = 0$, *where* $\delta B = \overline{B} - \mathring{B}$ *is the boundary of* B.
8. *(Lévy's continuity theorem) Let* $\{X_n\}$ *and* X *be r.v.s in* $\mathbb{R}^d$. *Then*
$$X_n \xrightarrow{d} X \text{ iif } \phi_{X_n}(t) \to \phi(t) \quad \forall t \in \mathbb{R}^d.$$

(i) $\Rightarrow$ (ii):

W.O.L.G. Assume $\sup|f(x)| \leq 1$. First assume the df of X, F_X is continuous. As $X_n \xrightarrow{d.} X$, we have:

$$\lim_{n\to\infty} \mathrm{P}(X_n \in I) = \mathrm{P}(X \in I), \quad \text{for every rectangle} \quad I \subset \mathbb{R}^k$$

On the other hand. $\forall \epsilon > 0$, choose I large enough s.t. $\mathrm{P}(X \in I^C) < \epsilon$. Then we partition I into many small and non-overlapped rectangles s.t. $I = \cup_{j=1}^K I_j$. Choose a $x_j \in I_j$. Now let:

$$f_\epsilon(x) = \sum_{j=1}^K f(x_j)\mathbb{I}(x \in I_j)$$

Obviously, we can choose K large enough to have $|f(x) - f_\epsilon(x)| < \epsilon, \quad x \in I$. Thus,

The Proof of Portmanteau Lemma

$$
\begin{aligned}
|\mathrm{E}f(X_n) - \mathrm{E}f_\epsilon(X_n)| &\le \mathrm{E}\left[|f(X_n) - f_\epsilon(X_n)|\mathbb{I}(X_n \in I)\right] \\
&\quad + \mathrm{E}\left[|f(X_n) - f_\epsilon(X_n)|\mathbb{I}(X_n \in I^C)\right] \\
&\le \epsilon + 2\mathrm{P}(X_n \in I^C) \qquad \text{(Recall that } \sup|f(x)| \le 1\text{)}
\end{aligned} \tag{1}
$$

Similarly,

$$
|\mathrm{E}f(X) - \mathrm{E}f_\epsilon(X)| \le \epsilon + 2\mathrm{P}(X \in I^C) < 3\epsilon \tag{2}
$$

Considering,

$$
\begin{aligned}
|\mathrm{E}f_\epsilon(X_n) - \mathrm{E}f_\epsilon(X)| &\le \left|\sum_{j=1}^{K} f(x_j)\mathrm{E}\left[\mathbb{I}(X_n \in I_j) - \mathbb{I}(X \in I_j)\right]\right| \\
&\le \sum_{j=1}^{K} |f(x_j)||\mathrm{P}(X_n \in I_j) - \mathrm{P}(X \in I_j)| \rightarrow 0
\end{aligned} \tag{3}
$$

due to $X_n \xrightarrow{d.} X$ and K is finite. Then considering (1), (2), and (3), we have $\lim_{n\to\infty} \mathrm{E}f(X_n) = \mathrm{E}f(X)$.

The Proof of Portmanteau Lemma

If F_X is not continuous:

As F_X is right continuous and monotonous,

$$C_{F_x} := \{F_X \text{ is continuous at } x\}$$

is dense in $\mathbb{R}^k$. Then we could choose the vertices of rectangles I_j in C_{F_x}, then repeat the early proof.

The Proof of Portmanteau Lemma

(iii) $\Rightarrow$ (v)*:

For every open set G there exists a sequence of Lipschitz functions with $0 \leq f_m \uparrow 1_G$. For instance $f_m(x) = (md(x, G^c)) \wedge 1$. For every fixed m,

$$\liminf_{n\to\infty} \mathrm{P}(X_n \in G) \geq \liminf_{n\to\infty} \mathrm{E} f_m(X_n) = \mathrm{E} f_m(X)$$

As $m \to \infty$ the right side increases to $\mathrm{P}(X \in G)$ by the monotone convergence theorem.

(v) $\Leftrightarrow$ (vi)*:

Because a set is open if and only if its complement is closed, this follows by taking complements.

The Proof of Portmanteau Lemma

(v) + (vi) $\Rightarrow$ (vii)*:

Let $\mathring{B}$ and $\bar{B}$ denote the interior and the closure of a set, respectively. By (iv),

$$\mathrm{P}(X \in \mathring{B}) \leq \liminf \mathrm{P}\left(X_n \in \mathring{B}\right) \leq \limsup \mathrm{P}\left(X_n \in \bar{B}\right) \leq \mathrm{P}(X \in \bar{B},$$

by (v). If $\mathrm{P}(X \in \delta B) = 0$, then left and right side are equal, whence all inequalities are equities. The probability $\mathrm{P}(X \in B)$ and the limit $\lim \mathrm{P}(X_n \in B)$ are between the expressions on left and right and hence equal to the common value.

(vii) $\Rightarrow$ (i)*:

Every cell $(-\infty, x]$ such that x is a continuity point of $x \mapsto \mathrm{P}(X \leq x)$ is a continuity set.

(ii) $\Leftrightarrow$ (iv): Exercise.

Mapping theorem

Theorem 2.7 (Mapping)

Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set $\mathcal{C}_g$ such that $\mathrm{P}(X \in \mathcal{C}_g) = 1$, then

1. *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$*
2. *If $X_n \xrightarrow{\mathrm{P}} X$, then $g(X_n) \xrightarrow{\mathrm{P}} g(X)$*
3. *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$*

The Proof of Mapping Theorem

(i). Consider using (vi) in Portmanteau Lemma: for a closed set F, define $g^{-1}(F) = \{x_n \in g^{-1}(F)\} = \{g(X_n) \in F\}$. We have:

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C_g^C \tag{4}$$

The last $\subset$ is because of $\forall x \in \overline{g^{-1}(F)}$, here exist a sequence $\{x_m\}_{m\geq 1} \subset g^{-1}(F)$ s.t. $x_m \to x$.

- If $x \in C_g$, then $g(x_m) \to g(x) \in F$ due to $g(x_m) \in F$ and F is a close set. Thus, $x \in g^{-1}(F)$.
- If $x \notin C_g$, (4) is evident.

Then,

$$\begin{aligned}
\limsup \mathrm{P}\left(g\left(X_n\right) \in F\right) &\leq \limsup \mathrm{P}\left(X_n \in \overline{g^{-1}(F)}\right) && \text{(by (4) left)} \\
&\leq \mathrm{P}(X \in \overline{g^{-1}(F)}) && \text{(by Portmanteau Lemma (iv))} \\
&\leq \mathrm{P}(X \in g^{-1}(F)) + \mathrm{P}(X \notin C_g) && \text{(by (4) again)} \\
&= \mathrm{P}(g(X) \in F)
\end{aligned}$$

Apply Portmanteau Lemma (iv) $\Rightarrow$ (i), we have $g_n(X_n) \xrightarrow{d.} g(X)$.

The Proof of Mapping Theorem

(ii)*. Fix arbitrary $\varepsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $|x - y| < \delta$, but $|g(x) - g(y)| > \varepsilon$. If $X \notin B_\delta$ and $|g(X_n) - g(X)| > \varepsilon$, then $|X_n - X| \geq \delta$. Consequently,

$$\mathrm{P}\left(|g(X_n) - g(X)| > \varepsilon\right) \leq \mathrm{P}\left(X \in B_\delta\right) + \mathrm{P}\left(|X_n - X| \geq \delta\right)$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Because $B_\delta \cap C \downarrow \emptyset$ by continuity of g, the first term converges to zero as $\delta \downarrow 0$.

Example of Mapping Theorem

- If $X_n \xrightarrow{d.} X \sim N(0,1)$, then $X_n \xrightarrow{d.} \chi_1^2$.

- If

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d.} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2\right)$$

then $X_n/Y_n \xrightarrow{d.}$ Cauchy, whose distribution has p.d.f.

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad f_{\mu,\sigma}(x) = \frac{1}{\pi\sigma}\frac{1}{1+((x-\mu)/\sigma)^2}$$

Example of Mapping Theorem

- $S_n^2 = n^{-1} \sum X_i^2 - \overline{X}^2$. Let $Y_i = (X_i, X_i^2)^\top$. By LLN,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{pmatrix} \to \begin{pmatrix} \mu \\ \mu_2 \end{pmatrix} \quad \text{w.p.1}$$

Let $g(x, y) = y - x^2$. Then by Mapping Theorem:

$$S_n^2 = g\left(\overline{X}, n^{-1}\sum_{i=1}^{n} X_i^2\right) \xrightarrow{p.} = \mu_2 - \mu^2 = \sigma^2$$

Apply Mapping Theorem again: $S_n \xrightarrow{p.} \sigma$.

Example of Mapping Theorem

- If $X_n \xrightarrow{d.} N_p(\mu, \Sigma)$, then for any constant matrix $C \in \mathbb{R}^{m \times p}$,

$$CX_n \xrightarrow{d.} N_m\left(C\mu, C\Sigma C^\top\right)$$

- If X_n is $AN(\mu, b_n^2\Sigma)$, then:

$$\frac{\|X_n - \mu\|}{b_n} \xrightarrow{d.} \text{ a limit r.v.}$$

In fact, since $(X_n - \mu)/b_n \xrightarrow{d.} N_p(0, \Sigma)$, by Mapping Theorem,

$$\frac{(X_n - \mu)^\top (X_n - \mu)}{b_n^2} \xrightarrow{d.} N_p^\top(0, \Sigma) N_p(0, \Sigma)$$

Therefore,

$$\frac{\|X_n - \mu\|}{b_n} \xrightarrow{d.} \sqrt{N_p^\top(0, \Sigma) N_p(0, \Sigma)}$$

Theorem 2.8

Let X_n, X and Y_n be random vectors. Then

1. $X_n \stackrel{a.s.}{\to} X$ *implies* $X_n \stackrel{\mathrm{P}}{\to} X$;
2. $X_n \stackrel{\mathrm{P}}{\to} X$ *implies* $X_n \stackrel{d}{\to} X$
3. $X_n \stackrel{\mathrm{P}}{\to} c$ *(c is a constant) if and only if* $X_n \stackrel{d}{\to} c$;
4. *if* $X_n \stackrel{d}{\to} X$ *and* $d(X_n, Y_n) \stackrel{\mathrm{P}}{\to} 0$, *then* $Y_n \stackrel{d}{\to} X$;
5. *if* $X_n \stackrel{d}{\to} X$ *and* $Y_n \stackrel{\mathrm{P}}{\to} c$ *for a constant* c, *then* $(X_n, Y_n) \stackrel{d}{\to} (X, c)$
6. *if* $X_n \stackrel{\mathrm{P}}{\to} X$ *and* $Y_n \stackrel{\mathrm{P}}{\to} Y$, *then* $(X_n, Y_n) \stackrel{\mathrm{P}}{\to} (X, Y)$

Proof

(i)*. The sequence of sets $A_n = \cup_{m \geq n} \{\|X_m - X\| > \varepsilon\}$ is decreasing for every $\varepsilon > 0$ and decreases to the empty set if $X_n(\omega) \to X(\omega)$ for every ω. If $X_n \xrightarrow{as} X$, then $\mathrm{P}(\|(X_n - X\| > \varepsilon) \leq \mathrm{P}(A_n) \to 0$.

(iv) We have:

$$A_n := |\mathrm{E}f(X_n) - \mathrm{E}f(Y_n)| \leq \mathrm{E}\{|f(X_n) - f(Y_n)|\mathbb{I}(\|X_n - Y_n\| \leq \epsilon)\} + \mathrm{E}\{|f(X_n) - f(Y_n)|\mathbb{I}(\|X_n - Y_n\| > \epsilon)\}$$

Only for bounded Lipschitz function $f \in C_{B,Lip}$ and $\epsilon > 0$,

$$A_n \leq L\epsilon \mathrm{P}(\|X_n - Y_n\| \leq \epsilon) + 2\sup\|f(x)\|\mathrm{P}(\|X_n - Y_n\| > \epsilon)$$

Thus, $\mathrm{E}f(X_n) - \mathrm{E}f(Y_n) \to 0$, and

$$\mathrm{E}f(Y_n) = \mathrm{E}f(X_n) + \mathrm{E}f(Y_n) - \mathrm{E}f(X_n) \to \mathrm{E}f(X)$$

which implies $Y_n \xrightarrow{d.} X$ due to Portmanteau Lemma (iii).

Proof

(ii) $X_n = X + (X_n - X)$. Since $X \xrightarrow{d.} X$ and $X_n - X \xrightarrow{p.} 0$, we have $X_n \xrightarrow{d.} X$ by using (iv).

(iii) "$\Rightarrow$" is from (ii). For "$\Leftarrow$", note that:

$$\{\|X_n - c\| \geq \epsilon\} = \{X_n \in \text{ball}(c, \epsilon)^C\}$$

where $\text{ball}(c, \epsilon) = \{x : \|x - c\| < \epsilon\}$ is open, so $\text{ball}(c, \epsilon)^C$ is a closed set. From Portmanteau Lemma,

$$\begin{aligned} \limsup \mathrm{P}(\|X_n - c\| \geq \epsilon) &= \limsup \mathrm{P}(X_n \in \text{ball}(c, \epsilon)^C) \\ &\leq \mathrm{P}\left(c \in \text{ball}(c, \epsilon)^C\right) = 0 \end{aligned}$$

Hence, $\mathrm{P}(\|X_n - c\| \geq \epsilon) \to 0$, and thus $X_n \xrightarrow{p.} c$.

Proof

(v) $(X_n, Y_n) = (X_n, c) + (X_n, Y_n) - (X_n, c)$. Since

$$(X_n, Y_n) - (X_n, c) = (0, Y_n - c) \xrightarrow{p.} (0, 0)$$

From (iv), we only need to show $(X_n, c) \xrightarrow{d.} (X, c)$.

For any bounded continuous function $f : (x, y) \mapsto f(x, y)$, the marginal function $f_m : x \mapsto f(x, c)$ is also bounded continuous. As $X_n \xrightarrow{d.} X$,

$$\mathrm{E}f(X_n, c) = \mathrm{E}f_m(X_n) \to \mathrm{E}f_m(X) = \mathrm{E}f(X, c)$$

Hence $(X_n, c) \xrightarrow{d.} (X, c)$.

(vi) As $\|(X_1, Y_1) - (X_2, Y_2)\| \leq \|X_1 - X_2\| + \|Y_1 - Y_2\|$,

$$\mathrm{P}\left(\|(X_n, Y_n) - (X, Y)\| > \epsilon\right) \leq \mathrm{P}(\|X_n - X\| > \epsilon/2) + \mathrm{P}(\|Y_n - Y\| > \epsilon/2) \longrightarrow 0$$

Remark

From this theorem, we know that:

$$\text{Marginal Convergence in Prob} \Rightarrow \text{Joint Convergence in Prob.}$$

The converse is also true due to Mapping Theorem. So,

$$\text{Marginal Convergence in Prob} \Leftrightarrow \text{Joint Convergence in Prob.}$$

Nonetheless, Marginal Convergence in Dist $\not\Rightarrow$ Joint Convergence in Dist, although the converse is true via Mapping Theorem.

Copula.

- If X_n is $AN(\mu, b_n^2\Sigma)$ with $b_n \to 0$, then $X_n \xrightarrow{d} \mu$ and $X_n \xrightarrow{p} \mu$.

Slutsky's theorem: algebraic operations for con. in dist.

Application of Theorem 2.8 and Continuous mapping theorem.

It is named after a Russian mathematical statistician/economist: Slutsky.

Lemma 2.9 (Slutsky)

Let X_n, X and Y_n be random vectors or variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ (or $Y_n \xrightarrow{P} c$) for a constant c, then

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $Y_n X_n \xrightarrow{d} cX$
3. $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$ *provided* $c \neq 0$

The (3) is valid for matrices Y_n and c and vectors X_n provided $c \neq 0$ is understood as c being invertible, because taking an inverse is also continuous.

Slusky's Theorem

From (v) of Theorem 2.8, if $X_n \xrightarrow{d.} X$ and $Y_n \xrightarrow{p.} c$, then

$$(X_n, Y_n) \xrightarrow{d.} (X, c)$$

Apply Mapping Theorem,

$$g(X_n, Y_n) \xrightarrow{d.} g(X, c)$$

for almost surely continuous g.

Choose $g(x, y) = x + y$, $x \times y$, x/y then we obtain Slusky's Theorem.

Example

T-statistic:

Considering $Y_1, \ldots, Y_n$ i.i.d. F with $\mathrm{E}Y_1 = 0$ and $\mathrm{E}Y_1^2 = 0$. Define T-statistic: $t_n := \sqrt{n}\overline{Y}/S_n$.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2\right)$$
$$\xrightarrow{p.} 1 \cdot \left(\mathrm{E}Y_i^2 - (\mathrm{E}Y_i)^2\right) = \mathrm{E}Y_i^2$$

Hence $S_n \xrightarrow{p,} \sqrt{\mathrm{E}Y_i^2}$.

From the CLT for IID data.

$$\sqrt{n}\overline{Y} \xrightarrow{d.} N(0, \mathrm{E}Y_i^2)$$

Hence,

$$\frac{\sqrt{n}\overline{Y}}{S_n} \xrightarrow[Slusky]{d.} \frac{N(0, \mathrm{E}Y_i^2)}{\sqrt{\mathrm{E}Y_i^2}} \overset{d.}{=\!=} N(0,1)$$

Tightness or Stochastic Boundedness

Definition 2.10

A sequence of random vectors $\{X_n\}$ is said to be stochastically bounded or tight if $\forall \varepsilon > 0$, $\exists M_\varepsilon > 0$, s.t. $\sup_n P(||X_n|| > M_\varepsilon) < \varepsilon$. Denote $X_n = O_p(1)$.

Theorem 2.11 (Prohorov's Theorem)

1. *If $X_n \xrightarrow{d} X$, then X_n is tight.*
2. *If X_n is tight, then $\exists$ a subsequence $\{X_{n_i}\}$ ($n_i > 1$), s.t. $X_{n_i} \xrightarrow{d} X$ as $n_i \to \infty$.*

Remark 3

A single random vector is tight.

Finite random vectors are tight.

Since the distribution function of $\|X\|$, $F(x)$, satisfies:

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1$$

So we can choose M_ϵ s.t. $F(M_\epsilon) - F(-M_\epsilon)$ is as close to 1 as possible, then,

$$\mathrm{P}(\|X\| > M_\epsilon) = 1 - F(M_\epsilon) + F(-M_\epsilon)$$

can as small as possible. So we could choose M_ϵ large enough s.t.

$$\mathrm{P}(\|X\| > M_\epsilon) < \epsilon$$

Similarly, we have:

Remark 4

Any finite collection of r.v. $\{X_i\}_{i=1}^K$ is tight.

The proof of Prohorov's Theorem

(i) only. As X is tight, one can choose M_ϵ properly s.t.

$$\mathrm{P}(\|X\| \geq M_\epsilon) < \epsilon, \quad \forall \epsilon > 0$$

By Portmanteau Lemma,

$$\limsup \mathrm{P}(\|X_n\| \geq M_\epsilon) \leq \mathrm{P}(\|X\| \geq M_\epsilon) < \epsilon$$

Hence $\exists\, N$, s.t. $\forall n \geq N$, $\mathrm{P}(\|X_n\| \geq M_\epsilon) \leq 2\epsilon$.

Note the $\{X_i\}_{i=1}^{N-1}$ is tight. By modifying M_ϵ, we can obtain

$$\mathrm{P}(\|X_n\| \geq M_\epsilon) < \epsilon, \quad \forall n \in \mathbb{N}_+$$

So $\{X_n\}_{n\geq 1}$ is tight.

Remark 5

(ii) is an extended vision of "Bounded sequences must have a convergent subsequence" in Mathematical Analysis.

Stochastic o and O

Definition 2.12

A sequence of random vectors $\{X_n\}$ is called

1. **stochastically of order $\{a_n\}$,** with $\{a_n\}$ being a sequence of constants,
   if $\frac{X_n}{a_n} = O_p(1)$.
2. **stochastically small,** if $X_n \xrightarrow{P} 0$ and denoted as $X_n = o_p(1)$.
3. **stochastically of smaller order $\{a_n\}$,** if $\frac{X_n}{a_n} = o_p(1)$, write as $X_n = o_p(a_n)$.

Stochastic o and O

In a previous example, $Y_n - \mu = \overline{X} - \mu = O_p(n^{-1/2})$, and also $Y_n - \mu = o_p(1)$. However, $Y_n - \mu = O_p(n^{-1/2})$ is more accurate.

In general, considering a set of stochastic quantity $\{T_n\}_{n\geq 1}$, n is typically the sample size. Let

$$\mu_n = \mathrm{E}\, T_n, \quad \sigma_n^2 = \mathrm{var}(T_n)$$

if they exist. From the Chebysev inequality,

$$\mathrm{P}\left(\sigma_n^{-1}|T_n - \mu_n| > M\right) \leq \frac{\mathrm{var}(T_n)}{(M\sigma_n)^2} = \frac{1}{M^2}$$

Then $\forall\, \epsilon > 0$, we can choose M_ϵ s.t. $M_\epsilon^{-2} < \epsilon$, and

$$\mathrm{P}\left(\sigma_n^{-1}|T_n - \mu_n| > M_\epsilon\right) < \epsilon$$

which implies $\sigma_n^{-1}(T_n - \mu_n) = O_p(1)$, and $T_n - \mu_n = O_p(\sigma_n)$. This is a typical way to find the stochastic order of a quantity if we can labour out σ_n^2.

Basic Rules of Stochastic o and O

There are rules of calculus on o and O symbols, which we apply without comment. For instance,

Some facts:

1. $o_P(1) + o_P(1) = o_P(1)$
2. $o_P(1) + O_P(1) = O_P(1)$
3. $O_P(1)o_P(1) = o_P(1)$
4. $(1 + o_P(1))^{-1} = O_P(1)$
5. $O_P(1) + O_P(1) = O_P(1)$
6. $o_P(O_P(1)) = O_P(o_P(1)) = o_P(1)$

Remark 6

The rules should be read from left to right.

Lemma of Stochastic Plug-in

Lemma 2.13

Let $R : \mathbb{R}^k \mapsto \mathbb{R}$ *be a real function with* $R(0) = 0$. *Let* $\{X_n\}$ *be a sequence of r.v.s with values in* $\mathrm{dom(R)}$ *s.t.* $X_n \xrightarrow{p} 0$. *Then,* $\forall p > 0$ *,*

1. *if* $R(h) = o\left(\|h\|^p\right)$ *as* $h \to 0$, *then* $R\left(X_n\right) = o_P\left(\|X_n\|^p\right)$*;*
2. *if* $R(h) = O\left(\|h\|^p\right)$ *as* $h \to 0$, *then* $R\left(X_n\right) = O_P\left(\|X_n\|^p\right)$*.*

Remark 7

The function $R(\cdot)$ *may not be continuous other than* $\{0\}$ *in* $\mathrm{dom(R)}$.

Proof

(i). Let:

$$g(h) = \begin{cases} R(h)/\|h\|^p & , \text{for } h \neq 0 \\ 0 & , \text{for } h = 0 \end{cases}$$

Then $R(X_n) = \|X_n\|^p g(X_n)$. We will show that $g(X_n) \xrightarrow{p.} 0$.

Note that g(h) is continuous at 0, and $\mathrm{P}\left(\lim_{n\to\infty} X_n = 0\right) = 1$, by Mapping Theorem:

$$g(X_n) \xrightarrow{p.} g(0) = 0$$

So $R(X_n) = o_p(\|X_n\|^p)$.

Proof

(ii). As $g(h) = R(h)/\|h\|^p$ is bounded near $x = 0$ (because $R(h) = O(\|h\|^p)$ and $g(0) = 0$),

$$\exists M, \delta > 0, \quad |g(h)| \leq M, \quad \forall |h| < \delta$$

so,

$$\{\omega : |g(X_n(\omega))| > M\} \subset \{\omega : \|X_n(\omega)\| > \delta\}$$

Thus,

$$\mathrm{P}(|g(X_n)| > M) \leq \mathrm{P}(\|X_n\| > \delta) \to 0$$

In fact, $\forall \epsilon > 0$, $\exists N$, s.t. $\forall n \geq N$, $\mathrm{P}(\|X_n\| > \delta) < \epsilon$. Thus

$$\mathrm{P}(|g(X_n)| > M) < \epsilon, \quad \forall n \geq N$$

This implies $g(X_n) = O_p(1)$.

An Applied Example

Considering $X_1, \ldots, X_n$ i.i.d. $F(\mu, \sigma^2)$ with $\mathrm{E}X^4 < \infty$. Let,

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu) - (\bar{X} - \mu)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2$$

Since $\mathrm{E}X^4 < \infty$, by a standard CLT,

$$\frac{n^{-1}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2}{\sqrt{\mathrm{var}((X_i - \mu)^2)/n}} \xrightarrow{d.} N(0, 1)$$

which implies,

$$\sqrt{n}\left(\tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right) \xrightarrow{d.} N(0, v^2)$$

where $v^2 = \mathrm{var}((X_i - \mu)^2)$. On the other hand, $\bar{X} - \mu = O_p(n^{-1/2})$, $\sqrt{n}(\bar{X} - \mu)^2 = O_p(n^{-1/2}) = o_p(1)$. By Slusky Theorem,

$$\sqrt{n}(S_n - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\bar{X} - \mu)^2 \xrightarrow{d.} N(0, v^2)$$

We need tools for derivating weak convergence.

Definition 3.1

For any random vector X with distribution function F, its cf is

$$\begin{aligned} \phi_X(t) &= \mathrm{E}[e^{itX}] = \int e^{itx} dF(x) \\ &= \mathrm{E}\left[\cos tX\right] + i\mathrm{E}\left[\sin tX\right] \quad \textit{for any } t \in \mathbb{R}. \end{aligned}$$

The moment generating function (MGF) is

$$M_X(t) = \mathrm{E}[e^{tX}].$$

The cf is a frequency domain view of a distribution F, it fully characterizes F.

Properties of Characteristic Functions

1. $|\phi_X(t)| \leq \phi_X(0) = 1$
2. $\overline{\phi_X(t)} = \phi_X(-t)$
3. $\phi_X(t)$ is uniformly continuous on $\mathbb{R}$
4. $\overline{\phi_X}$, $|\phi_X|^2$ and $Re(\phi_X)$ are c.f.s of $-X$, $X - Y$ for X, Y i.i.d.F, and $(F_X + F_{-X})/2$ respectively.
5. If $\exists t_0 \neq 0$ s.t. $|\phi_X(t_0)| = 1$, then $\exists a \in \mathbb{R}$ and $a \neq 0$ s.t. $P(X \in \{a + jh : j \in \mathbb{Z}\}) = 1$, so X is a lattice random vector.

Properties of Characteristic Functions (Cont.)

vi) If F is absolutely continuous, $\lim_{|t|\to\infty}|\phi_X(t)| = 0$

vii) Two random vectors X and Y in $\mathbb{R}^d$ are equal in distribution, denoted as $X \stackrel{d}{=} Y$, iff $\phi_X(t) = \phi_Y(t) \quad \forall t \in \mathbb{R}^d$.

viii) (Fourier Inversion) If ϕ_X is integrable, i.e. $\phi_X \in \mathcal{L}^1(\mathbb{R})$; then F is continuous with density:

$$f(x) \;=\; \frac{1}{2\pi}\int e^{-itx}\phi_X(t)dt.$$

Multivariate CF

The characteristic function (cf) of X a p-dimensional random vector is defined by

$$\phi_{\mathrm{X}}(\mathrm{t}^T) = \mathrm{E}e^{i\mathrm{t}^T\mathrm{X}} = \int_{\mathbb{R}^p} e^{it\mathrm{x}} dF_{\mathrm{X}}(\mathrm{x}) \quad \text{for any } t \in \mathbb{R}^d \tag{5}$$

where F_{X} is the cumulative distribution function.

Remark 8

The multivariate CFs inherit the properties of univariate CFs.

Properties of Multivariate Characteristic Functions

1. $\varphi_{\text{x}}(\boldsymbol{t})$ exists for all $\boldsymbol{t} \in \mathbb{R}^d$ and is continuous.
2. $\varphi_{\boldsymbol{X}}(0) = 1$ and $|\varphi_{\boldsymbol{X}}(\boldsymbol{t})| \leq 1$ for all $\boldsymbol{t} \in \mathbb{R}^d$.
3. For a scalar $b \neq 0, \varphi_{\boldsymbol{X}/b}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(\boldsymbol{t}/b)$;
   For a vector $\boldsymbol{c}, \varphi_{\boldsymbol{X}+\boldsymbol{c}}(\boldsymbol{t}) = \exp\{i\boldsymbol{t}^T\boldsymbol{c}\}\varphi_{\boldsymbol{X}}(\boldsymbol{t})$.
4. For $\boldsymbol{X}$ and $\boldsymbol{Y}$ independent, $\varphi_{\boldsymbol{X}+\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(\boldsymbol{t})\varphi_{\boldsymbol{Y}}(\boldsymbol{t})$.
5. If $E\|\boldsymbol{X}\| < \infty, \dot{\varphi}_{\boldsymbol{X}}(\boldsymbol{t})$ exists and is continuous and $\dot{\varphi}_{\boldsymbol{X}}(0) = i\boldsymbol{\mu}^T$, where $\boldsymbol{\mu} = E\boldsymbol{X}$.
6. If $E\|\boldsymbol{X}\|^2 < \infty, \ddot{\varphi}_{\boldsymbol{X}}(\boldsymbol{t})$ exists and is continuous and $\ddot{\varphi}_{\boldsymbol{X}}(0) = -E\boldsymbol{X}\boldsymbol{X}^T$.
7. If $\boldsymbol{X}$ is $N_d(\boldsymbol{\mu}, \Sigma), \varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\}$.

Lévy-Cramér's Theorem

Lévy's continuity theorem

Theorem 3.2

Let $\{X_n\}$ and X be random vectors in $\mathbb{R}^d$. Then $X_n \xrightarrow{d.} X$ iif $\phi_{X_n}(t) \to \phi(t) \quad \forall t \in \mathbb{R}^d$.

Proof: "$\Rightarrow$" is by Portmanteau Lemma (ii): $\mathrm{E}f(X_n) \to \mathrm{E}f(x)$ for $\forall f \in C_B$, "$\Leftarrow$" can be seen in P14 of vdv.

***Remark** 9*

It provides another way for establishing weak convergence!

Example

Suppose $X_1, \ldots, X_n$ i.i.d. Poisson(λ) for fixed $\lambda > 0$. Then the characteristic function of X_i is:

$$\phi_X(t) = \exp\left\{\lambda(e^{it} - 1)\right\}$$

Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$, check the c.f.

$$\begin{aligned}
\phi_{\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}}}(\lambda) &= \exp\{-it\sqrt{n\lambda}\}\phi_{\bar{X}}(t/\sqrt{\lambda/n}) = \exp\{-i\sqrt{n\lambda}\}\phi_X^n(t/\sqrt{n\lambda}) \\
&= \exp\{-it\sqrt{n\lambda}\}\exp\left\{n\lambda\left(e^{\frac{it}{\sqrt{n\lambda}}} - 1\right)\right\} \\
&= \exp\left\{-it\sqrt{n\lambda} + n\lambda\left(\frac{it}{\sqrt{n\lambda}} + \frac{i^2t^2}{2n\lambda} + o\left(\frac{1}{n\lambda}\right)\right)\right\} \\
&= \exp\{-t^2/2 + o(1)\} \to \exp\{-t^2/2\}
\end{aligned}$$

Therefore,

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{d.} N(0,1)$$

Example

Weak Law of Large Numbers (WLLN):

Let $Y_1, \ldots, Y_n$ be IID r.v. with $\phi_Y(t)$ being differential at $t = 0$ and $i\mu = \phi'(0)$, then:

$$\bar{Y} \xrightarrow{p.} \mu$$

Proof: As $\phi(0) = 1$ and $\phi'(0)$ exists at 0,

$$\phi_Y(t) = 1 + t\phi'(0) + o(t), \qquad \text{as } t \to 0$$

$$\begin{aligned} \phi_{\bar{Y}}(t) = \phi_Y^n(t/n) &= \left(1 + \frac{t}{n}\phi'(0) + o\left(\frac{t}{n}\right)\right)^n \\ &= \left(1 + \frac{it\mu}{n} + o\left(\frac{1}{n}\right)\right)^n \to e^{it\mu} = \phi_\mu(t) \end{aligned}$$

Hence, $\bar{Y} \xrightarrow{d.} \mu$, and $\bar{Y} \xrightarrow{p.} \mu$.

Moments and Expansion of CFs

If r.v. X's r-th moment exits, then $\phi_X(t)$ is r-th order differentiable and,

$$\phi_X^{(r)}(t) = \int (ix)^r e^{itx} dF(x) = \mathrm{E}\{(iX)^r e^{itX}\}$$

which implies $\phi_X^{(r)}(0) = i^r \mathrm{E} X^r$.

Conversely, if $\phi_X^{(r)}(0)$ exists for an even r, then X has finite r-th absolutely moment.

Theorem 3.3

If $\mathrm{E|X|^r} < \infty$, *then*

$$\phi_X(t) = \sum_{j=0}^{r} \frac{(it)^j}{j!} \mathrm{EX^j} + \mathrm{o(|t|^r)}.$$

Example: CLT

Suppose $X_1, \ldots, X_n$ i.i.d. F, with $\mu = \mathrm{E}X$ and $\sigma^2 = \mathrm{E}X^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i$, then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d.} N(0,1)$$

Proof:

$$\phi_{X-\mu}(t) = 1 + \frac{(it)^2\sigma^2}{2} + o(t^2)$$

$$\begin{aligned}\phi_{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}} &= \phi^n_{X-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \left(1 + \frac{1}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 \sigma^2 + o\left(\frac{t^2}{\sigma^2 n}\right)\right)^n \to e^{-t^2/2}\end{aligned}$$

Then the result comes from Lévy-Cramér's Theorem.

Some Remarks

Remark 10

$$\phi_{X-\mu}(t) = 1 + \frac{1}{2}(it)^2\sigma^2 + \cdots + \frac{(it)^r}{r!}m_r + o(|t|^r)$$

if $\mathrm{E}|X|^r < \infty$ *for* $r > 2$*, where* $m_j = \mathrm{E}(X-\mu)^j$ *is j-th central moment. Then,*

$$\phi_{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}} = \left(1 - \frac{1}{2}\frac{t^2}{n} - \frac{1}{6}\frac{it^3}{n^{3/2}}\left(\frac{m_3}{\sigma}\right)^3 + \frac{1}{24}\frac{t^4}{n^2}\left(\frac{m_4}{\sigma}\right)^4 + \cdots\right)^n$$

Higher order expansion of c.f. $\Rightarrow$ *Edgeworth Expansion.*

Remark 11

The c.f. determines all the moments of X*, but* $\{m_r := E(X)^r\}_{r=1}^n$ *cannot determine the law of* X*. This the famous* Moment problem. *Carleman's condition*

$$\sum_{r=1}^{\infty} m_{2r}^{-\frac{1}{2r}} = +\infty$$

gives a sufficient condition for the determinacy of X*.*

Cumulants (Semi-Invariants)

Definition 3.4

The cumulants κ_j's are obtained from a power series expansion of the cumulant generating function of a r.v. X:

$$K_X(t) := \log \phi_X(t) \;=\; \sum_{j\geq 1} \frac{(it)^j}{j!}\kappa_j =: \log\{1 + \sum_{j\geq 1} \frac{1}{j!} m_j (it)^j\}.$$

Matching, we have $\kappa_1 = m_1 = \mathrm{E}X$ and

$$\kappa_2 = m_2 - m_1^2 = \mathrm{E}(X - EX)^2 =: c_2,$$
$$\kappa_3 = m_3 - 3m_1 m_2 + 2m_1^3 = \mathrm{E}(X - EX)^3 =: c_3,$$
$$\kappa_4 = m_4 - 4m_1 m_3 - 3m_2^2 + 12 m_1^2 m_2 - 6m_1^4 = c_4 - 3c_2^2.$$

The higher order ($j > 3$) cumulants are different from central moment.

Cumulants (Semi-Invariants)

Considering $X_1, \ldots, X_n$ i.i.d. $F(\mu, \sigma^2)$, $Y_i = (X_i - \mu)/\sigma$. The cumulants for Y_i are:

$$\kappa_1 = 0, \quad \kappa_2 = 1, \quad \kappa_3 = \frac{\mathrm{E}(X_i - \mu)^3}{\sigma^3}, \quad \kappa_4 = \frac{\mathrm{E}(X_i - \mu)^4}{\sigma^4}$$

where κ_3 is called the Skewness of X, κ_4 is called Kurtosis. Then,

$$\phi_Y(t) = \exp\left\{\sum_{j\geq 1} \frac{(it)^j}{j!}\kappa_j\right\} = \exp\left\{-\frac{t^2}{2} + \sum_{j\geq 3} \frac{(it)^j}{j!}\kappa_j\right\}$$

$$\phi_{\frac{\bar{X}-\mu}{\sqrt{n\sigma^2}}} = \phi_Y^n(t) = \exp\left\{-\frac{t^2}{2} + \frac{(it)^3}{3!\sqrt{n}}\kappa_3 + \frac{(it)^4}{4!n}\kappa_4 + \cdots\right\}$$

Unlike the classical CLT, in this section, we will explore the general CLT when the random variables is independent but not identically distributed.

Definition 4.1

For each $n \geq 1$, let $\{X_{n1}, X_{n2}, \cdots, X_{nk_n}\}$ be a collection of random vectors on a probability space $(\Omega_n, \mathcal{F}_n, P_n)$ s.t. $X_{n1}, \cdots, X_{nk_n}$ are independent with $k_n \to \infty$ as $n \to \infty$. Then $\{X_{nj} : 1 \leq j \leq k_n\}_{n \geq 1}$ is called a double array of independent random vectors.

Notations:

$$S_n = \sum_{j=1}^{k_n} X_{nj}, \quad \alpha_{nj} = \mathrm{E}(X_{nj}), \quad \alpha_n = \sum_{j=1}^{k_n} \mathrm{E}(X_{nj}) = \sum_{j=1}^{k_n} \alpha_{nj},$$

$$\sigma_{nj}^2 = \mathrm{Var}(X_{nj}), \quad \sigma_n^2 = \sum_{j=1}^{k_n} \sigma_{nj}^2.$$

A Useful Lemma

A useful lemma in mathematical analysis.

Lemma 4.2

Let $\{\theta_{nj} : 1 \le j \le k_n\}_{n\ge1}$ *be a double array of complex numbers satisfying as* $n \to \infty$,

1. $\max_{1\le j\le k_n} |\theta_{nj}| \to 0$,
2. $\sum_{j=1}^{k_n} |\theta_{nj}| \le M < \infty$ *where* M *is free of* n,
3. $\sum_{j=1}^{k_n} \theta_{nj} \to \theta$ *for a finite complex* θ, *then* $\prod_{j=1}^{k_n}(1+\theta_{nj}) \to e^{\theta}$.

This is a generalized formula for $\lim_{n\to\infty}(1+\theta/n)^n \to e^{\theta}$ *with* $\theta_{nj} \equiv \theta/n$.

References for this chapter

Chung, K. L. (2001). A course in probability theory, 3rd. Academic press.

Preliminary

For any complex $z \neq 0$, the complex number w satisfying $e^w = z$ is $\operatorname{Log} z$, i.e. $\operatorname{Log} z = w$. Let $w = u + vi$, we have $z = e^u e^{iv}$, which means

$$|z| = e^u, \quad u = \log|z|, \quad v = \operatorname{Arg} z = \arg z + 2k\pi, \quad \arg z \in [-\pi, \pi]$$

then:

$$\operatorname{Log} z = \log|z| + i \operatorname{Arg} z, \quad \log z = \log|z| + i \arg z$$

Remark 12

For any $|z| < 1$,

$$\log(1+z) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{z^n}{n}$$

The Proof of Lemma 4.2

$$|\log(1+\theta_{nj}) - \theta_{nj}| = \left|(-1)^{m-1}\frac{\theta_{nj}^m}{m}\right| \le \frac{|\theta_{nj}|^m}{m}$$
$$\le \frac{|\theta_{nj}|^2}{2}\sum_{m=2}^{\infty}\left(\frac{1}{2}\right)^{m-2} = |\theta_{nj}|^2 < 1$$

Here the upper bound 1 is uniformly for all $\{\theta_{nj}\}$, hence,

$$\log(1+\theta_{nj}) = \theta_{nj} + \Lambda_{nj}|\theta_{nj}|^2, \quad \text{where } |\Lambda_{nj}| < 1$$

Now,

$$\sum_{j=1}^{k_n}\log(1+\theta_{nj}) = \sum_{j=1}^{k_n}\theta_j + \sum_{j=1}^{k_n}\Lambda_{nj}|\theta_{nj}|^2$$

The Proof of Lemma 4.2

From (i) and (ii):

$$\left|\sum_{j=1}^{k_n} \Lambda_{nj}|\theta_{nj}|^2\right| \le \max_{1\le j\le k_n} |\theta_{nj}| \sum_{j=1}^{k_n} |\theta_{nj}|$$
$$\overset{(ii)}{\le} \max_{1\le j\le k_n} |\theta_{nj}| M \xrightarrow{(i)} 0$$

And from (iii), we know that

$$\sum_{j=1}^{k_n} \log(1+\theta_{nj}) \to \theta$$

Liapounov's Theorem

Theorem 4.3

For a double array $\{X_{nj} : 1 \le j \le k_n\}_{n\ge 1}$, *let* $\Gamma_n = \sum_{j=1}^{k_n} \mathrm{E}|X_{nj} - \alpha_{nj}|^3$, *which is finite for every* n, *and if Liapounov's condition holds*

$$\frac{\Gamma_n}{\sigma_n^3} = \frac{1}{\sigma_n^3}\sum_{j=1}^{k_n} \mathrm{E}|X_{nj} - \alpha_{nj}|^3 \to 0 \quad \textit{as } n \to \infty, \textit{ then}$$

$$\frac{S_n - \alpha_n}{\sigma_n} \xrightarrow{d} N(0,1).$$

Remark 13

The constant 3 can be relaxed to $2 + \delta$ *some* $\delta > 0$.

The Proof of Liapounov's Theorem

Let $\gamma_{nj} = \mathrm{E}|X_{nj} - \alpha_{nj}|^3$. As:

$$\sigma_{nj} = \left(\mathrm{E}|X_{nj} - \alpha_{nj}|^2\right)^{1/2} \leq \left(\mathrm{E}|X_{nj} - \alpha_{nj}|^3\right)^{1/3}$$

We have $\sigma_{nj}^3 \leq \gamma_{nj}$, thus,

$$\max_{1 \leq j \leq k_n} \sigma_{nj}^3 \leq \max_{1 \leq j \leq k_n} \gamma_{nj} < \Gamma_n \tag{6}$$

Let ϕ_{nj} be the c.f. of $(X_{nj} - \alpha_{nj})/\sigma_n$. As γ_{nj} is finite, from Theorem 2.8,

$$\phi_{nj}(t) = 1 - \frac{\sigma_{nj}^2 t^2}{2\sigma^2} + \frac{\Lambda_{nj}}{6} \frac{\gamma_{nj} t^3}{\sigma_n^3}, \quad \text{where } |\Lambda_{nj}| < 1$$

$$\begin{aligned} \max_{1 \leq j \leq k_n} |\phi_{nj}(t) - 1| &\leq \frac{t^2}{2\sigma_n^2} \max_{1 \leq j \leq k_n} \sigma_{nj}^2 + \frac{t^3}{6\sigma_n^3} \max_{1 \leq j \leq k_n} \gamma_{nj} \\ &\leq \frac{t^2}{2} \left(\frac{\Gamma_n}{\sigma_n^3}\right)^{2/3} + \frac{t^3}{6\sigma_n^3} \max_{1 \leq j \leq k_n} \gamma_{nj} \xrightarrow{(6)} 0 \end{aligned} \tag{7}$$

which is the Assumption (i) of Lemma (4.2).

In fact, note that $\sigma_{nj}^2 = (\sigma_{nj}^3)^{2/3} \leq (\max_j \sigma_{nj}^3)^{2/3}$,

$$\frac{\max \sigma_{nj}^2}{\sigma_n^2} \leq \left(\frac{\max \sigma_{nj}^3}{\sigma_n^3}\right)^{2/3} \overset{(6)}{\leq} \left(\frac{\Gamma_n}{\sigma_n^3}\right)^{2/3} \to 0$$

and $\max \gamma_{nj}/\sigma_n^3 \to 0$ comes from the condition directly. On the other hand,

$$\sum_{j=1}^{k_n} |\phi_{nj}(t) - 1| \leq \frac{\sum \sigma_{nj}^2 t^2}{2\sigma^2} + \frac{t^3}{6}\frac{\Gamma_n}{\sigma_n^3} = \frac{t^2}{2} + \frac{t^3}{6}\frac{\Gamma_n}{\sigma_n^3} \leq M(t) \tag{8}$$

which is the Assumption (ii) of Lemma (4.2).

Finally, as

$$\left|\sum_{j=1}^{k_n} \frac{\Lambda_{nj}\gamma_{nj}}{\sigma_n^3}\right| \leq \frac{\Gamma_n}{\sigma_n^3} \to 0$$

we have,

$$\sum_{j=1}^{k_n}(\phi_{nj}(t)-1) = -\frac{t^2}{2} + t^3 \sum_{j=1}^{k_n} \frac{\Lambda_{nj}\gamma_{nj}}{\sigma_n^3} \to -\frac{t^2}{2} \tag{9}$$

which is the Assumption (iii) of Lemma (4.2). Then from (7) to (9) and Lemma (4.2), the c.f. of $(S_n - \alpha_n)/\sigma_n = \sum_{j=1}^{k_n}(X_{nj} - \alpha_{nj})/\sigma_n$ satisfying:

$$\prod_{j=1}^{k_n} \phi_{nj}(t) = \prod_{j=1}^{k_n}(1 + \phi_{nj}(t) - 1) \to e^{-t^2/2}$$

Apply Lévy-Cramér's Theorem., we obtain the result.

Corollary

Theorem 4.3 implies the following for a single array $\{X_n\}_{n\geq 1}$.

Corollary 4.4

Let $\{X_n\}_n \geq 1$ *be a sequence of independent random vectors,* $\alpha_j = \mathrm{E}(\mathrm{X_j})$, $\sigma_j^2 = \mathrm{Var}(\mathrm{X_j})$ *and* $\gamma_j = \mathrm{E}|\mathrm{X_j} - \alpha_\mathrm{j}|^3 < \infty$. *Let* $P_n = \sum_{j=1}^n \gamma_j$, *if* $\frac{P_n}{\sigma_n^3} \to 0$, *then*

$$\frac{S_n - \sum_{j=1}^n \alpha_j}{\sigma_n} \xrightarrow{d} N(0,1).$$

Proof

It can be proved by Lindeberg's method. Here we use another approach to verify it. W.L.O.G. assumimg $\alpha_j = 0$.

For any $f \in C^3 := \{g \mid g^{(3)} \text{ is continuous in } \mathbb{R}\}$. Let $Y_1, \ldots, Y_n$ are independent r.v. with $Y_j \sim N(0, \sigma_j^2)$ matching the first two moments of X_j, and $Y_0 = \sum_{i=1}^n Y_i/\sigma_n \sim N(0,1)$. We want to show $\forall f \in C^3$,

$$\mathrm{E}f\left(\frac{\sum_{i=1}^n X_i}{\sigma_n}\right) - \mathrm{E}f(Y_0) \to 0 \tag{10}$$

which implies $\sum_{i=1}^n X_i/\sigma_n \xrightarrow{d.} Y_0$ (Recall Portmanteau Lemma).

Proof

Let $Z_j = Y_1 + \ldots + Y_{j-1} + X_{j+1} + \ldots + X_n$ for $2 \le j \le n-1$, $Z_1 = X_2 + \ldots + X_n$, and $Z_n = Y_1 + \ldots + Y_{n-1}$, then:

$$\mathrm{E}f\left(\frac{\sum_{i=1}^n X_i}{\sigma_n}\right) - \mathrm{E}f\left(\frac{\sum_{i=1}^n Y_i}{\sigma_n}\right) = \sum_{i=1}^n \left[\mathrm{E}f\left(\frac{Z_i + X_i}{\sigma_n}\right) - \mathrm{E}f\left(\frac{Z_i + Y_i}{\sigma_n}\right)\right] \tag{11}$$

Note that:

$$f\left(\frac{Z_i + X_i}{\sigma_n}\right) = f\left(\frac{Z_i}{\sigma_n}\right) + f'\left(\frac{Z_i}{\sigma_n}\right)\frac{X_i}{\sigma_n} + \frac{1}{2}f''\left(\frac{Z_i}{\sigma_n}\right)\frac{X_i^2}{\sigma_n^2} + \theta_i^{(1)}\frac{X_i^3}{3!\sigma_n^3}$$
$$f\left(\frac{Z_i + Y_i}{\sigma_n}\right) = f\left(\frac{Z_i}{\sigma_n}\right) + f'\left(\frac{Z_i}{\sigma_n}\right)\frac{Y_i}{\sigma_n} + \frac{1}{2}f''\left(\frac{Y_i}{\sigma_n}\right)\frac{Y_i^2}{\sigma_n^2} + \theta_i^{(2)}\frac{Y_i^3}{3!\sigma_n^3}$$

where $|\theta_i^{(l)}| \le \|f^{(3)}\|_\infty < \infty$. As $\mathrm{E}X_i = \mathrm{E}Y_i = 0$, $\mathrm{E}X_i^2 = \mathrm{E}Y_i^2 = \sigma_i^2$, $\mathrm{E}Y_i^3 = \sqrt{8/\pi}\sigma_i^3$,

$$\begin{aligned}\left|\mathrm{E}f\left(\frac{Z_i + X_i}{\sigma_n}\right) - \mathrm{E}f\left(\frac{Z_i + Y_i}{\sigma_n}\right)\right| &\le \frac{1}{3!\sigma^3}|\mathrm{E}\left[\theta_i^{(1)}X_i^3\right] - \mathrm{E}\left[\theta_i^{(2)}Y_i^3\right]| \\ &\le \frac{M}{3!\sigma_n^3}\left(\gamma_i + \sqrt{\frac{8}{\pi}}\sigma_i^3\right)\end{aligned}$$

Proof

Then from (11),

$$\left|\mathrm{E}f\left(\frac{\sum_{i=1}^n X_i}{\sigma_n}\right) - \mathrm{E}f\left(\frac{\sum_{i=1}^n Y_i}{\sigma_n}\right)\right| \le \sum_{i=1}^n \left|\mathrm{E}f\left(\frac{Z_i + X_i}{\sigma_n}\right) - \mathrm{E}f\left(\frac{Z_i + Y_i}{\sigma_n}\right)\right|$$

$$(M = \max\{|\theta_i^{(1)}|, |\theta_i^{(2)}|\}) \quad \le \frac{M}{6}\left(\frac{\sum_{i=1}^n \gamma_i}{\sigma_n^3} + \sqrt{\frac{8}{\pi}}\frac{\sum_{i=1}^n \sigma_i^3}{\sigma_n^3}\right)$$

$$(\sum_{i=1}^n \sigma_i^3 \le \Gamma_n) \quad \le \frac{M_1}{6}\frac{\Gamma_n}{\sigma_n^3} \to 0$$

As a result, (10) is true.

Remark 14

The method of (11) is called Telescoping.

An immediate consequence of Liapounov CLT is

Corollary 4.5

For a double array (triangular array of independent variables) $\{X_{nj}, 1 \le j \le k_n\}_{n\ge 1}$*, if* $|X_{nj}| \le M_{nj}$ *a.e., and* $\lim_{n\to\infty} \max_{1\le j\le k_n} M_{nj} = 0$*. Let* $S_n = \sum_{j=1}^{k_n} X_{nj}$*, show that*

$$\frac{S_n - \mathrm{E}\,(S_n)}{\sigma_n} \xrightarrow{d} N(0,1).$$

Null Array

Four conditions:

(a) $\forall j, \lim_{n\to\infty} P(|X_{nj} - \alpha_{nj}| > \varepsilon\sigma_{nj}) = 0,$

(b) $\lim_{n\to\infty} \max_{1\le j\le k_n} P(|X_{nj} - \alpha_{nj}| > \varepsilon\sigma_{nj}) = 0,$

(c) $\lim_{n\to\infty} P(\max_{1\le j\le k_n} |X_{nj} - \alpha_{nj}| > \varepsilon\sigma_{nj}) = 0,$

(d) $\lim_{n\to\infty} \sum_{j=1}^{k_n} P(|X_{nj} - \alpha_{nj}| > \varepsilon\sigma_{nj}) = 0$

Homework: check $(d) \Rightarrow (c) \Rightarrow (b) \Rightarrow (a)$.

Definition 4.6

A double array satisfying condition (b) is called a null array.

Equivalent Form of Null Arrays

Proposition 4.7

A double array $\{X_{nj}, 1 \leq j \leq k_n\}_{n\geq 1}$ *is a null array iff*

$$\forall t \in \mathbb{R}, \lim_{n\to\infty} \max_{1\leq j\leq k_n} |\phi_{nj}(t) - 1| = 0 \qquad (e),$$

where ϕ_{nj} *is a c.f. of* $\frac{X_{nj}-\alpha_j}{\sigma_n}$. *Furthermore, the convergence in* (e) *is uniformly over any finite interval.*

Remark 15

1. *If* $\{X_{nj}\}$ *is a NA, then* $\frac{X_{nj}-\alpha_j}{\sigma_n} \xrightarrow{p} 0$.
2. *Prop 4.7 implies that each element of a null array* $\left\{\frac{X_{nj}-\alpha_j}{\sigma_n}\right\}_{j=1}^{k_n}$ *uniformly degenerate at 0 on* j *when* $n \to \infty$.

Proof

"$\Rightarrow$": WLOG assume $\alpha_j = 0$,

$$|\phi_{nj}(t) - 1| = \left|\mathrm{E}\left(e^{itX_{nj}/\sigma_n} - 1\right)\right| \leq \mathrm{E}\left[\left|e^{itX_{nj}/\sigma_n} - 1\right| \mathbb{I}(|X_{nj}| > \epsilon\sigma_n)\right] + \mathrm{E}\left[\left|e^{itX_{nj}/\sigma_n} - 1\right| \mathbb{I}(|X_{nj}| \leq \epsilon\sigma_n)\right]$$

$$(|e^{itu} - 1| = \sqrt{2(1 - \cos tu)} \leq |tu|) \leq 2\mathrm{P}\left(|X_{nj}| > \epsilon\sigma_n\right) + |t|\mathrm{E}\left[\left|\frac{X_{nj}}{\sigma_n}\right| \mathbb{I}\left(\left|\frac{X_{nj}}{\sigma_n}\right| \leq \epsilon\right)\right]$$

$$\leq 2\mathrm{P}\left(|X_{nj}| > \epsilon\sigma_n\right) + |t|\epsilon$$

Thus,

$$\max_j |\phi_{nj}(t) - 1| \leq 2\max_j \mathrm{P}\left(|X_{nj}| > \epsilon\sigma_n\right) + |t|\epsilon$$

(e) is veracious. In fact, for any $|t| \leq K$,

$$\sup_{|t|\leq K} \max_j |\phi_{nj}(t) - 1| \leq 2\max_j \mathrm{P}\left(|X_{nj}| > \epsilon\sigma_n\right) + K\epsilon$$

so the convergence is uniform over $t \in [-K, K]$ as dersired.

Proof

"$\Leftarrow$": It can be derived by using Lemma 2.3.2 (Homework).

$$\mathrm{P}\left(\left|\frac{X_{nj}}{\sigma_n}\right| > \frac{2}{\delta}\right) \overset{(*)}{\leq} \frac{1}{\delta}\int_{|t|\leq\delta}(1-\phi_{nj}(t))dt = \frac{1}{\delta}\left|\int_{|t|\leq\delta}(1-\phi_{nj}(t))dt\right|$$
$$\leq \frac{1}{\delta}\int_{|t|\leq\delta}|1-\phi_{nj}(t)|dt$$

This implies,

$$\max_j \mathrm{P}\left(\left|\frac{X_{nj}}{\sigma_n}\right| > \frac{2}{\delta}\right) \leq \max_j \frac{1}{\delta}\int_{|t|\leq\delta}|1-\phi_{nj}(t)|dt$$

From the BCT and (e), we know condition (h) holds.

Proof

$(*)$:

$$\mathbb{I}(|\delta x| > 2) \leq 2\left(1 - \frac{\sin \delta x}{\delta x}\right) = \frac{2}{\delta}\int_{-\delta}^{\delta}(1 - \cos tx)dt$$
$$= \frac{2}{\delta}\int_{|t|\leq\delta}(1 - \cos tx - i\sin tx)dt$$

Then $(*)$ comes from taking expectation on both sides.

Remark 16

Bounded Convergence Theorem (BCT): Suppose $f_n(t) \to f(t)$ for $\forall t$, and

$$|f_n(t)| \leq g(t), \quad \int g(t)dt \quad \textit{exists}$$

Then,

$$\int |f_n(t) - f(t)|dt \to 0, \quad \int f_n(t)dt \to \int f(t)dt$$

Lindeberg Condition (LC)

Definition 4.8

A double array $\{X_{nj,1\le j\le k_n}\}_{n\ge 1}$ is said to satisfy the Lindeberg condition, if $\forall\varepsilon > 0$,

$$\lim_{n\to\infty}\sigma_n^{-2}\sum_{j=1}^{k_n}\mathrm{E}\{(\mathrm{X_{nj}}-\alpha_{\mathrm{nj}})^2\mathrm{I}(|\mathrm{X_{nj}}-\alpha_{\mathrm{nj}}|>\varepsilon\sigma_{\mathrm{n}})\}=0,$$

where $\alpha_{nj}=\mathrm{E}(\mathrm{X_{nj}})$, $\sigma_n^2=\sum_{j=1}^{k_n}\mathrm{Var}(\mathrm{X_{nj}})$, implicitly assumed $\sigma_{nj}^2=\mathrm{E}(\mathrm{X_{nj}^2})<\infty$ for any n and j, .

Lindeberg-Feller CLT

Lemma 4.9

Let $u(m,n)$ *be a function of positive integers* m *and* n*, s.t.* $\forall m, \lim_{n\to\infty} u(m,n) = 0$*, then there exists a monotone increasing sequence* $\{m_n\}$*,* $m_n \to \infty$*, s.t.* $\lim_{n\to\infty} u(m_n, n) = 0$

Lindeberg's condition is a sufficient condition (and under certain conditions also a necessary condition) for the CLT to hold for a sequence of independent random variables.

Theorem 4.10

[Lindeberg-Feller] For a double array $\{X_{nj,1\le j\le k_n}\}_{n\ge1}$*, assume* $\mathrm{Var(X_{nj})} = \sigma^2_{\mathrm{nj}} < \infty$*, then*
(i)$\frac{S_n - \mathrm{ES_n}}{\sigma_n} \xrightarrow{d} N(0,1)$ *and (ii) the double array is a null array iff the* Lindeberg condition *is satisfied.*

The Proof of Lemma 4.9

As $\lim_{n\to\infty} u(m,n) = 0$ for each m. $\exists n_m$, s.t.

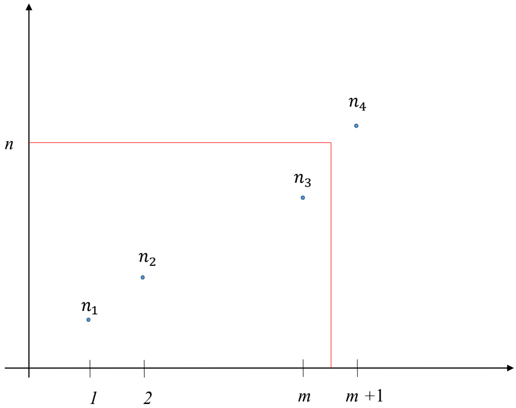
$$n \geq n_m, \quad u(m, n_m) \leq \frac{1}{m}$$

Here we obtain a sequence $\{n_m\}_{m\geq 1}$, and can make it strictly increase to ∞.
Now let $m_n = m$ s.t. $n_m \leq n \leq n_{m+1}$. When $n \geq n_m$,

$$u(m_n, n) = u(m, n) \leq \frac{1}{m}$$

As $n_m \uparrow \infty$, $m_n \uparrow \infty$ too, and $\lim_{n\to\infty} u(m_n, n) = 0$.

The Proof of Lemma 4.9



$$n_3 \le n \le n_4, \quad \Rightarrow \quad m_n = 3$$

WLOG, assume $\mathrm{E}X_{nj} = 0$ and $\sigma_n^2 = 1$, or we can redefine $X_{nj} = (X_{nj} - \mathrm{E}X_{nj})/\sigma_n$. Now truncate X_{nj} to X'_{nj} with a $\eta \in (0, 1)$:

$$X'_{nj} = \begin{cases} X_{nj} & \text{if } |X_{nj}| < \eta \\ 0 & \text{o.w.} \end{cases} \tag{12}$$

Denote $S'_n = \sum_{i=1}^{k_n} X'_{nj}$, $\sigma'^2_n = \sum_{i=1}^{k_n} \mathrm{var}(X'_{nj}) = \mathrm{var}(S'_n)$.

$$|\mathrm{E}X'_{nj}| = \left| \int_{|x|<\eta} x \, dF_{nj}(x) \right| \overset{\mathrm{E}X_{nj}=0}{=\!=\!=} \left| \int_{|x|\geq\eta} x \, dF_{nj}(x) \right| \leq \frac{1}{\eta} \int_{|x|\geq\eta} x^2 \, dF_{nj}(x)$$

Hence,

$$|\mathrm{E}S'_n| \leq \sum_{i=1}^{k_n} |\mathrm{E}X'_{nj}| \leq \frac{1}{\eta} \sum_{j=1}^{k_n} \int_{|x|\geq\eta} x^2 \, dF_{nj}(x) \xrightarrow{\text{Lind Con}} 0 \tag{13}$$

Similarly,

$$\begin{aligned}\sum_{j=1}^{k_n} \mathrm{E}X_{nj}'^2 &= \sum_{j=1}^{k_n} \int_{|x|<\eta} x^2\, dF_{nj}(x) \\ &= \sum_{j=1}^{k_n} \left[\int x^2 dF_{nj}(x) - \int_{|x|\geq\eta} x^2 dF_{nj}(x) \right] \xrightarrow{\sigma_n^2=1} 1\end{aligned} \tag{14}$$

Hence,

$$\sigma_n'^2 = \mathrm{var}(S_n') = \sum_{j=1}^{k_n} \mathrm{E}X_{nj}'^2 - \sum_{j=1}^{k_n} (\mathrm{E}X_{nj}')^2 \xrightarrow{(14)} 1 = \sigma_n^2$$

where we use the fact:

$$\sum_{j=1}^{k_n} (\mathrm{E}X_{nj}')^2 \leq \left(\sum_{j=1}^{k_n} |\mathrm{E}X_{nj}'| \right)^2 \xrightarrow{(12)} 0$$

As

$$\frac{S_n'}{\sigma_n'} = \frac{S_n' - \mathrm{E}S_n'}{\sigma_n'} + \frac{\mathrm{E}S_n'}{\sigma_n'},$$

$\mathrm{E}S_n' \to 0$, and $\sigma_n' \to 1$. From Slusky Theorem S_n'/σ_n' and $S_n' - \mathrm{E}S_n'/\sigma_n'$ would convergence to the same distribution.

Next to show:

$$\frac{S_n' - \mathrm{E}S_n'}{\sigma_n'} \xrightarrow{d.} N(0,1)$$

The Proof of Theorem 4.10 $\Rightarrow$

From the LC, for each fixed $m \geq 1$,

$$\lim_{n\to\infty} m^2 \sum_{j=1}^{k_n} \int_{|x|>1/m} x^2 dF_{nj}(x) = 0$$

From Lemma 4.9, there exists $\{m_n\} \uparrow \infty$ s.t.

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} m_n^2 \int_{|x|>1/m_n} x^2 dF_{nj}(x) = 0 \tag{15}$$

Let $\eta_n = m_n^{-1} \downarrow 0$, and use η_n to replace η in the definition of X'_{nj}, then

$$|X'_{nj}| \leq \eta_n := M_{nj}, \quad \lim_{n\to\infty} \max_{1\leq j\leq k_n} M_{nj} = \lim_{n\to\infty} \eta_n = 0$$

Then from corollary 4.5, $(S'_n - \mathrm{E}S'_n)/\sigma'_n \to N(0,1)$. So does S'_n/σ'_n. So $S'_n \xrightarrow{d.} N(0,1)$ as $\sigma'_n \to 1$.

Since

$$\mathrm{P}(S_n \neq S_n') \leq \mathrm{P}\left(\bigcup_{j=1}^{k_n}\{X_{nj} \neq X_{nj}'\}\right) \leq \sum_{j=1}^{k_n}\mathrm{P}(|X_{nj}| \geq \eta_n)$$

$$\leq \sum_{j=1}^{k_n}\frac{1}{\eta_n^2}\int_{|x|>\eta_n} x^2 dF_{nj}(x) \xrightarrow{LC} 0$$

So we have $S_n - S_n' \xrightarrow{p.} 0$. By Slusky, $S_n \xrightarrow{d.} N(0,1)$. Here we complete the proof of sufficiency.

Let ϕ_{nj} be the c.f of X_{nj} with $\mathrm{E}X_{nj} = 0$ and $\sigma_n^2 = 1$. As $S_n \xrightarrow{d.} N(0, 1)$ in (i),

$$\lim_{n\to\infty} \prod_{j=1}^{k_n} \phi_{nj}(t) = e^{-t^2/2}, \quad \lim_{n\to\infty} \sum_{j=1}^{k_n} \log \phi_{jn}(t) = -\frac{t^2}{2} \tag{16}$$

and it can be reinforced that it would be holed uniform over $t \in [-K, K]$ by using covering method.

On the other hand, (ii) and the properties of NA imply,

$$\lim_{n\to\infty} \sup_{|t|<K} \max_{1\le j\le k_n} |\phi_{nj}(t) - 1| = 0 \tag{17}$$

Let $\theta_{nj} = \phi_{nj}(t) - 1$, from the proof of Lemma 4.2, we know the following display holds.

The following display holds.

$$\log\phi_{nj}(t) = \log(1+\theta_{nj}) = \theta_{nj} + \Lambda_{nj}|\theta_{nj}|^2 = \phi_{nj}(t) - 1 + \Lambda_{nj}|\phi_{nj}(t)-1|^2 \quad (18)$$

where $|\Lambda_{nj}| < 1$ uniformly. Furthermore,

$$\sum_{j=1}^{k_n}|\phi_{nj}(t)-1|^2 \le \max_{1\le j\le k_n}|\phi_{nj}(t)-1|\sum_{j=1}^{k_n}|\phi_{nj}(t)-1| \xrightarrow{(17)\&(*)} 0 \quad (19)$$

where $(*)$ is the fact that,

$$\sum_{j=1}^{k_n}|\phi_{nj}(t)-1| = \sum_{j=1}^{k_n}\left|\int(e^{itx}-1)dF_{nj}(x)\right|$$

$$\overset{\text{Taylor Exp}}{\underset{\text{of } e^{itx}}{=\!=\!=\!=}} \sum_{j=1}^{k_n}\left|\int(itx+\kappa_t\frac{t^2x^2}{2})dF_{nj}(x)\right| \le \frac{t^2}{2} < \infty$$

in which $\kappa_t \in (0,1)$.

The Proof of Theorem 4.10 $\Rightarrow$

Then, from (16) and (18), we obtain:

$$\lim_{n\to\infty}\sum_{j=1}^{k_n}(\phi_{nj}(t)-1)=\lim_{n\to\infty}\sum_{j=1}^{k_n}\log\phi_{jn}(t)=-\frac{t^2}{2}$$

By taking the real part,

$$\lim_{n\to\infty}\sum_j\int_{-\infty}^{\infty}(1-\cos tx)dF_{nj}(x)=\frac{t^2}{2}$$

Hence for each $\eta>0$, split the integral into two parts, by Chebyshev's inequality,

$$\begin{aligned}\overline{\lim_{n\to\infty}}\left|\frac{t^2}{2}-\sum_j\int_{|x|\le\eta}(1-\cos tx)dF_{nj}(x)\right| &= \overline{\lim_{n\to\infty}}\left|\sum_j\int_{|x|>\eta}(1-\cos tx)dF_{nj}(x)\right|\\ &\le \overline{\lim_{n\to\infty}}\sum_j\int_{|x|>\eta}2dF_{nj}(x)\\ &\le \overline{\lim_{n\to\infty}}\,2\sum_j\frac{\sigma_{nj}^2}{\eta^2}=\frac{2}{\eta^2}\end{aligned}$$

Note that $0 \leq 1 - \cos\theta \leq \theta^2/2$ for every real θ, this implies,

$$\frac{2}{\eta^2} \geq \overline{\lim_{n\to\infty}} \left\{ \frac{t^2}{2} - \sum_j \frac{t^2}{2} \int_{|x|\leq n} x^2 dF_{nj}(x) \right\} \geq 0$$

Therefore, for fixed $\eta > 0$,

$$\overline{\lim_{n\to\infty}} \sum_{j=1}^{k_n} \mathrm{E}\left(|X_{nj}^2| \mathbb{I}\left(|X_{nj}| > \eta\right)\right) = \overline{\lim_{n\to\infty}} \left\{ 1 - \sum_{j=1}^{k_n} \int_{|x|\leq\eta} x^2 dF_{nj}(x) \right\} \leq \frac{4}{t^2\eta^2}$$

Let $t \to \infty$, we have:

$$\sum_{j=1}^{k_n} \mathrm{E}\left(|X_{nj}^2| \mathbb{I}\left(|X_{nj}| > \eta\right)\right) \longrightarrow 0$$

which exactly is the LC.

Example: Regression

$$y_j = x_j\beta + \epsilon_j, \quad \epsilon_j \text{ i.i.d. } N(0, \sigma_\epsilon^2), \quad j = 1, 2, \ldots$$

where $\{x_j\}_{j \geq 1}$ are fixed design points, s.t.

$$\max_{1 \leq j \leq n} \frac{|x_j|}{a_n} \to 0, \quad a_n^2 = \sum_{j=1}^n x_j^2$$

The ordinary least square estimate is $\hat{\beta}_{LS} = \sum_{j=1}^n x_j y_j / a_n^2$. We want to show that $a_n(\hat{\beta}_{LS} - \beta) \xrightarrow{d.} N(0, \sigma^2)$.

$$a_n(\hat{\beta}_{LS} - \beta) = \frac{\sum x_j y_j - \beta \sum x_j^2}{a_n} = \frac{\sum x_j \epsilon_j}{a_n} =: \sum_{j=1}^n X_{nj}$$

where:

$$X_{nj} = \frac{x_j \epsilon_j}{\sqrt{\sum x_j^2}}, \quad \alpha_{nj} = \mathrm{E} X_{nj} = 0, \quad \sigma_{nj}^2 = \frac{x_j^2 \sigma^2}{a_n^2}, \quad \sigma_n^2 = \sigma_\epsilon^2$$

Example: Regression

Here $\{X_{nj}, 1 \le j \le n\}_{n\ge1}$ is a triangular array, and

$$\begin{aligned}
\sigma_n^{-2}\sum_{j=1}^{n}\mathrm{E}\left[X_{nj}^2\mathbb{I}(|X_{nj}|>\delta\sigma_n)\right] &= \frac{1}{\sigma_n^2 a_n^2}\sum_{j=1}^{n}x_j^2\mathrm{E}\left[\epsilon_j^2\mathbb{I}\left(|x_j\epsilon_j/a_n|>\delta\sigma_n\right)\right] \\
(m_n=\max_j|x_j/a_n|) &\le \frac{1}{\sigma^2 a_n^2}\sum_{j=1}^{n}x_j^2\mathrm{E}\left[\epsilon_j^2\mathbb{I}(|\epsilon_j|>\delta\sigma_\epsilon m_n^{-1})\right] \\
&= \frac{1}{\sigma_\epsilon^2}\mathrm{E}\left[\epsilon_j^2\mathbb{I}(|\epsilon_j|>\delta\sigma_\epsilon m_n^{-1})\right] \longrightarrow 0
\end{aligned}$$

as $m_n \to 0$. From Theorem 4.10, we know that,

$$a_n(\hat\beta_{LS}-\beta)=\sum_{j=1}^{n}X_{nj} \xrightarrow{d.} N(0,\sigma_\epsilon^2)$$

A Sufficient Condition of LC

There is a sufficient condition to verify the Lindeberg-Feller condition.

Proposition 4.11

For a double array $\{X_{nj}\}_{j=1}^{k_n}$ with means $\{\mu_{nj}\}$ and variances $\{\sigma_{nj}^2\}$. If for some $\nu > 2$,

$$\sum_{j=1}^{k_n} \mathrm{E}|X_{nj} - \mu_{nj}|^\nu = o(\sigma_n^\nu)$$

Then, the Lindeberg condition is valid.

Proof

$$\begin{aligned}
\mathrm{E}\left[(X_{nj}-\mu_{nj})^2\,\mathbb{I}\left(|X_{nj}-\mu_{nj}|>\epsilon\sigma_n\right)\right] &= \int_{|t-\mu_{nj}|>\epsilon\sigma_n}(t-\mu_{nj})^2 dF_{nj}(t) \\
&\le (\epsilon\sigma_n)^{2-\nu}\int_{|t-\mu_{nj}|>\epsilon\sigma_n}(t-\mu_{nj})^\nu dF_{nj}(t) \\
&\le (\epsilon\sigma_n)^{2-\nu}\mathrm{E}|X_{nj}-\mu_{nj}|^\nu
\end{aligned}$$

Therefore,

$$\frac{1}{\sigma_n^2}\sum_{j=1}^{k_n}\mathrm{E}\left[(X_{nj}-\mu_{nj})^2\,\mathbb{I}\left(|X_{nj}-\mu_{nj}|>\epsilon\sigma_n\right)\right]$$

$$\le \epsilon^{2-\nu}\frac{\sum_{j=1}^{k_n}\mathrm{E}|X_{nj}-\mu_{nj}|^\nu}{\sigma_n^\nu}\longrightarrow 0$$

The LC holds.

CLT for m-dependent r.v.s

Definition 4.12

Def: A sequence of r.v. $\{X_n\}_{n\geq 1}$ istb m-dependent if $\exists$ a positive integer m s.t. for any $n \geq 1$ and $j \geq m, X_{n+j}$ is independent of $\mathcal{F}_n = \sigma\{X_j, 1 \leq j \leq n\}$, the σ-field generated by $\{X_j\}_{j=1}^n$.

Theorem 4.13

Let $\{X_n\}_{n\geq 1}$ be a sequence of m-dependent rv.s with uniformly bounded variance s.t. $\frac{\sigma_n}{mn^{1/3}} \triangleq \frac{\sqrt{Var(\sum_{i=1}^n X_i)}}{mn^{1/3}} \to \infty$ as $n \to \infty$ and $m = o(n^{1/3})$. Then

$$\frac{S_n - E(S_n)}{\sigma_n} \xrightarrow{d} N(0,1).$$

A CLT for more general dependent data, the so-called mixing dependent sequence, will be discussed later.

Proof: Use of Blocking Technique

Proof: As $\{X_n\}_{n\geq 1}$ has uniformly bounded variance. $\exists M$ s.t. $\sup_n |Var(X_n)| \leq M$.

WLOG, assume $E(X_j) = 0$.

We block the whole sequence by larger blocks followed by small blocks.

Let $k = [n^{1/3}]$ be the size of large blocks, m be the size of small blocks

$p = [\frac{n}{k+m}] = O(n^{2/3})$ be the number of blocks, $B_j = j(k+m)$,

Lay out of the Blocks

$Y_1 = X_1 + \cdots + X_k, \quad Z_1 = X_{k+1} + \cdots + X_{k+m}$

$\cdots$

$Y_p = X_{B_{p-1}+1} + \cdots + X_{B_{j-1}+k}, \, Z_p = X_{B_{p-1}+k+1} + \cdots + X_{B_p}$

$R_p = X_{B_p} + \cdots + X_n$ the residual block.

As $k >> m$ when n is large enough, then $\{Y_j\}_{j=1}^p$ and $\{Z_j\}_{j=1}^p$ are indpt rvs as they are at least m-apart.

The Detailed Proof **

Now, we have:

$$S_n = \sum_{j=1}^{p} Y_j + \sum_{j=1}^{p} Z_j + \sum_{l=1}^{n-p(k+m)} X_{B_l+l} := S_n' + S_n'' + S_n''' \tag{20}$$

As $\sup \operatorname{var}(X_j) \le M$, $|\mathrm{E}(X_j X_l)| \le M$, and

$$\begin{aligned} \operatorname{var}(S_n''') = \mathrm{E}(S_n''')^2 &= \left| \sum_{j,l=1}^{n-p(k+m)} \mathrm{E}\left(X_{p(k+m)+j} X_{p(k+m)+l}\right) \right| \\ &\le (n - p(k+m))^2 M \le (k+m)^2 M \end{aligned}$$

$$S_n''' = O_p(\sqrt{\operatorname{var}(S_n''')}) = O_p((n - p(k+m))) = O_p(k+m) = O_p(n^{1/3}) \tag{21}$$

The Detailed Proof **

Similarly,

$$\mathrm{E}Z_j^2 = \mathrm{E}\left(\sum_{j=1}^{m} X_{B_{j-1}+k+j}\right)^2 \le m^2 M$$

And,

$$\mathrm{var}(S_n'') \le pm^2 M, \quad S_n'' = O_p(p^{1/2}m) = O_p(n^{1/3}m)$$

As $\sigma_n/(mn^{1/3}) \to \infty$, we have:

$$\frac{S_n''}{\sigma_n} = \frac{S_n''}{mn^{1/3}} \times \frac{mn^{1/3}}{\sigma_n} = o_p(1)$$

Besides, since $k = O(n^{1/3})$, from (21), $S_n'''/\sigma_n = o_p(1)$. Now,

$$\frac{S_n}{\sigma_n} = \frac{S_n'}{\sigma_n} + \frac{S_n''}{\sigma_n} + \frac{S_n'''}{\sigma_n} = \frac{\sigma_n'}{\sigma_n}\frac{S_n'}{\sigma_n'} + o_p(1) \tag{22}$$

It remains to prove $\sigma_n'^2/\sigma_n^2 \to 1$ and $S_n'/\sigma_n' \xrightarrow{d.} N(0,1)$.

The Detailed Proof **

$\sigma_n'^2/\sigma_n^2 \to 1$:

$$\mathrm{E}S_n^2 = \mathrm{E}(S_n')^2 + \mathrm{E}(S_n'')^2 + \mathrm{E}(S_n''')^2 + 2\mathrm{E}\left(S_n'S_n''\right) + 2\mathrm{E}\left(S_n''S_n'''\right) + 2\mathrm{E}\left(S_n'''S_n'\right)$$

Then,

$$\begin{aligned} \left|\mathrm{E}S_n^2 - \mathrm{E}(S_n')^2\right| &\le \left|\mathrm{E}(S_n'')^2 + \mathrm{E}(S_n''')^2 + 2\mathrm{E}\left(S_n'S_n''\right) + 2\mathrm{E}\left(S_n''S_n'''\right) + 2\mathrm{E}\left(S_n'''S_n'\right)\right| \\ &\le pm^2M^2 + (k+m)^2M^2 + 4p(mM)^2 + 2m(k+m)M^2 \end{aligned}$$

Here we perceive that:

$$\mathrm{E}\left(S_n'S_n''\right) = \sum_{j,l=1}^{p} \mathrm{cov}(Y_j, Z_l) \xlongequal{\text{indep.}} \sum_{j=1}^{p} \left[\mathrm{cov}(Y_j, Z_j) + \mathrm{cov}(Y_j, Z_{j-1})\right] \le 2p(mM)^2$$

and, similarly,

$$\mathrm{E}\left(S_n''S_n'''\right) \le m(k+m)M^2, \quad \mathrm{E}\left(S_n'''S_n'\right) = \mathrm{cov}\left(S_n'', S_n'''\right) = 0$$

Therefore,

$$\left|1-\frac{\sigma_n'^2}{\sigma_n^2}\right| \le \frac{2pm^2M^2+2(k+m)^2M^2}{\sigma_n^2} = O\left(\frac{m^2n^{2/3}}{\sigma_n^2}\right) \to 0$$

Hence, $\frac{\sigma_n'^2}{\sigma_n^2} \to 1$.

$S_n'/\sigma_n' \xrightarrow{d.} N(0,1)$:

We use truncation method. As $k=[n^{1/3}]$, let $Y_{nj}=Y_j$, then $\{Y_{nj}, 1 \le j \le p\}$ is double array, $|Y_{nj}| \le km = O(n^{1/3}m) = o(\sigma_n')$, and

$$\frac{1}{\sigma_n'^2}\sum_{j=1}^{p} \mathrm{E}\left[Y_{nj}^2 \mathbb{I}(|Y_{nj}| \ge \eta\sigma_n')\right] \le \frac{k^2m^2}{\sigma_n'^2}\sum_{j=1}^{p}\mathrm{P}(|Y_{nj}| > \eta\sigma_n')$$

$$\le \frac{k^2m^2}{\sigma_n'^2}\frac{\sum_{j=1}^{p}\mathrm{var}(Y_{nj})}{\eta^2\sigma_n'^2} \le \frac{k^2m^2}{\eta^2\sigma_n'^2} \longrightarrow 0$$

as $(km)/\sigma_n' \to 0$, which implies the LC holds, then $S_n'/\sigma_n' \xrightarrow{d.} N(0,1)$ from Theorem 4.10.

Cramér-Wold device

The device allows the issue of convergence of multivariate distribution to be reduced to that of univariate ones.

Theorem 4.14 (see Serfling (1988, p18))

A sequence of random vectors $\boldsymbol{X}_n$ in $\mathbb{R}^d$ converges in distribution to the random vector $\boldsymbol{X}$ if and only if for any linear combination of the component of $\boldsymbol{X}_n$ converges in distribution to the same linear combination of the component of $\boldsymbol{X}$ as $n \to \infty$, i.e.,

$$X_n \xrightarrow{d} X \Leftrightarrow a^T X_n \xrightarrow{d} a^T X, \forall a \in \mathbb{R}^d.$$

Levy's theorem implies that weak convergence of vectors is equivalent to weak convergence of linear combinations.

Proof of Cramér-Wold

"$\Longleftarrow$" Let $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nd})^T$, $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ have a characteristic function ϕ_n and ϕ respectively. As for all $\boldsymbol{c} = (c_1, c_2, \ldots, c_d)^T$

$$c_1 X_{n1} + \cdots + c_k X_{nd} \xrightarrow{d} c_1 X_1 + \cdots + c_d X_d. \tag{23}$$

The characteristic function of $c_1 X_{n1} + \cdots + c_d X_{nd}$ is

$$\phi_n(tc_1, \cdots, tc_d) = E(e^{it(c_1 X_{n1} + \cdots + c_d X_{nd})}).$$

The characteristic function of $\lambda_1 X_1 + \cdots + \lambda_d X_d$ is $\phi(tc_1, \cdots, tc_d)$. Choose $t = 1$, then from (23)

$$\lim_{n \to \infty} \phi_n(c_1, \cdots, c_d) = \phi(c_1, \cdots, c_d)$$

implying $\boldsymbol{X_n} \xrightarrow{d} \boldsymbol{X}$.
The "$\Longrightarrow$" is obvious by mapping.

Multivariate CLT

The Cramér-Wold device provide another approach to show Multivariate CLT.

Theorem 4.15

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. random vectors with mean $\boldsymbol{\mu}$ and finite covariance matrix, Σ. Let $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i/n$, then

$$\sqrt{n}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{d} N_d(0, \Sigma).$$

by letting $\boldsymbol{Y}_n := \sqrt{n}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right)$, so

$$\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{Y} \quad \text{iff} \quad \boldsymbol{t}^T\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{t}^T\boldsymbol{Y} \quad \text{for all} \quad \boldsymbol{t} \in \mathbb{R}^d.$$

Let $Z_1, \ldots, Z_n \in \mathbb{R}^d$, where d is the dimension of Z_i, is the equally sampled time series, i.e. daily, weekly, or yearly data.

- Strictly Stationary: for any integers l and m,

$$(Z_{i_1}, \ldots, Z_{i_m})^\top \quad \text{and} \quad (Z_{i_1+l}, \ldots, Z_{i_m+l})^\top$$

  have the same distribution (strong shift invariance).

- Weak Stationary (or Second Order Stationary):

$$\mathrm{E}(Z_i) = \mathrm{E}(Z_{i+l}), \quad \mathrm{var}(Z_i) = \mathrm{var}(Z_{i+l}),$$
$$\mathrm{cov}(Z_i, Z_j) = \mathrm{cov}(Z_{i+l}, Z_{j+l})$$

  (weak shift invariance).

Ways to make a time series stationary: difference, square root, etc.

Definition 5.1 (ARMA models)

The sequence $\{Z_i\}_{i\in\mathbb{Z}}$ is said to be an $ARMA(p,q)$ if $\{Z_i\}$ is weakly stationary and for any t,

$$Z_t - \theta_1 Z_{t-1} - \cdots - \theta_p Z_{t-p} = \epsilon_t - \eta_1 \epsilon_{t-1} - \cdots - \eta_q \epsilon_{t-q}$$

where $\{\epsilon_t\}$ is an independent white noise process, defined as $WN(0,\sigma^2)$.

ARMA models

Let θ and η be p-th and q-th degree polynomials defined as:

$$\theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p$$
$$\eta(z) = 1 - \eta_1 z - \cdots - \eta_q z^q$$

Let B be the backward shift operator such that $B^j Z_t = Z_{t-j}$.

Definition 5.2

An $ARMA(p,q)$ process $\{Z_t\}$ is said to be causal if there exists $\{\psi_j\}_{j=0}^{\infty}$ such that

$$\sum_{j=0}^{\infty} |\psi_j| < \infty, \quad Z_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

ARMA models

Theorem 5.3

Let $\{Z_t\}$ be an ARMA(p,q) : $\theta(B)Z_t = \eta(B)\epsilon_t$. If $\theta(z)$ and $\eta(z)$ have no common zero roots, then $\{Z_t\}$ is causal iff

$$\theta(z) \neq 0, \quad \forall z \in \mathbb{C}, |z| \leq 1$$

and the coefficients $\{\psi_j\}$ are determined by $\psi(z) = \eta(z)/\theta(z)$.

Linear Process:

$$Z_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \quad \epsilon_t \text{ i.i.d. } F(0, \sigma^2)$$

ARMA process under certain conditions are special type of linear process!

ARCH(p): Auto-Regression Conditionally Heterogeneity

Model:

$$Z_t = m(\vec{Z}_{t,p}) + \sigma(\vec{Z}_{t,p})\epsilon_t, \quad \vec{Z}_{t,p} = (Z_{t-1}, \cdots, Z_{t-p})^\top$$

it generalizes $AR(p)$:

$$Z_t = \theta_0 + \theta_1 Z_{t-1} + \cdots + \theta_p Z_{t-p} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2)$$

in two aspects:

(i) From linear to non-linear conditional mean function.

(ii) From constant conditional variance to a function.

ARCH models (or non-linear time series models) in general are not necessarily stationary, but some conditions can guarantee they are "asymptotic stationary" (stationary after pre-burning the models for a period of "long" time). See Gouriéroux[1] for reference.

[1] Christian Gouriéroux. *ARCH models and financial applications.* Springer Science & Business Media, 2012.

Mixing Coefficients: measures of dependence

Let $Z_1, \cdots, Z_t, \cdots$ be a strictly stationary process and $\mathcal{F}_l^m$ be the sigma field generated by $\{Z_i\}_{i=l}^m$, where $m \geq l$ are positive integers.

(i) α-mixing (or strong mixing) coefficient:

$$\alpha(k) = \sup_{B \in \mathcal{F}_{-\infty}^t,\, C \in \mathcal{F}_{t+k}^\infty} |\mathrm{P}(B \cap C) - \mathrm{P}(B)\mathrm{P}(C)|, \quad k \geq 1$$

*(ii) β-mixing or absolute regularity coefficient:

$$\beta(k) = \mathrm{E} \sup_{c \in \mathcal{F}_{t+k}^\infty} |\mathrm{P}(C) - \mathrm{P}(C|\mathcal{F}_{-\infty}^t)|$$

(iii) ϕ-mixing:

$$\phi(k) = \sup_{B \in \mathcal{F}_{-\infty}^t,\, C \in \mathcal{F}_{t+k}^\infty} |\mathrm{P}(C) - \mathrm{P}(C|B)|$$

Mixing Coefficients: measures of dependence

(iv) ρ-mixing:

$$\rho(k) = \sup_{X \in L^2(\mathcal{F}_{-\infty}^{t}),\, Y \in L^2(\mathcal{F}_{t+k}^{\infty})} |\mathrm{corr}(X, Y)|$$

$$= \sup_{X \in L^2(\mathcal{F}_{-\infty}^{t}),\, Y \in L^2(\mathcal{F}_{t+k}^{\infty})} \left| \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} \right|$$

where $L^2(\mathcal{F})$ is the set of all r.vs defined on $\mathcal{F}$ which have finite second moments, i.e. $\forall X \in L^2(\mathcal{F})$, $\mathrm{E}X^2 < \infty$.

Relationship between these mixing coefficients

$$2\alpha(k) \le \beta(k) \le \phi(k), \quad 4\alpha(k) \le \rho(k) \le 2\phi^{1/2}(k)$$

The process $\{Z_t\}_{t\in\mathbb{Z}}$ is said to be α-mixing if $\lim_{k\to\infty} \alpha(k) = 0$. Similarly, ϕ-mixing process can be defined as $\lim_{k\to\infty} \phi(k) = 0$, etc.

Remark 17

Mixings are different descriptions of dependence between events in two sigma fields $(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+k}^{\infty})$. When the time gap between them goes to infinity, i.e. $k \to \infty$, mixings means asymptotic independence.

Relationship between these mixing coefficients

The inequality between different mixing coefficients means that,

$$\begin{array}{ccc} \phi\text{-mixing} & \Longrightarrow & \beta\text{-mixing} \\ \Downarrow & & \Downarrow \\ \rho\text{-mixing} & \Longrightarrow & \alpha\text{-mixing} \end{array}$$

So α-mixing is the weakest mixing coefficient, but ironically has been called strong mixing.

When a linear process is mixing?

For a linear causal process $Z_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}$, $\{\epsilon_t\}$ i.i.d. $F(0, \sigma^2)$. Gorodetskii[2] shows that $\{Z_k\}$ is α-mixing under certain conditions, and establish the rate of $\alpha(k)$.

Pham T. D. and Tran L. T.[3] show that if $\psi_j = O(r^j)$ for $0 < r < 1$ when $j \to \infty$, then the process is geometric α-mixing, i.e. there exists $C, \rho \in [0, 1)$ such that $\alpha(k) \leq C\rho^k$.

[2] VV Gorodetskii. "On the strong mixing properties for linear processes". In: *Theory of Probability and its Applications* 22 (1977), pp. 441–413.

[3] Tuan D Pham and Lanh T Tran. "Some mixing properties of time series models". In: *Stochastic processes and their applications* 19.2 (1985), pp. 297–303.

Lemma 5.4 (Billingley's Inequality)

If $\{Z_i\}$ *is* α*-mixing (NOT necessarily stationary),* $X \in \mathcal{F}_{-\infty}^{t}$ *and* $Y \in \mathcal{F}_{t+k}^{\infty}$, $|X| \leq C_1$, $|Y| \leq C_2$, *then,*

$$|\operatorname{cov}(X, Y)| \leq 4C_1C_2\alpha(k)$$

Proof

Since,

$$\begin{aligned} |\operatorname{cov}(X,Y)| = |\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)| &= \left|\mathrm{E}\left[X\left\{\mathrm{E}(Y|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}Y\right\}\right]\right| \\ &\le C_1 \mathrm{E}\left|\mathrm{E}(Y|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}Y\right| \\ &= C_1 \mathrm{E}\left[\xi\left(\mathrm{E}(Y|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}Y\right)\right] \end{aligned}$$

where $\xi = \operatorname{sgn}(\mathrm{E}(Y|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}Y) \in \mathcal{F}_{-\infty}^{t}$. Thus,

$$|\operatorname{cov}(X,Y)| \le C_1|\mathrm{E}(\xi Y) - \mathrm{E}\xi\mathrm{E}Y| = C_1|\operatorname{cov}(\xi,Y)| \tag{24}$$

By using the same approach, we have,

$$|\operatorname{cov}(\xi,Y)| \le C_2|\mathrm{E}(\xi\eta) - \mathrm{E}\xi\mathrm{E}\eta| \tag{25}$$

where $\eta = \operatorname{sgn}(\mathrm{E}(\xi|\mathcal{F}_{t+k}^{\infty}) - \mathrm{E}\xi) \in \mathcal{F}_{t+k}^{\infty}$.

Proof

Let:

$$A = \{\xi = 1\}, \quad B = \{\eta = 1\}$$
$$A^C = \{\xi = -1\}, \quad B^C = \{\eta = -1\}$$

Clearly, $A, A^C \in \mathcal{F}_{-\infty}^t$ and $B, B^C \in \mathcal{F}_{t+k}^{\infty}$. Then,

$$|\mathrm{E}(\xi\eta) - \mathrm{E}\xi\mathrm{E}\eta| = \big|\mathrm{P}(AB) + \mathrm{P}(A^C B^C) - \mathrm{P}(A^C B) - \mathrm{P}(AB^C) \\ -(\mathrm{P}(A) - \mathrm{P}(A^C))(\mathrm{P}(B) - \mathrm{P}(B^C))\big| \leq 4\alpha(k) \tag{26}$$

And the lemma is proved by combining (24), (25), and (26).

Lemma 5.5

If $\{Z_i\}$ *is* α*-mixing (NOT necessarily stationary),* $X \in \mathcal{F}_{-\infty}^{t}$ *and* $Y \in \mathcal{F}_{t+k}^{\infty}$, $\mathrm{E}|X|^p < \infty$ *for some* $p > 1$ *and* $|Y| \leq C$*, then*

$$|\operatorname{cov}(X, Y)| \leq 6C\|X\|_p \alpha^{1/q}(k)$$

where $\dfrac{1}{p} + \dfrac{1}{q} = 1$ *and* $\|X\|_p = (\mathrm{E}|X|^p)^{1/p}$.

Proof

For some $M > 0$, let $X_M = X\mathbb{I}(|X| \leq M)$, $X_M' = X - X_M = X\mathbb{I}(|X| > M)$. Then $X = X_M + X_M'$, and

$$|\operatorname{cov}(X, Y)| = |\operatorname{cov}(X_M, Y) + \operatorname{cov}(X_M', Y)| \leq |\operatorname{cov}(X_M, Y)| + |\operatorname{cov}(X_M', Y)|$$

From Lemma 5.4,

$$|\operatorname{cov}(X_M, Y)| \leq 4CM\alpha(k) \tag{27}$$

On the other hand, note that,

$$\mathrm{E}|X_M'| = \int_{|X|>M} |x| dF(x) \leq \int_{|X|>M} |x| \left(\frac{|x|}{M}\right)^{p-1} dF(x)$$

i.e. $\mathrm{E}|X_M'| \leq M^{-p+1}\mathrm{E}|X|^p$.

Proof

Thus,

$$
\begin{aligned}
|\operatorname{cov}(X'_M, Y)| = |\mathrm{E}(X'_M Y) - \mathrm{E}X'_M \mathrm{E}Y| &\leq \mathrm{E}|X'_M Y| + C\mathrm{E}|X'_M| \\
&\leq 2C\mathrm{E}|X'_M| \leq 2CM^{-p+1}\mathrm{E}|X|^p
\end{aligned}
\tag{28}
$$

Now, choose

$$
M = \|X\|_p \{\alpha(k)\}^{-1/p}
$$

and from (27) and (28), we prove the lemma.

Lemma 5.6 (Rio's Inequality)

Let X and Y be two integrable [a] *real-valued r.vs and let $Q_X(u) = \inf\{t : \mathrm{P}(|X| > t) \leq u\}$ be the quantile function of $|X|$. Then if $Q_X Q_Y$ us integrable over (0, 1). We have:*

$$|\operatorname{cov}(X, Y)| \leq 2 \int_0^{2\alpha} Q_X(u) Q_Y(u) du$$

where $\alpha = \alpha(\sigma(X), \sigma(Y))$ is the α-mixing coefficient between sigma fields $\sigma(X)$ and $\sigma(Y)$.

[a] $\lim_{c \to \infty} \mathrm{E}\{|X| \mathbb{I}(|X| > c)\} = 0$

Proof

Denote $X^+ = 0 \vee X$ and $X^- = 0 \vee (-X)$, then

$$\begin{aligned} \text{cov}(X, Y) = & \text{cov}(X^+, Y^+) + \text{cov}(X^-, Y^-) \\ & - \text{cov}(X^-, Y^+) - \text{cov}(X^+, Y^-) \end{aligned} \tag{29}$$

Note that:

$$\text{cov}\left(X^+, Y^+\right) = \int_{\mathbb{R}_+^2} [\mathrm{P}(X > u, Y > v) - \mathrm{P}(X > u)\mathrm{P}(Y > v)]dudv$$

Recall the definition of α, we have

$$\text{cov}\left(X^+, Y^+\right) \leq \int_{\mathbb{R}_+^2} \inf(\alpha, P(X > u), P(Y > v))dudv \tag{30}$$

Proof

Now apply (29), (30) ,and the elementary inequality

$$
\begin{aligned}
&(\alpha \wedge a \wedge c) + (\alpha \wedge a \wedge d) + (\alpha \wedge b \wedge c) + (\alpha \wedge b \wedge d) \\
&\le 2[(2\alpha) \wedge (a+b) \wedge (c+d)]
\end{aligned}
$$

to $a = \mathrm{P}(X > u), b = \mathrm{P}(-X > u), c = \mathrm{P}(Y > v), d = \mathrm{P}(-Y > v)$, we get

$$
|\operatorname{Cov}(X, Y)| \le 2 \int_{\mathbb{R}_+^2} \inf(2\alpha, P(|X| > u), P(|Y| > v) dudv =: I
$$

Then we only need to show,

$$
I = 2 \int_0^{2\alpha} Q_X(u) Q_Y(u) du \tag{31}
$$

Proof

Note the form of I, define a bivariate r.v. (Z, T),

$$(Z, T) = (0, 0)\mathbb{I}(U \geq 2\alpha) + (Q_X(U), Q_Y(U))\,\mathbb{I}(U < 2\alpha)$$

where U is a uniform distributed r.v. over $[0, 1]$. So for any $u, v > 0$,

$$\{Z > u, T > v\} = \{U < 2\alpha, U < \mathrm{P}(|X| > u), U < \mathrm{P}(|Y| > v)\}$$

by calculating the true integral value,

$$\int_0^{2\alpha} Q_X(u)Q_Y(u)du = \mathrm{E}(ZT) = \int_{\mathbb{R}_+^2} \mathrm{P}(Z > u, T > v)dudv$$
$$= \int_{\mathbb{R}_+^2} \inf(2\alpha, \mathrm{P}(|X| > u), \mathrm{P}(|Y| > v))dudv$$

which entails (31) and the proof is thus complete.

Lemma 5.7 (Davydov's Inequality)

Let X and Y be two real r.vs such that $X \in L^q(\mathcal{F}_{-\infty}^{t})$, $Y \in L^r(\mathcal{F}_{t+k}^{\infty})$ where $q > 1, r > 1$ and $\frac{1}{q} + \frac{1}{r} = 1 - \frac{1}{p}$, then

$$|\operatorname{cov}(X, Y)| \leq 2p(2\alpha(k))^{1/p}\|X\|_q\|Y\|_r$$

Proof

(i) Suppose first that q and r are finite. Then Markov's inequality yields,

$$\mathrm{P}\left(|X| > \frac{\|X\|_q}{u^{1/q}}\right) \leq \frac{\mathrm{E}|X|^q}{\left(\|X\|_q / u^{1/q}\right)^q} = u \quad , \quad 0 < u \leq 1$$

which implies,

$$Q_X(u) \leq \frac{\|X\|_q}{u^{1/q}} \quad , \quad 0 < u \leq 1$$

similarly, $Q_Y(u) \leq \|Y\|_r / u^{1/r}$. Using Rio's inequality,

$$\begin{aligned} |\operatorname{cov}(X, Y)| &\leq 2 \int_0^{2\alpha(k)} \frac{\|X\|_q}{u^{1/q}} \frac{\|Y\|_r}{u^{1/r}} du \\ &= 2\|X\|_q \|Y\|_r \int_0^{2\alpha(k)} u^{\frac{1}{p}-1} du \qquad \left(\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1\right) \\ &= 2p(2\alpha(k))^{1/p} \|X\|_q \|Y\|_r \end{aligned}$$

Proof

(ii) If $r = +\infty$, q is finite, then $\frac{1}{q} + \frac{1}{p} = 1$. Note that $Q_Y(u) \le Q_Y(0) = \|Y\|_\infty := \sup|Y|$. This is because,

$$\begin{aligned} &\{t \,|\, \mathrm{P}(|Y| > t) = 0\} \subset \{t \,|\, \mathrm{P}(|Y| > t) \le u\}, \quad \forall u \in [0,1] \\ \Longrightarrow\ &\inf\{t \,|\, \mathrm{P}(|Y| > t) = 0\} \ge \inf\{t \,|\, \mathrm{P}(|Y| > t) \le u\} \\ \Longrightarrow\ &Q_Y(u) \le Q_Y(0) \end{aligned}$$

Also, clearly $\inf\{t \,|\, \mathrm{P}(|Y| > t) = 0\} = \|Y\|_\infty$. Use Rio's inequality again,

$$\begin{aligned} |\operatorname{cov}(X,Y)| &\le 2\int_0^{2\alpha(k)} \frac{\|X\|_q}{u^{1/q}} \|Y\|_\infty du \\ &= 2\|X\|_q \|Y\|_\infty \int_0^{2\alpha(k)} u^{\frac{1}{p}-1} du \\ &= 2p(2\alpha(k))^{1/p} \|X\|_q \|Y\|_\infty \end{aligned}$$

which is similar to, but not exactly the same as Lemma 5.5.

Proof

(iii) If $r = +\infty$ and $q = +\infty$, then $p = 1$. Similarly, use Rio's inequality again,

$$|\operatorname{cov}(X, Y)| \leq 2 \times 2 \int_0^{2\alpha(k)} \|X\|_\infty \|Y\|_\infty du = 4\|X\|_\infty \|Y\|_\infty \alpha(k)$$

which is the same as Lemma 5.4. Now, we have completed all the proof of this lemma.

Weakly Dependent Stationary Process

Suppose $\{X_i\}$ be a weakly stationary process with finite second moments. Let $\gamma(j) = \text{cov}(X_i, X_{i+j})$.

Definition 5.8

The process is said to be weakly dependent if $\sum_{k=0}^{\infty} |\gamma(k)| < \infty$, or it would be said to be a long memory process.

Weakly Dependent Stationary Process

Let $\alpha(k)$ be the strong-mixing coefficient defined on the sigma fields generated by $\{X_i\}_{i\in\mathbb{Z}}$.

From Lemma 5.7 ($r = q$), if $\mathrm{E}|X_i|^q < \infty$ for $q > 2$, and $\sum_{k=0}^{\infty} \alpha^{1/p}(k) < \infty$ for $p = \frac{q}{q-2}$, then

$$\sum_{k=0}^{\infty} |\gamma(k)| \leq 2p\|X\|_q^2 \sum_{k=0}^{\infty} \alpha^{1/p}(k) < \infty$$

Thus, this process is weakly dependent (short-memory). In particular, if $\alpha(k) \leq C\rho^k$, i.e. geometric strong mixing, then

$$\sum_{k=0}^{\infty} \alpha^{1/p}(k) \leq C \sum_{k=0}^{\infty} \rho^{k/p} = \frac{C}{1-\rho^{1/p}} < \infty.$$

Remark 18

In general, to ensure $\sum_{k=0}^{\infty} \alpha^{1/p}(k) < \infty$*, we require* $\alpha^{1/p}(k) \sim k^{-(1+\eta)}$*, i.e.* $\alpha(k) \sim k^{-p(1+\eta)}$ *for* $\eta > 0$ *when* k *is sufficiently large, which implies* $\alpha(k) \to 0$ *as* $k \to \infty$ *sufficiently fast.*

Remark 19

Note that geometric strong mixing (GSM) means $\alpha(k) \leq C\rho^k = Ce^{-\beta k}$*, which entails* $\alpha(k) \to 0$ *at exponential rate.*

Spectral Density

Define:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ihv} dF(v) = \int_{-\pi}^{\pi} e^{ihv} f(v) dv$$

Then by Laplace transformation:

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \gamma(n)$$

Theorem 5.9

If $\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$, then $\{X_n\}$ has spectral density f, and we have $\sum_{h=\infty}^{\infty} \gamma(h) = 2\pi f(0)$.

Lemma 5.10

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a zero-mean real-valued weakly stationary process such that for some $r > 2$,

$$\sup_{t\in\mathbb{Z}} \mathrm{E}\,|X_t|^r < \infty, \quad \sum_{k\geq 1} \alpha(k)^{1-\frac{2}{r}} < +\infty$$

then the series $\sum_{k\in\mathbb{Z}} \gamma(k)$ is absolutely convergent, has a nonnegative sum σ^2 and,

$$n\,\mathrm{var}(S_n/n) \longrightarrow \sigma^2 \tag{32}$$

where $\gamma(k) = \mathrm{cov}(X_0, X_k)$.

Proof

First we study the series $\sum_{k\in\mathbb{Z}} \gamma(k)$, by using Lemma 5.7 with $q = r$ and $\frac{1}{p} = 1 - \frac{2}{r}$, we get

$$|\gamma(k)| \leq \frac{2r}{r-2} \left(\mathrm{E}\,|X_0|^r\right)^{2/r} (2\alpha(k))^{1-2/r}$$

which proves the absolute convergence of the series since $\sum_{k\geq 1} \alpha(k)^{1-2/r} < +\infty$. Now clearly,

$$n\,\mathrm{var}\left(\frac{S_n}{n}\right) = n^{-1} \sum_{0\leq s,t\leq n-1} \mathrm{cov}\,(X_s, X_t) = \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right) \gamma(k)$$

due to $\{X_t\}$ being weakly stationary, thus,

$$\lim_{n\to\infty} n\,\mathrm{var}\left(\frac{S_n}{n}\right) = \sigma^2 \geq 0$$

and the theorem is thus established.

CLT for strongly mixing processes

Theorem 5.11

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a zero-mean real-valued strictly stationary process such that for some $r > 2$ and some $\beta > 0$,

$$\mathrm{E}|X_t|^r < \infty, \quad \alpha(k) \le ak^{-\beta}$$

where a is a postive constant and $\beta > r/(r-2)$. Then if $\sigma^2 = \sum_{k=-\infty}^{\infty} \gamma(k) > 0$, we have

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

The proof can be seen in Theorem 1.7 of D. Bosq's book[4].

[4] Denis Bosq. *Nonparametric statistics for stochastic processes: estimation and prediction*. Vol. 110. Springer Science & Business Media, 2012.

Suppose we have a sequence of estimators $\{T_n\}_{n\geq 1}$ on $\mathbb{R}^k$ for a parameter $\theta \in \mathbb{R}^k$.

- For the $\phi(\theta)$, the parameter of interest, considering convergence of $\phi(T_n)$ to $\phi(\theta)$, where $\phi: \ \mathbb{R}^k \to \mathbb{R}^m$.
- From Mapping, $T_n \xrightarrow{p} \theta \ \Rightarrow \ \phi(T_n) \xrightarrow{p} \phi(\theta)$, if ϕ was continuous at θ.
- Does $\sqrt{n}(\phi(T_n) - \phi(\theta))$ have an asymptotic distribution, if suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$?
- Is the convergence in distribution persevered under smooth transformation?

The Derivative of Vector-valued Functions

Recall that $\phi(\cdot)$ is differentiable at θ if there exists a linear map (matrix) $\phi'_\theta : \mathbb{R}^k \mapsto \mathbb{R}^m$ such that

$$\begin{aligned}\phi(\theta + h) - \phi(\theta) &:= \phi'(\theta)h + R(h) \\ &= \phi'(\theta)h + o(\|h\|), \quad h \to 0.\end{aligned}$$

Define $\phi'_\theta(h) := \phi'(\theta)h$.

Derivative map (Jacobian matrix)

The derivative map $h \mapsto \phi'_\theta(h)$ is matrix multiplication by the matrix

$$\phi'(\theta) \triangleq \phi'_\theta = \begin{pmatrix} \frac{\partial \phi_1}{\partial \theta_1}(\theta) & \cdots & \frac{\partial \phi_1}{\partial \theta_k}(\theta) \\ \vdots & & \vdots \\ \frac{\partial \phi_m}{\partial \theta_1}(\theta) & \cdots & \frac{\partial \phi_m}{\partial \theta_k}(\theta) \end{pmatrix} = \left(\frac{\partial \phi_i(\theta)}{\partial \theta_j}\right)_{m \times k}.$$

Remark 20

If $m = 1, k > 1$, the derivative map is called the gradient *of the function.*

Main Result of Delta Method

Theorem 6.1 (First Order Delta Method)

If ϕ differentiable at θ, $\phi'(\theta) \neq 0$, and $r_n(T_n - \theta) \xrightarrow{d} T$ for a deterministic sequence of $\{r_n\}$, satisfied $r_n \to \infty$, then:

(i) $r_n\left(\phi(T_n) - \phi(\theta)\right) - \phi'(\theta)\left(r_n(T_n - \theta)\right) \xrightarrow{p} 0$;

(ii) $r_n\left(\phi(T_n) - \phi(\theta)\right) \xrightarrow{d} \phi'(\theta)\,T$.

The proof: (i)

- Since $r_n(T_n - \theta) \xrightarrow{d} T$, then $T_n - \theta \xrightarrow{p} 0$ by stochastic boundedness.
- By the differentiability of ϕ at θ, then

$$\phi(\theta + h) - \phi(\theta) - \phi'(\theta)h := R(h) = o\left(\|h\|\right).$$

- Replace the h by $T_n - \theta$, multiply r_n to get

$$r_n[\phi\left(T_n\right) - \phi(\theta) - \phi'_\theta\left(T_n - \theta\right)] = o_P\left(r_n \|T_n - \theta\|\right) = o_P(1).$$

  by Lemma of Stochastic Plug-in (since $T_n - \theta \xrightarrow{p} 0$).

The proof: (ii)

- Matrix multiplication is continuous, so $\phi'_\theta\left(r_n\left(T_n-\theta\right)\right) \xrightarrow{d} \phi'_\theta(T)$ by the continuous-mapping theorem.
- Apply Slutsky's lemma to conclude that

$$r_n\left(\phi(T_n)-\phi(\theta)\right) \xrightarrow{d} \phi'(\theta)T$$

which has the same weak limit as $\phi'_\theta\left(r_n\left(T_n-\theta\right)\right)$.

Example: Normal delta method

Let T_n be a sequence of statistics such that

$$\sqrt{n}\left(T_n-\theta\right) \xrightarrow{d} N\left(0,\sigma^2(\theta)\right)$$

Let $g:\mathbb{R}\to\mathbb{R}$ be once differentiable at θ with $g'(\theta)\neq 0$. Then

$$\sqrt{n}\left[g\left(T_n\right)-g(\theta)\right] \xrightarrow{d} N(0,[g'(\theta)]^2\sigma^2(\theta))$$

High Order Delta Method: what if $\phi'(\theta) = 0$?

First Order Delta Method is largely based on Taylor's expansions with $\phi`(\theta) \neq 0$. If $\phi'(\theta) = 0$ but $\phi''(\theta) \neq 0$, we have

$$\phi(T_n) = \phi(\theta) + \frac{1}{2}\phi''(\theta)(T_n - \theta)^2 + \cdots .$$

Then

$$n\left(\phi(T_n) - \phi(\theta)\right) = \frac{1}{2}\phi''(\theta)\left[\sqrt{n}(T_n - \theta)\right]^2 \xrightarrow{d} \cdots$$

If $\sqrt{n}\bar{X}_n \xrightarrow{d} N(0,1)$, then $n\phi(\bar{X}_n) \xrightarrow{d} \frac{1}{2}\phi''(\theta)\chi_1^2$.

Theorem 6.2

Suppose ϕ be a univ. m times differentiable at θ with $\phi^{(m)}(\theta) \neq 0$, $\phi^{(j)}(\theta) = 0,\ j < m$, then:

$$\frac{r_n^m\left(\phi(T_n) - \phi(\theta)\right)}{\frac{1}{m!}\phi^{(m)}(\theta)} \xrightarrow{d} T^m.$$

A multivariable version of this theorem is available in Serfling P124.

Examples: 2ed Delta Method

1. Suppose $X_1, \ldots, X_n$ are id with mean μ and known variance σ^2, and we want to test $H_0 : \mu = 0$. Under the null hypothesis $H_0 : \mu = 0$, the following statistic

$$T(\boldsymbol{X}) := n\overline{X}_n^2/\sigma^2 \xrightarrow{d} [N(0,1)]^2 = \chi_1^2.$$

2. Suppose that $\sqrt{n}\overline{X}_n$ converges in law to a standard normal distribution. Now consider the limiting behavior of $\cos(\overline{X}_n)$.
   - Because the derivative of $\cos(x)$ is zero at $x = 0$, we still use the proof of First Order Delta Method. It yields that
   $$\sqrt{n}(\cos(\overline{X}_n) - 1) \xrightarrow{d} \delta_0$$
   which implies that $\sqrt{n}(\cos(\overline{X}_n) - 1) \xrightarrow{P} \delta_0$.
   - Thus, it should be concluded that $\sqrt{n}$ is not the right norming rate for the random sequence $\cos(\overline{X}_n) - 1$. 2ed Order Delta Method
   $$\cos \overline{X} - \cos 0 = (\overline{X} - 0)0 + \frac{1}{2}(\overline{X} - 0)^2(\cos x)''_{|x=0} + \cdots$$
   implies
   $$-2n(\cos \overline{X} - 1) \xrightarrow{d} \chi_1^2.$$

Example: Variance

Let $X_1, \ldots, X_n$ i.i.d. F with finite 4-th moments. Let $\alpha_i = \mathrm{E}X_1^i$ for $i = 1, 2, 3, 4$ and $m_{nl} = n^{-1}\sum_{j=1}^n X_j^l$. Then,

$$S_n = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2 = n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \phi(m_{n1}, m_{n2})$$

where $\phi(x_1, x_2) = x_2 - x_1^2$. From MCLT,

$$\sqrt{n}\left[\begin{pmatrix} m_{n1} \\ m_{n2} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}\right] \xrightarrow{d.} N\left(0, \mathrm{var}\begin{pmatrix} X_1 \\ X_1^2 \end{pmatrix}\right)$$

and $\phi'(\alpha_1, \alpha_2) = (-2\alpha_1, 1)$.Hence,

$$\begin{aligned}\sqrt{n}(S_n - \sigma^2) &= \sqrt{n}(\phi(m_{n1}, m_{n2}) - \phi(\alpha_1, \alpha_2)) \\ &\xrightarrow{d.} (-2\alpha_1, 1)N\left(0, \mathrm{var}\begin{pmatrix} X_1 \\ X_1^2 \end{pmatrix}\right) = N(0, c_4 - c_2^2)\end{aligned}$$

where we use the fact that:

$$(-2\alpha_1, 1)\,\mathrm{var}\begin{pmatrix} X_1 \\ X_1^2 \end{pmatrix}\begin{pmatrix} -2\alpha_1 \\ 1 \end{pmatrix} = \mathrm{E}(X_1 - \alpha_1)^4 - \left(\mathrm{E}(X_1 - \alpha_1)^2\right)^2 = c_4 - c_2^2$$

Example: Standard Deviation

Considering the unbiased estimator:

$$S_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}S_n$$

So,

$$\begin{aligned}
\sqrt{n}(S_{n-1} - \sigma^2) &= \sqrt{n}(\frac{n}{n-1}S_n - \sigma^2) \\
&= \sqrt{n}\left(S_n - \sigma^2 + \left(\frac{n}{n-1} - 1\right)S_n\right) \\
&= \sqrt{n}(S_n - \sigma^2) + \sqrt{n}\left(\frac{n}{n-1} - 1\right)S_n \\
&\xrightarrow{d.} N(0, c_4 - c_2^2) \qquad \left(\sqrt{n}\left(\tfrac{n}{n-1} - 1\right)S_n = o_p(1)\right)
\end{aligned}$$

Furthermore, $S_n^{1/2} = \sqrt{S_n} = \phi(S_n)$, $\phi(x) = \sqrt{x}$, and $\phi'(x) = \frac{1}{2}x^{-1/2}$,

$$\sqrt{n}(S_n^{1/2} - \sigma) \xrightarrow{d.} N\left(0, \frac{c_4 - c_2^2}{4\sigma^2}\right)$$

More Examples

If X_n is $AN(\mu, \sigma_n^2)$ and $\sigma_n \to 0$. Then,

(i) X_n^2 is $AN(\mu^2, 4\mu^2\sigma_n^2)$ for $\mu \neq 0$.

(ii) $\dfrac{1}{X_n}$ is $AN\left(\mu^{-1}, \dfrac{\sigma_n^2}{\mu^4}\right)$ for $\mu \neq 0$.

(iii) e^{X_n} is $AN(e^{\mu}, e^{2\mu}\sigma_n^2)$ for any μ.

(iv) $\log|X_n|$ is $AN(\log|\mu|, \mu^{-2}\sigma_n^2)$ if $\mu \neq 0$; $\log|\sigma_n^{-1}X_n| \xrightarrow{d.} \log|N(0,1)|$ for $\mu = 0$.

(v) Suppose $X_1, \ldots, X_n$ i.i.d. F on $\mathbb{R}^p$ with (μ, Σ). Let

$$\theta = \mu^\top \mu, \qquad \hat{\theta} = \bar{X}^\top \bar{X} = \phi(\bar{X})$$

If $\mu \neq 0$, $\phi'(\mu) = 2\mu^\top$, $\phi''(\mu) = 2I_p$. As $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d.} N_p(0, \Sigma)$. So,

$$\sqrt{n}\left(\bar{X}^\top \bar{X} - \mu^\top \mu\right) \xrightarrow{d.} 2\mu^\top N_p(0, \Sigma) = N(0, 4\mu^\top \Sigma \mu)$$

Example: weighted χ_1^2 distribution

If $\mu = 0$, $\mu^\top \Sigma \mu = 0$, the above display $\xrightarrow{d.}$ 0 is not useful. In fact, as $\sqrt{n}\bar{X} \xrightarrow{d.} N_p(0, \Sigma)$,

$$n\bar{X}^\top \bar{X} \xrightarrow{d.} N_p^\top(0, \Sigma) N_p(0, \Sigma) \overset{d.}{=\!=} Z^\top \Sigma^{1/2} \Sigma^{1/2} Z = Z^\top \Sigma Z$$

where $Z \sim N(0, I_p)$. Suppose

$$\Sigma = U^\top \operatorname{diag}(\lambda_1, \ldots, \lambda_p) U, \quad \tilde{Z} = UZ \sim N_p(0, I_p)$$

Then,

$$Z^\top \Sigma Z \overset{d.}{=\!=} \sum_{i=1}^{p} \lambda_i \tilde{Z}_i^2 \overset{d.}{=\!=} \sum_{i=1}^{p} \lambda_i \chi_{1i}^2$$

where $\{\chi_{1i}^2\}_{i=1}^p$ i.i.d. χ_1^2.

So $n\bar{X}^\top \bar{X}$ converges to a weighted χ_1^2 distribution.

Thus $n\bar{X}^\top \bar{X} = O_p(1)$ if $\mu = 0$, and $\bar{X}^\top \bar{X} - \mu^\top \mu = O_p\left(\frac{1}{\sqrt{n}}\right)$ if $\mu \neq 0$.

Chi-Square Test for Variance

Suppose $X_1, \ldots, X_n$ i.i.d. F with $\mathrm{E}X_1^4 < \infty$.

$$H_0 : \sigma^2 \le 1 \quad \text{vs} \quad H_1 : \sigma^2 > 1$$

Denote $\Theta_0 = \{0 < \sigma^2 \le 1\}$ and $\Theta_1 = \{\sigma > 1\}$. If $F = N(\mu, \sigma^2)$, $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$.

Test statistic for $H_0 : \sigma^2 \le 1$ is nS_n by setting $\sigma^2 = 1$, and we reject H_0 if $nS_n > \chi^2_{n-1,\alpha}$. The size of this test is

$$\begin{aligned} \mathrm{P}_{\Theta_0}\left(nS_n > \chi^2_{n-1,\alpha} \,|\, \sigma^2 \in \Theta_0\right) &= \mathrm{P}_{\Theta_0}\left(\frac{nS_n}{\sigma^2} > \frac{1}{\sigma^2}\chi^2_{n-1,\alpha} \,\middle|\, \sigma^2 \le 1\right) \\ &\le \mathrm{P}(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha \end{aligned}$$

So the size $\le \alpha$ with the maximum size at $\sigma^2 = 1$ equals to α.

Chi-Square Test for Variance

If $F \neq$ Normal, the excessive kurtosis:

$$\kappa = \frac{\mathrm{E}(X-\mu)^4}{\sigma^4} - 3 \neq 0$$

From CLT and the fact that $\chi^2_{n-1} = \sum_{i=1}^{n-1} Z_i^2$ for $\{Z_i\}_{i=1}^{n-1}$ i.i.d. $N(0,1)$. Then,

$$\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{d.} N(0,1) \tag{33}$$

From the previous example, we know that:

$$\sqrt{n}\left(\frac{S_n}{\sigma^2} - 1\right) \xrightarrow{d.} N\left(0, \kappa - 1\right) \neq N(0,2)$$

And from (33),

$$\mathrm{P}\left(\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} \geq Z_\alpha\right) \longrightarrow \mathrm{P}(N(0,1) \geq Z_\alpha) = \alpha$$

Chi-Square Test for Variance

As $\mathrm{P}(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$, we have $\chi^2_{n-1,\alpha} \approx (n-1) + Z_\alpha\sqrt{2(n-1)}$. i.e.

$$\lim_{n\to\infty} \frac{\chi^2_{n-1,\alpha} - (n-1)}{\sqrt{2(n-1)}} = Z_\alpha$$

Consequently, the level of the Chi-Square Test is:

$$\begin{aligned}
\mathrm{P}_{\sigma^2=1}\left(\frac{nS_n}{\sigma^2} > \chi^2_{n-1,\alpha}\right) &= \mathrm{P}_{\sigma^2=1}\left(\sqrt{n}\left(\frac{S_n}{\sigma^2} - 1\right) > \frac{\chi^2_{n-1,\alpha} - n}{\sqrt{n}}\right) \\
&\approx \mathrm{P}_{\sigma^2=1}\left(\sqrt{n}\left(\frac{S_n}{\sigma^2} - 1\right) > \frac{(n-1) + Z_\alpha\sqrt{2(n-1)} - n}{\sqrt{n}}\right) \\
&\to \mathrm{P}(N(0,\kappa+2) > \sqrt{2}Z_\alpha) = 1 - \Phi\left(\frac{\sqrt{2}Z_\alpha}{\sqrt{\kappa+2}}\right)
\end{aligned}$$

For heavy-tail F, $\kappa > 0$, so that,

$$1 - \Phi\left(\frac{\sqrt{2}Z_\alpha}{\sqrt{\kappa+2}}\right) > 1 - \Phi(Z_\alpha) = \alpha$$

Chi-Square Test for Variance

Power of this test:

$$
\begin{aligned}
\mathrm{P}_{\sigma^2>1}\left(\frac{nS_n}{\sigma^2} > \chi^2_{n-1,\alpha}\right) &= \mathrm{P}_{\sigma^2>1}\left(\sqrt{n}\left(\frac{S_n}{\sigma^2}-1\right) > \frac{\sigma^{-2}\chi^2_{n-1,\alpha}-n}{\sqrt{n}}\right) \\
&\to 1-\mathrm{P}\left(N(0,\kappa+2) > \frac{\sigma^{-2}\{(n-1)+Z_\alpha\sqrt{2(n-1)}\}-n}{\sqrt{n}}\right) \\
&= 1-\Phi\left(\frac{(\sigma^{-2}-1)\sqrt{n}}{\sqrt{\kappa+2}} - \frac{\sigma^{-2}}{\sqrt{n(\kappa+2)}} + \frac{\sqrt{2}Z_\alpha}{\sqrt{\kappa+2}}\sqrt{\frac{n-1}{n}}\right) \\
&\approx 1-\Phi\left(\frac{(\sigma^{-2}-1)\sqrt{n}}{\sqrt{\kappa+2}} + \frac{\sqrt{2}}{\sqrt{\kappa+2}}Z_\alpha\right) \longrightarrow 1
\end{aligned}
$$

So the power $\to 1$ as $n \to \infty$, the test is consistent.

Multinominal Vectors and χ^2 statistic

Let $(n_1, \cdots, n_K)$ be multinominal $(n; p_1, \cdots, p_K)$ with each $p_i > 0$. Then

$$X_n = \sqrt{n}\left(\frac{n_1}{n} - p_1, \cdots, \frac{n_K}{n} - p_K\right) := (X_{n1}, \cdots, X_{nK}) \xrightarrow{d.} N(0, \Sigma)$$

where $\Sigma = (\sigma_{ij})$ with

$$\sigma_{ij} = \begin{cases} p_i(1-p_j) & i = j \\ -p_i p_j & i \neq j \end{cases}$$

A test statistic for goodness-of-fit is:

$$\begin{aligned} T_n &= \sum_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{K} \frac{1}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 \\ &= \sum_{i=1}^{K} \frac{1}{p_i} X_{ni}^2 = X_n^\top C X_n \end{aligned}$$

where $C = \text{diag}\left(p_1^{-1}, \cdots, p_K^{-1}\right)$.

Multinominal Vectors and χ^2 statistic

By mapping theorem,

$$T_n \xrightarrow{d.} Z^\top \Sigma^{1/2} C \Sigma^{1/2} Z \stackrel{d.}{=\!=} \chi^2_{n-1}$$

In fact, we can show that $A = \Sigma^{1/2} C \Sigma^{1/2}$ is an idempotent:

$$\sigma_{ij} = p_i(\delta_{ij} - p_j), \qquad C\Sigma = (p_i^{-1}\sigma_{ij}) = (\delta_{ij} - p_j)$$

$$(C\Sigma)^2 = \left(\sum_{l=1}^{K} (\delta_{il} - p_l)(\delta_{lj} - p_j) \right) = (\delta_{ij} - p_j) = C\Sigma$$

As a result, $C\Sigma$ is an idempotent, so is A, which entails:

$$\mathrm{tr}\left(\Sigma^{1/2} C \Sigma^{1/2}\right) = \mathrm{tr}(C\Sigma) = n - 1$$

Wald Test

Suppose $X_1, \cdots, X_n$ i.i.d. F on $\mathbb{R}^p$ with μ and $\Sigma > 0$ (p fixed).

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

Wald statistic is:

$$W_n = n(\bar{X} - \mu_0)^\top S_n^{-1}(\bar{X} - \mu_0), \quad S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$$

From LLN, $S_n \xrightarrow{p.} \Sigma$, $S_n^{-1} \xrightarrow{p.} \Sigma^{-1}$. And we note that:

$$\bar{X} - \mu = O_p(n^{-1/2})$$

Wald Test

Hence,

$$
\begin{aligned}
W_n &= n(\bar{X} - \mu_0)^\top \left(\Sigma^{-1} + S_n^{-1} - \Sigma^{-1}\right)(\bar{X} - \mu_0) \\
&= \sqrt{n}(\bar{X} - \mu_0)^\top \Sigma^{-1} \sqrt{n}(\bar{X} - \mu_0) \\
&+ \sqrt{n}(\bar{X} - \mu_0)^\top \left(S_n^{-1} - \Sigma^{-1}\right) \sqrt{n}(\bar{X} - \mu_0) \\
&= \sqrt{n}(\bar{X} - \mu_0)^\top \Sigma^{-1} \sqrt{n}(\bar{X} - \mu_0) + o_p(1) \xrightarrow{d.} \chi_p^2
\end{aligned}
$$

as $\sqrt{n}\Sigma^{-1/2}(\bar{X} - \mu_0) \xrightarrow{d.} N(0, I_p)$.

So Wald test requires H_0 if $W_n > \chi^2_{p,1-\alpha}$.

Variance Stablizing Transform (VST)

Typically, we have various means to obtain,

$$\sqrt{n}(T_n - \theta) \xrightarrow{d.} N\left(0, \sigma^2(\theta)\right)$$

where $\sigma^2(\theta)$ is the asymptotic variance depending on θ. The asymptotic CI for θ are:

$$T_n \pm Z_{1-\alpha/2}\frac{\sigma(\hat{\theta})}{\sqrt{n}}$$

So the width of the CIs varies with respect to $\sigma(\theta)$.

The purpose of VST is to transform T_n to $\phi(T_n)$ such that

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{d.} N(0, c^2)$$

where $c > 0$ is a constant.

Variance Stablizing Transform (VST)

From earlier result,

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{d.} \phi'(\theta) N\left(0, \sigma^2(\theta)\right) \stackrel{d.}{=\!=} N\left(0, \left(\phi'(\theta)\right)^2 \sigma^2(\theta)\right)$$

So $\phi'(\theta)\sigma(\theta) = c$, which implies:

$$\phi'(\theta) = \frac{c}{\sigma(\theta)}, \qquad \phi(\theta) = \int \frac{d\theta}{\sigma(\theta)}$$

is the VST.

Tukey's Hanging Rootogram

Let $X_1, \cdots, X_n$ i.i.d. the pdf f. The Kernel Density Estimator is:

$$\hat{f}_{nh}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where K is a symmetric pdf, $N(0,1)$'s pdf, for instance. It can be shown (Serfling[5], P114) that:

$$\sqrt{nh}(\hat{f}_{nh}(x) - f(x)) \xrightarrow{d.} N(0, f(x))$$

provided $nh^s \to 0$ and $nh \to \infty$ as $n \to \infty$. So $\hat{f}_{nh}(x)$ is $AN(f(x), \frac{f(x)}{nh})$. By Delta method and VST:

$$\phi(f) = \int \frac{df}{\sqrt{f}} = f^{1/2}$$

So we do "root-gram": $\hat{f}_{nh}^{1/2}(x)$ is:

$$AN(f^{1/2}(x), \frac{1}{4hn})$$

[5] Robert J Serfling. *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons, 2009.

Asymptotically Uniformly Integrable

Definition 6.3 (Uniform Integrability)

A sequence of random variables $\{Y_n\}_{n\geq 0}$ is called asymptotic uniformly integrable (u.i.) if:

$$\lim_{M\to\infty} \limsup_{n\to\infty} E\left[|Y_n|\mathbb{I}_{\{Y_n>M\}}\right] = 0$$

The uniform integrability is the missing link between convergence in distribution and convergence of moments.

Theorem 6.4

Let $f : \mathbb{R}^k \to \mathbb{R}$ be measurable and continuous at every point in a set C, $X_n \xrightarrow{d} X$ where X takes its values in C. Then $Ef(X_n) \to Ef(X)$ if and only if the sequence of r.v. $f(X_n)$ is asymptotically u.i.

Moment Approximation

If T_n has m-th moment exist. Knowing $\sqrt{n}(T_n - \theta) \xrightarrow{d} T \Rightarrow \sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$, if $\phi'(\theta) \neq 0$.

- Can we approximate $E\phi(T_n)$ by Taylor expansion?

$$\phi(T_n) = \phi(\theta) + \phi`(\theta)(T_n - \theta) + \frac{1}{2}\phi``(\theta)(T_n - \theta)^2 + \cdots$$

- So that, do we have the following equations:

$$E\phi(T_n) \approx \phi(\theta) + \phi`(\theta)\text{Bias}(T_n) + \frac{1}{2}\phi``(\theta)\text{MSE}(T_n)$$
$$\text{var}\,(\phi(T_n)) \approx (\phi`(\theta))^T \text{var}(T_n)\,(\phi`(\theta))$$

- We need $\phi(T_n) - \phi(\theta)$ being u.i. If $T_n - \theta$ is u.i. and ϕ is Lipschitz, then $\phi(T_n) - \phi(\theta)$ is u.i..
- See also Sargan, J.D. (1976, Econometrica).

Let $X_1, \ldots, X_n$ i.i.d. F_θ, where θ_0 is the true parameter, $f_1, \ldots, f_k$ be given known function.

- Moments:

$$E_\theta f_j(X) = \int f_j(x) dF_\theta(x), \quad j = 1, \ldots, k$$

- A popular or original choice is: $f_j(x) = x^j$. Let $f = (f_1, \ldots, f_k)'$.

Definition 7.1 (Moment Estimator (ME))

Match sample moments $\frac{1}{n}\sum_{i=1}^n f(x_i)$ with its population counterparts:

$$P_n f := \frac{1}{n}\sum_{i=1}^n f(x_i) = e(\theta) = E_\theta f(X) := P_\theta f$$

AN for ME

- If e is one to one, then the ME is $\hat{\theta}_n = e^{-1}(P_n f)$.
- If e^{-1} is differential and $E_{\theta_0} f(X) f^T(X) < \infty$ (which implies the AN of $P_n f$), then we can have the AN of $\hat{\theta}_n$.
- Note that:

$$\left(e^{-1}(x_0)\right)' = \left(e'(\theta_0)\right)^{-1}\Big|_{\theta_0 = e^{-1}(x_0)}$$

Theorem 7.2 (CLT for ME)

If $e(\theta) = P_\theta f =: E_\theta f(X)$ is 1-1 on an open set $\Theta \subset \mathbb{R}^k$ and is continuously differentiable at θ_0 with non-singular e'_θ and $P_{\theta_0}\|f\|^2 < \infty$, then $\hat{\theta}_n$ exists with prob approaching to 1 (wpa 1) and:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, [(e'(\theta_0))^{-1}](E_{\theta_0} ff^T - E_{\theta_0} f E_{\theta_0} f^T)[(e'(\theta_0))^{-1}]^T\right)$$

We need the inverse function in the proof.

Lemma 7.3 (The inverse function theorem)

Let A be open in $\mathbb{R}^n$; let $g : A \mapsto \mathrm{R}^k$ is continuously differentiable at $\boldsymbol{a}$ and differentiable in a neighborhood of $\boldsymbol{a} \in A$. If the Jacobi matrix

$$Dg(x) := \partial g(x)/\partial x^\tau$$

is non-singular at the point $\boldsymbol{a}$ and of A.

Then,

- *there is a neighborhood U of the point $\boldsymbol{a}$, such that*
$$g : U \mapsto V \text{ is one-to-one}$$
for an open set V of $\mathbb{R}^n$;
- *and there exists an inverse function $g^{-1} : V \mapsto U$ which is continuously differentiable with*

$$Dg^{-1}(y) := \partial g^{-1}(y)/\partial y^\tau = (Dg(x))^{-1}.$$

Proof of Theorem 7.2

- Continuous differentiability at θ_0 presumes differentiability in a neighborhood and the continuity of $\theta \mapsto e'_\theta$;
  the nonsingularity of e'_θ implies nonsingularity in a neighborhood.
- Therefore, by the **inverse function theorem** there exist open neighborhoods U of θ_0, and V of $P_{\theta_0} f$ such that

$$e : U \mapsto V \text{ is a differentiable bijection (one-to-one)}$$
$$\text{with a differentiable inverse } e^{-1} : V \mapsto U.$$

- By the LLN, $\mathbb{P}_n f \equiv \frac{1}{n} \sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} e(\theta_0) \in V$. Since $\mathbb{P}_n f \in V$,

$$\hat{\theta}_n := e^{-1}(\mathbb{P}_n f) \xrightarrow{a.s.} e^{-1}(e(\theta_0)) = \theta_0$$

  exist with probability tending to 1 by continous mapping theorem.
- The CLT guarantees asymptotic normality of the sequence $\sqrt{n}(\mathbb{P}_n f - P_{\theta_0} f)$. The proof is finished by Delta Method.

Example : ME for the Beta distribution

Let $X_1, X_2, \cdots X_n$ be a random sample from the Beta distribution which has a density function $f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $x \in (0,1)$ where $\alpha > 0$ and $\beta > 0$ are two unknown parameters, and $B(\alpha,\beta)$ is the Beta function. Note that

$$\mathbb{E}X^k = \frac{1}{B(\alpha,\beta)} \int x^{\alpha+k-1}(1-x)^{\beta-1} dx = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)}$$
$$= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)} = \prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}, \quad k = 1, 2, \ldots$$

then the moment estimator can be solved by the following equations:

$$\overline{X} = \frac{\alpha}{\alpha+\beta}, \quad \overline{X^2} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

namely, $\hat{\alpha} = (1-\overline{X}) \left[\frac{\overline{X}(\overline{X}-1)}{\overline{X^2}-\overline{X}^2} - 1 \right], \quad \hat{\beta} = \overline{X} \left[\frac{\overline{X}(\overline{X}-1)}{\overline{X^2}-\overline{X}^2} - 1 \right]$ is the solutions.

Example of Beta distribution: Con.

Let $\theta = (\alpha, \beta)$ be the set of true parameters vector. The moment function is $f(x) = (x, x^2)^T$ and the estimated function is:

$$e(\theta) = \mathrm{E}_\theta f(x) = \left(\frac{\alpha}{\alpha+\beta}, \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \right)^T$$

It is very easy to verify that $e^{-1} \in C^\infty(\mathbb{R}_+^2)$:

$$\hat{\theta} = e^{-1}\left(\mathbb{P}_n f\right) \xrightarrow{p} e^{-1}\left(P_n f\right) = \theta.$$

Note that $\frac{\partial e(\theta)}{\partial \theta^T} = \begin{bmatrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ -\frac{\beta(\alpha+1)(\alpha+\beta+1)+\alpha\beta(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)^2} & -\frac{\alpha(\alpha+1)(2\alpha+2\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)^2} \end{bmatrix}$, so

$$\mathrm{E}_\theta ff^T - \mathrm{E}_\theta f \mathrm{E}_\theta f^T = \begin{bmatrix} \mathbb{E}X^2 & \mathrm{E}X^3 \\ \mathrm{E}X^3 & \mathrm{E}X^4 \end{bmatrix} - \begin{bmatrix} (\mathrm{E}X)^2 & \mathrm{E}X\mathrm{E}X^2 \\ \mathrm{E}X\mathrm{E}X^2 & (\mathrm{E}X^2)^2 \end{bmatrix}.$$

as a result, we have:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, (\frac{\partial e(\theta)}{\partial \theta^T})^{-1}[\mathrm{E}_\theta ff^T - \mathrm{E}_\theta f \mathrm{E}_\theta f^T](\frac{\partial e(\theta)^T}{\partial \theta})^{-1}) =: N(0, \Sigma)$$

Exponential Family

Suppose $X_1, \ldots, X_n$ i.i.d. F_θ with density:

$$p_\theta(x) = c(\theta)h(x)\exp\{\theta^T t(x)\}$$

- Likelihood:

$$l_\theta(x) = \log p_\theta(x) = \log c(\theta) + \log h(x) + \theta^T t(x)$$

- Likelihood score:

$$\dot{l}_\theta(x) = \frac{\dot{c}(\theta)}{c(\theta)} + t(x) = t(x) - E_\theta t(X)$$

- $E\dot{l}_\theta(X) = 0$ implies $\dot{c}(\theta)/c(\theta) = -E_\theta t(X)$.

Hence, MLE are MEs:

$$\frac{1}{n}\sum_{i=1}^{n} t(x_i) = E_\theta t(X) = e(\theta)$$

Exponential Family

Furthermore, if $e(\theta)$ is continuous and differentiable and a moment condition $E_\theta \|t(X)\|^2 < \infty$ is satisfied, then:

- $\hat{\theta}_{mle} = \hat{\theta}_{me}$ exists wpa 1;
- and

$$\sqrt{n}\left(\hat{\theta}_{me} - \theta\right) \xrightarrow{d} N\left(0, [\left(e'(\theta)\right)^{-1}]\text{var}_\theta\left(t(x)\right)[\left(e'(\theta)\right)^{-1}]^T\right)$$
$$= N\left(0, I_\theta^{-1}\right)$$

- I_θ is Fisher Information matrix:

$$I_\theta = \text{var}(\dot{l}_\theta(X)) = E\left(\dot{l}_\theta(X)\dot{l}_\theta^T(X)\right) = -E\ddot{l}_\theta(X)$$

We can check the equal-sign in the preceding asymptotic distribution.

Generalized Method of Moments (GMM)

MLE requires parameter model specificates on:

MLE

(i) The full density of $w = (x, y)$, $f(w; \eta)$; or

(ii) The conditional density of y given x, $f(y|x; \theta)$; or

(iii) The partial density of y_t given x_t, $f_t(y_t|x_t; \theta)$ in the context of panel data.

GMM was developed by Lars Peter Hansen in 1982 as a generalization of the method of moments. GMM requires less model specifications:

GMM

(i) Instead of requiring densities, it asks for only moment restrictions;

(ii) GMM is largely "semi-parameter", where the parameter of interest is finite-dimensional [the full data's distribution function may not be known (infinite-dimensional)], and therefore MLE is not applicable.

Examples: Paremetric regressions

The regression data $\{w_i = (Y_i, x_)i\}_{i=1}^n$ ($Y_i \in \mathbb{R}$: response, x_i: covariate)

$$y_i = m(x_i, \theta) + \varepsilon_i, E(\varepsilon_i | x_i) = 0, \mathrm{Var}(\varepsilon_i | x) = \sigma^2(x_i) < \infty$$

where $\{\varepsilon_i\}_{i=1}^n$ are the indepedent error variable.
By least square method, the score functions is

$$g(w_i, \theta) = \frac{\partial m(x_i, \theta)}{\partial \theta}(y_i - m(x_i, \theta)).$$

and the weighted least square method leads to

$$g(w_i, \theta) = \frac{\partial m(x_i, \theta)}{\partial \theta} \cdot \frac{y_i - m(x_i, \theta)}{\sigma^2(x_i, r_0)}.$$

if is $\sigma^2(x_i, \gamma_0)$ known.

In both methods, $\gamma_0 = p$. And $\hat{\theta}$ is directly solved by estimating equation $\frac{1}{n}\sum_{i=1} g(w_i, \theta) = 0$.

Example: Poisson regressions

Consider the equal-dispersion assumption for Y_i being count data

$$E\left(Y_i|x_i\right) = \mu\left(x_i, \theta\right), \quad \sigma^2\left(Y_i|x_i\right) = \mu\left(x_i, \theta\right).$$

The $\mu\left(x_i, \theta\right)$ is a known funtion, for example: $\mu\left(x_i, \theta\right) = e^{x_i^\top \theta}$. Define

$$g\left(w_i, \theta\right) = \left(\begin{array}{c} g_1\left(w_i, \theta\right) \\ g_2\left(w_i, \theta\right) \end{array}\right) =: \left(\begin{array}{l} \frac{\partial \mu(x_i,\theta)}{\partial \theta}\left(y_i - \mu\left(x_i, \theta\right)\right) \\ a\left(x_i, \theta\right)\left\{\left[y_i - \mu\left(x_i, \theta_i\right]^2 - \mu\left(x_i, \theta_i\right)\right\}\right. \end{array}\right)$$

We can choose $a\left(x_i, \theta\right)$ almost freely to satisfy,, but we need choose one that $\hat{\theta}$ is most efficient.

Let $\{w_i\}_{i=1}^{n}$ be IID r.vs in $\mathbb{R}^m$, $g(w_i;\theta) \in \mathbb{R}^r$ be r-dimensional known function of w_i and $\theta \in \Theta \subset \mathbb{R}^p$. So that

$$\exists \theta_0 \in \Theta, \quad E\{g(w_i;\theta_0)\} = 0$$

- If $r = p$, we call it "just-identified";
- If $r = p$, the $\hat{\theta}$ can be made by solving directly:

$$\frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta) = 0$$

- When $r > p$, we call it "over-identified".

GMM

Definition 7.4

Given $g(w_i;\theta)$ s.t. $Eg(w_i;\theta_0) = 0$ for some $\theta_0 \in \Theta$. The GMM estimator $\hat{\theta}_{GMM}$ of θ is

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta)\right)^T \widehat{W}_n \left(\frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta)\right)$$

for contain $r \times r$ non-negative definite matrices $\widehat{W}_n$, which satisfied that $\widehat{W}_n \xrightarrow{p} W_0 > 0$, W_0 is deterministic and may depend on θ_0.

The GMM estimator above is asymptotically equivalent to

$$\hat{\theta}_{n0} = \underset{\theta\in\Theta}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} g^T(w_i;\theta) W_0 \frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta)\right)$$

which is a M-estimator.

GMM Identification

- To ensure identification of θ_0, we assume θ_0 is the" unique"

$$\theta \in \Theta, \quad s.t. \quad Eg(w_i;\theta) = 0$$

- As $W_0 > 0$, θ_0 is also the unique θ which minimizes

$$E\{g^T(w_i;\theta)\}W_0E\{g(w_i;\theta)\}$$

- Under certain conditions, we have $\hat{\theta}_{GMM} \xrightarrow{p} \theta_0$.

Asymptotic Normality

Suppose:

(i) $g(w,\cdot)$ is a continuous differentiable function on $\theta \in \text{Int}(\Theta)$;

(ii) $G_0 = E\left(\frac{\partial g(w,\theta_0)}{\partial \theta}\right)_{r\times p}$ exists and its has full rank p.

Then under the assumption $\hat{\theta}_n \xrightarrow{p} \theta_0$, we have

- AN:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \left(G_0^T W_0 G_0\right)^{-1} G_0^T W_0 \Lambda_0 W_0 G_0 \left(G_0^T W_0 G_0\right)^{-1}\right)$$

where

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(w_i;\theta_0) \xrightarrow{d} N\left(0, \Lambda_0\right), \quad \Lambda_0 = \text{var}\left\{g(w_i;\theta_0)\right\}$$

Proof

From the definition of $\hat{\theta}_n$ (Definition 6.4),

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g^\top(w_i,\hat{\theta}_n)}{\partial\theta}\widehat{W}_n\left(\frac{1}{n}\sum_{i=1}^{n}g(w_i,\hat{\theta}_n)\right)=0 \tag{34}$$

Note that:

$$\frac{\partial g^\top Wg}{\partial\theta_i}=\frac{\partial g^\top}{\partial\theta_i}\frac{\partial g^\top Wg}{\partial g}=\frac{\partial g^\top}{\partial\theta_i}2Wg$$

So,

$$\frac{\partial g^\top Wg}{\partial\theta}=\frac{\partial g^\top}{\partial\theta}\frac{\partial g^\top Wg}{\partial g}=2\frac{\partial g^\top}{\partial\theta_i}Wg$$

By using Taylor formula on (34) around θ_0 (assuming $\hat{\theta}_n \xrightarrow{p} \theta_0$):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g^\top(w_i,\hat{\theta}_n)}{\partial\theta}\widehat{W}_n\left\{\frac{1}{n}\sum_{i=1}^{n}g(w_i,\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(w_i,\hat{\theta}_n^*)}{\partial\theta}(\hat{\theta}_n-\theta_0)\right\}=0 \tag{35}$$

where $\hat{\theta}_n^*$ is between θ_0 and $\hat{\theta}_n$.

Proof

Since $\hat{\theta}_n \xrightarrow{p.} \theta_0$, it is easy to verify that:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g^\top(w_i,\hat{\theta}_n)}{\partial\theta},\ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g^\top(w_i,\hat{\theta}_n^*)}{\partial\theta} \xrightarrow{p.} G_0^\top = \mathrm{E}\left[\frac{\partial g(w,\theta_0)}{\partial\theta}\right]$$

Note that $\widehat{W}_n \xrightarrow{p.} W_0$, from (35),

$$A_0(\hat{\theta}_n-\theta_0) := \left(G_0^\top W_0 G_0\right)\left(\hat{\theta}_n-\theta_0\right) = -G_0 W_0 \frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta_0)\left\{1+o_p(1)\right\}$$

and,

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = -A_0^{-1}G_0^\top W_0 \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(w_i,\theta_0)+o_p(1)$$

As $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(w_i,\theta_0) \xrightarrow{d.} N(0,\Lambda_0)$ where $\Lambda_0 = \mathrm{var}(g(w_i,\theta_0))$, we obtain:

$$\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{d.} N\left(0, A_0^{-1}G_0^\top W_0\Lambda_0 W_0 G_0 A_0^{-1}\right)$$

Thus establish the asymptotic normality of $\hat{\theta}_n$.

If we use W_0 instead of W_n, then

- We can attain the same asymptotic normal distribution.
- Denote $A_0 = \left(G_0^T W_0 G_0\right)$ and $B_0 = G_0^T W_0 \Lambda_0 W_0 G_0$, Then $\text{Avar}(\hat{\theta}_{GMM}) = A_0^{-1} B_0 A_0^{-1}/n$.
- It can be estimated by $\widehat{A}_0^{-1}\widehat{B}_0\widehat{A}_0^{-1}/n$, where

$$\widehat{A}_0 = \widehat{G}^T \widehat{W_n} \widehat{G}, \quad \widehat{B}_0 = \widehat{G}^T \widehat{W_n} \widehat{\Lambda} \widehat{G} \widehat{W_n}$$

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(w_i; \hat{\theta})}{\partial \theta}, \quad \widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} g(w_i; \widehat{\theta}) g^T(w_i; \widehat{\theta})$$

Optional Weighting Matrix

The GMM estimator depends on the choice of $\widehat{W}_n$, the weighting matrix.

Question: which $\widehat{W}_n$ or W_0 is the best?

- If we choose $W_0 = \Lambda_0^{-1}$, then $B_0 = A_0$ and

$$\mathrm{Avar}(\hat{\theta}_{GMM}) = A_0^{-1} = \left(G_0^T \Lambda_0^{-1} G_0\right)^{-1}$$

- In Hansen(1982), it can be shown that for any $W_0 > 0$:

$$\begin{aligned}\mathrm{Avar}[\hat{\theta}_n(W_0)] &:= (G_0^T W_0^{-1} G_0)^{-1}(G_0^T W_0 \Lambda_0 W_0 G_0)(G_0^T W_0^{-1} G_0)^{-1} \\ &\geq (G_0^T \Lambda_0^{-1} G_0)^{-1} =: \mathrm{Avar}[\hat{\theta}_n(\Lambda_0^{-1})]\end{aligned}$$

- Hence, the choice of $W_0^* = \Lambda_0^{-1}$ is the optimal.

Weight Matrix

- The efficient GMM estimator will use $\widehat{W}_n = \widehat{\Lambda}^{-1}$ as the weight matrix, where

$$\widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} g(w_i;\hat{\hat{\theta}})g^T(w_i;\hat{\hat{\theta}})$$

- The $\hat{\hat{\theta}}$ is an initial estimator, any consistent estimator of θ can be used.
- For instance, it can be derived by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} g^T(w_i;\theta)\frac{1}{n}\sum_{i=1}^{n} g(w_i;\theta)$$

- So $\hat{\hat{\theta}}$ is a GMM with $\widehat{W}_n = I_r$.

The GMM estimator

Step 1: Construct an initial estimator $\hat{\hat{\theta}}$, which is a GMM with any weight $\widehat{W}_n > 0$, for example $\widehat{W}_n = I_r$.

Step 2: Obtain the optimal weight matirx

$$\widehat{W}_n^* = \left(\frac{1}{n} \sum_{i=1}^{n} g(w_i; \hat{\hat{\theta}}) g^T(w_i; \hat{\hat{\theta}}) \right)^{-1}$$

Then the GMM estimator with $\widehat{W}_n^*$ as the weight matrix satisfies:

$$\sqrt{n} \left(\hat{\theta}^*_{GMM} - \theta_0 \right) \xrightarrow{d} N \left(0, \left(G_0^T \Lambda_0^T G_0 \right)^{-1} \right)$$

Sargan-Hansen Test

It can be shown with $\hat{\theta} = \hat{\theta}_{GMM}$, the objective function satisfies (Homework):

$$T_n(\hat{\theta}) = n^{-1/2} \sum_{i=1}^{n} g^T(w_i; \hat{\theta}) \widehat{W_n^*} n^{-1/2} \sum_{i=1}^{n} g(w_i; \hat{\theta}) \xrightarrow{d} \chi^2_{r-p}$$

From this asymptotic distribution:

- We also need the condition $r - p \geq 1$.
- Hypothesis Testing: reject $H_0 : Eg(w_i; \theta_0) = 0$ if $T_n(\theta) > \chi^2_{r-p,1-\alpha}$.
- $\hat{\theta}_{GMM}$ also is named minimum χ^2-estimator.

Hypothesis Testing

Test for $H_0 : c(\theta_0) = 0$, where $c(\theta) \in \mathbb{R}^Q$ with $Q \leq q$.

Wald Test:

$$c^T(\hat{\theta})\Big(\hat{V}(\hat{\theta})\Big)^{-1}c(\hat{\theta}), \quad \hat{V}(\hat{\theta}) = \text{var}\Big(c(\hat{\theta})\Big)$$

LM Test:

$$\widetilde{T}_n = \frac{1}{n}\left[\sum_{i=1}^n g^T(w_i;\tilde{\theta}_n)\widehat{W}_n^* \sum_{i=1}^n g(w_i;\tilde{\theta}_n) - \sum_{i=1}^n g^T(w_i;\hat{\theta}_n)\widehat{W}_n^* \sum_{i=1}^n g(w_i;\hat{\theta}_n)\right]$$

We have $\widetilde{T}_n \xrightarrow{d} \chi^2_Q$ under H_0, where:

$$\tilde{\theta}_n = \underset{\theta\in\Theta, c(\theta)=0}{\text{argmin}} \sum_{i=1}^n g^T(w_i;\theta)\widehat{W}_n^* \sum_{i=1}^n g(w_i;\theta)$$

An important feature

If we have r restrictions, would more moment restrictions lead to more efficiency? To appreciate this question, let

$$g(w,\theta) = (\underbrace{g_{(r-1)}(w,\theta)}_{r}, \underbrace{g_r(w,\theta)}_{1})^T, Eg(w,\theta) = 0.$$

The asymptotic variance of $\hat{\theta}$ based on $g(\cdot,\cdot) \in \mathbb{R}^r$ is:

$$V_r^{-1} := \left[E\left(\frac{\partial g^T}{\partial \theta}\right) E^{-1}(gg^T) E\left(\frac{\partial g}{\partial \theta}\right) \right]^{-1}$$

And asymptotic variance of $\hat{\theta}_{(r-1)}$ based on $g_{(r-1)}$ is:

$$V_{r-1}^{-1} := \left[E\left(\frac{\partial g_{(r-1)}^T}{\partial \theta}\right) E^{-1}(g_{(r-1)}g_{(r-1)}^T) E\left(\frac{\partial g_{(r-1)}}{\partial \theta}\right) \right]^{-1}$$

An important feature in GMM:

- $V_r^{-1} \le V_{r-1}^{-1}$.
- Hence, GMM with r restrictions is at least as efficient as GMM with $r-1$ restrictions.

Proof

Note that at $\theta = \theta_0$

$$E(gg^T) = \begin{pmatrix} E\left(g_{(r-1)}g_{(r-1)}^T\right) & E\left(g_{(r-1)}g_r^T\right) \\ E\left(g_{(r-1)}^T g_r\right)\right) & E\left(g_r^2\right) \end{pmatrix} := \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$$

Let $D = C - B^T AB$. The inverse matrix of block matrices formula gives:

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} A^{-1}B \\ -I \end{pmatrix} D^{-1} \left(B^\top A, -I\right)$$

Write $E(gg^T) := \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ and $A := E\left(g_{(r-1)}g_{(r-1)}^T\right)$. We have

$$\left(E(gg^T)\right)^{-1} = \begin{bmatrix} \left(E\left(g_{(r-1)}g_{(r-1)}^T\right)\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} + H$$

where $H \geq 0$.

Proof

Therefore,

$$V_r = \begin{pmatrix} \mathrm{E}\left(\frac{\partial g_{(r-1)}}{\partial \theta}\right) \\ \mathrm{E}\left(\frac{\partial g_r}{\partial \theta}\right) \end{pmatrix}^\top \left\{ \begin{bmatrix} \left(\mathrm{E}\left(g_{(r-1)} g_{(r-1)}^T\right)\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix} + H \right\} \begin{pmatrix} \mathrm{E}\left(\frac{\partial g_{(r-1)}}{\partial \theta}\right) \\ \mathrm{E}\left(\frac{\partial g_r}{\partial \theta}\right) \end{pmatrix}$$

$$= \mathrm{E}^\top\left(\frac{\partial g_{(r-1)}}{\partial \theta}\right)\left(\mathrm{E}\left(g_{(r-1)} g_{(r-1)}^T\right)\right)^{-1} \mathrm{E}\left(\frac{\partial g_{(r-1)}}{\partial \theta}\right) + \widetilde{H}$$

$$= V_{r-1} + \widetilde{H} \geq V_{r-1}, \quad V_r^{-1} \leq V_{r-1}^{-1}$$

Remark 21

From the above inequality, we know that if

$$\mathrm{E}\left(\frac{\partial g^\top}{\partial \theta}\right) H \mathrm{E}\left(\frac{\partial g}{\partial \theta}\right) \neq 0$$

then there will be reduction of the asymptotic variance in using the moment restriction in some directions or combination of the parameter space. This is exactly the attraction of GMM.