Chapter 1

Time Series

1.1 Introduction

A random variable is a $(\Omega, \mathscr{F}) \setminus (\mathbb{R}^m, \mathscr{B})$ measure function. A time series is a sequence of random variables $(y_1(\omega), y_2(\omega), \ldots, y_n(\omega)) \in \mathbb{R}^{m \times n}$, and it can be extended to a doubly infinite sequence $(..., y_{t-1}, y_t, y_{t+1}, ...) \in \mathbb{R}^{m \times \infty}$. We consider discrete time series (instead of the continuous time series).

Example 1.1. For simplicity, consider $m = 1$ and $t = 1, 2, \ldots, n$. For each fixed ω, the sequence is a deterministic vector $(y_t(\omega))_{i=1}^n \in \mathbb{R}^n$; for each fixed t, the scalar $y_t(\omega)$ for $\omega \in \Omega$ is a plain random variable.

1.2 Stationarity

In reality, we have only one realized sequence, but statistics needs repeated observations. We introduce the concept *stationarity* to produce "repeated" observations.

Definition 1.1. (y_t) is **covariance stationarity** or **weakly stationarity** if the mean $\mu = E[y_t]$, covariance $\Sigma = E\left[(y_t - \mu)(y_t - \mu)'\right]$ and autocovariance $\Gamma(\ell) = E\left[(y_t - \mu)(y_{t-\ell} - \mu)'\right]$ are independent of t.

- For a vector-valued weakly stationarity time series, $\Sigma = \Gamma(0)$ is a positive-definite symmetric matrix. The autovariance $\Gamma(\ell)$ for $\ell \neq 0$ is not symmetric in general, and

$$\Gamma(-\ell) = E\left[(y_t - \mu)(y_{t+\ell} - \mu)'\right] = E\left[(y_{t-\ell} - \mu)(y_t - \mu)'\right] = \Gamma(\ell)'.$$

- When $m = 1$, we use $\gamma(0), \gamma(1), ...,$ for the autocovariance, and we define *autocorrelation* as $\rho(\ell) = \gamma(\ell)/\gamma(0)$. By the Cauchy-Schwarz inequality $\rho(\ell) \in [-1, 1]$.

Definition 1.2. (y_t) is **strictly stationarity**, if for every $\ell \in \mathbb{N}$, the joint distribution of $(y_t, y_{t+1}, \ldots, y_{t+\ell})$ is independent of t.

When one mentions "stationarity" without referring to a quantifier, in econometrics it means strictly stationarity by default.

- If (y_t) is i.i.d, it is a special case of strict stationarity.

- If (y_t) is strictly stationary, its transformation $x_t \in \phi(y_t, y_{t-1}, ...) \in \mathbb{R}^q$ is also strictly stationary. In other words, strict stationarity is preserved by transformation.

The infinite series x_t is **convergent** if the partial sum $\sum_{j=1}^{N} a_j y_{t-j}$ has a finite limit as $N \to \infty$ *almost surely.*

- If y_t is strictly stationarity, $E\|y\| < \infty$ and $\sum_{j=0}^{N} |a_j| < \infty$ (absolutely summable), then x_t is convergent and strictly stationarity.

1.3 Ergodicity

A time series (y_t) is **ergodic** if all invariant events are trivial.

Definition 1.3. Formal definitions

Let $\widetilde{y}_t = (..., y_{t-1}, y_t, y_{t+1}, ...)$, and the ℓ-th time shift is $\widetilde{y}_{t+\ell} = (..., y_{t-1+\ell}, y_{t+\ell}, y_{t+\ell+1}, ...)$.
Let an event $D \in \{\widetilde{y}_t \in G\}$ for some $G \subseteq \mathbb{R}^{m\times\infty}$, and a time shift of the event is $D_\ell \in \{\widetilde{y}_{t+\ell} \in G\}$.
An event is **invariant** if $D_\ell = D$ for all $\ell \in \mathbb{Z}$. An event is **trivial** if $P(D) = 0$ or $P(D) = 1$.

Ergodicty is preserved by transformation. If (y_t) is stationarity and erdogic, the same is for $x_t \in \phi(y_t, y_{t-1}, ...)$ (function with infinite terms).

Example 1.2. If $x_t = \sum_{j=0}^{\infty} a_j y_{t-j}$ is convergent and (y_t) is erdogic, then x_t is also erdogic.

Fact 1.1 (Cesaro mean)**.** *If $a_j \to a$ as $j \to \infty$, then $\frac{1}{n}\sum_{j=1}^{\infty} a_j \to a$ as $n \to \infty$.*

Theorem 1.1. *If $y_t \in \mathbb{R}^m$ is stationarity and erdogic, and var $(y_t) < \infty$, then $\frac{1}{n}\sum_{\ell=1}^{n} cov(y_t, y_{t+\ell}) \to 0$ as $n \to \infty$*

A stationarity (y_t) is erdogic if for all events A and B,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} P(A_\ell \cap B) = P(A)P(B)$$

Let $B = A$, and then we solve $P(A) = [P(A)]^2 \Rightarrow P(A) = 0$ or 1.

A "sufficient" condition for ergodicity is $P(A_\ell \cap B) \to P(A)P(B)$ as $\ell \to \infty$, according to Cesaro means. This sufficient condition is called "mixing".

- Mixing says that separate events (any A and B) are asymptotically independent when one of the event, say A, is shifted to A_ℓ as $\ell \to \infty$.
- Ergodicity is slightly weaker than mixing (weak dependence), in the sense that the independence is "on average" in the form of $\frac{1}{n}\sum_{\ell=1}^{n} P(A_\ell \cap B)$.

Theorem 1.2 (Ergodic Theorem)**.** *If $y_t \in \mathbb{R}^m$ is stationarity, erdogic, and $E\|y\| < \infty$, then $E\|\bar{y} - \mu\| \to 0$ and $\bar{y} \xrightarrow{p} \mu$.*

This is a version of LLN for time series.

1.4 Information Set

- For a univariate time series, definite $E_{t-1}[y_t] = E[y_t \mid y_{t-1}, y_{t-2}, ...]$ as the condition expectation of y_t given the past history $(y_{t-1}, y_{t-2}, ...)$
- More generally, we write $\mathscr{F}_t$ as the σ-field generated by some random variables (specified by the user) up to time t. $\mathscr{F}_t$ is called an **information set**. We can write $E_{t-1}[y_t] = E[y_t \mid \mathscr{F}_{t-1}]$.
- Information sets are nested: $\cdots \subseteq \mathscr{F}_{t-1} \subseteq \mathscr{F}_t \subseteq \mathscr{F}_{t+1} \subseteq ...$
- Information sets associate with the different generating variables are different. For example, in general $\sigma(y_t, y_{t-1}, ...) \neq \sigma(y_t, x_t, y_{t-1}, x_{t-1}, ...)$. The former is the information set for (y_t), whereas the latter is the information set for (y_t, x_t).

1.5 Martingale Difference Sequence (MDS)

- Let (e_t) be a time series, and $\mathscr{F}_t$ be an information set. We say (e_t) is **adapted** to $\mathscr{F}_t$ if $E[e_t \mid \mathscr{F}_t] = e_t$. It means that $\mathscr{F}_t$ contains the complete information of e_t. A **natural filtration** is $\mathscr{F}_t = \sigma(e_t, e_{t-1}, ...)$; it is the smallest information set to which (e_t) is adapted.

Definition 1.4 (MDS)**.** A process $\{e_t, \mathscr{F}_t\}$ is MDS if

1. e_t is adapted to $\mathscr{F}_t$
2. $E|e_t| < \infty$
3. $E[e_t \mid \mathscr{F}_{t-1}] = 0$

Interpretation: e_t is unforeseeable given the information $\mathscr{F}_{t-1}$. The definition of mds is about the mean independence. It does not rule our predictability in other moments.

Definition 1.5. A time series is a white noise if it is covariance stationarity with 0 autocovariance.

MDS implies that the series is a **white noise**, because

$$cov(e_t, e_{t-\ell}) = E[e_t e_{t-\ell}] = E[E[e_t e_{t-\ell} \mid \mathscr{F}_{t-1}]] = E[e_{t-\ell} E[e_t \mid \mathscr{F}_{t-1}]] = 0.$$

Example 1.3. Suppose $e_t = u_t u_{t-1}$, where $u_t \sim i.i.d.\, N(0,1)$. In this case, e_t is MDS. Consider the filtration $\mathscr{F}_t = \sigma(u_t, u_{t-1}, ...)$, which subsumes $\sigma(e_t, e_{t-1}, ...)$.

$$E[e_t|\mathscr{F}_{t-1}] = E[u_t u_{t-1}|\mathscr{F}_{t-1}] = u_{t-1} E[u_t|\mathscr{F}_{t-1}] = u_{t-1} \cdot 0 = 0.$$

On the other hand, the covariance of e_t^2 and e_{t-1}^2 is not 0:

$$\begin{aligned} cov\left(e_t^2, e_{t-1}^2\right) &= E\left[u_t^2 u_{t-2}^2 u_{t-1}^4\right] - E\left[u_t^2 u_{t-1}^2\right] E\left[u_{t-1}^2 u_{t-2}^2\right] \\ &= 1 \times 1 \times 3 - (1 \times 1)^2 = 2 \end{aligned}$$

as the kurtosis of $N(0,1)$ is 3. Therefore, (e_t) is an mds but not iid.

An MDS $(e_t, \mathscr{F}_t)$ is **conditional homoskedastic** if $E\left[e_t^2 \mid \mathscr{F}_{t-1}\right] = \sigma^2$. In the above example, $e_t = u_t u_{t-1}$ is MDS and stationary, but conditional heteroskedastic because

$$E\left[e_t^2|\mathscr{F}_{t-1}\right] = E\left[u_t^2 u_{t-1}^2|\mathscr{F}_{t-1}\right] = u_{t-1}^2 E\left[u_t^2|\mathscr{F}_{t-1}\right] = \sigma^2 u_{t-1}^2$$

varies over time.

In the real world, MDS is a benchmark model for the stock return. Indeed, mds is implied by the so-called *efficient market hypothesis*. On the other hand, empirical evidence shows that the conditional variance of stock return is much easier to predict. There are many models about conditional volatility, for example the well-known ARCH and GARCH models.

Theorem 1.3 (CLT for MDS)**.** *If (e_t) is strictly stationary, ergodic and MDS, then*

$$S_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} e_t \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = E\left[e_t e_t'\right]$.

There is the time series counterpart of the Lindeberg-Lévy CLT. Because (e_t) is strict stationary, its variance Σ must be a constant matrix. It does not rule out u_t being conditional heteroskedastic.

1.6 Mixing

MDS is useful, but too restrictive in that it rules out serial correlation. If we are reluctant to impose MDS, we will need stronger assumption on the dependence than ergodicity to establish large sample results.

We introduce some more definitions. The **alpha-coefficient for two events** is defined as

$$\alpha(A, B) = |P(AB) - P(A)P(B)|.$$

Denote two σ-fields be $\mathscr{F}_{-\infty}^{t} = \sigma(..., y_{t-1}, y_t)$ and $\mathscr{F}_t^{\infty} = \sigma(y_t, y_{t+1}, ...)$. The **strong mixing coefficient** (alpha-coefficient) is defined as

$$\alpha(\ell) = \sup_t \sup_{A \in \mathscr{F}_{-\infty}^{t-\ell}, B \in \mathscr{F}_t^{\infty}} \alpha(A, B)$$

We say (y_t) is **strong mixing** (alpha mixing) if $\alpha(\ell) \to 0$ as $\ell \to \infty$.

If the time series is strictly stationary, the definition of the alpha-coefficient can be simplified as $\alpha(\ell) = \sup_{A \in \mathscr{F}_{-\infty}^{t-\ell}, B \in \mathscr{F}_t^{\infty}} \alpha(A, B)$.

Fact 1.2. *An α-mixing process is ergodic.*

To use the α-mixing process, we usually need **rate conditions** (for example $\alpha(\ell) = O(\ell^{-r})$ gives the speed of decay) or **summation restriction** (for example $\sum_{\ell=0}^{\infty} [\alpha(\ell)]^r < \infty$ or $\sum_{\ell=0}^{\infty} \ell^s \alpha(\ell)^r < \infty$ for some positive constants s and r.)

Strong mixing is preserved by finite transformation.

Fact 1.3. *Suppose y_t has mixing coefficients $\alpha_y(\ell)$, and $x_t = \phi(y_t, y_{t-1}, ..., y_{t-q})$ is a finite transformation of (y_t). Then Then $\alpha_x(\ell) < \alpha_y(\ell - q)$ for $\ell \geq q$. The α-coefficients satisfy the same rate and summation properties.*

Another widely used measurement of dependence is **absolute regularity** (beta-coefficient)

$$\beta(\ell) = \sup_t \sup_{A \in \mathscr{F}_t^{\infty}} \left| P\left(A \mid \mathscr{F}_{-\infty}^{t-\ell}\right) - P(A) \right|.$$

β-mixing is stronger than α-mixing in that $\beta(\ell) \to 0$ implies $\alpha(\ell) \to 0$.

1.7 CLT for Correlated Variables

The scaled partial sum of a scalar zero-mean time series $(y_t)_{t=1}^n$ has variance

$$\begin{aligned}
var\left(S_n\right) &= var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} y_t\right)\\
&= \frac{1}{n}\mathbf{1}_n' E\left[YY'\right]\mathbf{1}_n\\
&= \frac{1}{n}\mathbf{1}_n'\begin{bmatrix} \sigma^2 & \gamma\left(1\right) & \gamma\left(2\right) & \cdots & \gamma\left(n-1\right)\\ \gamma\left(1\right) & \sigma^2 & \gamma\left(1\right) & \cdots & \gamma(n-2)\\ \gamma\left(2\right) & \gamma\left(1\right) & \sigma^2 & \cdots & \gamma(n-3)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \gamma\left(n-1\right) & \gamma(n-2) & \gamma(n-3) & \cdots & \sigma^2 \end{bmatrix}\mathbf{1}_n\\
&= \frac{1}{n}\left(n\sigma^2 + 2\left(n-1\right)\gamma\left(1\right) + 2\left(n-2\right)\gamma\left(2\right) + ... + 2\gamma\left(n-1\right) + 2\times 0\times\gamma\left(n\right)\right)\\
&= \sigma^2 + 2\sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\gamma\left(\ell\right)
\end{aligned}$$

Since $\gamma\left(-\ell\right) = \gamma\left(\ell\right)$, it is equivalent to write

$$var\left(S_n\right) = \sum_{\ell=-n}^{n}\left(1-\frac{|\ell|}{n}\right)\gamma\left(\ell\right).$$

If y_t is a vector time series, similarly

$$var\left(S_n\right) = \Gamma\left(0\right) + \sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\left(\Gamma\left(\ell\right)+\Gamma\left(\ell\right)'\right) = \sum_{\ell=-n}^{n}\left(1-\frac{|\ell|}{n}\right)\Gamma\left(\ell\right)$$

For any CLT to work, $var\left(S_n\right)$ must be convergent in the limit. Suppose $\sum_{\ell=1}^{\infty}\Gamma\left(\ell\right)$ is convergent, and then by the Cesaro mean the sum of the autocovariances is

$$\sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\Gamma\left(\ell\right) = \frac{1}{n}\sum_{\ell=1}^{n}\left(n-\ell\right)\Gamma\left(\ell\right) = \frac{1}{n}\sum_{\ell=1}^{n-1}\sum_{j=1}^{\ell}\Gamma\left(j\right)\to\sum_{j=1}^{\infty}\Gamma\left(j\right).$$

A necessary condition for $\sum_{\ell=1}^{\infty}\Gamma\left(\ell\right)$ to be convergent is that $\Gamma\left(\ell\right)\to 0$ as $\ell\to\infty$; and a sufficient condition is $\sum_{\ell=1}^{\infty}\left|\Gamma\left(\ell\right)\right| < \infty$.

Theorem 1.4. *(CLT) Suppose u_t is strictly stationarity with $E\left[u_t\right]=0$, and its α-mixing coefficient satisfies $\sum_{\ell=0}^{\infty}\alpha\left(\ell\right)^{1-2/r} < \infty$ and $E\left\|u_t\right\|^r < \infty$ for some $r > 2$. Then $S_n \xrightarrow{d} N\left(0,\Omega\right)$ where $\Omega = \sum_{\ell=-\infty}^{\infty}\Gamma\left(\ell\right)$ is the long-run variance.*

The most important difference from the iid CLT or the MDS CLT lies in the fact that if the summands have serial correlation ($\Gamma(\ell)\neq 0$ for some $\ell\neq 0$), here the asymptotic variance is no longer the variance of the individual variable $\Gamma(0)$, but the **long-run variance** Ω defined above.

1.8 Linear Projection

We have learned that in linear regression models

$$\mathscr{P}(y \mid X) = X\beta^* = X'\left(E\left[XX'\right]\right)^{-1} E\left[Xy\right].$$

Here $\mathscr{P}(\cdot|\cdot)$ is the linear projection operator (by default, an intercept is included in X).

Such a linear projection to a finite number of random variables can be extended to infinitely many variables. In our time series context, we are particularly interested in projecting y_t to the past history

$$\mathscr{P}_{t-1}(y_t) = \mathscr{P}(y_t \mid \tilde{y}_{t-1})$$

where $\tilde{y}_{t-1} := (y_{t-1}, y_{t-2}, ...)$. Define the projection error as $e_t := y_t - \mathscr{P}_{t-1}(y_t)$

Theorem 1.5 (Projection Theorem)**.** *If $y_t \in \mathbb{R}$ is covariance stationarity, then the projection error satisfies:*

(1) $E[e_t] = 0$
(2) $\sigma^2 = E\left[e_t^2\right] \leq E\left[y_t^2\right]$
(3) $E[e_t e_{t-j}] = 0$ for all $j \geq 1$.

Obviously, the covariance stationary $\{y_t\}$ produces covariance stationary $\{e_t\}$, and the latter is indeed a white noise. Furthermore, if $\{y_t\}$ is strictly stationarity, then $\{e_t\}$ is strictly stationarity. But recall that strictly stationary white noise does not implies temporal independence, as shown in Example 1.3.

To better understand the zero autocovariance of the residuals, it is helpful to imagine the projection as a linear combination

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + e_t = \alpha_0 + \sum_{k=1}^{\infty} \alpha_k y_{k-j} + e_t.$$

The nature of linear projection ensures e_t is uncorrelated with all regressions $(1, y_{t-1}, y_{t-2}, \ldots)$. On the other hand, after pushing back for $j \in \mathbb{N}$ lags, we have $e_{t-j} = y_{t-j} - \alpha_0 - \sum_{k=1}^{\infty} \alpha_k y_{t-j-k}$ is a linear combination of $(1, y_{t-j}, y_{t-j-1}, y_{t-j-2}, \ldots)$, which is a a subset of $(1, y_{t-1}, y_{t-2}, \ldots)$.

1.9 Wold Decomposition

If y_t is covariance stationarity, and the linear projection error has $\sigma^2 > 0$, then we can write

$$y_t = u_t + \sum_{j=0}^{\infty} b_j e_{t-j}$$

with $b_0 = 1$, and $u_t = \lim_{m\to\infty} \mathscr{P}_{t-m}(y_t)$. (For simplicity, we can consider the case $\mu_t = \mu$.) The Wold decomposition projects y_t onto the orthogonal basis spanned by $(e_t, e_{t-1,} e_{t-2}, ...)$.

`Zhentao Shi. March 13, 2025`