Time Series Analysis Notes

J S

2025.03.05

This is my personal study notes of Applied Time Series Analysis course, instructed by Dong Li, at Tsinghua University, during the spring semester of 2025. The course textbook is *Time Series Analysis With Applications in R*[a]. The notes is majorly inspired by the course slides. The last update is on 2025.06.11.

[a] Jonathan D. Cryer, Kung-Sik Chan.

Contents

0 Fundamentals of Probability and Measure Theory

Definition 0.1 (σ-algebra)**.** A **σ-algebra** ("sigma algebra") is part of the formalism for defining sets that can be measured. In formal terms, let Ω be some set, and let $P(\Omega)$ be its power set, i.e., the set of all subsets of Ω. Then a subset $\mathcal{F}$ is called a σ-algebra if and only if it satisfies the following properties:

- $\Omega \in \mathcal{F}$
- $\mathcal{F}$ is closed under complementation: If some set $A \in \mathcal{F}$, then so is its complement $\Omega \setminus A$.
- $\mathcal{F}$ is closed under countable unions: If some set $A_1, A_2, A_3, \ldots \in \mathcal{F}$, then so is $A = A_1 \cup A_2 \cup A_3 \cup \cdots$.

It also follows that the empty set $\emptyset$ is in $\mathcal{F}$, and thus Ω is in $\mathcal{F}$. The smallest possible σ-algebra on Ω is $\{\Omega, \emptyset\}$, while the largest is $P(\Omega)$.

Definition 0.2 (Measure)**. Measure** is a generalization and formalization of geometrical measures (length, area, volume) which represents the "**magnitude**" of a set. In formal terms, let Ω be a set and $\mathcal{F}$ a σ-algebra over Ω. A set function μ from $\mathcal{F}$ to the extended real line ($\mathbb{R} \cup \{-\infty, +\infty\}$) is called a measure if the following conditions hold:

1. **Non-negativity**: $\mu(E) \geq 0, \quad \forall E \in \mathcal{F}$
2. **Nullity**: $\mu(\emptyset) = 0$
3. **Countable Additivity**: For all countable collections $\{E_i\}_{i=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{F}$,

$$\mu\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mu(E_i)$$

Definition 0.3 (Measurable Set)**.** A set B is called a measurable set when it is the subset of a σ-algebra $\mathcal{F}$ on a set Ω.

Definition 0.4 (Measurable Space)**.** An ordered pair $(\Omega, \mathcal{F})$ is called a measurable space where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω.

Definition 0.5 (Measure Space)**.** An ordered pair $(\Omega, \mathcal{F}, \mu)$ is called a measure space where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and μ is a measure on $\mathcal{F}$.

Definition 0.6 (Probability Measure)**.** Let Ω be a set and $\mathcal{F}$ a σ-algebra over Ω. A probability measure on $\mathcal{F}$ is a measure with total measure one, i.e., $\mu(\Omega) = 1$. Or with a self-contained expression, a probability measure is a map from $\mathcal{F}$ to $[0, 1]$ which satisfies:

1. **Non-negativity**: $P(E) \geq 0, \quad \forall E \in \mathcal{F}$
2. **Unitarity**: $P(\Omega) = 1$
3. **Countable additivity**: For all countable collections $\{E_i\}_{i=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{F}$,

$$P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} P(E_i)$$

Definition 0.7 (Probability Space)**.** A probability space is a measure space with a probability measure. Formally, the triple $(\Omega, \mathcal{F}, P)$ is a probability space, where Ω is the sample space, $\mathcal{F}$ is the event space, and P is the probability measure.

Definition 0.8 (Random Variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(E, \mathcal{E})$ a measurable space. Then an $(E, \mathcal{E})$-valued random variable is a **measurable function** $X : \Omega \to E$, which means, for every subset $B \in \mathcal{E}$, its preimage is $\mathcal{F}$-measurable: $X^{-1}(B) \in \mathcal{F}$, where $X^{-1}(B) = \{\omega : X(\omega) \in B\}$.

In more intuitive terms, the random variable is a function from any outcome $\omega \in \Omega$ to a quantity $e \in E$. For any subset of possible quantities, $\{e_1, e_2, \ldots\} = B \in \mathcal{E}$, we can find its preimage $X^{-1}(B) = \{\omega_1, \omega_2, \ldots\} = \{\omega : X(\omega) \in B\} \in \mathcal{F}$. We can apply the probability measure $P : \mathcal{F} \to [0, 1]$ to this preimage, which gives a real number in $[0, 1]$. This is the probability of the event.

When E is a topological space, the most common choice for the σ-algebra $\mathcal{E}$ is the Borel σ-algebra $\mathcal{B}(E)$, which is the σ-algebra generated by the collection of all open sets in E. In such case the $(E, \mathcal{E})$-valued random variable is called an **E-valued random variable**. Moreover, when the space E is the real line $\mathbb{R}$, then such a real-valued random variable is called simply a **random variable**.

Definition 0.9 (Probability Distribution of Random Variable)**.** Consider a random variable $X : \Omega \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{F}, P)$, where $P : \mathcal{F} \to [0, 1]$ is a probability measure. The probability of the event $\{\omega : X(\omega) = x\}$ can be denoted as $P(X = x)$.

Recording all these probabilities of outputs of a random variable X yields the probability distribution of X, which can always be captured by its **cumulative distribution function (CDF)**,

$$F(x) = P(X \leq x)$$

In practice, the **probability density function (PDF)** is used more frequently, which is given by

$$f(x) = \frac{d}{dx} F_X(x)$$

In measure-theoretic terms, we push forward the measure P on Ω to a measure f on $\mathbb{R}$.

Note that both $F(x)$ and $f(x)$ are functions from $\mathbb{R}$ to $\mathbb{R}$. We have $F(-\infty) = 0$, $F(+\infty) = 1$, $\int_{-\infty}^{+\infty} f(x)dx = 1$, and $F(x)$ is always non-decreasing.

1 Fundamentals of Stochastic Processes

1.1 Stochastic Process

1.1.1 Basic Concepts of Stochastic Process

Definition 1.1 (Stochastic Process)**.** A stochastic process (random process) is a collection of random variables which are defined on a common probability space $(\Omega, \mathcal{F}, P)$, indexed by some index set T, and all take values in the same space E. In other words, for a given probability space $(\Omega, \mathcal{F}, P)$ and a measurable space $(E, \mathcal{E})$, a stochastic process is a collection of E-valued random variables, which can be written as $\{X_t : t \in T\}$ (or $\{X(\omega, t) : t \in T\}$). That is, $X : \Omega \times T \to E$.

- Often T will be a subset or an interval of $\mathbb{R}$.
- If $T = \mathbb{Z}$, $\{X_t\}$ is a discrete-time stochastic process (**time series**). If $T = [0, \infty)$, $\{X_t\}$ is a continuous-time stochastic process.
- Given a fixed $t \in T$, $X(\cdot, t)$ is an E-valued random variable.
- Given a fixed $\omega \in \Omega$, $X(\omega, \cdot)$, as a function of t, is a **realization** or **path** of stochastic process.

Time / State	discrete	continuous
discrete	Markov chain	time series
continuous	Poisson process	Brownian motion

Definition 1.2 (White Noise)**.** Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a stochastic process. If for any $s, t \in \mathbb{Z}$,

$$\mathbb{E}\varepsilon_t = \mu, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_s) = \sigma^2 \delta_{t-s} = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s, \end{cases}$$

then $\{\varepsilon_t\}$ is called a (weak) **white noise**, written as $\{\varepsilon_t\} \sim \mathrm{WN}(\mu, \sigma^2)$.

- If $\{\varepsilon_t\}$ is independent, then it is called **independent white noise**, denoted as $\mathrm{IWN}(\mu, \sigma^2)$.
- If $\mu = 0$, then ε_t is called **zero-mean white noise**, denoted as $\mathrm{WN}(0, \sigma^2)$. This is the most commonly used form.
- If $\mu = 0, \sigma^2 = 1$, then ε_t is called **standard white noise**, denoted as $\mathrm{WN}(0, 1)$.
- For $\mathrm{IWN}(\mu, \sigma^2)$, if ε_t is normally distributed, then ε_t is called **normal white noise** (i.e., $\varepsilon_t \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$).

Definition 1.3 (Martingale Difference Sequence)**.** A stochastic series $\{\varepsilon_t\}$ is an **martingale difference sequence** if its expectation with respect to the past is zero. Formally, consider an adapted sequence $\{\varepsilon_t, F_t\}_{-\infty}^{\infty}$ on a probability space $(\Omega, \mathcal{F}, P)$. $\{\varepsilon_t\}$ is an MDS if it satisfies the following two conditions:

- $\mathbb{E}|\varepsilon_t| < \infty$;
- $\mathbb{E}(\varepsilon_t|F_{t-1}) = 0$.

Definition 1.4 (Brownian motion/Wiener process)**.** If a continuous-time stochastic process $\{B(t) : t \in [0, \infty)\}$ satisfies

- $B(0) = 0$ a.s.
- $B(t)$ has independent increments, i.e., $B(t+u) - B(t)$ and $\{B(s) : s \leq t\}$ are independent for any $u \geq 0$,
- $B(t+u) - B(t) \sim \mathcal{N}(0, u\sigma^2)$,
- $B(t)$ is continuous in t with probability 1.

then $\{B(t) : t \in [0, \infty)\}$ is called a **Brownian motion** or **Wiener process**.

Facts: $\mathbb{E}[B(t)] = 0$ and $\mathrm{Var}\big(B(t)\big) = \sigma^2 t$.

Definition 1.5 (Gaussian process)**.** A stochastic process $\{X_t\}$ is **Gaussian** if and only if for every finite set of indices $t_1, t_2, \ldots, t_k$ in the index set T, $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is a multivariate Gaussian random variable.

Definition 1.6 (Orthogonality and Uncorrelation)**.** For $\{X_t\}$ and $\{Y_t\}$, for any $t, s \in T$, if $\mathbb{E}(X_s Y_t) = 0$, then call $\{X_t\}$ and $\{Y_t\}$ are **orthogonal**. If $\mathrm{Cov}(X_s, Y_t) = 0$, then call $\{X_t\}$ and $\{Y_t\}$ are **uncorrelated**.

1.1.2 Measure of Dependence

In this subsection, $\{X_t : t \in T\}$ is a given stochastic process.

Definition 1.7 (Mean Function)**.** The **mean function** is defined as

$$\mu(t) = \mathbb{E}(X_t) = \int x dF_t(x) = \int x F_t(dx), \quad t \in T$$

where $F_t(\cdot)$ is the cumulative distribution function of X_t, $t \in T$.

Definition 1.8 (Autocovariance and Autocorrelation)**.** For any $t, s \in \mathcal{T}$, $\gamma_{t,s} := \mathrm{Cov}(X_t, X_s)$ is called the **autocovariance function** (ACVF) of $\{X_t : t \in \mathcal{T}\}$. $\rho_{t,s} := \mathrm{Corr}(X_t, X_s)$ is called the **autocorrelation function** (ACF) of $\{X_t : t \in \mathcal{T}\}$.

Note: $\mathbb{E}(X_t^2) < \infty$ is prerequisite for each fixed $t \in \mathcal{T}$.

From the point of view of function:

- ACVF: $\gamma : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$,
- ACF: $\rho : \mathcal{T} \times \mathcal{T} \to [-1, 1]$.

Example 1.1 (ACF and ACVF of Brownian Motion)**.** Let $\{B(t) : t \in [0, \infty)\}$ be the standard Brownian motion, then its ACVF and ACF are

$$\gamma_{t,s} = \mathrm{Cov}(B(t), B(s)) = \min(t, s); \quad \rho_{t,s} = \frac{\min(t, s)}{\sqrt{ts}}, \quad s, t \in [0, \infty)$$

which both are continuous function on $T \times T$.

Proposition 1.1 (Properties of ACVF and ACF)**.** For a Weak stationary time series, γ_k and ρ_k have the following properties:

- $\gamma_0 = \mathrm{Var}(X_t)$; $\quad \rho_0 = 1$;
- $|\gamma_k| \leq \gamma_0$; $\quad |\rho_k| \leq 1$;
- $\gamma_k = \gamma_{-k}$; $\quad \rho_k = \rho_{-k}$;
- γ's and ρ's are **positive semidefinite** in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \gamma_{|t_i - t_j|} \geq 0, \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \rho_{|t_i - t_j|} \geq 0$$

for any $n \in \mathbb{N}$, any set of time point $\{t_1, t_2, \ldots, t_n\}$ and any real number $\alpha_1, \ldots, \alpha_n$.

Proposition 1.2 (Characterization of ACVF)**.** A real-valued function $\gamma(\cdot)$ defined on the integers is the ACVF of a weakly stationary time series **if and only if** it is even and **positive semidefinite** in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\gamma_{|t_i-t_j|}\geq 0$$

for any integer $n\geq 1$, any set of time point $\{t_1,t_2,\ldots,t_n\}$ and any real number $\alpha_1,\ldots,\alpha_n$.

Generally, for a weakly stationary time series, the following matrix (a special **Toeplitz matrix**) is **positive semidefinite** for any $n\geq 1$.

$$\mathbf{\Gamma}_n=\begin{pmatrix}\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-2} & \gamma_{n-1}\\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-3} & \gamma_{n-2}\\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-4} & \gamma_{n-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ \gamma_{n-2} & \gamma_{n-3} & \gamma_{n-4} & \cdots & \gamma_0 & \gamma_1\\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_1 & \gamma_0\end{pmatrix}_{n\times n}$$

Definition 1.9 (Partial Autocorrelation Function)**.** The partial correlation is the correlation between X and Y after their mutual linear dependency on the intervening variables Z has been removed. In time series, the partial autocorrelation is the correlation between X_t and X_{t+k} after their mutual linear dependency on the intervening variables $X_{t+1},\ldots,X_{t+k-1}$ has been removed.

In formal terms, let $\{X_t\}$ be a weakly stationary time series with $\mathbb{E}(X_t)=0$. The **partial autocorrelation** (PACF) is defined as $\phi_{11}=\rho_1$ and

$$\phi_{kk}=\mathrm{Corr}(R_{t|t+1,\ldots,t+k-1},R_{t+k|t+1,\ldots,t+k-1})\quad\text{for } k\geq 2$$

where $R_{j|t+1,\ldots,t+k-1}$ is the **residual** from the linear regression of X_j on $(X_{t+1},\ldots,X_{t+k-1})$, namely,

$$R_{j|t+1,\ldots,t+k-1}=X_j-(\alpha_{j1}X_{t+1}+\ldots+\alpha_{j,k-1}X_{t+k-1})$$

where

$$(\alpha_{j,1},\ldots,\alpha_{j,k-1})=\arg\min_{\{\beta_{j,i}\}}\mathbb{E}\Big[X_j-(\beta_{j1}X_{t+1}+\ldots+\beta_{j,k-1}X_{t+k-1})\Big]^2$$

Remark 1.1 (Remarks on Partial Autocorrelation Function)**.**

- Partial correlation can also be interpreted as the conditional correlation between X_t and X_{t-h} (conditioned on $X_{t-h+1},\cdots,X_{t-1}$).

  $$\phi_{kk}=\frac{\mathrm{Cov}(X_t,X_{t-k}|X_{t-k+1},\ldots,X_{t-1})}{\sqrt{\mathrm{Var}(X_t|X_{t-k+1},\ldots,X_{t-1})}\sqrt{\mathrm{Var}(X_{t-k}|X_{t-k+1},\ldots,X_{t-1})}}$$

  In the definition, we assume that $\mathbb{E}(X_t)=0$ to simplify the notation.
- For a **Gaussian** time series, the PACF is in fact equal to

  $$\phi_{kk}=\mathrm{Corr}(X_t,X_{t+k}\mid X_{t+1},\ldots,X_{t+k-1})$$

- Formula for ϕ_{kk} in vector form:

  $$\phi_{kk}=\frac{\gamma_k-\mathrm{Cov}(X_{t+k},\mathbf{X}_{k-1:1}^\tau)\mathbf{\Sigma}_{k-1:1}^{-1}\mathrm{Cov}(\mathbf{X}_{k-1:1},X_{t+k})}{\gamma_0-\mathrm{Cov}(X_t,\mathbf{X}_{k-1:1}^\tau)\mathbf{\Sigma}_{k-1:1}^{-1}\mathrm{Cov}(\mathbf{X}_{k-1:1},X_t)}$$

  where $\mathbf{X}_{k-1:1}=(X_{t+k-1},X_{t+k-2},\ldots,X_{t+1})^\tau$ and $\mathbf{\Sigma}_{k-1:1}=\mathrm{Var}(\mathbf{X}_{k-1:1})\coloneqq\mathrm{Cov}(\mathbf{X}_{k-1:1},\mathbf{X}_{k-1:1})$.
- Because by definition $\rho_0=\phi_{00}=1$ for any process, when we talk about the ACF and PACF, we refer only to ρ_k and ϕ_{kk} for $k\neq 0$.

Proposition 1.3 (Simple formula for Partial Autocorrelation)**.**

$$\phi_{11} = \rho_1$$

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & 1 & \rho_{k-1} \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & \rho_k \end{vmatrix}_{k\times k}}{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & 1 & \rho_1 \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_1 & 1 \end{vmatrix}_{k\times k}}, \quad k \geq 2$$

Specifically,

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}, \qquad \phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

1.1.3 Stationarity

Definition 1.10 (Weak Stationarity)**.** A stochastic process $\{X_t : t \in T\}$ is **weakly stationary** if $\mathbb{E}(X_t^2) < \infty$ for each $t \in T$, and

- $\mathbb{E}(X_t) \equiv \mu$, e.g., $\mathbb{E}(X_t)$ is a constant independent of t,
- $\mathrm{Cov}(X_t, X_{t+k})$ is independent of t for each k.

Write: $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$ and $\rho_k = \gamma_k/\gamma_0$.

Definition 1.11 (Strict Stationarity)**.** A stochastic process $\{X_t : t \in T\}$ is **strictly stationary** if $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$ have the same joint distribution for all $n \geq 1$, for all h, and for all $\{t_1 + h, \ldots, t_n + h\} \subseteq T$.

Example 1.2 (Weakly Stationary but not Strictly Stationary)**.** Let $\{X_t : t \in \mathbb{N}\} = \{\varepsilon, \eta, \varepsilon, \eta, \ldots\}$, i.e., $X_{2t-1} = \varepsilon$, $X_{2t} = \eta$, where $\varepsilon \sim \mathcal{N}(0, 1)$, $\eta \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$, and ε and η are independent. Then $\{X_t\}$ is weakly stationary but not strictly stationary.

Solution: Clearly, $\mathbb{E}(X_t^2) = 1 < \infty$, $\mathbb{E}(X_t) = 0$ and

$$\gamma_{t,t+s} = \mathrm{cov}(X_t, X_{t+s}) = \begin{cases} \mathrm{cov}(X_t, X_t) = 1, & \text{if } s \text{ is even,} \\ \mathrm{cov}(\varepsilon, \eta) = 0, & \text{if } s \text{ is odd.} \end{cases}$$

Thus, $\gamma_{t,t+s}$ is independent of t and only depends on s, and $\{X_t\}$ is weakly stationary. However, $\{X_t\}$ is not strictly stationary since ε and η do not have the same distribution functions.

Example 1.3 (Strictly Stationary but not Weakly Stationary)**.** Consider a time series $\{X_t : t \in T\} = \{\varepsilon_t : t \in T\}$ which is i.i.d. standard **Cauchy** random variables with the density $f(x) = \dfrac{1}{\pi(1 + x^2)}$. Then, $\{X_t\}$ is strictly stationary but not weakly stationary in that $\mathbb{E}(X_t^2) = \infty$.

1.2 Representation and Decomposition*

1.2.1 Wold Decomposition

Definition 1.12 (Vector Space View of Time Series)**.** Denote the closed linear span of $\{X_t : t \leq n\}$ as

$$\mathcal{M}_n = \overline{\text{sp}}\{X_t : t \leq n\}$$

In Hilbert space framework,

- $\mathcal{M} = \overline{\text{sp}}\{X_t : t \in \mathbb{Z}\}$ is defined as the overall space.
- $\mathbf{P}_{\mathcal{M}_n}$ is defined as the projection operator from $\mathcal{M}$ to $\mathcal{M}_n$.
- $\mathcal{M}_{-\infty}$ is defined as the intersection of all $\mathcal{M}_n$:

$$\mathcal{M}_{-\infty} = \bigcap_{n=-\infty}^{\infty} \mathcal{M}_n$$

  $\mathcal{M}_{-\infty}$ captures the "infinitely persistent" part of $\{X_t\}$. It represents information that remains relevant across all time horizons. If $\mathcal{M}_{-\infty}$ is non-trivial (i.e. not just zero), the process has a deterministic component that cannot be reduced to randomness.

Definition 1.13 (Deterministic Process)**.** A process is **deterministic** if its future values can be perfectly predicted (in mean-square sense) using its past. Formally, consider a stochastic process $\{X_t, t \in \mathbb{Z}\}$. The 1-step mean squared error can be written as

$$\sigma^2 = \mathbb{E}|X_{n+1} - \mathbf{P}_{\mathcal{M}_n} X_{n+1}|^2$$

$\{X_t\}$ is called deterministic if $\sigma^2 = 0$.

Definition 1.14 (Wold Decomposition)**.** For any zero-mean weakly stationary time series $\{X_t : t \in \mathbb{Z}\}$, if $\sigma^2 > 0$, then X_t can be expressed as

$$X_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j} + V_t$$

where

1. $\phi_0 = 1$ and $\sum_{j=0}^{\infty} \phi_j^2 < \infty$;
2. $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$;
3. $\varepsilon_t \in \mathcal{M}_t$ for each $t \in \mathbb{Z}$;
4. $E(\varepsilon_t V_s) = 0$ for all $s, t \in \mathbb{Z}$;
5. $V_t \in \mathcal{M}_{-\infty}$ for each $t \in \mathbb{Z}$;
6. $\{V_t\}$ is deterministic.

Remark:

- (5) and (6) are not the same since $\mathcal{M}_{-\infty}$ is defined in terms of $\{X_t\}$, not $\{V_t\}$.
- $\{\phi_j\}$, $\{\varepsilon_t\}$ and $\{V_t\}$ are **uniquely** determined by $X_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j} + V_t$ and the conditions (1)-(6).

Summary: The Wold Decomposition separates a stationary process into:

- A noise-driven component, which is purely random and unpredictable;
- A deterministic component, which can be perfectly forecasted using its infinite past.

1.2.2 AR and MA Representation

Definition 1.15 (General Linear Process)**.** A general linear process $\{X_t\}$ is one that can be represented as a weighted linear combination of present and past white noise terms as

$$X_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

where $\psi_0 = 1$, $\{\varepsilon_t\} \sim \text{WN}(0, \sigma_\varepsilon^2)$, and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Proposition 1.4 (Properties of General Linear Process)**.**

1. $\mathbb{E}(X_t) = \mu$;
2. $\text{Var}(X_t) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j^2$;
3. $\gamma_k = \text{Cov}(Y_t, Y_{t-k}) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}, \quad k \geq 0.$
4. $\rho_k = \text{Corr}(Y_t, Y_{t-k}) = \dfrac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}, \quad k \geq 0.$

Example 1.4 (Example of General Linear Process)**.** An important nontrivial example is the case where the ψ's form an exponentially decaying sequence

$$\psi_j = \phi^j$$

where ϕ is a number strictly between -1 and 1. So $\{X_t\}$ can be written as

$$X_t = \mu + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

For this example,

$$\mathbb{E}(Y_t) = \mu, \qquad \text{Var}(Y_t) = \sigma_\varepsilon^2 \frac{1}{1-\phi^2}$$

$$\text{Cov}(Y_t, Y_{t-k}) = \sigma_\varepsilon^2 \frac{\phi^k}{1-\phi^2}, \qquad \text{Corr}(Y_t, Y_{t-k}) = \phi^k$$

Definition 1.16 (Impulse Response)**.** For a general linear process $X_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$, the ψ_j's are called impulse responses, since

$$\frac{\partial X_t}{\partial \varepsilon_{t-s}} = \psi_s, \quad s = 1, 2, \ldots$$

- For a weakly stationary and ergodic time series, $\lim_{s\to\infty} \psi_s = 0$, and the long-run cumulative impulse response $\sum_{s=0}^{\infty} \psi_s < \infty$.
- A plot of ψ_s against s is called the **impulse response function (IRF)**.

Definition 1.17 (AR(∞) representation)**.** In AR(∞) form, a stochastic process $\{X_t\}$ can be written as

$$X_t - \mu = \pi_1(X_{t-1} - \mu) + \pi_2(X_{t-2} - \mu) + \cdots + \varepsilon_t$$

or

$$\pi(B)(X_t - \mu) = \varepsilon_t$$

where $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$, and $\sum_{j=1}^{\infty} |\pi_j| < \infty$.

- Explanation:
  1. Regard X_t on its own past values plus a random shock.
  2. The random shock can be recovered by the values of X's.
- Usefulness: understanding the mechanism of forecasting.

Definition 1.18 (Invertible Process)**.** If a time series $\{X_t\}$ can be written in AR(∞) form, it is called **invertible process**. Invertibility allows for representation of errors as a linear combination of past observations, which is crucial for forecasting.

1.3 Estimation

A weak stationary time series is characterized by its mean μ, variance σ^2, ACVF γ_k, ACF ρ_k, and PACF ϕ_{kk}. The exact values of these parameters can be calculated if the ensemble of all possible realizations is known.

Question: Assume only one realization $\{y_1, y_2, ..., y_n\}$ is available for a weakly stationary time series $\{y_t\}$, how to estimate μ, σ^2, ρ_k, γ_k and ϕ_{kk}?

1.3.1 Estimation of Mean

Definition 1.19 (Sample mean)**.**

$$\bar{y}_n = \frac{1}{n}\sum_{t=1}^{n} y_t$$

Proposition 1.5 (Properties of Sample mean)**.** If $\{y_t\}$ is weakly stationary, then the sample mean $\bar{y}_n$ satisfies:

1. **Unbiasedness**: $E(\bar{y}_n) = \frac{1}{n}\sum_{t=1}^{n} E(y_t) = \mu$.
2. **Consistency**: If $\mathrm{Var}(\bar{y}_n) \to 0$, then $\bar{y}_n \to \mu$ in L_2.

1.3.2 Estimation of ACVF γ_k

Definition 1.20 (Sample ACVF)**.**

$$\widehat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y}_n)(y_{t+k} - \bar{y}_n)$$

or

$$\widehat{\widehat{\gamma}}_k = \frac{1}{n-k}\sum_{t=1}^{n-k}(y_t - \bar{y}_n)(y_{t+k} - \bar{y}_n)$$

It can be derived that

$$\mathbb{E}(\widehat{\gamma}_k) \approx (1 - \frac{k}{n})\Big(\gamma_k - \mathrm{Var}(\bar{y}_n)\Big)$$

$$\mathbb{E}(\widehat{\widehat{\gamma}}_k) \approx \gamma_k - \mathrm{Var}(\bar{y}_n)$$

1.3.3 Estimation of ACF ρ_k

Definition 1.21 (Sample ACF)**.**

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y}_n)(y_{t+k} - \bar{y}_n)}{\sum_{t=1}^{n}(y_t - \bar{y}_n)^2}, \quad k \geq 1$$

Remark 1.2 (Remark on Sample ACF*)**.** Further, if a stochastic process $\{y_t\}$ is **Gaussian**, Bartlett (1946) has shown that for $k > 0$ and $k + j > 0$,

$$\mathrm{Cov}(\widehat{\rho}_k, \widehat{\rho}_{k+j}) \approx \frac{1}{n}\sum_{i=-\infty}^{\infty}\Big(\rho_i\rho_{i+j} + \rho_{i+k+j}\rho_{i-k} - 2\rho_k\rho_i\rho_{i-k-j} - 2\rho_{k+j}\rho_i\rho_{i-k} + 2\rho_k\rho_{k+j}\rho_i^2\Big)$$

For large n, $\sqrt{n}(\widehat{\rho}_k - \rho_k) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where

$$\sigma^2 = \sum_{i=-\infty}^{\infty}\Big(\rho_i^2 + \rho_{i+k}\rho_{i-k} - 4\rho_k\rho_i\rho_{i-k} + 2\rho_k^2\rho_i^2\Big). \quad (\approx \mathrm{Var}(\widehat{\rho}_k))$$

For processes in which $\rho_k = 0$ for $k > q$, (e.g. MA(q) model), Bartlett's approximation becomes

$$\mathrm{Var}(\widehat{\rho}_k) \approx \frac{1}{n}(1 + 2\rho_1^2 + \ldots + 2\rho_q^2)$$

In practice, $\rho_i (i = 1, \ldots, q)$ are unknown and are replaced by their sample estimates $\widehat{\rho}_i$ and we have the following large-lag standard error of $\widehat{\rho}_k$:

$$s_{\widehat{\rho}_k} \approx \sqrt{\frac{1}{n}(1 + 2\widehat{\rho}_1^2 + \ldots + 2\widehat{\rho}_q^2)}$$

Particularly, to test a white noise process, we use

$$s_{\widehat{\rho}_k} \approx \sqrt{\frac{1}{n}}$$

Definition 1.22 (Sample PACF)**.**

$$\widehat{\phi}_{11} = \widehat{\rho}_1$$

$$\widehat{\phi}_{kk} = \frac{\begin{vmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_{k-2} & \widehat{\rho}_1 \\ \widehat{\rho}_1 & 1 & \widehat{\rho}_1 & \cdots & \widehat{\rho}_{k-3} & \widehat{\rho}_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \widehat{\rho}_{k-1} & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-3} & \cdots & \widehat{\rho}_1 & \widehat{\rho}_k \end{vmatrix}_{k\times k}}{\begin{vmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-1} \\ \widehat{\rho}_1 & 1 & \widehat{\rho}_1 & \cdots & \widehat{\rho}_{k-3} & \widehat{\rho}_{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \widehat{\rho}_{k-1} & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-3} & \cdots & \widehat{\rho}_1 & 1 \end{vmatrix}_{k\times k}}, \quad k \geq 2$$

It was shown by Quenouille (1949) that on the hypothesis that underlying process is a white noise sequence,

$$\mathrm{Var}(\widehat{\phi}_{kk}) \approx \frac{1}{n}$$

Hence, $\pm 2/\sqrt{n}$ can be used as critical limits on ϕ_{kk} to test the hypothesis of a white noise process.

1.4 Testing

1.4.1 Tests for ACF

Proposition 1.6 (Tests for ACF)**.** If $\{y_t\}$ are i.i.d. sequence with finite **fourth moments**, then

- For any fixed $j \neq 0$, we have $\sqrt{n}\hat{\rho}_j \overset{A}{\sim} \mathcal{N}(0,1)$ (Fuller, 1976). To test $H_0 : \rho_j = 0$ at the α level, reject if $\sqrt{n} \cdot |\hat{\rho}_j| \geq z_{\alpha/2}$.
- **Box-Pierce Q-statistic** (Box and Pierce, 1970): Consider testing $H_0 : \rho_1 = \cdots = \rho_m = 0$. Under H_0 of IID sequence, the portmanteau statistic

$$Q_m = n \sum_{k=1}^{m} \hat{\rho}_k^2 \rightsquigarrow \chi_m^2$$

- **Modified Q-statistic** or Ljung-Box test statistic (Ljung and Box, 1978): (finite sample performance)

$$Q_m^* = n(n+2) \sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{n-k} \rightsquigarrow \chi_m^2$$

- Simulation studies suggest that the choice of $m \approx \ln(n)$ provides better power performance.

1.5 Data Handling

1.5.1 Operators in Time Series Representation

Definition 1.23 (Backshift Operator)**.** An operator B which satisfies $B^k X_t = X_{t-k}$ is defined as **backshift operator**.

In compact form, a general linear process $X_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ can be written as

$$X_t - \mu = \psi(B)\varepsilon_t$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

Definition 1.24 (Difference Operator)**.** An operator Δ which satisfies $\Delta X_t = X_t - X_{t-1} = (1-B)X_t$ is defined as **difference operator**.

- This is a special case of lag polynomial.

- The difference operator can generalize to the k-th order:

$$\Delta^k y_t = (1-B)^k y_t$$

- Log-first-difference / log-difference: $\Delta \ln y_t = \ln y_t - \ln y_{t-1}$.
- Similarly, we have the following results:
- **Seasonal difference operator**: If consider the change between y_t and y_{t-k}, we can use notation of difference operator by $\Delta_k y_t = (1-B^k)y_t$. For monthly data, use $\Delta_{12} y_t = (1-B^{12})y_t$; for quarterly data, use $\Delta_3 y_t = (1-B^3)y_t$.

1.5.2 Seasonality and Seasonal Adjustment

Remark 1.3 (Remark on Seasonal Models)**.** The investigation of many (economic) time series becomes problematic due to seasonal fluctuations. Time series are made up of **four components**:

- S_t: The **seasonal** component;
- T_t: The **trend** component;
- C_t: The **cyclical** component;
- I_t: The **error**, or irregular component.

In time series data, **seasonality** is the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly. Seasonality may be caused by various factors, such as weather, vacation, and holidays and consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series.

A **cycle** occurs when the data exhibit rises and falls that are not of a fixed period. These fluctuations are usually due to economic conditions and are often related to the "business cycle." It is important to distinguish **cyclic patterns** and **seasonal patterns**. Seasonal patterns have a fixed and known length, while cyclic patterns have variable and unknown length. The average length of a cycle is usually longer than that of seasonality, and the magnitude of cyclic variation is usually more variable than that of seasonal variation.

2 ARMA Models

2.1 Auto Regression (AR)

Definition 2.1 (Autoregressive Model)**.** If a stochastic process $\{y_t, t \in \mathbb{Z}\}$ satisfies

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

where c is an intercept term, and ϕ_i's are coefficients, then $\{y_t\}$ is called an **autoregressive model** of order p, denoted by **AR(p)**.

In backshift operator notation, let $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, then AR($p$) model can be written as

$$\phi(B) y_t = c + \varepsilon_t$$

Definition 2.2 (Causal Time Series)**.** A time series/process $\{y_t : t \in \mathbb{Z}\}$ is **causal** if $y_t = f(\varepsilon_{t-j} : j \geq 0)$ for all t, where f is a measurable function.

Remark 2.1 (Discussion on Causal Time Series)**.**

1. Causality means that y_t is caused by the white noise process (from the past) up to time t. For AR(1) model (or in general, ARMA (p, q) process), causality is equivalent to that $\phi(z) \neq 0$ for all $|z| \leq 1$, and therefore it implies weak stationarity. But the converse is not true.
2. In fact, a **unique weakly stationary solution** is attained if and only if $\phi(z) \neq 0$ for all complex numbers z on the unit cycle $|z| = 1$.

3. Under the condition $\phi(z) \neq 0$ for all $|z| > 1$, the weakly stationary solution of AR(1) is of the form

$$y_t = \sum_{k=0}^{\infty} d_k \varepsilon_{t+k}$$

which is not causal. One may argue whether such a process should be called a time series since y_t depends on 'future' noise ε_{t+k} for $k \geq 1$.

4. Any weakly stationary noncausal ARMA process can be represented as a causal ARMA process (with the same orders) in terms of a newly defined white noise, and both processes have identical first two moments (Proposition 3.5.1 of Brockwell and Davis 1991).

2.1.1 AR(1) Model

Definition 2.3 (AR(1) Model)**.** A simplest AR(1) model is given as

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

Proposition 2.1 (Weak Stationarity of AR(1) Model)**.** By iteration, we have

$$y_t = \varepsilon_t + \sum_{i=1}^{m} \phi^i \varepsilon_{t-i} + \phi^{m+1} y_{t-m-1}$$

Consider four cases:

1. if $|\phi| < 1$, then $y_t = \varepsilon_t + \sum_{j=1}^{\infty} \phi^j \varepsilon_{t-j}$.
2. if $\phi = 1$, then $y_t = \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1 + y_0$, the variance grows with t.
3. if $\phi = -1$, then $y_t = \varepsilon_t + \sum_{j=1}^{t-1} (-1)^j \varepsilon_{t-j} + (-1)^t y_0$, the variance grows with t.
4. if $|\phi| > 1$, then $y_t = -\frac{\varepsilon_{t+1}}{\phi} + \frac{y_{t+1}}{\phi} = -\sum_{j=1}^{m} \frac{\varepsilon_{t+j}}{\phi^j} + \frac{y_{t+m}}{\phi^m} = -\sum_{j=1}^{\infty} \frac{\varepsilon_{t+j}}{\phi^j}$.

Thus, AR(1) model is weakly stationary if and only if $|\phi| \neq 1$.

Remark 2.2 (Causal Assumption of AR(1) Model)**.** An AR(1) model $y_t = \phi y_{t-1} + \varepsilon_t$ is noncausal if $|\phi| > 1$. In the scope of this file, we assume $|\phi| < 1$ for AR(1) models.

Proposition 2.2 (Other Properties of AR(1) Model)**.**

1. **Characteristic equation**:

$$\phi(z) = 1 - \phi z$$

So the root of $\phi(z) = 0$ is $z = 1/\phi$.

2. **ACVF and ACF**: It is easy to show that

$$\gamma_0 = \frac{\sigma^2}{1 - \phi^2}, \quad \gamma_j = \phi \gamma_{j-1} = \phi^j \frac{\sigma^2}{1 - \phi^2}, \quad \rho_j = \phi^j, \quad \text{for } j \in \mathbb{N}.$$

The ACFs decay at a **geometric rate**.

3. **The Wold decomposition**:

$$y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

4. **IMF**:

$$\frac{\partial y_t}{\partial \varepsilon_{t-j}} = \phi^j, \quad j \in \mathbb{N}$$

5. **Mean-reverting behavior**:

$$\Delta y_t = (\phi - 1) y_{t-1} + \varepsilon_t \implies \begin{cases} E[\Delta y_t | \mathcal{F}_{t-1}] < 0, & \text{if } y_{t-1} > 0; \\ E[\Delta y_t | \mathcal{F}_{t-1}] > 0, & \text{if } y_{t-1} < 0, \end{cases}$$

where $\mathcal{F}_t = \sigma(y_j : j \leq t)$, the σ-algebra that contains all information about the process $\{y_t\}$ up to time t.

2.1.2 AR(2) Model

Definition 2.4 (AR(2) Model)**.** An AR(2) model is given by

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

The Characteristic equation is $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 = 0$, and the mean-corrected form is given by

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

Proposition 2.3 (Weak stationarity of AR(2) Model)**.** An AR(2) model is weakly stationary if and only if all roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 = 0$ lie **outside** the unit circle. Equivalently,

$$-1 < \phi_2 < 1, \quad \phi_2 + \phi_1 < 1, \quad \phi_2 - \phi_1 < 1$$

Proposition 2.4 (Moment Equations of AR(2) Model)**.** . Given an AR(2) model $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$, the moment equations are:

- **Mean equation**: Assume weak stationarity, we have

$$\begin{aligned} & \mathbb{E}(y_t) = c + \phi_1 \mathbb{E}(y_{t-1}) + \phi_2 \mathbb{E}(y_{t-2}) \\ \Longrightarrow \quad & \mu = c + \phi_1 \mu + \phi_2 \mu \\ \Longrightarrow \quad & \mu = \frac{c}{1 - \phi_1 - \phi_2} \end{aligned}$$

- **Variance equation**: Assume zero mean ($c = 0$), we have:

$$\begin{aligned} & \mathbb{E}(y_t y_{t-j}) = \phi_1 \mathbb{E}(y_{t-1} y_{t-j}) + \phi_2 \mathbb{E}(y_{t-2} y_{t-j}) \\ \Longrightarrow \quad & \mathrm{Cov}(y_t, y_{t-j}) = \phi_1 \mathrm{Cov}(y_{t-1}, y_{t-j}) + \phi_2 \mathrm{Cov}(y_{t-2}, y_{t-j}) \\ \Longrightarrow \quad & \gamma_j = \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} \\ \Longrightarrow \quad & \phi(B)\gamma_j = 0 \end{aligned}$$

  By $\rho_k = \gamma_k / \gamma_0$, we have

$$\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2}, \quad \phi(B)\rho_j = 0, \quad (j \geq 1)$$

  In particular, by $\rho_1 = \phi_1 + \phi_2 \rho_1$ and $\rho_2 = \phi_1 \rho_1 + \phi_2$, we get the **Yule-Walker equation**:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}$$

Proposition 2.5 (MA(∞) representation of AR(2) Model)**.**

$$\begin{aligned} & (y_t - \mu) - \phi_1(y_{t-1} - \mu) - \phi_2(y_{t-2} - \mu) = \varepsilon_t \\ \Longrightarrow \quad & (1 - \phi_1 B - \phi_2 B^2)(y_{t-1} - \mu) = \varepsilon_t \\ \Longrightarrow \quad & \phi(B)(y_t - \mu) = \varepsilon_t \\ \Longrightarrow \quad & y_t - \mu = [\phi(B)]^{-1} \varepsilon_t \\ \Longrightarrow \quad & y_t - \frac{c}{1 - \phi_1 - \phi_2} = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \psi_3 \varepsilon_{t-3} + \cdots \end{aligned}$$

Here $1 + \psi_1 B + \psi_2 B^2 + \cdots = [\phi(B)]^{-1} = \dfrac{1}{1 - \phi_1 B - \phi_2 B^2}$.

Example 2.1 (Example of AR(2) Process)**.** Assume the AR(2) process is

$$y_t = 0.3 y_{t-1} + 0.4 y_{t-2} + \varepsilon_t \quad \text{or} \quad (1 - 0.3B - 0.4B^2) y_t = \varepsilon_t$$

Firstly, solve the characteristic equation

$$\phi(z) = 1 - 0.3z - 0.4z^2 = (1 - 0.8z)(1 + 0.5z) = 0$$

The two roots are $z_1 = 1.25$ and $z_2 = -2$. Thus,

$$\begin{aligned} y_t &= \frac{1}{(1-0.8B)(1+0.5B)}\varepsilon_t \\ &= \frac{1}{1.3B}\left[\frac{1}{1-0.8B} - \frac{1}{1+0.5B}\right]\varepsilon_t \\ &= \frac{1}{1.3B}\left\{\left[1+(0.8B)+(0.8B)^2+\cdots\right] - \left[1+(-0.5B)+(-0.5B)^2+\cdots\right]\right\}\varepsilon_t \\ &= \frac{1}{1.3B}\left[\sum_{j=0}^{\infty}(0.8B)^{j+1} - \sum_{j=0}^{\infty}(-0.5B)^{j+1}\right]\varepsilon_t \\ &= \sum_{j=0}^{\infty}\frac{0.8^{j+1}-(-0.5)^{j+1}}{1.3}\varepsilon_{t-j} \end{aligned}$$

Generally, if an AR(2) process can be expressed as $(1-\lambda_1 B)(1-\lambda_2 B)y_t = \varepsilon_t$, its MA($\infty$) representation can be written as

$$y_t = \begin{cases} \displaystyle\sum_{j=0}^{\infty}\frac{\lambda_1^{j+1}-\lambda_2^{j+1}}{\lambda_1-\lambda_2}\varepsilon_{t-j}, & \lambda_1 \neq \lambda_2 \\ \displaystyle\sum_{j=0}^{\infty}(j+1)\lambda_1^j\varepsilon_{t-j}, & \lambda_1 = \lambda_2 \end{cases}$$

Proposition 2.6 (ACF of AR(2) Model)**.** Denote λ_1, λ_2 the reciprocal roots of the characteristic equation $1-\phi_1 B - \phi_2 B^2 = 0$, i.e. $(1-\lambda_1 B)(1-\lambda_2 B) = 0$. By the moment equation, we have derived that $(1-\phi_1 B - \phi_2 B^2)\rho_j = 0$, so the ACF has the following form:

$$\rho_j = c_1\lambda_1^j + c_2\lambda_2^j$$

where c_1 and c_2 are unknown parameters depending on the initial conditions.

Consider the following two cases:

- If both roots (λ_1, λ_2) are real, then the ACF of AR(2) is a mixture of two exponential decay.
- If the roots are a complex conjugate pair, i.e., $\phi_1^2 + 4\phi_2 < 0$, then ACF would show a picture of damping sine and cosine waves; such roots usually give rise to the behavior of seasonal patterns.

Proposition 2.7 (Average length of stochastic cycle of AR(2) Model)**.** For a weakly stationary AR(2) model with a pair of complex characteristic roots ($\phi_1^2 + 4\phi_2 < 0$), the average length of the stochastic cycle is

$$T_0 = \frac{2\pi}{\arccos\left[\phi_1/(2\sqrt{-\phi_2})\right]}$$

2.1.3 AR(p) Model

Definition 2.5 (AR(p) model)**.** An AR(p) model is given by

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

or

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)y_t = c + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

Proposition 2.8 (Weak stationarity of AR(p) Model)**.** It can be shown that the AR(p) is weakly stationary provided all roots of the characteristic equation

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$$

lie **outside the complex unit circle** (have modulus greater than one).

- A necessary condition for weak stationarity is $\phi_1 + \cdots + \phi_p \neq 1$.
- When there exists a root 1 in the characteristic equation, we say that the AR(p) process has a **unit root**.

Proposition 2.9 (Other Properties of AR(p) Model)**.**

- **Unconditional mean** (given stationarity): $\mu = c/(1 - \phi_1 - \cdots - \phi_p)$ and thus $\phi(B)(y_t - \mu) = \varepsilon_t$.
- **ACVF**: For simplicity, assume that $c = \mu = 0$. Then, multiplying both sides of the AR model by y_{t-j}, and taking expectation, we have

$$\begin{aligned} & y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = \varepsilon_t \\ \Longrightarrow \quad & \mathbb{E}[(y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p}) y_{t-j}] = \mathbb{E}(\varepsilon_t y_{t-j}) \\ \Longrightarrow \quad & \mathbb{E}[(y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p}) y_{t-j}] = 0 \\ \Longrightarrow \quad & \gamma_j - \phi_1 \gamma_{j-1} - \cdots - \phi_p \gamma_{j-p} = 0 \\ \Longrightarrow \quad & \phi(B)\gamma_j = 0 \quad (j \geq 1) \end{aligned}$$

- **ACF**: By moment equation of ACVF, we have

$$\phi(B)\rho_j = \rho_j - \phi_1 \rho_{j-1} - \cdots - \phi_p \rho_{j-p} = \begin{cases} \sigma_\varepsilon^2/\gamma_0, & \text{for } j = 0; \\ 0, & \text{for } j > 0 \end{cases}$$

- **MA(∞) representation**:

$$y_t - \mu = \frac{1}{\phi(B)}\varepsilon_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

  where the ψ_i's are referred to as the ψ-weight of the model and also called the impulse response function.

- **Partial Autocorrelation of AR (p) Process**: The partial autocorrelation of AR(p) process **truncates** after order p, i.e.,

$$\psi_{kk} = 0 \text{ for } k = p+1,\ p+2,\ p+3, \cdots$$

Remark 2.3 (How to Calculate ψ_j for General Linear Process Representation?)**.**

$$\begin{aligned} \frac{1}{\phi(B)} =& 1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots \\ 1 =& (1 - \phi_1 B - \cdots - \phi_p B^p)(1 + \psi_1 B + \psi_2 B^2 + \cdots) \\ \Longrightarrow & \begin{cases} \psi_1 &= \phi_1 \\ \psi_2 &= \phi_1 \psi_1 + \phi_2 \\ \psi_3 &= \phi_1 \psi_2 + \phi_2 \psi_1 + \phi_3 \\ \vdots & \\ \psi_p &= \phi_1 \psi_{p-1} + \phi_2 \psi_{p-2} + \cdots + \phi_{p-1}\psi_1 + \phi_p \\ \psi_j &= \sum_{i=1}^{p} \psi_{j-i}, \quad \text{for } j > p \end{cases} \end{aligned}$$

So, for AR process $\phi(B)y_t = \varepsilon_t$, the general linear process representation is

$$y_t = [\phi(B)]^{-1}\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\psi_0 = 0$, and other ψ_j's are calculated as above.

2.1.4 Parameter Estimation

Proposition 2.10 (Yule-Walker equation of AR(p) Model)**.** For the AR(p) model, the first $p+1$ moment equations of ACVF are:

$$\gamma_j = \phi_1 \gamma_{j-1} + \cdots + \phi_j \gamma_{j-p}, \quad j = 1, \ldots, p$$

$$\sigma_\varepsilon^2 = \gamma_0 - \phi_1\gamma_1 - \cdots - \phi_p\gamma_p$$

The derivation has been shown in proposition 2.4. Written in matrix notation, we have the **Yule-Walker equations**:

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{p-1} \\ \gamma_p \end{bmatrix} = \begin{bmatrix} 1 & \gamma_1 & \cdots & \gamma_{p-2} & \gamma_{p-1} \\ \gamma_1 & 1 & \cdots & \gamma_{p-3} & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & 1 & \gamma_1 \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p-1} \\ \phi_p \end{bmatrix}$$

Denote $\boldsymbol{\Gamma}_p = (\gamma_{k-j})_{j,k=1}^p \in \mathbb{R}^{p\times p}$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T \in \mathbb{R}^{p\times 1}$, and $\boldsymbol{\gamma}_p = (\gamma_1, \ldots, \gamma_p)^T \in \mathbb{R}^{p\times 1}$, the Yule-Walker equations can be expressed as

$$\boldsymbol{\Gamma}_p\boldsymbol{\phi} = \boldsymbol{\gamma}_p, \quad \sigma_\varepsilon^2 = \gamma_0 - \boldsymbol{\phi}^T\boldsymbol{\gamma}_p$$

Replace with sample values, we have

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p, \quad \hat{\sigma}_\varepsilon^2 = \hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_p^T\hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p$$

These estimators are typically called the **Yule-Walker estimators**.

By the same rules, the Yule-Walker equations for ACF are

$$\rho_j = \phi_1\rho_{j-1} + \cdots + \phi_j\rho_{j-p}, \quad j = 1, \ldots, p$$

$$\sigma_\varepsilon^2 = \gamma_0(1 - \phi_1\rho_1 - \cdots - \phi_p\rho_p)$$

In matrix notation,

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-3} & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{p-2} & \rho_{p-3} & \cdots & \rho_2 & \rho_1 \\ \rho_{p-1} & \rho_{p-2} & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p-1} \\ \phi_p \end{bmatrix}$$

The Yule-Walker estimates with sample ACF are

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{R}}_p^{-1}\hat{\boldsymbol{\rho}}_p, \quad \hat{\sigma}_\varepsilon^2 = \hat{\gamma}_0\left(1 - \hat{\boldsymbol{\rho}}_p^T\hat{\mathbf{R}}_p^{-1}\hat{\boldsymbol{\rho}}_p\right)$$

where $\hat{\mathbf{R}}_p = (\hat{\rho}_{k-j})_{j,k=1}^p \in \mathbb{R}^{p\times p}$ and $\hat{\boldsymbol{\rho}}_p = (\hat{\rho}_1, \ldots, \hat{\rho}_p)^T \in$ ı × 𝓚. If the sample size n is large, then

$$\sqrt{n}(\hat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2\boldsymbol{\Gamma}_p^{-1}), \quad \hat{\sigma}_\varepsilon^2 \xrightarrow{p} \sigma_\varepsilon^2$$

In particular, when $p = 1$, we have $\boldsymbol{\Gamma}_p = \gamma_0 = \frac{\sigma_\varepsilon^2}{1-\phi^2}$. Thus,

$$\sqrt{n}(\hat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(0, 1 - \phi^2) \quad \text{when } |\phi| < 1$$

Proposition 2.11 (Maximum Likelihood Estimation of AR(1) Model)**.** To be completed...

2.2 Moving Average (MA)

2.2.1 MA(1) Model

Definition 2.6 (MA(1) Model)**.** A typical MA(1) model is given by

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

Proposition 2.12 (Properties of MA(1) Model)**.**

1. **Invertibility**: $|\theta| < 1$.

2. **ACVF**:
$$\gamma_j = \begin{cases} \sigma^2(1+\theta^2), & \text{for } j = 0 \\ \sigma^2\theta, & \text{for } j = 1 \\ 0, & \text{for } j \geq 2 \end{cases}$$

3. **ACF**:
$$\rho_j = \begin{cases} 1, & \text{for } j = 0; \\ \dfrac{\theta}{1+\theta^2}, & \text{for } j = 1; \\ 0, & \text{for } j \geq 2 \end{cases}$$
   - The ACF **cuts off** at lag one, and $|\rho_k| \leq 0.5$, $(k \geq 1)$;
   - Identification problem: θ and $1/\theta$ produce the same value of ρ_1.

4. **PACF**:
$$\phi_{kk} = \frac{-(-\theta)^k(1-\theta^2)}{1-\theta^2(k+1)}, \quad \text{for } k \geq 1$$

5. **AR(∞) representation**:
$$\varepsilon_t = \sum_{j=0}^{\infty}(-\theta)^j y_{t-j}$$

6. **Moment estimation**: An estimator $\hat{\theta}$ of θ:
$$\frac{\hat{\theta}}{1+\hat{\theta}^2} = \hat{\rho}_1$$

2.2.2 MA(q) Model

Definition 2.7 (MA(q) Model)**.** If a stochastic process $\{y_t, t \in \mathbb{Z}\}$ satisfies

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}, \quad \varepsilon_t \sim \text{WN}(0, \sigma^2), \quad t \in \mathbb{Z}$$

where μ is the mean of the series, and the θ_i's are the parameters of the model, then $\{y_t\}$ is called a **moving average model of order** q, denoted by MA(q).

In backshift operator notation, let $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$, then $y_t - \mu = \theta(B)\varepsilon_t$.

Proposition 2.13 (Properties of MA(q) Model)**.**

1. **Mean**: The constant term μ.
2. **Weak stationarity**: Finite order MA models are always weakly stationary.
3. **Invertibility**: If and only if all of the roots of the MA characteristic equation $\theta(z) = 1+\theta_1 z+\cdots+\theta_q z^q = 0$ lie **outside the complex unit circle**.
4. **AR(∞) representation**:
$$\varepsilon_t = \frac{1}{\theta(B)}(y_t - \mu) = \frac{\mu}{1+\theta_1+\cdots+\theta_q} + \sum_{j=0}^{\infty}\psi_j y_{t-j}$$
5. **ACVF**:
$$\gamma_j = \begin{cases} \sigma^2(1+\theta_1^2+\cdots+\theta_q^2), & \text{for } j = 0, \\ \sigma^2\sum_{i=0}^{q-j}\theta_i\theta_{i+j}, & \text{for } j = 1, \ldots, q, \\ 0, & \text{for } j > q, \end{cases}$$
   where $\theta_0 = 1$.

6. **ACF**:

$$\rho_j = \begin{cases} \frac{\sum_{i=0}^{q-j} \theta_i \theta_{i+j}}{1+\theta_1^2+\cdots+\theta_q^2}, & \text{for } j = 0, \ldots, q, \\ 0, & \text{for } j > q \end{cases}$$

In other words, the ACF of an MA(q) is non-zero up to lag q and is **zero** afterwards. This is a special feature of MA processes and it provides a convenient way to identify an MA model in practice.

2.3 Auto Regressive Moving Average (ARMA)

2.3.1 ARMA(1,1) Model

Definition 2.8 (ARMA(1,1) Model)**.** An ARMA(1,1) model is defined as

$$y_t = c + \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{WN}(0, \sigma^2)$$

Proposition 2.14 (Properties of ARMA(1,1) Model)**.**

1. Weak stationarity: If and only if $|\phi| < 1$.
2. Invertibility: if and only if $|\theta| < 1$.
3. Mean: $\mu = \frac{c}{1-\phi}$.

Proposition 2.15 (Moment Equations of ARMA(1,1) Model)**.** Without loss of generality, assume zero mean and stationarity for ARMA(1,1) process. Multiplying the model by y_{t-j} and taking expectation yields:

$$\begin{aligned} & y_t = c + \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \\ \implies & \mathbb{E}(y_t y_{t-j}) = c\mathbb{E}(y_{t-j}) + \phi \mathbb{E}(y_{t-1} y_{t-j}) + \mathbb{E}(\varepsilon_t y_{t-j}) + \theta \mathbb{E}(\varepsilon_{t-1} y_{t-j}) \\ \implies & \text{Cov}(y_t, y_{t-j}) = 0 + \phi \text{Cov}(y_{t-1}, y_{t-j}) + \mathbb{E}(\varepsilon_t y_{t-j}) + \theta \mathbb{E}(\varepsilon_{t-1} y_{t-j}) \\ \implies & \gamma_j - \phi \gamma_{j-1} = \mathbb{E}(\varepsilon_t y_{t-j}) + \theta \mathbb{E}(\varepsilon_{t-1} y_{t-j}) \end{aligned}$$

By the fact that

$$\begin{aligned} \mathbb{E}(\varepsilon_t y_t) &= \mathbb{E}[\varepsilon_t(\phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1})] \\ &= \phi \mathbb{E}(\varepsilon_t y_{t-1}) + \mathbb{E}(\varepsilon_t^2) + \theta \mathbb{E}(\varepsilon_t, \varepsilon_{t-1}) \\ &= 0 + \sigma^2 + 0 = \sigma^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\varepsilon_{t-1} y_t) &= \mathbb{E}[\varepsilon_{t-1}(\phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1})] \\ &= \phi \mathbb{E}(\varepsilon_{t-1} y_{t-1}) + \mathbb{E}(\varepsilon_{t-1} \varepsilon_t) + \theta \mathbb{E}(\varepsilon_{t-1}^2) \\ &= \phi \sigma^2 + 0 + \theta \sigma^2 = (\phi + \theta)\sigma^2 \end{aligned}$$

We get

$$\gamma_j - \phi \gamma_{j-1} = \begin{cases} [1 + \theta(\phi + \theta)]\sigma^2, & \text{if } j = 0; \\ \theta \sigma^2, & \text{if } j = 1; \\ 0, & \text{if } j > 1 \end{cases}$$

Solving the first two equations produces

$$\gamma_0 = \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}\sigma^2, \quad \gamma_1 = \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}\sigma^2$$

and using the last recursively shows

$$\gamma_j = \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}\phi^{j-1}\sigma^2 \quad \text{for } j \geq 1$$

The ACF can then be computed as

$$\rho_j = \frac{(1 + \theta\phi)(\phi + \theta)}{1 + 2\theta\phi + \theta^2}\phi^{j-1} \quad \text{for } j \geq 1$$

The pattern here is similar to that for AR(1) model, except for the first term.

2.3.2 ARMA(p,q) Model

Definition 2.9 (ARMA(p,q) Model)**.** An ARMA(p, q) model is given by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

In lag operator notation,

$$\phi(B) y_t = \theta(B) \varepsilon_t$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$. It is assumed that the polynomials $\phi(B)$ and $\theta(B)$ do not have common factors, and $\phi_p \theta_q \neq 0$.

Proposition 2.16 (Properties of ARMA(p,q) Model)**.** The Properties of ARMA model can be summarized as follows:

1. **Weak stationarity**: If and only if all roots of the AR characteristic equation $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$ lie outside the complex unit circle.
2. **Invertibility**: If and only if all roots of the MA characteristic equation $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q = 0$ lie outside the complex unit circle.
3. **MA representation**: $y_t = \psi(B)\varepsilon_t$, where $\psi(B) = \frac{\theta(B)}{\phi(B)}$. The ψ-weight $\{\psi_i\}$ can be obtained by $\psi(B)\phi(B) \equiv \theta(B)$.
4. **AR representation**: $\varepsilon_t = \pi(B) y_t$, where $\pi(B) = \frac{\phi(B)}{\theta(B)}$. The π-weight $\{\pi_i\}$ can be obtained by $\pi(B)\theta(B) \equiv \phi(B)$.

Proposition 2.17 (Moment Equations of ARMA(p,q) Model)**.** Without loss of generality, assume zero mean and stationarity for ARMA(p,q) process.

- ACVF: Using the MA(∞) representation $y_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$,

$$\mathbb{E}(y_t \varepsilon_{t-j}) = \begin{cases} \sigma^2, & \text{for } j = 0; \\ \psi_j \sigma^2, & \text{for } j > 0; \\ 0, & \text{for } j < 0 \end{cases}$$

  Using the same technique as in ARMA(1,1), we have

$$\phi(B)\gamma_j = \begin{cases} (1 + \theta_1 \psi_1 + \cdots + \theta_q \psi_q)\sigma^2, & \text{for } j = 0; \\ (\theta_j + \theta_{j+1}\psi_1 + \cdots + \theta_q \psi_{q-j})\sigma^2, & \text{for } j = 1, \ldots, q; \\ 0, & \text{for } j > q, \end{cases}$$

  where $\psi_0 = 1$, and $\theta_j = 0$ for $j > q$.

- ACF: The correlation coefficient ρ_j satisfies the difference equation

$$\rho_j - \phi_1 \rho_{j-1} - \cdots - \phi_p \rho_{j-p} = 0, \text{ i.e,} \quad \phi(B)\rho_j = 0, \quad \text{for } j > q$$

  With $\rho_1, \ldots, \rho_q$ as the initial conditions, $\rho_j (j > q)$ can be recursively solved.

- Generalized Yule-Walker equation: Consider the above equations of ACF for $j = q+1, q+2, \ldots, q+p$, we have

$$\begin{bmatrix} \rho_{q+1} \\ \rho_{q+2} \\ \vdots \\ \rho_{q+p} \end{bmatrix} = \begin{bmatrix} \rho_q & \rho_{q-1} & \cdots & \rho_{q+2-p} & \rho_{q+1-p} \\ \rho_{q+1} & \rho_q & \cdots & \rho_{q+3-p} & \rho_{q+2-p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{q+p-1} & \rho_{q+p-2} & \cdots & \rho_{q+1} & \rho_q \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}$$

  which is referred to as a p-order generalized Yule-Walker equation for the ARMA(p, q) process. It can be used to solve for ϕ_i's given ρ_i's.

Remark 2.4 (Parsimony of ARMA(p,q) Model)**.** Low order ARMA(p,q) models with p and q less than 3 are generally sufficient for real world data analysis.

2.3.3 ARIMA(p,d,q) Model

Definition 2.10 (ARIMA(p,d,q) Model)**.** The specification of the ARMA(p,q) model assumes that y_t is **weakly stationary** and **ergodic**.

If y_t is a trending variable like an asset price or a macroeconomic aggregate like real GDP, then y_t must be transformed to stationary form by eliminating the trend.

Box and Jenkins (1976) advocated removal of trends by **differencing**. Let $\Delta = 1 - B$ denote the difference operator,

- If there is a linear trend in y_t then the first difference $\Delta y_t = y_t - y_{t-1}$ will not have a trend.
- If there is a quadratic trend in y_t, then Δy_t will contain a linear trend, but the second difference $\Delta^2 y_t = (1 - 2B + B^2)y_t = y_t - 2y_{t-1} + y_{t-2}$ will not have a trend.

The class of ARMA(p,q) models where the trends have been transformed by differencing d times is denoted ARIMA(p,d,q).

2.4 Seasonal Time Series Model

Definition 2.11 ($(\text{ARIMA}(p,d,q) \times (P,D,Q)_s$ Model)**.** A time series $\{y_t\}$ is said to be a multiplicative seasonal ARIMA model if it satisfies

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D(y_t-\mu) = \theta_q(B)\Theta_Q(B^s)\varepsilon_t, \quad \varepsilon_t \sim \text{WN}(0,\sigma^2)$$

where

$$\begin{array}{lll}
\phi_p(B) & = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, & \text{(regular AR factor/polynomial)} \\
\theta_q(B) & = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q, & \text{(regular MA factor/polynomial)} \\
\Phi_p(B^s) & = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{Ps}, & \text{(seasonal AR factor/polynomial)} \\
\Theta_Q(B^s) & = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}, & \text{(seasonal MA factor/polynomial)}
\end{array}$$

The process is denoted as $y_t \sim \text{ARIMA}(p,d,q) \times (P,D,Q)_s$.

Example 2.2 (Example of ARIMA$(0,1,1) \times (0,1,1)_{12}$ Model)**.** A time series $\{y_t\}$ is called the **airline model** if it satisfies

$$(1-B)(1-B^{12})y_t = (1-\theta B)(1-\Theta B^{12})\varepsilon_t$$

where $\{\varepsilon_t\} \sim \text{WN}(0,\sigma^2)$.

Example 2.3 (Example of ARIMA$(0,0,1) \times (1,0,0)_{12}$ Model)**.** Consider a seasonal model:

$$y_t = \Phi y_{t-12} + \varepsilon_t - \theta\varepsilon_{t-1}, \quad |\Phi| < 1, \quad \{\varepsilon_t\} \sim \text{WN}(0,\sigma^2)$$

Using the standard techniques, it follows that

$$\gamma_0 = \left(\frac{1+\theta^2}{1-\Phi^2}\right)\sigma^2, \quad \gamma_1 = \Phi\gamma_{11} - \theta\sigma^2, \quad \gamma_k = \Phi\gamma_{k-12} \quad \text{for } k \geq 2$$

Thus,

$$\rho_{12k} = \Phi^k \quad \text{for } k \geq 1$$

$$\rho_{12k-1} = \rho_{12k+1} = -\frac{\theta}{1+\theta^2}\Phi^k \quad \text{for } k = 0,1,2,\ldots$$

3 Forecasting

Objective: **predict future values** of a time series, y_{t+h}, $h = 1,2,\ldots$, based on the data collected to the present, $I_t = \{y_t, y_{t-1}, \ldots\}$.

Let $y_{t+h|t}$ denote a forecast of y_{t+h} made at time t, which has an associated forecast error or prediction error $\varepsilon_{t+h|t}$, i.e.,

$$\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t}$$

3.1 Forecasting Criterions

Definition 3.1 (MSE Criterion)**.**

$$\text{MSE}(\varepsilon_{t+h|t}) \equiv \mathbb{E}[\varepsilon_{t+h|t}^2] = \mathbb{E}[(y_{t+h} - y_{t+h|t})^2]$$

Proposition 3.1 (Minimum MSE Forecast)**.** The **minimum MSE forecast** (optimal forecast) of y_{t+h} based on I_t is

$$y_{t+h|t} = \mathbb{E}(y_{t+h}|I_t)$$

$y_{t+h|t}$ is used to denote the minimum MSE forecast in the literature.

Proof:

$$\begin{aligned}
\mathbb{E}[(y_{t+h} - y_{t+h|t})^2] &= \mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t) + \mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]^2\Big\} \\
&= \mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big]^2\Big\} + \mathbb{E}\Big\{\Big[\mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]^2\Big\} \\
&\quad + 2\mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big] \underbrace{\Big[\mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]}_{\text{no randomness to } y_t}\Big\} \\
&= \mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big]^2\Big\} + \mathbb{E}\Big\{\Big[\mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]^2\Big\} \\
&\quad + 2\mathbb{E}\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big] \underbrace{\Big[\mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]}_{=0} \\
&= \mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big]^2\Big\} + \mathbb{E}\Big\{\Big[\mathbb{E}(y_{t+h}|I_t) - y_{t+h|t}\Big]^2\Big\} \\
&\geq \mathbb{E}\Big\{\Big[y_{t+h} - \mathbb{E}(y_{t+h}|I_t)\Big]^2\Big\}
\end{aligned}$$

Here the equality holds if and only if $y_{t+h|t} = \mathbb{E}(y_{t+h}|I_t)$.

Definition 3.2 (Forecast Based on Linear Projection)**.** Using the linear combination of past observations, the forecast value is given by

$$y_{t+h|t}^* = \sum_{j=0}^{\infty} c_j y_{t-j}$$

where $\{c_j\}$ is determined by

$$\{c_j\} = \arg\min_{\{\beta_k\}} \mathbb{E}\Big[\Big(y_{t+h} - \sum_{k=0}^{\infty} \beta_k y_{t-k}\Big)^2\Big]$$

This is the linear projection of the forecast on the space spanned by past observations $y_t, y_{t-1}, y_{t-2}, \ldots$.

Definition 3.3 (Innovation Sequences*)**.**

$$\{e_t\} = \{y_t - y_{t|t-1}\}$$

$$\{e_t^*\} = \{y_t - y_{t|t-1}^*\}$$

Proposition 3.2 (Properties of Innovation Sequences*)**.** Suppose $\{y_t\}$ is weakly stationary with zero mean and finite variance. Then

1. $\{e_t\}$ and $\{e_t^*\}$ are weakly stationary with zero mean and finite variance.
2. $\mathbb{E}(e_t|I_{t-1}) = 0$ a.s.
3. $\mathbb{E}(e_t^* \cdot \sum_{j=1}^{k} c_j y_{t-j}) = 0$ for any $k \geq 1$ and any real numbers $\{c_j\}$.
4. $\mathbb{E}[(y_t - y_{t|t-1})^2] \leq \mathbb{E}[(y_t - y_{t|t-1}^*)^2]$.

Remark 3.1 (Linear and Nonlinear Optimal Forecast)**.** In some models, the minimum MSE forecast is exactly the best linear forecast, i.e., $y_{t|t-1} = y^*_{t|t-1}$. AR, MA, ARMA, ARIMA models all have this nice property, since their conditional expectations are the linear combination of past observations. But this doesn't hold true for all models. For instance, ARCH and GARCH's best linear forecast are not necessarily the minumum MSE forecast.

Remark 3.2 (Richness of Models: A Point of View from Forecasting)**.**

$$y_{t|t-1} = \mathbb{E}(y_t|I_{t-1}) = \mathbb{E}(y_t|y_{t-1}, y_{t-2}, \ldots) = \varphi(y_{t-1}, y_{t-2}, \ldots)$$

$$\mathrm{Var}(y_t|I_{t-1}) = \mathbb{E}\{(y_t - y_{t|t-1})^2|I_{t-1}\} = \sigma^2(y_{t-1}, y_{t-2}, \ldots)$$

That is,

$$y_t = \varphi(y_{t-1}, y_{t-2}, \ldots) + \varepsilon_t\sigma(y_{t-1}, y_{t-2}, \ldots)$$

3.2 Forecasting of AR

Proposition 3.3 (1-step-ahead Forecast of AR(p) Model)**.** From the AR(p) model, we have

$$y_{t+1} = c + \phi_1 y_t + \cdots + \phi_p y_{t+1-p} + \varepsilon_{t+1}, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma^2)$$

1. Under the MSE loss function, the point forecast of y_{t+1} given $I_t = \{y_t, y_{t-1}, \ldots\}$ is

$$y_{t+1|t} = \mathbb{E}[y_{t+1}|I_t] = c + \phi_1 y_t + \cdots + \phi_p y_{t+1-p}$$

2. The 1-step-ahead forecast error is

$$\varepsilon_{t+1|t} = y_{t+1} - y_{t+1|t} = \varepsilon_{t+1}$$

3. The variance of 1-step-ahead forecast error is

$$\mathrm{Var}(\varepsilon_{t+1|t}) = \mathrm{Var}(\varepsilon_{t+1}) = \sigma^2$$

Proposition 3.4 (Multistep-ahead Forecast of AR(p) Model)**.** In general, the h-step ahead expression of AR(p) process is

$$y_{t+h} = c + \phi_1 y_{t+h-1} + \cdots + \phi_p y_{t+h-p} + \varepsilon_{t+h}$$

1. Taking conditional expectation, we have the h-step ahead forecast of AR(p) model:

$$y_{t+h|t} = \mathbb{E}(y_{t+h}|I_t) = c + \sum_{i=1}^{p} \phi_i y_{t+h-i|t}$$

   Here we define $y_{t+\ell|t} = y_{t+\ell}$ if $\ell \leq 0$. This forecast can be computed recursively using forecasts $y_{t+i|t}$ for $i = 1, \ldots, h-1$.

2. The h-step-ahead forecast error is

$$\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t}$$

3. The forecasting interval is not straightforward for the AR(p) model. We will use the method similar to that of the ARMA model.

Example 3.1 (Multistep-ahead Forecast of AR(1) Model)**.** The AR(1) model can be written as

$$y_t - \mu = \phi(y_{t-1} - \mu) + \varepsilon_t$$

The general form of the forecast equation is

$$y_{t+h|t} = \mu + \phi^h(y_t - \mu), \quad h \geq 1$$

Example 3.2 (Multistep-ahead Forecast of AR(2) Model)**.** The AR(2) model can be written as

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \varepsilon_t$$

The general form of the forecast equation is given recursively as

$$\begin{aligned}
y_{t+1|t} - \mu &= \phi_1(y_t - \mu) && +\phi_2(y_{t-1} - \mu) \\
y_{t+2|t} - \mu &= \phi_1(y_{t+1|t} - \mu) && +\phi_2(y_t - \mu) && = (\phi_1^2 + \phi_2)(y_t - \mu) && +\phi_1\phi_2(y_{t-1} - \mu) \\
y_{t+3|t} - \mu &= \phi_1(y_{t+2|t} - \mu) && +\phi_2(y_{t+1|t} - \mu) && = (\phi_1^3 + 2\phi_1\phi_2)(y_t - \mu) + (\phi_1^2\phi_2 + \phi_2^2)(y_{t-1} - \mu) \\
&\vdots \\
y_{t+h|t} - \mu &= \phi_1(y_{t+h-1|t} - \mu) + \phi_2(y_{t+h-2|t} - \mu)
\end{aligned}$$

3.3 Forecasting of MA

Proposition 3.5 (Forecast of MA(q) Model)**.** Consider the MA(q) model with $\theta_0 \equiv 1$:

$$y_{t+h} = \sum_{i=0}^{q} \theta_i \varepsilon_{t+h-i}$$

1. Using the IID properties of ε_t, the optimal h-step forecast is

$$y_{t+h|t} = \mathbb{E}(y_{t+h}|I_t) = \begin{cases} \sum_{i=h}^{q} \theta_i \varepsilon_{t+h-i}, & \text{for } h = 1, \dots, q, \\ 0, & \text{for } h > q, \end{cases}$$

2. The h-step forecast error is

$$\begin{aligned}
\varepsilon_{t+h|t} &= \begin{cases} \sum_{i=0}^{h-1} \theta_i \varepsilon_{t+h-i}, & \text{for } h = 1, \dots, q, \\ \sum_{i=0}^{q} \theta_i \varepsilon_{t+h-i}, & \text{for } h > q, \end{cases} \\
&= \sum_{i=0}^{h-1} \theta_i \varepsilon_{t+h-i}, \qquad \text{define } \theta_i = 0 \text{ for } i > q
\end{aligned}$$

3. The MSE of the h-step forecast error is

$$\text{MSE}(\varepsilon_{t+h|t}) = \mathbb{E}(\varepsilon_{t+h|t}^2) = \sigma^2 \sum_{i=0}^{h-1} \theta_i^2$$

4. Assuming normality, the 95% confidence interval for predicted y_{t+h} is

$$\left(y_{t+h|t} - 1.96\sqrt{\text{MSE}(\varepsilon_{t+h|t})}, \quad y_{t+h|t} + 1.96\sqrt{\text{MSE}(\varepsilon_{t+h|t})}\right)$$

3.4 Forecasting of ARMA

Proposition 3.6 (Forecast of ARMA(p,q) Model)**.** The ARMA(p,q) model for y_t is

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

or, in lag operator form, $\phi_p(B)y_t = \theta_q(B)\varepsilon_t$.

- The true value of h-step ahead forecast is

$$y_{t+h} = \phi_1 y_{t+h-1} + \cdots + \phi_p y_{t+h-p} + \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + \cdots + \theta_q \varepsilon_{t+h-q}.$$

  The optimal h-step ahead forecast is

$$y_{t+h|t} = \phi_1 y_{t+h-1|t} + \cdots + \phi_p y_{t+h-p|t} + \theta_1 \varepsilon_{t+h-1|t} + \cdots + \theta_q \varepsilon_{t+h-q|t}$$

  We define $y_{t+\ell|t} = y_{t+\ell}$ if $\ell \leq 0$; $\varepsilon_{t+\ell|t} = 0$ if $\ell > 0$, otherwise $\varepsilon_{t+\ell|t} = \varepsilon_{t+\ell}$. We can compute the optimal h-step forecast by the above formula recursively.

- It is convenient to rewrite the model as an MA(∞) model, that is,

$$y_t = \phi_p(B)^{-1}\theta_q(B)\varepsilon_t = \varepsilon_t + \eta_1\varepsilon_{t-1} + \eta_2\varepsilon_{t-2} + \eta_3\varepsilon_{t-3} + \cdots$$

  The h-step ahead forecast error is given by

$$\varepsilon_{t+h|t} = \varepsilon_{t+h} + \eta_1\varepsilon_{t+h-1} + \cdots + \eta_{h-1}\varepsilon_{t+1} = \sum_{i=0}^{h-1}\eta_i\varepsilon_{t+h-i}$$

- The MSE of the h-step ahead forecast error is

$$\text{MSE}(\varepsilon_{t+h|t}) = \sigma^2\sum_{i=0}^{h-1}\eta_i^2$$

  Note that as $h \to \infty$, $\text{MSE}(\varepsilon_{t+h|t}) \to \sigma^2\sum_{i=0}^{\infty}\eta_i^2 = \text{Var}(y_t)$

3.5 Forecasting of ARIMA

Proposition 3.7 (Forecast of ARIMA(p,d,q) Model)**.** The ARIMA(p,d,q) model is:

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\varepsilon_t$$

If invertible, then write the process as $\pi(B)y_t = \varepsilon_t$, where

$$\pi(B) = \frac{\phi_p(B)(1-B)^d}{\theta_q(B)} = 1 - \sum_{j=1}^{\infty}\pi_j B^j$$

The h-step ahead expression of ARIMA process can be written as

$$y_{t+h} = \sum_{j=1}^{\infty}\pi_j y_{t+h-j} + \varepsilon_{t+h}$$

Applying the operator $1 + \psi_1 B + \ldots + \psi_{h-1}B^{h-1}$ to both sides of the above equation, we have

$$\sum_{j=0}^{\infty}\sum_{k=0}^{h-1}\pi_j\psi_k y_{t+h-j-k} + \sum_{k=0}^{h-1}\psi_k\varepsilon_{t+h-k} = 0$$

where $\pi_0 = -1$ and $\psi_0 = 1$. By changing the order of summation[a], it can be shown that

$$\sum_{j=0}^{\infty}\sum_{k=0}^{h-1}\pi_j\psi_k y_{t+h-j-k} = \pi_0 y_{t+h} + \sum_{m=1}^{h-1}\Big(\sum_{i=0}^{m}\pi_{m-i}\psi_i\Big)y_{t+h-m} + \sum_{j=1}^{\infty}\Big(\sum_{i=0}^{h-1}\pi_{h-1+j-i}\psi_i\Big)y_{t+1-j}$$

We can choose $\{\psi_i\}$ weights so that

$$\sum_{i=0}^{m}\pi_{m-i}\psi_i = 0, \quad m = 1, 2, \ldots, h-1$$

The ψ_i weights is given recursively by

$$\psi_0 = 0, \quad \psi_j = \sum_{i=0}^{j-1}\pi_{j-i}\psi_i, \quad j = 1, 2, \ldots, h-1$$

[a] We can sum the expression in a diagonal manner. First we draw the term where $i = 0, k = 0$, which yields $\pi_0\psi_0 y_{t+h}$ ($\pi_0 = -1, \psi_0 = 1$). Then we sum the terms where the indices of π and ψ sum to $1, 2, \ldots, h-1$, hence we get the term $\sum_{m=1}^{h-1}\Big(\sum_{i=0}^{m}\pi_{m-i}\psi_i\Big)y_{t+h-m}$. After that, we sum the terms where the indices of π and ψ sum to $h, h+1, h+2, \ldots$, thus we get the term $\sum_{j=1}^{\infty}\Big(\sum_{i=0}^{h-1}\pi_{h-1+j-i}\psi_i\Big)y_{t+1-j}$.

Define

$$\pi_j^{(h)} = \sum_{i=0}^{h-1} \pi_{h-1+j-i}\psi_i, \quad j = 1, 2, 3, \ldots$$

we have

$$y_{t+h} = \sum_{j=1}^{\infty} \pi_j^{(h)} y_{t-j+1} + \sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-i}$$

1. The h-step optimal forecast is
$$y_{t+h|t} = \sum_{j=1}^{\infty} \pi_j^{(h)} y_{t-j+1}$$
2. The forecast error is
$$\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t} = \sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-i}$$
3. The MSE of the forecast error is
$$\mathrm{Var}(\varepsilon_{t+h|t}) = \sigma^2 \sum_{i=0}^{h-1} \psi_i^2$$

   Note that as $h \to \infty$, $\mathrm{Var}(\varepsilon_{t+h|t}) \to \infty$.

Proposition 3.8 (Forecasts Updating Equation)**.** The h-step forecast error at time t is

$$e_t(h) = y_{t+h} - y_{t+h|t} = \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j}$$

Substitute t by $t-1$, h by $h+1$, it's easily to derive

$$e_{t-1}(h+1) = e_t(h) + \psi_h \varepsilon_t$$

Hence the updating equation is given by

$$y_{t+h|t} = y_{t+h|t-1} + \psi_h(y_t - y_{t|t-1})$$

$$y_{t+h|t+1} = y_{t+h|t} + \psi_{h-1}(y_{t+1} - y_{t+1|t})$$

If a new observation at time $t+1$ is observed, it can be used to update the h-step ahead forecast as above.

4 Model Building

4.1 Box-Jenkins Approach

Definition 4.1 (Box-Jenkins Approach)**.** An effective procedure for building empirical time series models is the Box-Jenkins Approach (Box and Jenkins, 1976):

1. Model specification,
2. Parameter estimation,
3. Statistical model checking (Diagnostics).

Forecasts follow directly from the fitted model.

Remark 4.1 (Box-Jenkins Approach for ARIMA Models)**.**

1. Input and plot the time series data and choose proper transformation (e.g., logarithm, logit, and variance stabilization transformation).
2. Compute and examine the sample ACF and the sample PACF of the data to further confirm a necessary degree of differencing.

- Check stationarity of the transformed time series data. If the transformed data has unit roots, then difference it until the differenced transformed data do not have unit root anymore. Denote d as the differencing times.

3. Determine the order values of p and q preliminarily for the differenced series. For ARMA models, there are two main approaches to model specification.
   - The first approach is called the "correlation" approach such as the ACF, PACF and extended autocorrelation function (EACF).
   - The second approach is called the information criterion approach such as AIC, BIC, HQ.
4. Estimate the ARIMA(p, d, q) model for the transformed data.
5. Perform diagnostic analysis to confirm whether the transformed ARIMA(p,d,q) model adequately describes the data (e.g. examine residuals from the fitted model).
6. Respecify the ARIMA(p, d, q) model if necessary.
7. Use the fitted model for descriptive or forecasting purposes.

4.2 Model Specification

4.2.1 Correlation approach

Definition 4.2 (Extended ACF (EACF))**.** For an observed time series $\{y_t\}$, the EACF is given through the following steps:

1. **Iterative Regression:** For each candidate AR order i and MA order j:
   - First, regress y_t on $y_{t-1}, \ldots, y_{t-i}$ to get residuals $\tilde{e}_t$.
   - Then, regress y_t on $y_{t-1}, \ldots, y_{t-i}$ and $\tilde{e}_{t-1}, \ldots, \tilde{e}_{t-j}$ to get consistent estimators of AR coefficients, $\tilde{\phi}_1, \ldots, \tilde{\phi}_i$.
2. **Filter Series:** Compute the filtered series $\{w_t\}$ as
$$w_t = y_t - \tilde{\phi}_1 y_{t-1} - \cdots - \tilde{\phi}_k y_{t-i}$$
If the hypothesized AR order (i) is actually the correct AR order (p), and if the hypothesized MA order $j \geq q$, then $\{w_t\}$ is an MA(q) process. In that case, the true autocorrelation of $\{w_t\}$ of lag $q+1$ or higher should be zero.
3. **Sample Autocorrelation Significance Test:** Calculate the sample autocorrelation of $\{w_t\}$ at lag $m = j + 1$:
$$\hat{\rho}_{i,j}(m) = \frac{\sum_{t=m+1}^{n}(w_t - \bar{w})(w_{t-m} - \bar{w})}{\sum_{t=1}^{n}(w_t - \bar{w})^2}$$
Under the null hypothesis that $\{w_t\}$ is a MA(q) process ($j \geq q$):
$$\rho_{i,j}(m) \sim \mathcal{N}(0, 1/\sqrt{n - i - j})$$
If $|\hat{\rho}_{i,j}(m)| > 1.96/\sqrt{n - i - j}$, reject $H_0 : \rho_{i,j}(m) = 0$.
4. **EACF Table**: Run grid search for AR orders $i = 0, 1, 2, \ldots$ and MA orders $q = 0, 1, 2, \ldots$, and fill the table entry (i,j) with
$$\begin{cases} \text{X}, & \text{if } \hat{\rho}_{i,j}(j+1) \text{ is significant} \\ \text{O}, & \text{if } \hat{\rho}_{i,j}(j+1) \text{ is not significant} \end{cases}$$
A typical EACF table has the form:

p/q	0	1	2	3	4	$\cdots$
0	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	$\cdots$
1	$\rho_{1,1}$	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$	$\rho_{1,5}$	$\cdots$
2	$\rho_{2,1}$	$\rho_{2,2}$	$\rho_{2,3}$	$\rho_{2,4}$	$\rho_{2,5}$	$\cdots$
3	$\rho_{3,1}$	$\rho_{3,2}$	$\rho_{3,3}$	$\rho_{3,4}$	$\rho_{3,5}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

5. **Order Identification:** For a true ARMA(p,q) process, entries form a triangle of zeros where the top-left ○ appears at $(k, j) = (p, q)$. By locating such triangles, the AR and MA orders can be determined.

p/q	0	1	2	3	4	5	⋯
0	X	X	X	X	X	X	⋯
1	X	○	○	○	○	○	⋯
2	X	X	○	○	○	○	⋯
3	X	X	X	○	○	○	⋯
4	X	X	X	X	○	○	⋯
5	X	X	X	X	X	○	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 1: ARMA(1,1)

p/q	0	1	2	3	4	5	⋯
0	X	X	X	X	X	X	⋯
1	X	X	X	X	X	X	⋯
2	X	X	X	○	○	○	⋯
3	X	X	X	X	○	○	⋯
4	X	X	X	X	X	○	⋯
5	X	X	X	X	X	X	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 2: ARMA(2,3)

4.2.2 Model Selection Criteria

Remark 4.2 (Model Selection Criteria)**.** The idea of model selection is to fit all ARMA(p, q) models with orders $p \leq p_{\text{max}}$ and $q \leq q_{\text{max}}$ and choose the values of p and q which minimizes some model selection criteria.

Model selection criteria for ARMA(p, q) models have the form

$$\text{MSC}(p, q) = \ln(\tilde{\sigma}^2(p, q)) + c_n \cdot \psi(p, q)$$

- $\tilde{\sigma}^2(p, q)$ is the LSE of $\text{Var}(\varepsilon_t) = \sigma^2$ without a degrees of freedom correction from the ARMA(p,q) model.
- c_n is a sequence indexed by the sample size n.
- $\psi(p, q)$ is a penalty function which penalizes ARMA(p, q) models with a high order.

The two most common information criteria are the **Akaike Information Criteria (AIC)** and **Schwarz-Bayesian Information Criteria (BIC)**:

$$\text{AIC}(p, q) = \ln(\tilde{\sigma}^2(p, q)) + \frac{2}{n}(p + q)$$

$$\text{BIC}(p, q) = \ln(\tilde{\sigma}^2(p, q)) + \frac{\ln n}{n}(p + q)$$

Lower AIC or BIC value indicates better performance.

4.3 Residual Diagnostics

4.3.1 Residual Autocorrelation

Remark 4.3 (Methods to Test Residual ACF)**.** Under the null hypothesis H_0 where the residuals have no autocorrelation:

1. The sample ACF of the residuals has the distribution

$$r_k(\hat{\varepsilon}) = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=k+1}^{n} \hat{\varepsilon}_t^2} \overset{H_0}{\sim} \mathcal{N}(0, 1/n).$$

2. The joint significance of the first m residual ACF can be measured by the test-statistic developed by Ljung and Box (1978), which is given as

$$\text{LB}(m) = n(n + 2) \sum_{k=1}^{m} \frac{[r_k(\hat{\varepsilon})]^2}{n - k} \overset{H_0}{\sim} \chi^2_{m-p-q}$$

4.3.2 Residual Homoscedasticity

Remark 4.4 (Importance of Residual Heteroscedasticity)**.** Neglecting heteroscedasticity of the residuals leads to

- Ordinary t-statistics cannot be used.
- Many diagnostic tests, such as tests for nonlinearity, are affected.
- Confidence intervals for forecasts can no longer be computed in the usual manner.

Definition 4.3 (Portmanteau Test)**.** Portmanteau test statistic on (standardized) squared residuals developed by McLeod and Li (1983) is:

$$\mathrm{McL}(m) = n(n+2)\sum_{k=1}^{m}\frac{[r_k(\hat{\varepsilon}^2)]^2}{n-k}$$

When applied to the residuals from an ARMA(p,q) model, the McL test has an asymptotic χ^2_{m-p-q} distribution, provided that m/n is small and m is moderately large.

4.3.3 Residual Normality

Remark 4.5 (Skewness and Kurtosis Test of Residual Normality)**.** Defining the jth moment of the estimated (standardized) residuals:

$$\hat{m}_j = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^j$$

The skewness and kurtosis of $\hat{\varepsilon}_t$ are

$$\hat{S}_{\hat{\varepsilon}} = \frac{\hat{m}_3}{\hat{m}_2^{3/2}}, \qquad \hat{K}_{\hat{\varepsilon}} = \frac{\hat{m}_4}{\hat{m}_2^2}$$

Under the null hypothesis of normality, we have

$$\sqrt{n/6}\cdot\hat{S}_{\hat{\varepsilon}} \sim \mathcal{N}(0,1), \qquad \sqrt{n/24}\cdot(\hat{K}_{\hat{\varepsilon}}-3) \sim \mathcal{N}(0,1)$$

A joint test for normality (Jarque and Bera, 1987) is then given by

$$\mathrm{JB} = \frac{n}{6}\hat{S}_{\hat{\varepsilon}}^2 + \frac{n}{24}(\hat{K}_{\hat{\varepsilon}}-3)^2 \sim \chi_2^2$$

4.4 Intervention Analysis*

Introduced by Box and Tiao (1975), intervention analysis provides a framework for assessing the effect of an intervention on a time series. Assumption: A stochastic process is affected by an intervention by changing the mean or trend.

Definition 4.4 (Intervention Models)**.** The general model for the time series $\{y_t\}$, perhaps after suitable transformation, is given by

$$y_t = m_t + N_t$$

- m_t is the change in the mean function;
- N_t is modeled as some ARIMA process, possibly seasonal;

$\{N_t\}$ represents the underlying time series where there is **no intervention**. It is referred to as the natural or unperturbed process, and may be stationary or nonstationary, seasonal or nonseasonal.

Suppose the time series is subject to an intervention that takes place at time T. Before T, m_t is assumed to be identically zero. The time series $\{y_t, t<T\}$ is referred to as the **preintervention data** and can be used to specify the model for the unperturbed process $\{N_t\}$.

Definition 4.5 (Two Common Types of Intervention Variables)**.**

- **Step function**:

$$S_t^{(T)} = \begin{cases} 1, & t \geq T, \\ 0, & \text{otherwise} \end{cases}$$

- **Pulse function**:

$$P_t^{(T)} = \begin{cases} 1, & t = T, \\ 0, & t \neq T \end{cases}$$

Relationship: $P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} \equiv (1-B)S_t^{(T)}$.

Proposition 4.1 (Modeling by the Step Function)**.**

- If the intervention results in an immediate and permanent shift in the mean function, the shift can be modeled as
$$m_t = \omega S_t^{(T)}$$
where ω is the unknown permanent change in the mean due to the intervention.
- If there is a delay of d time units before the intervention takes effect and d is known, then we can specify
$$_t = \omega S_{t-d}^{(T)}$$
- The intervention may affect the mean function gradually, with its full force reflected only in the long run. This can be modeled by specifying m_t as an AR(1)-type model with the error term replaced by a multiple of the lag 1 of $S_{t-1}^{(T)}$:

$$m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)} = \sum_{j=0}^{\infty} \delta^j \omega S_{t-1-j}^{(T)} = \begin{cases} \dfrac{\omega}{1-\delta}(1-\delta^{t-T}), & t > T, \\ 0, & \text{otherwise.} \end{cases}, \quad \delta \in (0,1], \quad m_0 = 0$$

  - Often, $\delta \in (0,1)$. In this case, $m_t \to \frac{\omega}{1-\delta}$, which is the ultimate change (gain or loss) for the mean function.
  - If $\delta = 1$, then $m_t = \omega(t-T)$ for $t \geq T$ and 0 otherwise. The time sequence plot of m_t displays the shape of a ramp with slope ω. This specification implies that the intervention changes the mean function linearly in the post-intervention period.

Proposition 4.2 (Modeling by the Pulse Function)**.**

- An immediate intervention may be modeled as:
$$m_t = \omega P_t^{(T)}$$
The intervention impacts the mean function only at $t = T$.
- Intervention effects that die out gradually may be specified via the AR(1)-type specification
$$m_t = \delta m_{t-1} + \omega P_t^{(T)}, \quad m_0 = 0$$
i.e., $m_t = \omega\delta^{t-T}$ for $t \geq T$.
- Delayed changes can be incorporated by lagging the pulse function:
$$m_t = \delta m_{t-1} + \omega P_{t-1}^{(T)}, \quad m_0 = 0$$
or written in lag operator form,
$$m_t = \frac{\omega B}{1-\delta B} P_t^{(T)}$$
- More sophisticated intervention effects can be modeled in a mixed way. For example,
$$m_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)}$$
More generally, we can model the change in the mean function by an ARMA-type specification
$$m_t = \frac{\omega(B)}{\delta(B)} P_t^{(T)}$$

Remark 4.6 (Difference between Mean Test and Intervention Analysis)**.** Similar to testing whether the population means are the same with data in the form of two independent random samples from the two populations. However, the major difference here is that the pre- and post-intervention data cannot generally be assumed to be independent and identically distributed. The inherent serial correlation in the data makes the problem more interesting but at the same time more difficult.

4.5 Time Series Outliers*

Outliers refer to atypical observations that may arise because of measurement and/or copying errors or because of abrupt, short-term changes in the underlying process.

Definition 4.6 (Additive Outliers and Innovative Outliers)**.**

- **Additive Outlier**: An AO occurs at time T if the underlying process is perturbed additively at time T so that the data equal
$$y_t' = y_t + \omega_A P_t^{(T)}$$
where $\{y_t\}$ is the unperturbed process (no outliers), $\{y_t'\}$ the observed process that may be affected by some outliers. That is,
$$y_T' = y_T + \omega_A, \quad y_t' = y_t, \quad \text{for} \quad t \neq T$$
An AO can also be treated as an intervention that has a pulse response at T so that $m_t = \omega_A P_t^{(T)}$.
- **Innovative Outlier**: An innovative outlier occurs at time T if the error (also known as an innovation) is perturbed, so $e_T' = e_T + \omega_I$ and $e_t' = e_t$ for $t \neq T$. Suppose that the unperturbed process is stationary and admits an MA(∞) representation
$$y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots$$
Consequently, the perturbed process can be written as
$$\begin{aligned} y_t' &= \varepsilon_t' + \psi_1 \varepsilon_{t-1}' + \psi_2 \varepsilon_{t-2}' + \cdots \\ &= \{\varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots\} + \psi_{t-T}\omega_1 \\ &= y_t + \psi_{t-T}\omega_1 \end{aligned}$$
where $\psi_0 = 1$ and $\psi_j = 0$ for $j < 0$. Thus, an innovative outlier at T perturbs all observations on and after T, although with diminishing effect.

Proposition 4.3 (Detection of Time Series Outliers)**.** To detect whether an observation is an AO or IO, use AR(∞) representation to define the residuals
$$a_t = y_t' - \pi_1 y_{t-1}' - \pi_2 y_{t-2}' - \cdots$$
For simplicity, we assume the process has zero mean and that the parameters are known.

- **Innovative Outlier**: If the series has exactly one IO at time T, then the residual $a_T = \omega_I + \varepsilon_T$ and $a_t = \varepsilon_t$ for $t \neq T$. So ω_I can be estimated by $\hat{\omega}_I = a_T$ with variance σ^2.

  A test statistic for testing for an IO at a known T is
$$\lambda_{I,T} = \frac{a_T}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$
under H_0 that there are no outliers. At the 5% significance level, if $|\lambda_{I,T}| > 1.96$, reject H_0, otherwise, accept H_0.

  Generally T is unknown, and a simple conservative procedure is to use the Bonferroni rule for controlling the overall error rate of multiple tests. Let
$$\lambda_I = \max_{1 \leq t \leq n} \lambda_{I,t}$$

be attained at $t = T$, then the Tth observation is deemed an IO if λ_I exceeds the upper $0.025/n \times 100$ percentile of $\mathcal{N}(0, 1)$. This procedure guarantees that there is at most a 5% probability of a false detection of an IO.

Note that σ can be more robustly estimated by the mean absolute residual times $\sqrt{2/\pi}$.

- **Additive Outlier**: Suppose that the process admits an AO at T and is otherwise free of outliers. Then the residual has the representation

$$a_t = -\omega_A \pi_{t-T} + \varepsilon_t, \quad \pi_0 = -1, \quad \pi_j = 0, \quad j < 0$$

A LSE of ω_A is

$$\hat{\omega}_{A,T} = -\rho^2 \sum_{t=1}^{n} \pi_{t-T} a_t$$

where $\rho^2 = (1 + \pi_1^2 + \ldots + \pi_{n-T}^2)^{-1}$, with the variance of the estimate being equal to $\rho^2\sigma^2$.

A test statistic for testing the H_0 that the time series has no outliers versus the alternative hypothesis of an AO at T is defined as

$$\lambda_{A,T} = \frac{\hat{\omega}_{A,T}}{\rho\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

If T is unknown, similar to the above discussion, we can use Bonferroni rule to control the overall error rate of multiple tests.

In the case where an outlier is detected at T, it may be classified as an IO if $|\lambda_{I,T}| > |\lambda_{A,T}|$ and an AO otherwise.

4.6 Spurious Correlation and Prewhitening*

Definition 4.7 (Cross Correlation Function)**.** Let $Y = \{y_t\}$ be the time series of the response variable, $X = \{x_t\}$ be a covariate time series that we hope will help explain or forecast Y. The cross-covariance function is defined as $\gamma_{X,Y}(t, s) = \text{Cov}(x_t, y_s)$, and the cross-correlation function (CCF) between X and Y at lag k can then be defined by

$$\rho_{X,Y}(k) = \text{corr}(x_t, y_{t-k}) = \text{corr}(x_{t+k}, y_t)$$

which measures the linear association between x_t and y_{t-k}.

The CCF can be estimated by the sample CCF (SCCF) defined by

$$r_{X,Y}(k) = \frac{\sum(x_t - \bar{x})(y_{t-k} - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2 \sum(y_{t-k} - \bar{y})^2}}$$

The covariate X is independent of Y if and only if $\beta_1 = 0$, in which case the SCCF $r_{X,Y}(k)$ is approximately $\mathcal{N}(0, \frac{1}{n})$, where n is the sample size — the number of pairs of (x_t, y_t) available.

Definition 4.8 (Spurious Correlation)**.** Consider a regression model:

$$y_t = \beta_0 + \beta_1 x_{t-d} + z_t$$

where $\{z_t\}$ may follow some ARIMA(p,d,q) model. The variance of $\sqrt{n} r_{X,Y}(k)$ is approximately

$$1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)$$

Particularly, if X and Y are both AR(1) processes with AR(1) coefficients ϕ_X and ϕ_Y, respectively. Then

$$\sqrt{n} r_{X,Y}(k) \xrightarrow{d} \mathcal{N}\big(0, \frac{1 + \phi_X\phi_Y}{1 - \phi_X\phi_Y}\big)$$

When both AR(1) coefficients are close to 1, the ratio of the sampling variance of $r_{X,Y}(k)$ to the nominal value of $1/n$ approaches infinity. Thus, the unquestioned use of the $1/n$ rule in deciding the significance of the SCCF may lead to many more false positives than the nominal 5% error rate, even though the response and covariate time series are independent of each other. The problem of inflated variance of the SCCF becomes more acute for nonstationary data.

Definition 4.9 (Prewhitening)**.** **Prewhitening** is a useful technique to model the cross correlation of two time series $X = \{x_t\}$ and $Y = \{y_t\}$ with higher fidelity.

Idea:

1. Fit X by some time series model, e.g., ARIMA(p,d,q).
2. Obtain an AR(∞) representation, i.e., $a_t = \pi(B)x_t$.
3. Transform Y by the same filter $\pi(B)$, i.e., $\tilde{y}_t = \pi(B)y_t$.
4. Consider the sample cross correlation between $\{a_t\}$ and $\{\tilde{y}_t\}$.

By applying prewhitening, the spurious correlation between to time series can be reduced, which helps identify the true level of cross correlation.

5 Conditional Heteroscedastic Models

Remark 5.1 (Necessity for Volatility Models)**.** Wold's decomposition theorem states that any covariance stationary $\{y_t\}$ may be written as

$$y_t = \mu_t + u_t$$

where

$$u_t = \theta(B)\varepsilon_t = \Big(\sum_{i=0}^{\infty} \theta_i B^i\Big)\varepsilon_t, \quad \sum_{i=0}^{\infty} \theta_i^2 < \infty, \quad \theta_0 = 1$$

$$\mathbb{E}(\varepsilon_t) = 0, \quad \mathbb{E}(\varepsilon_t \varepsilon_s) = \begin{cases} \sigma^2 < \infty, & \text{if } t = s \\ 0, & \text{otherwise} \end{cases}$$

The innovation sequence $\{\varepsilon_t\}$ is not necessarily independent.

Now suppose that y_t is a linear stationary process with i.i.d. innovations:

$$y_t = \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}$$

The unconditional mean and unconditional variance are:

$$\mathbb{E}[y_t] = 0, \quad \mathbb{E}[y_t^2] = \sigma^2 \sum_{i=0}^{\infty} \theta_i^2$$

The conditional mean of $\{y_t\}$ is time-varying:

$$\mathbb{E}[y_t|\mathcal{F}_{t-1}] = \sum_{i=1}^{\infty} \theta_i \varepsilon_{t-i}$$

The conditional variance of $\{y_t\}$ is constant:

$$\mathbb{E}[(y_t - \mathbb{E}[y_t|\mathcal{F}_{t-1}])^2|\mathcal{F}_{t-1}] = \sigma^2$$

This model is **unable to capture the conditional variance dynamics**.

Definition 5.1 (Random variance models)**.** Random variance models are generally written as

$$y_t = \mu_t(\theta) + \varepsilon_t$$

$$\varepsilon_t = \sigma_t(\theta) z_t, \quad z_t \sim \text{IID}(0, 1)$$

where $\sigma_t := \sigma_t(\theta) > 0$, and

$$\mu_t(\theta) = \mathbb{E}[y_t|\mathcal{F}_{t-1}] := \mathbb{E}_{t-1}[y_t]$$

$$\sigma_t^2(\theta) = \mathbb{E}[(y_t - \mu_t(\theta))^2|\mathcal{F}_{t-1}] := \mathbb{E}_{t-1}[\varepsilon_t^2]$$

The dynamics of the conditional mean is captured by $\mu_t(\theta)$, which may be an ARMA(p, q) process or could consist of seasonality features. The dynamics of the conditional variance is captured by $\sigma_t(\theta)$, and may also be modeled in an ARMA way.

5.1 ARCH

Definition 5.2 (ARCH(q) Model)**.** ARCH(q) model is originally introduced by Engle (1982). It models the conditional volatility as a linear function (moving average) of past squared disturbances:

$$\varepsilon_t = \sigma_t z_t, \quad z_t \sim \text{IID}(0,1)$$

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2$$

Proposition 5.1 (Common Properties of ARCH(q) Model)**.**

- The model parameters must satisfy: $\omega > 0$, and $\alpha_i \geq 0$, $i = 1, \cdots, q$.
- Defining $v_t = \varepsilon_t^2 - \sigma_t^2$, where $\mathbb{E}_{t-1}[v_t] = 0$, we can construct an AR(q) model for the squared innovation sequence $\{\varepsilon_t^2\}$:
$$\varepsilon_t^2 = \omega + \alpha(B)\varepsilon_t^2 + v_t$$
where $\alpha(B) = \alpha_1 B + \alpha_2 B^2 + \cdots + \alpha_q B^q$.
- **Stationarity condition**: The process $\{\varepsilon_t\}$ is weakly stationary if and only if $\sum_{i=1}^q \alpha_i < 1$.
- The unconditional variance of innovation is
$$\mathbb{E}[\varepsilon_t^2] = \frac{\omega}{1 - \alpha_1 - \cdots - \alpha_q} =: \sigma_\varepsilon^2$$

Proposition 5.2 (Excess Kurtosis of ARCH Model)**.** ARCH models are able to generate **excess kurtosis**. Indeed, even if the standardized innovations z_t is assumed to be normal, the unconditional distribution for ε_t has fatter tails than the normal distribution.

Example 5.1 (Kurtosis of ARCH(1) Model with Normal Innovations)**.** Consider an ARCH(1) model

$$\varepsilon_t = \sigma_t z_t, \quad \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2, \quad z_t \overset{i.i.d}{\sim} \mathcal{N}(0,1)$$

The **unconditional variance** is given by (if $0 < \alpha < 1$)

$$\mathbb{E}[\varepsilon_t^2] = \omega/(1-\alpha) =: \sigma_\varepsilon^2$$

Under normality, the **conditional fourth moment** is given by

$$\begin{aligned}\mathbb{E}[\varepsilon_t^4 | \mathcal{F}_{t-1}] &= \mathbb{E}[\sigma_t^4 z_t^4 | \mathcal{F}_{t-1}] = \mathbb{E}[\sigma_t^4 | \mathcal{F}_{t-1}]\mathbb{E}[z_t^4 | \mathcal{F}_{t-1}] \\ &= 3[\sigma_t^4 | \mathcal{F}_{t-1}] = 3(\omega + \alpha \varepsilon_{t-1}^2)^2\end{aligned}$$

The **unconditional fourth moment** (if exist) is given by

$$\begin{aligned} m_4 = \mathbb{E}[\varepsilon_t^4] &= \mathbb{E}[\mathbb{E}[\sigma_t^4 z_t^4 | \mathcal{F}_{t-1}]] = \mathbb{E}[3(\omega + \alpha \varepsilon_{t-1}^2)^2] \\ &= 3\Big(\omega^2 + 2\omega\alpha \mathbb{E}(\varepsilon_{t-1}^2) + \alpha^2 \mathbb{E}(\varepsilon_{t-1}^4)\Big) = 3\Big(\omega^2 + 2\omega\alpha \frac{\omega}{1-\alpha} + \alpha^2 m_4\Big) \\ \implies \quad m_4 &= \frac{3\omega^2(1+\alpha)}{(1-\alpha)(1-3\alpha^2)}\end{aligned}$$

Since the fourth moment of ε_t is positive, we must have that $\alpha^2 < 1/3$.

The **unconditional kurtosis** is

$$\kappa_\varepsilon = \frac{m_4}{\sigma_\varepsilon^4} = \frac{3\omega^2(1+\alpha)}{(1-\alpha)(1-3\alpha^2)} \times \frac{(1-\alpha)^2}{\omega^2} = 3\frac{1-\alpha^2}{1-3\alpha^2} \geq 3$$

So, the excess kurtosis is always positive and the tails of the distribution of ε_t are fatter than that of a normal distribution, even if the conditional distribution is normal.

Proposition 5.3 (Forecasting Volatility of ARCH Model)**.**

- The **1-step ahead forecast** for σ^2_{t+1} is

$$\sigma^2_{t+1|t} = \hat{\omega} + \hat{\alpha}_1\hat{\varepsilon}^2_t + \cdots + \hat{\alpha}_q\hat{\varepsilon}^2_{t+1-q} = \hat{\omega} + \sum_{i=1}^{q}\hat{\alpha}_i\hat{\varepsilon}^2_{t+1-i}$$

- The **2-step ahead forecast** for σ^2_{t+2} is built on the basis of the 1-step case:

$$\sigma^2_{t+2|t} = \hat{\omega} + \hat{\alpha}_1\sigma^2_{t+1|t} + \hat{\alpha}_2\hat{\varepsilon}^2_t + \cdots + \hat{\alpha}_q\hat{\varepsilon}^2_{t+2-q}$$

- The **h-step ahead forecast** for σ^2_{t+h} is also given recursively as

$$\sigma^2_{t+h|t} = \hat{\omega} + \hat{\alpha}_1\hat{\sigma}^2_{t+h-1|t} + \cdots + \hat{\alpha}_p\hat{\sigma}^2_{t+h-q|t} = \hat{\omega} + \sum_{i=1}^{q}\hat{\alpha}_i\hat{\sigma}^2_{t+h-i|t},$$

  with $\sigma^2_{t+\ell|t} = \hat{\varepsilon}^2_{t+\ell}$ if $\ell \leq 0$.

Remark 5.2 (Model checking)**.** Let $\hat{\varepsilon}_t$ be the residual of the mean equation of an ARMA(p, q) model,

$$\hat{\varepsilon}_t = y_t - \sum_{i=1}^{p}\hat{\phi}_i y_{t-i} - \sum_{j=1}^{q}\hat{\theta}_j\hat{\varepsilon}_{t-j}, \quad \hat{\varepsilon}_t = \hat{\sigma}_t z_t$$

Let $\hat{\sigma}^2_t = \hat{\omega} + \sum_{j=1}^{\tilde{q}}\hat{\alpha}_j\hat{\varepsilon}^2_{t-j}$ be the estimated conditional variance. Define the standardized residual as

$$\hat{e}_t = \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t}$$

- The ACF of $\hat{e}_t$ can be used to check the adequacy of the mean equation of y_t, i.e. the ARMA model.
- the ACF of the squared residuals $\hat{\varepsilon}^2_t$ can be used to check the adequacy of the volatility equation, i.e. the ARCH model.

5.2 GARCH

Due to the large persistence in volatility, the ARCH model often requires a large p to fit the data. In such cases, it is more parsimonious to use the GARCH (Generalized ARCH) model proposed by Bollerslev (1986).

Definition 5.3 (GARCH(p,q) process)**.** The process ε_t is called a (strong) GARCH(p,q) (with respect to the sequence $\{\eta_t\}$) if

$$\begin{cases} \varepsilon_t = \sigma_t\eta_t, \quad \eta_t \sim i.i.d.(0,1), \\ \sigma^2_t = \omega + \sum_{i=1}^{q}\alpha_i\varepsilon^2_{t-i} + \sum_{j=1}^{p}\beta_j\sigma^2_{t-j} \end{cases}$$

where the α_i and β_j are nonnegative constants, and ω is a (strictly) positive constant.

Proposition 5.4 (The implied condition of ARCH(∞) model)**.** Rewriting the GARCH(p, q) model as an ARCH(∞):

$$\begin{aligned}\sigma^2_t = [1-\beta(B)]^{-1}[\omega + \alpha(B)\varepsilon^2_t] &= \omega^* + \psi(B)\varepsilon^2_t \\ &= \omega^* + \sum_{i=1}^{\infty}\psi_i\varepsilon^2_{t-i}\end{aligned}$$

where $\psi(B) = \psi_1 B + \psi_2 B^2 + \ldots = \sum_{i=1}^{\infty}\psi_i B^i$.

- **Nonnegativity**: $\sigma^2_t \geq 0$, if $\omega^* \geq 0$ and all $\psi_i \geq 0$. The non-negativity of ω^* and ψ_i is also a necessary condition for the non negativity of σ^2_t.
- **Invertibility of** $1-\beta(B)$: To make ω^* and $\{\psi_i\}_{i=1}^{\infty}$ well defined, assume that:
  - The roots of the polynomial $\beta(B) = 1$ lie outside the unit circle, and that $\omega \geq 0$. This is a condition for ω^* to be finite and positive.

  - $\alpha(z)$ and $1-\beta(z)$ have no common roots.
- Note that these conditions are establishing neither that $\{\sigma_t^2\}_{t=-\infty}^{\infty}$ is stationary nor that $\sigma_t^2 < \infty$.

Proposition 5.5 (Weak Stationarity of GARCH(p,q)). A process $\{\varepsilon_t\}$ which satisfies a GARCH(p,q) model with positive coefficient $\omega > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, q$, $\beta_j \geq 0$, $j = 1, \ldots, p$ is weakly stationary if and only if:

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$$

Proposition 5.6 (Strict Stationarity of GARCH(1,1)). As proposed by Nelson (1990), GARCH(1,1) process with $\omega > 0$ has a unique strictly stationary solution if and only if

$$\gamma := \mathbb{E}[\ln(\beta + \alpha\eta_t^2)] < 0$$

A sufficient but not necessary strictly stationary condition is $\alpha + \beta < 1$ by the following result (using the Jensen's Inequality)

$$\mathbb{E}[\ln(\beta + \alpha\eta_t^2)] \leq \ln[\mathbb{E}[\beta + \alpha\eta_t^2]] = \ln(\alpha + \beta)$$

So $\alpha + \beta < 1$ can lead to $\gamma < 0$. But note that when $\alpha + \beta = 1$, the model is still strictly stationary.

When $\gamma \geq 0$, the behavior of GARCH(1,1) process is:

- $\gamma > 0$: $\sigma_t^2/\rho^t \to \infty$ a.s. as $t \to \infty$ for some $\rho > 1$.
- $\gamma = 0$: $\sigma_t^2 \to \infty$ in probability as $t \to \infty$.

Remark 5.3 (Drawbacks of GARCH).

- Volatility tends to rise in response to "bad news" and to fall in response to "good news" (leverage effect);
- The volatility in the GARCH process is **symmetric**, determined only by the magnitude of the previous return and shock, not by its sign.
- The parameters in GARCH are restricted to be positive to ensure nonnegativity of σ_t^2. When estimating, however, sometimes best fits are achieved for negative parameters.

6 Multivariate Time Series Analysis

6.1 Basic Concepts

Definition 6.1 (Weak Stationarity).

- Denote $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^T \in \mathbb{R}^m$. The series $\{\mathbf{y}_t : t \in \mathcal{T}\}$ is weakly stationary if its first and second moments are time-invariant, i.e., $\mathbb{E}[\mathbf{y}_t]$ is constant and $\mathbb{E}[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-k} - \boldsymbol{\mu})^T]$ only depends on k.
- For a weakly stationary time series $\{\mathbf{y}_t\}$, define its **mean** as

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{y}_t) = \begin{bmatrix} \mathbb{E}(y_{1t}) \\ \mathbb{E}(y_{2t}) \\ \vdots \\ \mathbb{E}(y_{mt}) \end{bmatrix} := \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$$

  and the **covariance matrix** as

$$\boldsymbol{\Gamma}_0 = \mathbb{E}[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'] = \begin{bmatrix} \Gamma_{11}(0) & \Gamma_{12}(0) & \cdots & \Gamma_{1m}(0) \\ \Gamma_{21}(0) & \Gamma_{22}(0) & \cdots & \Gamma_{m2}(0) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{m1}(0) & \Gamma_{m2}(0) & \cdots & \Gamma_{mm}(0) \end{bmatrix}$$

  where

$$\Gamma_{ii}(0) = \mathbb{E}[(y_{it} - \mu_i)^2]$$

  and

$$\Gamma_{ij}(0) = \mathbb{E}[(y_{it} - \mu_i)(y_{jt} - \mu_j)], \quad i, j = 1, \ldots, n$$

Definition 6.2 (Cross-Correlation Matrices)**.** Let $\mathbf{D} = \text{diag}\{\sqrt{\Gamma_{11}(0)}, \ldots, \sqrt{\Gamma_{mm}(0)}\}$.

- The concurrent, or lag-zero, **cross-correlation matrix** is defined as

$$\boldsymbol{\rho}_0 \equiv [\rho_{ij}(0)] = \mathbf{D}^{-1}\boldsymbol{\Gamma}_0\mathbf{D}^{-1} = \begin{bmatrix} \rho_{11}(0) & \rho_{12}(0) & \cdots & \rho_{1m}(0) \\ \rho_{21}(0) & \rho_{22}(0) & \cdots & \rho_{2m}(0) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}(0) & \rho_{m2}(0) & \cdots & \rho_{mm}(0) \end{bmatrix}$$

  where the correlation coefficient between y_{it} and y_{jt} is

$$\rho_{ij}(0) = \text{corr}(y_{it}, y_{jt}) = \frac{\Gamma_{ij}(0)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(y_{it}, y_{jt})}{\text{std}(y_{it}) \cdot \text{std}(y_{jt})}$$

- The elements of $\boldsymbol{\rho}_0$ satisfies

$$\rho_{ij}(0) = \text{corr}(y_{it}, y_{jt}) = \text{corr}(y_{jt}, y_{it}) = \rho_{ji}(0), \quad -1 \leq \rho_{ij}(0) \leq 1, \quad \rho_{ii}(0) = 1, \quad 1 \leq i, j \leq m$$

  Thus, $\boldsymbol{\rho}_0$ is a symmetric matrix with unit diagonal elements,

$$\boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho_{21}(0) & \cdots & \rho_{m1}(0) \\ \rho_{21}(0) & 1 & \cdots & \rho_{m2}(0) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}(0) & \rho_{m2}(0) & \cdots & 1 \end{bmatrix}$$

Definition 6.3 (Lead-lag relationships)**.**

- The **lagged cross-correlation matrices** are used to measure the strength of linear dependence between time series. Under weak stationarity, the **lag-ℓ cross-covariance matrix** is defined as

$$\begin{aligned} \boldsymbol{\Gamma}_\ell \equiv [\Gamma_{ij}(\ell)] &= \mathbb{E}[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-\ell} - \boldsymbol{\mu})^T] \\ &= \begin{bmatrix} \text{Cov}(y_{1t}, y_{1,t-\ell}) & \text{Cov}(y_{1t}, y_{2,t-\ell}) & \cdots & \text{Cov}(y_{1t}, y_{m,t-\ell}) \\ \text{Cov}(y_{2t}, y_{1,t-\ell}) & \text{Cov}(y_{2t}, y_{2,t-\ell}) & \cdots & \text{Cov}(y_{2t}, y_{m,t-\ell}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(y_{mt}, y_{1,t-\ell}) & \text{Cov}(y_{mt}, y_{2,t-\ell}) & \cdots & \text{Cov}(y_{mt}, y_{m,t-\ell}) \end{bmatrix} \\ &= \begin{bmatrix} \Gamma_{11}(\ell) & \Gamma_{12}(\ell) & \cdots & \Gamma_{1m}(\ell) \\ \Gamma_{21}(\ell) & \Gamma_{22}(\ell) & \cdots & \Gamma_{m2}(\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{m1}(\ell) & \Gamma_{m2}(\ell) & \cdots & \Gamma_{mm}(\ell) \end{bmatrix} \end{aligned}$$

- The **lag-ℓ cross-correlation matrix** (CCM) is defined as

$$\boldsymbol{\rho}_\ell \equiv [\rho_{ij}(\ell)] = \mathbf{D}^{-1}\boldsymbol{\Gamma}_\ell\mathbf{D}^{-1} = \begin{bmatrix} \rho_{11}(\ell) & \rho_{12}(\ell) & \cdots & \rho_{1m}(\ell) \\ \rho_{21}(\ell) & \rho_{22}(\ell) & \cdots & \rho_{2m}(\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}(\ell) & \rho_{m2}(\ell) & \cdots & \rho_{mm}(\ell) \end{bmatrix}$$

  where

$$\rho_{ij}(\ell) = \text{corr}(y_{it}, y_{j,t-\ell}) = \frac{\Gamma_{ij}(\ell)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(y_{it}, y_{j,t-\ell})}{\text{std}(y_{it}) \cdot \text{std}(y_{jt})}$$

  - The diagonal element $\rho_{ii}(\ell)$ is simply the lag-ℓ autocorrelation coefficient of y_{it}. In general, $\rho_{ii}(\ell) \neq 1$.
  - When $\ell > 0$, the correlation coefficient $\rho_{ij}(\ell)$ measures the linear dependence of y_{it} on $y_{j,t-\ell}$. If $\rho_{ij}(\ell) \neq 0$ and $\ell > 0$, we say that the series y_{jt} leads the series y_{it} at lag ℓ.

– Similarly, $\rho_{ji}(\ell)$ measures the linear dependence of y_{jt} and $y_{i,t-\ell}$, and we say that the series y_{it} leads the series y_{jt} at lag ℓ if $\rho_{ji}(\ell) \neq 0$ and $\ell > 0$.

Definition 6.4 (Linear Dependence)**.** Considered jointly, the cross-correlation matrices $\{\boldsymbol{\rho}_\ell | \ell = 0, 1, \ldots\}$ of a weakly stationary vector time series contain the following information:

- The diagonal elements $\rho_{ii}(\ell)$, $\ell = 0, 1, \ldots$ are the autocorrelation function (ACF) of y_{it}.
- The off-diagonal element $\rho_{ij}(0)$ measures the concurrent linear relationship between y_{it} and y_{jt}.
- For $\ell > 0$, the off-diagonal element $\rho_{ij}(\ell)$ measures the linear dependence of y_{it} on the past value $y_{j,t-\ell}$.

In general, the linear relationship between two time series $\{y_{it}\}$ and $\{y_{jt}\}$ can be summarized as follows:

1. y_{it} and y_{jt} have **no linear relationship** if $\rho_{ij}(\ell) = \rho_{ji}(\ell) = 0$ for all $\ell \geq 0$.
2. y_{it} and y_{jt} are **concurrently correlated** if $\rho_{ij}(0) \neq 0$.
3. y_{it} and y_{jt} have **no lead-lag relationship** if $\rho_{ij}(\ell) = 0$ and $\rho_{ji}(\ell) = 0$ for all $\ell > 0$.
4. There is a **unidirectional relationship** from y_{it} to y_{jt} if $\rho_{ij}(\ell) = 0$ for all $\ell > 0$, but $\rho_{ji}(v) \neq 0$ for some $v > 0$. In this case, y_{it} does not depend on any past value of y_{jt}, but y_{jt} depends on some past values of y_{it}.
5. There is a **feedback relationship** between y_{it} and y_{jt} if $\rho_{ij}(\ell) \neq 0$ for some $\ell > 0$ and $\rho_{ji}(v) \neq 0$ for some $v > 0$.

Definition 6.5 (Sample Cross-Correlation Matrices)**.** Given the data $\{\mathbf{y}_t\}_{t=1}^T$, the **cross-covariance matrix** Γ_ℓ can be estimated by

$$\widehat{\Gamma}_\ell = \frac{1}{T} \sum_{t=\ell+1}^{T} (\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_{t-\ell} - \bar{\mathbf{y}})', \quad \ell \geq 0$$

where $\bar{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^T \mathbf{y}_t$ is the vector of sample means.

The **cross-correlation matrix** ρ_ℓ is estimated by

$$\widehat{\rho}_\ell = \widehat{\mathbf{D}}^{-1}\widehat{\Gamma}_\ell\widehat{\mathbf{D}}^{-1}, \quad \ell \geq 0$$

where $\widehat{\mathbf{D}} = \text{diag}\{\widehat{\Gamma}_{11}^{1/2}(0), \ldots, \widehat{\Gamma}_{mm}^{1/2}(0)\}$.

Definition 6.6 (CCF test between two series)**.** Consider the linear dependence of $\{y_t\}$ on $\{x_t\}$ and test for $\rho_{xy}(\ell) = 0$, $\ell > 0$. Under the hypothesis that there is no relation at time ℓ and that at least one of the two series are independent and identically distributed, we have the result:

$$\sqrt{n}\hat{\rho}_{xy}(\ell) \overset{A}{\sim} \mathcal{N}(0, 1)$$

Definition 6.7 (Multivariate Wold Representation)**.** Any m-dimensional weakly stationary multivariate time series $\{\mathbf{y}_t\}$ has a Wold or linear process representation of the form

$$\begin{aligned}\mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots \\ &= \boldsymbol{\mu} + \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k\boldsymbol{\varepsilon}_{t-k}\end{aligned}$$

where $\boldsymbol{\Psi}_0 = \mathbf{I}_m$, $\boldsymbol{\varepsilon}_t \sim \text{WN}(0, \boldsymbol{\Sigma})$, and $\boldsymbol{\Psi}_k$ is an $m \times m$ matrix with (i, j)-th element $\psi_{ij}^{(k)}$:

$$\boldsymbol{\Psi}_k = [\psi_{ij}^{(k)}] = \begin{bmatrix} \psi_{11}^{(k)} & \cdots & \psi_{1m}^{(k)} \\ \vdots & \ddots & \vdots \\ \psi_{m1}^{(k)} & \cdots & \psi_{mm}^{(k)} \end{bmatrix}$$

In lag operator notation, the Wold form is

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\Psi}(B) = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k B^k$$

The moments of $\mathbf{y}_t$ are given by

$$\mathbb{E}(\mathbf{y}_t) = \boldsymbol{\mu}, \quad \text{Var}(\mathbf{y}_t) = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k\boldsymbol{\Sigma}\boldsymbol{\Psi}_k'$$

6.2 Vector Autoregression Models

6.2.1 Vector AR(1) Models

Definition 6.8 (Vector AR(1) Model)**.** A multivariate time series $\mathbf{y}_t$ is a VAR(1) process, if it follows the model

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

$$\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0, \text{ and } \mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s) = \begin{cases} \boldsymbol{\Sigma}, & \text{if } t = s; \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\phi}_0$ is a m-dimensional vector, $\boldsymbol{\Phi}$ is an $m \times m$ matrix. In empirical applications, it is often assumed that $\boldsymbol{\varepsilon}_t$ is multivariate normal, i.e., $\boldsymbol{\varepsilon}_t \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Proposition 6.1 (Stationarity Condition of VAR(1) Model)**.** Consider the VAR(1) model

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Stationarity: Similar to the scalar AR process, for the VAR(1) process to be weakly stationary, we must have that the roots in the **characteristic equation**

$$\det(\mathbf{I} - \boldsymbol{\Phi}z) = \|\mathbf{I} - \boldsymbol{\Phi}z\| = 0$$

all **lie outside the unit circle**. In other ways, it is equivalent to say that the roots in the polynomial equation

$$\det(\lambda\mathbf{I} - \boldsymbol{\Phi}) = |\lambda\mathbf{I} - \boldsymbol{\Phi}| = 0$$

all lie **inside** the unit circle, where $\lambda = z^{-1}$.

Alternatively, consider the spectral radius of $\boldsymbol{\Phi}$, defined as $\rho(\boldsymbol{\Phi}) \equiv \max\{|\lambda_1|, \ldots, |\lambda_n|\}$, where $\lambda_1, \ldots, \lambda_n$ are the (real or complex) eigenvalues of the matrix $\boldsymbol{\Phi}$. The stationary condition is further equivalent to

$$\rho(\boldsymbol{\Phi}) < 1$$

This condition provides an easy tool for checking the stability of a VAR process.

Proposition 6.2 (Moments of VAR(1) Model)**.** Consider a weakly stationary VAR(1) model:

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Taking expectation of the model and using $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$, we obtain

$$\begin{aligned} &\mathbb{E}(\mathbf{y}_t) = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbb{E}(\mathbf{y}_{t-1}) \\ &\Longrightarrow \boldsymbol{\mu} \equiv \mathbb{E}[\mathbf{y}_t] = (\mathbf{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{\phi}_0 \end{aligned}$$

Let $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \boldsymbol{\mu}$ be the mean-corrected time series, i.e., $\tilde{\mathbf{y}}_t = \boldsymbol{\Phi}\tilde{\mathbf{y}}_{t-1} + \boldsymbol{\varepsilon}_t$.]

The MA(∞) form is

$$\tilde{\mathbf{y}}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{\Phi}\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Phi}^2\boldsymbol{\varepsilon}_{t-2} + \boldsymbol{\Phi}^3\boldsymbol{\varepsilon}_{t-3} + \cdots$$

This expression shows several moment characteristics of a VAR(1) process:

1. $\mathrm{Cov}(\mathbf{y}_t, \boldsymbol{\varepsilon}_t) = \boldsymbol{\Sigma}$;
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}_t, \mathbf{y}_{t-j}) = 0, \ j = 1, 2, 3, \ldots$;
3. $\mathrm{Corr}(\mathbf{y}_t, \boldsymbol{\varepsilon}_{t-j}) = \boldsymbol{\Phi}^j, \ j = 1, 2, 3, \ldots$;
4. $\mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t-\ell}) = \boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}\boldsymbol{\Gamma}_{\ell-1} = \boldsymbol{\Phi}^\ell\boldsymbol{\Gamma}_0$ for $\ell > 0$.
5. $\mathrm{Corr}(\mathbf{y}_t, \mathbf{y}_{t-\ell}) = \boldsymbol{\rho}_\ell = \boldsymbol{\Upsilon}\boldsymbol{\rho}_{\ell-1} = \boldsymbol{\Upsilon}^\ell\boldsymbol{\rho}_0$ for $\ell > 0$, where $\boldsymbol{\Upsilon} = \mathbf{D}^{-1}\boldsymbol{\Phi}\mathbf{D}$, $\mathbf{D} = \mathrm{diag}\{\Gamma_{11}^{1/2}(0), \ldots, \Gamma_{mm}^{1/2}(0)\}$.

6.

$$\text{Cov}(\mathbf{y}_t) = \mathbf{\Gamma}_0 = \mathbf{\Sigma} + \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}' + \mathbf{\Phi}^2\mathbf{\Sigma}(\mathbf{\Phi}^2)' + \cdots = \sum_{i=0}^{\infty} \mathbf{\Phi}^i \mathbf{\Sigma}(\mathbf{\Phi}^i)'$$

or in closed form,

$$\text{Cov}(\mathbf{y}_t) = \mathbf{\Phi}\text{Cov}(\mathbf{y}_{t-1})\mathbf{\Phi}' + \mathbf{\Sigma}$$
$$\text{vec}(\mathbf{\Gamma}_0) = (\mathbf{\Phi} \otimes \mathbf{\Phi}) \cdot \text{vec}(\mathbf{\Gamma}_0) + \text{vec}(\mathbf{\Sigma})$$
$$\text{vec}(\mathbf{\Gamma}_0) = [\mathbf{I} - (\mathbf{\Phi} \otimes \mathbf{\Phi})]^{-1}\text{vec}(\mathbf{\Sigma})$$

Remark 6.1 (Vector Operations)**.** Let $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ be matrices whose dimensions are such that the product $\mathbf{ABC}$ exists. Then

$$\text{vec}(\mathbf{A} + \mathbf{B}) = \text{vec}(\mathbf{A}) + \text{vec}(\mathbf{B}), \quad \text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\text{vec}(\mathbf{B})$$

where the symbol $\otimes$ denotes the **Kronecker product**. Here for $\text{vec}(\cdot)$ and $\otimes$, the rule is given as

$$\text{vec}\left(\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \\ a_5 & a_6 \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ a_3 \\ a_5 \\ a_2 \\ a_4 \\ a_6 \end{bmatrix}, \quad \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nn}\mathbf{B} \end{bmatrix}_{(mp \times nq)}$$

6.2.2 Vector AR(p) Models

Definition 6.9 (Vector AR(p) Models)**.** The time series $\{\mathbf{y}_t\}$ follows a VAR(p) model if it satisfies

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad p > 0$$

$$\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0, \quad \mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_\tau) = \begin{cases} \mathbf{\Sigma}, & \text{if } t = \tau; \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\phi}_0$ is a m-dimensional vector, and $\mathbf{\Phi}_j$ are $n \times n$ matrices.

Using the back-shift operator B, the VAR(p) model can be written as

$$(\mathbf{I} - \mathbf{\Phi}_1 B - \cdots - \mathbf{\Phi}_p B^p)\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\varepsilon}_t$$
$$\implies \mathbf{\Phi}(B)\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\varepsilon}_t$$

where $\mathbf{\Phi}(B) = \mathbf{I} - \mathbf{\Phi}_1 B - \cdots - \mathbf{\Phi}_p B^p$ is a matrix polynomial.

Proposition 6.3 (Moment properties of VAR(p) Model)**.** If $\mathbf{y}_t$ is **weakly stationary**, then we have

$$\boldsymbol{\mu} = E(\mathbf{y}_t) = (\mathbf{I} - \mathbf{\Phi}_1 - \cdots - \mathbf{\Phi}_p)^{-1}\boldsymbol{\phi}_0 = [\mathbf{\Phi}(1)]^{-1}\boldsymbol{\phi}_0$$

provided that the inverse exists.

Let $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \boldsymbol{\mu}$. The VAR($p$) model becomes

$$\tilde{\mathbf{y}}_t = \mathbf{\Phi}_1\tilde{\mathbf{y}}_{t-1} + \cdots + \mathbf{\Phi}_p\tilde{\mathbf{y}}_{t-p} + \boldsymbol{\varepsilon}_t$$

Under weak stationarity of $\{\mathbf{y}_t\}$, we have the following Wold representation

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - \boldsymbol{\mu}_t = \boldsymbol{\varepsilon}_t + \mathbf{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \mathbf{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots$$

Using this equation and the same techniques as those for VAR(1) models, we obtain that:

- $\text{Cov}(\mathbf{y}_t, \boldsymbol{\varepsilon}_t) = \mathbf{\Sigma}$, the covariance matrix of $\boldsymbol{\varepsilon}_t$;
- $\text{Cov}(\mathbf{y}_{t-\ell}, \boldsymbol{\varepsilon}_t) = \mathbf{0}$ for $\ell > 0$;

- $\boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{\ell-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_{\ell-p}$ for $\ell > 0$, since

$$\mathbb{E}(\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}'_{t-\ell}) = \boldsymbol{\Phi}_1 E(\tilde{\mathbf{y}}_{t-1}\tilde{\mathbf{y}}'_{t-\ell}) + \cdots + \boldsymbol{\Phi}_p E(\tilde{\mathbf{y}}_{t-p}\tilde{\mathbf{y}}'_{t-\ell}), \quad \ell > 0$$

The last property is called the **moment equations** of a VAR(p) model. It is a multivariate version of the **Yule-Walker equation** of AR(p) model.

In terms of cross correlation matrix (CCM), the moment equations become

$$\begin{aligned}\boldsymbol{\rho}_\ell &= \mathbf{D}^{-1}\boldsymbol{\Gamma}_\ell\mathbf{D}^{-1} \\ &= \mathbf{D}^{-1}\boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{\ell-1}\mathbf{D}^{-1} + \cdots + \mathbf{D}^{-1}\boldsymbol{\Phi}_p\boldsymbol{\Gamma}_{\ell-p}\mathbf{D}^{-1} \\ &= \boldsymbol{\Upsilon}_1\boldsymbol{\rho}_{\ell-1} + \cdots + \boldsymbol{\Upsilon}_p\boldsymbol{\rho}_{\ell-p}\end{aligned}$$

for $\ell > 0$, where $\boldsymbol{\Upsilon}_i = \mathbf{D}^{-1}\boldsymbol{\Phi}_i\mathbf{D}$, $\mathbf{D} = \text{diag}\{\Gamma_{11}^{1/2}(0), \ldots, \Gamma_{mm}^{1/2}(0)\}$.

Definition 6.10 (State-space representation)**.** Sometimes, it is more convenient to write a scalar valued time series, e.g., AR(p) process, in vector form.

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

So we have rewrite an scalar-valued AR(p) process as an VAR(1) process.

Similarly, we could also transform a VAR(p) process to a **"companion form"** mp-dimensional VAR(1) process:

$$\boldsymbol{\xi}_t = \boldsymbol{\Phi}^*\boldsymbol{\xi}_{t-1} + \mathbf{u}_t$$

where

$$\boldsymbol{\xi}_t = \begin{bmatrix} \tilde{\mathbf{y}}_t \\ \tilde{\mathbf{y}}_{t-1} \\ \tilde{\mathbf{y}}_{t-2} \\ \vdots \\ \tilde{\mathbf{y}}_{t-p+1} \end{bmatrix}, \quad \underset{mp\times mp}{\boldsymbol{\Phi}^*} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p \\ \mathbf{I}_m & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_m & \mathbf{0} \end{bmatrix}, \quad \mathbf{u}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

Furthermore, $\mathbf{u}_t \sim \mathcal{N}_{mp}(\mathbf{0}, \boldsymbol{\Omega})$, where

$$\underset{mp\times mp}{\boldsymbol{\Omega}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

Proposition 6.4 (Stationarity Condition of VAR(p) Model)**.** Using the stationarity condition for VAR(1) process and the state-space representation, we can derive the stationarity condition for VAR(p) process: A VAR(p) process is covariance-stationary, if all the values of $\mathbf{z}$ satisfying

$$\det(\mathbf{I}_{mp\times mp} - \boldsymbol{\Phi}^*_{mp\times mp}\mathbf{z}) = \|\mathbf{I}_m - \boldsymbol{\Phi}_1\mathbf{z} - \boldsymbol{\Phi}_2\mathbf{z}^2 - \cdots - \boldsymbol{\Phi}_p\mathbf{z}^p\| = 0$$

lie outside the unit circle, or equivalently, if the eigenvalues of $\boldsymbol{\Phi}^*_{mp\times mp}$ satisfy

$$\det(\lambda\mathbf{I}_{mp\times mp} - \boldsymbol{\Phi}^*_{mp\times mp}) = \|\mathbf{I}_m\lambda^p - \boldsymbol{\Phi}_1\lambda^{p-1} - \boldsymbol{\Phi}_2\lambda^{p-2} - \cdots - \boldsymbol{\Phi}_p\| = 0$$

Hence, the stationary condition is also required such that $\rho(\boldsymbol{\Phi}^*) < 1$.

Definition 6.11 (Vector moving average representation)**.** Recall that we could invert a scalar stationary AR(p) process, $\phi(B)y_t = \varepsilon_t$, to a MA($\infty$) process, $y_t = \psi(B)\varepsilon_t$, where $\psi(B) = \phi(B)^{-1}$.

Invertibility: For a covariance-stationary VAR(p) process, $\Phi(B)\mathbf{y}_t = \boldsymbol{\varepsilon}_t$ (for simplicity set $\mathbf{c} = \mathbf{0}$), we could invert it to

$$\mathbf{y}_t = \Psi(B)\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots$$

where

$$\Psi(B) = \Phi(B)^{-1}$$

The coefficients of $\boldsymbol{\Psi}$ can be solved in the same way as in the scalar case, i.e., if $\Phi^{-1}(B) = \Psi(B)$, then $\Phi(B)\Psi(B) = \mathbf{I}_m$:

$$(\mathbf{I}_m - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B - \cdots - \boldsymbol{\Phi}_p B^p)(\mathbf{I}_m + \boldsymbol{\Psi}_1 B + \boldsymbol{\Psi}_2 B^2 + \cdots) = \mathbf{I}_m$$

Equating the coefficients of B^j, we have

$$\begin{aligned}
\boldsymbol{\Psi}_0 &= \mathbf{I}_m \\
\boldsymbol{\Psi}_1 &= \boldsymbol{\Phi}_1 \\
\boldsymbol{\Psi}_2 &= \boldsymbol{\Phi}_1\boldsymbol{\Psi}_1 + \boldsymbol{\Phi}_2 \\
&\vdots
\end{aligned}$$

In general, $\boldsymbol{\Psi}_s = \boldsymbol{\Phi_1}\boldsymbol{\Psi}_{s-1} + \boldsymbol{\Phi}_2\boldsymbol{\Psi}_{s-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{s-p}$, with $\boldsymbol{\Psi}_j = 0$ for $j < 0$.

6.3 Estimation and Model Specification

Remark 6.2 (Lag Length Selection)**.** Model selection criteria for VAR(p) models have the form

$$\text{IC}(p) = \ln|\widehat{\boldsymbol{\Sigma}}(p)| + c_T \cdot \psi(m, p)$$

where $\widehat{\boldsymbol{\Sigma}}(p) = T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'$ is the residual covariance matrix without a degrees of freedom correction from a VAR(p) model, c_T is a sequence indexed by the sample size T, and $\psi(m, p)$ is a penalty function which penalizes large VAR(p) models.

The three most common information criteria are the Akaike (AIC), Schwarz-Bayesian (BIC) and Hannan-Quinn (HQ):

$$\text{AIC}(p) = \ln|\widehat{\boldsymbol{\Sigma}}(p)| + \frac{2}{T}pm^2$$

$$\text{BIC}(p) = \ln|\widehat{\boldsymbol{\Sigma}}(p)| + \frac{\ln(T)}{T}pm^2$$

$$\text{HQ}(p) = \ln|\widehat{\boldsymbol{\Sigma}}(p)| + \frac{2\ln\ln(T)}{T}pm^2$$

Proposition 6.5 (VAR Forecasting)**.** The best linear predictor of $\mathbf{y}_{t+1}$(in terms of minimum mean squared error), or 1-step forecast based on information available at time t is

$$\hat{\mathbf{y}}_{t+1|t} = \mathbf{c} + \boldsymbol{\Phi}_1\mathbf{y}_t + \cdots + \boldsymbol{\Phi}_p\mathbf{y}_{t-p+1}$$

Forecasts for longer horizons h (h-step forecasts) may be obtained using the chain-rule of forecasting as

$$\hat{\mathbf{y}}_{t+h|t} = \mathbf{c} + \boldsymbol{\Phi}_1\hat{\mathbf{y}}_{t+h-1|t} + \cdots + \boldsymbol{\Phi}_p\hat{\mathbf{y}}_{t+h-p|t}$$

where $\hat{\mathbf{y}}_{t+j|t} = \mathbf{y}_{t+j}$ if $j < 0$. The h-step forecast errors may be expressed as

$$\hat{\boldsymbol{\varepsilon}}_{t+h|t} = \mathbf{y}_{t+h} - \hat{\mathbf{y}}_{t+h|t} = \sum_{s=0}^{h-1}\boldsymbol{\Psi}_s\boldsymbol{\varepsilon}_{t+h-s}$$

where the matrices $\mathbf{\Psi}_s$ are determined by recursive substitution

$$\mathbf{\Psi}_s = \mathbf{\Phi}_1\mathbf{\Psi}_{s-1} + \mathbf{\Phi}_2\mathbf{\Psi}_{s-2} + \cdots + \mathbf{\Phi}_p\mathbf{\Psi}_{s-p}$$

with $\mathbf{\Psi}_0 = \mathbf{I}_n$ and $\mathbf{\Psi}_j = \mathbf{0}$ for $j < 0$.

The forecasts are unbiased since all of the forecast errors have expectation zero and the MSE matrix for $\hat{\mathbf{y}}_{t+h|t}$ is

$$\mathbf{\Sigma}(h) = \text{MSE}(\hat{\boldsymbol{\varepsilon}}_{t+h|t}) = \sum_{s=0}^{h-1} \mathbf{\Psi}_s\mathbf{\Sigma}\mathbf{\Psi}_s'$$

Asymptotic $100(1-\alpha)\%$ confidence intervals for the individual elements of $\hat{\mathbf{y}}_{t+h|t}$ are then computed as

$$[\hat{y}_{k,t+h|t} - z_{\alpha/2}\hat{\sigma}_k(h), \quad \hat{y}_{k,t+h|t} + z_{\alpha/2}\hat{\sigma}_k(h)]$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution and $\hat{\sigma}_k(h)$ denotes the square root of the diagonal element of $\hat{\mathbf{\Sigma}}(h)$.

7 Appendix

Remark 7.1 (Comparison of Regression Analysis and Time Series Analysis)**.**

- Regression is a static concept, describing a still world, i.e., a dead world. Time series is a dynamic concept, describing a living world, i.e., a live world.
  - Without time, there is no history.
  - Without time, there is no dynamics.

  Bringing time into the picture is a revolution.
- Time series analysis is rooted in physical sciences, especially physics in the Newtonian sense. Time derivatives are at the center of its attention. Regression analysis is probably rooted in social sciences, which until fairly recently paid scant attention to phenomena that change over time.
- Time series analysis is a powerful instrument dealing with random processes. Regression analysis is a powerful instrument dealing with static relations between random variables, not processes.

Remark 7.2 (Discussion on White Noise)**.** In signal processing, white noise is a random signal having equal intensity at different frequencies, giving it a constant power spectral density. White noise refers to a statistical model for signals and signal sources, rather than to any specific signal.

Remark 7.3 (Discussion on AIC Criterions for Model Selection)**.** The AIC has been criticized because it does not yield a consistent estimator with respect to the selection of orders. Such an argument is frequently misunderstood, and we attempt to clarify these misunderstandings in the following.

1. First, the objective of our modeling is to obtain a "good" model, rather than a "true" model. If one recalls that statistical models are approximations of complex systems toward certain objectives, the task of estimating the true order is obviously not an appropriate goal. A true model or order can be defined explicitly only in a limited number of situations, such as when running simulation experiments. From the standpoint that a model is an approximation of a complex phenomenon, the true order can be infinitely large.
2. Even if a true finite order exists, the order of a good model is not necessarily equal to the true order. In situations where there are only a small number of observations, considering the instability of the parameters being estimated, the AIC reveals the possibility that a higher prediction accuracy can be obtained using models having lower orders.
3. Shibata's (1976) results described in the previous section indicate that if the true order is assumed, the asymptotic distribution of orders selected by the AIC can be a fixed distribution that is determined solely by the maximum order and the true order of a family of models. This indicates that the AIC does not

provide a consistent estimator of orders. It should be noted, however, that when the true order is finite, the distribution of orders that is selected does not vary when the number of observations is increased. It should also be noted that in this case, even if a higher order is selected, when the number of observations is large, each coefficient estimate of a regressor with an order greater than the true order converges to the true value 0 and that a consistent estimator can be obtained as a model.

4. Although the information criterion makes automatic model selection possible, it should be noted that the model evaluation criterion is a relative evaluation criterion. This means that selecting a model using an information criterion is only a selection from a family of models that we have specified. Therefore, the critical task for us is to set up more appropriate models by making use of knowledge regarding that object.