Lecture 4: Asymptotic Distribution Theory[*]

In time series analysis, we usually use *asymptotic* theories to derive joint distributions of the estimators for parameters in a model. Asymptotic distribution is a distribution we obtain by letting the time horizon (sample size) go to infinity. We can simplify the analysis by doing so (as we know that some terms converge to zero in the limit), but we may also have a finite sample error. Hopefully, when the sample size is large enough, the error becomes small and we can have a satisfactory approximation to the *true* or *exact* distribution. The reason that we use asymptotic distribution instead of exact distribution is that the exact finite sample distribution in many cases are too complicated to derive, even for Gaussian processes. Therefore, we use asymptotic distributions as alternatives.

1 Review

I think that this lecture may contain more propositions and definitions than any other lecture for this course. In summary, we are interested in two type of asymptotic results. The first result is about convergence to a constant. For example, we are interested in whether the sample moments converge to the population moments, and law of large numbers (LLN) is a famous result on this. The second type of results is about convergence to a random variable, say, Z, and in many cases, Z follows a standard normal distribution. Central limit theorem (CLT) provides a tool in establishing asymptotic normality.

The confusing part in this lecture might be that we have several versions of LLN and CLT. The results may look similar, but the assumptions are different. We will start from the strongest assumption, *i.i.d.*, then we will show how to obtain similar results when *i.i.d.* is violated. Before we come to the major part on LLN and CLT, we review some basic concepts first.

1.1 Convergence in Probability and Convergence Almost Surely

Definition 1 (Convergence in probability) *X_n is said to be convergent in probability to X if for every $\epsilon > 0$,*

$$P(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

If $X = 0$, we say that X_n converges in probability to zero, written $X_n = o_p(1)$, or $X_n \to_p 0$.

Definition 2 (Boundedness in probability) *X_n is said to be bounded in probability, written $X_n = O_p(1)$ if for every $\epsilon > 0$, there exists $\delta(\epsilon) \in (0, \infty)$ such that*

$$P(|X_n| > \delta(\epsilon)) < \epsilon \quad \forall\ n$$

[*] Copyright 2002-2006 by Ling Hu.

We can similarly define order in probability: $X_n = o_p(n^{-r})$ if and only if $n^r X_n = o_p(1)$; and $X_n = O_p(n^{-r})$ if and only if $n^r X_n = O_p(1)$.

Proposition 1 *if X_n and Y_n are random variables defined in the same probability space and $a_n > 0$, $b_n > 0$, then*

(i) If $X_n = o_p(a_n)$ and $Y_n = o_p(b_n)$, we have

$$\begin{aligned} &X_n Y_n = o_p(a_n b_n) \\ &X_n + Y_n = o_p(\max(a_n, b_n)) \\ &|X_n|^r = o_p(a_n^r) \quad \text{for} \quad r > 0. \end{aligned}$$

(ii) If $X_n = o_p(a_n)$ and $Y_n = O_p(b_n)$, we have $X_n Y_n = o_p(a_n b_n)$.

Proof of (i): If $|X_n Y_n|/(a_n b_n) > \epsilon$ then either $|Y_n|/b_n \leq 1$ and $|X_n|/a_n > \epsilon$ or $|Y_n|/b_n > 1$ and $|X_n Y_n|/(a_n b_n) > \epsilon$, hence

$$\begin{aligned} P(|X_n Y_n|/(a_n b_n) > \epsilon) &\leq P(|X_n|/a_n > \epsilon) + P(|Y_n|/b_n > 1) \\ &\to 0 \end{aligned}$$

If $|X_n + Y_n|/\max(a_n, b_n) > \epsilon$, then either $|X_n|/a_n > \epsilon/2$ or $|Y_n|/b_n > \epsilon/2$.

$$\begin{aligned} P(|X_n + Y_n|/\max(a_n, b_n) > \epsilon) &\leq P(|X_n|/a_n > \epsilon/2) + P(|Y_n|/b_n > \epsilon/2) \\ &\to 0. \end{aligned}$$

Finally,

$$P(|X_n|^r/a_n^r > \epsilon) = P(|X_n|/a_n > \epsilon^{1/r}) \to 0.$$

Proof of (ii): If $|X_n Y_n|/(a_n b_n) > \epsilon$, then either $|Y_n|/b_n > \delta(\epsilon)$ and $|X_n Y_n|/(a_n b_n) > \epsilon$ or $|Y_n|/b_n \leq \delta(\epsilon)$ and $|X_n|/a_n > \epsilon/\delta(\epsilon)$, then

$$\begin{aligned} P(|X_n Y_n|/(a_n b_n) > \epsilon) &\leq P(|X_n|/a_n > \epsilon/\delta(\epsilon)) + P(|Y_n|/b_n > \delta(\epsilon)) \\ &\to 0 \end{aligned}$$

This proposition is very useful. For example, if $X_n = o_p(n^{-1})$ and $Y_n = o_p(n^{-2})$, then $X_n + Y_n = o_p(n^{-1})$, which tells that the slowest convergent rate 'dominates'. Later on, we will see sum of several terms, and to study the asymptotics of the sum, we can start from judging the convergent rates of each term and pick the terms that converge slowerest. In many cases, the terms that converges faster can be omitted, such as Y_n in this example.

The results also hold if we replace o_p in (i) with O_p. The notations above can be naturally extended from sequence of scalar to sequence of vector or matrix. In particular, $\mathbf{X}_n = o_p(n^{-r})$ if and only if all elements in $\mathbf{X}$ converge to zero at order n^r. Using Euclidean distance $|\mathbf{X}_n - \mathbf{X}| = \left(\sum_{j=1}^{k}(X_{nj} - X_j)^2\right)^{1/2}$, where k is the dimension of X_n, we also have

Proposition 2 $\mathbf{X}_n - \mathbf{X} = o_p(1)$ *if and only if* $|\mathbf{X}_n - \mathbf{X}| = o_p(1)$.

Proposition 3 (Preservations of convergence of continuous transformations) *If $\{\mathbf{X}_n\}$ is a sequence of k-dimensional random vectors such that $\mathbf{X}_n \to \mathbf{X}$ and if $g : \mathbb{R}^k \to \mathbb{R}^m$ is a continuous mapping, then $g(\mathbf{X}_n) \to g(\mathbf{X})$.*

Proof: let M be a positive real number. Then $\forall\ \epsilon > 0$, we have

$$\begin{aligned} P(|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon) &\leq P(|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon, |\mathbf{X}_n| \leq M, |\mathbf{X}| \leq M) \\ &\quad +P(\{|\mathbf{X}_n| > M\} \cup \{|\mathbf{X}| > M\}). \end{aligned}$$

(the above inequality uses

$$P(A \cup B) \leq P(A) + P(B)$$

where

$$A = \{|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon, |\mathbf{X}_n| \leq M, |\mathbf{X}| \leq M)\},$$
$$B = \{|\mathbf{X}_n| > M, |\mathbf{X}| > M\}.$$

) Recall that if a function g is uniformly continuous on $\{\mathbf{x} : |\mathbf{x}| \leq M\}$, then $\forall \epsilon > 0$, $\exists \eta(\epsilon)$, $|\mathbf{X}_n - \mathbf{X}| < \eta(\epsilon)$, so that $|g(\mathbf{X}_n) - g(\mathbf{X})| < \epsilon$. Then

$$\{|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon, |\mathbf{X}_n| \leq M, |\mathbf{X}| \leq M)\} \subseteq \{|\mathbf{X}_n - \mathbf{X}| > \eta(\epsilon).\}$$

Therefore,

$$\begin{aligned} P(|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon) &\leq P(|\mathbf{X}_n - \mathbf{X}| > \eta(\epsilon)) + P(|\mathbf{X}_n| > M) + P(|\mathbf{X}| > M) \\ &\leq P(|\mathbf{X}_n - \mathbf{X}| > \eta(\epsilon)) + P(|\mathbf{X}| > M) \\ &\quad +P(|\mathbf{X}| > M/2) + P(|\mathbf{X}_n - \mathbf{X}| > M/2) \end{aligned}$$

Given any $\delta > 0$, we can choose M to make the second and third terms each less than $\delta/4$. Since $\mathbf{X}_n \to \mathbf{X}$, the first and fourth terms will each be less than $\delta/4$. Therefore, we have

$$P(|g(\mathbf{X}_n) - g(\mathbf{X})| > \epsilon) \leq \delta.$$

Then $g(\mathbf{X}_n) \to g(\mathbf{X})$.

Definition 3 (Convergence almost surely) *A sequence $\{X_n\}$ is said to converge to X almost surely or with probability one if*

$$P(\lim_{n\to\infty} |X_n - X| > \epsilon) = 0.$$

If X_n converges to X almost surely, we write $X_n \to_{a.s.} X$. Almost sure convergence is stronger than convergence in probability. In fact, we have

Proposition 4 *If $X_n \to_{a.s.} X$, $X_n \to_p X$.*

However, the converse is not true. Below is an example.

Example 1 (Convergence in probability but not almost surely) Let the sample space $S = [0,1]$, a closed interval. Define the sequence $\{X_n\}$ as

$$X_1(s) = s + \mathbf{1}_{[0,1]}(s) \quad X_2(s) = s + \mathbf{1}_{[0,1/2]}(s) \quad X_3(s) = s + \mathbf{1}_{[1/2,1]}(s)$$

$$X_4(s) = s + \mathbf{1}_{[0,1/3]}(s) \quad X_5(s) = s + \mathbf{1}_{[1/3,2/3]}(s) \quad X_6(s) = s + \mathbf{1}_{[2/3,1]}(s)$$

etc, where $\mathbf{1}$ is the indicator function, i.e., it equals to 1 if the statement is true and equals to 0 otherwise. Let $X(s) = s$. Then $X_n \to_p X$, as $P(|X_n - X| \geq \epsilon)$ is equal to the probability of the length of the interval of s values whose length is going to zero as $n \to \infty$. However, X_n does not converge to X almost surely, Actually there is no $s \in S$ for which $X_n(s) \to s = X(s)$. For every s, the value of $X_n(s)$ alternates between the values of s and $s+1$ infinately often.

1.2 Convergence in L_p Norm

When $E(|X_n|^p) < \infty$ with $p > 0$, X_n is said to be *L_p-bounded*. Define that the L_p norm of X is $\|X\|_p = (E|X|^p)^{1/p}$. Before we define L_p convergence, we first review some useful inequalities.

Proposition 5 (Markov's Inequality) *If $E|X|^p < \infty$, $p \geq 0$ and $\epsilon > 0$, then*

$$P(|X| \geq \epsilon) \leq \epsilon^{-p} E|X|^p$$

Proof:

$$\begin{aligned} P(|X| \geq \epsilon) &= P(|X|^p \epsilon^{-p} \geq 1) \\ &= E\mathbf{1}_{[1,\infty)}(|X|^p \epsilon^{-p}) \\ &\leq E[|X|^p \epsilon^{-p} \mathbf{1}_{[1,\infty)}(|X|^p \epsilon^{-p})] \\ &\leq \epsilon^{-p} E|X|^p \end{aligned}$$

In the Markov's inequality, we can also replace $|X|$ with $|X - c|$, where c can be any real number. When $p = 2$, the inequality is also known as *Chebyshev's inequality*. If X is L_p bounded, then Markov's inequality tells that the tail probabilities converge to zero at the rate ϵ^p as $\epsilon \to \infty$. Therefore, the order of L_p boundedness measures the tendency of a distribution to generate outliers.

Proposition 6 (Holder's inequality) *For any $p \geq 1$,*

$$E|XY| \leq \|X\|_p \|Y\|_q,$$

where $q = p/(p-1)$ if $p > 1$ and $q = \infty$ if $p = 1$.

Proposition 7 (Liapunov's inequality) *If $p > q > 0$, then $\|X\|_p \geq \|X\|_q$.*

Proof: Let $Z = |X|^q$, $Y = 1$, $s = p/q$, Then by Holder's inequality, $E|ZY| \leq \|Z\|_s \|Y\|_{s/(s-1)}$, or

$$E(|X|^q) \leq E(|X|^{qs})^{1/s} = E(|X|^p)^{q/p}.$$

Definition 4 (L_p convergence) *If $\|X_n\|_p < \infty$ for all n with $p > 0$, and $\lim_{n\to\infty} \|X_n - X\|_p = 0$, then X_n is said to converge in L_p norm to X, written $X_n \to_{Lp} X$. When $p = 2$, we say it converges in mean square, written as $X_n \to_{m.s.} X$.*

For any $p > q > 0$, L_p convergences implies L_q convergence by Liaponov's inequality. We can take convergence in probability as an L_0 convergence, therefore, L_p convergence implies convergence in probability:

Proposition 8 (L_p convergence implies convergence in probability) *If $X_n \to_{Lp} X$ then $X_n \to_p X$.*

Proof:

$$\begin{array}{ll} & P(|X_n - X| > \epsilon) \\ \leq & \epsilon^{-p} E|X_n - X|^p \quad \text{by Markov's inequality} \\ \to & 0 \end{array}$$

1.3 Convergence in Distribution

Definition 5 (Convergence in distribution) *The sequence $\{X_n\}_{n=0}^{\infty}$ of random variables with distribution functions $\{F_{X_n}(x)\}$ is said to converge in distribution to X, written as $X_n \to_d X$ if there exists a distribution function $F_X(x)$ such that*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Again, we can naturally extend the definition and related results about scalar random variable X to vector valued random variable $\mathbf{X}$. To verify convergence in distribution of a k by 1 vector, if the scalar $(\lambda_1 X_{1n} + \lambda_2 X_{2n} + \ldots + \lambda_k X_{kn})$ converges in distribution to $(\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k)$ for *any* real values of $(\lambda_1, \lambda_2, \ldots, \lambda_k)$, then the vector $(X_{1n}, X_{2n}, \ldots, X_{kn})$ converges in distribution to the vector $(X_1, X_2, \ldots, X_k)$.

We also have the continuous mapping theorem for convergence in distribution.

Proposition 9 *If $\{\mathbf{X}_n\}$ is a sequence of random k-vectors with $\mathbf{X}_n \to_d \mathbf{X}$ and if $g : \mathbb{R}^k \to \mathbb{R}^m$ is a continuous function. Then $g(\mathbf{X}_n) \to_d g(\mathbf{X})$.*

In the special case that that the limit is a constant scalar or vector, convergence in distribution implies convergence in probability.

Proposition 10 *If $X_n \to_d c$ where c is a constant, then $X_n \to_p c$.*

Proof:. If $X_n \to_d c$, then $F_{X_n}(x) \to \mathbf{1}_{[c,\infty)}(x)$ for all $x \neq c$. For any $\epsilon > 0$,

$$\begin{array}{rcl} P(|X_n - c| \leq \epsilon) & = & P(c - \epsilon \leq X_n \leq c + \epsilon) \\ & \to & \mathbf{1}_{[c,\infty)}(c + \epsilon) - \mathbf{1}_{[c,\infty)}(c - \epsilon) \\ & = & 1 \end{array}$$

On the other side, for a sequence $\{X_n\}$, if the limit of convergence in probability or convergence almost sure is a random variable X, then the sequence also converges in distribution to x.

1.4 Law of Large Numbers

Theorem 1 (Chebychev's Weak LLN) *Let X be a random variable with $E(X) = \mu$ and* $\lim_{n\to\infty} Var(\bar{X}_n) = 0$*, then*

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t \to_p \mu.$$

The proof follow readily from Chebychev's inequality.

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} \to 0.$$

WLLN tells that the sample mean is a consistent estimate for the population mean and the variance goes away as $n \to \infty$. Since $E(\bar{X}_n - \mu)^2 = Var(\bar{X}_n) \to 0$, we also know that $\bar{X}_n$ converges to the population mean in mean square.

Theorem 2 (Kolmogorov's Strong LLN) *Let X_t be i.i.d and $E(|X|) < \infty$, then*

$$\bar{X}_n \to_{a.s.} \mu.$$

Note that Kolmogorov's LLN does not require finite variance. Next we consider the LLN for an heterogeneous process without serial correlations, say, $E(X_t) = \mu_t$ and $Var(X_t) = \sigma_t^2$, and assume that $\bar{\mu}_n = n^{-1}\sum_{t=1}^n \mu_t \to \mu$. Then we know that $E(\bar{X}_n) = \bar{\mu}_n \to \mu$, and

$$Var(\bar{X}_n) = E\left(n^{-1}\sum_{t=1}^{n}(X_t - \mu_t)\right)^2 = n^{-2}\sum_{t=1}^{n}\sigma_t^2.$$

To prove the condition for $Var(\bar{X}_n) \to 0$, we need another fundamental tool in asymptotic theory, Kronecker's lemma.

Theorem 3 (Kronecker's lemma) *Let X_n be a sequence of real numbers and Let $\{b_n\}$ be a monotone increasing sequence with $b_n \to \infty$, and $\sum_{t=1}^{\infty} X_t$ convergent. then*

$$\frac{1}{b_n}\sum_{t=1}^{n} b_t X_t \to 0.$$

Theorem 4 *Let $\{X_t\}$ be a serially uncorrelated sequence, and $\sum_{t=1}^{\infty} t^{-2}\sigma_t^2 < \infty$, then*

$$\bar{X}_n \to_{m.s.} \mu.$$

Proof: take $b_t = t^2$, then by Kronecker's lemma, $Var(\bar{X}_n) = n^{-2}\sum_{t=1}^{n}\sigma_t^2 \to 0$. Then we have $E(\bar{X}_n - \mu)^2 \to 0$, therefore, $\bar{X}_n \to_{m.s.} \mu$.

1.5 Classical Central Limit Theory

Finally, central limit theorem (CLT) provides a tool to establish asymptotic normality of an estimator.

Definition 6 (Asymptotic Normality) *A sequence of random variables $\{X_n\}$ is said to be asymptotic normal with mean μ_n and standard deviation σ_n if $\sigma_n > 0$ for n sufficiently large and*

$$(X_n - \mu_n)/\sigma_n \rightarrow_d Z, \quad \text{where} \quad Z \sim N(0,1).$$

Theorem 5 (Lindeberg-Levy Central Limit Theorem) *If $\{X_n\} \sim iid(\mu, \sigma^2)$, and $\bar{X}_n = (X_1 + \ldots + X_n)/n$, then*

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \rightarrow_d N(0,1).$$

Note that in CLT, we obtain normality results about $\bar{X}_n$ without assuming normality for the distribution of X_n. Here we only require that X_n follows some *i.i.d.* We will see a moment later that central limit theorem also holds for more general cases. Another useful tool which can be used together with LLN and CLT is known as Slutsky's theorem.

Theorem 6 (Slutsky's theorem) *If $X_n \rightarrow X$ in distribution and $Y_n \rightarrow c$, a constant, then*

(a) $Y_n X_n \rightarrow cX$ in distribution.

(b) $X_n + Y_n \rightarrow X + c$ in distribution.

If we know the distribution of a random variable, we can derive the distribution of a function of this random variable using the so called 'δ-method'.

Proposition 11 (δ-method) *Let $\{X_n\}$ be a sequence of random variables such that $\sqrt{n}(X_n - \mu) \rightarrow_d N(0, \sigma^2)$, and if g is a function which is differentiable at μ, then*

$$\sqrt{n}[g(X_n) - g(\mu)] \rightarrow_d N(0, g'(\mu)^2\sigma^2).$$

Proof: The Taylor expansion of $g(X_n)$ around $X_n = \mu$ is

$$g(X_n) = g(\mu) + g'(\mu)(X_n - \mu) + o_p(n^{-1}).$$

as $X_n \rightarrow_p \mu$. Applying the Slutsky's theorem to

$$\sqrt{n}[g(X_n) - g(\mu)] = g'(\mu)\sqrt{n}(X_n - \mu),$$

where we know that $\sqrt{n}(X_n - \mu) \rightarrow N(0, \sigma^2)$, then

$$\sqrt{n}[g(X_n) - g(\mu)] = g'(\mu)\sqrt{n}(X_n - \mu) \rightarrow N(0, g'(\mu)^2\sigma^2).$$

For example, let $g(X_n) = 1/X_n$, and $\sqrt{n}(X_n - \mu) \rightarrow_d N(0, \sigma^2)$, then we have $\sqrt{n}(1/X_n - 1/\mu) \rightarrow_d N(0, \sigma^2/\mu^4)$.

Lindeberg-Levy CLT assumes *i.i.d.*, which is too strong in practice. Now we retain the assumption of independence but allow heterogeneous distributions (*i.ni.d*), and in the next section, we will show versions of CLT for serial dependent sequence.

In the following analysis, it is more convenient to work with normalized variables. We also need to use triangular arrays in the analysis. An array X_{nt} is a double-indexed collection of numbers and each sample size n can be associated with a different sequence. We use $\{\{X_{nt}\}_{t=0}^{n}\}_{n=1}^{\infty}$, or just $\{X_{nt}\}$ to denote an array. Let $\{Y_t\}$ be the sequence of the raw sequence with $E(Y_t) = \mu_t$. Define $s_n^2 = \sum_{t=1}^n E(Y_t - \mu_t)^2$, $\sigma_{nt}^2 = E(Y_t - \mu_t)^2/s_n^2$, and

$$X_{nt} = \frac{Y_t - \mu_t}{s_n}.$$

Then $E(X_{nt}) = 0$ and $Var(X_{nt}) = \sigma_{nt}^2$. Define

$$S_n = \sum_{t=1}^{n} X_{nt},$$

then $E(S_n) = 0$ and

$$E(S_n^2) = \sum_{t=1}^{n} \sigma_{nt}^2 = 1. \tag{1}$$

Definition 7 (Lindeberg CLT) *Let the array $\{X_{nt}\}$ be independent with zero mean and variance sequence $\{\sigma_{nt}^2\}$ satisfying (1). If the following condition holds,*

$$\lim_{n\to\infty} \sum_{t=1}^{n} \int_{\{|X_{nt}|>\epsilon\}} X_{nt}^2 dP = 0 \quad \text{for all} \quad \epsilon > 0, \tag{2}$$

then $S_n \to_d N(0,1)$.

Equation (2) is known as the *Lindeberg condition.* What Lindeberg condition rules out are the cases where some sequences exhibit extreme behavior as to influence the distribution of the sum in the limit. Only finite variances are not sufficient to rule out these kind of situations with non-identically distributed observations. The following is a popular version of the CLT for independent processes.

Definition 8 (Liapunov CLT) *A sufficient condition for Lindeberg condition (2) is*

$$\lim_{n\to\infty} \sum_{t=1}^{n} E|X_{nt}|^{2+\delta} = 0, \quad \text{for some} \quad \delta > 0 \tag{3}$$

(3) is known as *Liapunov condition.* It is stronger than Lindeberg condition, but it is more easily checkable. Therefore it is more frequently used in practice.

2 Limit Theorems for Serially Dependent Observations

We have seen that if the data $\{X_n\}$ are generated by an ARMA process, then the observations are not *i.i.d*, but serially correlated. In this section, we will discuss how to derive asymptotic theories for stationary and serially dependent process.

2.1 LLN for a Covariance Stationary Process

Consider a covariance stationary process $\{X_n\}$. Without loss of generality, let $E(X_n) = 0$, so $E(X_t X_{t-h}) = \gamma(h)$, where $\sum_{h=0}^{\infty} |\gamma(h)| < \infty$. Now we will consider the the properties of the sample mean: $\bar{X}_n = (X_1 + \ldots + X_n)/n$. First we see that it is an unbiased estimate for the population mean, $E(\bar{X}_n) = E(X_t) = 0$. Next, the variance of this estimate is:

$$\begin{aligned}
E(\bar{X}_n^2) &= E[(X_1 + \ldots + X_n)/n]^2 \\
&= (1/n^2)E(X_1 + \ldots + X_n)^2 \\
&= (1/n^2)\sum_{i,j=1}^{n} E(X_i X_j) \\
&= (1/n^2)\sum_{i,j=1}^{n} \gamma_x(i-j) \\
&= (1/n)\left(\gamma_0 + 2\sum_{h=1}^{n-1}\left(1 - \frac{h}{n}\right)\gamma(h)\right) \\
&\ \text{or} \\
&= (1/n)\sum_{|h|<n}(1 - n^{-1}|h|)\gamma(h)
\end{aligned}$$

First we can see that

$$\begin{aligned}
nE(\bar{X}_n^2) &= \gamma_0 + 2\sum_{h=1}^{n-1}\left(1 - \frac{h}{n}\right)\gamma(h) \\
&= \gamma(0) + \left(1 - \frac{1}{n}\right)2\gamma(2) + \left(1 - \frac{2}{n}\right)2\gamma(3) + \ldots + \left(1 - \frac{m}{n}\right)2\gamma(m) + \ldots \\
&\leq |\gamma(0)| + 2|\gamma(1)| + 2|\gamma(2)| + \ldots < \infty
\end{aligned}$$

by our assumption on the absolute summability of $\gamma(h)$. Now $nE(\bar{X}_n^2)$ is bounded, then we know that $E(\bar{X}_n^2) \to 0$, which means that $\bar{X}_n \to_{m.s.} 0$, the population mean.

Next, we consider the limit of $nE(\bar{X}_n^2) = \gamma(0) + 2\sum_{h=1}^{n}\left(1 - \frac{h}{n}\right)\gamma(h)$. First we know that if a series is summable, then its tails must go to zero. So with large h, those autocovariance does not affect the sum; and with small h, the weight approaches 1 when $n \to \infty$. Therefore, we have

$$\lim_{n\to\infty} nE(\bar{X}_n^2) = \sum_{h=-\infty}^{\infty} \gamma(h) = \gamma(0) + 2\gamma(1) + 2\gamma(2) + \ldots$$

We summarize out results in the following proposition

Proposition 12 (LLN for covariance stationary process) *Let X_t be a zero-mean covariance stationary process with $E(X_t X_{t-h}) = \gamma(h)$ and absolutely summable autocovariances, then the sample mean satisfies $\bar{X}_n \to 0$ and $lim_{n\to\infty} nE(\bar{X}_n^2) = \sum_{h=-\infty}^{\infty} \gamma(h)$.*

If the process has population mean μ, then accordingly we have $\bar{X}_n \to \mu$ and the limit of $nE(\bar{X}_n^2)$ remain the same. A covariance stationary process is said to *ergodic for the mean* if the time series average converges to the population mean. Similarly, if the sample average provides an consistent estimate for the second moment, then the process is said to be *ergodic for the second moment*. In this section, we see that a sufficient condition for a covariance stationary process to be ergodic for the mean is that $\sum_{h=0}^{\infty} |\gamma(h)| < \infty$. Further, if the process is Gaussian, then absolute summable autocovariances also ensure that the process is ergodic for all moments.

Recall that in spectrum analysis, we have

$$\sum_{h=-\infty}^{\infty} \gamma_x(h) = 2\pi S_x(0),$$

therefore the limit of $nE(\bar{X}_n^2)$ can be equivalently expressed as $2\pi S_x(0)$.

2.2 Ergodic Theorem*

Ergodic theorem is a law of large number for a strictly stationary and ergodic process. We need a few concepts to define ergodic stationarity, and those concepts can be found in the appendix. Given a probability space $(\Omega, \mathcal{F}, P)$, an event $E \in \mathcal{F}$ is *invariant* under transformation T if $E = T^{-1}E$. Now, a measure-preserving transformation T is *ergodic* if for any invariant event E, we have $P(E) = 1$ or $P(E) = 0$. In other words, events that are invariant under ergodic transformations either occur almost surely, or do not occur almost surely. Let T be a shift operator, then a strictly stationary process $\{X_t\}$ is said to be ergodic if $X_t = T^{t-1}X_1$ for any t where T is measure-preserving and ergodic.

Below is an alternative way to define ergodicity,

Theorem 7 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{X_t\}$ be a strictly stationary process, $X_t(\omega) = X_1(T^{t-1}\omega)$. Then this process is ergodic if and only if for any pair of events A, $B \in \mathcal{F}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P(T^k A \cap B) = P(A)P(B). \tag{4}$$

To understand this result, if event A is not invariant and T is measure preserving, then $TA \cap A^c$ is not empty. Therefore repeated iterations of the transformation generate a sequence of sets $\{T^k A\}$ containing different mixtures of the elements of A and A^c. A positive dependence of B on A implies a negative dependence of B on A^c, i.e.

$$P(A \cap B) > P(A)P(B) \Rightarrow P(A^c \cap B) = P(B) - P(A \cap B) < P(B) - P(A)P(B) = P(A^c)P(B).$$

So the average dependence of B on a mixtures of A and A^c should tend to zero as $k \to \infty$.

Example 2 (Absence of ergodicity) Let $X_t = U_t + Z$, where $U_t \sim i.i.d.\text{Uniform}(0,1)$ and $Z \sim N(0,1)$. Then X_t is stationary, as each observation follows the same distribution. However, this process is not ergodic, because

$$X_t = U_t + Z = T^{t-1}U_1 + Z,$$

so Z is an invariant event under the shift operator. If we compute the autocovariance, $\gamma_X(h) = E(X_t X_{t+h}) = 1$, no matter how large h is. This means that the dependence is too persistent. Recall that in lecture one we have proposed that the time series average of a stationary converges to its population mean only when it is ergodic. In this example, the series is not ergodic. We can compute that the true expectation of the process is 1/2, while the sample average $\bar{X}_n = (1/n)\sum_{t=1}^n U_t + Z$ does not converge to 1/2, but to $Z + 1/2$.

In Example 2 we can see that in order for X_t to be ergodic, Z has to be a constant almost surely. In practice, ergodicity is usually assumed theoretically, and it is impossible to test it empirically. If a process is stationary and ergodic, we have the following LLN:

Theorem 8 (Ergodic theorem) *Let X_t be a strictly stationary and ergodic process and $E(X_t) = \mu$, then*

$$\bar{X}_n = \sum_{t=1}^n X_t \to_{a.s.} \mu.$$

Recall that when a process is strictly stationary, then a measurable function of this process is also strictly stationary. Similar property holds for ergodicity. Also, if the process is ergodic stationary, then all its moment, given that they exist and are finite, can also be consistently estimated by the sample moment. For instance, if X_t and strictly stationary and ergodic, $E(X_t^2) = \sigma^2$, then $(1/n)\sum_{t=1}^n X_t^2 \to \sigma^2$.

2.3 Mixing Sequences*

Application of ergodic theorem is restricted in applications since it requires strict stationary, which is a too strong assumption in many cases. Now, we introduce another condition on dependence: mixing.

A mixing transformation T implies that repeated application of T to event A mix up A and A^c, so that when k is large, T^kA provides no information about the original event A. A classical example about 'mixing' is due to Halmos (1956) (draw a picture here).

Consider that to make a dry martini, we pour a layer of vermouth (10% of the volume) on top of the gin (90% of the volume). let G denote the gin, and F an arbitrary small region of the fluid, so that $F \cap G$ is the gin contained in F. If $P(\cdot)$ denotes the volume of a set as a proportion of the whole, $P(G) = 0.9$. The proportion of gin in F, denoted by $P(F \cap G)/P(F)$ is initially either 0 or 1. Let T denote the operation of stirring the martini with a swizzle stick, so that $P(T^kF \cap G)/P(F)$ is the proportion of gin in F after k stirs. If the stirring mixes the martini we would expect the proportion of gin in T^kF, which is $P(T^kF \cap G)/P(F)$ tends to $P(G)$, so that each region F of the martini eventually contains 90% gin.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{G}$, $\mathcal{H}$ be σ subfields of $\mathcal{F}$, define

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G)P(H)|, \tag{5}$$

and

$$\phi(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}; P(G) > 0} |P(H|G) - P(H)|. \tag{6}$$

Clearly, $\alpha(\mathcal{G}, \mathcal{H}) \leq \phi(\mathcal{G}, \mathcal{H})$. The events in $\mathcal{G}$ and $\mathcal{H}$ are independent iff α and ϕ are zero.

For a sequence, $\{X_t\}_{-\infty}^{\infty}$, let $\mathcal{F}_{-\infty}^t = \sigma(...X_{t-1}, X_t)$, $\mathcal{F}_{t+m}^{\infty} = \sigma(X_{t+m}, X_{t+m+1}, ...)$. Define the strong mixing coefficient $\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^{\infty})$ and the uniform mixing coefficient to be $\phi_m = \sup_t \phi(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^{\infty})$.

Next, the sequence is said to be α-mixing or strong mixing if $\lim_{m\to\infty}\alpha_m = 0$ and it is said to be ϕ-mixing or uniform mixing if $\lim_{m\to\infty}\phi_m = 0$. Since $\alpha \leq \phi$, ϕ-mixing implies α-mixing.

A mixing sequence is not necessarily stationary, and it could be hetergeneous. However if a strictly stationary process is mixing, it must be ergodic. As you can see from (4), ergodicity implies 'average asymptotic independence'. However, ergodicity does not imply that any two parts will eventually become independent. On the other hand, a mixing sequence has this property (asymptotic independence). Hence mixing is a stronger condition than ergodicity. A stationary and ergodic sequence needs not be mixing.

We usually use a statistics called *size* to characterize the rate of convergence of α_m or ϕ_m. A sequence is said to be α-mixing of size $-\gamma_0$ if $\alpha_m = O(m^{-\gamma})$ for some $\gamma > \gamma_0$. If X_t is a α-mixing sequence of size $-\gamma_0$, and if $Y_t = g(X_t, X_{t-1}, \ldots, X_{t-k})$ is a measurable function and k be finite, then Y is also α-mixing of size $-\gamma_0$. All above statements can also be applied to ϕ-mixing.

When a sequence is stationary and mixing, then $Cov(X_1, X_m) \to 0$ as $m \to \infty$. Consider the ARMA processes. If it is MA(q), then the process must be mixing since any two events with time interval larger than q are independent, i.e., $\alpha(m) = \phi(m) = 0$ for $m > q$. We will not discuss sufficient conditions for a MA(∞) to be strong or uniform mixing, but note that if the innovations are *i.i.d.* Gaussian, then absolute summability of the moving average coefficients is sufficient to ensure strong mixing.

The following LLN (McLeish (1975)) applies to hetergeneous and temporarily dependent (mixing) sequences. We will only consider strong mixing.

Proposition 13 *(LLN for heterogeneous mixing sequences) Let $\{X_t\}$ be strong mixing with size $-r/(r-1)$ for some $r > 1$, with finite means $\mu_t = E(X_t)$. If for some δ, $0 < \delta \leq r$,*

$$\sum_{t=1}^{\infty} \left(\frac{E|Z_t - \mu_t|^{r+\delta}}{t^{r+\delta}} \right)^{1/r} < \infty, \tag{7}$$

then $\bar{X}_n - \bar{\mu}_n \to_{a.s.} 0$.

2.4 Martingale, Martingale Difference Sequence, and Mixingale

In time series observations, we know the past but we do not know the future. Therefore, a very important way in time series modeling is to condition sequentially on past events. In a probability space $(\Omega, \mathcal{F}, P)$, we characterize partial knowledge by specifying a σ-subfield of events from $\mathcal{F}$, for which it is known whether each of the events belonging to it has occurred or not. The accumulation of information over time is represented by an increasing sequence of σ-field, $\{\mathcal{F}\}_{-\infty}^{\infty}$, with $\ldots \subseteq \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}$. Here the set $\mathcal{F}$ has also been referred as the universal information set. If X_t is known given $\mathcal{F}_t$ for each t, then $\{\mathcal{F}_t\}_{-\infty}^{\infty}$ is said to be *adapted* to the sequence $\{X_t\}_{-\infty}^{\infty}$. The pair $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ are called an adapted sequence. Setting $\mathcal{F}_t = \sigma(X_s, -\infty < s \leq t)$, i.e.,$\mathcal{F}_t$ generated by all lagged observations of X, we obtain the minimum adapted sequence. And $\mathcal{F}_t$ defined in this way is also known as the *natural filtration.*

Given an adapted sequence $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$, if we have

$$E|X_t| < \infty,$$

$$E(X_t|\mathcal{F}_{t-1}) = X_{t-1},$$

for all t, then the sequence is called a *martingale.* A simple example of martingale is a random walk.

Example 3 (Random walk) Let

$$X_t = X_{t-1} + \epsilon_t, X_0 = 0, \epsilon_t \sim i.i.d.(0, \sigma^2)$$

$$\mathcal{F}_t = \{\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_1\}$$

Then we know that X_t is a martingale as $E|X_t| \leq \sum_{k=1}^t E|\epsilon_k| < \infty$ and $E(X_t|\mathcal{F}_{t-1}) = X_{t-1}$.

Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an adapted sequence, two concepts that are related to martingales are *submartingales*, which means $E(X_{t+1}|\mathcal{F}_t) \geq X_t$ and *supermartingales*, which means $E(X_{t+1}|\mathcal{F}_t) \leq X_t$.

A sequence $\{Z_t\}$ is known as a *martingale difference sequence* if $E(Z_t|\mathcal{F}_{t-1}) = 0$. As you can see, a mds can be constructed using martingales. For example, let $Z_t = X_t - X_{t-1}$ where $\{X_t\}$ is a martingale. Then this sequence of Z_t is an mds. On the other hand, the sum of mds is a martingale, i.e., $\{X_t\}$ will be a martingale if $X_t = \sum_{i=1}^t Z_i$ where Z_i is an mds.

Proposition 14 *If X_t is an mds, then $E(X_t X_{t-h}) = 0$ for all t and $h \neq 0$.*

Proof: $E(X_t X_{t-h}) = E(E_{t-h}(X_t X_{t-h})) = E(X_{t-h} E_{t-h}(X_t)) = 0$.

Remark: 1. mds is a stronger condition than being serially uncorrelated. If X_t is an mds, then we cannot forecast X_t as a linear or nonlinear function of its past realizations. 2. mds is a weaker condition than independence, since it does not rule out the possibility that higher moments such as $E(X_t^2|\mathcal{F}_{t-1})$ depends on lagged value of X_t.

Example 4 (mds but not independent) Let $\epsilon_t \sim i.i.d.(0, \sigma^2)$, then $X_t = \epsilon_t \epsilon_{t-1}$ is a an mds but not serially independent.

Another example is Garch model. In Garch model, the error terms are mds, but the variance of the error depends on past values. Although mds is weaker than independence, it behaves in many ways just like independent sequence. In cases where independence is violated, if the sequence is an mds, then we will find that many asymptotic results which hold for independent sequence also hold for mds.

One of the fundamental results in martingale theory is the martingale convergence theorem.

Theorem 9 (Martingale convergence theorem) *If $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is an L_1-bounded submartingale, then $X_n \to_{a.s.} X$ where $E|X| < \infty$. Further, let $1 < p < \infty$. If $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is a martingale and $\sup_t E|X_t|^p < \infty$, then X_t converges in L_p as well as with probability one.*

This is an existence theorem and it tells that X_n converges to X, but it does not tell what X is. But martingale convergence theorem (MGCT) is still a very powerful result.

Example 5 (LLN for heterogeneous mds) Let $\epsilon_t \sim mds(0, \sigma_t^2)$ with $\sup_t \sigma_t^2 = M < \infty$. Define $S_n = \sum_{t=1}^n \epsilon_t/t$, then S_n is a martingale with $E(S_n^2) = \sum_{t=1}^n \sigma_t^2/t^2$. Verify that $\sup_n E(|S_n|^2) \leq \sup_t \sigma_t^2 (\sum_{t=1}^n (1/t^2)) < \infty$. Therefore, $S_n = \sum_{t=1}^n \epsilon_t/t$ converges by MGCT. Next, let $b_n = n$, then by Kronecker's lemma

$$\frac{1}{n}\sum_{t=1}^n t\frac{\epsilon_t}{t} = \frac{1}{n}\sum_{t=1}^n \epsilon_t \to 0.$$

A concept similar to martingale (mds) is mixingales, which can be regarded as asymptotic martingales. A sequence of random variables $\{X_t\}$ with $E(X_t) = 0$ is called a L^p mixingale ($p \geq 1$) with respect to $\{\mathcal{F}_t\}$ if for sequence of nonnegative constants c_t and ξ_m, where $\xi_m \to 0$ as $m \to \infty$, we have

$$\|E(X_t|\mathcal{F}_{t-m})\|_p \leq c_t\xi_m \tag{8}$$

$$\|X_t - E(X_t|\mathcal{F}_{t+m})\|_p \leq c_t\xi_{m+1} \tag{9}$$

for all $t \geq 1$ and $m \geq 0$. Intuitively, mixingale captures the idea that the sequence $\{\mathcal{F}_s\}$ contains progressively more information about X_t as s increases. In the remote past nothing is known according to (8), or any past event eventually became useless in predicting event that will happen today (t). While in the future, everything will eventually be known according to (9). When X_t is $\mathcal{F}_t$-measurable, as in most of the cases we will be interested in, condition (9) always holds (since $E(X_t|\mathcal{F}_{t+m}) = X_t$). So to test if a sequence is mixingale, in many cases we only need to test condition (8). In what follows, we will mostly use L_1-mixingale. Condition (8) can then be written as

$$E|E(X_t|\mathcal{F}_{t-m})| \leq c_t\xi_m. \tag{10}$$

As you can see, mixingales are even more general than mds, in fact, a mds is a special kind of mixingale and you can set $c_t = E|X_t|$ and set $\xi_0 = 1$ and $\xi_m = 0$ for $m \geq 1$.

Example 6 Consider a two-sided MA(∞) process,

$$X_t = \sum_{j=-\infty}^{\infty} \theta_j \epsilon_{t-j},$$

where ϵ_t is an mds with $E(|\epsilon_t|) < \infty$. Then

$$E(X_t|\mathcal{F}_{t-m}) = \sum_{j=m}^{\infty} \theta_j \epsilon_{t-j}.$$

Take $c_t = \sup_t E|\epsilon_t|$ and take $\xi_m = \sum_{j=m}^{\infty} |\theta_j|$. Then if the moving average coefficients are absolutely summable, i.e., $\sum_{j=-\infty}^{\infty} |\theta_j| < \infty$, then its tails has to go to zero, i.e., $\xi_m \to 0$. Then condition (10) is satisfied and X_t is an L_1-mixingale.

In this example, first, we specify an MA process as generated by mds errors, which is a more generalized class of stochastic processes than *i.i.d* and white noise. Second, if $E(|\epsilon_t|) < \infty$ (which controls the tails of ϵ_t), then the condition of absolutely summable coefficients makes X_t a L_1-mixingale.

2.5 Law of Large Numbers for L_1-Mixingales

To derive the law of large numbers for L_1-mixingales, we need the notion of *uniformly integrable.*

Definition 9 (Uniformly integrable sequence) *A sequence $\{X_t\}$ is said to be uniformly integrable if for every $\epsilon > 0$ there exists a number $c > 0$ for all t such that*

$$E(|X_t|\mathbf{1}_{[c,\infty)}(|X_t|) < \epsilon$$

We will see how to make use of this notion in a moment. First, we introduce the following two conditions for uniform integrability.

Proposition 15 (Conditions for uniform integrability) *(a) A sequence $\{X_t\}$ is uniformly integrable if there exits an $r > 1$ and an $M < \infty$ such that $E(|X_t|^r) < M$ for all t. (b) Let $\{X_t\}$ be a uniformly integrable sequence and if $Y_t = \sum_{k=-\infty}^{\infty} \theta_k X_{t-k}$ with $\sum_{k=-\infty}^{\infty} |\theta_k| < \infty$, then the sequence $\{Y_t\}$ is also uniformly integrable.*

To derive inference for a uniformly integrable sequence, we have the following proposition.

Proposition 16 (Law of large numbers for L_1-mixingale) *Let $\{X_t\}$ be an L_1-mixingale. If $\{X_t\}$ is uniformly integrable and there exists a sequence of $\{c_t\}$ such that*

$$\lim_{n\to\infty} (1/n) \sum_{t=1}^{n} c_t < \infty,$$

then $\bar{X}_n = (1/n) \sum_{t=1}^{n} X_t \to_p 0$.

Example 7 (LLN for mds with finite variance) Let $\{X_t\}$ be a mds with $E|X_t|^2 = M < \infty$, then it is uniformly integrable and we can take $c_t = M$, and since $(1/n) \sum_{t=1}^{n} c_t = M < \infty$, by proposition 16, $\bar{X}_n \to_p 0$.

We can naturally generalize mixingale sequence to mixingale arrays. An array $\{X_{nt}\}$ is said to be L_1 mixingale with respect to $\{\mathcal{F}_{nt}\}$ if there exists nonnegative constant constants $\{c_{nt}\}$ and non-negative sequence $\{\xi_m\}$ such that $\xi_m \to 0$ as $m \to$ and

$$\|E(X_{nt}|\mathcal{F}_{n,t-m})\|_p \le c_{nt}\xi_m \tag{11}$$

$$\|X_{nt} - E(X_{nt}|\mathcal{F}_{n,t+m})\|_p \le c_{nt}\xi_{m+1} \tag{12}$$

for all $t \ge 1$ and $m \ge 0$. If the array is uniformly integrable with $\lim_{n\to\infty}(1/n) \sum_{t=1}^{n} c_{nt} < \infty$, then $\bar{X}_n = (1/n) \sum_{t=1}^{n} X_{nt} \to_p 0$.

Example 8 Let $\{\epsilon_t\}_{t=1}^{\infty}$ be an mds with $E|\epsilon|^r < M$ for some $r > 1$ and $M < \infty$ (i.e. ϵ_t is L_r*-bounded*). Let $X_{nt} = (t/n)\epsilon_t$. Then $\{X_{nt}\}$ is a uniformly integrable L_1-mixingale with $c_{nt} = \sup_t E|\epsilon_t|$, $\xi_0 = 1$ and $\xi_m = 0$ for $m > 0$. Then applying LLN for L_1-mixingales, we have $\bar{X}_n \to 0$.

2.6 Consistent Estimate of Second Moment

In this section, we will show how to prove the consistency of the estimate of second moments using the LLN of L_1-mixingales. There are two steps in the proof: first, we need to construct an L_1-mixingales; second, we need to verify that the conditions for applying the LLN is satisfied. This kind of methodology is very useful in many applications. Out following proof can also be found on page 192-192 in Hamilton.

First, we want to construct a mixingale. Out problem is outlined as follows. Let $X_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$, where $\sum_{j=0}^{\infty} |\theta_j| < \infty$ and ϵ_t is *i.i.d.* with $E|\epsilon_t|^r < \infty$ for some $r > 2$. We what to prove that

$$(1/n) \sum_{t=1}^{n} X_t X_{t-k} \to_p E(X_t X_{t-k}).$$

Define $X_{tk} = X_t X_{t-k} - E(X_t X_{t-k})$, then we want to show that X_{tk} is an L_1-mixingale.

$$\begin{aligned} X_t X_{t-k} &= \left(\sum_{i=0}^{\infty} \theta_i \epsilon_{t-i}\right)\left(\sum_{j=0}^{\infty} \theta_j \epsilon_{t-k-j}\right) \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \theta_i \theta_j \epsilon_{t-i}\epsilon_{t-k-j} \end{aligned}$$

$$\begin{aligned} E(X_t X_{t-k}) &= E\left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \theta_i \theta_j \epsilon_{t-i}\epsilon_{t-k-j}\right) \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \theta_i \theta_j E(\epsilon_{t-i}\epsilon_{t-k-j}) \end{aligned}$$

then

$$X_{tk} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \theta_i \theta_j (\epsilon_{t-i}\epsilon_{t-k-j} - E(\epsilon_{t-i}\epsilon_{t-k-j})).$$

Let $\mathcal{F}_t = \{\epsilon_t, \epsilon_{t-1}, \ldots\}$, then

$$E(X_{tk}|\mathcal{F}_{t-m}) = \sum_{i=m}^{\infty}\sum_{j=m-k}^{\infty} \theta_i \theta_j (\epsilon_{t-i}\epsilon_{t-k-j} - E(\epsilon_{t-i}\epsilon_{t-k-j})).$$

Now, we want to find c_t and ξ_m so that condition (10) holds.

$$\begin{aligned} E\left|E(X_{tk}|\mathcal{F}_{t-m})\right| &= E\left|\sum_{i=m}^{\infty}\sum_{j=m-k}^{\infty} \theta_i \theta_j (\epsilon_{t-i}\epsilon_{t-k-j} - E(\epsilon_{t-i}\epsilon_{t-k-j}))\right| \\ &\leq E\left(\sum_{i=m}^{\infty}\sum_{j=m-k}^{\infty} |\theta_i \theta_j| |\epsilon_{t-i}\epsilon_{t-k-j} - E(\epsilon_{t-i}\epsilon_{t-k-j})|\right) \\ &\leq \sum_{i=m}^{\infty}\sum_{j=m-k}^{\infty} |\theta_i \theta_j| M \end{aligned}$$

for some $M < \infty$. We take $c_t = M$ and

$$\xi_m = \sum_{i=m}^{\infty}\sum_{j=m-k}^{\infty} |\theta_i \theta_j| = \sum_{i=m}^{\infty} |\theta_i| \sum_{j=m-k}^{\infty} |\theta_j|.$$

Since θ_j is absolutely summable, its tails goes to zero, i.e., $\sum_{i=m}^{\infty} \theta_i \to 0$ as $m \to 0$, therefore, $\xi_m \to 0$.

Now, we have shown that X_{tk} is an L_1-mixingale. Next, we want to show that it is uniformly integrable and $(1/n)\sum_{t=1}^{n} c_t < \infty$. Since $c_t = M < \infty$, this latter condition holds. The uniform integrability can also be easily verified using the second part (b) of proposition 15. Therefore, applying the LLN, we have

$$(1/n)\sum_{t=1}^{n} X_{tk} = (1/n)\sum_{t=1}^{n}(X_t X_{t-k} - E(X_t X_{t-k})) \to_p 0,$$

therefore,

$$(1/n)\sum_{t=1}^{n} X_t X_{t-k} \to_p E(X_t X_{t-k}). \tag{13}$$

2.7 Central Limit Theorem for Martingale Difference Sequence

We have already learned several versions of CLT: (1) CLT for independently identically distributed sequence (Lindeberg-Levy CLT), (2) CLT for independently non-identically distributed sequence (Lindeberg CLT, Liapunov CLT). Now, we will consider the conditions for CLT to hold for a martingale difference sequence. Actually we can have CLT for any stationary ergodic mds with finite variance:

Proposition 17 *Let $\{X_t\}$ be stationary and ergodic martingale difference sequences with $E(X_t^2) = \sigma^2 < \infty$, then*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t \to N(0, \sigma^2). \tag{14}$$

Let $S_n = S_{n-1} + X_n$ with $E(S_n) = 0$, which is a martingale with stationary and ergodic differences, then from the above proposition we can have $n^{-1/2}S_n \to N(0, \sigma^2)$.

The conditions in the following version of CLT is usually easy to check in applications:

Proposition 18 (Central Limit Theorem for mds) *Let $\{X_t\}$ be a mds with $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$. Suppose that (a) $E(X_t^2) = \sigma_t^2 > 0$ with $n^{-1}\sum_{t=1}^{n}\sigma_t^2 \to \sigma^2 > 0$, (b) $E|X_t|^r < \infty$ for some $r > 2$ and all t, and (c), $n^{-1}\sum_{t=1}^{n} X_t^2 \to_p \sigma^2$. Then $\sqrt{n}\bar{X}_n \to N(0, \sigma^2)$.*

Again, this proposition can be extended from sequence $\{X_t\}$ to mds array $\{X_{nt}\}$ with $E(X_{nt}^2) = \sigma_{nt}^2$. In our last example in this lecture, we will use the next proposition, which is also a very useful tool.

Proposition 19 *Let X_t be a strictly stationary process with $E(X_t^4) < \infty$. Let $Y_t = \sum_{j=0}^{\infty}\theta_j X_{t-j}$, where $\sum_{j=0}^{\infty}|\theta_j| < \infty$. Then Y_t is a strictly stationary process with $E|Y_t Y_s Y_i Y_j| < \infty$ for all t, s, i and j.*

Example 9 (Example 7.15 in Hamilton) Let $Y_t = \sum_{j=0}^{\infty}\theta_j\epsilon_{t-j}$ with $\sum_{j=0}^{\infty}|\theta_j| < \infty$, $\epsilon_t \sim iid(0, \sigma^2)$ and $E(\epsilon^4) < \infty$. Then we see that $E(Y_t) = 0$ and $E(Y_t^2) = \sigma^2\sum_{j=0}^{\infty}\theta_j^2$. Define $X_t = \epsilon_t Y_{t-k}$ for $k > 0$, then X_t is an mds with respect to $\{\epsilon_t, \epsilon_{t-1}, \ldots\}$, with $E(X_t^2) = \sigma^2 E(Y_t^2) = \sigma^4\sum_{j=0}^{\infty}\theta_j^2$ (so condition (a) in proposition 18 is satisfied), $E(X_t^4) = E(\epsilon_t^4 Y_{t-k}^4) = E(\epsilon^4)E(Y_t^4) < \infty$. Here $E(\epsilon_t^4) < \infty$ by assumption and $E(Y_t^4) < \infty$ by proposition 19. So condition (b) in proposition 18 is also satisfied, and the remaining condition we need to verify to apply CLT is condition (c),

$$(1/n)\sum_{t=1}^{n} X_t^2 \to_p E(X_t^2).$$

Write

$$\begin{aligned}(1/n)\sum_{t=1}^{n} X_t^2 &= (1/n)\sum_{t=1}^{n} \epsilon_t^2 Y_{t-k}^2 \\ &= (1/n)\sum_{t=1}^{n}(\epsilon_t^2-\sigma^2)Y_{t-k}^2 + (1/n)\sum_{t=1}^{n}\sigma^2 Y_{t-k}^2\end{aligned}$$

The first term is a normed sum of mds with finite variance To see this,

$$E_{t-1}[(\epsilon_t^2-\sigma^2)Y_{t-k}^2] = Y_{t-k}^2(E_{t-1}(\epsilon_t^2)-\sigma^2) = 0$$

and

$$E[(\epsilon_t^2-\sigma^2)^2 Y_{t-k}^4] = E(\epsilon_t^4-\sigma^4)E(Y_t^4) < \infty.$$

Then $(1/n)\sum_{t=1}^{n}(\epsilon_t^2-\sigma^2)Y_{t-k}^2 \to 0$ (example 7).

By (13), we have

$$(1/n)\sum_{t=1}^{n}\sigma^2 Y_{t-k}^2 \to_p \sigma^2 E(Y_t^2).$$

Therefore, we have

$$(1/n)\sum_{t=1}^{n} X_t^2 \to_p \sigma^2 E(Y_t^2).$$

Finally, by proposition 18, we have

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t \to_d N(0, E(X_t^2)) = N\left(0, \sigma^4\sum_{j=0}^{\infty}\theta_j^2\right).$$

2.8 Central limit theorem for serially correlated sequence

Finally we present a CLT for a serially correlated sequence.

Proposition 20 *Let*

$$X_t = \mu + \sum_{j=0}^{\infty} c_j\epsilon_{t-j}.$$

where ϵ_t is i.i.d. with $E(\epsilon_t^2) < \infty$ and $\sum_{j=0}^{\infty} j\cdot|c_j| < \infty$. Then

$$\sqrt{n}(\bar{X}_n-\mu) \to_d N(0, \sum_{h=-\infty}^{\infty}\gamma(h)).$$

To prove the results, we can use a tool known as BN Decomposition and Phillips-Solo Device. Let

$$u_t = C(L)\epsilon_t = \sum_{j=0}^{\infty} c_j\epsilon_{t-j}, \tag{15}$$

where (a) $\epsilon_t \sim iid(0, \sigma^2)$ and (b) $\sum_{j=0}^{\infty} j \cdot |c_j| < \infty$. The BN-decomposition tells that we could rewrite the lag operator as

$$C(L) = C(1) + (L-1)\tilde{C}(L)$$

where $C(1) = \sum_{j=0}^{\infty} c_j$, $\tilde{C}(L) = \sum_{j=0}^{\infty} \tilde{c}_j L^j$, and $\tilde{c}_j = \sum_{j+1}^{\infty} c_k$. Since we assume that $\sum_{j=0}^{\infty} j \cdot |c_j| < \infty$, we have $\sum_{j=0}^{\infty} |\tilde{c}_j| < \infty$. When $C(1) > 0$ (the assumption ensured that $C(1) < \infty$), we can rewrite u_t as

$$\begin{aligned} u_t &= (C(1) + (L-1)\tilde{C}(L))\epsilon_t \\ &= C(1)\epsilon_t - \tilde{C}(L)(\epsilon_t - \epsilon_{t-1}) \\ &= C(1)\epsilon_t - (\tilde{u}_t - \tilde{u}_{t-1}). \end{aligned}$$

For example, let $u_t = \epsilon_t + \theta\epsilon_{t-1}$, then it can be written as $u_t = (1+\theta)\epsilon_t - \theta(\epsilon_t - \epsilon_{t-1})$. In this case, $C(1) = 1 + \theta$, $\tilde{c}_0 = c_1 = \theta$, and $\tilde{u}_t = \theta\epsilon_t$.

Therefore,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t = C(1) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \epsilon_t - \frac{1}{\sqrt{n}} (\tilde{u}_n - \tilde{u}_0).$$

Clearly, $C(1)\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \to N(0, C(1)^2\sigma_\epsilon^2)$. The variance, denoted by $\lambda_u^2 = C(1)^2\sigma_\epsilon^2$, is called the *long run variance* of u_t. When u_t is i.i.d., then $c_0 = 1$ and $c_j = 0$ for $j > 0$. Hence $\tilde{c}_j = 0$, for $j \geq 0$. In that case, the variance and the long run variance are equal. But in general, they are different. Take MA(1) as another example. Write

$$u_t = \epsilon_t + \theta\epsilon_{t-1} = (1+\theta)\epsilon_t - \theta(\epsilon_t - \epsilon_{t-1}).$$

Hence for this process, $C(1) = 1 + \theta$, $\tilde{c}_0 = \theta$, and $\tilde{c}_j = 0$ for $j > 0$. Note that the variance of u_t is that $\gamma_0 = (1 + \theta^2)\sigma^2$ while the long run variance of u_t is $\lambda^2 = \sigma^2 C(1)^2 = (1+\theta)^2\sigma^2$.

Note that since $\tilde{c}_j$ is absolutely summable, then $\tilde{u}_n - \tilde{u}_0$ is bounded in probability, hence

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t = C(1) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \epsilon_t + o_p(1) \to N(0, C(1)^2\sigma_\epsilon^2). \tag{16}$$

You can verify that $\sum_{h=-\infty}^{\infty} \gamma_x(h) = \left(\sum_{j=0}^{\infty} c_j\right)^2 \sigma_\epsilon^2 = C(1)^2\sigma_\epsilon^2$.

This result also applies when ϵ_t is a martingale difference sequence satisfying certain moment conditions (Phillips and Solo 1992).

Readings: Hamilton (Ch. 7) Davidson (Part IV and Part V)

Appendix: Some concepts

Set theory is trivial when it has finite number of elements. When a set has infinite number of elements, how to measure its 'size' becomes an interesting problem. Let X denote a set we are interested in and we want to investigate the classes of its subsets. If X has n elements, then the total number of its subsets is 2^n, which could be huge when n is large. And if X includes infinite number of elements, specifying the classes of all its subsets is more difficult. Therefore, we need to introduce some notations for study of these subsets.

Definition 10 (σ-Field) *A σ-field $\mathcal{F}$ is a class of subsets of X satisfying*

(a) $X \in \mathcal{F}$.

(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(c) If $\{A_n, n \in \mathbb{N}\}$ is a sequence of $\mathcal{F}$-sets, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

So a σ-Field is closed under the operations of complementation and countable unions and intersections. The smallest σ-field for a set X is $\{X, \emptyset\}$. Let A be subset of X, the smallest σ-field that contains A is $\{X, A, A^c, \emptyset\}$. So given any set or a collection of sets, we can write down the smallest σ-field that contains it. Let $\mathcal{C}$ denote a collection of sets, then the smallest field containing $\mathcal{C}$ is called 'the σ-field generated by $\mathcal{C}$'.

A measure is a nonnegative countably additive set function and it associates a real number with a set.

Definition 11 (Measure) *Given a class $\mathcal{F}$ of a subsets of a set Ω, a measure $\mu : \mathcal{F} \mapsto \mathbb{R}$ is a function satisfying*

(a) $\mu(A) \geq 0$, for all $A \in \mathcal{F}$.

(b) $\mu(\emptyset) = 0$.

(c) For a countable collection $\{A_j \in \mathcal{F}, j \in \mathbb{N}\}$ with $A_j \cap A_l = \emptyset$ for $j \neq l$ and $\bigcup_j A_j \in \mathcal{F}$,

$$\mu\left(\bigcup_j A_j\right) = \sum_j \mu(A_j).$$

A *measurable space* is a pair $(\Omega, \mathcal{F})$ where Ω is any collection of objects, and $\mathcal{F}$ is a σ-field of subsets of Ω. Let $(\Omega, \mathcal{F})$ and $(\Psi, \mathcal{G})$ be two measurable spaces and let transformation T be a mapping $T : \Omega \mapsto \Psi$. T is said to be *measurable* if $T^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{G}$. The idea is that a measure μ defined on $(\Omega, \mathcal{F})$ can be mapped into $(\Psi, \mathcal{G})$. Every event $B \in \mathcal{G}$ is assigned a measure, denoted by ν, with $\nu(B) = \mu(T^{-1}(B))$.

When the set we are interested in is the real line, $\mathbb{R}$, the σ-field generated by the open sets is called the Borel set, denoted by $\mathcal{B}$. Let λ denote the Lebesgue measure, and it is the only measure on $\mathbb{R}$ with $\lambda((a, b]) = b - a$.

We usually use $(\Omega, \mathcal{F}, P)$ to denote a probability space. Ω is the *sample space*, the set of all the possible outcomes of the experiment, and each of the individual elements is denoted by ω. $\mathcal{F}$ is the σ-field of subsets of Ω. The event $A \in \mathcal{F}$ is said to have occurred if the outcome of the experiment is an element of A. A measure P is assigned to elements of $\mathcal{F}$ with $P(\Omega) = 1$, and $P(A)$ is the probability of A. For example, in an experiment of tossing a coin, we can define $\Omega = \{\text{head}, \text{tail}\}$, $\mathcal{F} = \{\emptyset, \{\text{head}\}, \{\text{tail}\}, \{\text{head}, \text{tail}\}\}$, and we can assign probability to each element in $\mathcal{F}$, $P(\emptyset) = 0, P(\{\text{head}\}) = 1/2, P(\{\text{tail}\}) = 1/2$, and $P(\{\text{head}, \text{tail}\}) = 1$. Formally, the probability measure is defined as

Definition 12 *A probability measure on a measurable space $(\Omega, \mathcal{F})$ is a set function $P :\ \mathcal{F} \mapsto [0, 1]$ satisfying axioms of probability:*

(a) $P(A) \geq 0$ *for all* $A \in \mathcal{F}$.

(b) $P(\Omega) = 1$.

(c) Countable additivity: for a disjoint collection $\{A_j \in \mathcal{F}, j \in \mathbb{N}\}$,

$$P\left(\bigcup_j A_j\right) = \sum_j P(A_j).$$

We can define a random variable in a probability space. If the mapping $X : \Omega \mapsto \mathbb{R}$ is $\mathcal{F}$-measurable then X is a real valued random variable on Ω. For example, if Ω is a discrete probability space, as in our example of tossing a coin, then any function $X : \Omega \mapsto \mathbb{R}$ is a random variable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, the transformation $T : \Omega \mapsto \Omega$ is *measure-preserving* if it is measurable and $P(A) = P(TA)$ for all $A \in \mathcal{F}$. A *shift transformation* T for a sequence $\{X_t(\omega)\}$ is defined by $X_t(T\omega) = X_{t+1}(\omega)$. So a shift transformation works like a lag operator. If the shift transformation T is measure-preserving, then the sequences $\{X_t\}_{t=1}^{\infty}$ and $\{X_{t+k}\}_{t=1}^{\infty}$ have the same joint distribution for every $k > 0$. Therefore we can see that when the shift transformation T is measure-preserving, the process is strictly stationary.