Lectures 5-6

This week we will cover two important theorems: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). You have probably heard about these two theorems before, at least in the context of i.i.d. data. Today I will re-state the theorems for you (and mention how they extend to our time series framework).

1 LLN and CLT for i.i.d. data

1.1 LLN

Definition 1: A sequence of real-valued random variables $\widehat{\mu}_T$ converges in probability to a constant μ if for any $\epsilon > 0$

$$P(|\widehat{\mu}_T - \mu| > \epsilon) \to 0$$

as $T \to \infty$. We denote convergence in probability as:

$$\widehat{\mu}_T \xrightarrow{p} \mu.$$

Theorem (Law of Large numbers for i.i.d. data): Let $\{X_1, X_2, \ldots\}$ be a collection of i.i.d. random variables each with distribution P and let $\mu \equiv E_P[X_t]$. If $E(|X_t|) < \infty$ then:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{p} \mu.$$

That is, the sample mean is 'consistent' for the population mean, $E_P[X_t]$.[1]

The result might look a bit abstract, but we will have a very concrete use for the LLN: approximation of population means.

Example 1: We have already computed the autocovariance function of the Gaussian MA(1) model: $X_t = \epsilon_t + \theta\epsilon_{t-1}, \epsilon_t$ i.i.d. $(0, \sigma^2)$. We have shown that $\gamma(0) = (1+\theta^2)\sigma^2$, $\gamma(1) = \theta\sigma^2$, and the rest of the function is zero. Since

$$\gamma(1) = \text{Cov}(X_t, X_{t+1}) = \mathbb{E}[X_t X_{t+1}] - \mathbb{E}[X_t]\mathbb{E}[X_{t+1}],$$

we could generate I draws from X_t, X_{t+1} using the Gaussian MA(1) model. Denote these draws as $X_t(i), X_{t+1}(i)$, respectively. The law of large numbers suggests that

$$\frac{1}{I}\sum_{i=1}^{I} X_t(i)X_{t+1}(i) \xrightarrow{p} \mathbb{E}[X_t X_{t+1}]$$

Example 2: Suppose that $X \sim \mathcal{N}(\theta, 1)$ for some $\theta \in \mathbb{R}$. How would you use Python to approximate the mean of $E_\theta[\exp(X)]$? If you are a Matlab user, run

[1] A proof of this result can be found on Theorem 22.1 of the book "Probability and Measure" [1995] by Patrick Billingsley.

the first two sections of the script `Example 1.m` to see how we can apply the LLN to approximate $E_\theta[\exp(X)]$?

1.2 CLT

Definition 2: A sequence of real-valued random variables $\widehat{\mu}_T$ converges in distribution to a random variable Z with distribution $\mathbb{P}$ if for any $x \in \mathbb{R}$:

$$P(\widehat{\mu}_T \leq x) \to \mathbb{P}(Z \leq x)$$

as $T \to \infty$. We denote convergence in distribution as:

$$\mu_T \xrightarrow{d} Z$$

Theorem (Central Limit Theorem for i.i.d data): Let $\{X_1, X_2, \ldots\}$ be a collection of i.i.d. random variables each with distribution P and let $\mu \equiv V_P[X_t] \equiv \sigma^2 < \infty$. Then:

$$\sqrt{T}\Big(\frac{1}{T}\sum_{t=1}^{T} X_t - \mu\Big) \xrightarrow{d} N(0, \sigma^2),$$

This means that the approximation error between the *sample mean* and the *population mean* is random, and approximately $N(0, \sigma^2/\sqrt{T})$.[2]

Thus, in a way, the Central Limit Theorem tell us how good (or bad) a LLN-type approximation with T i.i.d draws is. More specifically, the CLT usually allow us to conclude that

$$\widehat{\mu}_T \text{ approx } \mathcal{N}(\mu, \sigma^2).$$

This means that with 95% probability the population mean is in between

$$\frac{1}{T}\sum_{t=1}^{T} X_t \pm 2\frac{\sigma}{\sqrt{T}}$$

[2]The statement of this result can be found in Theorem 27.1 in Billingsley's "Probability and measure".

Example 1: We suggested to estimate $Cov(X_t, X_{t+1})$ as

$$\frac{1}{I}\sum_{i=1}^{I} X_t(i)X_{t+1}(i).$$

We know that the true covariance is 5. This means that by the CLT

$$\sqrt{I}\left(\frac{1}{I}\sum_{i=1}^{I} X_t(i)X_{t+1}(i) - 5\right)$$

is approximately a normal distribution. How would you check this in the computer?

Example 2: Run the last sections of the script `Example 1.m` to see how CLT is used to provide a confidence interval for the population mean.

2 LLN and CLT for weakly stationary data

2.1 LLN

The first question that we want to ask is whether the LLN also obtains in the presence of time series correlation. We make this argument for the MA(1) model so that you get some intuition. The generalization to MA(q) models is straightforward.

Consider the following MA(1) model with i.i.d. noise :

$$X_t = \mu + \theta_0 \epsilon_t + \theta_1 \epsilon_{t-1}, \quad E[\epsilon_t] = 0 \quad E[\epsilon_t^2] = \sigma^2.$$

The argument goes as follows:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} X_t - \mu &= \frac{1}{T}\sum_{t=1}^{T}(\mu + \theta_0\epsilon_t + \theta_0\epsilon_{t-1}) - \mu \\
&\quad \text{(using the MA(1) assumption)} \\
&= \frac{1}{T}\sum_{t=1}^{T}(\theta_0\epsilon_t + \theta_1\epsilon_{t-1}) \\
&= \theta_0 \frac{1}{T}\sum_{t=1}^{T}\epsilon_t + \theta_1 \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}
\end{aligned}$$

Therefore, for any $\epsilon > 0$

$$\begin{aligned}
P\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_t - \mu\right| > \epsilon\right) &= P\left(\left|\theta_0 \frac{1}{T}\sum_{t=1}^{T}\epsilon_t + \theta_1 \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right| > \epsilon\right) \\
&\leq P\left(\left|\theta_0 \frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right| + \left|\theta_1 \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right| > \epsilon\right) \\
&\quad \text{(since the triangle ineq. implies that whenever} \\
&\quad |x+y| > \epsilon \text{ then } |x| + |y| > \epsilon)
\end{aligned}$$

Moreover:

$$P\left(\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|+\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\epsilon\right)\leq P\left(\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|>\frac{\epsilon}{2}\text{ or }\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\frac{\epsilon}{2}\right)$$

As the only way in which

$$\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|+\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\epsilon$$

can happen is if either:

$$\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|>\frac{\epsilon}{2}\text{ or }\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\frac{\epsilon}{2}.$$

Finally, note that

$$P\left(\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|>\frac{\epsilon}{2}\text{ or }\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\frac{\epsilon}{2}\right)$$

is less than or equal:

$$P\left(\left|\theta_0\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\right|>\frac{\epsilon}{2}\right)+P\left(\left|\theta_1\frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}\right|>\frac{\epsilon}{2}\right),$$

As for any events A and B, $P(A\text{ or }B)$ is smaller than $P(A)+P(B)$. Since both of these probabilities become arbitrarily small as T grows large, then we conclude that:

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T}X_t-\mu\right|>\epsilon\right)\to 0.$$

Main Take Away from this algebra: The Law of Large numbers will hold even under some time series correlation.

Theorem (Laws of Large Numbers for MA(q) data): Suppose $\{X_t\}$ is an MA(q) process with i.i.d. noise ε_t ($\mathbb{E}[\varepsilon_t]=0,\mathbb{V}[\varepsilon_t]=\sigma^2$). Then:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} X_t &\xrightarrow{p} E[X_t] \\
\frac{1}{T}\sum_{t=1}^{T-h} X_{t+h}X_t &\xrightarrow{p} E[X_{t+h}X_t] \\
\frac{\frac{1}{T}\sum_{t=1}^{T-h} X_{t+h}X_t}{\frac{1}{T}\sum_{t=1}^{T} X_t^2} &\xrightarrow{p} \frac{E[X_{t+h}X_t]}{E[X_t^2]}
\end{aligned}
$$

This means that the "sample mean", "sample covariances", and "sample autocorrelations" of a MA(q) processes will be consistent for their "population" counterparts. Over the next lectures I will provide references for this result (for example, Theorems 3.4 and 3.7 in the Annals of Statistics paper of Peter Phillips and Victor Solo: "Asymptotics for Linear Processes").

2.2 CLT

Time series dependence does not really compromise the consistency property of sample averages of MA(q) processes. How about the accuracy of the approximation? We have already argued that if we have data:

$$\{X_1, X_2, \ldots, X_n\}$$

with no temporal dependence, then the CLT implies that:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \approx \mathcal{N}\left(E[X_t], \frac{\text{Var}(X_t)}{T}\right)$$

Do we get a similar approximation for the sample mean if X_t exhibits time series dependence?

To get some intuition for the main result, consider the Gaussian MA(1) model that we discussed before, but normalize $\theta_0 = 1$:

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Note that the standard i.i.d. Central Limit Theorem approximation would imply that:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \approx \mathcal{N}\left(0, \frac{(1+\theta_1^2)\sigma^2}{T}\right).$$

However:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} X_t &\sim \mu + \frac{1}{T}\sum_{t=1}^{T}(\epsilon_t + \theta_1 \epsilon_{t-1}) \\ &= \mu + \frac{1}{T}\sum_{t=1}^{T}\epsilon_t + \theta_1 \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1}. \end{aligned}$$

Since:

$$\begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T}\epsilon_t \\ \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t-1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{T}\begin{pmatrix} \sigma^2 & (1-(1/T))\sigma^2 \\ (1-(1/T))\sigma^2 & \sigma^2 \end{pmatrix} \right),$$

then:

$$\frac{1}{T}\sum_{t=1}^{T} X_t \sim \mathcal{N}\left(\mu, \frac{(1+\theta_1^2+2\theta_1)\sigma^2}{T} + o(1)\right).$$

This means that the approximation error for the MA(1) model will not coincide with the typical approximation error (and the difference between the standard i.i.d. approximation and the variance that shows up in the true distribution can be positive or negative). The intuition is that not only the variance of X_t influences the approximation error, but also the first-order covariance.

How general is this result? We can show—see section 2.4 in the book—that in general if X_t is an MA(q) process the sample mean is approximately distributed as:

$$\mathcal{N}\left(E[X_t], \frac{\text{Var}(X_t) + 2\sum_{j=1}^{q}\gamma(j)}{T}\right)$$

The term:

$$\text{Var}(X_t) + 2\sum_{j=1}^{q}\gamma(j),$$

which equals $\sum_{j=-\infty}^{\infty}\gamma(j)$ is referred to as the long-run variance of X_t.