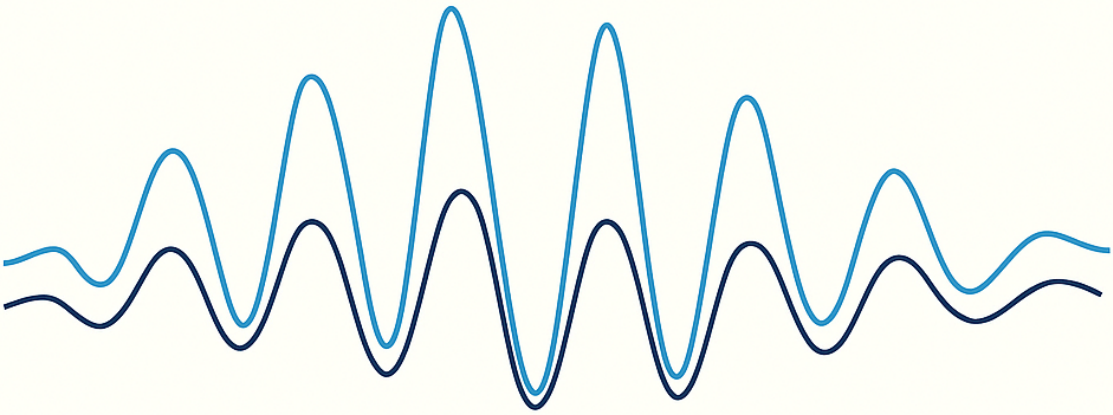
Introduction to Econometric Time Series



Domenico Sartore

Department of Economics –
Ca' Foscari University of Venice

Introduction to Econometric Time Series

Domenico Sartore

Department of Economics — Ca' Foscari University of Venice

Published and archived by Zenodo, Geneva

ISBN 979-12-243-1114-0

Version 2.0 — February 2026
This work is archived at the permanent Concept DOI:
https://doi.org/10.5281/zenodo.17572969

For comments, corrections, or suggestions, please contact the author at sartore@unive.it.

Contents

Preface

The purpose of this book is to provide a rigorous yet accessible introduction to econometric time series analysis. It originates from lecture notes prepared for courses in Econometrics taught at the International Master in Economics and Finance, held at Ca' Foscari University of Venice. Over time, these notes have been revised and expanded with the aim of offering a systematic treatment of both theoretical foundations and applied aspects, guiding the reader from the specification of the linear regression model to the analysis of nonstationary and cointegrated time series.

Alongside the exposition of the fundamental results, special attention has been devoted to those parts of the econometric theory of time series that help to motivate and clarify various aspects of econometric practice. This approach aims to reduce the distance between mathematical formalism and empirical applications, providing a logical framework that makes the connection between theoretical assumptions, estimation procedures, and the interpretation of results more transparent.

Forecasting patterns are further analyzed in their dynamics in the case of nonstationary and cointegrated series, interpreted through models incorporating an error-correction mechanism, discussed in *Chapter 5*. Particular attention has also been given to the attempt to connect the multivariate case to the univariate one, highlighting both the conceptual analogies and the operational differences. This makes it possible to better understand the transition from simple to more complex model structures, without losing sight of the basic principles that guide inference and forecasting in time series analysis.

For this reason, the text is not intended to be exhaustive of all the topics in time series analysis, and readers who wish to explore further may refer to the bibliography provided at the end of the book.

In line with its didactic purpose, intermediate steps in the derivations have been preserved, even when these are sometimes omitted in more advanced texts. This choice is meant to support students in following the logical development of the results.

The presentation is structured to guide the reader through the essential stages of econometric reasoning.Throughout the book, emphasis is placed on conceptual coherence. Connections between algebraic structure, stochastic assumptions, and inferential procedures are made explicit, so that theoretical results can be interpreted within a unified econometric framework.

The first chapters introduce the basic probabilistic tools and the main stationary processes, followed by the general representation theorems and their implications. In particular, *Chapter 2* discusses the concept of ergodicity, considered fundamental for justifying inference in time series, while *Chapter 3* gives ample space to Wold's decomposition, which provides a comprehensive view of stationary processes. Subsequent chapters deal with multivariate processes, dynamic properties, and forecasting methods. The last part of the book focuses on nonstationary processes, unit root tests, cointegration, and the

Kalman filter, highlighting their role in modern econometric practice.
The approach adopted is deliberately didactic: formal definitions and propositions are always accompanied by examples, intuitive explanations, and, in some cases, numerical simulations. The aim is not only to present results in their formal rigor, but also to clarify their interpretation in an econometric context. Graphical illustrations and empirical applications are included to support intuition and facilitate learning.
Although the text is primarily intended for students in economics and statistics, it may also be useful to researchers and practitioners who wish to consolidate their understanding of time series methods. The book does not assume advanced mathematical knowledge beyond basic probability and linear algebra, but it gradually introduces the tools required for a deeper study of econometrics.
The scope of this book is therefore introductory by design. Topics such as volatility modeling with ARCH and GARCH processes, though crucial in financial econometrics, fall outside its boundaries. Their omission reinforces the intended focus: to provide a clear and structured introduction to the probabilistic and econometric foundations of time series analysis.

1 Introduction to Stochastic Processes

Definition 1.1. *A stochastic process* $\{X_t : t \in \mathcal{T}\}$ *is an ordered collection of random variables indexed by a parameter* t *(often time) and defined on a common sample space.*

If $\mathcal{T}$ is an interval of $\mathbb{R}$, the process is a *continuous-time* process, denoted $\{X_t : t \in \mathcal{T}\}$. If $\mathcal{T}$ is countable (e.g., $\mathbb{Z}$ or $\mathbb{N}$), it is a *discrete-time* process.

By setting a generic point in time, for example, $t = t_0$, one element of the ordered set is selected, i.e. a real random variable X_{t_0} is chosen. It is known that more *outcomes* can be associated with each random variable (named also *measures* or *observations*). An outcome is realized according to the probability or probability density law that characterizes the random variable.

The set of all possible outcomes forms the *sample space*, denoted here by Ω. Accordingly, the random variable can be seen as a *set function* $\{X_{t_0}(\omega); \omega \in \Omega\}$. Fixing a specific element of the sample space, for example $\omega = \omega_0$, yields the value $X_{t_0}(\omega_0)$, which represents a single numerical value (measurement or observation) associated with both the time t and the outcome ω. It is therefore evident that the notation $\{\mathrm{X_t; t} \in \mathcal{T}\}$ could more completely be written as $\{X_t(\omega); \omega \in \Omega\}$, but the explicit reference to the sample space is usually implicit and omitted.

To better understand the intuition behind the expression $\{X_t(\omega_0); t \in \mathcal{T}\}$: by fixing a single element of the sample space $\omega = \omega_0$, this set represents a sequence of specific realizations, each associated with its corresponding random variable within the index set $\mathcal{T}$. This sequence is known as a *realization of the stochastic process* and it is represented as a continuous or discrete curve along the time axis (or index axis t) depending on whether $\mathcal{T}$ is continuous or discrete.

Figure 1.1 illustrates segments of 10 realizations of the stochastic process $X_t(\omega)$. *Figure 1.2* shows the values taken by six random variables across the same realizations from *Figure 1.1*.

1.1 Probabilistic Characterization of Stochastic Processes

To distinguish stochastic processes unambiguously, it is necessary to fully characterize them from a probabilistic perspective. To illustrate the importance of such characterization, consider the simplest case of two continuous random variables X and Y, for which the distribution functions $F_X(x)$ and $F_Y(y)$ exist. We say that $X = Y$, that is, they are identically distributed if $F_X(x) = F_Y(y)$ with probability 1, meaning their distributions are almost surely the same. Otherwise, one may write $X \neq Y$.

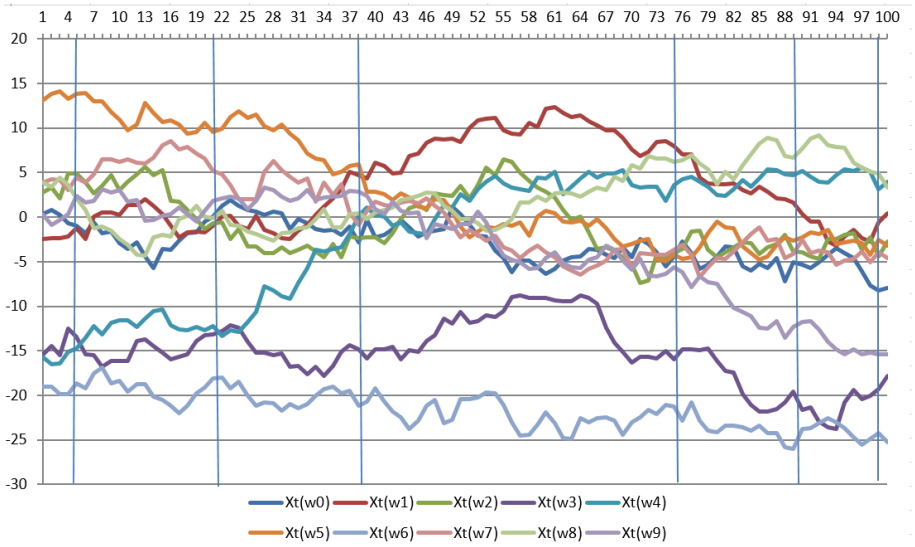


Figure 1.1: *Segments of 10 realizations of a stochastic process* $X_t(\omega)$

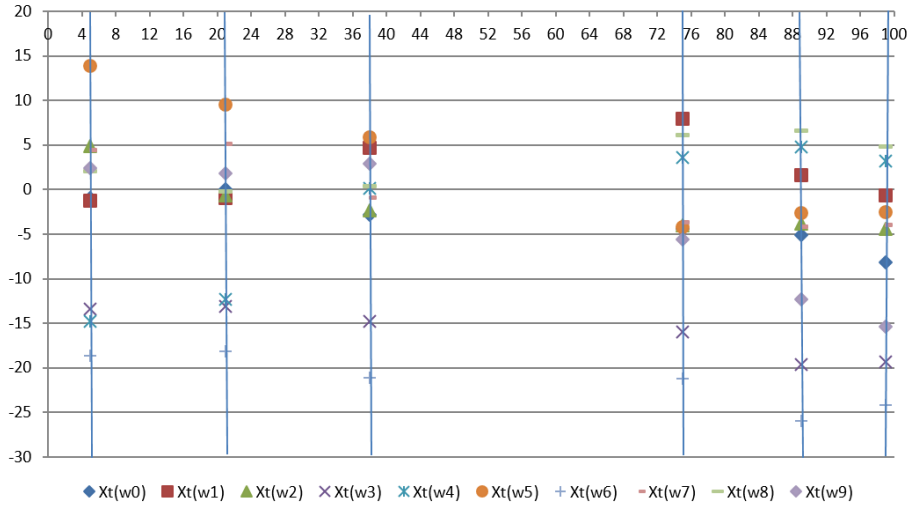


Figure 1.2: *Values taken by six random variables -* $X_5, X_{21}, X_{38}, X_{75}, X_{89}, X_{99}$ *- across the same realizations from Figure 1.1. (The points in Figure 1.2 preserve the same color coding as the realizations in Figure 1.1 and represent the exact intersection of vertical lines with those realizations)*

It is important to note that probabilistic equality does not coincide with mathematical equality. The following example highlights this point.

Suppose both X and Y are uniformly distributed, with densities given respectively by:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & if\ a < x < b \\ 0 & elsewhere \end{cases},$$

and:

$$f_Y(y) = \begin{cases} \dfrac{1}{b-a}, & if\ a \leq y \leq b \\ 0 & elsewhere \end{cases}.$$

The only difference between the two distributions lies in the inclusion or exclusion of the endpoints a and b. Variable X is defined in the open interval (a, b), while Y is defined in the closed interval $[a, b]$. From probabilistic standpoint, X and Y can be considered identically distributed, i.e., $X = Y$. However, mathematically they are different because their domains are not identical. Indeed, $\Pr(Y = a) = \Pr(Y = b) = 0$, as Y is a continuous random variable. We can also say that X and Y are equal in distribution except on a finite set of zero probability. This type of equality in distribution is sometimes denoted as $X \overset{d}{=} Y$, where the symbol $\overset{d}{=}$ represents distributional equality.

The same reasoning can be extended to vectors of random variables. In such cases, we refer to multivariate distributions.
Since a stochastic process is an infinite collection of random variables, its probabilistic characterization would require defining distributions over infinite dimensions. This is mathematically intractable.
Kolmogorov (1933) resolved this issue through his *extension theorem* (often referred to as the *fundamental theorem of stochastic processes*)[1].
The theorem states that a stochastic process is well-defined, in the probabilistic sense, if and only if there exists a *family of finite-dimensional distributions* of the form:

$$F_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n) \tag{1.1}$$

where the finite dimension n can be selected arbitrarily.
Distributions (1.1) must satisfy the following conditions:

i) *Symmetry*:

$$F_{X_{i_1},\ldots,X_{i_n}}(x_{i_1}, \ldots, x_{i_n}) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \tag{1.2}$$

for any permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$.

[1] The reference cited is the 1956 English translation Kolmogorov (1956): *Foundations of the Theory of Probability*. The original formulation appeared in the 1933 monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung*, published in German, as part of a series of mathematical results founded in 1933 by a group of internationally renowned mathematicians (including the Italian *Tullio Levi-Civita*).

ii) *Compatibility*:

$$F_{X_1,\ldots,X_m,X_{m+1},\ldots,X_n}(x_1,\ldots,x_m,+\infty,\ldots,+\infty) = F_{X_1,\ldots,X_m}(x_1,\ldots,x_m), \qquad (1.3)$$

for $m < n$.

The condition of symmetry means that the finite-dimensional family of distributions must be invariant under any permutation of X_{i_t} and x_{i_t}; that is, the specific labels assigned to the variables X_1, X_2, ... are irrelevant.
The compatibility condition refers to the limit of the joint distribution:

$$\begin{aligned} &F_{X_1,\ldots,X_m,X_{m+1},\ldots,X_n}(x_1,\ldots,x_m,+\infty,\ldots,+\infty) \\ &= \lim_{x_{m+1}\to+\infty} \lim_{x_{m+2}\to+\infty} \cdots \lim_{x_n\to+\infty} F_{X_1,\ldots,X_m,X_{m+1},\ldots,X_n}(x_1,\ldots,x_m,x_{m+1},\ldots,x_n). \end{aligned}$$

This limit must be equal to the marginal distribution[2] of $X_1, X_2, \ldots, X_m$.

Generally, it is difficult to use this characterization in terms of families of dimensionally finite distributions and it is preferred to build stochastic processes by defining their characteristics in terms of their moments or to use elementary processes,[3] as i.i.d. or white noise processes, defining on them more complex functional structures.

1.2 Moments of Stochastic Processes

The moments are synthetic numerical characteristics, or theoretical values, that provide a partial description of the behavior of a stochastic process. These synthetic values are calculated as expected values, using the linear expectation operator E.

Definition 1.2. *(r-th moment)*
For a univariate stochastic process, the r-th order moment is defined as:

$$\mu_{r,t} = E(X_t^r)$$

Example 1.1. Suppose that the stochastic process is $X_t \sim N(\mu_t, \sigma_t^2)$, $\forall t \in T$, then

$$\mu_{r,t} = \int_{-\infty}^{\infty} x_t^r f(x_t) dx_t$$

[2] A multidimensional distribution must satisfy the following properties:
(1) it is non-decreasing in each of its arguments;
(2) it is right-continuous in each of its arguments; and
(3) it satisfies the following conditions:
(3.1) $F_{X_1,X_2,\ldots,X_m}(+\infty,+\infty,\ldots,+\infty) = 1$;
(3.2) $\lim_{x_k\to-\infty} F_{X_1,X_2,\ldots,X_m}(x_1,\ldots,x_k,\ldots,x_m) = 0$ for any $1 \le k \le m$, and for any arbitrary values of the other arguments.

[3] See next sections § 2.1 and §2.2.

where $f(x_t) = \dfrac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{ -\dfrac{(x_t - \mu_t)^2}{2\sigma_t^2} \right\}$.

In particular, when $r = 1$ the moment $\mu_{1,t} = \mu_t$ coincides with the theoretical mean of the process at time t. This value appears as a parameter in the definition of the normal probability density $f(x_t)$. The range of odd-order moments $\mu_{2r+1,t}, \quad r = 0, 1, 2, \cdots$ is $[-\infty, +\infty]$, while for even-order moments $\mu_{2r,t}, \quad r = 1, 2, \cdots$, the range is $[0, +\infty]$.
A relevant numerical characteristic of the stochastic process is the *variance* defined by:

Definition 1.3. *(Variance of a stochastic process)*
The variance of a stochastic process is defined by:

$$\sigma_t^2 = E(X_t - \mu_t)^2, .$$

It is easy to verify that $\sigma_t^2 = \mu_{2,t} - \mu_{1,t}^2,$ with $0 \leq \sigma_t^2 \leq \infty$.
Up to this point, there is a strong analogy with the moments defined for random variables. However, for stochastic processes, it is important to define other peculiar characteristics, such as the *autocovariance function* and the *autocorrelation function*. The autocovariance function (ACF) is a measure of the linear time dependence of the stochastic process.

Definition 1.4. *(Autocovariance function)*
The autocovariance function is defined as:

$$\begin{aligned} \gamma_{t,k} &= Cov(X_t, X_{t+k}) \\ &= E(X_t - EX_t)(X_{t+k} - EX_{t+k}) \quad , \\ &= EX_tX_{t+k} - EX_tEX_{t+k}, \ \ k \in \mathbb{Z} \end{aligned} \tag{1.4}$$

where $\mathbb{Z}$ is the set of integers.

The range of the autocovariance function is $[-\infty, +\infty]$.
The autocorrelation function of the stochastic process $\{X_t, t \in \mathcal{T}\}$, is defined as follows:

Definition 1.5. *(Autocorrelation function)*

$$\rho_{t,k} = \frac{\gamma_{t,k}}{\sqrt{\gamma_{t,t}}\sqrt{\gamma_{t+k,t+k}}} \tag{1.5}$$

The terms $\gamma_{t,t}$ and $\gamma_{t+k,t+k}$ represent the variances of the process at time t and $t + k$.
The range of $\rho_{t,k}$ is $[-1, \ 1]$, Since the values of autocorrelation are independent of the scale of measurement of the stochastic process, the autocorrelation function can be used to compare the autocorrelation structure of different stochastic processes. Note that $\rho_{t,t} = 1$.
In general, $\gamma_{t,k} \neq \gamma_{t,-k}$ and $\rho_{t,k} \neq \rho_{t,-k}$.

1.3 Stationarity

Definition 1.6. *(strictly stationary process)*
A stochastic process $\{X_t, t \in \mathcal{T}\}$ is a strictly stationary process if it is characterized by a family of finite-dimensional distributions that are invariant with respect to any shift of the process along the time axis. That is:

$$F_{t_1+k,t_2+k,\ldots t_n+k}(x_1,\, x_2, \ldots,\, x_n) = F_{t_1,t_2,\ldots t_n}(x_1,\, x_2, \ldots,\, x_n), \quad \forall k \in \mathbb{Z}, \tag{1.6}$$

Strict stationarity has a direct implication on the moments of the stochastic process: all moments, of any order, are time-invariant. In particular, the first and second moments—i.e., the mean and variance—are constant over time.
The autocovariance function is also time-invariant, though not invariant with respect to the time lag. This follows from the identity:

$$\gamma_{t+\tau,k} = \gamma_{t,k}. \;\; \forall \tau \in \mathcal{T}$$

Since τ can take any positive or negative integer value, this identity shows that the autocovariance function is independent of the specific point in time to which it refers.
Consequently, the autocovariance function depends only on the time-lag k and can be written as:

$$Cov(X_t,\; X_{t+k}) = \gamma_k. \;\; \forall k \in \mathbb{Z}\,.$$

The same reasoning applies to the autocorrelation function.

Definition 1.7. *(weakly stationary or covariance stationary process)*
A stochastic process is weakly stationary or covariance stationary if:

1. $E(X_t) = \mu \;\; \forall t \in \mathcal{T}$,
2. $E(X_t - \mu)^2 = \sigma^2 < \infty \;\; \forall t \in \mathcal{T}$,
3. $E[(X_t - \mu)(X_{t+k} - \mu)] = \gamma_k \;\; \forall t \in T, \;\; \forall k \in \mathbb{Z}$

Strict stationarity implies these three properties, but they must be seen as consequences of the behavior of the finite-dimensional distributions, not as defining conditions. Furthermore, weak stationarity does not require the existence of moments of order higher than two or their time-invariance.
If a stationary stochastic process in covariance is Gaussian, then it is also strict stationary.

From the assumption of stationarity, the autocovariance function satisfies the following properties:

1. $\gamma_k = \gamma_{-k}$ *(Symmetry, the autocovariance is an even function*[4] *)*
   In fact, $\gamma_{-k} = E\left[X_t - E(X_t)\right]\left[X_{t-k} - E(X_{t-k})\right]$, defining $s = t - k$, with a substitution we obtain $\gamma_{-k} = E\left[X_{s+k} - E(X_{s+k})\right]\left[X_s - E(X_s)\right] = \gamma_k$.
2. $|\gamma_k| \leq \gamma_0$ *(Cauchy-Schwartz inequality* [5] *)*
3. $\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i\delta_j\gamma_{|i-j|} \geq 0$ *(Semi-definite positive function)*

   If we consider the following linear combination $Y_t = \sum_{j=1}^{n} \delta_j X_{t-j}$, then its variance is:

$$Var(Y_t) = Var\left(\sum_{j=1}^{n} \delta_j X_{t-j}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i\delta_j Cov(X_{t-i}, X_{t-j}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i\delta_j\gamma_{|i-j|} \geq 0$$

The same properties hold for the autocorrelation function:

1. $\rho_k = \rho_{-k}$
2. $|\rho_k| \leq 1$
3. $\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i\delta_j\rho_{|i-j|} \geq 0$
   with $\rho_0 = 1$.

Every strictly stationary process with finite second moments is covariance stationary. If a covariance-stationary process is Gaussian, then it is also strictly stationary.

[4] An even function satisfies the property: $f(x) = f(-x)$. An odd function satisfies the property $f(x) = -f(x)$.
[5] See Appendix 1.A.

Appendix 1.A *(Cauchy-Schwarz Inequality)*

Proposition 1.A1. *(Cauchy-Schwarz Inequality)*
The following inequality, known as the Cauchy–Schwarz inequality, holds:

$$\big(E[XY]\big)^2 \;\le\; E[X^2]\,E[Y^2].$$

Proof. Since the expected value of a non-negative random variable is always non-negative, for every $\lambda \in \mathbb{R}$ we have:

$$0 \le E\left[(\lambda X + Y)^2\right] = \lambda^2 E\left[X^2\right] + 2\lambda E\left[XY\right] + E\left[Y^2\right]$$

Viewing the right-hand side as a quadratic function of λ, we observe that it must be non-negative for all $\lambda \in \mathbb{R}$. Therefore the function $E\left[(\lambda X + Y)^2\right]$ describes a parabolic curve, that becomes tangent to λ real axis if and only if its discriminant assumes zero value. A quadratic function is non-negative for all λ if and only if its discriminant is non-positive, that is:

$$\begin{aligned} 4\big(E[XY]\big)^2 - 4E[X^2]E[Y^2] &\le 0 \\ \big(E[XY]\big)^2 &\le E[X^2]E[Y^2] \end{aligned}$$

□

Proposition 1.A1 when applied to random variables X_t and X_{t+k} leads to the following result:

$$\begin{aligned} [E(X_t - EX_t)(X_{t+k} - EX_{t+k})]^2 &\le E(X_t - EX_t)^2 E(X_{t+k} - EX_{t+k})^2 \\ \gamma_k^2 &\le \gamma_0^2 \end{aligned}$$

Dividing both sides by γ_0^2 and taking square roots yields $|\rho_k| \le 1$.

2 Relevant Stationary Processes

2.1 Independent and Identically Distributed Process (i.i.d.)

Definition 2.1. *(i.i.d. process)*[6]
The stochastic process $\{\varepsilon_t,\, t \in \mathcal{T}\}$ *is called an i.i.d. process if each random variable in the collection has the same probability distribution as the others and all are mutually independent.*

Due to independence and identical distribution, for any selection of variables $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ we have:

$$F_{t_1,t_2,\ldots t_n}(\varepsilon_1,\, \varepsilon_2, \ldots,\, \varepsilon_n) = F(\varepsilon_1)F(\varepsilon_2)\cdots F(\varepsilon_n) = \prod_{j=1}^{n} F(\varepsilon_j) \tag{2.1}$$

The same result applies to a shifted selection of variables, such as $\varepsilon_{1+k}, \varepsilon_{2+k}, \cdots, \varepsilon_{n+k}$:

$$F_{t_1+k,t_2+k,\ldots,t_n+k}(\varepsilon_1,\, \varepsilon_2, \ldots,\, \varepsilon_n) = F(\varepsilon_1)F(\varepsilon_2)\cdots F(\varepsilon_n) = \prod_{j=1}^{n} F(\varepsilon_j)$$

In conclusion, it is evident that the i.i.d. process is strictly stationary.

Moments of i.i.d. process
Given strict stationarity, all the distribution moments are independent of time. In particular, the mean and variance are invariant over time. The r-th moment is given by:

$$\mu_{r,t} = E(\varepsilon_t^r) = \mu_r \tag{2.2}$$

By convention, we assume $E(\varepsilon_t) = 0$. The autocovariance function is:

$$\gamma_k = \begin{cases} \sigma_\varepsilon^2 \geq 0, & k = 0 \\ 0, & k \neq 0 \end{cases} \tag{2.3}$$

and the autocorrelation function is:

$$\rho_k = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases} \tag{2.4}$$

This implies that an i.i.d. process has zero autocorrelation for all non-zero lags, serves as a useful benchmark for detecting serial correlation in time series data.

[6] The symbol ε_t is conventionally used to denote random errors in regression models.

2.2 White Noise and Other Related Stationary Processes

Definition 2.2. *(White noise process)*
The stochastic process $\{\varepsilon_t,\, t \in \mathcal{T}\}$ *is called a white noise process if the following properties are satisfied:*

1) $E(\varepsilon_t) = \mu, \quad \forall t \in \mathcal{T}$. *By convention,* $\mu = 0$ *is assumed.*

2) $E(\varepsilon_t^2) = \sigma^2, \quad \forall t \in \mathcal{T}$

3) $E(\varepsilon_t \varepsilon_{t+k}) = 0, \forall t \in \mathcal{T}$ *and* $k \neq 0$,

These three properties are also satisfied by the i.i.d. process. However, unlike the i.i.d. process, white noise does not impose any condition on moments higher than the second. For example, a white noise process may have a time-varying fourth moment. If the white noise process is Gaussian, it is also an i.i.d. process, since a Gaussian process is fully characterized by its first two moments and autocovariance function. White noise plays a central role in linear stochastic processes, as it forms the elementary building block on which they are constructed.

Example 2.1. [7] Let y_t's be a collection of i.i.d. random variables such that:

$$E(y_t) = \mu \text{ and } Var(y_t) = \sigma^2 \tag{2.5}$$

then:

$$\gamma_{t,s} = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s. \end{cases} \tag{2.6}$$

This process is strictly stationary, but if we drop the property of identical distribution [while the properties (2.5) and (2.6) are retained], the resulting process is covariance stationary, but not strictly stationary.

Example 2.2. Let all the y_t's be identically equal to a random variable y.
If there are the first two moments:

$$E(y_t) = E(y) = \mu \tag{2.7}$$

$$\gamma_{t,s} = Var(y) = \sigma^2, \ \forall t, s\,, \tag{2.8}$$

then this process is strictly stationary .

[7] This example, along with the next two, is adapted from Anderson (1971), p. 374-376.

Example 2.3. [8] Define the process $\{y_t\}$ as follows:

$$y_t = \sum_{j=1}^{q} (A_j \cos \omega_j t + B_j \sin \omega_j t), t = \cdots, -1, 0, 1, \cdots \tag{2.9}$$

where the ω_j are constant and $A_1, \cdots, A_q, B_1, \cdots, B_q$ are random variables such that:

$$EA_j = EB_j = 0, j = 1, \cdots, q, \tag{2.10}$$

$$EA_j^2 = EB_j^2 = \sigma_j^2, j = 1, \cdots, q, \tag{2.11}$$

$$EA_iA_j = EB_iB_j = 0, i \neq j, i, j = 1, \cdots, q, \tag{2.12}$$

$$EA_iB_j = 0, i, j = 1, \cdots, q, \tag{2.13}$$

then:

$$Ey_t = 0, \tag{2.14}$$

$$\begin{aligned} Ey_t y_s &= \sum_{j=1}^{q} E(A_j \cos \omega_j t + B_j \sin \omega_j t)(A_j \cos \omega_j s + B_j \sin \omega_j s) \\ &= \sum_{j=1}^{q} (EA_j^2 \cos \omega_j t \cos \omega_j s + EB_j^2 \sin \omega_j t \sin \omega_j s) \\ &= \sum_{j=1}^{q} \sigma_j^2 (\cos \omega_j t \cos \omega_j s + \sin \omega_j t \sin \omega_j s) \\ &= \sum_{j=1}^{q} \sigma_j^2 \cos \omega_j (t - s) \end{aligned} \tag{2.15}$$

Hence, the process is covariance stationary. If A_j and B_j are normally distributed, then y_t is also normally distributed as it is a linear combination of jointly normal variables. Consequently, the process is strictly stationary

The process can also be written as:

$$y_t = \sum_{j=1}^{q} R_j \cos(\omega_j t - \theta_j), t = \cdots, -1, 0, 1, \cdots \tag{2.16}$$

where:

$$R_j^2 = A_j^2 + B_j^2, j = 1, \cdots, q, \tag{2.17}$$

$$\tan \theta_j = \frac{B_j}{A_j}, j = 1, \cdots, q, \tag{2.18}$$

and $0 < \theta_j < \pi$ if $B_j > 0$ and $\pi < \theta_j < 2\pi$ if $B_j < 0$.

[8] This type of process, known as a periodic or harmonic process, is considered in the Wold decomposition theorem (see Chapter 3).

If A_j and B_j are normally distributed, then R_j^2 is proportional to a chi-square variable with 2 degrees of freedom and θ_j is uniformly distributed between 0 and 2π (by the symmetry of the normal distribution) and is independent of R_j^2.
If A_j and B_j are not normally distributed, $\{y_t\}$ is not necessarily stationary in the strict sense.

This example is important[9] because, in a sense, every stochastic process stationary in covariance with finite variance can be approximated by a linear combination such as the right-hand side of (2.9).

Terminology of periodic processes
The constants ω_j are called *angular frequencies*.
The variables R_j are the *harmonic amplitudes*.
The variables θ_j are the *angular phases*.
The sinusoidal component with frequency ω_j is called the *j-th harmonic* of the periodic stochastic process.

Example 2.4. Graphic representation of harmonics with different amplitude, phase, and frequency:
(See YouTube) `https://www.youtube.com/watch?v=cUD1gMAl6W4`
In this graphic representation:

The amplitude is represented by the radius (pendulum) of the circles.

The phase by the starting point of the pendulum.

The frequency is the speed of the pendulum completing 360° around the circle.

The fifth graph is the sum of the first four harmonics.

2.3 Moving Average Process (*MA*)

Definition 2.3. *(Moving average - MA(q))*
A moving average process of order q is defined by the following linear combination:

$$X_t = \sum_{j=0}^{q} \beta_j \varepsilon_{t-j}, \quad \beta_0 = 1, \tag{2.19}$$

where $\{\varepsilon_t\}$ is a white noise process with zero mean. The condition $\beta_0 = 1$ is conventional and does not restrict generality.

[9] See §3.1 on Wold Decomposition Theorem.

The order q implies that $\beta_q \neq 0$. However, the values of β_j for $j < q$ may be zero or not, without affecting the order.
Using the linear lag operator L (see *Appendix 2.A*), the process can be rewritten more compactly as:

$$\begin{cases} X_t = \sum_{j=0}^{q} \beta_j L^j \varepsilon_t = \beta(L)\varepsilon_t \\ \beta(L) = (1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q) \end{cases} \tag{2.20}$$

Is MA(q) a stationary process in covariance?

In general, the linear combination of covariance stationary processes results in a covariance stationary process[10]

To verify this, we analyze the moment properties of the simplest case, *MA(1)*.

***MA(1)* Process**
From the general definition $MA(1): \; X_t = \varepsilon_t + \beta\varepsilon_{t-1}$.
Its properties are:

1) $E(X_t) = E(\varepsilon_t + \beta\varepsilon_{t-1}) = E(\varepsilon_t) + \beta E(\varepsilon_{t-1}) = 0$;

2) $$\begin{aligned} Var(X_t) &= Var(\varepsilon_t + \beta\varepsilon_{t-1}) = Var(\varepsilon_t) + \beta^2 Var(\varepsilon_{t-1}) + 2\beta Cov(\varepsilon_t,\ \varepsilon_{t-1}) \\ &= \gamma_0 = (1 + \beta^2)\sigma_\varepsilon^2 \end{aligned}$$

3) $$\begin{aligned} Cov(X_t, X_{t+k}) &= Cov(\varepsilon_t,\ \varepsilon_{t+k}) + \beta Cov(\varepsilon_{t-1},\ \varepsilon_{t+k}) \\ &+ \beta Cov(\varepsilon_t,\ \varepsilon_{t+k-1}) + \beta^2 Cov(\varepsilon_{t-1},\ \varepsilon_{t+k-1}) \\ &= \gamma_k = \begin{cases} \beta\sigma_\varepsilon^2 & k = 1 \\ 0 & k > 1 \end{cases} \end{aligned}$$

Since mean and variance are constant and the autocovariance depends only on the lag k, the process is covariance stationary.
The same logic extends to the general case of *MA(q)*.

***MA(q)* Process**
The properties are:

1) $E(X_t) = E\left(\sum_{j=0}^{q} \beta_j \varepsilon_{t-j}\right) = \sum_{j=0}^{q} \beta_j E(\varepsilon_{t-j}) = 0$;

2) $Var(X_t) = Var\left(\sum_{j=0}^{q} \beta_j \varepsilon_{t-j}\right) = \gamma_0 = \sum_{j=0}^{q} \beta_j^2 Var(\varepsilon_{t-j}) = \sigma_\varepsilon^2 \sum_{j=0}^{q} \beta_j^2$;

[10] There are examples of linear combinations of covariance stationary processes that are not themselves stationary. See Kemp (1997).

$$3)\ Cov(X_t,\, X_{t+k}) = Cov\left(\sum_{j=0}^{q}\beta_j\varepsilon_{t-j}, \sum_{j=0}^{q}\beta_j\varepsilon_{t+k-j}\right)$$

$$= \gamma_k = \begin{cases} \sigma_\varepsilon^2 \sum_{j=0}^{q-k}\beta_j\beta_{j+k}, & k = 1,\, 2,\, \cdots,\, q \\ 0, & k > q \\ \gamma_{-k}, & k < 0 \end{cases}$$

The result in 2) comes from the fact that for the white noise process, the variance of the sum is equal to the sum of the variances, being all the covariances zero.
The result in 3) is obtained considering the sum of all the expected values:

$$E(\beta_i\beta_j\varepsilon_i\varepsilon_j),\ \ i = 0,\, \cdots q-k;\ j = 0,\, \cdots,\, q;\ k \le q,$$

calculated within the following table of dimensions $(q-k+1)\times(q+1)$:

	ε_{t+k}	$\beta_1\varepsilon_{t+k-1}$	$\cdots$	$\beta_k\varepsilon_t$	$\beta_{k+1}\varepsilon_{t-1}$	$\cdots$	$\beta_q\varepsilon_{t+k-q}$
ε_t	0	0	$\cdots$	$\beta_k\sigma_\varepsilon^2$	0	$\cdots$	0
$\beta_1\varepsilon_{t-1}$	0	0	$\cdots$	0	$\beta_1\beta_{k+1}\sigma_\varepsilon^2$	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\beta_{q-k}\varepsilon_{t+k-q}$	0	0	$\cdots$	0	0	$\cdots$	$\beta_{q-k}\beta_q\sigma_\varepsilon^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_q\varepsilon_{t-q}$	0	0	$\cdots$	0	0	$\cdots$	0

When $k < q$ grows, the table has fewer rows and the diagonal of non-null values moves to the right. The sum of the calculated values inside the table is:

$$\sigma_\varepsilon^2(\beta_k + \beta_1\beta_{k+1} + \cdots + \beta_{q-k}\beta_q),$$

which is the desired result.
The three properties mentioned above show that the $MA(q),\ \ q \ge 1$ process is covariance stationary.
The autocorrelation function is:

$$
\begin{aligned}
Corr(X_t,\, X_{t+k}) &= \frac{\gamma_k}{\gamma_0} \\
&= \rho_k = \begin{cases} \dfrac{\sum_{j=0}^{q-k} \beta_j \beta_{j+k}}{\sum_{j=0}^{q} \beta_j^2} & k = 1,\, 2,\, \cdots,\, q \\ 0, & k > q \\ \rho_{-k}, & k < 0 \end{cases}
\end{aligned} \tag{2.21}
$$

The term σ_ε^2, which appears in both the variance and the covariance functions, cancels out in the expression of the autocorrelation function since the latter is independent of the scale of measurement of the process.

Graphically[11], the autocorrelation curve generally exhibits nonzero values for some lags $k < q$, and it necessarily has a nonzero value at lag q, where q is the order of the process. For all lags $k > q$ the autocorrelation is identically zero.

In the graph shown in *Figure 2.1*, which corresponds to an *MA(4)* process, we observe that $\rho_2 = 0$. However, the relevant information lies in the fact that all values for $k > q$ are zero, clearly indicating that the last nonzero autocorrelation occurs at lag $k = 4$, thereby identifying the order of the process within the *MA(q)* class.

In conclusion, *the autocorrelation function* is a key diagnostic tool for determining the order of an *MA* process

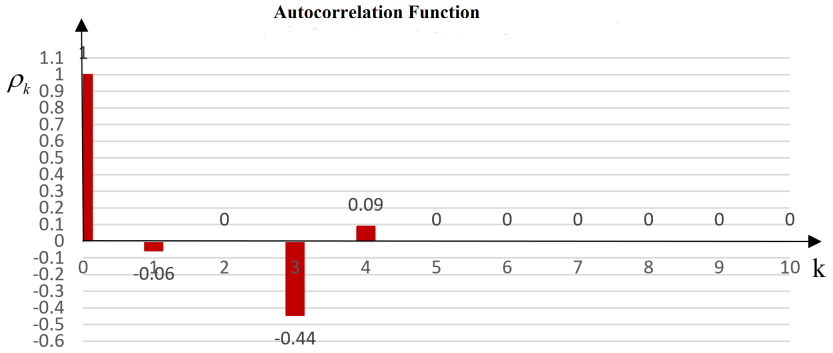


Figure 2.1: *Autocorrelation function for the process* $MA(4)$: $X_t = \varepsilon_t - 5\varepsilon_{t-1} + 8\varepsilon_{t-4}$

[11] The graph of the autocovariance function is plotted only for non-negative lags because, in the stationary case, it is an *even function*.

2.4 Autoregressive Process (AR)

Definition 2.4. *An autoregressive process of order p is defined as a stochastic process $\{X_t\}$ satisfying the linear equation:*

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \varepsilon_t, \tag{2.22}$$

where $\{\varepsilon_t\}$ is a white noise process with zero mean and constant variance, and $\alpha_1, ..., \alpha_p$ are real coefficients.

The term *autoregressive* is used because the current value of the process depends on its own past values, up to p lags.
The order p of the process implies that the coefficient $\alpha_p \neq 0$, while the values of the other coefficients—whether zero or not—do not affect the order.
Using the lag operator L, the process can be written as:

$$\begin{array}{l} X_t = \sum_{j=1}^{p} \alpha_j L^j X_t + \varepsilon_t \\ X_t - \sum_{j=1}^{p} \alpha_j L^j X_t = \varepsilon_t \\ \left(1 - \sum_{j=1}^{p} \alpha_j L^j\right) X_t = \varepsilon_t \end{array},$$

or, compactly, as:

$$\begin{cases} \alpha(L) X_t = \varepsilon_t \\ \alpha(L) = (1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p) \ . \end{cases} \tag{2.23}$$

We now examine whether the AR process is covariance stationary.
This analysis begins with the simplest member of the AR class: the *AR(1)* process.

AR(1) process
From the general definition of AR processes, we have:

$$X_t = \alpha X_{t-1} + \varepsilon_t \tag{2.24}$$

To verify covariance stationarity, we compute the mean, variance, and autocovariance function.
We begin with the mean. Taking expectations on both sides of equation (2.24) yields:

$$E(X_t) = \alpha E(X_{t-1}),$$

This is a recursive equation that cannot be resolved directly without knowing the initial condition.
One way to overcome this limitation is to express the AR process solely in terms of the white noise process.
Since $X_{t-1} = \alpha X_{t-2} + \varepsilon_{t-1}$, we substitute into X_t:

$$\begin{aligned} X_t &= \alpha(\alpha X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \alpha^2 X_{t-2} + \alpha\varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

Repeating the substitution:

$$\begin{aligned} X_t &= \alpha^2(\alpha X_{t-3} + \varepsilon_{t-2}) + \alpha\varepsilon_{t-1} + \varepsilon_t \\ &= \alpha^3 X_{t-3} + \alpha^2\varepsilon_{t-2} + \alpha\varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

Continuing this process recursively, after replacing X_{t-k+1} we obtain:

$$\begin{aligned} X_t &= \alpha^k X_{t-k} + \alpha^{k-1}\varepsilon_{t-k+1} + \alpha^{k-2}\varepsilon_{t-k+2} + \cdots + \alpha\varepsilon_{t-1} + \varepsilon_t \\ &= \alpha^k X_{t-k} + \sum_{j=0}^{k-1} \alpha^j \varepsilon_{t-j} \end{aligned} \tag{2.25}$$

The goal is to reach the infinite moving average representation:

$$X_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}. \tag{2.26}$$

This representation is valid only if the term $\alpha^k X_{t-k}$ vanishes as $k \to \infty$ removing dependence on the initial condition.
This requires the following two conditions:

1. The infinite sum involving white noise must converge, which holds if $|\alpha| < 1$. This is known as the *stationarity condition*.

2. The remainder term[12] $\alpha^k X_{t-k}$ must vanish in the mean square sense.

   From (2.25) we write:

   $$\alpha^k X_{t-k} = X_t - \sum_{j=0}^{k-1} \alpha^j \varepsilon_{t-j} \ ,$$

   If the second moment $E(\alpha^k X_{t-k})^2 = \alpha^{2k} E(X_{t-k})^2$ exists for all t, k, that is, if $E(X_{t-k})^2 < \infty$, for all t, k, then, given $\alpha < 1$, follows that

   $$\lim_{k\to\infty} \alpha^{2k} E(X_{t-k})^2 = 0$$

[12] This second condition is usually assumed and often omitted in textbooks.

Consequently,

$$\lim_{k\to\infty} E\left(X_t - \sum_{j=0}^{k-1} \alpha^j \varepsilon_{t-j}\right)^2 = 0, \tag{2.27}$$

meaning that the process X_t is approximated in the *mean square sense* by the infinite sum of past white noise terms. Therefore the relation (2.26) is justified.

The same result is derived using the lag operator L.
The AR(1) model is:

$$X_t = \alpha X_{t-1} + \varepsilon_t \;\; \rightarrow X_t - \alpha L X_t = \varepsilon_t \;\; \rightarrow (1-\alpha L) X_t = \varepsilon_t$$

or

$$\begin{cases} \alpha(L) X_t &= \varepsilon_t \\ \alpha(L) &= 1 - \alpha L \end{cases} \tag{2.28}$$

Inverting the polynomial $\alpha(L)$, we have:

$$X_t = \alpha(L)^{-1} \varepsilon_t = \sum_{j=0}^{\infty} \alpha^j L^j \varepsilon_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \tag{2.29}$$

because:

$$\alpha(L)^{-1} = \frac{1}{1-\alpha L} = \sum_{j=0}^{\infty} \alpha^j L^j \,. \tag{2.30}$$

This equality holds because the expression $1/(1-\alpha L)$ is the sum of a geometric series[13] with common ratio $\alpha\, L$, valid when $|\alpha| < 1$.
The stationarity condition can alternatively be expressed in term of the *characteristic equation* associated with (2.24):

$$\alpha(L) = 1 - \alpha L = 0 \tag{2.31}$$

The root of this equation is[14] $L = 1/\alpha$; stationarity condition $|\alpha| < 1$ is thus equivalent to requiring that the *root of the polynomial equation* $\alpha(L) = 0$ *must lie outside the unit interval.*

[13] A geometric series is defined as: $r^0 + r^1 + r^2 + \cdots + r^n = \sum_{j=0}^{n} r^j = (1 - r^{n+1})/(1-r)$, where r is the *common ratio* of the series. If $|r| < 1$ then $r^{n+1} \to 0$, as $n \to \infty$, so the sum becomes $\dfrac{1}{1-r}$.

[14] Here, the lag operator L is treated as if it were a real-valued variable. Strictly speaking, this is a slight abuse of notation. However, this practice is justified by the fact that lag polynomials are isomorphic to ordinary algebraic polynomials.

Expression (2.26) has the form of an $MA(\infty)$ process, and can be written as:

$$X_t = \begin{cases} \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j} \\ \beta_j = \alpha^j \end{cases} \tag{2.32}$$

From a mathematical and probabilistic point of view, equations (2.24) and (2.26) are equivalent. While (2.24) describes the process in terms of its dependence on its immediate past, (2.26) expresses it through an infinite linear combination of white noise.
This representation is useful for computing the moments of the process:

1) $E(X_t) = E\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}\right) = \sum_{j=0}^{\infty} \alpha^j E\varepsilon_{t-j} = 0$

2) $Var(X_t) = Var\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}\right) = \gamma_0 = \sum_{j=0}^{\infty} \alpha^{2j} Var(\varepsilon_{t-j}) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \alpha^{2j} = \frac{\sigma_\varepsilon^2}{1-\alpha^2}$

3)
$$\begin{aligned} Cov(X_t, X_{t+k}) &= Cov\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}, \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t+k-j}\right) \\ &= \gamma_k = \begin{cases} \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \alpha^j \alpha^{j+k} = \sigma_\varepsilon^2 \alpha^k \sum_{j=0}^{\infty} \alpha^{2j} = \frac{\sigma_\varepsilon^2 \alpha^k}{1-\alpha^2} & k > 0 \\ \gamma_{-k} & k < 0 \end{cases} \end{aligned}$$

These results confirm that the AR(1) process is covariance stationary when $|\alpha| < 1$. The autocorrelation function is:

$$\begin{aligned} Corr(X_t, X_{t+k}) &= \frac{\gamma_k}{\gamma_0} \\ &= \rho_k = \begin{cases} \frac{\sigma_\varepsilon^2 \alpha^k}{1-\alpha^2} \frac{1-\alpha^2}{\sigma_\varepsilon^2} = \alpha^k, & k > 0 \\ \rho_{-k} & k < 0 \end{cases} \end{aligned} \tag{2.33}$$

which decays exponentially and tends to zero only as $k \to \infty$.
Figures below show two examples of autocorrelation functions for different values of α. In both figures, the autocorrelation functions are exponentially decreasing, a typical behavior for each AR(1) process if the coefficient is $0 < \alpha < 1$ *(Figure 2.2)* or $-1 < \alpha < 0$ *(Figure 2.3)*.

2.5 *AR(2)* Process

Before examining the stationarity condition for the general *AR(p)* process, it is helpful to consider it in the simpler case of an *AR(2)* process.

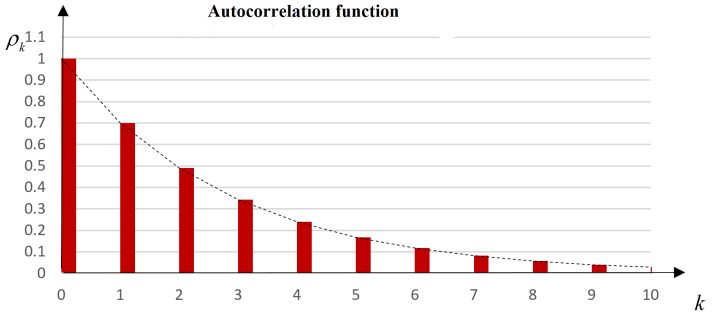


Figure 2.2: *Autocorrelation function for the process* $AR(1): \quad X_t = 0.7X_{t-1} + \varepsilon_t$

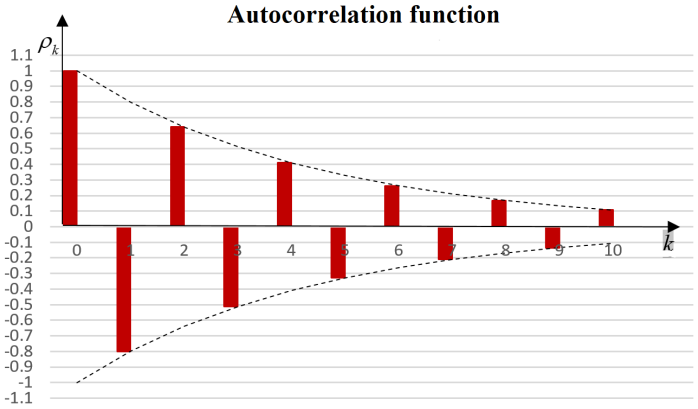


Figure 2.3: *Autocorrelation function for the process* $AR(1): \quad X_t = -0.8X_{t-1} + \varepsilon_t$

From the general definition, the AR(2) process is given by:

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t \\ X_t &- \alpha_1 L X_t - \alpha_2 L^2 X_t = \varepsilon_t \end{aligned}$$

or more compactly:

$$\begin{aligned} &X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t \\ &X_t - \alpha_1 L X_t - \alpha_2 L^2 X_t = \varepsilon_t \\ &\begin{cases} \alpha(L) X_t = \varepsilon_t \\ \alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 \end{cases} \end{aligned} \tag{2.34}$$

The characteristic equation associated with this process is:

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 = 0 \tag{2.35}$$

Any second-degree polynomial can be factored into the product of two binomials, so equation (2.35) can be written as:

$$(1 - w_1 L)(1 - w_2 L) = 0 \tag{2.36}$$

This factorization shows the connection between the values w_i, $i = 1, 2$, and the corresponding roots $r_i = 1/w_i$ of the characteristic equation in L given in (2.36).
It also establishes a relationship between the parameters α_1, α_2 of the process and the values w_1, w_2. In fact, expanding the product of the two factors gives:

$$1 - (w_1 + w_2)L + w_1 w_2 L^2 = 0 \tag{2.37}$$

from which we obtain:

$$\alpha_1 = w_1 + w_2, \quad \alpha_2 = -w_1 w_2$$

As in the *AR(1)* case, the condition for stationarity requires[15] that $|r_i| > 1$, $i = 1, 2$. That is, *both roots must lie outside the unit circle.*[16]
In the AR(2) case, the condition requires that all roots lie outside the unit circle, not just outside the unit interval, since the roots may be complex (see *Appendix 2.B*).
It is therefore useful to examine the stationarity condition of the AR(2) process in more detail.

2.5.1 Stationarity Condition for an *AR(2)* Process

This sub-section analyzes the stationarity conditions in terms of coefficients α_j, $j = 1, 2$.
It is well known that the roots of a second-degree polynomial equation:

$$ax^2 + bx + c = 0\,,$$

are given by:

$$r_1,\, r_2 = \frac{1}{2a}(-b \pm \sqrt{b^2 - 4ac})$$

Using the relations between α_j, $j = 1, 2$ and the inverse roots z_i, $i = 1, 2$, we obtain $z_2 = \alpha_1 - z_1$ and $\alpha_2 = -z_1(\alpha_1 - z_1)$, that is:

$$z_1^2 - \alpha_1 z_1 - \alpha_2 = 0$$

---

[15] This is an intuitive statement; a formal proof follows from Theorem 5.2.1 in Anderson (1971), p. 170.

[16] An equivalent condition is $|z_i| < 1$, $i = 1, 2$. The values z_i are the roots of the reciprocal polynomial associated with the process in (2.34), that is, $z^2 - \alpha_1 z - \alpha_2 = 0$. The relationships between the parameters and the roots are the same as for the characteristic equation (2.36), the two sets of roots being reciprocals of each other.
Throughout the book, the term *characteristic equation* refers to the autoregressive polynomial, while the term *reciprocal polynomial* is used for its inverse representation.

The roots are:

$$z_{1i} = \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2}, \quad i = 1, 2$$

The same result is obtained for $f(z_2) = 0$.
The stationarity condition is:

$$\left| \frac{1}{2}\alpha_1 \pm \frac{1}{2}\sqrt{\alpha_1^2 + 4\alpha_2} \right| < 1 \tag{2.38}$$

Solving[17] this inequality yields the triangular region[18]:

$$\begin{cases} \alpha_2 + \alpha_1 < 1 \\ \alpha_2 - \alpha_1 < 1 \\ -1 < \alpha_2 < 1 \end{cases} \tag{2.39}$$

that is well represented in *Figure 2.4*:

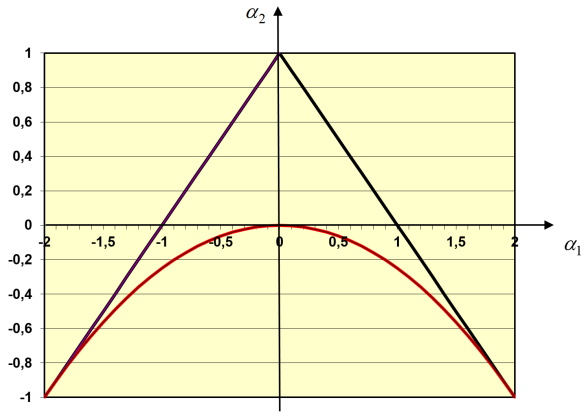


Figure 2.4: *Admissibility stationarity region of* α_1 *and* α_2 *parameters*

The red curve corresponds to the parabola $\alpha_1^2 + 4\alpha_2 = 0$. For values $\alpha_1^2 + 4\alpha_2 \geq 0$, that is, above or on the curve,the roots are real, below it, for values $\alpha_1^2 + 4\alpha_2 < 0$, the roots are complex.
Note that $\alpha_1^2 + 4\alpha_2 < 0$ is only one of the three inequalities defining the system (2.39), and therefore it represents a *necessary but not sufficient condition* for stationarity.

[17] See *Appendix 2.C: Admissibility Region of AR(2) coefficients for stationarity.*
[18] The triangular region is sometimes referred to as *Stralkowski's region* after the author who most thoroughly investigated the characteristics of the AR(2) process with respect to the values of its parameters. See Stralkowski, Wu, and DeVor (1970).

Example 2.5. Consider the process:

$$x_t = -1.9x_{t-1} - 0.9x_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, \sigma_\varepsilon^2).$$

In this case, $\alpha_2 + \alpha_1 = -2.8 < 1$ and $|\alpha_2| < 1$, so both conditions are satisfied. However, $\alpha_2 - \alpha_1 = 1$ which is no less than 1, and therefore the process is not stationary. The roots of the polynomial equation:

$$1 + 1.9w + 0.9w^2 = 0$$

are $w_1 = -1.111...$ and $w_2 = -1.$, the second root is on the unit circle.
Note that the condition:

$$|\alpha_1| + |\alpha_2| < 1,$$

is sufficient for stationarity. The region defined by this inequality corresponds to the square inscribed within the triangular stationarity region shown in *(Figure 2.5)*.

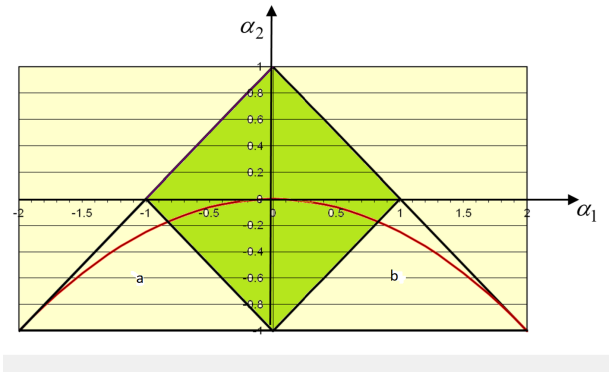


Figure 2.5: *Stationarity sufficient region of* α_1 *and* α_2 *parameters*

It is important to note that the sufficient condition $|\alpha_1| + |\alpha_2| < 1$ is not a necessary one. A process can still be stationary even when this condition is violated, provided that the coefficient pair lies inside triangle **a** or **b**.

Extending these results to the general *AR(p)* process:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, \sigma_\varepsilon^2)$$

We can state the following:

- $\sum_{j=1}^{p} \alpha_j < 1$ is a *necessary condition* for stationarity[19]

[19] See also *Appendix 2.F.*

- $\sum_{j=1}^{p} |\alpha_j| < 1$ is *sufficient condition* for stationarity[19]
- $\sum_{j=1}^{p} \alpha_j = 1$ indicates the *presence of a unit root* in the process[20]

The *AR(2)* process can be represented in its dual form as an $MA(\infty)$ process: $X_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}$, by computing the coefficients $\beta_j = f(\alpha_1, \alpha_2), \ j = 1, 2, \cdots$.
Under the stationarity condition, the AR(2) process can be written as:

$$\begin{array}{l} \alpha(L)X_t = \varepsilon_t \\ X_t = \alpha(L)^{-1}\varepsilon_t \end{array} ,$$

or

$$\left\{ \begin{array}{l} X_t = \beta(L)\varepsilon_t \\ \beta(L) = \alpha(L)^{-1} \end{array} \right. , \tag{2.40}$$

from which it follows that:

$$\begin{array}{l} \beta(L)\alpha(L) = 1 \\ (1 + \beta_1 L + \beta_2 L^2 + \cdots)(1 - \alpha_1 L - \alpha_2 L^2) = 1 \ . \end{array} \tag{2.41}$$

This equality holds if each power of the lag operator L, the coefficients resulting from the polynomial product on the left-hand side match those on the right-hand side.
Expanding the left-hand side yields:

$$\begin{array}{r} 1 + \beta_1 L \quad + \beta_2 L^2 \quad + \beta_3 L^3 \quad + \beta_4 L^4 + \cdots \\ -\alpha_1 L - \alpha_1\beta_1 L^2 - \alpha_1\beta_2 L^3 - \alpha_1\beta_3 L^4 - \cdots \\ -\alpha_2 L^2 - \alpha_2\beta_1 L^3 - \alpha_2\beta_2 L^4 - \cdots = 1 \end{array} .$$

Since the right-hand side of equation (2.41) contains no polynomial terms (only the constant 1), we equate to zero the sum of the coefficients of each power of L in the expanded left-hand side.
Eventually, the relation between the MA(∞) coefficients β_j and the AR(2) coefficients α_1, α_2 is given by the recursive formula:

$$\beta_k = \alpha_1\beta_{k-1} + \alpha_2\beta_{k-2}, \ \ k = 2, 3, \cdots , \tag{2.42}$$

with the initial conditions: $\beta_0 = 1$ and $\beta_1 = \alpha_1$.
Once the coefficients β_j are computed, the properties of the *AR(2)* process follow:

---

[20] Also for *AR(p)* process, the polynomial equation $1 - \sum_{j=1}^{p} \alpha_j w^j = 0$, is satisfied at $w = 1$ if and only if $\sum_{j=1}^{p} \alpha_j = 1$. Thus, the presence of a unit root can be detected by verifying whether the sum of the autoregressive coefficients equals one.

1) $E(X_t) = E\left(\sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}\right) = \sum_{j=0}^{\infty} \beta_j E\varepsilon_{t-j} = 0$

2) $Var(X_t) = Var\left(\sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}\right) = \gamma_0 = \sum_{j=0}^{\infty} \beta_j Var(\varepsilon_{t-j}) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \beta_j^2$

3)
$$\begin{aligned} Cov(X_t, X_{t+k}) &= Cov\left(\sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}, \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+k-j}\right) \\ &= \gamma_k = \begin{cases} \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \beta_j \beta_{j+k} & k > 0 \\ \gamma_{-k} & k < 0 \end{cases} \end{aligned}$$

Given that the stationarity condition holds, the three properties above confirm that the *AR(2)* process is covariance stationary.
The autocorrelation function is:

$$\begin{aligned} Corr(X_t, X_{t+k}) &= \frac{\gamma_k}{\gamma_0} \\ &= \rho_k = \begin{cases} \dfrac{\sum_{j=0}^{\infty} \beta_j \beta_{j+k}}{\sum_{j=0}^{\infty} \beta_j^2} & k > 0 \\ \rho_{-k} & k < 0 \end{cases} \end{aligned} \tag{2.43}$$

The autocorrelation function approaches zero only as $k \to \infty$. A graphical example up to lag 10 is shown in *Figure 2.6*, where the autocorrelation rapidly decays as a result of the stationarity condition.

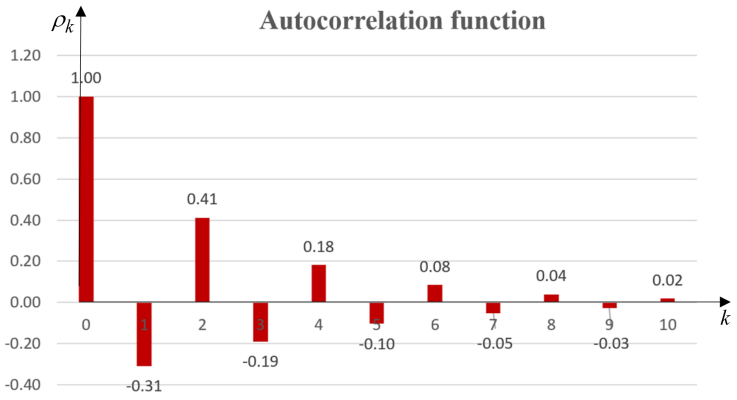


Figure 2.6: *Autocorrelation function for the process* $AR(2)$: $X_t = -0.2X_{t-1} + 0.35X_{t-2} + \varepsilon_t$

2.6 *AR(p)* Process as *MA* of Infinite Order

In this section, we consider the duality of the *AR(p)* process with the $MA(\infty)$ representation.
To this end, it is useful to refer to the well-known *fundamental theorem of algebra*, which states that *every non-constant single-variable polynomial with complex coefficients has at least one complex root.* This includes polynomials with real coefficients, since every real number is a complex number with zero imaginary part.
From this theorem it follows that any polynomial of degree p :

$$f(z) = \sum_{j=0}^{p} a_j z^j,$$

with $a_j \in \mathbb{C}$, the set of complex numbers and $a_p \neq 0$, has at least one complex root $z_0 \in \mathbb{C}$ such that $f(z_0) = 0$. Once a (complex) root is found, the polynomial can be factored into a product of first-degree linear terms[21]:

$$f(z) = \prod_{j=0}^{p-1} (z - z_j), \tag{2.44}$$

with $z_j \in \mathbb{C}$, where each $z_j \in \mathbb{C}$ is a root of the polynomial.
Taking multiplicities into account, any polynomial with complex coefficients has exactly p (possibly repeated) complex roots.
The stationarity condition applies to the characteristic polynomial:

$$\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p = 0\,, \tag{2.45}$$

which can be factored as:

$$(1 - z_1 L)(1 - z_2 L) \cdots (1 - z_p L) = 0\,, \tag{2.46}$$

where $z_i = \dfrac{1}{r_i}, \quad i = 1, 2, \ldots, p$, and r_i are the roots of the polynomial, generally belonging to $\mathbb{C}$.
As in the AR(2) case, stationarity requires that all roots satisfy $|r_i| > 1, \quad i = 1, \ldots, p$, that is, that *all the roots of the characteristic equation lie outside the unit circle.*

Example 2.6. Consider the *AR(4)* process:

$$X_t = 0.2X_{t-1} + 0.1X_{t-2} - 0.2X_{t-3} + 0.4X_{t-4} + \varepsilon_t$$

The associated characteristic equation is:

$$1 - 0.2w - 0.1w^2 + 0.2w^3 - 0.4w^4 = 0$$

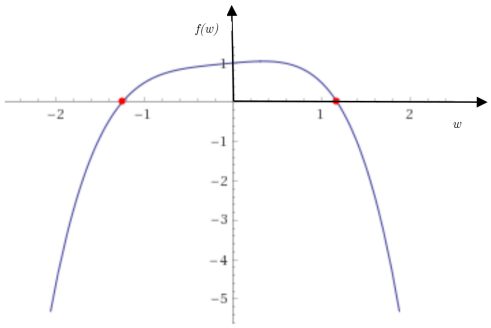


Figure 2.7: *Fourth degree polynomial:* $1 - 0.2w - 0.1w^2 + 0.2w^3 - 0.4w^4 = 0$

The curve described by this polynomial intersects the real axis in only two points (*Figure 2.7*).

Thus, there are two real roots and two complex roots.

The solutions are:

$$w_1 = -1.25295,\ w_2 = 1.17789,\ w_3 = -0.21247 - 1.28406\,i,\ w_4 = -0.21247 + 1.28406\,i$$

All the roots are outside the unit circle, so the stationarity condition of the *AR(4)* process is satisfied. Their position in the complex plane is shown in *Figure2.8.*

Without referring to the graph, for the real roots it is evident that $w_1 < -1,\ w_2 > 1$. For the complex roots, it is sufficient to verify whether the sum of the squares of the real and imaginary parts of the complex roots exceeds one, since $(-0.21247)^2 + (\pm 1.28406)^2 = 1.693954 > 1$.

The moment properties can be obtained by rewriting the *AR(p)* process in its dual form $MA(\infty)$ in a way completely analogous in what has been done for the *AR(2)* process.

However, in the next section, we present a simpler method to compute the autocovariance and autocorrelation functions.

[21] This result can be obtained using the *factor theorem* recursively.

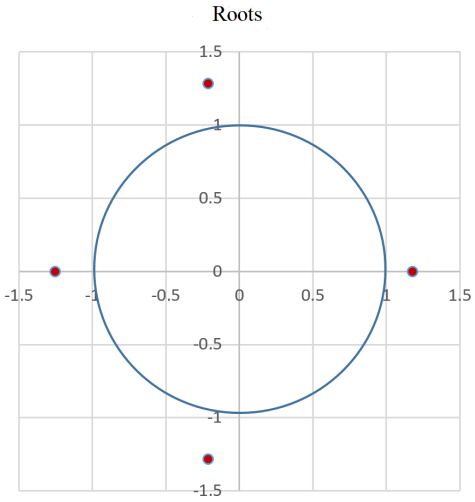


Figure 2.8: *Roots of the fourth-degree polynomial* $1 - 0.2w - 0.1w^2 + 0.2w^3 - 0.4w^4 = 0$

2.7 Recursive Formula for the Autocovariance and Autocorrelation of AR Processes

We can obtain a recursive formula to calculate the autocovariance of autoregressive processes, by multiplying the *AR(p)* process in equation (2.22) by X_{t-k} on both sides and taking the expectations:

$$E(X_t X_{t-k}) = \sum_{j=1}^{p} \alpha_j E(X_{t-j} X_{t-k}) + E(\varepsilon_t X_{t-k}), \quad k > 0\,.$$

From which the recursive formula for the autocovariance function follows:

$$\gamma_k = \sum_{j=1}^{p} \alpha_j \gamma_{k-j}, \quad k > 0. \tag{2.47}$$

The second term $E(\varepsilon_t X_{t-k}) = 0$, for $k > 0$. In fact, considering the dual *MA* form of X_{t-k}, we have:

$$E\left(\varepsilon_t \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-k-j}\right) = \sum_{j=0}^{\infty} \beta_j E(\varepsilon_t \varepsilon_{t-k-j}) = 0, \quad k > 0\,.$$

The expression (2.47) can be used as a recursive formula for computing the autocovariance function, provided that the initial condition given by the variance γ_0 is known.

A recursive formula for the autocorrelation function is readily obtained from (2.47) by dividing both sides of the equation by γ_0. We obtain:

$$\begin{aligned} \frac{\gamma_k}{\gamma_0} &= \sum_{j=1}^{p} \alpha_j \frac{\gamma_{k-j}}{\gamma_0}, \quad k > 0 \\ \rho_k &= \sum_{j=1}^{p} \alpha_j \rho_{k-j}, \quad k > 0 \end{aligned} \tag{2.48}$$

In this case, the initial condition becomes simply $\rho_0 = 1$.

Example 2.7. The recursive formula can be applied to the process in *Figure 2.6*. We have:

$$\rho_k = -0.2\rho_{k-1} + 0.35\rho_{k-2}\,.$$

From which:

$$\begin{aligned} &\rho_0 = 1 \\ &\rho_1 = -0.2\rho_0 + 0.35\rho_{-1} \rightarrow (1 - 0.35)\rho_1 = -0.2 \rightarrow \rho_1 = -0.30769 \\ &\rho_2 = -0.2\rho_1 + 0.35\rho_0 \rightarrow \rho_2 = -0.2 \times -0.30769 + 0.35 = 0.411538 \\ &\rho_3 = -0.2\rho_2 + 0.35\rho_1 \rightarrow \rho_3 = -0.2 \times 0.411538 - 0.35 \times 0.30769 = -0.19 \\ &\dots \end{aligned}$$

These values correspond to those plotted in *Figure 2.6*.

Since the autocorrelation function for an AR process of any finite order approaches zero only as $k \rightarrow \infty$, it follows that for the class of AR processes, the *autocorrelation function is not a reliable indicator of the process order.*
An alternative tool proposed for identifying the order is the *partial autocorrelation function.*

2.8 Partial Autocorrelation Function

A partial autocorrelation function can be built using the simple autocorrelation function to determine the values of the coefficients of the AR process. The order of the process is determined by the $k - th$ lag such that $\alpha_j = 0, \ \forall j > k$.
Not knowing the actual order of the process, we proceed by repeated steps verifying at each lag k which coefficients are zero.
For example, if the process were of order $p = 1$, using the recursive form for the autocorrelation function we would have $\rho_1 = \alpha\rho_0$ from which $\alpha = \rho_1$.
If we assume p = 2, then the recursive formula gives: $\rho_1 = \alpha_1\rho_0 + \alpha_2\rho_{-1}$, in such a case only one equation is not enough to determine the two unknown values of the coefficients.

However, we can use again the recursive form to write the second equation $\rho_2 = \alpha_1\rho_1+\alpha_2\rho_0$ and find the solution of the system[22]:

$$\begin{cases} \alpha_1 + \alpha_2\rho_{-1} = \rho_1 \\ \alpha_1\rho_1 + \alpha_2 = \rho_2 \end{cases}$$

Using the matrix notation:

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix},$$

or:

$$\mathbf{P}\boldsymbol{\alpha} = \boldsymbol{\rho} \tag{2.49}$$

The correlation matrix $\mathbf{P}$ is certainly positive definite for $|\rho_1| < 1$, therefore, it is invertible, so the solutions for coefficients are:

$$\boldsymbol{\alpha} = \mathbf{P}^{-1}\boldsymbol{\rho}\,, \tag{2.50}$$

i.e.:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2} \\ \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2} \end{bmatrix} \tag{2.51}$$

If the order of the process is $p = 2$, then $\alpha_2 \neq 0$. Instead, if the process were $p = 1$, then $\alpha_2 = 0$, in fact $\rho_2 - \rho_1^2 = 0$ it is true thus, for the first-order AR process $\rho_2 = \alpha^2 = \rho_1^2$.
We introduce the notation:

$$\begin{bmatrix} \alpha_{21} \\ \alpha_{22} \end{bmatrix} = \begin{bmatrix} \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2} \\ \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2} \end{bmatrix}, \tag{2.52}$$

where the first index indicates the choice of order of the vector $\boldsymbol{\alpha}$, and the second index denotes the element inside the vector.
More generally, for any k, the solution to equation (2.50) takes the following form:

$$\begin{bmatrix} \alpha_{k1} \\ \alpha_{k2} \\ \vdots \\ \alpha_{kk} \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_k \end{bmatrix} \tag{2.53}$$

[22] The equation system is known as the *Yule-Walker equations*,from the authors who have proposed this system in their works published in 1927 and 1931.

We focus on the behavior of the coefficient α_{kk}. In the sequence α_{11}, $\alpha_{22}, \cdots$ it is important to note the last value $\alpha_{kk} \neq 0$, beyond which all coefficients become zero. Hence, it is evident that the order of the AR process is $p = k$. Conversely, the values of the coefficients α_{kj}, $j < k$ are not informative about the order p, they may be either zero or non-zero.
The sequence $\{\alpha_{kk},\ k = 1, 2, \cdots\}$ *defines the partial autocorrelation function.*
The partial autocorrelation function serves as an indicator of the process order within the class of autoregressive processes.

An equivalent formulation is given by the correlation between X_t and X_{t-k}, after deducing from both of them the linear dependence with the intermediate random variables X_{t-1}, X_{t-2}, $\cdots$, X_{t-k+1}, that is:

$$\alpha_{kk} = Corr\left\{\left[X_t - E(X_t|X_{t-1}, \cdots, X_{t-k+1})\right]\left[X_{t-k} - E(X_{t-k}|X_{t-1}, \cdots, X_{t-k+1})\right]\right\} \tag{2.54}$$

The (2.54) motivates the partial autocorrelation name.

The Durbin-Levinson algorithm
The application of both the previous procedures (2.53) or (2.54) may be cumbersome or computationally intensive. For this reason, it is useful to refer to the recursive procedure proposed by Durbin and Levinson[23] which is based on the autocovariance function γ_k of the AR(p) process:

$$\alpha_{kk} = \left[\gamma_k - \sum_{j=1}^{k-1} \alpha_{k-1,j}\gamma_{k-j}\right] \nu_{k-1}^{-1}$$

$$\begin{bmatrix} \alpha_{k1} \\ \vdots \\ \alpha_{k,k-1} \end{bmatrix} = \begin{bmatrix} \alpha_{k-1,1} \\ \vdots \\ \alpha_{k-1,k-1} \end{bmatrix} - \alpha_{kk} \begin{bmatrix} \alpha_{k-1,k-1} \\ \vdots \\ \alpha_{k-1,1} \end{bmatrix}$$

$$\nu_k = \nu_{k-1}\left[1 - \alpha_{kk}^2\right],$$

with the initial conditions: $\alpha_{11} = \gamma_1/\gamma_0$, $\nu_0 = \gamma_0$.

Example 2.8. By applying the Durbin-Levinson algorithm to the *AR(2)* model of *Figure 2.6* , we obtain the graph of *Figure 2.9*. The last nonzero value occurs for the lag that identifies the order of the AR process.

Example 2.9. Considering the *AR(4)* process in *Example 2.6*, the autocorrelation function shown in *Figure 2.10* and the partial autocorrelation function shown in *Figure 2.11* are calculated.
When the lags increase, the simple autocorrelation function decreases with an oscillatory behavior due to the presence of real and complex negative roots. The partial autocorrelation function becomes zero after the fourth lag indicating the order of the AR process.

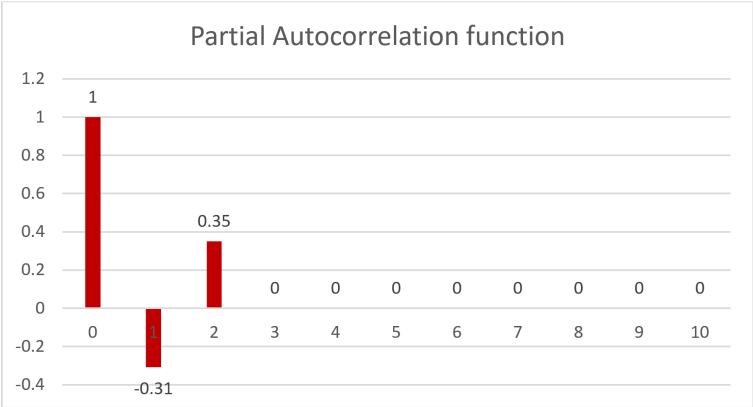


Figure 2.9: *Partial autocorrelation function of AR(2) process:* $X_t = -0.2X_{t-1} + 0.35X_{t-2} + \varepsilon_t$

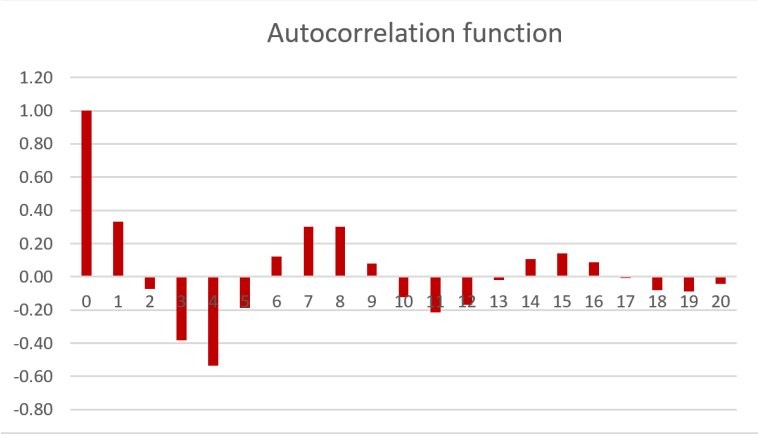


Figure 2.10: *Autocorrelation function of AR(4) process:* $X_t = 0.2X_{t-1} - 0.1X_{t-2} - 0.2X_{t-3} - 0.4X_{t-4} + \varepsilon_t$

Note the concordance between the signs of partial autocorrelation values and those of the process coefficients.

2.9 *MA(q)* as *AR* Process of Infinite Order

Given the $MA(q)$ process:

$$X_t = \beta(L)\varepsilon_t,$$

What has already been said regarding the duality between $AR(p)$ and $MA(\infty)$ can also be reformulated to show the duality between $MA(q)$ and $AR(\infty)$. it is sufficient to invert the polynomial $\beta(L)$ yielding:

[23] For the proof see: Brockwell and Davis (2016), § 2.5.3.

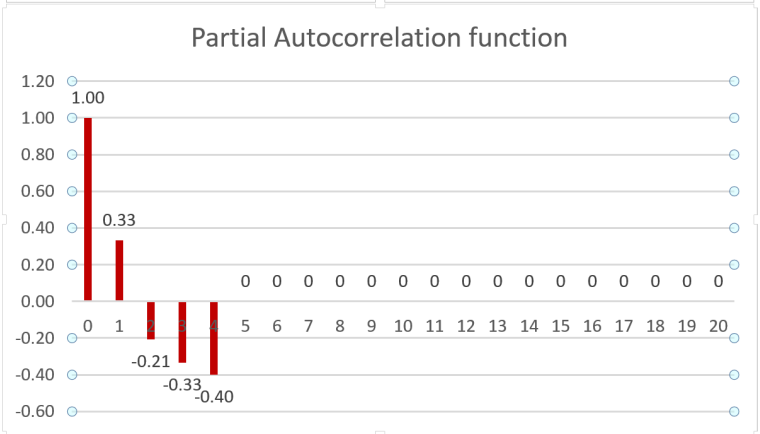


Figure 2.11: *Partial autocorrelation function of AR(4) process:* $X_t = 0.2X_{t-1} - 0.1X_{t-2} - 0.2X_{t-3} - 0.4X_{t-4} + \varepsilon_t$

$$\beta(L)^{-1} X_t = \varepsilon_t \,. \tag{2.55}$$

The inversion of the polynomial gives rise to an infinite sum of coefficients, denoted by α_j, $j = 1, 2, \cdots$. Therefore:

$$\beta(L)^{-1} = \sum_{j=0}^{\infty} \alpha_j L^j = \alpha(L) \tag{2.56}$$

In equation (2.56) it is assumed that the series $\sum_{j=0}^{\infty} \alpha_j$ is convergent; otherwise the inversion of the polynomial $\beta(L)$ is meaningless. The conditions for the convergence of this series are called the *invertibility conditions* and they are satisfied if *all the roots of the characteristic equation associated with the polynomial* $\beta(L)$*,i.e.:*

$$1 + \beta_1 w + \beta_2 w^2 + \cdots + \beta_q w^q = 0 \,,$$

lie outside the unit circle.
Note the analogy with the stationarity conditions for the *AR* processes. However, in the case of the *MA* processes the above condition cannot be regarded as a condition for stationarity, since *MA* processes are always stationary by construction, being linear combinations of a white noise process.
This can be illustrated by considering the simple *MA(1)*. The inversion of the polynomial $\beta(L)$ leads to:

$$\beta(L)^{-1} = \frac{1}{1+\beta L} = \frac{1}{1-(-\beta L)} = \sum_{j=0}^{\infty} (-\beta)^j L^j$$

which is a convergent series under the invertibility condition $|\beta| < 1$. However, stationarity is preserved even when $|\beta| > 1$ due to the alternating signs.

For example, consider the process: $X_t = 2\varepsilon_{t-1} + \varepsilon_t$. The inversion leads to:

$$X_t - 2X_{t-1} + 4X_{t-2} - 16X_{t-3} + 32X_{t-4} - \cdots = \varepsilon_t.$$

As the lags increase, the coefficients become very large in absolute value, alternating in sign. These alternating signs tend to compensate for each other, which helps preserve stationarity. However, in this case the process X_t cannot be represented as an $AR(\infty)$ process because the condition of invertibility is not satisfied: the inverse polynomial does not yield a convergent series.

Just a simple *AR(1)* process can be represented (under the stationarity condition) as a weighted sum of all past values of ε_t, similarly the simple *MA(1)* process can be explained (under the invertibility condition) as a weighted sum of all past values of X_t.

This duality extends to any *MA(q)* process, which can be expressed as an $AR(\infty)$. As a result of this duality, *MA* have a partial autocorrelation function that becomes zero only for $k \to \infty$.

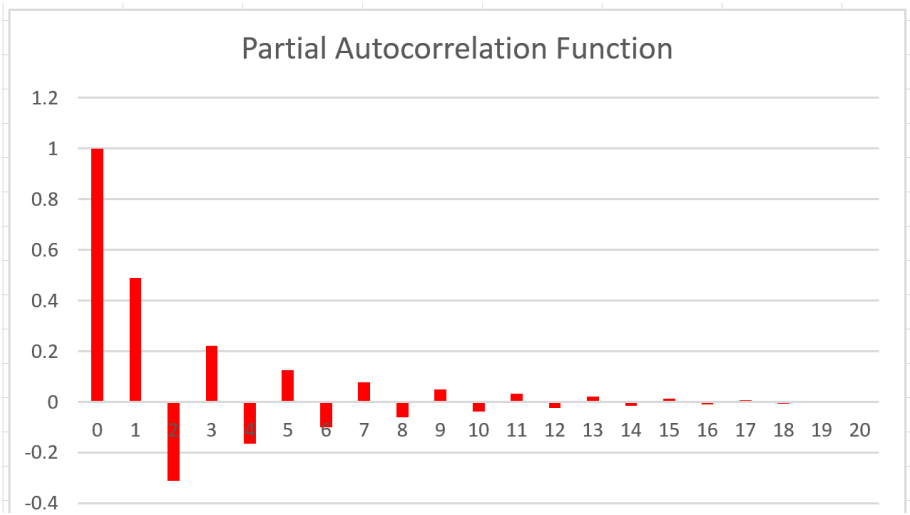


Figure 2.12: *Partial autocorrelation function of the MA(1) process:* $X_t = 0.8\varepsilon_{t-1} + \varepsilon_t$

Examples are given in *Figure 2.12* for an *MA(1)* with a positive coefficient, and in the *Figure 2.13* for an *MA(1)* process with a negative coefficient.

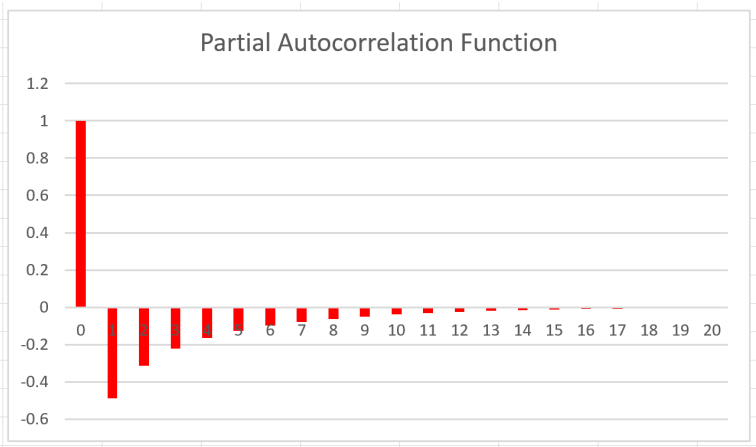


Figure 2.13: *Partial autocorrelation function of the MA(1) process:* $X_t = -0.8\varepsilon_{t-1} + \varepsilon_t$

2.10 *ARMA(p, q)* Process

A more general class of linear stochastic processes is the class of *ARMA(p, q)* processes, defined as:

$$\begin{cases} \alpha(L)X_t = \beta(L)\varepsilon_t \\ \alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p \\ \beta(L) = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q \end{cases}, \tag{2.57}$$

where $\alpha(L)$ is the lag polynomial of the autoregressive component and $\beta(L)$ is the lag polynomial of the moving average component. This process admits a dual representation in both autoregressive and moving average forms.
The autoregressive duality corresponds to the $AR(\infty)$ process obtained by inverting the polynomial $\beta(L)$. Under the invertibility conditions, we have:

$$\begin{gathered} \beta(L)^{-1}\alpha(L)X_t = \delta(L)X_t = \varepsilon_t \\ \begin{cases} \delta(L)X_t = \varepsilon_t \\ \delta(L) = \sum_{j=0}^{\infty} \delta_j L^j \end{cases} \end{gathered} . \tag{2.58}$$

Moving average duality corresponds to the $MA(\infty)$ process obtained, under stationarity conditions, by inverting the polynomial $\alpha(L)$:

$$\begin{gathered} X_t = \alpha(L)^{-1}\beta(L)\varepsilon_t = \phi(L)\varepsilon_t \\ \begin{cases} X_t = \phi(L)\varepsilon_t \\ \phi(L) = \sum_{j=0}^{\infty} \phi_j L^j \end{cases} \end{gathered} \tag{2.59}$$

This duality clearly implies that neither the simple autocorrelation function nor the partial autocorrelation function is a reliable indicator of the order of the *ARMA(p, q)* process.

In the econometrics literature, various methods have been proposed to determine the orders p and q, but none of them are fully satisfactory[24].
An example of a simulated *ARMA(p, q)* process generated with EViews software is provided in *Appendix 2.E.*

2.11 Time Series and Ergodicity

Definition 2.5. *(Time series)*
A time series is a finite portion of a realization of a stochastic process, i.e., a segment of a trajectory.

For economic variables, only one among the potentially infinite realizations is available. Consequently, inference about the future evolution of the process can only be made by assuming specific properties and intertemporal dependencies. Among these, a fundamental property is *stationarity*, which is a necessary condition for the validity of the *ergodic theorems* that underpin inference based on time series.

Ergodicity concerns the asymptotic behavior of time averages of functions of the process. In this section, we focus on the function defined as the time average of the variables:

$$\bar{x}_T = \frac{1}{T}\left(x_1 + x_2 + \cdots + x_T\right) \tag{2.60}$$

The asymptotic properties of this average can be framed in terms of the Weak Law of Large Numbers (WLLN). In stochastic process theory, some results concerning the WLLN are commonly referred to as ergodic propositions.

For convenience, we recall the definition of the WLLN.

Definition 2.6. ***(Weak Law of Large Numbers - WLLN)***
A sequence of random variables $\{x_t,\, t = 1, 2, \ldots\}$ *with finite first moments* μ_t *satisfies the weak law of large numbers*[25] *if:*

$$\operatorname*{plim}_{T\to\infty}\left(\frac{\omega_T}{T} - \frac{m_T}{T}\right) = 0$$

here, $\omega_T = \sum_{t=1}^{T} x_t$ *and* $m_T = \sum_{t=1}^{T} \mu_t$

We now present the first ergodic propositions concerning the mean in equation (2.60).

[24] A survey of these methods is provided in de Gooijer et al. (1985).

[25] If the deterministic limit $\lim_{T\to\infty} \frac{m_T}{T}$ exists and is finite, then the sample mean converges in probability to that limit of the theoretical mean.

Proposition 2.1. ***(Khinchin[26])***
Let $\{x_t,\, t = 1, 2, \ldots\}$ be a sequence of mutually independent and identically distributed (i.i.d.) random variables with finite mean μ. Then the sequence satisfies the weak law of large numbers:

$$plim\bar{x}_T = plim\frac{1}{T}\sum_{t=1}^{T} x_t = \mu.$$

Proof. See Dhrymes (1974), p. 101. □

Khinchin showed that for i.i.d. random variables, the existence of the mean μ is a sufficient condition to apply the law of large numbers.
To understand the relevance of ergodicity, consider the calculation of the mean in the case of non-stationary processes.
In such cases, the mean is not constant over time, and estimating it requires computing the sample mean at each time t. This, in turn, requires a set of N independent realizations, denoted by:

$$\begin{gathered}\{x_{1t},\ t = 1,\, 2,\, \ldots\} \\ \{x_{2t},\ t = 1,\, 2,\, \ldots\} \\ \vdots \\ \{x_{Nt},\ t = 1,\, 2,\, \ldots\}\end{gathered} \quad .$$

The sample mean at each time t is:

$$\hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N} x_{it}$$

In the case of stationary processes, the mean is constant over time, so it makes sense to compute the average of the sample means over time:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_t\,. \tag{2.61}$$

We refer to expression (2.61) as the *ensemble average*.
Even for stationary processes, ensemble average estimation requires N stochastically independent realizations. In practice, especially in econometrics, we typically have only one realization ($N = 1$), making the ensemble average impractical. Ergodic theorems allow us to replace the ensemble average with the time average (2.60) computed from the single available realization.
A stochastic process need not be an i.i.d. sequence of random variables. Hence, Khinchin's proposition imposes quite strict conditions. A second proposition, due to Chebyshev, relaxes the requirement of identical distributions.

[26] Aleksandr Yakovlevich Khinchin, 1894–1959.

Proposition 2.2. ***(Chebyshev[27])***
Let $\{x_t,\, t = 1, 2, \ldots\}$ *be a sequence of mutually independent random variables with finite means* μ_t *and variances* σ_t^2*. If:*

$$\lim_{T\to\infty} \frac{1}{T^2} \sum_{t=1}^{T} \sigma_t^2 = 0 \tag{2.62}$$

then the sequence satisfies the WLLN:

$$plim \bar{x}_T = \lim \mu_T = \lim \frac{1}{T} \sum_{t=1}^{T} \mu_t$$

Proof. See Feller (1968), p. 254. □

If the mean and variance are constant, this proposition reduces to Khinchin's. Note that Chebyshev's proposition requires the existence of the second moment, whereas Khinchin's does not. Furthermore, condition (2.62) is sufficient but not necessary.
Chebyshev's proposition applies to a broad class of non-stationary processes. In general, it is valid for processes that are *uniformly bounded*, i.e., those for which there exists a constant A such that $|x_t| < A$ for all t.

In most economic time series applications, we observe a single realization of a stochastic process, and the variables are not mutually independent. In this context, provided that the process is stationary, a more general result can be applied:

Proposition 2.3. ***(Birkhoff[28]–Khinchin)***
Let $\{x_t,\, t = 1, 2, \ldots\}$ *be a sequence of random variables with finite constant mean* μ*, finite constant variance* σ^2*, and autocovariance function* $\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]$*. If:*

$$\lim_{k\to\infty} \gamma_k = 0 \tag{2.63}$$

then the sequence satisfies the WLLN:

$$plim \bar{x}_T = \mu$$

Proof. See Gnedenko (1969), p. 337. □

Proposition 2.3 states a condition only sufficient for the ergodicity of the covariance stationary process. The following proposition states a necessary and sufficient condition.

[27] Pafnuty Lvovich Chebyshev, 1821–1894.
[28] George David Birkhoff (1884–1944), American mathematician.

Proposition 2.4. *(Ergodicity for Covariance Stationary Processes)*
Let $\{x_t, t = 1, 2, \ldots\}$ *be a covariance stationary process with autocovariance function defined as:*

$$\gamma_k = E(x_t - \mu)(x_{t+k} - \mu)\,,$$

then

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \gamma_k = 0\,, \tag{2.64}$$

if, and only if

$$\lim_{T\to\infty} E(\bar{x}_T - \mu)^2 = 0 \tag{2.65}$$

Proof. See Karlin and H. M. Taylor (1975), Theorem 5.1, p. 476. □

Remark 2.1. Beyond the proof of this proposition, note that by expanding the expected value, that is, the argument of the limit in (2.64), we obtain:

$$\begin{aligned} E(\bar{x}_T - \mu)^2 &= E\left(\frac{1}{T}\sum_{t=1}^{T} x_t - \mu\right)^2 \\ &= \frac{1}{T^2} E\left\{\sum_{t=1}^{T} (x_t - \mu)^2 + \sum_{t=1}^{T}\sum_{s\neq t=1}^{T} (x_t - \mu)(x_s - \mu)\right\} \\ &= \frac{1}{T^2}\left\{\sum_{t=1}^{T} \sigma_t^2 + \sum_{t=1}^{T}\sum_{s\neq t=1}^{T} \gamma_{t,s}\right\} \end{aligned}$$

The double summation involves the autocovariances of the process x_t. Due to the symmetry of the autocovariance matrix, can be written as $2\sum_{k=0}^{T-1} k\gamma_{T-k}$.
Therefore, we obtain:

$$E(\bar{x}_T - \mu)^2 = \frac{1}{T^2}\left\{\sum_{t=1}^{T} \sigma_t^2 + 2\sum_{k=0}^{T-1} k\gamma_{T-k}\right\}\,.$$

The condition stated in *Proposition 2.4* requires not only that $\lim \frac{1}{T^2}\sum_{t=1}^{T} \sigma_t^2 = 0$, which is the same condition as in the Chebyshev[29] *Proposition 2.2*, but also that

$$\lim \frac{2}{T^2} \sum_{k=0}^{T-1} k\gamma_{T-k} = 0\,.$$

[29] In the case of uncorrelated random variables, that is, when $\gamma_k = 0, \forall k \neq 0$, the *Proposition 2.4* reduces to Chebishev*Proposition 2.2*, considering that in the case of a covariance stationary process, condition (2.62) is automatically satisfied.

It can be shown that the assumption (2.64) ensures this second limit holds. Moreover, the sufficient condition in (2.63) implies (2.64).
In conclusion, *Proposition 2.4* establishes the equivalence between

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \gamma_k = 0 \text{ and } \lim_{T\to\infty} E(\bar{x}_T - \mu)^2 = 0\,,$$

which means convergence in mean square of the random variable $\bar{x}_T$ to its mean μ. This convergence does not require independence among the random variables of the process $\{x_t, t = 1, 2, \ldots\}$, but only the covariance stationarity. This requirement has implications for the behavior of the autocovariance function as shown by *Proposition 2.3*.

Ergodic Proposition 2.3 *is relevant because, in the case of a covariance stationary process, it allows replacing the ensemble average (2.61)—which is generally unfeasible—with the time average (2.60).*

The ergodic property can also be extended[30] to second moments of the process. In particular the variance and the autocovariance function.

Remark 2.2. It should be emphasized that ergodicity should not be confused with stationarity. An ergodic process is certainly stationary, but a stationary process is not necessarily ergodic. To clarify this aspect, consider the following examples.

Example 2.10. If we refer to the harmonic process in *Example 2.3*, in general, it is not ergodic[31]. In particular, we consider a single harmonic:

$$y_t = A\cos 2\pi ft + B\sin 2\pi ft, t = \cdots, -1, 0, 1, \cdots, \tag{2.66}$$

where f is the frequency (in cycles per unit of time), and A and B are random variables with $E(A) = E(B) = 0$, $Var(A) = Var(B) = \sigma^2$, and $E(AB) = 0$. Hence, $E(y_t) = 0$ and the autocovariance is $\gamma_k = Cov(y_t, y_{t+k}) = \sigma^2 \cos(2\pi fk)$.
A realization of the process y_t is determined by specific values for A and B.
Suppose $A \sim N(0, 4)$ and $B \sim N(0, 4)$. Two possible outcomes might be: $A_1 = -2.43445$, $B_1 = -5.49087$ and $A_2 = 1.654937$, $B_2 = -3.59948$.
Hence

$$y_{1t} = A_1\cos 2\pi ft + B_1\sin 2\pi ft, t = \cdots, -1, 0, 1, \cdots$$

and:

$$y_{2t} = A_2\cos 2\pi ft + B_2\sin 2\pi ft, t = \cdots, -1, 0, 1, \cdots$$

[30] This extension requires some assumptions regarding on fourth moments of the process. See Karlin and H. M. Taylor (1975), Theorem 5.2, p. 479.
[31] See Wold (1953), p.167.

with (chosen) frequency $f = 0.01$.
Figure 2.14 shows the two series over the interval $(-100, 100)$.

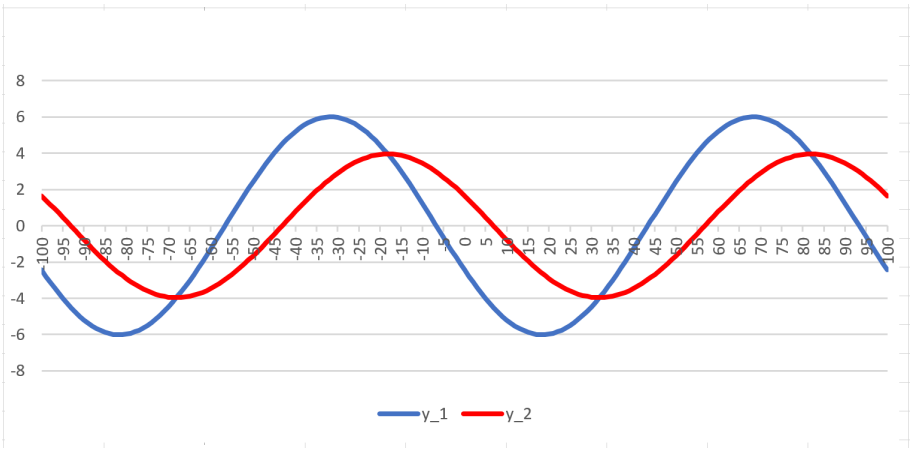


Figure 2.14: *Periodic waves with* $t \in (-100\,, 100)$, $f = 0.01$, $A_1 = -2.43445$, $B_1 = -5.49087$ *(blue curve), and* $A_2 = 1.654937$, $B_2 = -3.59948$ *(red curve)*

Let us discuss the ergodicity of the process y_t.
For the realization y_{1t}, the time average is:

$$\bar{y}_{1T} = \frac{1}{T}\sum_{t=1}^{T} y_{1t} = \frac{1}{T}\sum_{t=1}^{T} \left(A_1 \cos 2\pi f t + B_1 \sin 2\pi f t\right),$$

and $\lim\limits_{T\to\infty} \bar{y}_{1T} = 0$ since $\lim\limits_{T\to\infty} \frac{1}{T}\sum\limits_{t=0}^{T} \cos(2\pi f t) = 0$ and $\lim\limits_{T\to\infty} \frac{1}{T}\sum\limits_{t=0}^{T} \sin(2\pi f t) = 0$.
This might suggest ergodicity with respect to the mean, since the time average equals the ensemble mean $E(y_t) = 0$. This occurs because the waves oscillate symmetrically above and below the zero axis.
However, the process y_t is not ergodic with respect to the second moment (variance).
The ensemble variance limit is $Ey_t^2 = 4$, while the time variance limits are:

$$\lim_{T\to\infty} \mathrm{Var}_T(y_1) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} y_{1t}^2 = 18.0381$$
$$\lim_{T\to\infty} \mathrm{Var}_T(y_2) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} y_{2t}^2 = 7.84754$$

These results are explained by the following formula:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \left(A\cos(2\pi f t) + B\sin(2\pi f t)\right)^2 = \frac{1}{2}\left(A^2 + B^2\right), \tag{2.67}$$

in which the expansion:

$$\begin{aligned}&(A\cos(2\pi ft) + B\sin(2\pi ft))^2\\&= A^2\cos^2(2\pi ft) + 2AB\sin(2\pi ft)\cos(2\pi ft) + B^2\sin^2(2\pi ft)\\&= \frac{1}{2}(A^2 + B^2) + \frac{1}{2}(A^2 - B^2)\cos(4\pi ft) + AB\sin(4\pi ft).\end{aligned}$$

leads to:

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(A\cos(2\pi ft) + B\sin(2\pi ft))^2\\&= \frac{1}{2}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(2ABsin(4f\pi t) - B^2cos(4f\pi t) + A^2cos(4f\pi t) + B^2 + A^2)\right].\end{aligned}$$

All terms inside the summation tend to zero except the constant terms A^2 and B^2.

The variances of y_1 and y_2 clearly show that different realizations, corresponding to different values of the random coefficients A and B, yield different time variances.
To complete the evidence of the lack of ergodicity, consider the following simulation based on model (2.66). The simulation is performed with 5,000 observations in the interval $-2500 \le t < 2500$, assuming $A \sim N(0,4)$ and $B \sim N(0,4)$, stochastically independent, with $f = 0.01$.
Over the 5,000 observations, the time averages are $\bar{y}_1 = -2.4\times10^{-16}$ and $\bar{y}_2 = 5.59\times10^{-17}$. The time variances are $\mathrm{Var}(y_1) = 18.04174$ and $\mathrm{Var}(y_2) = 7.849106$.
The computed variances for y_1 and y_2 differ from the ensemble average and are consistent with the corresponding limits of the time variances.
The simulation also allows plotting the behavior of $\lim_{T\to\infty} Var_T(y_i)$ for the first 200 random realizations of y_i based on the random A_i and B_i (see the *Figure 2.15*), where $\mathrm{Var}_T(y_i)$ denotes the time variance computed along a single realization, that is,

$$\mathrm{Var}_T(y_i) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} y_{i,t}^2.$$

Figure 2.15 provides empirical evidence that the true variance of the stochastic process y_t can be obtained only through the calculation of the ensemble average:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\lim_{T\to\infty} Var_T(y_i),\ i = 1, 2, \cdots N$$

With $N = 5,000$, the sample analogue of ensemble average is 4.00355, similar to the theoretical ensemble average $Var(y_t) = 4$.
This example shows that the process is ergodic with respect to the mean, but not with respect to the second moment.

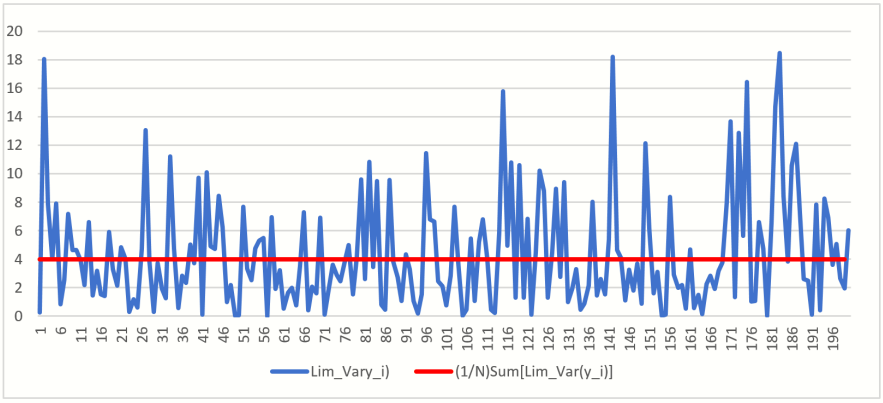


Figure 2.15: *Limit of time variance for the first 200 realizations. The red line is 4.00355, i.e., the sample ensemble average of all 5,000 time variance limits*

Remark 2.3. The ergodicity of the periodic curve (2.66) can be restored if the amplitude is constant and the randomness of the various realizations in the sample period arises from a particular choice of the random phase. To clarify this point, consider the following example.

Example 2.11. Consider the periodic process[32]:

$$y_t = R\sin(2\pi f t + \theta) \tag{2.68}$$

where R is the maximum amplitude and θ is a random phase. The value of y_t is random due to the stochastic nature of the phase θ.
The value of the sinusoidal wave at $t = t_1$ is:

$$y_1 = R\sin(2\pi f t_1 + \theta)\,, \tag{2.69}$$

where θ has a density defined on $(0\,, 2\pi)$.
If $f(t)$ is a realization of a periodic random process, ergodicity implies:

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f(t)dt = \int_{-\infty}^{\infty} xP_y(x)dx\,, \tag{2.70}$$

where $P_y(x)$ is the probability density function of the random amplitude y. For discrete realizations z_t, (2.70)becomes:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} z_t = \int_{-\infty}^{\infty} xP_y(x)dx\,. \tag{2.71}$$

[32] This is an equivalent form of $y_t = R\cos(2\pi f t - \theta)$, where R is the amplitude and θ is the phase, as defined in model (2.16).

Stationarity implies that $P_y(x)$ does not depend on time, so the ensemble average $\int_{-\infty}^{\infty} xP_y(x)dx$ remains constant. For the random value y_1 defined in (2.69), it can be shown[33] that its density is:

$$P_y(y_1) = \begin{cases} \dfrac{P_\theta(\theta_1 - 2\pi f_1 t_1) + P_\theta(\theta_2 - 2\pi f_1 t_1)}{\sqrt{R^2 - y_1^2}} & for \quad -R < y_1 < R \\ 0 & elsewhere \end{cases}, \tag{2.72}$$

where $P_\theta(\cdot)$ is the density function of the random phase θ, and θ_1 and θ_2 are the two solutions of equation (2.69) corresponding to a fixed value $y_1 \in (-R, R)$.
For example, let $f_1 = 0.01$, $t_1 = 50$, and $R = 4$, so:

$$\begin{aligned} 4\sin(2\pi f_1 t_1 + \theta) &= 4\sin(\pi + \theta) \\ &= 4\left[\cos(\pi)\sin(\theta) + \cos(\theta)\sin(\pi)\right] \\ &= -4\sin(\theta) \end{aligned} \tag{2.73}$$

Choosing $y_1 = 2$, the solutions of $-4\sin(\theta) = 2$ are $\theta_1 = 3.665$ and $\theta_2 = 5.760$, as shown in *Figure 2.16.*

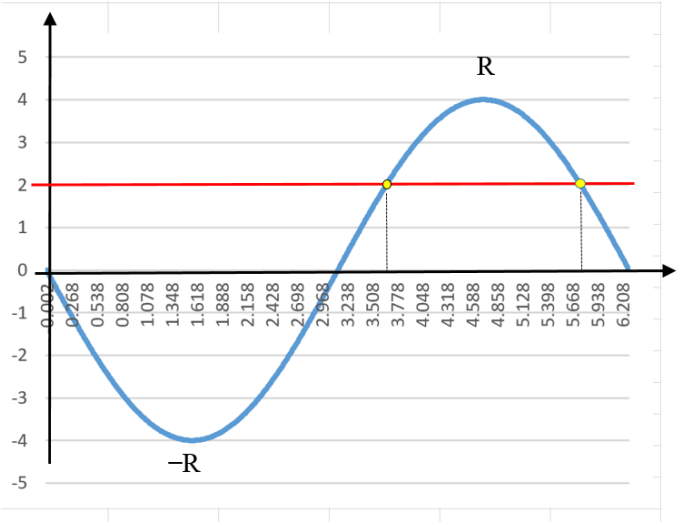


Figure 2.16: *Periodic wave defined by (2.73) with* $y_1 = 2$ *(red line)*

In general $P_y(y_1)$ in (2.72) depends on t_1, making the process generally non-stationary. To ensure stationarity, assume a rectangular (uniform) distribution for the phase:

$$P_\theta(x) = \begin{cases} \dfrac{1}{2\pi} & for\ 0 < x < 2\pi \\ 0 & elsewhere \end{cases}. \tag{2.74}$$

[33] See Lee (1960), p. 196.

Under this assumption, equation (2.72) simplifies to:

$$P_y(y_1) = \begin{cases} \dfrac{1}{2\pi} + \dfrac{1}{2\pi} = \dfrac{1}{\pi\sqrt{R^2 - y_1^2}} & for \ -R < y_1 < R \\ 0 & elsewhere \end{cases} . \tag{2.75}$$

Then one can verify that equation (2.71) holds for the variance of a realization.
The time average of the variance is given by:

$$\lim_{T\to\infty} Var_T(y_1) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} R^2\cos^2(2\pi ft) = \frac{R^2}{2} \tag{2.76}$$

This result follows from the same procedure used in equation (2.66).
The ensemble variance is:

$$\begin{aligned} \int_{-R}^{R} y_1^2 P_y(y_1)dy_1 &= \int_{-R}^{R} y_1^2 \frac{1}{\pi\sqrt{R^2 - y_1^2}}dy_1 \\ &= \frac{1}{2\pi}\left[R^2 \arcsin\left(\frac{y_1}{R}\right) - y_1\sqrt{R^2 - y_1^2}\right]_{-R}^{R} \\ &= \frac{1}{2\pi}\left[R^2\arcsin(1) - R^2\arcsin(-1)\right] \\ &= \frac{1}{2\pi}\left[R^2\frac{\pi}{2} + R^2\frac{\pi}{2}\right] = \frac{R^2}{2} \end{aligned} . \tag{2.77}$$

so time and ensemble variances coincide.
In conclusion, assuming a uniform phase distribution recovers ergodicity for the periodic process. The time average can be computed because the realization is periodic and determined once the initial angle is fixed. Meanwhile, the ensemble variance does not depend on that initial angle. This equivalence between time and ensemble averages is especially relevant in econometrics, where multiple realizations are typically unavailable. In contrast, other disciplines—such as statistical mechanics—may face measurement difficulties in describing how realizations evolve over time, for example when attempting to observe the instantaneous motion of particles.

Example 2.12. Hamilton provides another simple example[34] of a covariance stationary process that is not ergodic.
Suppose the mean μ_i, $i = 1, 2, \cdots$ for the ith realization $y_{i,t}$, $t = 1, 2, \cdots$, is drawn from a distribution $N(0, \sigma_\mu^2)$, so that:

$$y_{i,t} = \mu_i + \varepsilon_t, \ \ \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) . \tag{2.78}$$

[34] Hamilton (1994) pp. 46-47.

Here, the innovation process ε_t is independent of μ_i for all i and t.
This process is covariance stationary, since:

$$Ey_{i,t} = E\mu_i + E\varepsilon_t = 0, \ \forall i, t\,, \tag{2.79}$$

and its autocovariance function is given by:

$$\gamma_k = E(\mu_i + \varepsilon_t)(\mu_i + \varepsilon_{t+k}) = \begin{cases} \sigma_\mu^2 + \sigma_\varepsilon^2 & for\ k = 0 \\ \sigma_\mu^2 & for\ k \neq 0 \end{cases}$$

However, the process is not ergodic. The time average:

$$\frac{1}{T}\sum_{t=1}^{T} y_{i,t} = \frac{1}{T}\sum_{t=1}^{T}(\mu_i + \varepsilon_t) = \mu_i + \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t \tag{2.80}$$

converges to μ_i as $T \to \infty$, rather than to the ensemble mean (which is zero). Thus, the time average does not converge to the ensemble mean in (2.79).

2.11.1 Inference in Time Series

Let $\{x_t;\, t = 1,\, 2,\, \ldots,\, T\}$ be a time series, where T denotes the sample size. The following expression provides an estimate of the autocovariance function[35]:

$$\hat{\gamma}_k = \frac{1}{T}\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x}), \ \ k = 0,\, 1,\, \cdots \tag{2.81}$$

where $\bar{x} = \dfrac{1}{T}\sum_{t=1}^{T} x_t$ is the sample mean.
In principle, the summation in (2.81) should be divided by $T-k$ rather than T. However, for large samples the difference becomes negligible, whereas for small samples it may be non-trivial and should be taken into account.
The corresponding estimate of the autocorrelation function is given by the ratio:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}\,. \tag{2.82}$$

Sample autocorrelations can be tested for statistical significance. To this end, one needs the standard errors of the estimated autocorrelations. An approximation of the standard error is provided by a formula due to Bartlett (1946):

$$\hat{\sigma}(\hat{\rho}_k) \approx \frac{1}{T^{1/2}}\left(1 + 2\sum_{j=1}^{q}\hat{\rho}_j^2\right)^{1/2}, \ k > q\,, \tag{2.83}$$

[35] This is also the formula used by the EViews software.

where q denotes the lag beyond which the autocorrelations are assumed to be zero.
In many applications, it is of interest to test whether all autocorrelations are zero, as in the case of a white noise process. In such a case, the standard error formula simplifies to:

$$\hat{\sigma}(\hat{\rho}_k) \approx \frac{1}{T^{1/2}}, \quad k > 0\,. \tag{2.84}$$

Thus, a 95% two-sided confidence interval is given by:

$$\pm\frac{2}{\sqrt{T}}\,, \tag{2.85}$$

where the value 2 approximates the critical value 1.96.
This interval, $\left[-\frac{2}{\sqrt{T}}\,, \frac{2}{\sqrt{T}}\right]$, can also be interpreted as a significance test: values of $\hat{\rho}_k$ falling outside this range are statistically significant at the 5% level.
The same procedure applies to the partial autocorrelation function. It can be estimated using the Durbin-Levinson algorithm, replacing theoretical autocorrelations with their sample counterparts[36].
Alternatively, the partial autocorrelation can be computed from the following regressions, according to definition (2.54):

$$x_t = \hat{\alpha}_{k1}x_{t-1} + \hat{\alpha}_{k2}x_{t-2} + \cdots + \hat{\alpha}_{kk}x_{t-k} + \hat{\varepsilon}_t, \quad k = 1,\, 2,\, \cdots$$

Series iid
Sample: 1 200
Included observations: 200

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.099	-0.099	2.0027	0.157
		2	-0.010	-0.020	2.0236	0.364
		3	0.079	0.077	3.3077	0.347
		4	-0.077	-0.063	4.5310	0.339
		5	0.083	0.072	5.9516	0.311
		6	-0.004	0.003	5.9549	0.428
		7	0.042	0.056	6.3272	0.502
		8	-0.107	-0.117	8.7327	0.365
		9	0.021	0.014	8.8221	0.454
		10	0.088	0.077	10.481	0.399
		11	0.016	0.057	10.535	0.483
		12	0.001	-0.017	10.536	0.569
		13	-0.048	-0.044	11.036	0.608
		14	0.037	0.033	11.337	0.659
		15	-0.026	-0.019	11.487	0.717
		16	-0.022	-0.041	11.592	0.772
		17	-0.039	-0.063	11.934	0.804
		18	-0.083	-0.063	13.468	0.763
		19	0.010	-0.001	13.492	0.812
		20	-0.069	-0.074	14.547	0.802
		21	-0.032	-0.058	14.781	0.834
		22	0.108	0.114	17.429	0.739
		23	-0.023	0.027	17.553	0.781
		24	-0.021	-0.030	17.650	0.820

Figure 2.17: *Correlogram of i.i.d. series*

Example 2.13. A time series of 200 observations is generated through random draws from a normal distribution $N(0,\, 4)$.

[36] This is the method used, for example, by EViews.

Given its probabilistic characteristics, the series is i.i.d. The autocorrelation and partial autocorrelation functions are computed using EViews, producing the correlogram shown in *Figure 2.17*.
The 95% confidence interval for testing the significance of the autocorrelations is:

$$\pm 2 \Big/ \sqrt{200} = \pm 2/14.141 = \pm 0.141.$$

All values of the autocorrelation function (AC) and partial autocorrelation function (PACF) fall within these bounds. Interpreting the interval as critical values, we conclude that none of the estimated coefficients are statistically significant.
The last two columns of *Figure 2.17* report the *Q-Statistic* and associated p-values, as proposed by Ljung and Box. This statistic tests the joint null hypothesis that autocorrelations up to lag k are jointly zero. It is computed as:

$$Q_k = T(T+2) \sum_{j=1}^{k} \frac{\hat{\rho}_j^2}{T-j} \tag{2.86}$$

The results confirm that for no value of k are the autocorrelations statistically significant. For an i.i.d. process, the correlogram of the squared series—relevant for fourth-moment properties—should also show no significant autocorrelations. This is because the i.i.d. series is a realization of a strictly stationary process. Verification can be performed by squaring the observations (denoted iid^2) and examining the correlogram in *Figure 2.18*. The result aligns with expectations: no autocorrelation value is significant.

Example 2.14. As an exercise, we construct a process that qualifies as white noise but is not i.i.d.
Consider the following process: $x_t = \varepsilon_t \varepsilon_{t-1}$, where $\varepsilon_t \sim \text{NID}(0\,, \sigma^2)$.
For this process (see *Appendix 2.D*), we have:

a) $Ex_t = 0\ Var(x_t) = \sigma^4\ Corr(x_t,\, x_{t+k}) = 0, \quad for \quad k > 0$.
Hence the process is white noise.

b) $Ex_t^2 = \sigma^4$
$Var(x_t^2) = 8\sigma^8$
$$Cov(x_t^2, x_{t+k}^2) = \begin{cases} 2\sigma^8, & if \quad k = 1 \\ 0, & if \quad k > 1 \end{cases}$$
$$Corr(x_t^2, x_{t+k}^2) = \begin{cases} 0.25, & if \quad k = 1 \\ 0, & if \quad k > 1 \end{cases}$$
Thus, the process is covariance stationary but not i.i.d.

The results concerning the autocorrelation function are general and do not depend on the magnitude of the variance. *Figure 2.19* displays the correlogram of the series $wn_t = \varepsilon_t \varepsilon_{t-1}$.

As expected, the autocorrelation coefficients suggest a white noise behavior. To illustrate this, a sample of 10,000 observations is used.
Figure 2.20 shows the correlogram of the squared series wn_t^2. As predicted by theory, the autocorrelation at lag $k = 1$ is not zero, and the series is therefore not i.i.d. The estimated value $\hat{\rho}_1 = 0.24$ is statistically significant under the null hypothesis $\rho = 0$ ($t_stat = 24$, $t_prob = 0.00001$), but loses significance when the null hypothesis is set at $\rho = 0.25$ ($t_stat = -1$, $t_prob = 0.3173$).
All other estimated autocorrelation coefficients for $k \neq 1$ lie within the ± 0.02 band and are thus not statistically significant, in line with the theoretical findings.

Series iid^2
Sample: 1 200
Included observations: 200

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.014	0.014	0.0404	0.841
		2	-0.013	-0.013	0.0754	0.963
		3	0.015	0.016	0.1227	0.989
		4	-0.044	-0.045	0.5242	0.971
		5	-0.045	-0.043	0.9457	0.967
		6	-0.067	-0.067	1.8762	0.931
		7	-0.022	-0.020	1.9749	0.961
		8	0.093	0.091	3.7800	0.876
		9	-0.021	-0.026	3.8757	0.919
		10	0.043	0.039	4.2657	0.935
		11	0.148	0.138	8.9438	0.627
		12	-0.027	-0.028	9.1015	0.694
		13	0.125	0.137	12.468	0.490
		14	0.019	0.026	12.548	0.562
		15	-0.073	-0.056	13.713	0.547
		16	-0.036	-0.030	13.996	0.599
		17	-0.103	-0.081	16.356	0.499
		18	-0.095	-0.095	18.371	0.431
		19	-0.084	-0.106	19.938	0.398
		20	-0.016	-0.014	19.996	0.458
		21	0.092	0.038	21.893	0.406
		22	0.044	0.011	22.337	0.440
		23	-0.053	-0.064	22.989	0.461
		24	0.041	-0.011	23.368	0.498

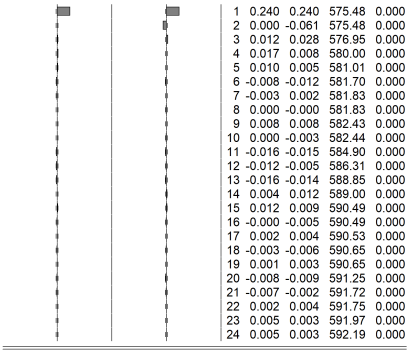
Figure 2.18: *Correlogram of i.i.d.^{2} series*

Serie wn=eps*eps(-1)
Sample: 1 10000
Included observations: 9999

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.012	-0.012	1.5511	0.213
		2	0.004	0.003	1.6779	0.432
		3	-0.009	-0.008	2.4013	0.493
		4	0.011	0.011	3.6218	0.460
		5	-0.009	-0.009	4.4199	0.491
		6	0.013	0.012	6.0723	0.415
		7	0.016	0.017	8.6863	0.276
		8	0.007	0.007	9.1730	0.328
		9	0.001	0.001	9.1829	0.421
		10	0.002	0.002	9.2119	0.512
		11	0.003	0.003	9.2989	0.594
		12	-0.002	-0.002	9.3236	0.675
		13	0.006	0.006	9.6774	0.720
		14	-0.003	-0.003	9.7624	0.779
		15	-0.010	-0.010	10.735	0.771
		16	-0.019	-0.019	14.378	0.571
		17	-0.002	-0.003	14.432	0.636
		18	0.000	0.000	14.433	0.700
		19	-0.000	-0.001	14.435	0.758
		20	0.008	0.008	15.123	0.769
		21	0.009	0.009	15.925	0.774
		22	0.000	0.001	15.927	0.819
		23	-0.011	-0.010	17.050	0.807
		24	-0.008	-0.008	17.648	0.820

Figure 2.19: *Correlogram of wn_t series*

Serie wn^2=wn*wn
Sample: 1 10000
Included observations: 9999



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.240	0.240	575.48	0.000
		2	0.000	-0.061	575.48	0.000
		3	0.012	0.028	576.95	0.000
		4	0.017	0.008	580.00	0.000
		5	0.010	0.005	581.01	0.000
		6	-0.008	-0.012	581.70	0.000
		7	-0.003	0.002	581.83	0.000
		8	0.000	-0.000	581.83	0.000
		9	0.008	0.008	582.43	0.000
		10	0.000	-0.003	582.44	0.000
		11	-0.016	-0.015	584.90	0.000
		12	-0.012	-0.005	586.31	0.000
		13	-0.016	-0.014	588.85	0.000
		14	0.004	0.012	589.00	0.000
		15	0.012	0.009	590.49	0.000
		16	-0.000	-0.005	590.49	0.000
		17	0.002	0.004	590.53	0.000
		18	-0.003	-0.006	590.65	0.000
		19	0.001	0.003	590.65	0.000
		20	-0.008	-0.009	591.25	0.000
		21	-0.007	-0.002	591.72	0.000
		22	0.002	0.004	591.75	0.000
		23	0.005	0.003	591.97	0.000
		24	0.005	0.003	592.19	0.000

Figure 2.20: *Correlogram of* $wn_t{}^2$ *series*

Appendix 2.A *(Lag operator L)*

The lag operator L acts on a stochastic process by shifting it one period back, i.e.:

$$LX_t = X_{t-1}$$

The lag operator is a linear operator.
Linearity follows directly from the definition. For the stochastic process $Z_t = X_t + Y_t$, we have:

$$LZ_t = Z_{t-1} = X_{t-1} + Y_{t-1} = LX_t + LY_t$$

Similarly, if $Z_t = aX_t$, then

$$LZ_t = Z_{t-1} = aX_{t-1} = aLX_t.$$

When applied to a constant, the operator has no effect[37], that is $Lk = k$.
The operator can be applied iteratively to the stochastic process. We have:

$$X_{t-2} = LX_{t-1} = L(LX_t) = L^2X_t\,.$$

In general $L^kX_t = X_{t-k}$.
Using these properties, any linear combination of stochastic processes can be expressed as follows:

$$Z_t = \sum_{j=1}^{n} a_j X_{t-j} = \left(\sum_{j=1}^{n} a_j L^j\right) X_t$$

Moreover, by analogy with polynomials in a real variables, we define:

$$a(L) = \sum_{j=1}^{n} a_j L^j$$

This notation is particularly useful, as it allows the usual algebraic operations on polynomials to be applied directly.

[37] The equality $Lk = k$ expresses that the lag operator leaves constant sequences invariant. Formally, L coincides with the identity operator only when restricted to the subspace of constant processes.

Appendix 2.B *(Roots of a second-degree equation and complex numbers)*

The roots of a second-degree equation:

$$ax^2 + bx + c = 0\,,$$

are obtained using the formula:

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

The nature of the roots depends on the discriminant $b^2 - 4ac$ in the above expression:

1) If $b^2 - 4ac > 0$, the roots are real and distinct.

2) If $b^2 - 4ac = 0$, the roots are real and equal.

3) If $b^2 - 4ac < 0$, the roots are complex and distinct.

In the third case, the solutions involve the quantity[38] $\sqrt{-1}$, and the roots can be written as:

$$\begin{aligned} r_1,\, r_2 &= \frac{1}{2a}\left(-b \pm \sqrt{-(4ac - b^2)}\right) = \frac{1}{2a}\left(-b \pm i\sqrt{4ac - b^2}\right) \\ &= f \pm i\, g \end{aligned} \tag{2.B1}$$

where:
$f = -\dfrac{b}{2a}$,
$g = \dfrac{\sqrt{4ac - b^2}}{2a}$,
and $i = \sqrt{-1}$, known as the *imaginary unit*[39].
If we define $r_1 = f + ig$, then $r_2 = f - ig$ is called the *complex conjugate* of r_1, denoted by $\bar{r}_1$.
There is a one-to-one correspondence between real numbers and points on the real line. Similarly, the geometric interpretation of complex numbers relies on a one-to-one correspondence with points in the Cartesian plane: any point (f, g) corresponds to the complex number $f + ig$ (see *Figure 2.B1*).

[38] We use the definition $i = \sqrt{-1}$, which satisfies $i^2 = -1$. The use of imaginary units appeared in the 16th century in the work of the Italian mathematician Niccolò Tartaglia (1499–1557) for solving cubic equations.

[39] The term "imaginary" was first used by Descartes in the 17th century. Mathematicians such as Abraham de Moivre (1667–1754) and Leonhard Euler (1707–1783) gave a theoretical foundation to complex numbers. Their acceptance in mathematics became complete with the geometric interpretation introduced in 1799 by Caspar Wessel (1745–1818).

Real numbers are a special case of complex numbers, represented as pointsof the form $(f, 0)$ on the Cartesian plane.

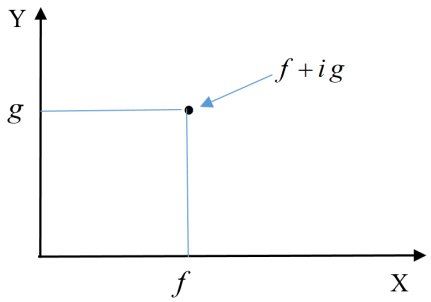


Figure 2.B1: *Representation of a complex number on the Cartesian plane*

The product of a complex number and its conjugate yields a real positive number:

$$(f + i\,g)(f - i\,g) = f^2 + i\,gf - i\,gf - i^2\,g^2 = f^2 + g^2$$

The equation of a circle centered at the origin of the Cartesian plane is:

$$x^2 + y^2 = r^2.$$

If the radius is one, the equation becomes:

$$x^2 + y^2 = 1$$

This justifies the expression *roots outside the unit circle*, meaning that

$$|r| = \sqrt{f^2 + g^2} > 1.$$

If $f > 0$, $g > 0$, and $f^2 + g^2 > 1$, the complex root and its conjugate lie outside the unit circle, as illustrated in *Figure 2.B2*.

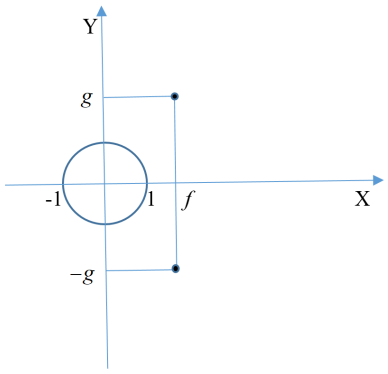


Figure 2.B2: *Representation of complex roots outside the unit circle*

Appendix 2.C

(Admissibility Region of AR(2) Coefficients for Stationarity)

To establish inequality (2.38), we first determine the admissible region of the coefficients α_1 and α_2 implied by the location of the characteristic roots z_i.

Here we work with the reciprocal polynomial

$$z^2 - \alpha_1 z - \alpha_2 = 0,$$

whose roots z_i are the inverses of the roots r_i of the autoregressive polynomial. Since stationarity requires $|r_i| > 1$, this condition is equivalent to

$$|z_i| < 1.$$

The condition $|z_i| < 1$ yields the extreme values of the coefficients α_i as shown in *Table 2.C1.*

z_1 \ z_2	1	-1
1	2	0
-1	0	-2

$\alpha_1 = z_1 + z_2$

z_1 \ z_2	1	-1
1	-1	1
-1	1	-1

$\alpha_2 = -z_1 z_2$

Table 2.C1: *Admissibility region of AR(2) coefficients: boundaries defined by the lines* $\alpha_2+\alpha_1 = 1$ *and* $\alpha_2 - \alpha_1 = 1$

Therefore, the admissibility range for α_1 is the open interval $(-2, 2)$, and for α_2 it is $(-1, 1)$. Within this region, the expressions $(1 \pm \frac{1}{2}\alpha_1)$ are positive, which allows the inequalities to be squared without reversing their direction.

$$\left|\tfrac{1}{2}\alpha_1 \pm \tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2}\right| < 1$$

$$-1 < \tfrac{1}{2}\alpha_1 \pm \tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2} < 1$$

$$-1-\tfrac{1}{2}\alpha_1 < \pm\tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2} < 1-\tfrac{1}{2}\alpha_1$$

$$-1-\tfrac{1}{2}\alpha_1 < -\tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2} < \tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2} < 1-\tfrac{1}{2}\alpha_1$$

$$1+\tfrac{1}{2}\alpha_1 > \tfrac{1}{2}\sqrt{\alpha_1^2+4\alpha_2} \quad \text{and} \quad \tfrac{1}{4}(\alpha_1^2+4\alpha_2) < \left(1-\tfrac{1}{2}\alpha_1\right)^2$$

$$\left(1+\tfrac{1}{2}\alpha_1\right)^2 > \tfrac{1}{4}(\alpha_1^2+4\alpha_2) \quad \text{and} \quad \tfrac{1}{4}\alpha_1^2+\alpha_2 < 1+\tfrac{1}{4}\alpha_1^2-\alpha_1$$

$$1+\tfrac{1}{4}\alpha_1^2+\alpha_1 > \tfrac{1}{4}\alpha_1^2+\alpha_2 \quad \text{and} \quad \alpha_2 < 1-\alpha_1$$

$$1+\alpha_1 > \alpha_2 \quad \text{and} \quad \alpha_2+\alpha_1 < 1$$

$$\alpha_2-\alpha_1 < 1 \quad \text{and} \quad \alpha_2+\alpha_1 < 1$$

The equations $\alpha_2+\alpha_1=1$ and $\alpha_2-\alpha_1=1$, viewed as linear functions of α_1, represent two straight lines with opposite slopes, as illustrated in *Figure 2.C1*.

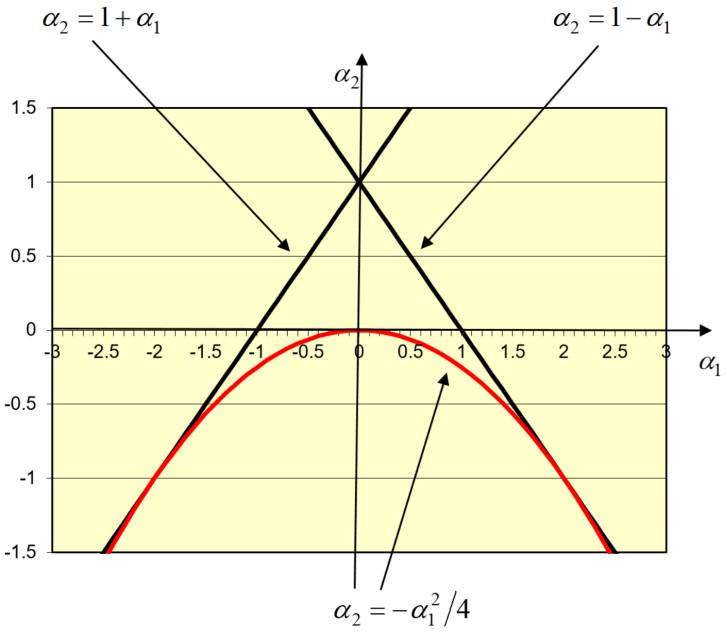


Figure 2.C1: *Lines delimiting the admissible region for stationarity*

The admissible region for stationarity is the triangular area bounded by these two lines and the line $\alpha_2 = -1$, with vertex at $(0, 1)$.
The red curve represents the parabola defined by the equation $\alpha_1^2 + 4\alpha_2 = 0$. It follows that when $\alpha_1^2 + 4\alpha_2 \geq 0$ the characteristic roots are real, whereas for $\alpha_1^2 + 4\alpha_2 < 0$ they are complex.

The lower boundary of the admissible region corresponds to the range of the coefficient α_2, which extends from -1 to 1.

Hence, the triangular region of stationarity is defined by:

$$\begin{cases} \alpha_2 + \alpha_1 < 1, \\ \alpha_2 - \alpha_1 < 1, \\ -1 < \alpha_2 < 1. \end{cases}$$

Appendix 2.D *(Solution of Example 2.14)*

We now provide the detailed solution of *Example 2.14*, following the steps required to compute the autocovariances. Consider the process $x_t = \varepsilon_t \varepsilon_{t-1}$, where $\varepsilon_t \sim NID(0, \sigma^2)$.

a) The process $\{x_t\}$ is white noise (uncorrelated), but it is not i.i.d.

We have

$$E(x_t) = E(\varepsilon_t)E(\varepsilon_{t-1}) = 0, \qquad Var(x_t) = E(x_t^2) = E(\varepsilon_t^2)E(\varepsilon_{t-1}^2) = \sigma^4.$$

For $k \geq 1$,

$$\gamma_k = Cov(x_t, x_{t+k}) = E(x_t x_{t+k}) = E\big(\varepsilon_t \varepsilon_{t-1} \varepsilon_{t+k} \varepsilon_{t+k-1}\big) = 0,$$

because at least one factor has zero mean and is independent of the remaining terms (the only possible overlap occurs at $k = 1$, where one still obtains $E(\varepsilon_{t-1}\varepsilon_{t+1}) = 0$). Therefore, all autocorrelations for $k \geq 1$ are zero.

b) The process $\{x_t^2\}$ is covariance stationary, but its autocorrelations are not all zero.

First,

$$E(x_t^2) = E(\varepsilon_t^2)E(\varepsilon_{t-1}^2) = \sigma^4.$$

Moreover,

$$\gamma_0 = Var(x_t^2) = E(x_t^4) - \big(E(x_t^2)\big)^2 = E(\varepsilon_t^4)E(\varepsilon_{t-1}^4) - \sigma^8.$$

Since $\varepsilon_t \sim N(0, \sigma^2)$, we have $E(\varepsilon_t^4) = 3\sigma^4$, hence

$$\gamma_0 = (3\sigma^4)(3\sigma^4) - \sigma^8 = 8\sigma^8.$$

(Equivalently, $(\varepsilon_t/\sigma)^2 \sim \chi_1^2$, so $Var(\varepsilon_t^2) = 2\sigma^4$.)

Now consider, for $k \geq 1$,

$$\gamma_k = Cov(x_t^2, x_{t+k}^2) = E(x_t^2 x_{t+k}^2) - E(x_t^2)E(x_{t+k}^2) = E(\varepsilon_t^2 \varepsilon_{t-1}^2 \varepsilon_{t+k}^2 \varepsilon_{t+k-1}^2) - \sigma^8.$$

For $k = 1$,

$$\gamma_1 = E(\varepsilon_t^4 \varepsilon_{t-1}^2 \varepsilon_{t+1}^2) - \sigma^8 = E(\varepsilon_t^4)\, E(\varepsilon_{t-1}^2)\, E(\varepsilon_{t+1}^2) - \sigma^8 = 3\sigma^4 \cdot \sigma^2 \cdot \sigma^2 - \sigma^8 = 2\sigma^8,$$

because ε_{t-1} and ε_{t+1} are independent. For all $k \geq 2$, the sets of indices $\{t, t-1\}$ and $\{t+k, t+k-1\}$ do not overlap, hence

$$E(x_t^2 x_{t+k}^2) = E(x_t^2)E(x_{t+k}^2) = \sigma^8, \qquad \Rightarrow \qquad \gamma_k = 0.$$

Therefore,

$$\rho_k = Corr(x_t^2, x_{t+k}^2) = \frac{\gamma_k}{\gamma_0} = \begin{cases} \dfrac{2\sigma^8}{8\sigma^8} = 0.25, & k = 1, \\ 0, & k \geq 2. \end{cases}$$

Hence, $\{x_t\}$ is covariance stationary and uncorrelated, yet not independent.

Appendix 2.E *(Simulation of ARMA processes in EViews)*

The following *AR(3)*, *MA(2)* and *ARMA(3,2)* processes are simulated:

$$\begin{aligned} X_{1t} &= 0.8X_{1t-1} + 0.15X_{1t-2} - 0.1X_{1t-3} + \varepsilon_t, \;\; \varepsilon_t \sim N(0,\, 0.25) \\ X_{2t} &= \varepsilon_t - 0.95\varepsilon_{t-1} + 0.3\varepsilon_{t-2} \\ X_{3t} &= 0.8X_{3t-1} + 0.15X_{3t-2} - 0.1X_{3t-3} + \varepsilon_t - 0.95\varepsilon_{t-1} + 0.3\varepsilon_{t-2} \end{aligned} \quad . \qquad \text{(2.E1)}$$

The first step is to define a new Workfile in EViews with an assigned name (e.g., `Simulation_ARMA`) and select the appropriate *Workfile structure type*. (*See Figure 2.E1.*)
For simplicity, no specific date structure with regular frequencies is defined; thus, EViews prompts only for the number of observations (*Data range*).

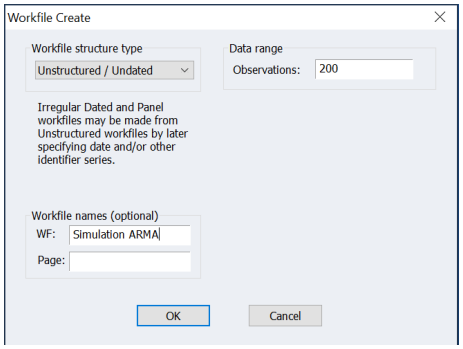


Figure 2.E1: *'Workfile create' dialog box*

Using the command line in EViews, we first generate 200 random numbers from $N(0,\, 0.25)$, saved as the series `eps`:

```
series eps = 0.5 * nrnd
```

Since `nrnd` generates $N(0, 1)$ draws, multiplying by 0.5 yields variance 0.25.
Next, we set the initial values of the three series to zero. Since the maximum lag is 3, initial values must be assigned to the first three observations.

```
smpl @first @first+2
series x1 = 0
series x2 = 0
series x3 = 0
```

Now we define the sample range for the remaining observations and generate the three series according to equations (2.E1):

```
smpl @first+3 @last
x1 = 0.8*x1(-1) + 0.15*x1(-2) - 0.1*x1(-3) + eps
x2 = eps - 0.95*eps(-1) + 0.3*eps(-2)
x3 = 0.8*x3(-1) + 0.15*x3(-2) - 0.1*x3(-3) + eps - 0.95*eps(-1) + 0.3*eps(-2)
```

To compute the correlogram for `x1`, double-click the series name and select the menu sequence: `Views > Show > Correlogram`
This will open the correlogram settings window shown in *Figure 2.E2*.

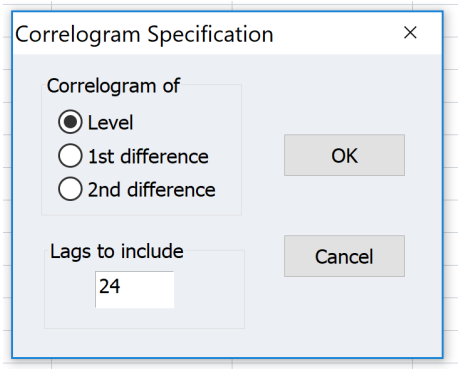
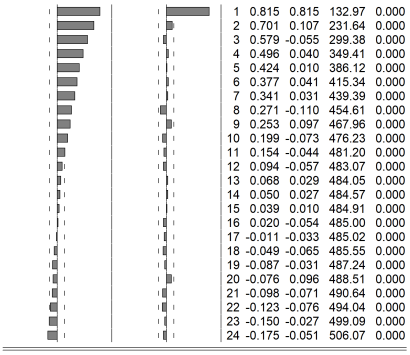


Figure 2.E2: *Correlogram specification window*

We select 24 lags for display. The resulting correlogram of the AR(3) process is shown in *Figure 2.E3* and is consistent with theoretical expectations.
Repeating the procedure for X_{2t} yields the correlogram in *Figure 2.E4*, which again reflects theoretical behavior.
For X_{3t}, the correlogram displays a mixed pattern characteristic of ARMA processes, without the clear cutoff typical of pure AR or MA models, as shown in *Figure 2.E5*.

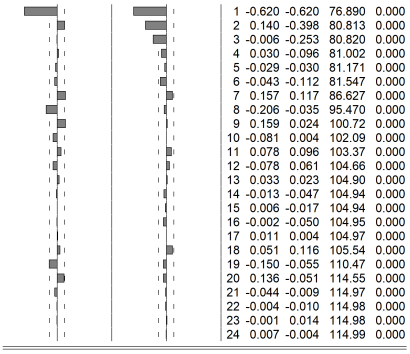
Series AR(3)
Sample: 4 200
Included observations: 197



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.815	0.815	132.97	0.000
		2	0.701	0.107	231.64	0.000
		3	0.579	-0.055	299.38	0.000
		4	0.496	0.040	349.41	0.000
		5	0.424	0.010	386.12	0.000
		6	0.377	0.041	415.34	0.000
		7	0.341	0.031	439.39	0.000
		8	0.271	-0.110	454.61	0.000
		9	0.253	0.097	467.96	0.000
		10	0.199	-0.073	476.23	0.000
		11	0.154	-0.044	481.20	0.000
		12	0.094	-0.057	483.07	0.000
		13	0.068	0.029	484.05	0.000
		14	0.050	0.027	484.57	0.000
		15	0.039	0.010	484.91	0.000
		16	0.020	-0.054	485.00	0.000
		17	-0.011	-0.033	485.02	0.000
		18	-0.049	-0.065	485.55	0.000
		19	-0.087	-0.031	487.24	0.000
		20	-0.076	0.096	488.51	0.000
		21	-0.098	-0.071	490.64	0.000
		22	-0.123	-0.076	494.04	0.000
		23	-0.150	-0.027	499.09	0.000
		24	-0.175	-0.051	506.07	0.000

Figure 2.E3: *Correlogram of the AR(3) process*

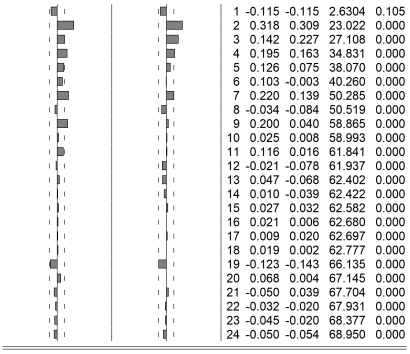
Series MA(2)
Sample: 4 200
Included observations: 197



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.620	-0.620	76.890	0.000
		2	0.140	-0.398	80.813	0.000
		3	-0.006	-0.253	80.820	0.000
		4	0.030	-0.096	81.002	0.000
		5	-0.029	-0.030	81.171	0.000
		6	-0.043	-0.112	81.547	0.000
		7	0.157	0.117	86.627	0.000
		8	-0.206	-0.035	95.470	0.000
		9	0.159	0.024	100.72	0.000
		10	-0.081	0.004	102.09	0.000
		11	0.078	0.096	103.37	0.000
		12	-0.078	0.061	104.66	0.000
		13	0.033	0.023	104.90	0.000
		14	-0.013	-0.047	104.94	0.000
		15	0.006	-0.017	104.94	0.000
		16	-0.002	-0.050	104.95	0.000
		17	0.011	0.004	104.97	0.000
		18	0.051	0.116	105.54	0.000
		19	-0.150	-0.055	110.47	0.000
		20	0.136	-0.051	114.55	0.000
		21	-0.044	-0.009	114.97	0.000
		22	-0.004	-0.010	114.98	0.000
		23	-0.001	0.014	114.98	0.000
		24	0.007	-0.004	114.99	0.000

Figure 2.E4: *Correlogram of the MA(2) process*

Series ARMA(3,2)
Sample: 4 200
Included observations: 197



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.115	-0.115	2.6304	0.105
		2	0.318	0.309	23.022	0.000
		3	0.142	0.227	27.108	0.000
		4	0.195	0.163	34.831	0.000
		5	0.126	0.075	38.070	0.000
		6	0.103	-0.003	40.260	0.000
		7	0.220	0.139	50.285	0.000
		8	-0.034	-0.084	50.519	0.000
		9	0.200	0.040	58.865	0.000
		10	0.025	0.008	58.993	0.000
		11	0.116	0.016	61.841	0.000
		12	-0.021	-0.078	61.937	0.000
		13	0.047	-0.068	62.402	0.000
		14	0.010	-0.039	62.422	0.000
		15	0.027	0.032	62.582	0.000
		16	0.021	0.006	62.680	0.000
		17	0.009	0.020	62.697	0.000
		18	0.019	0.002	62.777	0.000
		19	-0.123	-0.143	66.135	0.000
		20	0.068	0.004	67.145	0.000
		21	-0.050	0.039	67.704	0.000
		22	-0.032	-0.020	67.931	0.000
		23	-0.045	-0.020	68.377	0.000
		24	-0.050	-0.054	68.950	0.000

Figure 2.E5: *Correlogram of the ARMA(3,2) process*

Appendix 2.F *(Coefficient-Based Necessary and Sufficient Conditions for Covariance Stationarity of an AR(p) Process)*

The following *Proposition 2.F1* provides a simple and quick diagnostic rule for detecting non-stationarity in an autoregressive model. If the condition stated below is violated, the $AR(p)$ process is certainly not covariance stationary. However, if the condition is satisfied, stationarity is not guaranteed, since the condition is necessary but not sufficient.
Proposition 2.F2, on the other hand, introduces an inequality rule that defines a conservative stationarity region. Whenever it holds, stationarity is ensured; however, stationarity may still hold even when the inequality is violated.

Proposition 2.F1. *Let the autoregressive process $AR(p)$ be defined as*

$$\alpha(L)X_t = \varepsilon_t, \qquad \alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p.$$

A necessary condition[40] *for covariance stationarity is*

$$\sum_{j=1}^{p} \alpha_j < 1.$$

Proof. Covariance stationarity requires that all roots of the characteristic equation

$$\alpha(L) = 0$$

lie outside the unit circle, i.e. $|L_i| > 1$ for all i.
Consider the polynomial $\alpha(L)$ evaluated at $L = 0$ and $L = 1$. We have

$$\alpha(0) = 1 > 0,$$

and

$$\alpha(1) = 1 - \sum_{j=1}^{p} \alpha_j.$$

Suppose that

$$\sum_{j=1}^{p} \alpha_j > 1.$$

Then

$$\alpha(1) < 0.$$

[40] This condition is a direct corollary of the root-based stationarity criterion.

Since $\alpha(L)$ is a continuous function of L, and since

$$\alpha(0) > 0 \quad \text{and} \quad \alpha(1) < 0,$$

the *Intermediate Value Theorem* implies the existence of at least one root

$$L^* \in (0, 1)$$

such that $\alpha(L^*) = 0$.
But then $|L^*| < 1$, meaning that the characteristic equation has a root inside the unit circle. This contradicts the stationarity requirement.
Therefore, for stationarity it is necessary that

$$\alpha(1) > 0,$$

that is,

$$1 - \sum_{j=1}^{p} \alpha_j > 0 \quad \Longleftrightarrow \quad \sum_{j=1}^{p} \alpha_j < 1.$$

If

$$\sum_{j=1}^{p} \alpha_j = 1,$$

then $\alpha(1) = 0$, so $L = 1$ is a root and the process has a unit root, hence it is not stationary. □

Proposition 2.F2. *Let the autoregressive process $AR(p)$ be defined as*

$$\alpha(L)X_t = \varepsilon_t, \qquad \alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p,$$

where $\{\varepsilon_t\}$ is white noise with finite variance. If

$$\sum_{j=1}^{p} |\alpha_j| < 1,$$

then the process is covariance stationary.

Proof. Covariance stationarity of an $AR(p)$ process holds if and only if all roots of the characteristic equation $\alpha(L) = 0$ lie outside the unit circle, i.e. $\alpha(L) \neq 0$ for all L such that $|L| \leq 1$.
Let $|L| \leq 1$. By the triangle inequality,

$$\left| \sum_{j=1}^{p} \alpha_j L^j \right| \leq \sum_{j=1}^{p} |\alpha_j| \, |L|^j \leq \sum_{j=1}^{p} |\alpha_j|.$$

If $\sum_{j=1}^{p} |\alpha_j| < 1$, then

$$\left| \sum_{j=1}^{p} \alpha_j L^j \right| < 1.$$

Therefore,

$$\alpha(L) = 1 - \sum_{j=1}^{p} \alpha_j L^j \neq 0 \qquad \text{for all } |L| \leq 1,$$

because the equality $\alpha(L) = 0$ would require $\left| \sum_{j=1}^{p} \alpha_j L^j \right| = 1$, which is impossible under the above bound.

Hence $\alpha(L)$ has no zeros in the closed unit disk, and consequently all roots of $\alpha(L) = 0$ satisfy $|L_i| > 1$. It follows that the $AR(p)$ process admits a causal $MA(\infty)$ representation with absolutely summable coefficients and is therefore covariance stationary.[41]

□

[41] For the relation between root conditions, causality, and the $MA(\infty)$ representation, see Brockwell and Davis (2016), Chapter 3.

3 Wold Decomposition Theorem and General Linear Stationary Processes

The *ARMA(p,q)* process represents a general class of linear stationary stochastic processes. A natural question is whether this class exhaustively describes all covariance stationary stochastic processes. The Wold Decomposition Theorem provides an answer.

Theorem 3.1. *(Wold Decomposition Theorem)*
Let y_t be an arbitrary discrete covariance stationary process with finite variance. Then there exists a three-component stationary process $\{z_t, x_t, \varepsilon_t\}$ satisfying the following properties:

A) $y_t = z_t + x_t$,

B) z_t *and* x_t *are uncorrelated,*

C) z_t *is a* singular process,

D) ε_t *is a white noise process (uncorrelated),*

E) $x_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}, \beta_0 = 1$, *where the coefficients represent real numbers such that* $\sum_{j=0}^{\infty} \beta_j^2 < \infty$.

Proof. See Theorem 7, p. 89 in Wold (1953)[42]. □

Remark 3.1. The process x_t described in E) is called a *regular process*. The process ε_t in D) is a white noise process, conventionally with $E\varepsilon_t = 0$.

Remark 3.2. The theorem is relevant because it identifies a broader class of linear processes than those defined by *AR(p)*, *MA(q)*, or *ARMA(p,q)* models. This broader class is known as *General Linear Stationary Processes (GLSP)* and is represented by the process x_t in part E) of the theorem.

Remark 3.3. The GLSP class is more comprehensive than the *ARMA(p,q)* class because every stationary *ARMA(p,q)* process has an equivalent $MA(\infty)$ representation. However, the converse is not true: an $MA(\infty)$ process does not necessarily have a finite representation as an *AR(p)* or *ARMA(p,q)* process. Thus, the parameters of a GLSP may not correspond to any finite set of parameters $\{\alpha_j, j = 1, 2, \ldots, p\}$ as defined within the classes of *AR(p)* or *ARMA(p,q)* processes.

[42] Herman Wold proved this result in his Ph.D. thesis, published in 1938. The importance of this theorem is clarified in the following remarks.

Remark 3.4. The only condition required for y_t in the theorem is covariance stationarity. Therefore, even non-linear stationary processes can be represented as a linear combination of white noise. However, when the original process is non-linear and defined by a small number of parameters, its linear representation will generally involve an infinite number. In practice, such processes can often be well approximated by a truncated *MA(q)* process. For reasons of parsimony, it may be preferable to specify the model using the *AR(p)* or *ARMA(p,q)* classes. As a result, the *ARMA(p,q)* class is sufficiently broad to describe many stationary time series.

Remark 3.5. The decomposition theorem can also be interpreted in light of prediction theory for stochastic processes[43]. Let us consider the process:

$$x_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}. \tag{3.1}$$

For simplicity, assume this corresponds exactly to an *AR(1)* process:

$$x_t = \alpha x_{t-1} + \varepsilon_t. \tag{3.2}$$

The *predictor* of x_t is the conditional expectation based on past values of the process (in this case, just x_{t-1}):

$$x_t^e = E(x_t \mid x_{t-1}) = \alpha x_{t-1}. \tag{3.3}$$

The *prediction error* is:

$$x_t - x_t^e = \varepsilon_t, \tag{3.4}$$

with prediction error variance $\sigma^2 \neq 0$.
Now consider the general case of an $AR(\infty)$ process, which exactly corresponds to (3.1):

$$x_t = \sum_{j=1}^{\infty} \alpha_j x_{t-j} + \varepsilon_t. \tag{3.5}$$

The *predictor* is:

$$x_t^e = E(x_t \mid x_{t-1}, x_{t-2}, \ldots) = \sum_{j=1}^{\infty} \alpha_j x_{t-j}, \tag{3.6}$$

and again, the *prediction error* variance is $\sigma^2 \neq 0$.
In summary, *for the regular process, it is not possible to obtain an exact predictor in the sense that the prediction error variance is zero.*
By contrast, the singular process is characterized by a zero prediction error variance. The following remark clarifies this distinction.

[43] See also §6.5.

Remark 3.6. The class of singular processes is well defined and analyzed by Wold[44]. The discussion of singular processes goes beyond the scope of this lecture note; however, in §6.3.2, the singular process is defined as a *perfectly predictable process*, in the sense that *the variance of the forecast error tends to zero in the limit.*
We clarify the concept of singularity through the following example.

Example 3.1. Consider the class of periodic stochastic processes, already defined in *Example 2.3*:

$$z_t = \sum_{j=1}^{q} (A_j \cos \omega_j t + B_j \sin \omega_j t), \qquad t = \ldots, -1, 0, 1, \ldots \tag{3.7}$$

Under the assumptions on the amplitudes $A_1, \ldots, A_q, B_1, \ldots, B_q$ given in equations (2.10) to (2.13), we know that $E[z_t] = 0$ and:

$$\gamma_k = E z_t z_{t+k} = \sum_{j=1}^{q} \sigma_j^2 \cos \omega_j k$$

For simplicity, consider a single harmonic:

$$z_t = A \cos \omega t + B \sin \omega t, \qquad t = \ldots, -1, 0, 1, \ldots \tag{3.8}$$

The predictor is:

$$z_t^e = E(z_t | z_{t-1}, z_{t-2}, \ldots, z_{t-p}) \tag{3.9}$$

Determining the appropriate lag p in (3.9) is crucial to specifying the best predictor for the process in (3.8).
Step 1: Consider an $AR(1)$ process:

$$z_t = \phi z_{t-1} + \varepsilon_t \tag{3.10}$$

Note:

$$z_0 = A \cos 0 + B \sin 0 = A \tag{3.11}$$

$$z_1 = A \cos \omega + B \sin \omega \tag{3.12}$$

The regression coefficient ϕ is:

$$\phi = \frac{\mathrm{Cov}(z_t, z_{t-1})}{\mathrm{Var}(z_{t-1})} = \frac{\gamma_1}{\gamma_0} = \frac{\sigma^2 \cos \omega}{\sigma^2} = \cos \omega \tag{3.13}$$

[44] See Wold (1953), §14, p. 41.

Hence, the predicted values are:

$$\begin{aligned} z_1^e &= A\cos\omega \\ z_2^e &= A\cos^2\omega \\ &\vdots \\ z_t^e &= A\cos^t\omega \end{aligned} \tag{3.14}$$

and the prediction error becomes:

$$\begin{aligned} z_1 - z_1^e &= A\cos\omega + B\sin\omega - A\cos\omega = B\sin\omega \\ z_2 - z_2^e &= A\cos 2\omega + B\sin 2\omega - A\cos^2\omega = A(\cos 2\omega - \cos^2\omega) + B\sin 2\omega \\ &\vdots \\ z_t - z_t^e &= A\cos(\omega t) + B\sin(\omega t) - A\cos^t\omega = A\big(\cos(\omega t) - \cos^t\omega\big) + B\sin(\omega t) \end{aligned} \quad . \tag{3.15}$$

In conclusion, the prediction error contains a periodic function with frequency ω.

Step 2: Now consider the AR(2) model:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \varepsilon_t \tag{3.16}$$

The parameters ϕ_1 and ϕ_2 are determined from the recursive relationship (2.48) for the autocorrelation function. The application of this formula leads to the system:

$$\begin{cases} \rho_1 = \phi_1\rho_0 + \phi_2\rho_{-1} \\ \rho_2 = \phi_1\rho_1 + \phi_2\rho_0 \end{cases} \rightarrow \begin{cases} \rho_1 = \phi_1 + \phi_2\rho_1 \\ \rho_2 = \phi_1\rho_1 + \phi_2 \end{cases} \tag{3.17}$$

Substitute the values for the harmonic process (3.8), i.e., $\rho_1 = \cos\omega$ and $\rho_2 = \cos 2\omega$, the system becomes:

$$\begin{cases} \cos\omega = \phi_1 + \phi_2\cos\omega \\ \cos 2\omega = \phi_1\cos\omega + \phi_2 \end{cases} \quad . \tag{3.18}$$

The solution is given by:

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} 1 & \cos\omega \\ \cos\omega & 1 \end{bmatrix}^{-1} \begin{bmatrix} \cos\omega \\ \cos 2\omega \end{bmatrix}, \tag{3.19}$$

that is:

$$\begin{aligned}
\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} &= \frac{1}{1-\cos^2\omega} \begin{bmatrix} 1 & -\cos\omega \\ -\cos\omega & 1 \end{bmatrix} \begin{bmatrix} \cos\omega \\ \cos 2\omega \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\cos\omega(1-\cos 2\omega)}{1-\cos^2\omega} \\ \dfrac{-\cos^2\omega+\cos 2\omega}{1-\cos^2\omega} \end{bmatrix} = \begin{bmatrix} \dfrac{2\cos\omega(1-\cos 2\omega)}{1-\cos 2\omega} \\ \dfrac{\cos^2\omega \;-\; 1}{1-\cos^2\omega} \end{bmatrix} \qquad (3.20) \\
&= \begin{bmatrix} 2\cos\omega \\ -1 \end{bmatrix} .
\end{aligned}$$

This result is obtained considering the *trigonometric power-reducing formulas*:

$$\cos^2\omega = \frac{1+\cos 2\omega}{2}.$$

From (3.20) the predictor becomes:

$$z_t^e = 2z_{t-1}\cos\omega - z_{t-2}\,. \qquad (3.21)$$

Considering the first two values:

$$\begin{aligned} z_0 &= A \\ z_1 &= A\cos\omega + B\sin\omega \end{aligned} \quad , \qquad (3.22)$$

we obtain:

$$\begin{aligned}
z_2^e &= (2\cos\omega)(A\cos\omega + B\sin\omega) - A \\
&= A\cos 2\omega + B\sin 2\omega \\
z_3^e &= (2\cos\omega)(A\cos 2\omega + B\sin 2\omega) - (A\cos\omega + B\sin\omega) \\
&= A\cos 3\omega + B\sin 3\omega \\
&\vdots \\
z_t^e &= (2\cos\omega)\left[A\cos\omega(t-1) + B\sin\omega(t-1)\right] - \left[A\cos\omega(t-2) + B\sin\omega(t-2)\right] \\
&= A\cos\omega t + B\sin\omega t \qquad (3.23) \\
&= z_t
\end{aligned}$$

The result in (3.23) shows that the process is perfectly predictable using an AR(2) model, in the sense that $Var(z_t - z_t^e) = 0$.

Therefore, the process (3.8) is both a stationary process and a singular process.

This result also holds for a specific realization of the stochastic process, as shown by the following numerical example.

Example 3.2. Let $\omega = \frac{2}{3}\pi$ be the angular frequency of a harmonic process, and suppose that a particular realization has amplitudes $A = 3$ and $B = 2$. Then, from equation (3.8), the realization is given by:

$$z_t = 3\cos\left(\frac{2}{3}\pi t\right) + 2\sin\left(\frac{2}{3}\pi t\right) \tag{3.24}$$

In *Figure 3.1*, we highlight the values: $(0, 3)$, $(1, 0.23205)$, $(2, -3.23205)$, $(3, 3)$. These same values can be calculated using the predictor with parameters:

$$\phi_1 = 2\cos\omega = 2\cos\left(\frac{2}{3}\pi\right) = -1,$$
$$\phi_2 = -1.$$

with the predictor defined as:

$$\hat{z}_t = \phi_1\hat{z}_{t-1} + \phi_2\hat{z}_{t-2} = -\hat{z}_{t-1} - \hat{z}_{t-2}.$$

Starting with the initial values:

$$\hat{z}_0 = A = 3,$$
$$\hat{z}_1 = A\cos\omega + B\sin\omega = 3\cos\left(\frac{2}{3}\pi\right) + 2\sin\left(\frac{2}{3}\pi\right) = 0.23205,$$

we obtain:

$$\hat{z}_2 = -\hat{z}_1 - \hat{z}_0 = -0.23205 - 3 = -3.23205,$$
$$\hat{z}_3 = -\hat{z}_2 - \hat{z}_1 = -(-3.23205) - 0.23205 = 3.$$

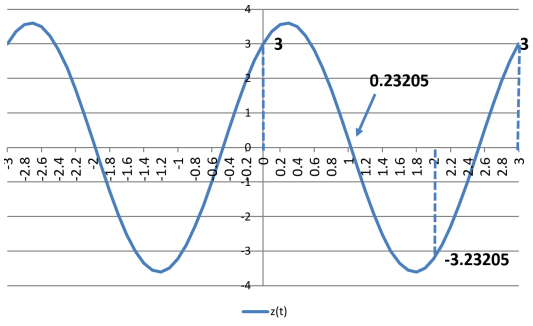


Figure 3.1: *Harmonic realization* $z_t = 3\cos\left(\frac{2}{3}\pi t\right) + 2\sin\left(\frac{2}{3}\pi t\right)$ *in the* $(-3, 3)$ *interval*

An interesting question is:
What happens if we construct a third-order predictor for this harmonic process? That is, if we specify the AR(3) model:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \phi_3 z_{t-3} + \varepsilon_t, \tag{3.25}$$

Using again the recursive formula for autocorrelations, we get the system:

$$\begin{bmatrix} 1 & \cos\omega & \cos 2\omega \\ \cos\omega & 1 & \cos\omega \\ \cos 2\omega & \cos\omega & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} = \begin{bmatrix} \cos\omega \\ \cos 2\omega \\ \cos 3\omega \end{bmatrix} \tag{3.26}$$

This system has no solution because the determinant of the 3×3 matrix in (3.26) is zero[45]. This result confirms that knowing only two values (or two random variables) is sufficient to determine the entire realization (or stochastic process) exactly. Including a third autoregressive term adds no further information: the regressor z_{t-3} is *collinear* with z_{t-1} and z_{t-2} and thus redundant.

Remark 3.7. Also, property B) of the theorem is important. The non-correlation between regular and singular components allows us to concentrate the analysis on the regular component. The singular component is less interesting from the econometric point of view because it is easily predictable. Once this easily predictable component has been determined, we can delete it from the stationary process and concentrate the analysis on the regular part. The regular part is more challenging because we must find the specification of the model that minimizes the variance of the prediction error.

Example 3.3. *(Wold Decomposition Theorem applied to the tide level in Venice –Punta della Salute)*
An interesting application of the Wold decomposition theorem is provided by the tide level recorded by the tide gauge at Punta della Salute in Venice.
Figure 3.2 shows the tide level time series from 7 to 28 November 2019.
This period was chosen because it includes the most recent exceptional tide in Venice (*Aqua Granda 2019*), the second highest ever recorded (189 cm), after that of 4 November

[45] This singularity of the matrix further justifies the term *singular process* used by Wold. According to Theorem 3 in his book (p. 47), process (3.8) is said to be *singular of rank 3* because the autocorrelation matrix in (3.26) is the first to have a vanishing determinant in the sequence of autocorrelation matrices of order 1, 2, ...
In 1944, the mathematician Joseph Doob proposed the term *deterministic process* instead of singular process, a convention now widely accepted. In the literature, the component z_t is referred to as a *purely deterministic stochastic process*, while the component x_t is known as a *purely non-deterministic stochastic process*. It is important not to confuse the term purely deterministic process with the deterministic trend in regression, which refers to a non-stochastic component.

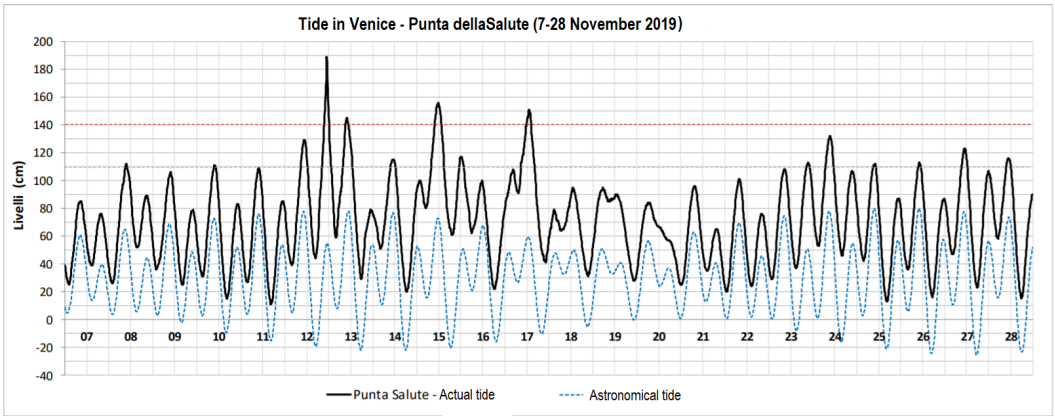


Figure 3.2: *Tide levels recorded by the tide gauge of Punta della Salute – Canale della Giudecca. The first red dotted line (110 cm) indicates the threshold beyond which the alarm sirens are activated. The second red dotted line (140 cm) marks the threshold for an exceptional tide. Measurements are relative to the Mareographic Zero of Punta della Salute – MZPS 1897.*

1966 (*Aqua Granda 1966*), which reached[46] 194 cm. In October 2020, the MOSE system[47] prevented further exceptional tides.
Figure 3.2 displays both the *observed tide* and the *astronomical tide*. The difference between the two defines the *atmospheric tide*.
Applying the Wold decomposition theorem, the tide level in Venice can be modeled as:

$$y_t = x_t + z_t \,,$$

where:

- y_t is the total tide level,
- x_t is the atmospheric tide,
- z_t is the astronomical tide.

The astronomical tide results from the sum of seven harmonic constituents[48].

[46] The period from 7 to 28 November 2019 was truly exceptional. Three events exceeding or equal to 150 cm occurred in November within just six days (from 12/11 to 17/11). In the last 150 years, no other year recorded more than one such event.
[47] MOSE is the acronym for *Modulo Sperimentale Elettromeccanico*, or *Experimental Electromechanical Module*. It consists of a series of mobile gates located at the inlets of the Venetian Lagoon.
[48] In this example, the astronomical tide is determined using tide level data provided by CNR-ISMAR (Institute of Marine Sciences of the National Research Council of Italy). The estimates were obtained using the tidal analysis software Polifemo (see the technical note by Tomasin (2005).

Denoting by z_t the hourly average astronomical level at time t, it is computed using the following formula:

$$z_t = R_0 + \sum_{i=1}^{7} R_i \cos(\omega_i t - \varphi_i)$$

where:

- $R_0 = 31$ cm is the current mean sea level with respect to the 1897 average, as measured by the Punta della Salute tide gauge;
- R_i is the amplitude of the i-th tidal component, in centimeters;
- ω_i is the angular frequency of that component, expressed in degrees per hour;
- Time t is measured in hours, and the phase lag φ_i is expressed in degrees. The convention is that $t = 1$ refers to 1 January 2019 at 1:00 AM (Italian time), and $t = 8760$ (since 2019 was not a leap year) corresponds to 31 December 2019 at midnight.

The harmonic constituents of the process are reported in *Table 3.1.*[49]

Harmonic Constituent	ω_i (deg/hour)	R_i (cm)	φ_i (°)
M2 – Principal lunar semidiurnal	28.9841042	24.1	187.5
S2 – Principal solar semidiurnal	30.0000000	13.8	324.0
N2 – Larger lunar elliptic semidiurnal	28.4397295	3.9	264.8
K2 – Lunisolar declinational semidiurnal	30.0821373	3.8	134.9
K1 – Lunisolar declinational diurnal	15.0410686	17.0	88.7
O1 – Principal lunar diurnal	13.9430356	5.0	308.5
P1 – Principal solar diurnal	14.9589314	5.5	97.7

Table 3.1: *Harmonic constituents in Venice (Punta della Salute), year 2019*

To illustrate the difference between the observed tide and the astronomical tide, the tide forecast issued by the *Centro Previsioni e Segnalazioni Maree di Venezia* for 12/11/2019 is shown in *Figure 3.3.*
At the same times during the day, the differences are quite evident, as shown[50] in *Table 3.2.*

[49] The *Centro Previsioni e Segnalazioni Maree di Venezia – CPSM (Tide Forecast and Alert Center of Venice)* includes an additional component called *S1* – the "meteorological solar wave." It is a diurnal constituent with angular frequency 15.0000020, amplitude 1.4 cm, and phase 275°.

[50] The tide levels observed at Punta della Salute on 12 November 2019 (astronomical and total observed) are based on hourly mareographic records from the RMLV network and CPSM tide forecasts. The atmospheric contribution is obtained as the difference between observed and astronomical components

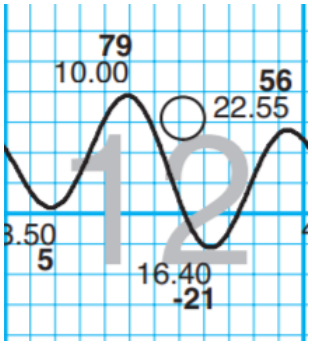


Figure 3.3: *Tide forecast issued by CPSM for 12/11/2019*

Hour	Astronomical tide (cm)	Observed tide (cm)	Atmospheric tide (cm)
03:50	5	40	35
10:00	79	130	51
16:40	-21	45	66
22:55	56	189	133

Table 3.2: *Comparison between astronomical and observed tide levels on 12/11/2019*

Both the astronomical and the observed tides exhibit a clear cyclical pattern over the day. In contrast, the atmospheric tide displays a non-cyclical upward movement: from the early morning until 4:40 PM the increments are nearly constant, while after that time they increase significantly.

Following *Remark 3.7*, from an econometric point of view it is not convenient to model directly the observed tide level time series. The absence of correlation between the astronomical tide (interpreted as a singular process) and the atmospheric tide (interpreted as a regular process) suggests subtracting the astronomical component from the observed data. The same procedure applies whenever the observed series contains a *perfectly predictable singular component*, such as a deterministic or harmonic seasonal pattern.

Furthermore, after subtracting the astronomical and seasonal components, the resulting time series may still be affected by other factors, such as *subsidence* and *eustatism*. For this reason, the atmospheric tide is denoted by v_t and is modeled as the sum of a non-stochastic trend and a stochastic component driven by atmospheric variables. In this

context, we consider the following specification[51]:

$$\begin{cases} v_t = d_t + s_t, \\ s_t = \sum_{j=1}^{k} \beta_j w_{jt} + x_t, \\ d_t = \alpha_0 + \alpha_1 t, \\ x_t = \rho x_{t-1} + \varepsilon_t, \qquad |\rho| < 1, \end{cases}$$

- $\{w_{jt},\, j = 1, \ldots, k\}$ are atmospheric variables such as barometric pressure, temperature, precipitation, and wind direction and intensity[52];

- d_t is a linear non-stochastic trend. This deterministic non-stochastic component is essential for estimating, through the coefficient α_1, the effects of *subsidence* (vertical land sinking) and *eustatism* (long-run sea level variation in the Adriatic Sea) over the sample period[53];

- x_t is a stationary autoregressive process with $|\rho| < 1$ and therefore represents the regular component in the sense of the Wold decomposition theorem.

- ε_t is a *white noise* process.

[51] A model with this specification can be found in Sartore (1975).

[52] Wind direction and intensity must be parameterized to be included as regressors. In Sartore (1975), parameterizations were adopted for the eight main directions: N, NE, E, SE, S, SW, W, NW.

[53] In Sartore (1975), subsidence was estimated using monthly data for the period 1947–1971, yielding a value of 2.479 mm per year. Remarkably, this coincides with the value obtained in 1970 by the Italian National Research Council (CNR) using geometric precision leveling (2.48 mm/year).
More recently, De Biasio, Baldin, and Vignudelli (2020) distinguish between Vertical Land Motion (VLM) and Absolute Sea Level (ASL). VLM was estimated at –1.59 mm/year using GPS data, while ASL increased at an average rate of 2.43 mm/year between 1974 and 2018 across six tide gauge stations.

4 Multivariate Stochastic Processes

Definition 4.1. *(Multivariate Stochastic Process)*
A multivariate (or multidimensional) stochastic process $\{\mathbf{X}_t;\, t \in \mathcal{T}\}$ *is a family of random vectors indexed by time* t*. For each* t*, we have* $\mathbf{X}_t \in \mathbb{R}^n$*.*

If the index t is fixed, then the vector $\mathbf{X}_t$ consists of n components X_{jt}, for $j = 1,\, 2,\, \ldots,\, n$. In what follows, it is assumed that each component is square-integrable, that is,

$$E[X_{jt}^2] < \infty, \quad \forall t,\, j.$$

A necessary and sufficient condition for this is:

$$E\left\|\mathbf{X}_t\right\|^2 = E\left(\sum_{j=1}^{n} X_{jt}^2\right) < \infty.$$

The expected value of the vector is defined as:

$$E[\mathbf{X}_t] = \begin{bmatrix} E[X_{1t}] \\ \vdots \\ E[X_{nt}] \end{bmatrix}. \tag{4.1}$$

The variance matrix is:

$$\begin{aligned} \mathbf{\Gamma}_{t,t} &= \mathrm{Var}(\mathbf{X}_t) = E\left[(\mathbf{X}_t - E[\mathbf{X}_t])\,(\mathbf{X}_t - E[\mathbf{X}_t])'\right] \\ &= E[\mathbf{X}_t\mathbf{X}_t'] - E[\mathbf{X}_t]E[\mathbf{X}_t]' \;=\; \begin{bmatrix} \mathrm{Var}(X_{1t}) & \mathrm{Cov}(X_{1t}, X_{2t}) & \cdots & \mathrm{Cov}(X_{1t}, X_{nt}) \\ \mathrm{Cov}(X_{2t}, X_{1t}) & \mathrm{Var}(X_{2t}) & \cdots & \mathrm{Cov}(X_{2t}, X_{nt}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{nt}, X_{1t}) & \mathrm{Cov}(X_{nt}, X_{2t}) & \cdots & \mathrm{Var}(X_{nt}) \end{bmatrix}. \end{aligned} \tag{4.2}$$

The covariance function is:

$$\begin{aligned} \mathbf{\Gamma}_{t,t+k} &= \mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_{t+k}) = E\left[(\mathbf{X}_t - E[\mathbf{X}_t])\,(\mathbf{X}_{t+k} - E[\mathbf{X}_{t+k}])'\right] \\ &= E[\mathbf{X}_t\mathbf{X}_{t+k}'] - E[\mathbf{X}_t]E[\mathbf{X}_{t+k}]' \\ &= \begin{bmatrix} \mathrm{Cov}(X_{1t}, X_{1,t+k}) & \cdots & \mathrm{Cov}(X_{1t}, X_{n,t+k}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{nt}, X_{1,t+k}) & \cdots & \mathrm{Cov}(X_{nt}, X_{n,t+k}) \end{bmatrix}. \end{aligned} \tag{4.3}$$

The correlation function is defined element-wise as the well-known ratio:

$$\rho_{ij,t,t+k} = \frac{\mathrm{Cov}(X_{it}, X_{j,t+k})}{\sqrt{\mathrm{Var}(X_{it})}\sqrt{\mathrm{Var}(X_{j,t+k})}}, \quad i, j = 1, \ldots, n;\ \forall t, k\,. \tag{4.4}$$

4.1 Multivariate Stationary Process

As in the univariate case, multivariate processes also allow for two types of stationarity: strict-sense stationarity and covariance stationarity.
A multivariate process is said to be strictly stationary when the family of finite-dimensional distributions is invariant under time shifts.

Definition 4.2. (Covariance Stationarity)
A multivariate process is stationary in covariance if:

$$\begin{aligned} &E[\mathbf{X}_t] = \boldsymbol{\mu}, \quad \forall t \\ &\mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_{t+k}) = E\left[(\mathbf{X}_t - E[\mathbf{X}_t])(\mathbf{X}_{t+k} - E[\mathbf{X}_{t+k}])'\right] = \boldsymbol{\Gamma}_k, \quad \forall t, k\,. \end{aligned} \tag{4.5}$$

The covariance function $\boldsymbol{\Gamma}_k$ has the following property:

$$\begin{aligned} \boldsymbol{\Gamma}_{-k} &= \mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_{t-k}) = E\left[(\mathbf{X}_t - E[\mathbf{X}_t])(\mathbf{X}_{t-k} - E[\mathbf{X}_{t-k}])'\right] \\ &= E\left[(\mathbf{X}_t - \boldsymbol{\mu})(\mathbf{X}_{t-k} - \boldsymbol{\mu})'\right] = \left(E\left[(\mathbf{X}_s - \boldsymbol{\mu})(\mathbf{X}_{s+k} - \boldsymbol{\mu})'\right]\right)' = \boldsymbol{\Gamma}_k' \end{aligned} \tag{4.6}$$

The penultimate equality is obtained by substituting $s = t-k$ and applying the transpose rule for matrix products: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
Unlike the univariate case, the covariance function is not an even function. This asymmetry is not surprising and is clarified in the following example.

Example 4.1. Consider the case $k = -1$ and $n = 2$. The covariance matrix is:

$$\begin{aligned} \boldsymbol{\Gamma}_1 &= \mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_{t-1}) = E\left[(\mathbf{X}_t - E[\mathbf{X}_t])(\mathbf{X}_{t-1} - E[\mathbf{X}_{t-1}])'\right] \\ &= \begin{bmatrix} E[(X_{1t} - E[X_{1t}])(X_{1,t-1} - E[X_{1,t-1}])] & E[(X_{1t} - E[X_{1t}])(X_{2,t-1} - E[X_{2,t-1}])] \\ E[(X_{2t} - E[X_{2t}])(X_{1,t-1} - E[X_{1,t-1}])] & E[(X_{2t} - E[X_{2t}])(X_{2,t-1} - E[X_{2,t-1}])] \end{bmatrix} \\ &= \begin{bmatrix} \gamma_{11,1} & \gamma_{12,1} \\ \gamma_{21,1} & \gamma_{22,1} \end{bmatrix} \end{aligned}$$

The third index (after the comma) indicates the lag of the process that precedes it. Therefore, it is clear that $\gamma_{12,1} \neq \gamma_{21,1}$. The cross-covariance $\gamma_{12,1}$ refers to the relationship between process X_{1t} and the past of process X_{2t} (red arrow in *Figure 4.1*), while $\gamma_{21,1}$ refers to the relationship between X_{2t} and the past of X_{1t} (black arrow).
Example 4.1 shows that the matrix $\boldsymbol{\Gamma}_1$ is not symmetric, and in fact, this applies to all $\boldsymbol{\Gamma}_k$ with $k \neq 0$.
A second property is that $\boldsymbol{\Gamma}_k$ is a positive semi-definite function, that is:

$$\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{a}_i'\boldsymbol{\Gamma}_{|i-j|}\mathbf{a}_j \geq 0\,, \tag{4.7}$$

which follows from the fact that the variance of the linear combination $\sum_{j=1}^{p} \mathbf{a}_j'\mathbf{X}_{t-j}$ cannot be negative.

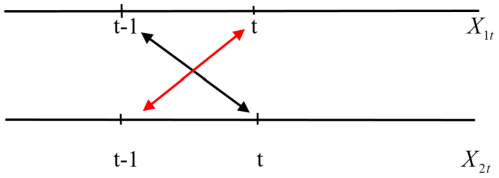


Figure 4.1: *Interrelations between X_{1t} and X_{2t} processes*

Definition 4.3. (Multivariate White Noise)
A process $\boldsymbol{\varepsilon}_t$ is a white noise process if:

$$\begin{aligned} E[\boldsymbol{\varepsilon}_t] &= \mathbf{0} \\ \boldsymbol{\Gamma}_k &= \delta_k \boldsymbol{\Sigma}, \quad \forall k \,, \end{aligned} \tag{4.8}$$

where $\boldsymbol{\Sigma} = E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t']$ is the variance matrix of the process, and δ_k is the Kronecker delta[54] *defined as:*

$$\delta_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \tag{4.9}$$

The components of a white noise process can be contemporaneously correlated, but are uncorrelated across time.

Definition 4.4. *(VARMA(p, q) Process)*
The multivariate process $\mathbf{X}_t$ is called a Vector Autoregressive Moving Average process, denoted by VARMA(p, q), if it has the following representation:

$$\mathbf{A}(L)\mathbf{X}_t = \mathbf{B}(L)\boldsymbol{\varepsilon}_t \,, \tag{4.10}$$

where $\boldsymbol{\varepsilon}_t$ is a white noise process. The matrices $\mathbf{A}(L)$ and $\mathbf{B}(L)$ are matrix polynomials in the lag operator L, defined as:

$$\begin{aligned} \mathbf{A}(L) &= \mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p \\ \mathbf{B}(L) &= \mathbf{I} + \mathbf{B}_1 L + \mathbf{B}_2 L^2 + \cdots + \mathbf{B}_q L^q \,, \end{aligned} \tag{4.11}$$

where $\mathbf{I}$ is the identity matrix, and the matrices $\mathbf{A}_i$, $\mathbf{B}_j$, with $i = 1, \ldots, p$ and $j = 1, \ldots, q$, are square coefficient matrices of the same dimension.

An alternative representation of (4.11) is:

$$\mathbf{A}(L) = \begin{bmatrix} a_{11}(L) & \cdots & a_{1n}(L) \\ \vdots & \ddots & \vdots \\ a_{n1}(L) & \cdots & a_{nn}(L) \end{bmatrix}, \quad \mathbf{B}(L) = \begin{bmatrix} b_{11}(L) & \cdots & b_{1n}(L) \\ \vdots & \ddots & \vdots \\ b_{n1}(L) & \cdots & b_{nn}(L) \end{bmatrix}, \tag{4.12}$$

[54] Named after the German mathematician Leopold Kronecker (1823–1891).

where:

$$a_{ij}(L) = \begin{cases} 1 - \sum_{s=1}^{p} a_{ijs} L^s, & \text{if } i = j \\ -\sum_{s=1}^{p} a_{ijs} L^s, & \text{if } i \neq j \end{cases} \quad \text{and} \quad b_{ij}(L) = \begin{cases} 1 + \sum_{s=1}^{q} b_{ijs} L^s, & \text{if } i = j \\ \sum_{s=1}^{q} b_{ijs} L^s, & \text{if } i \neq j \end{cases}$$

Example 4.2. Representations (4.11) and (4.12) can be used to define an *ARMA(2,1)* model for a stochastic process.
According to (4.11), the polynomial matrices are:

$$\mathbf{A}(L) = \mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2,$$
$$\mathbf{A}_1 = \begin{bmatrix} a_{111} & a_{121} & a_{131} \\ a_{211} & a_{221} & a_{231} \\ a_{311} & a_{321} & a_{331} \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} a_{112} & a_{122} & a_{132} \\ a_{212} & a_{222} & a_{232} \\ a_{312} & a_{322} & a_{332} \end{bmatrix}$$

and:

$$\mathbf{B}(L) = \mathbf{I} + \mathbf{B}_1 L, \quad \mathbf{B}_1 = \begin{bmatrix} b_{111} & b_{121} & b_{131} \\ b_{211} & b_{221} & b_{231} \\ b_{311} & b_{321} & b_{331} \end{bmatrix}$$

Thus, the VARMA(2,1) process becomes:

$$\mathbf{A}(L)\mathbf{X}_t = \mathbf{B}(L)\boldsymbol{\varepsilon}_t \quad \Rightarrow \quad \mathbf{X}_t = \mathbf{A}_1\mathbf{X}_{t-1} + \mathbf{A}_2\mathbf{X}_{t-2} + \mathbf{B}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\varepsilon}_t$$

In expanded form:

$$\begin{aligned} \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} &= \mathbf{A}_1 \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{bmatrix} + \mathbf{A}_2 \begin{bmatrix} X_{1,t-2} \\ X_{2,t-2} \\ X_{3,t-2} \end{bmatrix} + \mathbf{B}_1 \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \\ \varepsilon_{3,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \\ &= \begin{bmatrix} a_{111} & a_{121} & a_{131} \\ a_{211} & a_{221} & a_{231} \\ a_{311} & a_{321} & a_{331} \end{bmatrix} \begin{bmatrix} X_{1t-1} \\ X_{2t-1} \\ X_{3t-1} \end{bmatrix} + \begin{bmatrix} a_{112} & a_{122} & a_{132} \\ a_{212} & a_{222} & a_{232} \\ a_{312} & a_{322} & a_{332} \end{bmatrix} \begin{bmatrix} X_{1t-2} \\ X_{2t-2} \\ X_{3t-2} \end{bmatrix} \\ &+ \begin{bmatrix} b_{111} & b_{121} & b_{131} \\ b_{211} & b_{221} & b_{231} \\ b_{311} & b_{321} & b_{331} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t-1} \\ \varepsilon_{2t-1} \\ \varepsilon_{3t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \end{aligned}$$

More explicitly:

$$\begin{cases} X_{1t} = a_{111}X_{1t-1} + a_{121}X_{2t-1} + a_{131}X_{3t-1} + a_{112}X_{1t-2} + a_{122}X_{2t-2} + a_{132}X_{3t-2} \\ \quad\quad +b_{111}\varepsilon_{1t-1} + b_{121}\varepsilon_{2t-1} + b_{131}\varepsilon_{3t-1} + \varepsilon_{1t} \\ X_{2t} = a_{211}X_{1t-1} + a_{221}X_{2t-1} + a_{231}X_{3t-1} + a_{212}X_{1t-2} + a_{222}X_{2t-2} + a_{232}X_{3t-2} \\ \quad\quad +b_{211}\varepsilon_{1t-1} + b_{221}\varepsilon_{2t-1} + b_{231}\varepsilon_{3t-1} + \varepsilon_{2t} \\ X_{3t} = a_{311}X_{1t-1} + a_{321}X_{2t-1} + a_{331}X_{3t-1} + a_{312}X_{1t-2} + a_{322}X_{2t-2} + a_{332}X_{3t-2} \\ \quad\quad +b_{311}\varepsilon_{1t-1} + b_{321}\varepsilon_{2t-1} + b_{331}\varepsilon_{3t-1} + \varepsilon_{3t} \end{cases}$$

Some observations are relevant to this structure:

- Each process includes autoregressive components and depends on the lagged values of the other processes; the same applies to the error term ε.

- Regarding the parameter indexing: from a mnemonic point of view, the order of the indices can be interpreted as follows. The first index refers to the process on the left-hand side of the equation (i.e., it identifies the equation); the second index refers to the process whose coefficient is being specified; the third index indicates the lag of that process.

- Not all coefficients need to be different from zero. At least one of the coefficients in the matrix $\mathbf{A}_2$ must be nonzero to determine the autoregressive order of the process, i.e., $p = 2$. Similarly, at least one of the coefficients in the matrix $\mathbf{B}_1$ must be nonzero to determine the moving average order, i.e., $q = 1$.

- In econometric terminology, the processes are often referred to as "variables." The variables on the left-hand side of the equation are called *endogenous* (or *dependent*) variables. The variables on the right-hand side are called *explanatory* (or *independent*) variables and may include both endogenous and exogenous terms. If any of the explanatory variables include contemporaneous endogenous variables, we are dealing with a *simultaneous equations linear system*. The *ARMA(p, q)* structure, however, does not include contemporaneous endogenous variables among the explanatory variables and, therefore, cannot be interpreted as a simultaneous equations model[55].

According to representation (4.12), the polynomial matrices of the *ARMA(2, 1)* model are:

$$\mathbf{A}(L) = \begin{bmatrix} a_{11}(L) & a_{12}(L) & a_{13}(L) \\ a_{21}(L) & a_{22}(L) & a_{23}(L) \\ a_{31}(L) & a_{32}(L) & a_{33}(L) \end{bmatrix},$$

[55] In the terminology of simultaneous equation systems, ARMA models are referred to as *reduced form models*.

where the individual polynomials are defined as:

$$\begin{aligned}
a_{11}(L) &= 1 - a_{111}L - a_{112}L^2, & a_{12}(L) &= -a_{121}L - a_{122}L^2, & a_{13}(L) &= -a_{131}L - a_{132}L^2 \\
a_{21}(L) &= -a_{211}L - a_{212}L^2, & a_{22}(L) &= 1 - a_{221}L - a_{222}L^2, & a_{23}(L) &= -a_{231}L - a_{232}L^2 \\
a_{31}(L) &= -a_{311}L - a_{312}L^2, & a_{32}(L) &= -a_{321}L - a_{322}L^2, & a_{33}(L) &= 1 - a_{331}L - a_{332}L^2
\end{aligned}$$

Similarly, for the polynomial matrix $\mathbf{B}(L)$ we have:

$$\mathbf{B}(L) = \begin{bmatrix} b_{11}(L) & b_{12}(L) & b_{13}(L) \\ b_{21}(L) & b_{22}(L) & b_{23}(L) \\ b_{31}(L) & b_{32}(L) & b_{33}(L) \end{bmatrix},$$

where the individual polynomials are defined as:

$$\begin{aligned}
b_{11}(L) &= 1 + b_{111}L, & b_{12}(L) &= b_{121}L, & b_{13}(L) &= b_{131}L \\
b_{21}(L) &= -b_{211}L, & b_{22}(L) &= 1 + b_{221}L, & b_{23}(L) &= b_{231}L \\
b_{31}(L) &= b_{311}L, & b_{32}(L) &= b_{321}L, & b_{33}(L) &= 1 + b_{331}L\,.
\end{aligned}$$

Each univariate polynomial has its own degree depending on whether its coefficients are zero or nonzero. The requirement that at least one of the coefficients a_{132}, a_{232}, a_{332} must be nonzero ensures that the autoregressive order of the multivariate process is indeed $p = 2$. Similarly, at least one of the coefficients b_{131}, b_{231}, b_{331} must be nonzero to ensure that the moving average order is $q = 1$.
In this representation, instead of defining the order of the multivariate process through two scalar values, p and q, a matrix definition is used:

$$\mathbf{p} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}.$$

Although apparently more complex, representation (4.12) is useful when verifying the stationarity of an *ARMA* process.

4.1.1 Stationarity and Invertibility of the *VARMA* Process

The VARMA process $\mathbf{X}_t$ is stationary in covariance if all the roots of the determinantal polynomial equation

$$|\mathbf{A}(L)| = 0 \tag{4.13}$$

lie outside the unit circle[56].

[56] See the discussion in Chapter 2 on the interpretation of lag polynomials.

The process is invertible if all the roots of the determinantal polynomial equation

$$|\mathbf{B}(L)| = 0 \tag{4.14}$$

lie outside the unit circle.
If the stationarity condition is satisfied, the $ARMA(p, q)$ process admits a dual $MA(\infty)$ representation. In this case, the inverse of the matrix polynomial $\mathbf{A}(L)$ exists and the process can be written as:

$$\begin{aligned} \mathbf{X}_t &= \mathbf{A}^{-1}(L)\,\mathbf{B}(L)\,\boldsymbol{\varepsilon}_t \\ &= \mathbf{C}(L)\,\boldsymbol{\varepsilon}_t, \end{aligned} \tag{4.15}$$

where

$$\mathbf{C}(L) = \sum_{s=0}^{\infty} \mathbf{C}_s\, L^s, \quad \mathbf{C}_0 = \mathbf{I}, \quad \sum_{s=1}^{\infty} \|\mathbf{C}_s\| < \infty. \tag{4.16}$$

Here, $\|\mathbf{Q}\|$ denotes the *spectral norm* of a square matrix $\mathbf{Q}$, defined as the square root of the largest eigenvalue of $\mathbf{Q}'\mathbf{Q}$, that is, $\|\mathbf{Q}\| = \sqrt{\lambda_{\max}(\mathbf{Q}'\mathbf{Q})}$.
If the sequence of coefficient matrices $\mathbf{C}_k$ is *absolutely summable*, i.e.,

$$\sum_{s=1}^{\infty} \|\mathbf{C}_s\| < \infty,$$

then the process $\mathbf{X}_t = \sum_{k=0}^{\infty} \mathbf{C}_k\, L^k\, \boldsymbol{\varepsilon}_t$ is stationary in covariance.
In this sense, absolute summability provides a condition for stationarity equivalent to the one based on the root behavior in equation (4.13)[57].

Example 4.3. Consider the bivariate process $\{y_t,\, x_t\}$:

$$\begin{cases} y_t = -y_{t-1} + x_{t-1} + \varepsilon_t \\ x_t = x_{t-1} - 0.5\, y_{t-1} + \eta_t, \end{cases}$$

where $\{\varepsilon_t,\, \eta_t\}$ is a white noise process. To verify stationarity, we write the system in matrix form, using representation (4.12):

$$\begin{bmatrix} 1+L & -L \\ 0.5L & 1-L \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix}.$$

The determinantal polynomial equation becomes:

$$(1+L)(1-L) + 0.5L^2 = 0 \quad \Rightarrow \quad 1 - 0.5L^2 = 0.$$

Its roots are $\sqrt{2}$ and $-\sqrt{2}$, both real, distinct, and outside the unit circle. Therefore, the process is stationary in covariance.
Note that the process coefficients are not necessarily constrained to have modulus less than one.

[57] In the univariate case, stationarity requires that the series $\sum_{s=1}^{\infty} |c_s|$ be convergent.

Example 4.4. Consider the bivariate process $\{y_t,\, x_t\}$:

$$\begin{cases} y_t = -0.4\, y_{t-1} + x_{t-1} + \varepsilon_t \\ x_t = x_{t-1} - 1.2\, y_{t-1} + \eta_t, \end{cases}$$

where $\{\varepsilon_t,\, \eta_t\}$ is a white noise process.
The determinantal polynomial equation becomes:

$$(1 + 0.4L)(1 - L) + 1.2L^2 = 0 \quad \Rightarrow \quad 1 - 0.6L + 0.8L^2 = 0.$$

This quadratic equation has complex roots:

$$\frac{3}{8} \pm i\frac{\sqrt{71}}{8},$$

which lie outside the unit circle. This can be seen since the squared modulus of the roots is

$$\left(\frac{3}{8}\right)^2 + \left(\frac{\sqrt{71}}{8}\right)^2 = \frac{9 + 71}{64} = \frac{80}{64} = 1.25 > 1.$$

Therefore, the process is stationary in covariance.

Example 4.5. The two previous examples suggest a generalization of the bivariate model with symbolic (unspecified) coefficients:

$$\begin{cases} y_t = a\, y_{t-1} + x_{t-1} + \varepsilon_t \\ x_t = x_{t-1} + b\, y_{t-1} + \eta_t \end{cases}, \tag{4.17}$$

where $\{\varepsilon_t,\, \eta_t\}$ is a white noise process.
The purpose of this generalization is to determine the admissible region for the parameters a and b such that the bivariate process is stationary in covariance.
Using the matrix representation, we have:

$$\begin{bmatrix} 1 - aL & -L \\ -bL & 1 - L \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix}.$$

The determinantal polynomial equation becomes:

$$(1 - aL)(1 - L) - bL^2 = 0 \quad \Rightarrow \quad 1 - (a + 1)L - (b - a)L^2 = 0.$$

We recall that the stationarity condition for a univariate $AR(2)$ process requires the roots of the characteristic equation

$$1 - \alpha_1 L - \alpha_2 L^2 = 0$$

to lie outside the unit circle. This condition defines the following triangular region in the (α_1, α_2) plane (see § 2.5.1):

$$\begin{cases} \alpha_2 + \alpha_1 < 1 \\ \alpha_2 - \alpha_1 < 1 \\ -1 < \alpha_2 < 1 \end{cases} .$$

By setting $\alpha_1 = a + 1$ and $\alpha_2 = b - a$, the inequalities become:

$$\begin{cases} b - a + (a + 1) < 1 \\ b - a - (a + 1) < 1 \\ -1 < b - a < 1 \end{cases} . \tag{4.18}$$

Rewriting these conditions in terms of a and b, we obtain:

$$\begin{cases} b < 0 \\ b < 2(1 + a) \\ b < a + 1 \\ b > a - 1 \end{cases} . \tag{4.19}$$

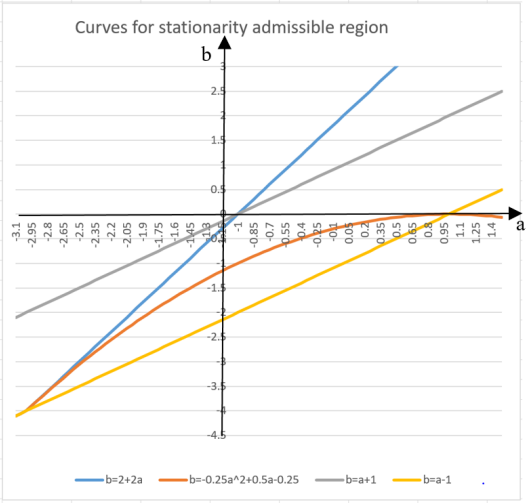


Figure 4.2: *Curves for determining the stationarity region*

In the (a, b) Cartesian plane:

- The second inequality in (4.19) defines the region below the line $b = 2(1 + a)$ (blue line in Figure 4.2).
- The third and fourth inequalities correspond to the region between the lines $b = a+1$ and $b = a - 1$ (grey and yellow lines).

- When $b = 0$, the third inequality of (4.18) implies $-1 < a < 1$.
- The third inequality of (4.19) is redundant since it conflicts with both the first and second inequalities and does not further restrict the admissible region.

The discriminant of the second-order polynomial is given by:

$$\Delta = (1 + a)^2 + 4(b - a) = a^2 - 2a + 4b + 1.$$

Solving for b in terms of a, we obtain the parabola:

$$b = -\frac{1}{4}a^2 + \frac{1}{2}a - \frac{1}{4}.$$

This parabola is concave and reaches its maximum at the point (1, 0).
The discriminant is positive (real roots) when:

$$b > -\frac{1}{4}a^2 + \frac{1}{2}a - \frac{1}{4},$$

i.e., above the red parabola in *Figure 4.2*. The origin lies below this curve. By combining the first, second, and fourth inequalities from system (4.19), the admissible stationarity region is the triangular area shown in *Figure 4.3*.

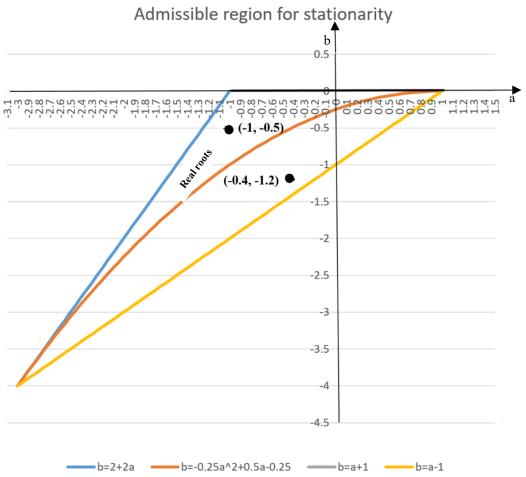


Figure 4.3: *Stationarity admissible region for bivariate process (4.17)*

In conclusion, for the bivariate process to be stationary in covariance, the values of the parameters a and b must lie strictly within the triangular region shown in *Figure 4.3*.
The values chosen in *Examples 4.3* and *4.4* are consistent with this result. In particular:

- The point $(-1, -0.5)$ lies above the parabola and inside the admissible region (real roots).
- The point $(-0.4, -1.2)$ lies below the parabola but still inside the admissible region (complex roots).

4.2 Covariance Function of the *VARMA* Process

We consider the *VARMA(p, q)* process defined in (4.10), which can be rewritten as:

$$\mathbf{X}_t - \sum_{j=1}^{p} \mathbf{A}_j \mathbf{X}_{t-j} = \sum_{j=0}^{q} \mathbf{B}_j \boldsymbol{\varepsilon}_{t-j}. \tag{4.20}$$

Taking the transpose of (4.20), multiplying it on the left by the lagged process, and then computing the expected value, we obtain:

$$E[\mathbf{X}_{t-k}\mathbf{X}_t'] - E\left[\mathbf{X}_{t-k} \sum_{j=1}^{p} \mathbf{X}_{t-j}'\mathbf{A}_j'\right] = E\left[\mathbf{X}_{t-k} \sum_{j=0}^{q} \boldsymbol{\varepsilon}_{t-j}'\mathbf{B}_j'\right]. \tag{4.21}$$

Recalling from (4.6) that $\boldsymbol{\Gamma}_k = Cov(\mathbf{X}_t, \mathbf{X}_{t+k}) = [Cov(\mathbf{X}_{t-k}, \mathbf{X}_t)]'$, we rewrite:

$$\begin{aligned} \boldsymbol{\Gamma}_k - \sum_{j=1}^{p} E\left(\mathbf{X}_{t-k}\mathbf{X'}_{t-j}\right)\mathbf{A'}_j &= \sum_{j=0}^{q} E\left(\mathbf{X}_{t-k}\boldsymbol{\varepsilon'}_{t-j}\right)\mathbf{B'}_j \\ \boldsymbol{\Gamma}_k - \sum_{j=1}^{p} \boldsymbol{\Gamma}_{k-j}\mathbf{A'}_j &= \sum_{j=0}^{q} \left(\sum_{s=0}^{\infty} \mathbf{C}_s E\boldsymbol{\varepsilon}_{t-k-s}\boldsymbol{\varepsilon'}_{t-j}\right)\mathbf{B'}_j, \end{aligned} \tag{4.22}$$

where, on the right-hand side of the second line, $\mathbf{X}_{t-k}$ has been replaced by its dual MA(∞) representation, as in (4.15). Since $\boldsymbol{\Sigma}$ is the variance matrix of the white noise process, the expectation in the right-hand side takes specific values:

- For $j = 0$:

$$\sum_{s=0}^{\infty} \mathbf{C}_s E[\boldsymbol{\varepsilon}_{t-k-s}\boldsymbol{\varepsilon}_t']\mathbf{B}_0' = \begin{cases} \mathbf{C}_0\boldsymbol{\Sigma}\mathbf{B}_0', & \text{if } k = 0,\ s = 0 \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

- For $j = 1$:

$$\sum_{s=0}^{\infty} \mathbf{C}_s E[\boldsymbol{\varepsilon}_{t-k-s}\boldsymbol{\varepsilon}_{t-1}']\mathbf{B}_1' = \begin{cases} \mathbf{C}_0\boldsymbol{\Sigma}\mathbf{B}_1', & \text{if } k = 1,\ s = 0 \\ \mathbf{C}_1\boldsymbol{\Sigma}\mathbf{B}_1', & \text{if } k = 0,\ s = 1 \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

- ...

- For $j = q$:

$$\sum_{s=0}^{\infty} \mathbf{C}_s E[\boldsymbol{\varepsilon}_{t-k-s}\boldsymbol{\varepsilon}_{t-q}']\mathbf{B}_q' = \begin{cases} \mathbf{C}_0\boldsymbol{\Sigma}\mathbf{B}_q', & \text{if } k = q,\ s = 0 \\ \mathbf{C}_1\boldsymbol{\Sigma}\mathbf{B}_q', & \text{if } k = q-1,\ s = 1 \\ \vdots & \\ \mathbf{C}_q\boldsymbol{\Sigma}\mathbf{B}_q', & \text{if } k = 0,\ s = q \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

The total sum over j becomes:

$$\begin{cases} \sum_{j=0}^{q} \mathbf{C}_j \boldsymbol{\Sigma} \mathbf{B}'_j, & k = 0 \\ \sum_{j=1}^{q} \mathbf{C}_j \boldsymbol{\Sigma} \mathbf{B}'_j, & k = 1 \\ \vdots & \\ \mathbf{C}_0 \boldsymbol{\Sigma} \mathbf{B}'_q, & k = q \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

Substituting into equation (4.22), we obtain the expression for the covariance function:

$$\boldsymbol{\Gamma}_k = \sum_{j=1}^{p} \boldsymbol{\Gamma}_{k-j} \mathbf{A}'_j + \begin{cases} \sum_{j=k}^{q} \mathbf{C}_{j-k} \boldsymbol{\Sigma} \mathbf{B}'_j, & 0 \leq k \leq q \\ \mathbf{0}, & k > q \end{cases} \tag{4.23}$$

Equivalently, by changing index $s = j - k$:

$$\boldsymbol{\Gamma}_k = \sum_{j=1}^{p} \boldsymbol{\Gamma}_{k-j} \mathbf{A}'_j + \begin{cases} \sum_{s=0}^{q-k} \mathbf{C}_s \boldsymbol{\Sigma} \mathbf{B}'_{s+k}, & 0 \leq k \leq q \\ \mathbf{0}, & k > q \end{cases} \tag{4.24}$$

The analogy with the univariate case is evident. As in the univariate case, the covariance function depends on the process parameters. This relationship is particularly useful when the theoretical $\boldsymbol{\Gamma}_k$ are replaced with estimated values $\hat{\boldsymbol{\Gamma}}_k$ to obtain approximate estimates of the process parameters.
In the simpler case of a vector MA(q) process, the covariance function (4.24) reduces to:

$$\boldsymbol{\Gamma}_k = \begin{cases} \sum_{s=0}^{q-k} \mathbf{C}_s \boldsymbol{\Sigma} \mathbf{B}'_{s+k}, & 0 \leq k \leq q \\ \mathbf{0}, & k > q \end{cases} \tag{4.25}$$

Also in the multivariate case, *the covariance function is an indicator of the order of the process*, since its values are null for $k > q$.
This is not the case for a VAR(p) process, for which the covariance function is given by:

$$\boldsymbol{\Gamma}_k = \begin{cases} \sum_{j=1}^{p} \boldsymbol{\Gamma}_{k-j} \mathbf{A}'_j + \boldsymbol{\Sigma}, & k = 0 \\ \sum_{j=1}^{p} \boldsymbol{\Gamma}_{k-j} \mathbf{A}'_j, & k > 0 \end{cases} \tag{4.26}$$

In this case, $\boldsymbol{\Gamma}_k$ tends to zero only as $k \to \infty$.
In econometric applications, VAR models are widely used.

5 Dynamic Properties of Steady-State Systems

A particular formulation of *linear dynamic systems*, represented through *Autoregressive Distributed Lag (ADL) models*, allows the analysis of s*teady-state equilibrium* and dynamic properties.
An $ADL(p, q)$ model is defined as:

$$\begin{cases} \alpha(L)y_t &= \beta(L)x_t + \varepsilon_t \\ \alpha(L) &= 1 - \alpha_1 L - \cdots - \alpha_p L^p \\ \beta(L) &= \beta_0 + \beta_1 L + \cdots + \beta_q L^q \end{cases} \tag{5.1}$$

The following assumptions are made:

- $E(\varepsilon_t | x_t, x_{t-1}, \ldots) = 0$, which implies $E(\varepsilon_t x_s) = 0$ for all $s \leq t$, i.e., the *strict exogeneity* of x_t;

- The polynomial equation $\alpha(L) = 0$ has all roots outside the unit circle. This condition is referred to as the *stability condition*[58] of the ADL model, and allows the model to be rewritten as a function of the exogenous variable only, by inverting the polynomial $\alpha(L)$:

$$y_t = \alpha(L)^{-1}\beta(L)x_t + \alpha(L)^{-1}\varepsilon_t \tag{5.2}$$

  The polynomial ratio $\alpha(L)^{-1}\beta(L)$ is referred to as the *transfer function*.

- The coefficient β_0 is not necessarily equal to 1, as it would be in the case of ARMA(p, q) processes, since $\beta(L)$ multiplies the exogenous variable x_t rather than the error term. Moreover, β_0 is not necessarily zero, thus allowing x_t to have a contemporaneous effect on y_t.

Model (5.2) can be interpreted as a *linear economic system*, where the process x_t represents the *input* flow and y_t the *output* flow. In general, such an *input-output system* can be graphically represented using a *block diagram*, as shown in *Figure 5.1.*

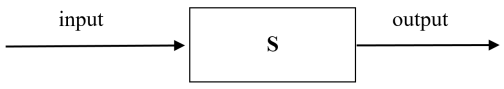


Figure 5.1: *Input-Output System*

In *Figure 5.1*, the block **S** represents the system that transforms the input into the output.

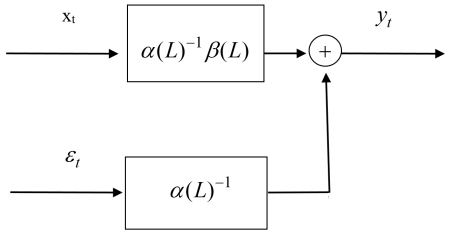


Figure 5.2: *Block diagram of system (5.2)*

The block diagram of system (5.2) has a more complex structure, as shown in *Figure 5.2*: System (5.2) has two inputs: the exogenous variable x_t and the disturbance ε_t, which are combined to produce the output.
Based on this formulation, two key features of the dynamic ADL(p, q) system are typically analyzed:

a) Long-run coefficients (total multipliers);

b) Impulse response functions (dynamic multipliers).

Both the long-run coefficients and the impulse response functions are properties of the *steady state* of the system, which we now define formally.

Definition 5.1. (Steady-State System)
A dynamic system described by model (5.2) is said to be in steady state if the input x_t and the output y_t take constant values over time and no further shocks affect the system (i.e., disturbances are set to zero for the purpose of defining the steady-state relation). In steady state, the dynamic structure of the model remains unchanged, but all transient dynamics have died out, so that the output remains constant as long as the input remains constant.

5.1 Long-run Coefficients

Consider a steady-state system, where the *constant input* and *constant output* are denoted by x and y, respectively. Even in steady state, the system's structure remains unchanged, and the input and output still follow the relations described by the model. The steady-state condition implies a situation of *quiet* or *equilibrium*, and for this reason, the disturbance term is set to zero to avoid perturbing the output. Hence, in steady state, system (5.2) becomes:

$$\alpha(1)y = \beta(1)x. \tag{5.3}$$

[58] Note the similarity with the stationarity condition of ARMA processes; however, the term "stability condition" is preferred here since the model may involve non-stationary variables.

In (5.3), the operator L is replaced by 1, since[59] $Lc = c$ for any constant c.
Solving equation (5.3) yields:

$$y = \frac{\beta(1)}{\alpha(1)}x = \frac{\sum_{i=0}^{q}\beta_i}{1-\sum_{i=1}^{p}\alpha_i}x. \tag{5.4}$$

The ratio $\beta(1)/\alpha(1)$ is referred to as the *long-run coefficient*, and it represents the long-run change in output resulting from a unit change in the input, when the system is at steady state. Letting y^* denote the new output following a unit increase in the input, we have:

$$\begin{aligned} y^* &= \frac{\beta(1)}{\alpha(1)}(x+1) \\ &= \frac{\beta(1)}{\alpha(1)}x + \frac{\beta(1)}{\alpha(1)} \\ &= y + \frac{\beta(1)}{\alpha(1)}. \end{aligned} \tag{5.5}$$

Thus, the long-run coefficient quantifies the ultimate effect on the steady-state output of a unit shock to the steady-state input.[60]
If y_t and x_t are stationary processes, the econometric interpretation of the long-run coefficient and steady-state relationship is straightforward.

In the case of non-stationary processes, we must assume that a linear long-run relation such as (5.3) characterizes the equilibrium configuration of the system, namely

$$y = \frac{\beta(1)}{\alpha(1)}x.$$

In this situation, the deviation

$$u_t = y_t - \frac{\beta(1)}{\alpha(1)}x_t$$

must represent a stationary process with zero mean. In other words, although the individual variables may be non-stationary, their long-run combination remains stable around a constant mean. This can be illustrated by considering the simple representation

$$y_t = \frac{\beta(1)}{\alpha(1)}x_t + u_t, \tag{5.6}$$

[59] This involves an abuse of notation since the lag operator L is treated as a numeric value, but this substitution is operationally valid.

[60] It is important to note that the long-run coefficient is sensitive to the measurement scales of x and y. Therefore, the unit increase should be interpreted according to the scale of the input variable and not as a 1% increase. Likewise, the corresponding change in output is not to be interpreted as a percentage variation but in terms of the output's measurement scale.

where u_t is stationary with $E(u_t) = 0$. The term u_t measures the short-run deviation from the long-run equilibrium.
When the variables are stationary, this relation simply characterizes the steady-state configuration of the dynamic system. In the case of non-stationary variables, additional conditions are required for this relation to represent a valid stochastic equilibrium, as discussed in *Chapter 8*.

5.2 Impulse Response Function

In a steady-state system, *if the input experiences a unit increase at a specific point in time, does the output immediately reach the new steady-state value?*
If not, how long does it take for the output to reach the new steady state?
To answer these questions, we consider again the transfer function in equation (5.2), which can be rewritten as:

$$\begin{aligned} y_t &= \alpha(L)^{-1}\beta(L)x_t + \alpha(L)^{-1}\varepsilon_t \\ &= h(L)x_t + \alpha(L)^{-1}\varepsilon_t \end{aligned} \tag{5.7}$$

If we are interested in the dynamic effects of a unit shock to the exogenous variable on the transition to a new steady state, we neglect the effect of the white noise term and consider the system in the absence of random disturbances:

$$\begin{aligned} y_t^* &= h(L)(x_t + 1) \\ &= h(L)x_t + h(L) \end{aligned} \tag{5.8}$$

The polynomial $h(L)$ may be of finite or infinite order. The latter occurs when $\alpha(L) \neq 1$, i.e., when the system has *feedback effects*. In that case:

$$h(L) = \sum_{k=0}^{\infty} h_k L^k. \tag{5.9}$$

The polynomial $h(L)$ expresses the dynamic effects on the output caused by a unit shock at each lag of the exogenous variable. The coefficients h_k describe how a unit shock at time $t-k$ ($k \geq 0$) propagates to affect the current and future values of the output.
The sequence $\{h_k,\, k = 0, 1, 2, \ldots\}$ is called the *impulse response function*, and it can be represented graphically for each lag k.
The *cumulative impulse response function* is defined as:

$$H_k = \sum_{j=0}^{k} h_j\,, \tag{5.10}$$

which measures the cumulative effect of a unit shock to the input at time t on the output over the lags $\{x_t, x_{t-1}, \ldots, x_{t-k}\}$.

Taking the limit of the cumulative effect yields:

$$\lim_{k\to\infty} H_k = \sum_{j=0}^{\infty} h_j = h(1). \tag{5.11}$$

Note that $h(1) = \frac{\beta(1)}{\alpha(1)}$ is the long-run coefficient. Therefore, the long-run coefficient is the *asymptote* of the cumulative impulse response function (or the *integral function* in continuous time).
The sequence $\{h_k,\, k = 0, 1, 2, \ldots\}$ can be computed assuming that the coefficients of the polynomials are known (or estimated). This is done by solving the identity:

$$(1-\alpha_1 L-\alpha_2 L^2-\cdots-\alpha_p L^p)(h_0+h_1 L+h_2 L^2+\cdots) = \beta_0+\beta_1 L+\beta_2 L^2+\cdots+\beta_q L^q. \tag{5.12}$$

This equality holds if the coefficients of each power of L on both sides of the equation are identical. The h_k coefficients can therefore be obtained recursively:

$$\begin{array}{rccccccccc}
h_0 & + & h_1 L & + & h_2 L^2 & + & h_3 L^3 & + & h_4 L^4 + \cdots \\
 & - & \alpha_1 h_0 L & - & \alpha_1 h_1 L^2 & - & \alpha_1 h_2 L^3 & - & \alpha_1 h_3 L^4 - \cdots \\
 & & & - & \alpha_2 h_0 L^2 & - & \alpha_2 h_1 L^3 & - & \alpha_2 h_2 L^4 - \cdots \\
= & \beta_0 & +\beta_1 L & +\beta_2 L^2 & +\beta_3 L^3 & +\beta_4 L^4 & +\cdots & &
\end{array}$$

From which:

$$\begin{aligned}
h_0 &= \beta_0 \\
h_1 - \alpha_1 h_0 &= \beta_1 \Rightarrow h_1 = \beta_1 + \alpha_1 h_0 \\
h_2 - \alpha_1 h_1 - \alpha_2 h_0 &= \beta_2 \Rightarrow h_2 = \beta_2 + \alpha_1 h_1 + \alpha_2 h_0 \\
h_3 - \alpha_1 h_2 - \alpha_2 h_1 - \alpha_3 h_0 &= \beta_3 \Rightarrow h_3 = \beta_3 + \alpha_1 h_2 + \alpha_2 h_1 + \alpha_3 h_0 \\
&\cdots \\
h_k &= \beta_k + \sum_{j=1}^{k} \alpha_j h_{k-j}
\end{aligned}$$

Orders p and q of the polynomials must be taken into account, so the final formulation of the recursive algorithm becomes:

$$\begin{cases}
h_0 = \beta_0 \\
h_k = \beta_k + \sum_{j=1}^{m} \alpha_j h_{k-j}, \quad k = 1, \ldots, q;\ m = \min(k,\, p) \\
h_k = \sum_{j=1}^{m} \alpha_j h_{k-j}, \qquad\qquad k > q
\end{cases} \tag{5.13}$$

Example 5.1. Let the following model *ADL (2, 1)* be defined by:

$$y_t = 0.4\, y_{t-1} + 0.1\, y_{t-2} + 2\, x_t - 0.5\, x_{t-1} + \varepsilon_t\,. \tag{5.14}$$

The long-run coefficient is 3, since:

$$y = \frac{\beta(1)}{\alpha(1)}x = \frac{\sum_{i=0}^{q}\beta_i}{1-\sum_{i=1}^{p}\alpha_i}x = \frac{2-0.5}{1-0.4-0.1}x = 3x$$

By using the recursive algorithm (5.13), we calculate the $h_k,\ k = 0, 1, 2, \cdots$ values of the impulse response function and the $H_k,\ k = 0, 1, 2, \cdots$ values of the cumulative response function.
The values of these two functions are shown in *Table 5.1* for the first 12 lags. *Figure 5.3* and *Figure 5.4* show their graphs.

k	0	1	2	3	4	5	6
h_k	2	0.3	0.32	0.158	0.0952	0.0539	0.0311
H_k	2	2.3	2.62	2.778	2.8732	2.9271	2.9582
k	7		8	9	10	11	12
h_k	0.0178		0.0102	0.0059	0.0034	0.0019	0.0011
H_k	2.9760		2.9862	2.9921	2.9955	2.9974	2.9985

Table 5.1: First 12 values of h_k and H_k of the model (5.14)

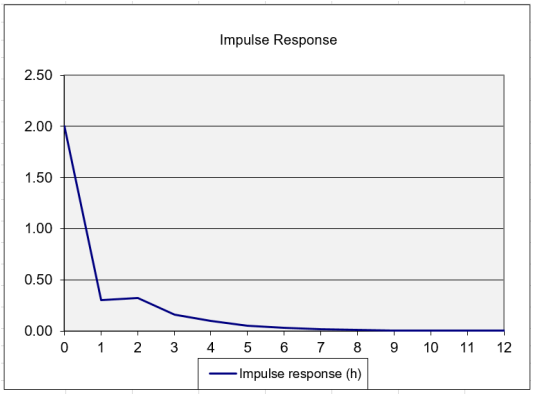


Figure 5.3: *Impulse response of y_t with respect to variable x_t in the model (5.14)*

The values of the impulse response function rapidly converge to zero, while those of the cumulative response converge to the value of the long-run coefficient.

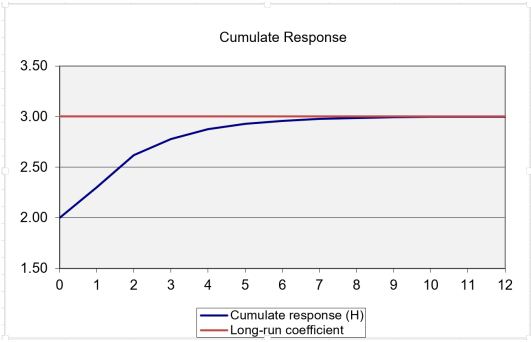


Figure 5.4: *Cumulative response of y_t with respect to variable x_t in the model (5.14)*

The *transition* from one steady state to the next can be illustrated using the same model (5.14).
Suppose, in the first phase, that the system has reached a steady state for both input and output with values $x = \frac{1}{3}$ and $y = 3 \times \frac{1}{3} = 1$.
In the second phase, at time k_H, the input receives a unit step (permanent) shock. Consequently, the new steady-state input becomes $x* = \frac{1}{3} + 1 = \frac{4}{3}$, and the steady-state output should eventually reach $y* = 3 \times \frac{4}{3} = 4$ at time k_H.
However, this new value is not attained immediately at time k_H but only after a sequence of dynamic adjustments described by the impulse and cumulative response functions. The autoregressive structure of the output induces a feedback effect that dies out only asymptotically as the lag increases. In other words, the entire infinite history of the input is required for the output to complete its transition to the new steady state. By contrast, in the absence of an autoregressive component in the ADL representation—that is, when $\alpha(L) = 1$ and the model does not embed feedback dynamics—the new steady state is reached within a finite number of periods, since $h(L) = \beta(L)$ is then a finite-order polynomial. When feedback effects are present (i.e. $\alpha(L) \neq 1$), the adjustment typically unfolds over an infinite horizon and convergence to the new steady state occurs only asymptotically, provided that the system is stable. This distinction will become particularly relevant when non-stationary variables are considered (*Chapter 8*).
From time k_H onward, the red curve traces the path of the output as it moves from the initial steady state to the new one. This trajectory corresponds exactly to the cumulative response curve depicted in *Figure 5.4*. Theoretically, the output reaches the new steady-state value of 4 only as $k \to \infty$ (i.e., as the horizon increases after the shock).
In *Figure 5.5*, the unit impulse to the input at $k_H = 5$ increases its level from 1/3 to 4/3 (blue curve). The immediate impact on the output is given by adding the cumulative value $H_0 = 2$ (from *Table 5.1*) to the pre-shock steady-state output $y = 1$, resulting in a new level of 3 (red curve).

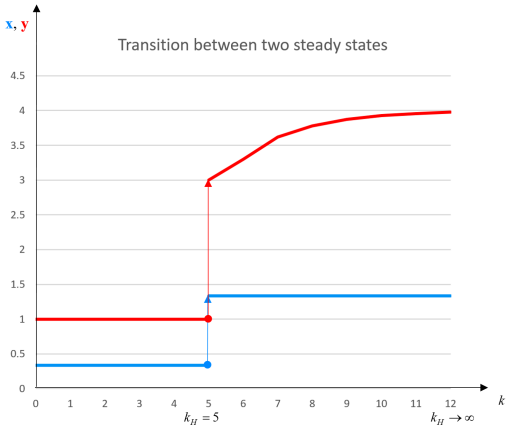


Figure 5.5: *Transition between two steady states for model (5.14)*

Remark 5.1. In concluding this chapter, it is useful to clarify that the notion of a *steady-state system* does not imply that the underlying processes (input and output) are stationary. The stability condition imposed on the dynamic system is compatible with the presence of non-stationary variables. Stability is therefore a necessary, but not sufficient, condition for the existence of a steady state. This distinction will become particularly relevant in *Chapter 7*.
According to *Definition 5.1*, the statement that *the dynamic structure of the model remains unchanged* means that the system reacts to shocks through a dynamic adjustment mechanism. The effects of a shock are not exhausted instantaneously, nor does the system jump directly to its final long-run outcome. Instead, dynamics describe the smooth transition path through which the system converges to its new steady state, as captured by impulse response functions and cumulative responses.

6 Prediction

Forecasting[61] methods are manifold. We do not aim to provide an exhaustive overview, but rather highlight some aspects, with particular attention given to quantitative methods.[62]
A useful distinction can be made between *adaptive* methods and *regression-based* forecasting methods.
Adaptive methods are characterized by simplicity and immediacy, and they are often used as benchmarks for more complex methods such as those based on regression.
Among adaptive methods, examples include the so-called *naive* methods and *exponential smoothing*.
As for regression-based forecasting methods, we include predictions derived from econometric models and those based on stochastic processes. The *Kalman filter* can also be classified among regression-based methods, as it refers to models formulated in *state-space* form.

Example 6.1. *(Naive method)*
Let $\hat{x}_t$ be the predicted value at time t for the variable x_t, such as a company's sales. This value may be defined as:

$$\hat{x}_t = x_{t-4}\frac{x_{t-1}}{x_{t-5}}\,, \tag{6.1}$$

where:

x_{t-4} represents the sales from the same quarter of the previous year;

$\frac{x_{t-1}}{x_{t-5}}$ is an adjustment factor applied to x_{t-4} under the assumption of proportionality between the previous quarter and the same quarter of the previous year.

Example 6.2. *(Exponential Smoothing)*
The exponential smoothing method is based on the hypothesis that recent values of x_t have a stronger influence on $\hat{x}_t$ than more distant ones[63]:

$$\begin{aligned}\hat{x}_t &= \alpha x_{t-1} + \alpha(1-\alpha)x_{t-2} + \alpha(1-\alpha)^2 x_{t-3} + \cdots \\ &= \sum_{k=0}^{\infty} \alpha(1-\alpha)^k x_{t-k-1}\end{aligned}\,.$$

[61] Here, we use the terms *forecast* and *prediction* interchangeably.
[62] For a general overview of forecasting methods, see for instance the open-access textbook Hyndman and Athanasopoulos (2021).
For a more quantitative approach, see: Chatfield (2000).
[63] This model is widely known in the literature by the acronym *EWMA (Exponentially Weighted Moving Average)*.

This formula represents a weighted arithmetic average with exponentially decreasing weights. The value of α must lie in the open interval (0, 1), ensuring:

$$\sum_{k=0}^{\infty} \alpha(1-\alpha)^k = \frac{\alpha}{1-(1-\alpha)} = 1\,.$$

Observe that:

$$\hat{x}_{t-1} = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k x_{t-k-2}\,,$$

and multiplying both sides by $(1-\alpha)$ gives:

$$(1-\alpha)\hat{x}_{t-1} = \sum_{k=0}^{\infty} \alpha(1-\alpha)^{k+1} x_{t-k-2}\,,$$

thus, the difference yields:

$$\begin{aligned}
\hat{x}_t - (1-\alpha)\hat{x}_{t-1} &= \sum_{k=0}^{\infty} \alpha(1-\alpha)^k x_{t-k-1} - \sum_{k=0}^{\infty} \alpha(1-\alpha)^{k+1} x_{t-k-2} \\
&= \alpha x_{t-1} + \sum_{k=0}^{\infty} \alpha(1-\alpha)^{k+1} x_{t-k-2} - \sum_{k=0}^{\infty} \alpha(1-\alpha)^{k+1} x_{t-k-2}\,, \\
&= \alpha x_{t-1}
\end{aligned}$$

therefore:

$$\hat{x}_t = \alpha x_{t-1} + (1-\alpha)\hat{x}_{t-1}\,. \tag{6.2}$$

This latter expression defines a recursive relationship between predicted values and operates under the initial condition $\hat{x}_1 = x_1$.
The choice of α is crucial. In fact, if $\alpha = 1$, then $\hat{x}_t = x_{t-1}$. On the other hand, as $\alpha \to 0$, more weight is given to the entire history of the x_t process.
Forecasting via exponential smoothingis not in contrast with prediction based on stochastic models: it coincides with the prediction of an *ARIMA(0,1,1)* model, i.e., a random walk model with a first-order moving average component. The *ARIMA(0,1,1)* model is defined as:

$$\Delta x_t = \varepsilon_t + \theta\varepsilon_{t-1}\,,$$

where ε_t is a white noise process.
The optimal linear predictor[64] is:

$$\hat{x}_t = E(x_t|\Im_{t-1}) = x_{t-1} + \theta\varepsilon_{t-1}\,,$$

where $\Im_{t-1}$ denotes the available *information set* on the process x_t up to time $t-1$.

[64] The concept of optimality refers to the minimization of the prediction-error variance. See §6.3.1.

From the definition of the optimal predictor, we obtain with a few algebraic steps:

$$x_t - \hat{x}_t = \varepsilon_t \,,$$

and therefore:

$$\hat{x}_t = x_{t-1} + \theta(x_{t-1} - \hat{x}_{t-1}) = (1+\theta)x_{t-1} - \theta\hat{x}_{t-1} \,.$$

By defining $\alpha = 1 + \theta$, we finally obtain:

$$\hat{x}_t = \alpha x_{t-1} + (1-\alpha)\hat{x}_{t-1} \,.$$

6.1 Prediction for Econometric Models

Consider the single-equation regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,,$$

where $\boldsymbol{\varepsilon} \sim WN(\mathbf{0},\, \sigma^2\mathbf{I})$.
It is well known that, under the stated assumptions on $\boldsymbol{\varepsilon}$, the best linear unbiased estimator (BLUE) of the parameters $\boldsymbol{\beta}$ is obtained via the ordinary least squares (OLS) method. We denote these estimates as $\hat{\boldsymbol{\beta}}$. The best-fitted regression curve is thus given by: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
Now suppose that the future values of the explanatory variables $\mathbf{X}$ are known and denote them by $\mathbf{X}_f$. If the model assumptions continue to hold in the future, we may assume that the future values of the dependent variable follow the same probabilistic structure:

$$\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \boldsymbol{\varepsilon}_f \,, \tag{6.3}$$

where $\boldsymbol{\varepsilon}_f \sim WN(\mathbf{0},\, \sigma^2\mathbf{I})$ is the theoretical prediction error.
Operationally, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}_f$ are unknown.
Since, in general, forecasts are given by the conditional expected value with respect to the available information set, the optimal theoretical prediction[65] is:

$$E(\mathbf{y}_f|\Im) = \mathbf{X}_f\boldsymbol{\beta} \,,$$

where $\Im = \{\mathbf{X}, \mathbf{X}_f\}$ is the information set[66], and $E(\boldsymbol{\varepsilon}_f|\Im) = 0$ by assumption.
Replacing the unknown parameter $\boldsymbol{\beta}$ with its OLS estimate $\hat{\boldsymbol{\beta}}$, the practical prediction becomes:

$$\hat{\mathbf{y}}_f = \mathbf{X}_f\hat{\boldsymbol{\beta}} \,,$$

[65] Optimality refers to the minimization of the prediction-error variance. See §6.3.1.
[66] Here, the information set has a stacked data structure rather than a time-sequential one, and therefore it is not indexed by t.

which is unbiased in the sense that[67]:

$$E(\hat{\boldsymbol{\varepsilon}}_f|\Im) = E(\mathbf{y}_f - \hat{\mathbf{y}}_f|\Im) = E(\mathbf{X}_f\boldsymbol{\beta} + \boldsymbol{\varepsilon}_f - \mathbf{X}_f\hat{\boldsymbol{\beta}}|\Im)$$
$$= E(\boldsymbol{\varepsilon}_f|\Im) + E\left[\mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|\Im\right] = \mathbf{X}_f E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|\Im) = 0\,.$$

The last equality holds due to the unbiasedness of the OLS estimator.
The computed prediction error $\hat{\boldsymbol{\varepsilon}}_f$ consists of two components:

$$\hat{\boldsymbol{\varepsilon}}_f = \mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}_f\,.$$

The first term reflects the error introduced by using estimated instead of true parameters; the second is the inherent stochastic disturbance.
The prediction error covariance matrix is:

$$E(\hat{\boldsymbol{\varepsilon}}_f\hat{\boldsymbol{\varepsilon}}'_f|\Im) = E\left[\left(\mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}_f\right)\left(\mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}_f\right)'|\Im\right]$$
$$= E\left[\mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'_f\right] + E\left[\boldsymbol{\varepsilon}_f\boldsymbol{\varepsilon}'_f\right]$$
$$= \sigma^2\left(\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_f + \mathbf{I}\right)$$

Cross-product terms vanish under the assumption $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'_f|\Im) = 0$, since:

$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\,,$$

and thus:

$$E\left[\mathbf{X}_f(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{\varepsilon}'_f|\Im\right] = \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'_f|\Im) = \mathbf{0}\,.$$

6.2 Prediction Intervals

Focusing on a single-step-ahead forecast , let $\mathbf{x}_f$ denote the column vector of the future values of the explanatory variables. The associated prediction error is:

$$\hat{\varepsilon}_f = y_f - \mathbf{x}'_\mathbf{f}\hat{\boldsymbol{\beta}}\,,$$

with variance:

$$\sigma^2\left(\mathbf{x}'_\mathbf{f}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f + 1\right)\,.$$

Assuming normally distributed errors, the standardized prediction error

$$\frac{y_f - \mathbf{x}'_\mathbf{f}\hat{\boldsymbol{\beta}}}{\sigma\sqrt{\mathbf{x}'_\mathbf{f}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f + 1}} \tag{6.4}$$

[67] Note that the predictor $\hat{\mathbf{y}}_f$ is unbiased not because it matches the future value $\mathbf{y}_f$, but because the expected prediction error is zero.

follows a standard normal distribution.
Since σ is unknown, it is replaced by its estimator $\hat{\sigma}$, which alters the distribution of the above ratio. It no longer follows the standard normal, but rather a Student's t-distribution with $T-k$ degrees of freedom (here k represents the number of estimated parameters). This result stems from the fact that the quantity

$$\frac{(T-k)\hat{\sigma}^2}{\sigma^2}$$

follows a chi-squared distribution with $T-k$ degrees of freedom and is independent of the standard normal variable in (6.4).
Given that

$$t_{T-k} = \frac{Z}{\sqrt{\chi^2_{T-k}/(T-k)}}, \quad \text{with } Z \sim N(0,1)\,,$$

and considering:

a) $Z = \dfrac{y_f - \mathbf{x'_f}\hat{\boldsymbol{\beta}}}{\sigma\sqrt{\mathbf{x'_f}(\mathbf{X'X})^{-1}\mathbf{x}_f + 1}}$,

b) $\dfrac{(T-k)\hat{\sigma}^2}{\sigma^2} = \chi^2_{T-k}$,

we obtain:

$$t_{T-k} = \frac{Z}{\sqrt{\dfrac{\chi^2_{T-k}}{T-k}}} = \frac{\dfrac{y_f - \mathbf{x'_f}\hat{\boldsymbol{\beta}}}{\sigma\sqrt{\mathbf{x'_f}(\mathbf{X'X})^{-1}\mathbf{x}_f + 1}}}{\sqrt{\dfrac{(T-k)\hat{\sigma}^2}{\sigma^2(T-k)}}}$$

$$= \frac{y_f - \mathbf{x'_f}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{\mathbf{x'_f}(\mathbf{X'X})^{-1}\mathbf{x}_f + 1}}$$

Let α be the probability that the random variable t_{T-k} falls outside the interval $\left[-t_{\alpha/2},\, t_{\alpha/2}\right]$. Then, the prediction intervals at the $(1-\alpha)$ confidence level are:

$$\begin{aligned}&\Pr\left[\mathbf{x'_f}\hat{\boldsymbol{\beta}} - t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x'_f}(\mathbf{X'X})^{-1}\mathbf{x}_f + 1} \le y_f\right.\\ &\left.\le \mathbf{x'_f}\hat{\boldsymbol{\beta}} + t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x'_f}(\mathbf{X'X})^{-1}\mathbf{x}_f + 1}\right] = 1-\alpha\,.\end{aligned} \tag{6.5}$$

Unlike confidence intervals —in which only the bounds of the inequality are random —prediction intervals also treat the value y_f as a random variable, since it contains a stochastic error component, as defined in equation (6.3)[68].

[68] Prediction intervals for univariate regressions in the *Error Correction Mechanism (ECM)* form are discussed in §8.10.

6.3 Prediction for Stochastic Linear Processes

Let us define:

x_t: a discrete stochastic process, stationary in covariance;

x_{t+h}: a continuous random variable characterized by a probability density function;

$\hat{x}_{t+h}$: the expected value of the variable x_{t+h} at time t for horizon h;

$\Im_t$: the information setavailable at time t (typically $\Im_t = \{x_{t-j}\,;\, j = 0,\, 1,\, \ldots,\, N\}$).

Suppose that the conditional probability density function of x_{t+h}, given the information set $\Im_t$, is known and defined by:

$$\Pr\{x < x_{t+h} \leq x + dx \,|\, \Im_t\} = f_{t,h}(x)\, dx\,,$$

then, in general, all its moments are also defined. For example, the first moment is given (over the domain D of the random variable x_{t+h}) by:

$$E_t(x_{t+h}) = E(x_{t+h} \mid \Im_t) = \int_D x\, f_{t,h}(x)\, dx\,.$$

The definition of the variance of x_{t+h} follows accordingly.

6.3.1 Optimal Prediction

To obtain an optimal prediction, it is useful to introduce the concept of a *Loss Function (or Cost Function)*[69], which is a function of the prediction error.
Let the prediction error be denoted by:

$$\hat{\varepsilon}_{t+h} = x_{t+h} - \hat{x}_{t+h}\,.$$

The loss function $C(\hat{\varepsilon}_{t+h})$ is assumed to satisfy the following properties:

a) $C(0) = 0$

b) $C(\hat{\varepsilon}_{t+h_1}) \geq C(\hat{\varepsilon}_{t+h_2})$ if $\hat{\varepsilon}_{t+h_1} > \hat{\varepsilon}_{t+h_2} > 0$

c) $C(\hat{\varepsilon}_{t+h_1}) \geq C(\hat{\varepsilon}_{t+h_2})$ if $\hat{\varepsilon}_{t+h_1} < \hat{\varepsilon}_{t+h_2} < 0$

Note that the loss function always takes non-negative values and is not necessarily symmetric, as illustrated in *Figure 6.1.*
In the literature, a *quadratic loss function* is often chosen due to its useful analytical properties. It is defined by:

$$C(\hat{\varepsilon}) = k\hat{\varepsilon}^2, \quad k > 0\,,$$

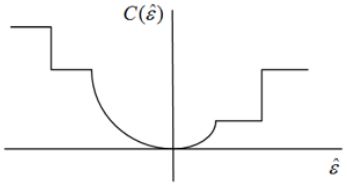


Figure 6.1: Loss Function

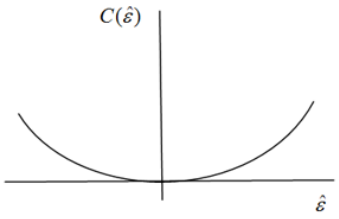


Figure 6.2: Quadratic Loss Function

and its graph is a parabola, as shown in *Figure 6.2.*
This function is symmetric and satisfies properties a), b), and c). The following decision rule is adopted:

We choose the prediction that minimizes the loss function.

The prediction that satisfies this rule is called the *optimal prediction.*
In the case of the quadratic loss function, the optimal prediction satisfies:

$$\begin{aligned} E_t\left[k(x_{t+h}-\hat{x}_{t+h})^2\right] &= \min\,, \quad k>0\,, \\ E_t\left[(x_{t+h}-\hat{x}_{t+h})^2\right] &= \min\,. \end{aligned} \tag{6.6}$$

The second condition follows from the fact that the minimum of the expectation does not depend on the positive constant k.
We now show that $E_t(x_{t+h})$ satisfies this condition. Consider:

$$\begin{aligned} E_t\left[x_{t+h}-E_t(x_{t+h})+E_t(x_{t+h})-\hat{x}_{t+h}\right]^2 = {} & E_t\left[x_{t+h}-E_t(x_{t+h})\right]^2 \\ & + E_t\left[E_t(x_{t+h})-\hat{x}_{t+h}\right]^2 \\ & + 2E_t\left[(x_{t+h}-E_t(x_{t+h}))\,(E_t(x_{t+h})-\hat{x}_{t+h})\right]\,. \end{aligned}$$

Now,

$$E_t\left[(x_{t+h}-E_t(x_{t+h}))\,(E_t(x_{t+h})-\hat{x}_{t+h})\right] = (E_t(x_{t+h})-\hat{x}_{t+h})\,E_t\left[x_{t+h}-E_t(x_{t+h})\right] = 0\,,$$

[69] This section follows the setting of §4.2 in Granger and Newbold (1977).

because the factor $(E_t(x_{t+h}) - \hat{x}_{t+h})$ does not depend on the expected conditional operator and $E_t\left[x_{t+h} - E_t(x_{t+h})\right] = 0$ is true by the property of the mean.
Hence,

$$E_t\left[(x_{t+h} - E_t(x_{t+h}))^2\right] + E_t\left[(E_t(x_{t+h}) - \hat{x}_{t+h})^2\right]$$

is minimized if and only if $\hat{x}_{t+h} = E_t(x_{t+h})$, since the second term is zero and the first is the minimum variance attainable for a random variable.

Therefore, if the loss function is quadratic, the optimal predictor is the conditional expectation of the variable given the past of the process.

Granger (1969) demonstrates that this result holds even for many non-quadratic loss functions, provided both the loss function and the probability distribution of x_{t+h} are symmetric.
In general, $E_t(x_{t+h})$ is not necessarily a linear function of the values in $\Im_t$, but it is if the stochastic process is Gaussian.
When the process is not Gaussian, linearity may be imposed, leading to sub-optimal predictions[70].

6.3.2 Linear Prediction with Minimum Mean Square Error

Suppose that the random variable x_{t+h} depends linearly on the past k values of the process up to time t:

$$x_{t+h} = \alpha_0 x_t + \alpha_1 x_{t-1} + \cdots + \alpha_k x_{t-k} + \varepsilon_{t+h,\,k}\,.$$

The linear predictor is defined by:

$$\hat{x}_{t+h,\,k} = \hat{\alpha}_0 x_t + \hat{\alpha}_1 x_{t-1} + \cdots + \hat{\alpha}_k x_{t-k}\,.$$

The coefficients $\hat{\alpha}_j$, for $j = 0, 1, \ldots, k$, are chosen to minimize the mean square error:

$$\min_{\alpha} E_t\left[(x_{t+h} - \hat{x}_{t+h,\,k})^2\right] = \min_{\alpha} E_t[\hat{\varepsilon}^2_{t+h,\,k}]\,.$$

These coefficients[71] generally depend on:

- the time index t;
- the forecast horizon h;
- the number of regressors k.

[70] According to the Wold decomposition theorem , a predictor that is linear in past values can always be constructed for a covariance-stationary process with an arbitrarily small approximation error.
[71] We refer to the theoretical coefficients that minimize the expected squared error, not to estimated parameters.

It is interesting to observe the asymptotic behavior of the forecast error when $k \to \infty$ for a covariance-stationary process.
Denoting the limit forecast and forecast error by $\hat{x}_{t+h}$ and $\hat{\varepsilon}_{t+h}$ respectively, we write:

$$\hat{x}_{t+h} = \lim_{k\to\infty} \sum_{j=0}^{k} \hat{\alpha}_j x_{t-j}\,, \quad \text{so that} \quad x_{t+h} = \hat{x}_{t+h} + \hat{\varepsilon}_{t+h}\,.$$

The variance of the forecast error in the limit is:

$$\mathrm{Var}(\hat{\varepsilon}_{t+h}) = E_t\left[(x_{t+h} - \hat{x}_{t+h})^2\right] = \lim_{k\to\infty} E_t\left(\hat{\varepsilon}^2_{t+h,\,k}\right) \geq 0\,. \tag{6.7}$$

For covariance-stationary processes, we introduce the following definitions[72]:

- *Purely deterministic (or singular) process* if $\mathrm{Var}(\hat{\varepsilon}_{t+h}) = 0$, in which case the process is *perfectly predictable*;
- *Purely non-deterministic stochastic process* if $\mathrm{Var}(\hat{\varepsilon}_{t+h}) > 0$, meaning that the forecast error retains a nonzero variance even with access to the infinite past.

6.3.3 Optimal Prediction for Purely Non-Deterministic Linear Processes

According to the Wold decomposition theorem , a purely non-deterministic stochastic process can be written as:

$$x_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j} = \sum_{j=0}^{\infty} \beta_j L^j \varepsilon_t = \beta(L)\varepsilon_t, \quad \beta_0 = 1\,, \tag{6.8}$$

where ε_t is a white noise process with zero mean and variance σ^2_ε.
Assuming the polynomial $\beta(L)$ is invertible, we define the infinite-order polynomial $\alpha(L) = \beta^{-1}(L) = \sum_{j=0}^{\infty} \alpha_j L^j$ with $\alpha_0 = 1$, so that the process admits the autoregressive representation:

$$\alpha(L)x_t = \varepsilon_t\,.$$

For a given *forecast horizon* h, the assumption of linearity and optimality leads to a linear predictor of the form:

$$\hat{x}_{t,h} = \sum_{j=0}^{\infty} c_j x_{t-j} = c(L)x_t\,, \tag{6.9}$$

such that:

$$S = E\left[(x_{t+h} - c(L)x_t)^2\right] = \min\,.$$

[72] See also the Wold Decomposition Theorem in §3.1.

Using the definition of the process in (6.8), we can express the predictor as:

$$\hat{x}_{t,h} = c(L)\beta(L)\varepsilon_t = d(L)\varepsilon_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} \,, \tag{6.10}$$

so that:

$$S = E\left[\sum_{j=0}^{\infty} \beta_j \varepsilon_{t+h-j} - \sum_{j=0}^{\infty} d_j \varepsilon_{t-j}\right]^2 = \min \,. \tag{6.11}$$

If the coefficients β_j are known, the coefficients d_j of the linear predictor can be determined.
For $h = 0$, the minimum of (6.11) is clearly attained when $\beta_j = d_j$ for all j.
For $h > 0$, the process can be decomposed as:

$$x_{t+h} = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+h-j} = \sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j} + \sum_{j=0}^{\infty} \beta_{j+h} \varepsilon_{t-j} \,. \tag{6.12}$$

Substituting into the expression for S gives:

$$\begin{aligned} S &= E\left[\sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j} + \sum_{j=0}^{\infty} \beta_{j+h} \varepsilon_{t-j} - \sum_{j=0}^{\infty} d_j \varepsilon_{t-j}\right]^2 \\ &= E\left[\sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j}\right]^2 + E\left[\sum_{j=0}^{\infty} (\beta_{j+h} - d_j) \varepsilon_{t-j}\right]^2 \\ &= \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \beta_j^2 + \sigma_\varepsilon^2 \sum_{j=0}^{\infty} (\beta_{j+h} - d_j)^2 = \min \,. \end{aligned} \tag{6.13}$$

The cross-products cancel out because the autocovariances of white noise are zero. The expression (6.13) is minimized if and only if:

$$d_j = \beta_{j+h} \,. \tag{6.14}$$

Thus, the optimal linear predictor is:

$$\hat{x}_{t,h} = \sum_{j=0}^{\infty} \beta_{j+h} \varepsilon_{t-j} \,. \tag{6.15}$$

Using the decomposition in (6.12), the forecast error becomes:

$$\hat{\varepsilon}_{t,h} = x_{t+h} - \hat{x}_{t,h} = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+h-j} - \sum_{j=0}^{\infty} \beta_{j+h} \varepsilon_{t-j} = \sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j} \,. \tag{6.16}$$

This shows that the forecast error follows an $MA(h-1)$ process. Its mean and variance are:

$$\begin{aligned} E(\hat{\varepsilon}_{t,h}) &= 0\,, \\ Var(\hat{\varepsilon}_{t,h}) &= \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \beta_j^2\,. \end{aligned} \tag{6.17}$$

For $h = 1$, the one-step-ahead forecast error reduces to:

$$\hat{\varepsilon}_{t,1} = x_{t+1} - \hat{x}_{t,1} = \varepsilon_{t+1}\,, \tag{6.18}$$

that is, a white noise process with zero mean and constant variance. Even with infinite past information, the forecast error variance cannot be reduced further for $h = 1$.
As the forecast horizon increases, the variance of the forecast error increases:

$$Var(\hat{\varepsilon}_{t,h+1}) - Var(\hat{\varepsilon}_{t,h}) = \sigma_\varepsilon^2 \beta_h^2 \geq 0\,, \tag{6.19}$$

which shows that the *sequence of forecast error variances is monotonically non-decreasing.* Moreover,

$$\lim_{h\to\infty} Var(\hat{\varepsilon}_{t,h}) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \beta_j^2 = \sigma_x^2\,,$$

i.e., for $h \to \infty$ the unexplained variance coincides with the total variance of the process. The covariance between forecast errors can be computed in two ways:

a) For errors at the same horizon but from different starting points:

$$\begin{aligned} E(\hat{\varepsilon}_{t,h}\hat{\varepsilon}_{t+k,h}) &= E\left[\sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j} \sum_{i=0}^{h-1} \beta_i \varepsilon_{t+h+k-i}\right] \\ &= \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \beta_j \beta_{j+k}\,, \quad k > 0 \end{aligned} \tag{6.20}$$

b) For errors from the same starting point but with different horizons: forecast horizons and equal time origin:

$$\begin{aligned} E(\hat{\varepsilon}_{t,h}\hat{\varepsilon}_{t,h+k}) &= E\left[\sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j} \sum_{i=0}^{h+k-1} \beta_i \varepsilon_{t+h+k-i}\right] \\ &= \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \beta_j \beta_{j+k}\,, \quad k > 0 \end{aligned} \tag{6.21}$$

These results highlight that forecast errors are generally correlated across time; this may cause predictions to systematically overestimate or underestimate actual values.

Since prediction is a conditional expectation $E(x_{t+h}|\Im_t)$, the variance of the process at time $t+h$ is always at least as large as the variance of the prediction. In fact:

$$\begin{aligned} Var(x_{t+h}) &= Var(\hat{x}_{t,h} + \hat{\varepsilon}_{t,h}) \\ &= Var(\hat{x}_{t,h}) + Var(\hat{\varepsilon}_{t,h}) + 2\,Cov(\hat{x}_{t,h}, \hat{\varepsilon}_{t,h})\,. \end{aligned} \tag{6.22}$$

For $h > 0$, we have:

$$Cov(\hat{x}_{t,h}, \hat{\varepsilon}_{t,h}) = E\left(\sum_{j=0}^{\infty} \beta_{j+h}\varepsilon_{t-j} \cdot \sum_{j=0}^{h-1} \beta_j \varepsilon_{t+h-j}\right) = 0\,,$$

because the two sums involve uncorrelated white noise terms.
Therefore:

$$Var(x_{t+h}) \geq Var(\hat{x}_{t,h})\,, \tag{6.23}$$

with equality only if $Var(\hat{\varepsilon}_{t,h}) = 0$, i.e., in the purely deterministic case.
If we further assume that the white noise process ε_t is Gaussian, then:

$$x_{t+h} - \hat{x}_{t,h} \sim \mathcal{N}\left(0,\, Var(\hat{\varepsilon}_{t,h})\right)\,. \tag{6.24}$$

As a result, the prediction interval is given by:

$$x_{t+h}(\pm) = \hat{x}_{t,h} \pm z_{\alpha/2}\left\{\sum_{j=0}^{h-1} \beta_j^2\right\}^{1/2} \sigma_\varepsilon\,, \tag{6.25}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Example 6.3. ***(Prediction intervals for an AR(2) stationary process)***
Consider the *AR(2)* process:

$$x_t = 0.2x_{t-1} - 0.35x_{t-2} + \varepsilon_t\,, \quad \varepsilon_t \sim \text{NID}(0,\, 4)\,.$$

The process is stationary in covariance since the roots of the characteristic equation

$$1 - 0.2L + 0.35L^2 = 0$$

are $-10/7$ and 2, both of which lie outside the unit circle[73].
The β_j coefficients can be computed using the following recursive formula:

$$\beta_j = \alpha_1\beta_{j-1} + \alpha_2\beta_{j-2}, \quad j = 2,\, 3,\, \ldots\,, \tag{6.26}$$

with initial conditions $\beta_0 = 1$ and $\beta_1 = \alpha_1$.
Table 6.1 reports the first 30 computed values of the β_j coefficients.

[73] In this case, since the roots are real (rather than complex), it suffices to verify they lie outside the unit interval $(-1,\, 1)$.

Table 6.1: β_j coefficients for the *AR(2)* process of *Example 6.3*

β_j Coefficients					
j	β_j	j	β_j	j	β_j
0	1	11	-0.00281	21	-9.3E-06
1	0.2	12	-0.00111	22	7.07E-06
2	-0.31	13	0.000763	23	4.66E-06
3	-0.132	14	0.000541	24	-1.5E-06
4	0.0821	15	-0.00016	25	-1.9E-06
5	0.06262	16	-0.00022	26	1.51E-07
6	-0.01621	17	1.14E-05	27	7.09E-07
7	-0.02516	18	7.97E-05	28	8.88E-08
8	0.000642	19	1.19E-05	29	-2.3E-07
9	0.008934	20	-2.5E-05	30	-7.7E-08
10	0.001562				

Given the stationarity of the process, the values of β_j decrease rapidly and tend to zero as j increases. This behavior is clearly visible in *Figure 6.3*, which displays the first 50 values.
The calculated β_j coefficients allow for the construction of theoretical prediction intervals using formula (6.25) for a chosen confidence level α. These are shown in *Figure 6.4*.
We observe that the 95% prediction intervals are wider than the 90% intervals. Furthermore, the intervals widen as the forecast horizon increases. For instance, when $h = 1$, the difference between the two interval widths is $15.68 - 13.2 = 2.48$. When $h = 11$, the difference increases to $17.49 - 14.81 = 2.78$.

Example 6.4. *(Prediction intervals for an AR(2) process with a unit root)*[74]
It is interesting to analyze how the presence of a unit root, i.e., non-stationarity, affects prediction intervals.
Consider the process:

$$(1 - L)(1 + \alpha L)x_t = \varepsilon_t \,,$$

which expands to:

$$\left[1 - (1 - \alpha)L - \alpha L^2\right] x_t = \varepsilon_t \,,$$

or:

$$x_t = (1 - \alpha)x_{t-1} + \alpha x_{t-2} + \varepsilon_t \,.$$

[74] This example is included for coherence with the topic of prediction and anticipates concepts that will be formally introduced in Chapter 7. The reader may skip this example and return to it after reading that chapter.

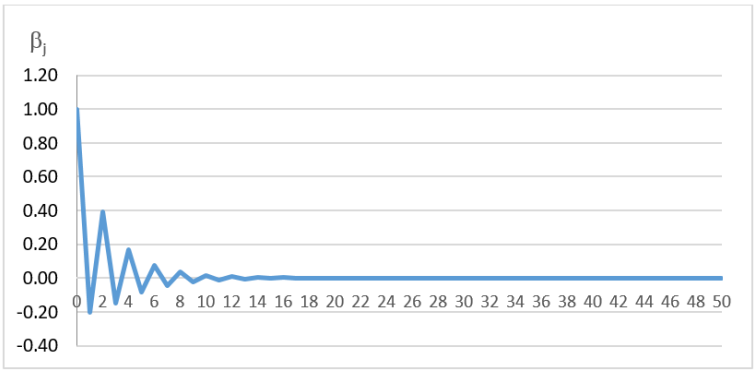


Figure 6.3: β_j coefficients for the *AR(2)* process of *Example 6.3*

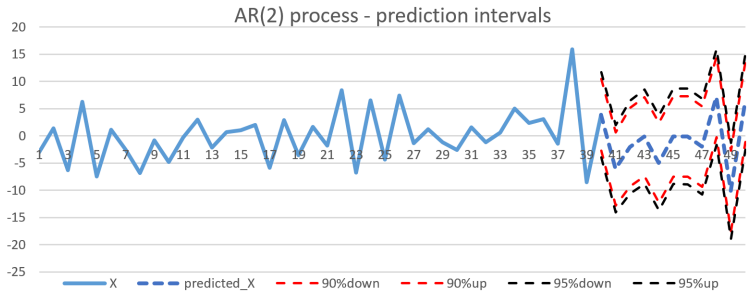


Figure 6.4: Prediction intervals for the *AR(2)* process of Example 6.3

The presence of a unit root is evident. The second root is $-1/\alpha$, which lies outside the unit interval for $|\alpha| < 1$.

The coefficients β_j can be calculated using the recursive formula (6.26). In this case, the sequence does not converge to zero, but instead to a constant value, which can be computed theoretically[75].

[75] For an $AR(p)$ process with a unit roots, the behavior of the impulse response coefficients $\{\beta_j\}$ differs sharply from the stationary case. If the process has *one unit root*, then $\{\beta_j\}$ converges to a constant, equal to $\dfrac{1}{1-\sum_{i=1}^{p}\alpha_i}$, so that shocks have permanent but bounded effects. If the process has *two unit roots*, then $\{\beta_j\}$ grows linearly with j, implying that shocks generate effects that increase without bound over time.
See Hamilton (1994), p. 50, Beveridge and Nelson (1981), p. 154, and Proietti (2006), p. 2235 for formal derivations and discussions of the long-run effect of shocks in the presence of one or two unit roots.

From the recurrence:

$$\begin{aligned}
\beta_0 &= 1\,,\\
\beta_1 &= 1-\alpha\,,\\
\beta_2 &= 1-\alpha+\alpha^2\,,\\
\beta_3 &= 1-\alpha+\alpha^2-\alpha^3\,,\\
&\vdots\\
\beta_k &= \sum_{j=0}^{k}(-\alpha)^j\,,
\end{aligned}$$

we see that as $k \to \infty$, the sum converges to:

$$\sum_{j=0}^{\infty}(-\alpha)^j = \frac{1}{1+\alpha}\,.$$

For example, with $\alpha = 0.4$, the *AR(2) process with a unit root* becomes[76]:

$$x_t = 0.6x_{t-1} + 0.4x_{t-2} + \varepsilon_t\,.$$

In this case, the β_j sequence converges to the limit $\dfrac{1}{1.4} \approx 0.625$ as $j \to \infty$. The convergence is illustrated in *Figure 6.5*, where a good approximation of the limit is reached around $j = 30$.

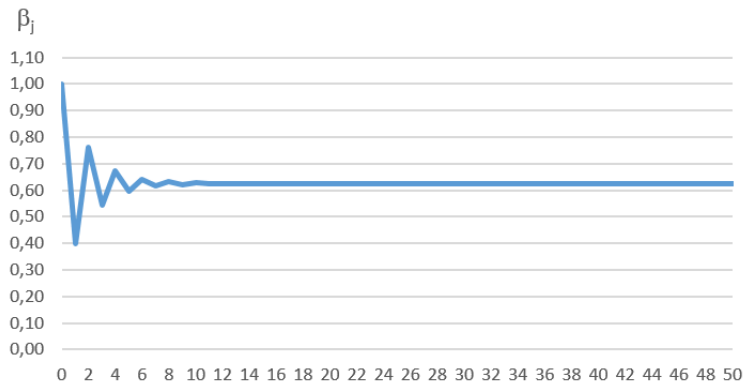


Figure 6.5: Coefficients β_j for $j \to \infty$

Figure 6.6 illustrates the prediction intervals for the *AR(2)* process with a unit root, clearly showing the substantial widening of the prediction intervals caused by the presence of the unit root.

[76] This process is referred to in §7.2 as an *ARIMA(2,1,0)* process.

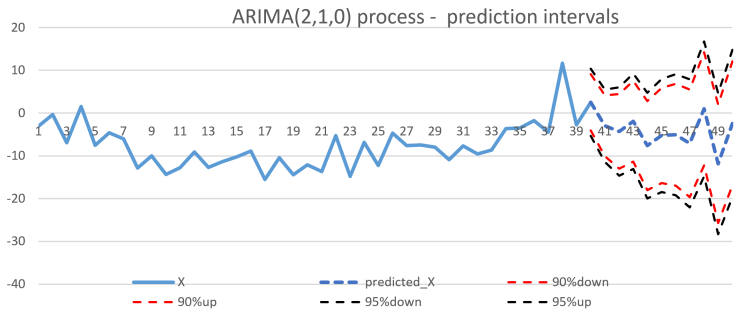


Figure 6.6: Prediction intervals for the *AR(2)* process with a unit root

6.3.4 Prediction Memory

For a stationary stochastic process, it is important to define the memory of its predictor[77]. As shown in equation (6.7), as the number of regressors increases, the variance of the forecast error tends to a non-zero lower bound if the process is purely non-deterministic (or regular).
For convenience of notation, we define:

$$\begin{aligned} V_{h,k} &= E_t\left[x_{t+h} - \hat{x}_{t+h,k}\right]^2; \\ V_h &= \lim_{k\to\infty} V_{h,k} \end{aligned} \tag{6.27}$$

Here, $V_{h,k}$ is the variance of the forecast error at horizon h obtained when the predictor uses the history of the process from time t back to $t-k$. Note that the information set $\{x_{t-k}, \ldots, x_t\}$ may be a truncated subset of the full conditional information set $\Im_t$. For example, for an $MA(1)$ process, we have $\Im_t = \{x_{-\infty}, \ldots, x_t\}$, while the predictor might use only a truncated past up to $t-k$.

Definition 6.1 (Backward δ-memory)**.** *Given a small number $\delta > 0$, the backward δ-memory is the integer M_δ such that:*

$$V_{1,k} - V_1 \leq \delta\,, \quad \forall k \geq M_\delta \tag{6.28}$$

This means that extending the information set beyond M_δ does not significantly improve the one-step-ahead forecast.
In *Figure 6.7*, we see that for an $AR(p)$ process, $M_\delta \leq p$ for any $\delta \geq 0$, and $M_0 = p$. On the other hand, for an $MA(q)$ process, $M_0 = \infty$ since the dual representation is $AR(\infty)$. However, for $\delta > 0$, a finite value of M_δ may still be found.
In conclusion, *a one-step-ahead forecast reaches its maximum accuracy with a finite number of lags if the process belongs to the AR class, whereas the entire infinite past is required if it belongs to the MA class.*

[77] This section is inspired by Granger, C.W.J. and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York, p. 115.

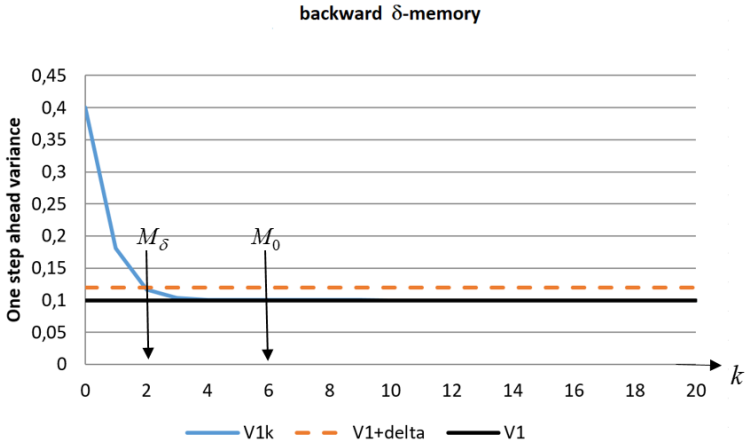


Figure 6.7: Backward-memory for an *AR(6)* process

Definition 6.2 (Forward δ-memory)**.** *Given a small number $\delta > 0$, the forward δ-memory is the integer M_δ^* such that:*

$$V_\infty - V_h < \delta\,, \quad \forall h \geq M_\delta^* \tag{6.29}$$

The forward δ-memory provides a measure of the reliability of an h-step-ahead forecast. As discussed, the forecast error variance increases with the forecast horizon and converges to the variance of the process, denoted V_∞. The value $h = M_\delta^*$ thus corresponds to the maximum acceptable imprecision, where $V_\infty - V_h < \delta$.
If $V_\infty - V_h = 0$, then for an $MA(q)$ process we have $M_\delta^* = q$, while for an $AR(p)$ process we get $M_\delta^* = \infty$. This reflects the well-known duality between these two classes of processes.
In conclusion, *the forecast error variance converges to the variance of the process over a finite horizon for MA processes, while it takes an infinite horizon for AR processes.*
For clarity, we summarize:

- *An AR process requires only a finite number of past values equal to the process order, but it loses predictive power only over an infinite forecast horizon.*
- *An MA process requires the entire infinite past for forecasting, but its predictive power vanishes after a finite forecast horizon equal to the process order.*

In the case of non-stationary processes, forward δ-memory is not meaningful because V_∞ is not finite. However, the definition of backward δ-memory remains applicable.

Example 6.5. *(Forward δ-memory for both AR and MA processes)*
Consider the following processes, already seen in *Example 2.6*:

$$AR(4): \quad X_t = 0.2X_{t-1} - 0.1X_{t-2} - 0.2X_{t-3} - 0.4X_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, 4)$$

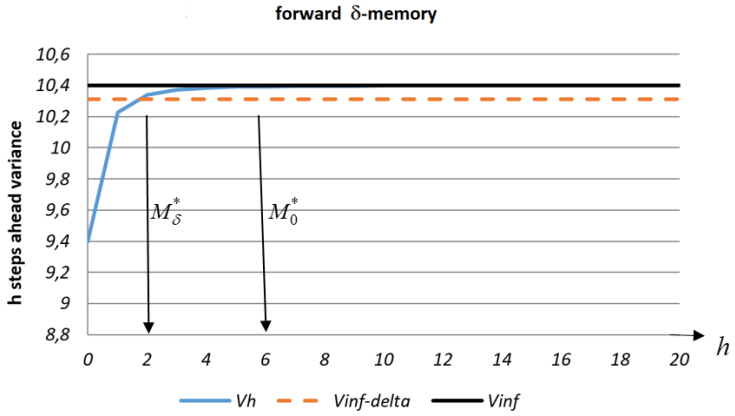


Figure 6.8: Forward δ-memory for an *MA(6)* process

$$MA(4): \quad Y_t = 0.3\varepsilon_{t-1} - 0.2\varepsilon_{t-2} + 0.5\varepsilon_{t-3} + 0.1\varepsilon_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, 4)$$

The AR process is stationary and the MA process is invertible. Their theoretical variances are:

$$Var(X_t) \approx 5.0755\,, \qquad Var(Y_t) = 5.56$$

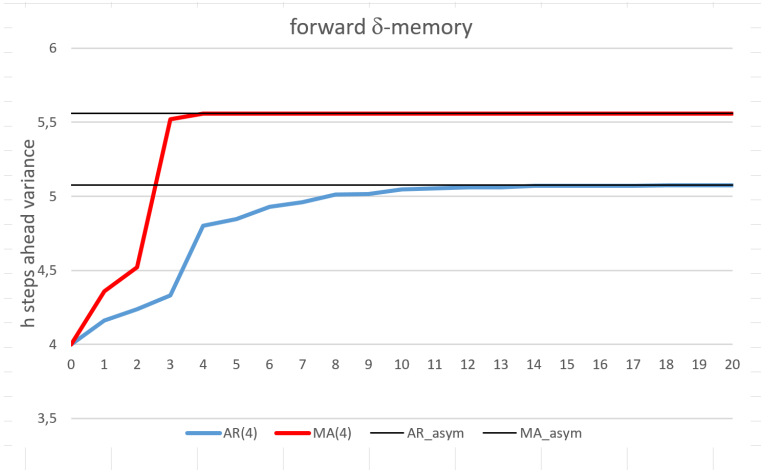


Figure 6.9: Forward δ-memory for *AR(4)* and *MA(4)* processes

In *Figure 6.9*, the prediction-error variance for the *MA(4)* process reaches $Var(Y_t) = 5.56$ when $h = 4$. In contrast, the error variance for the *AR(4)* process is still far from its asymptotic value $Var(X_t) \approx 5.0755$, which is attained only as $h \to \infty$.
A similar plot could be drawn to illustrate the behavior of the backward δ-memory for these two models.

The memory behavior of AR and MA processes with respect to their past and future completes the picture of their duality, already discussed in terms of autocorrelation and partial autocorrelation functions.

7 Non-stationary Processes

Many economic time series exhibit persistent growth or decline. It is therefore essential to study processes that violate stationarity. The simplest example of a non-stationary process arises when the stationarity condition is violated in a first-order autoregressive process. Consider the AR(1) process:

$$X_t = \alpha X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2), \tag{7.1}$$

If we set $\alpha = 1$, equation (7.1) becomes:

$$X_t = X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2), \tag{7.2}$$

which is no longer stationary.
This process is known in the literature as a *random walk (RW)*[78]. The process is assumed to originate from an initial value X_0, which may be either a deterministic constant or a random variable. In the latter case, it is assumed that X_0 has mean μ_0 and variance σ_0^2, both finite.
By repeated substitution, we obtain:

$$\begin{aligned} X_1 &= X_0 + \varepsilon_1 \\ X_2 &= X_1 + \varepsilon_2 = X_0 + \varepsilon_1 + \varepsilon_2 \\ &\vdots \\ X_t &= X_{t-1} + \varepsilon_t = X_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t \\ &= X_0 + \sum_{j=1}^{t} \varepsilon_j \end{aligned} \tag{7.3}$$

The expected value of the process is:

$$E(X_t) = E\left(X_0 + \sum_{j=1}^{t} \varepsilon_j\right) = \begin{cases} X_0 & if\ X_0\ is\ a\ deterministic\ value \\ \mu_0 & if\ X_0\ is\ a\ random\ variable \end{cases} \tag{7.4}$$

For simplicity, it is common to assume that the mean is zero in both cases. Then equation (7.3) becomes:

$$X_t = \sum_{j=1}^{t} \varepsilon_j \tag{7.5}$$

Equation (7.5) shows that the random walk is the sum of t white noise terms. As with continuous-time stochastic processes, where integration replaces summation, this is referred to as an *integrated process*.

[78] Sometimes vividly described as a *drunkard's walk*. The term "random walk" was first introduced by Pearson (1905).

From equation (7.5), the variance is:

$$Var(X_t) = Var\left(\sum_{j=1}^{t} \varepsilon_j\right) = \sum_{j=1}^{t} Var(\varepsilon_j) = t\sigma_\varepsilon^2 \tag{7.6}$$

Since the variance increases with time, we have $\lim_{t\to\infty} Var(X_t) = \infty$.
For non-stationary processes, the covariance function no longer satisfies the even function property, as both variances and covariances may depend on time. As a result, the autocovariance function must be computed separately for positive and negative lags.
For $k > 0$:

$$\begin{aligned} Cov(X_t, X_{t+k}) &= E\left(\sum_{j=1}^{t} \varepsilon_j \sum_{i=1}^{t+k} \varepsilon_i\right) = E\left[\sum_{j=1}^{t} \varepsilon_j \left(\sum_{i=1}^{t} \varepsilon_i + \sum_{s=t+1}^{t+k} \varepsilon_s\right)\right] \\ &= E\left(\sum_{j=1}^{t} \varepsilon_j \sum_{i=1}^{t} \varepsilon_i + \sum_{j=1}^{t} \varepsilon_j \sum_{s=t+1}^{t+k} \varepsilon_s\right) = t\sigma_\varepsilon^2 \end{aligned} \tag{7.7}$$

Also for $k > 0$:

$$\begin{aligned} Cov(X_t, X_{t-k}) &= E\left(\sum_{j=1}^{t} \varepsilon_j \sum_{i=1}^{t-k} \varepsilon_i\right) = E\left[\left(\sum_{j=1}^{t-k} \varepsilon_j + \sum_{s=t-k+1}^{t} \varepsilon_s\right) \sum_{i=1}^{t-k} \varepsilon_i\right] \\ &= E\left(\sum_{j=1}^{t-k} \varepsilon_j \sum_{i=1}^{t-k} \varepsilon_i + \sum_{s=t-k+1}^{t} \varepsilon_s \sum_{i=1}^{t-k} \varepsilon_i\right) = (t-k)\sigma_\varepsilon^2 \end{aligned} \tag{7.8}$$

The corresponding autocorrelation functions are:

$$Corr(X_t, X_{t+k}) = \rho_{t,k} = \frac{t\sigma_\varepsilon^2}{\sqrt{t\sigma_\varepsilon^2}\sqrt{(t+k)\sigma_\varepsilon^2}} = \frac{\sqrt{t}}{\sqrt{(t+k)}} = \sqrt{\frac{t}{t+k}} \tag{7.9}$$

$$Corr(X_t, X_{t-k}) = \rho_{t,-k} = \frac{(t-k)\sigma_\varepsilon^2}{\sqrt{t\sigma_\varepsilon^2}\sqrt{(t-k)\sigma_\varepsilon^2}} = \frac{\sqrt{t-k}}{\sqrt{t}} = \sqrt{\frac{t-k}{t}} \tag{7.10}$$

In an alternative formulation, assuming $s \neq t$, we have:

$$Corr(X_t, X_s) = \rho_{t,s} = \frac{\min(t,s)}{\sqrt{t}\sqrt{s}} = \frac{\min(t,s)}{\sqrt{t\,s}} \tag{7.11}$$

For any finite lag k, it holds that $\lim_{t\to\infty} \rho_{t,\pm k} = 1$. This implies that the random walk process remains highly autocorrelated even at very long lags.
Figure 7.1 displays the theoretical autocorrelation values of the random walk process for the first 20 lags and for various values of time t. As time increases, the autocorrelation curve tends to remain close to 1.

Figure 7.2 illustrates the simulation of a random walk process starting from a zero initial value, where the error terms are i.i.d. with zero mean and variance equal to 4. The simulation is extended over 2000 observations to show that the process diverges from its theoretical mean (which is zero), and it becomes impossible to identify the turning points of the curve or to determine when the process returns to its mean[79].

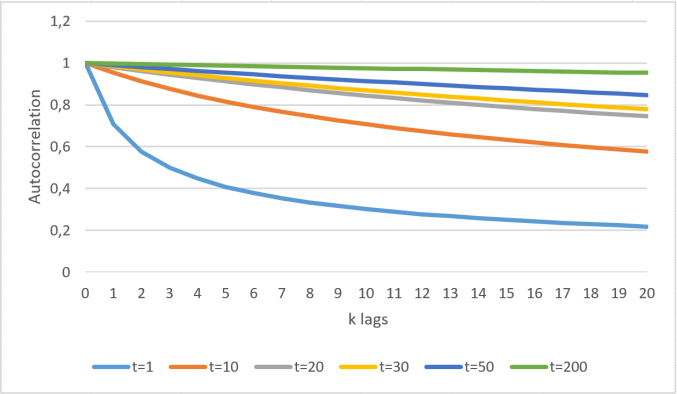


Figure 7.1: Autocorrelation function of a *RW* process

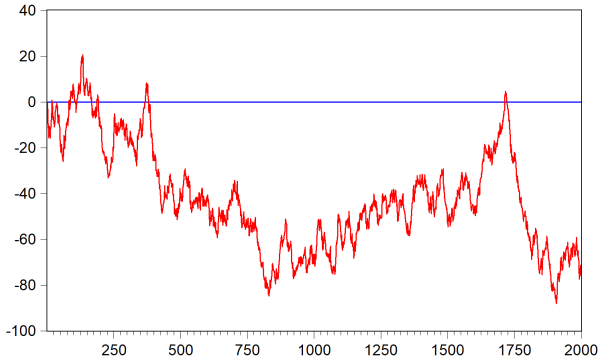


Figure 7.2: Simulation of a *RW* process: $X_t = X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 4)$ with initial value $X_0 = 0$

Applying the first difference to the random walk process yields a stationary process, specifically, a white noise process. In fact:

$$X_t - X_{t-1} = \varepsilon_t \,. \tag{7.12}$$

This transformation can be described using the *first difference operator*, denoted by the symbol Δ.

[79] In the financial literature, this behavior is often referred to as a *non-mean-reverting process*.

The Δ operator is a linear operator and is defined in terms of the lag operator L as follows:

$$\Delta = 1 - L \tag{7.13}$$

Rewriting expression (7.12) using the first difference operator, we obtain:

$$\Delta X_t = \varepsilon_t \,, \tag{7.14}$$

where ε_t is a white noise process.
The first difference operator can be applied to any stochastic process. For instance, if we apply it to a stochastic process Z_t, we obtain:

$$\Delta Z_t = X_t \,,$$

where X_t is not necessarily stationary—it could be, for example, a random walk. In such a case, to obtain a stationary process, the differencing operator must be applied to X_t itself. Applying the operator again to both sides yields:

$$\Delta(\Delta Z_t) = \Delta X_t = \varepsilon_t \,,$$

which implies:

$$\Delta^2 Z_t = (1 - L)^2 Z_t = (1 - 2L + L^2) Z_t = Z_t - 2Z_{t-1} + Z_{t-2} = \varepsilon_t \,. \tag{7.15}$$

The process Z_t is said to be *integrated of order 2*, since two applications of the differencing operator are required to obtain a stationary process. The notation $Z_t \sim I(2)$ is used to indicate this.
In general, the letter d denotes the order of integration, and the general notation for integrated processes is $I(d)$.
The random walk is an integrated process of order 1. From equation (7.13), we obtain:

$$(1 - L) X_t = \varepsilon_t \,.$$

Hence, the characteristic equation is $(1 - L) = 0$, from which it follows that there is a single root, which is a unit root. This observation motivates the statement that the *random walk is a process with a unit root*. By contrast, the process Z_t from the previous example has two unit roots, as it can be written as $(1 - L)(1 - L) Z_t = \varepsilon_t$, with the associated characteristic equation $(1 - L)^2 = 0$.
In general, an integrated process with d unit roots is denoted as $I(d)$. A process of type $I(0)$ is said to be without unit roots and is therefore stationary[80]. In empirical econometrics, integration orders commonly encountered are 0 or 1, while time series of order 2 are rarely observed.

[80] In the literature, there is no unanimous agreement on the definition of an $I(0)$ process. For example, James Davidson, in one of his recent papers (`http://people.exeter.ac.uk/jehd201/WhenisI0.pdf`), lists five different definitions from various authors and adds a sixth of his own.

Random Walk
Sample: 1 200
Included observations: 200

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.968	0.968	190.37	0.000
		2	0.926	-0.181	365.49	0.000
		3	0.885	0.017	526.19	0.000
		4	0.845	-0.010	673.45	0.000
		5	0.802	-0.078	806.75	0.000
		6	0.762	0.043	927.63	0.000
		7	0.725	0.015	1037.7	0.000
		8	0.691	-0.003	1138.1	0.000
		9	0.663	0.090	1230.9	0.000
		10	0.632	-0.093	1315.9	0.000
		11	0.603	0.020	1393.7	0.000
		12	0.579	0.064	1465.9	0.000
		13	0.559	0.005	1533.4	0.000
		14	0.543	0.061	1597.3	0.000
		15	0.527	-0.020	1657.9	0.000
		16	0.501	-0.180	1713.0	0.000
		17	0.468	-0.073	1761.4	0.000
		18	0.440	0.090	1804.4	0.000
		19	0.411	-0.075	1842.2	0.000
		20	0.376	-0.060	1873.9	0.000
		21	0.344	0.044	1900.6	0.000
		22	0.314	-0.028	1923.0	0.000
		23	0.295	0.173	1942.9	0.000
		24	0.281	-0.001	1961.1	0.000

Figure 7.3: Correlogram of a simulated *RW* with 200 observations.

Figure 7.3 shows the correlogram for the first 200 observations of the time series depicted in *Figure 7.2*. The autocorrelation pattern is typical of non-stationary processes, displaying a slow decay in autocorrelation values and a partial autocorrelation at lag 1 close to unity. This latter feature may be interpreted as an indication of the presence of a unit root[81].

7.1 Random Walk in Finance

The random walk process plays an important role in finance. Many financial and macroeconomic time series appear to be well represented by this process, such as stock prices and certain macroeconomic indicators.

From a forecasting perspective, the random walk is not attractive, since the optimal forecast[82] is given by: $E_t(X_{t+1}) = X_t$. For example, if X_t represents the time series of a stock price, then the best possible forecast for tomorrow's price is today's price—information that is already available to all economic agents, and thus useless for any profitable trading strategy.

Based on this empirical evidence, the so-called *"Random Walk Theory"* emerged. The principal advocate of this theory is Burton Malkiel (2003), of Princeton University, who authored the book *"A Random Walk Down Wall Street"* in which he argues that investors

[81] One should avoid the temptation to attribute this type of correlogram to a stationary *AR(1)* process merely because the estimated autocorrelation coefficient is less than one, i.e., $\hat{\alpha}_{11} = \hat{\alpha} = 0.968 < 1$. In Section 8.3, the estimated value of $\hat{\alpha}$ will be compared to the true value of the α parameter using appropriate unit root tests.

[82] See Section 6.4.

are better off holding an index fund tracking a broad market index than attempting to pick individual stocks or actively manage mutual funds.
This theory is closely linked to the *Efficient Market Hypothesis (EMH)*. An "efficient" market is defined[83], as a market in which numerous rational economic agents maximize profits and actively compete by attempting to predict the future market value of individual securities. In such a market, relevant current information is almost freely and simultaneously available to all participants.
As a result, competition among many rational agents leads to a situation in which, at any given moment, current stock prices reflect the effects of all past and expected future events. In other words, in an efficient market, the actual price of a security at any given time is a good estimate of its intrinsic value.
The *"Non-Random Walk Theory"* is supported by those—investors, economists, and academics—who believe that the market exhibits some degree of predictability, in contrast to the Random Walk Theory.
Among the most prominent exponents of this view are Lo and Lo and MacKinlay (2002), who compiled a collection of previously published studies in their book *A Non-Random Walk Down Wall Street*. In this work, they cite numerous authors to argue that the Random Walk Hypothesis is neither a necessary nor a sufficient condition for the rational pricing of financial assets. In other words, price unpredictability does not imply a well-functioning financial market with rational investors, and price predictability does not imply the opposite. Their book presents a series of methods for detecting predictability and evaluating its statistical and economic significance, as well as the prospects for future methodological developments.

7.2 ARIMA(p,d,q) Processes

Given that $Z_t \sim I(d)$, by definition we have:

$$\Delta^d Z_t = X_t \sim I(0) \tag{7.16}$$

Since X_t is a stationary process, it can be represented[84] as belonging to the $ARMA(p,\, q)$ family:

$$\alpha(L)X_t = \beta(L)\varepsilon_t\,.$$

Substituting from equation (7.16), we obtain:

$$\alpha(L)\Delta^d Z_t = \beta(L)\varepsilon_t\,, \tag{7.17}$$

which defines the general form of an *Autoregressive Integrated Moving Average* - ($ARIMA(p,d,q)$) - process, where p is the order of the autoregressive polynomial $\alpha(L)$, d is the order of integration, and q is the order of the moving average polynomial $\beta(L)$.

[83] The definition used here is taken from Fama (1965).
[84] As noted in Section 3.1 (*Remark 3.3*), the *ARMA* class does not exhaust all stationary linear processes, although it represents an extremely broad class.

ARIMA models constitute a highly general framework for modeling non-stationary stochastic processes.
Before estimating an *ARIMA* model, it is necessary to determine the order of integration of the time series, apply differencing accordingly, and then estimate the corresponding *ARMA* process.

7.3 Cointegrated Processes

It has been noted that many economic time series can be modeled as $I(1)$ processes. However, there are cases in which a linear combination of such processes results in an $I(0)$ process. In this case, the component processes of the linear combination are referred to as *cointegrated processes*.
More specifically, let $\mathbf{X}_t$ be a vector of stochastic processes sharing the same integration order, for instance $\mathbf{X}_t \sim I(d)$. The vector $\mathbf{X}_t$ is said to be cointegrated if there exists a coefficient vector $\mathbf{a}$ and an integer $b > 0$ such that the following linear combination holds:

$$Y_t = \mathbf{a}'\mathbf{X}_t \sim I(d-b) \tag{7.18}$$

Expression (7.18) implies that there exists a linear combination which reduces the degree of integration with respect to the individual processes involved.

Remark 7.1. In empirical applications, cointegrating relationships may include deterministic components, such as an intercept or a linear time trend. Accordingly, cointegration can be classified into different cases depending on the deterministic specification of the long-run equilibrium relationship. Common cases include:

(i) cointegration without deterministic terms, when the equilibrium relation has zero mean;

(ii) cointegration with an intercept, when the equilibrium relation fluctuates around a non-zero constant;

(iii) cointegration with a linear time trend, when the equilibrium relation includes a deterministic trend.

These distinctions are important because they affect both the specification of cointegration tests and the form of the corresponding error correction representation.

Example 7.1. Let $\{\varepsilon_t\}$ and $\{\eta_t\}$ be two white noise processes that are stochastically independent and identically distributed with equal variance. Define two new processes $\{Y_t\}$ and $\{X_t\}$ as follows:

$$\begin{cases} Y_t = Y_{t-1} + \varepsilon_t \\ X_t = X_{t-1} + \eta_t \end{cases} \tag{7.19}$$

By construction, $\{Y_t\}$ and $\{X_t\}$ are two independent random walk processes. Since they are stochastically independent, they are not cointegrated: there exists no linear combination $aY_t + bX_t = Z_t$ such that $Z_t \sim I(0)$.

Example 7.2. Consider the following bivariate process[85]:

$$\begin{cases} u_{1t} = \rho_1 u_{1t-1} + \varepsilon_t \\ u_{2t} = \rho_2 u_{2t-1} + \eta_t \,, \end{cases}$$

where $\begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \sim NID\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \gamma \\ \gamma & \sigma_\eta^2 \end{pmatrix}\right]$.
If we assume that $\rho_1 = 1$ and $|\rho_2| < 1$, then the process u_{1t} is a random walk, while u_{2t} is stationary.
Now consider the following linear combinations:

$$\begin{cases} Y_t + \alpha X_t = u_{1t} \\ Y_t - \beta X_t = u_{2t} \end{cases} . \tag{7.20}$$

What is the order of integration of the processes $\{Y_t\}$ and $\{X_t\}$?
The system in equation (7.20) can be written in matrix form as:

$$\begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix},$$

By inverting the coefficient matrix, we obtain:

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \frac{1}{\alpha + \beta} \begin{bmatrix} -\beta & -\alpha \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$Y_t = \frac{1}{\alpha + \beta} (-\beta u_{1t} - \alpha u_{2t}) \quad .$$

$$X_t = \frac{1}{\alpha + \beta} (u_{2t} - u_{1t})$$

Hence, both processes are $X_t \sim I(1)$ and $Y_t \sim I(1)$, as they are obtained from a linear combination of an $I(1)$ process (u_{1t}) and an $I(0)$ process (u_{2t}). However, the vector

[85] Inspired by the example in Engle and Granger (1987).

$\begin{bmatrix} Y_t \\ X_t \end{bmatrix}$ is cointegrated, since there exists a linear combination $Y_t - \beta X_t = u_{2t}$ with a stationary error. The corresponding cointegrating vector is $\mathbf{a}' = \begin{bmatrix} 1 & -\beta \end{bmatrix}$.
From an interpretive standpoint, the second equation in system (7.20) can be rewritten as:

$$Y_t = \beta X_t + u_{2t}\,, \tag{7.21}$$

which represents a regression of process Y_t on process X_t with a stationary error term. It can thus be interpreted, for instance, as an *Autoregressive Distributed Lag — ADL(0,0)* representation[86]. In other words, the process Y_t exhibits a unit root because it is *induced* by the combination of a conditioning process X_t (which itself has a unit root) and a stationary component u_{2t}.

In general, economic variables are expressed in *levels of measurement*, i.e., their observed value at a given point in time—for example, the level of prices, wages, or income. The analysis of such series is directly linked to their *long-run* behavior. The long run is a theoretical concept used in economics to describe the equilibrium relationship between variables[87]. This concept stands in contrast to the *short-run*, during which dynamic fluctuations and market disequilibria may occur. Typically, short-run dynamics are described in terms of incremental or decremental changes, or rates of variation[88].
From an economic viewpoint *equilibrium* is well-defined; econometrically this corresponds to *stability* of certain linear combinations of the involved stochastic processes.
The concept of cointegration, introduced by Clive Granger (1981), had a major impact on the development of econometric models. In the statistical literature on multivariate ARIMA models, estimation has traditionally been conducted by differencing non-stationary series to achieve stationarity, as in equation (7.16). However, this approach removes a substantial portion of the original series' variability—sometimes as much as 99%. Economists, in contrast, are often interested in the relationships between variables in levels—that is, precisely in the part of variability eliminated by differencing.
The idea of cointegration has the advantage of restoring the analysis to focus on the relationships between variables expressed in levels. Moreover, it allows for a distinction between long-run and short-run behavior by linking the latter to the dynamics of the former.
In Chapter 8 we will see how cointegration underpins the ECM representation, linking long-run equilibria to short-run dynamics.

[86] See Chapter 5.
[87] In Chapter 5, the concept of the steady-state system is discussed.
[88] For further discussion on the long-run concept and equilibrium relationships between variables, see Banerjee et al. (1993).

7.4 Trend-Stationary (TS) versus Difference-Stationary (DS) processes

Up to this point, $MA(q)$ and $AR(p)$ stochastic processes have been considered under the assumption of zero mean. The process can be generalized by allowing the theoretical mean to differ from zero. Two examples, based on the $AR(1)$ process, are presented to highlight the role of the autoregressive coefficient α in the presence of deterministic, non-stochastic components such as a constant term or a time trend.

Remark 7.2. In empirical econometrics, it is common to distinguish between evolutionary patterns that are interpreted as deterministic trends (trend-stationary, TS processes) and those interpreted as stochastic trends (difference-stationary, DS processes), that is, as integrated processes. In finite samples, however, this distinction is often difficult to establish.

The distinction has both interpretative and econometric implications. From an interpretative perspective, describing a time series as driven by a deterministic trend is a very strong assumption, since it implies that the long-run evolution of the series will never change direction and will follow a predetermined path. Such an assumption may raise doubts as to whether truly deterministic trends exist in economic time series, or whether the inclusion of a deterministic trend should instead be regarded as an econometric device aimed at improving the fit of the model.

By contrast, interpreting a series as driven by a stochastic trend is more consistent with the idea that economic time series are subject to persistent but potentially reversible changes in their long-run trajectory. This interpretation allows for shifts in the direction of evolution as the result of accumulated shocks.

From an econometric standpoint, the distinction between TS and DS behavior determines the appropriate transformation of the data and the validity of statistical inference. Over-differencing a stationary process may inflate variance and induce unnecessary serial correlation, while under-differencing a non-stationary process leads to spurious regression results.

Example 7.3. *(Process with intercept or drift)*

A process with intercept is defined as follows:

$$X_t = c + \alpha X_{t-1} + \varepsilon_t; \quad c \neq 0, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)\,, \tag{7.22}$$

where $X_0 \neq 0$ is the initial value of the process. By iterated substitution, we obtain:

$$\begin{aligned} X_1 &= c + \alpha X_0 + \varepsilon_1 \\ X_2 &= c + \alpha X_1 + \varepsilon_2 = c + \alpha c + \alpha^2 X_0 + \alpha\varepsilon_1 + \varepsilon_2 \\ X_3 &= c \sum_{j=0}^{2} \alpha^j + \alpha^3 X_0 + \sum_{j=0}^{2} \alpha^j \varepsilon_{t-j} \\ &\vdots \\ X_t &= c \sum_{j=0}^{t-1} \alpha^j + \alpha^t X_0 + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j} \\ &= c \frac{1-\alpha^t}{1-\alpha} + \alpha^t X_0 + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j} \end{aligned} \tag{7.23}$$

In expression (7.23), the first term derives from the formula for the sum of a geometric progression. As time increases, the parameter α plays a critical role. We consider two cases: $|\alpha| < 1$ (stationary, i.e., $I(0)$ process) and $\alpha = 1$ (unit root process).

a) If $|\alpha| < 1$, then the process converges to:

$$X_t = \frac{c}{1-\alpha} + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \,. \tag{7.24}$$

In this case, the limit process has the following features:

1) The $AR(1)$ process can be expressed as an $MA(\infty)$ process with $E(X_t) = \dfrac{c}{1-\alpha}$;
2) The initial condition X_0 is asymptotically irrelevant.

The variance is:

$$\begin{aligned} Var(X_t) &= Var\left(\frac{c}{1-\alpha} + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \right) = Var\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j} \right) \\ &= \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \alpha^{2j} = \frac{\sigma_\varepsilon^2}{1-\alpha^2} \,. \end{aligned} \tag{7.25}$$

b) If $\alpha = 1$, expression (7.23) is no longer valid, and it is preferable to proceed by

iterative substitution:

$$\begin{aligned} X_1 &= c + X_0 + \varepsilon_1 \\ X_2 &= c + X_1 + \varepsilon_2 = 2c + X_0 + \varepsilon_1 + \varepsilon_2 \\ X_3 &= 3c + X_0 + \sum_{j=1}^{3} \varepsilon_j \\ &\vdots \\ X_t &= ct + X_0 + \sum_{j=1}^{t} \varepsilon_j \,. \end{aligned} \tag{7.26}$$

In this case, we obtain an $I(1)$ process whose mean depends on time. Moreover, the process retains memory of its initial value X_0, as $E(X_t) = ct + X_0$ and $Var(X_t) = t\sigma_\varepsilon^2$. Therefore, the limiting process has both infinite mean and infinite variance.

It is evident that the level process defined in equation (7.26) has a deterministic trend when $c \neq 0$. Applying the first difference operator yields:

$$\begin{aligned} \Delta X_t &= \Delta\left(ct + X_0 + \sum_{j=1}^{t} \varepsilon_j\right) = c\Delta t + \Delta X_0 + \Delta\left(\sum_{j=1}^{t} \varepsilon_j\right) \\ &= c[t - (t-1)] + 0 + \left(\sum_{j=1}^{t} \varepsilon_j - \sum_{j=1}^{t-1} \varepsilon_j\right) \\ &= c + \varepsilon_t \,. \end{aligned} \tag{7.27}$$

The differenced process is stationary and, as such, loses memory of its origin. The original process in levels is thus called an $I(1)$ *process with drift*.
In the previous case a), involving a stationary process, applying the first difference operator yields:

$$\Delta X_t = \alpha \Delta X_{t-1} + \Delta\varepsilon_t \,, \tag{7.28}$$

since the difference of constant terms is zero. The differenced process remains stationary, but over-differencing may lead to increased variance compared to the original stationary process. Indeed:

$$\begin{aligned} \Delta X_t &= \alpha \Delta X_{t-1} + \Delta\varepsilon_t \\ (1 - \alpha L)\Delta X_t &= \Delta\varepsilon_t \\ \Delta X_t &= \sum_{j=0}^{\infty} \alpha^j \Delta\varepsilon_{t-j} \,. \end{aligned} \tag{7.29}$$

The process $\Delta\varepsilon_t = (1 - L)\varepsilon_t$ is stationary but not invertible, since it can be interpreted as an $MA(1)$ process of the form $\varepsilon_t - \theta\varepsilon_{t-1}$ with parameter $\theta = -1$ (boundary case). As a consequence, $\Delta\varepsilon_t$ is no longer white noise. Therefore, although ΔX_t is stationary, it is a linear combination of error terms that are not white noise.

Note that $Var(\Delta\varepsilon_t) = 2\sigma_\varepsilon^2$ is constant over time, and thus:

$$\begin{aligned} Var(\Delta X_t) &= Var(X_t) + Var(X_{t-1}) - 2\,Cov(X_t, X_{t-1}) \\ &= \frac{2\sigma_\varepsilon^2}{1-\alpha^2} - \frac{2\alpha\sigma_\varepsilon^2}{1-\alpha^2} = \frac{2\sigma_\varepsilon^2}{1+\alpha}\,. \end{aligned} \tag{7.30}$$

Given the stationarity condition $|\alpha| < 1$, the result in equation (7.30) is always positive. It can be noted that if $-1 < \alpha < 0.5$, then:

$$Var(\Delta X_t) > Var(X_t).$$

Indeed:

$$\begin{aligned} \frac{2\sigma_\varepsilon^2}{1+\alpha} > \frac{\sigma_\varepsilon^2}{1-\alpha^2} \quad &\Rightarrow \quad \frac{2\sigma_\varepsilon^2}{1+\alpha} - \frac{\sigma_\varepsilon^2}{1-\alpha^2} > 0 \\ &\Rightarrow \quad \frac{2\sigma_\varepsilon^2(1-\alpha) - \sigma_\varepsilon^2}{1-\alpha^2} > 0 \quad \Rightarrow \quad \sigma_\varepsilon^2(1-2\alpha) > 0 \\ &\Rightarrow \quad \alpha < \frac{1}{2} \end{aligned}$$

In conclusion, over-differencing a stationary process is not always advisable, for at least two reasons: it introduces autocorrelation in the error term and may increase the process's variance.

Example 7.4. *(Process with intercept and trend)*
An $AR(1)$ process with intercept and trend is given by:

$$X_t = \beta_0 + \beta_1 t + \alpha X_{t-1} + \varepsilon_t; \quad \varepsilon_t \sim WN(0,\, \sigma_\varepsilon^2)\,. \tag{7.31}$$

Assume that the initial value is $X_0 \neq 0$. The iterated substitution procedure gives:

$$\begin{aligned} X_1 &= \beta_0 + \beta_1 + \alpha X_0 + \varepsilon_1 \\ X_2 &= \beta_0 + 2\beta_1 + \alpha X_1 + \varepsilon_2 = \beta_0 + 2\beta_1 + \alpha(\beta_0 + \beta_1 + \alpha X_0 + \varepsilon_1) + \varepsilon_2 \\ &= (1+\alpha)\beta_0 + (2+\alpha)\beta_1 + \alpha^2 X_0 + \alpha\varepsilon_1 + \varepsilon_2 \\ X_3 &= \beta_0 + 3\beta_1 + \alpha X_2 + \varepsilon_3 \\ &= (1+\alpha+\alpha^2)\beta_0 + (3+2\alpha+\alpha^2)\beta_1 + \alpha^3 X_0 + \sum_{j=0}^{2} \alpha^j \varepsilon_{3-j} \\ &\vdots \\ X_t &= \alpha^t X_0 + \beta_0 \sum_{j=1}^{t} \alpha^{j-1} + \beta_1 \sum_{j=1}^{t} \alpha^{t-j} j + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j} \end{aligned} \tag{7.32}$$

By evaluating the sums, we obtain:

$$X_t = \alpha^t X_0 + \beta_0 \frac{1-\alpha^t}{1-\alpha} + \beta_1 \frac{\alpha^{t+1} + t(1-\alpha) - \alpha}{(1-\alpha)^2} + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j}\,. \tag{7.33}$$

The third term arises from the known summation formula[89]:

$$\sum_{i=0}^{n-1} ia^i = \frac{a - na^n + (n-1)a^{n+1}}{(1-a)^2}\,.$$

Setting $n = t$ and $i = t - k$ in the third term of equation (7.33), we obtain:

$$\begin{aligned} \sum_{k=1}^{t} (t-k)a^{t-k} &= \sum_{k=1}^{t} ta^{t-k} - \sum_{k=1}^{t} ka^{t-k} \\ &= \frac{a - ta^t + (t-1)a^{t+1}}{(1-a)^2} \end{aligned}\,,$$

Thus:

$$\begin{aligned} \sum_{k=1}^{t} ka^{(t-k)} &= t\sum_{k=1}^{t} a^{(t-k)} - \frac{a - ta^t + (t-1)a^{t+1}}{(1-a)^2} \\ &= t\sum_{k=0}^{t-1} a^{k} - \frac{a - ta^t + (t-1)a^{t+1}}{(1-a)^2} \\ &= t\frac{1-a^t}{1-a} - \frac{a - ta^t + (t-1)a^{t+1}}{(1-a)^2} \\ &= \frac{t(1-a) - a + a^{t+1}}{(1-a)^2} \end{aligned}$$

If $|\alpha| < 1$, the stationarity condition ensures that the process gradually loses memory of its initial value. Nevertheless, the trend component remains in expression (7.33), so as $t \to \infty$, we also have $E(X_t) \to \infty$. However, an approximate evaluation of its value can be provided for finite time intervals[90].
If $\alpha = 1$, the presence of a unit root transforms equation (7.32) into:

$$\begin{aligned} X_t &= X_0 + \beta_0 \sum_{j=1}^{t} 1 + \beta_1 \sum_{j=1}^{t} j + \sum_{j=0}^{t-1} \varepsilon_{t-j} \\ &= X_0 + \beta_0 t + \beta_1 \left[(t+1)\tfrac{t}{2}\right] + \sum_{j=0}^{t-1} \varepsilon_{t-j} \\ &= X_0 + \left(\beta_0 + \tfrac{\beta_1}{2}\right) t + \tfrac{\beta_1}{2} t^2 + \sum_{j=0}^{t-1} \varepsilon_{t-j}\,, \end{aligned} \tag{7.34}$$

[89] See Graham, Knuth, and Patashnik (1990), formula (2.26), p. 33.
[90] See *Appendix 7.A.*

where, as time increases, the presence of a parabolic trend and the retention of memory of the initial value become evident.
By applying the difference operator to equation (7.34), the unit root is removed, and the process becomes:

$$\Delta X_t = \left(\beta_0 + \tfrac{\beta_1}{2}\right) + \beta_1 t + \varepsilon_t \,. \tag{7.35}$$

which is consistent with the expression in (7.31).

The estimator of the regression coefficient in a trend-stationary process is said to be *superconsistent*[91].
To motivate this result, consider the slope coefficient in the following model:

$$X_t = \beta_0 + \beta_1 t + u_t, \quad u_t \sim WN(0, \sigma_u^2) \,. \tag{7.36}$$

Equation (7.36) defines an $I(0)$ process with a deterministic trend.
The OLS estimator of β_1 is given by:

$$\hat{\beta}_{1,T} = \frac{\sum_{t=1}^{T}(t-\bar{t})X_t}{\sum_{t=1}^{T}(t-\bar{t})^2} \,, \tag{7.37}$$

with expectation and variance:

$$\begin{aligned} E(\hat{\beta}_{1,T}) &= \beta_1 \\ Var(\hat{\beta}_{1,T}) &= \frac{\sigma_u^2}{\sum_{t=1}^{T}(t-\bar{t})^2} \,. \end{aligned} \tag{7.38}$$

The sum in the denominator of the variance is:

$$\sum_{t=1}^{T}(t-\bar{t})^2 = \sum_{t=1}^{T} t^2 - T\bar{t}^2 = \frac{T(T+1)(2T+1)}{6} - T\Big(\frac{T+1}{2}\Big)^2 = \frac{T(T^2-1)}{12}. \tag{7.39}$$

Applying the central limit theorem yields the following result:

$$\lim_{T\to\infty}\left[\sqrt{T^3}\left(\hat{\beta}_{1,T} - \beta_1\right)\right] \xrightarrow{d} \mathcal{N}\big(0,\ 12\sigma_u^2\big) \,. \tag{7.40}$$

By multiplying the difference between the estimator and the parameter by the factor $\sqrt{T^3}$, the limiting variance does not approach zero but tends to the finite value $12\sigma_u^2$. Indeed:

[91] See § 8.5.

$$\begin{aligned}
\lim_{T\to\infty} Var\left[\sqrt{T^3}\left(\hat{\beta}_{1,T}-\beta_1\right)\right] &= \lim_{T\to\infty} T^3 \cdot Var(\hat{\beta}_{1,T}) \\
&= \lim_{T\to\infty} \frac{T^3\sigma_u^2}{\sum_{t=1}^{T}(t-\bar{t})^2} \\
&= \sigma_u^2 \cdot \lim_{T\to\infty} \frac{T^3}{\frac{T}{12}(T^2-1)} \\
&= \sigma_u^2 \cdot \lim_{T\to\infty} \frac{12T^3}{T^3-T} = 12\sigma_u^2 .
\end{aligned} \tag{7.41}$$

If, instead, the scaling factor were $\sqrt{T}$ or T, then the limiting variance would tend to zero, and the transformed distribution would degenerate asymptotically.
In conclusion, the presence of a trend implies a higher degree of superconsistency than the presence of a unit root, for which a T factor is sufficient to avoid asymptotic degeneration of the distribution[92].
We can summarize the convergence rates for consistency by considering the multiplicative factor (mf) used to avoid degeneracy of the estimator's asymptotic distribution:

- $I(0)$ process $\qquad mf = T^{1/2}$
- $I(1)$ process $\qquad mf = T$
- Trend-stationary process $\qquad mf = T^{3/2}$

This result aligns with the intuitive behavior of time series. When errors are autocorrelated, the estimators converge slowly to their theoretical values, with convergence speed measured by $\sqrt{T}$. The speed increases in the presence of unit roots, being proportional to T. In stationary series, the process fluctuates within confidence bands centered around a constant mean.
By contrast, unit root processes drift away from their unconditional mean along persistently increasing or decreasing paths, with confidence bands widening over time due to variance growing toward infinity.
Finally, for trend-stationary processes, the behavior fluctuates around a deterministic linear trend. The estimator of β_1 rapidly captures the trend direction, as the series does not deviate from its deterministic path. Hence, the estimate is highly precise, and convergence to the true parameter is exceptionally fast.

Remark 7.3. In the presence of autocorrelated residuals, the rate of convergence of estimators becomes particularly relevant. Autocorrelation typically reduces the efficiency of ordinary least squares estimators, making faster convergence rates desirable in order to mitigate this loss of efficiency.

[92] See also §8.5 for superconsistency in the presence of unit roots.

However, combining estimators with different rates of convergence within the same specification may be problematic. Since estimators are generally correlated, the presence of slowly convergent estimators can act as a brake on faster convergent ones, potentially undermining the benefits of *superconsistency.* For this reason, model specification should take into account not only consistency, but also the relative convergence rates of the estimators involved.

Appendix 7.A

(Approximate Theoretical and Sample Mean in Trend-Stationary Processes)

Considering the model of *Example 7.4*, it has been seen that, starting from an initial value X_0 and using iterated substitutions, the expression (7.33) is obtained.
If $|\alpha| < 1$, for sufficiently large values of t, we have:

$$E(X_t) \approx \frac{\beta_0}{1-\alpha} + \beta_1 \left[\frac{t}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} \right] , \tag{7.A1}$$

where the symbol $\approx$ denotes approximation to the true value, obtained by neglecting terms that vanish exponentially as $t \to \infty$.
Evidently, as $t \to \infty$, $E(X_t) \to \infty$.
Expression (7.A1) is useful for approximating the expected value $E(X_t)$ over a finite time interval.
For values of t such that α^t is negligible, the approximation in (7.A1) also applies to the sample mean of X_t, provided that the sample size is sufficiently large and the variance of the error term is finite.
Rewriting (7.33) as:

$$X_t = \alpha^t X_0 + \beta_0 \frac{1-\alpha^t}{1-\alpha} + \beta_1 \frac{\alpha^{t+1} + t(1-\alpha) - \alpha}{(1-\alpha)^2} + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j} ,$$

we observe that the mean of the cumulative error is negligible, since it tends to zero (more quickly if the error variance is small).
Therefore, the approximate sample mean, denoted by $\bar{X}_{a,T}$, can be written as:

$$\begin{aligned} \bar{X}_{a,T} &\approx \frac{1}{T} \sum_{t=1}^{T} \left\{ \frac{\beta_0}{1-\alpha} + \beta_1 \left[\frac{t}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} \right] \right\} \\ &= \frac{\beta_0}{1-\alpha} + \frac{\beta_1 \sum_{t=1}^{T} t}{T(1-\alpha)} - \frac{\beta_1 \alpha}{(1-\alpha)^2} \\ &= \frac{\beta_0}{1-\alpha} + \frac{\beta_1 (1+T)T/2}{T(1-\alpha)} - \frac{\beta_1 \alpha}{(1-\alpha)^2} \\ &= \frac{\beta_0}{1-\alpha} + \frac{\beta_1 (1+T)}{2(1-\alpha)} - \frac{\beta_1 \alpha}{(1-\alpha)^2} . \end{aligned}$$

As we can see, the approximate sample mean is a linear function of the sample size T.
If $\beta_1 = 0$, the estimated approximate sample mean coincides with the asymptotic mean of the process X_t without trend.

When $\beta_1 \neq 0$, the sample mean $\bar{X}_{a,T}$ diverges from $E(X_t)$. Their difference, a function of t and T, is given by:

$$\begin{aligned}
\bar{X}_{a,T} - E(X_t) &= \frac{\beta_0}{1-\alpha} + \frac{\beta_1(1+T)}{2(1-\alpha)} - \frac{\beta_1\alpha}{(1-\alpha)^2} \\
&\quad - \left[\frac{\beta_0}{1-\alpha} + \beta_1\left(\frac{t}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2}\right)\right] \\
&= \frac{\beta_1}{2(1-\alpha)} + \frac{\beta_1 T}{2(1-\alpha)} - \frac{\beta_1 t}{1-\alpha} \\
&= \frac{\beta_1}{2(1-\alpha)} + \frac{\beta_1(T-2t)}{2(1-\alpha)}\,.
\end{aligned}$$

If the sample size coincides with the time index t (i.e., $T = t$), the difference becomes:

$$\begin{aligned}
\bar{X}_{a,T} - E(X_t) &= \frac{\beta_1}{2(1-\alpha)} + \frac{\beta_1(t-2t)}{2(1-\alpha)} \\
&= \frac{\beta_1}{2(1-\alpha)} - \frac{\beta_1 t}{2(1-\alpha)}\,.
\end{aligned}$$

Thus, the discrepancy between the approximate sample mean and the theoretical mean follows a linear trend, with intercept equal in absolute value and opposite in sign to its slope.

8 Effects of the Presence of Unit Roots in Regression Estimations

In this chapter, we examine some key aspects essential for the correct specification of regression models.
The first aspect concerns spurious regression also known as nonsense regression, due to the fact that empirical results on time series can lead to accepting the presence of relationships between stochastic processes actually inexistent (§ 8.1).
The second aspect is closely linked to the first and concerns the estimation of autoregressive coefficients in an AR process using OLS, as a way to test for the presence of unit roots in the data-generating process. (§ 8.2). The usual Student's t-statistic is not the appropriate statistic in detecting the presence of unit roots; its use leads the test to decide in favor of their absence when this is not true. Therefore, it is necessary to refer to non-standard distributions to correct the testing procedure (§ 8.3).
The third aspect is equally relevant and concerns the reduction of estimation errors in regressions between cointegrated processes in the presence of unit roots (§8.5) In the subsequent paragraphs (from § 8.6 to §8.12) we introduce a particular form of regression, known as regression with Error Correction Mechanism (ECM), proposed in the literature to overcome some critical issues analyzed in the previous sections.

8.1 Spurious Regression

A spurious regression happens when performing a regression between time series generated by two stochastically independent processes Y_t and X_t , regression such as:

$$Y_t = \alpha + \beta X_t + u_t \,, \tag{8.1}$$

where u_t is assumed to be white noise, and we obtain significant estimates for the β coefficient and R^2 values substantially above zero.
This result happens very often when the series have unit roots, but there can be a spurious regression also for stationary time series, as is evident from the simulations reported below.

Consider the simulations of time series using the following data-generating process (DGP):

$$\begin{cases} Y_t = aY_{t-1} + \varepsilon_t \\ X_t = bX_{t-1} + \eta_t \end{cases} , \tag{8.2}$$

where $\varepsilon_t \sim NID(0,\, 0.49)$ is stochastically independent of the process $\eta_t \sim NID(0,\, 0.25)$. The condition of independence between errors implies stochastic independence between Y_t and X_t.
The following examples simulating data generated by the model (8.2) and applying the regression (8.1) seem to contradict this evidence.

Example 8.1. *($a = b = 0.8$, n = 10,000 simulations, sample size T=150)*
Having generated 10,000 series of $\{X_t, Y_t\}$ pairs from the model (8.2), regressions are carried out according to the model (8.1) and for each regression, the significance of the OLS estimates of the β parameter is assessed. Since the processes $\{Y_t\}$ and $\{X_t\}$ are stochastically independent, the value of parameter estimates should be not significant in any regression except for a small percentage of samples not representative of generating processes. The test is performed using the Student's t-statistic by calculating the ratio $\hat{t} = \dfrac{\hat{\beta}}{se(\hat{\beta})}$.
If a significance level of 5% has been fixed, the critical region w, corresponding to the random variable $\hat{t}$, consists of the interval $w = \left\{ \hat{t} : \ |\hat{t}| > t_{g,\alpha/2} \right\}$, where g are the degrees of freedom and $t_{148,0.025} = 1.976$. Out of 10,000 regressions, the null hypothesis $H_0 : \ \beta = 0$ should be rejected about 500 times, obtaining R^2 values near zero.

The test is bilateral, therefore the following regions are defined:

- acceptance region: if $\hat{t}$ belongs to the $(-t_{g,\alpha/2},\ t_{g,\alpha/2})$ interval;
- rejection region: if $\hat{t}$ belongs to the $(-\infty, -t_{g,\alpha/2})$ or $(\ t_{g,\alpha/2},\ \infty)$ interval.

The correct acceptance region for the null hypothesis should be determined by the theoretical distribution that reflects the empirical behavior of the $\hat{\beta}_1$ estimates (blue curve in *Figure 8.1*). However, in econometric practice it is common to rely on the standard Student's t distribution (red curve in *Figure 8.1*), which does not capture the actual distribution of the test statistic in this context. This discrepancy explains why the test tends to reject the null hypothesis too often.

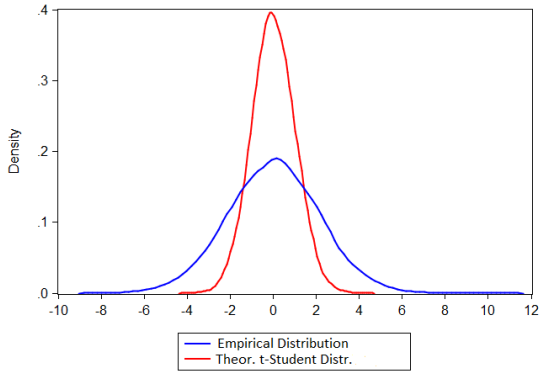


Figure 8.1: Comparison between empirical and theoretical distribution (T=150, n=10.000, a=b=0.8)

With the simulation, we obtain the following results:

- The values for which $|\hat{t}| > t_{148,0.025} = 1.976$ are 1,828, corresponding to 18.28% of the cases, a percentage well above the 5% threshold.

- The comparison between the empirical distribution of $\hat{t}$ values and the standard theoretical distribution of t_{148} is shown in *Figure 8.1*. The empirical distribution exhibits heavier tails than the theoretical one, which explains the exceedance beyond 5%.

- The average value of all R^2 indices, calculated for each regression, was equal to 0.028828.

In this example, based on the R^2 index, it can be concluded that the time series are independent: the index is only marginally affected by the independence between the series. Conversely, based on the Student's t-test, the hypothesis of linear dependence between the series is incorrectly accepted in a non-negligible number of cases.

Example 8.2. *($a = b = 1$, $n = 10{,}000$ simulations, sample size $T = 150$)*
By setting $a = b = 1$, two stochastically independent random walk processes are generated.
With the simulation, we obtain the following results:

- The values for which $|\hat{t}| > t_{148,0.025} = 1.976$ are 4,047, corresponding to 40.47% of the cases, a percentage greatly exceeding the expected 5%.

- The comparison between the empirical distribution of $\hat{t}$ values and the standard theoretical t_{148} distribution is shown in *Figure 8.2*. The empirical distribution is considerably flattened along the real axis compared to the standard theoretical Student's t-distribution.

- The average value of all R^2 indices, calculated for each regression, was equal to 0.245165. This value is significantly different from zero.

In conclusion, it can be said that, based on both the Student's t-statistic and the R^2 index, the hypothesis of linear dependence between series is erroneously accepted in an excessive number of cases.
If we compare the two examples, we observe that spurious regression can arise even in the stationary case, but it becomes particularly evident when time series originate from processes with unit roots.
The authors who most notably brought the issue of spurious regression in the presence of unit roots to the attention of econometricians and applied economists were Granger and Newbold (1974). In their work, they highlighted this issue using simulations in a way similar to what has been done in *Example 8.2* above.

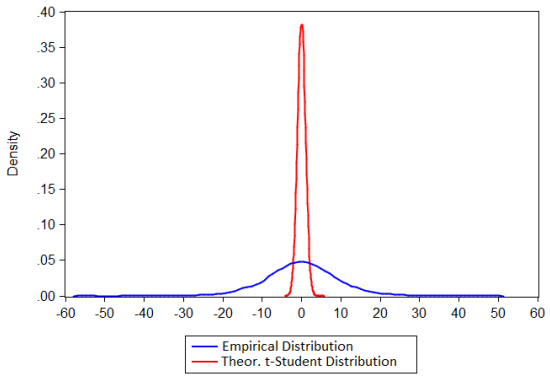


Figure 8.2: Comparison between empirical and theoretical distribution (T = 150, n = 10,000, a = b = 1)

However, important theoretical results are due to Phillips (1986), who developed an asymptotic theory of regressions concerning integrated processes in general, including the cases of spurious regression considered by Granger and Newbold.
Phillips starts from the observation that an appropriate transformation of a random walk has an asymptotic distribution as a process in continuous time, known as the *Wiener process*[93], and demonstrates a series of results.
The first result concerns the behavior of the Student's t-statistic, which does not have an asymptotic limit distribution and diverges as the sample size increases. Consequently, in asymptotic terms, there are no correct critical values for the usual significance levels (e.g., 5%). The significance levels continue to increase with the sample size.
A second result shows that the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ of the regression coefficients in (8.1) do not converge in probability to a constant as $T \to \infty$. In particular, the limit distribution as $T \to \infty$ of $\hat{\alpha}$ diverges, and that of $\hat{\beta}$ does not degenerate. Phillips identifies the reason for this behavior in the fact that such time series originate from stochastic processes that are not ergodic; therefore, the sample moments of these time series and their relative sample moments do not converge to constants, as they should for time series from ergodic processes. If the processes had no unit roots, both $\hat{\alpha}$ and $\hat{\beta}$ would converge in probability to zero.
A third result concerns the asymptotic behavior of the Durbin–Watson (DW) statistic[94] and the coefficient of determination R^2. As the sample size increases, the probability limit $\mathrm{DW}_T \to 0$, while R_T^2 has a non-degenerate limit distribution. For this reason, in cases of spurious regression, we can expect low values of the DW statistic and moderate values of the R_T^2 coefficient.

[93] For an introduction to the Wiener process, see *Appendix 8.A.*
[94] For an explanation of the DW test, see for example Verbeek (2017), §4.7.2, pp. 120–121 .

Another interesting consideration is the following.
Acceptance of the hypothesis $\beta_1 = 0$ would reduce model (8.1) to:

$$Y_t = \alpha + u_t \tag{8.3}$$

Since $Y_t \sim I(1)$ by hypothesis, from (8.3) it would necessarily follow that $u_t \sim I(1)$. This outcome violates the maintained hypothesis $u_t \sim I(0)$ and therefore contradicts the assumption underlying model (8.1). This shows an internal inconsistency in the usual hypothesis testing procedure: it is irreconcilable that $\beta_1 = 0$ and $u_t \sim I(0)$ hold simultaneously.
One reason for the biased behavior when using the Student's t-statistic is that the process generating Y_t is a random walk, but in estimating regression (8.1) a twofold specification error is made, given that the DGP comes from (8.2):

a) The relevant variable Y_{t-1} is omitted;

b) The irrelevant variable X_t is included.

From econometric theory, error (a) has more serious consequences than error (b). The consequences of (a) are biased parameter estimates and a biased estimate of the regression error variance, and therefore also a biased Student's t-statistic.
A different simulation was conducted, maintaining the generation of the series with the models defined in (8.2), but modifying the estimated regressions with the following alternative specification:

$$Y_t = \alpha + \rho Y_{t-1} + \beta X_t + u_t \tag{8.4}$$

In this way, specification error (a) is eliminated, but not error (b).
In any case, the presence of unit roots leads to an OLS estimate of the ρ parameter that is less than one in a significant number of cases (see §8.2), which also biases the estimate of β, since these estimators are correlated. However, eliminating the most serious specification error allows us to obtain the distributions shown in *Figure 8.3*:

Davidson and MacKinnon also carried out simulations, summarized in *Figure 8.4*.
In this figure, the term "Valid regression" refers to the estimation of parameters using specification (8.4). The horizontal axis shows the number of observations (n) from 20 to 20,000, while the vertical axis shows the proportion of times that the Student's t-statistic for $\beta = 0$ rejected the null hypothesis at the 5% level, as a function of n.
The graph presents four curves corresponding to the following cases:

- *Estimated model: Spurious regression; DGP: random walk ($a = b = 1.0$)* Simulation using model (8.2) with $a = b = 1.0$, while estimation uses model (8.1). *Figure 8.4* (curve: Spurious regression, random walk) shows that, as the sample size increases, the rejection frequency of the null hypothesis does not tend to zero but approaches one. One explanation, given by the authors, is that it is easy to reject a false hypothesis (the model in (8.1) is false) when the alternative is also false.

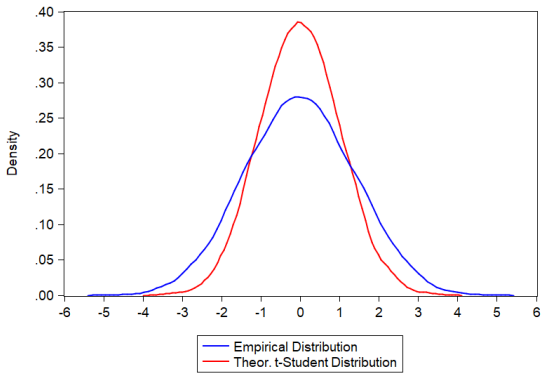


Figure 8.3: Comparison between empirical and theoretical distribution (T = 150, n = 10,000) using model (8.4)

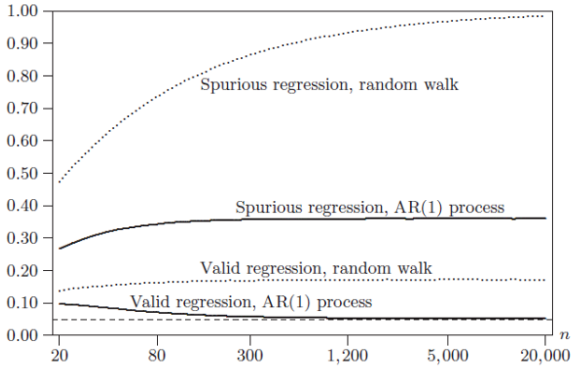


Figure 8.4: Rejection frequencies for spurious and valid regressions (n = 1,000,000)

- *Estimated model: Spurious regression; DGP: AR(1) ($a = b = 0.8$)* Simulation using model (8.2) with $a = b = 0.8$ (stationary series, no unit roots), while estimation uses model (8.1). Here, *Figure 8.4* (curve: Spurious regression, AR(1) process) shows that the rejection frequency does not tend to one, but remains substantially higher than 0.05. This is mainly due to model specification error: in (8.1), neither the constant nor X_t have explanatory power for the dependent variable. Moreover, under the null hypothesis, the model error is not white noise but an AR(1) process, preventing OLS from providing correct estimates of the variance–covariance matrix of the estimators (especially of β) and, consequently, of the Student's t-statistic. In such cases, it is preferable to estimate the covariance matrix using the Newey–West procedure[95], also known as *HAC (Heteroscedasticity and Autocorrela-*

[95] Newey and West (1987).

tion Consistent) estimates. Stationarity, therefore, does not rule out the possibility of a spurious regression.

- *Estimated model: Valid regression; DGP: random walk (a = b = 1.0)* Simulation using model (8.2) with $a = b = 1.0$, while estimation uses model (8.4). Here, the rejection frequency no longer tends to one but remains substantially higher than 0.05. This behavior must be attributed to unit roots, whose presence alters the asymptotic distribution of the Student's t-statistic for $\beta = 0$.

- *Estimated model: Valid regression; DGP: AR(1) (a = b = 0.8)* In this case, the rejection frequencies tend to converge to 0.05 as the sample size increases.

Conclusions
Spurious regression can occur both in the presence and in the absence of unit roots. In both cases, the Student's t-statistic is affected by:

a) Testing a false null hypothesis against an equally false alternative;

b) The need to correct the covariance matrix of the estimators (HAC estimates).

Moreover, the presence of unit roots alters the Student's t-statistic even when the regression is correctly specified, because its distribution is theoretically non-standard (*Figure 8.3* is an empirical example).
In the past, many authors have highlighted the absurdity of certain regressions, such as the proportion of Church of England marriages to all marriages (1866–1911) and the standardised mortality per 1,000 people in the same period in England[96].
However, only with the work of Granger and Newbold (1974) did the community of econometricians and applied economists fully recognise the relevance of the problem. In the absence of unit roots, it is not easy to identify a spurious regression; we must rely on procedures that lead to a correct specification, assuming that economic theory supports a meaningful relationship between dependent and explanatory variables. In the presence of unit roots, the risk of spurious regression is very high: we can find apparent linear relationships even when none exist. To avoid this incorrect conclusion, it is not always necessary to eliminate the unit roots from the series, as shown by the case of cointegration between stochastic processes (and, consequently, between their realisations[97]).

8.2 Unit Root Tests: An Inappropriate Test

Another important aspect that should be underlined is the following. Consider the AR(1) model: $Y_t = \rho Y_{t-1} + \varepsilon_t$, where ε_t is a white noise process. In the presence of unit roots,

[96] Cited in Yule (1926).
[97] See §8.6.

it has been shown that the Student's t-statistic for testing the hypothesis $\rho = 1$ does not converge in distribution to the standard normal distribution. Furthermore, the Student's t-statistic is not appropriate for testing the presence of unit roots. Consider the following example.

Example 8.3. Generation of 10,000 replications of Y_t with 150 observations, using the model:

$$Y_t = \rho Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0,\, 0.25) \tag{8.5}$$

The hypotheses to be tested are:

$$\begin{cases} H_0 : \rho = 1 \\ H_1 : \rho < 1 \end{cases} \tag{8.6}$$

The null hypothesis H_0 indicates the presence of a unit root, while the alternative hypothesis H_1 refers to its absence.
The distribution associated with the $\rho = 1$ test considers the transformation $T(\hat{\rho} - \rho)$ and should therefore be centered at zero. Indeed, from *Figure 8.5* we can see that, compared with the theoretical distribution of the Student's t-statistic with 148 degrees of freedom, the empirical distribution of the $\hat{\rho}$ estimates is clearly shifted to the left.

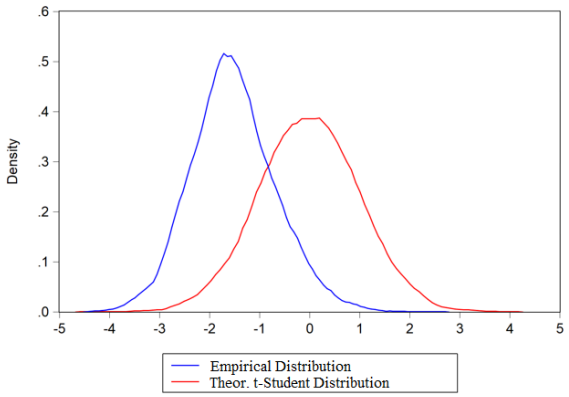


Figure 8.5: Regression of Y_t (random walk) on Y_{t-1} (150 observations, 10,000 simulations).

For example, if we choose a 5% significance level, the corresponding critical region $w = t_{148} < -1.655$ would be used to test the hypothesis $\rho = 1$ against the alternative $\rho < 1$. In this simulation, however, the null hypothesis is rejected 4,541 times instead of the expected 500 times, that is, with a frequency about nine times higher. In fact, in this case the significance level should be no lower than 45.41%.
In conclusion, the test is biased in favor of the alternative hypothesis of absence of a unit root (stationarity). *Figure 8.5* clearly shows this bias, as the empirical curve is shifted to the left with respect to the theoretical curve. Furthermore, in this case we are dealing with a one-sided (left-tail) test, so the critical regions are:

- Acceptance region for H_0: $(-t_\alpha, \infty)$
- Rejection region for H_0: $(-\infty, -t_\alpha)$

Therefore, also in this case, the use of the standard test (red curve) leads to a narrower acceptance region than that which would be determined by the theoretical distribution consistent with the empirical curve[98].

8.3 Unit Root Tests: DF, ADF

The econometric literature has produced several proposals for testing the presence of unit roots in time series. As seen in the previous section on spurious regression, one of the reasons for test failure is assuming that H_0 is true when it is actually false, together with an alternative hypothesis H_1 that is also false.
For the moment, we will assume that either the null hypothesis H_0 is true or the alternative hypothesis H_1 is true.

Dickey–Fuller (DF) and Augmented Dickey–Fuller (ADF) Tests

To avoid the failure of the Student's t-statistic test for the hypothesis system (8.6), based on regression (8.5), Dickey and Fuller (1979) proposed modifying expression (8.5) as follows.
In expression (8.5), subtract Y_{t-1} from both sides, obtaining:

$$\begin{aligned} Y_t - Y_{t-1} &= \rho Y_{t-1} - Y_{t-1} + \varepsilon_t \\ \Delta Y_t &= (\rho - 1)\, Y_{t-1} + \varepsilon_t \quad , \\ \Delta Y_t &= \alpha Y_{t-1} + \varepsilon_t \end{aligned} \tag{8.7}$$

where $\alpha = \rho - 1$.
The hypothesis system becomes:

$$\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha < 0\,. \end{cases} \tag{8.8}$$

If the test favors H_0, we accept the presence of a unit root in the stochastic process generating the series. If we reject H_0 in favor of H_1, we accept the absence of a unit root. Note that if H_0 is true, regression (8.7) has a dependent variable $I(0)$ and an explanatory variable that remains $I(1)$. This is an *unbalanced regression*, which becomes *balanced* under the alternative hypothesis. This balance allows analysis of the asymptotic distribution of the test, as shown by Dickey and Fuller using the Wiener process[99]. Their test is no

[98] A theoretical explanation of the failure of the test in the presence of a unit root is in Hamilton (1994), p. 488.
[99] See, for example, Fuller (1995), Corollary 10.1.1.2, p. 554.

longer referred to as a Student's t-statistic but as a τ-statistic, although it has the same functional form:

$$\hat{\tau} = \frac{\hat{\alpha}}{se(\hat{\alpha})}\,, \tag{8.9}$$

where $se(\hat{\alpha})$ is the standard error of $\hat{\alpha}$. This statistic has a well-defined asymptotic distribution, which is no longer normal and is asymmetric.
The asymptotic distribution is derived assuming no autocorrelation in the disturbances of (8.7). If the OLS residuals are autocorrelated, the number of regressors is increased:

$$\Delta Y_t = \alpha Y_{t-1} + \beta_1 \Delta Y_{t-1} + \cdots + \beta_p \Delta Y_{t-p} + \varepsilon_t\,, \tag{8.10}$$

where lag p must be large enough to remove autocorrelation in the residuals but not so large as to reduce test power.
The test statistic remains (8.9) but is now referred to as the ADF test. The choice of p may depend on data frequency: for quarterly data, $p = 4$ or a multiple of 4 may be chosen to capture possible residual seasonality (even if seasonally adjusted). A *general-to-specific* strategy can then be applied, removing non-significant lagged differences ΔY_{t-j} (starting from the largest lag) based on Student's t-tests, and stopping when the residuals show no evidence of autocorrelation (e.g., when the Durbin–Watson statistic is close to 2 and/or the Ljung–Box test[100] fails to reject the null of no autocorrelation).
When deterministic components such as a constant or linear trend are present, the DF reference regressions are:

$$\begin{aligned} &a)\ \Delta Y_t = c_0 + \alpha Y_{t-1} + \varepsilon_t \\ &b)\ \Delta Y_t = c_0 + c_1 t + \alpha Y_{t-1} + \varepsilon_t\,, \end{aligned} \tag{8.11}$$

where c_0 is the intercept and $c_1 t$ the trend.
Deterministic components affect the asymptotic distribution of the ADF statistic[101].
Dickey and Fuller simulated the critical values for (8.9) under (8.7) and (8.11), reported in *Table 8.2*. The corresponding test cases are summarized in *Table 8.1*[102].
For example, with $T = 100$ observations and a regression including a constant and trend, the 1%, 5%, and 10% critical values are -4.04, -3.45, and -3.15, respectively.
Some software (e.g., EViews) also report an approximate p-value $\Pr(\tau \geq \hat{\tau})$ based on MacKinnon's (1996)[103] response surface estimates. This allows testing without fixing a

[100] For the Durbin–Watson statistic see, for example, Verbeek (2017), §4.7.2, pp. 120–121. and for Ljung–Box test, see §8.7.3, p.319.
[101] In some special cases, the distribution is unaffected, leading to *similar tests* (distribution independent of *nuisance parameters*). See Banerjee et al. (1993), p. 104.
[102] See Hamilton (1994), p. 502.
[103] MacKinnon, J.G. (1996), "Numerical Distribution Functions for Unit Root and Cointegration Tests", *Journal of Applied Econometrics*, 11, 601–618.

Case 1		
Estimated regression:	$\Delta Y_t = \alpha Y_{t-1} + \varepsilon_t$	$\alpha = \rho - 1$
DGP of process:	$Y_t = Y_{t-1} + \varepsilon_t$	$\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$
DF critical values are described under *Case 1* in *Table 8.2*		
Case 2		
Estimated regression:	$\Delta Y_t = c_0 + \alpha Y_{t-1} + \varepsilon_t$	$\alpha = \rho - 1$
DGP of process:	$Y_t = Y_{t-1} + \varepsilon_t$	$\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$
DF critical values are described under *Case 2* in *Table 8.2*		
Case 3		
Estimated regression:	$\Delta Y_t = c_0 + \alpha Y_{t-1} + \varepsilon_t$	$\alpha = \rho - 1$
DGP of process:	$Y_t = c_0 + Y_{t-1} + \varepsilon_t,\ c_0 \neq 0$	$\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$
In this case, the asymptotic distribution of the DF τ-statistic is the same as in *Case 2* (i.e., the presence of drift in the DGP does not change the DF critical values when a constant is included in the test regression).		
Case 4		
Estimated regression:	$\Delta Y_t = c_0 + c_1 t + \alpha Y_{t-1} + \varepsilon_t$	$\alpha = \rho - 1$
DGP of process:	$Y_t = c_0 + Y_{t-1} + \varepsilon_t$	$\forall c_0,\ \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$
DF critical values are described under *Case 4* in *Table 8.2*		

Table 8.1: Summary of Dickey-Fuller Test for unit roots in the absence of serial correlation

significance level in advance: the closer the p-value is to zero, the more H_0 (unit root) is rejected; the closer to one, the more it is accepted.
The values in *Table 8.2* are unchanged when moving from DF to ADF tests, i.e., in regressions:

$$\begin{aligned}
\Delta Y_t &= \alpha Y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta Y_{t-j} + \varepsilon_t \\
\Delta Y_t &= c_0 + \alpha Y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta Y_{t-j} + \varepsilon_t \\
\Delta Y_t &= c_0 + c_1 t + \alpha Y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta Y_{t-j} + \varepsilon_t
\end{aligned} \tag{8.12}$$

In practice, the researcher may not know whether H_0 or H_1 is true, and in many cases both may be false. There is no universally accepted rule for choosing among the different

Sample Size T	Probability that $\hat{\tau} = \hat{\alpha}/se(\hat{\alpha})$ is less than the value shown in the cells							
	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
Case 1								
25	-2.66	-2.26	-1.95	-1.60	0.92	1.33	1.70	2.16
50	-2.62	-2.25	-1.95	-1.61	0.91	1.31	1.66	2.08
100	-2.60	-2.24	-1.95	-1.61	0.90	1.29	1.64	2.03
250	-2.58	-2.23	-1.95	-1.62	0.89	1.29	1.63	2.01
500	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.62	2.00
∞	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.62	2.00
Case 2								
25	-3.75	-3.33	-3.00	-2.63	-0.37	0.00	0.34	0.72
50	-3.58	-3.22	-2.93	-2.60	-0.40	-0.03	0.29	0.66
100	-3.51	-3.17	-2.89	-2.58	-0.42	-0.05	0.26	0.63
250	-3.46	-3.14	-2.88	-2.57	-0.42	-0.06	0.24	0.62
500	-3.44	-3.13	-2.87	-2.57	-0.43	-0.07	0.24	0.61
∞	-3.43	-3.12	-2.86	-2.57	-0.44	-0.07	0.23	0.60
Case 4								
25	-4.38	-3.95	-3.60	-3.24	-1.14	-0.80	-0.50	.-0.15
50	-4.15	-3.80	-3.50	-3.18	-1.19	-0.87	-0.58	-0.24
100	-4.04	-3.73	-3.45	-3.15	-1.22	-0.90	-0.62	-0.28
250	-3.99	-3.69	-3.43	-3.13	-1.23	-0.92	-0.64	-0.31
500	-3.98	-3.68	-3.42	-3.13	-1.24	-0.93	-0.65	-0.32
∞	-3.96	-3.66	-3.41	-3.12	-1.25	-0.94	-0.66	-0.33

Table 8.2: Critical values for Dickey–Fuller (DF) and Phillips–Perron (PP) unit root tests, based on the OLS t-statistic.
(*The probabilities marked at the head of each column are the left-tail probabilities at the indicated critical values*)

specifications. A commonly adopted general-to-specific procedure proceeds as follows:

1. Start from the most general model (Case 4), which includes both a constant and a deterministic trend.
2. Test the significance of the trend coefficient (c_1). If it is not statistically significant, re-estimate the model without the trend term (Case 2).
3. Then test the significance of the constant term (c_0). If it is also not significant, estimate the model without it (Case 1).

Including irrelevant deterministic terms reduces the test's power and increases the risk that both H_0 and H_1 are misspecified.

Several other tests for unit roots are implemented in econometric software. For example, EViews includes:

- *Dickey–Fuller Test with GLS Detrending (DFGLS)*;
- *Phillips–Perron (PP)*;
- *Kwiatkowski–Phillips–Schmidt–Shin (KPSS)*;
- *Elliot, Rothenberg, and Stock Point Optimal (ERS)*;
- *Ng–Perron (NP) Tests.*

If the *ADF* test result is uncertain, a second test (e.g., *PP*) can be used to confirm or reject the presence of a unit root and possible deterministic components.

8.4 Multiple Unit Roots

The case in which the test favors the H_0 hypothesis may occur even if the series is characterized by the presence of multiple unit roots. It is relevant for econometric analysis to verify whether the generating process of the series is $I(1)$ or $I(2)$. For this reason, it is suggested to repeat the ADF test on the first differences of the original series to test whether ΔY_t is $I(1)$.
If the original series exhibits a deterministic trend, the differenced series will remove it only if the trend is linear. If the trend is a polynomial of degree $k > 1$, first differencing reduces it to a polynomial of degree $k - 1$ (for example, a quadratic trend becomes linear after first differencing). Similarly, if the original series contains only a constant term, first differencing will remove it entirely; if it contains both a constant and a trend, the constant will disappear but the trend component may remain and must be treated accordingly.
The presence of multiple unit roots in economic time series is very rare, and therefore conclusions in favor of a double presence must be drawn with caution.
It is also advisable to examine the correlogram of the differenced series to assess whether applying the second-order differencing operator leads to excessive over-differencing of the series[104].

8.5 Superconsistency

The presence of unit roots can have some advantages from the point of view of regression estimation. An advantage, outside the cases of spurious regression, is known as *superconsistency*. This property is illustrated in the following example.

[104] As shown in Chapter 7, over-differencing can lead to an increase in the variance of the process and may induce a non-invertible MA component in the model.

Example 8.4. Consider the generation of Y_t data based on the following model:

$$\begin{cases} Y_t = 1 + 2X_t + u_t, \\ u_t = 0.7u_{t-1} + \eta_t, \quad \eta_t \sim NID(0,\, 0.4), \\ X_t = 0.6X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0,\, 2.25), \end{cases} \tag{8.13}$$

and a second series of data based on the alternative model:

$$\begin{cases} Y_t = 1 + 2X_t + u_t, \\ u_t = 0.7u_{t-1} + \eta_t, \quad \eta_t \sim NID(0,\, 0.4), \\ X_t = X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0,\, 2.25), \end{cases} \tag{8.14}$$

where in both models ε_t and η_t are normally distributed, stochastically independent errors. In (8.13) the Y_t and X_t processes are both $I(0)$, while in (8.14) they are $I(1)$ and cointegrated[105].
If the model is specified as:

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \tag{8.15}$$

and Y_t is regressed on X_t, the residuals will be autocorrelated whether the data come from (8.13) or from (8.14). In both cases, the regression will yield inefficient OLS estimates of (8.15), particularly for the slope parameter β_1.
Figure 8.6 shows the average value of $(\hat{\beta}_1 - 2)$ as the sample size increases, computed over 300 simulations for each sample size from $T = 1$ to $T = 500$. The bias converges quickly to zero when the processes are cointegrated with unit roots, but much more slowly when the processes are stationary.

We now provide a theoretical explanation of this result.
To simplify the notation without loss of generality, replace (8.15) with:

$$Y_t = \beta_1 X_t + u_t. \tag{8.16}$$

The OLS estimator of β_1 is:

$$\hat{\beta}_{1T} = \frac{\sum_{t=1}^{T} Y_t X_t}{\sum_{t=1}^{T} X_t^2}. \tag{8.17}$$

Substituting (8.16) into (8.17):

$$\hat{\beta}_{1T} = \beta_1 + \frac{\sum_{t=1}^{T} u_t X_t}{\sum_{t=1}^{T} X_t^2}. \tag{8.18}$$

[105] In this example, Y_t and X_t are cointegrated because there exists a linear combination that is an $I(0)$ process, i.e., without a unit root. This combination is, by construction, $u_t = Y_t - 1 - 2X_t$, which is stationary. In model (8.14) the unit root in Y_t is induced by that in X_t.

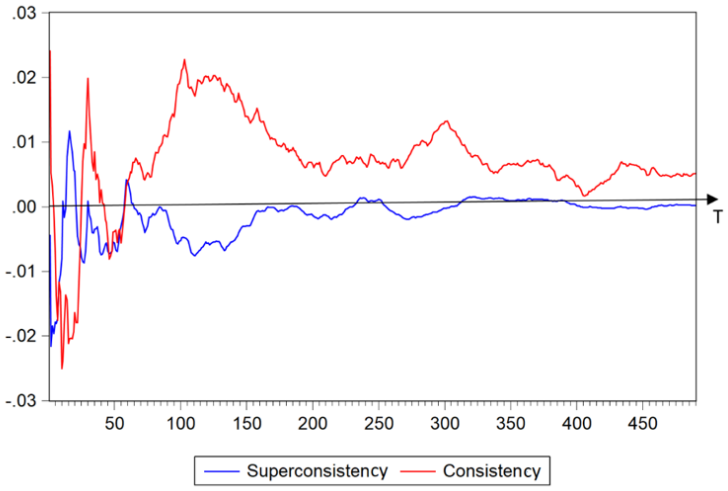


Figure 8.6: Regression of Y_t on X_t, with autocorrelated errors. Average bias $(\hat{\beta}_1 - 2)$ from 300 simulations.

Thus:

$$\hat{\beta}_{1T} - \beta_1 = \frac{\sum_{t=1}^{T} u_t X_t}{\sum_{t=1}^{T} X_t^2}. \tag{8.19}$$

If u_t is stationary and X_t is also $I(0)$, then[106]:

$$z_T = \sqrt{T}(\hat{\beta}_{1T} - \beta_1) \xrightarrow[T\to\infty]{\mathcal{D}} N\left(0,\, \sigma_u^2 \frac{1-\rho^2}{\sigma_\varepsilon^2}\right), \tag{8.20}$$

since the variance of the AR(1) process is $\sigma_\varepsilon^2/(1-\rho^2)$. In (8.13), Example 8.4, this variance equals:

$$\frac{2.25}{1-0.6^2} = 3.5156.$$

If $\rho = 1$, as in (8.14), $\sqrt{T}(\hat{\beta}_{1T} - \beta_1)$ converges in probability to zero. This is useless for significance tests that require non-degenerate distributions.
Referring to the Wiener process, it can be shown[107] that a non-degenerate distribution is obtained by considering $T(\hat{\beta}_{1T} - \beta_1)$. This is the well-known *Superconsistency Theorem* of Stock (1987).

[106] Relation (8.20) follows from the general convergence in distribution of an OLS estimator:

$$\mathbf{z}_T = \sqrt{T}\left(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\right) \xrightarrow[T\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0},\, \sigma^2 \boldsymbol{\Sigma}_{xx}^{-1}\right),$$

where $\sigma^2 \boldsymbol{\Sigma}_{xx}^{-1}$ is the asymptotic variance of the regressors.
For simplicity, this asymptotic result is stated under standard regularity conditions (including weak dependence of u_t and appropriate exogeneity of X_t).
[107] See Hamilton (1994), p. 483.

In the presence of a unit root, the convergence rate is T rather than $\sqrt{T}$. Hence, for integrated and cointegrated processes, the estimators are called *superconsistent*.
This important property will be used in the next section on the Engle–Granger two-step estimation approach.

8.6 *ADL(1,1)* and ECM Model: Engle–Granger Two-Step Estimation Procedure

The third aspect discussed in the previous section concerning superconsistency allows a more in-depth analysis of the static and dynamic relationships between stochastic processes. It is useful to illustrate this through a simple dynamic model of the *ADL(1,1)* class:

$$Y_t = \alpha Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t, \tag{8.21}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ and $X_t \sim I(1)$.
Under the stability condition $|\alpha| < 1$, it is possible to find an equivalent transformation of (8.21). First, subtract Y_{t-1} from both sides:

$$\begin{aligned} Y_t - Y_{t-1} &= \alpha Y_{t-1} - Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t \\ \Delta Y_t &= (\alpha - 1) Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t. \end{aligned}$$

Next, add and subtract $\beta_0 X_{t-1}$ on the right-hand side:

$$\Delta Y_t = \beta_0 \Delta X_t + (\alpha - 1) Y_{t-1} + (\beta_0 + \beta_1) X_{t-1} + \varepsilon_t.$$

Factor out $(\alpha - 1)$:

$$\Delta Y_t = \beta_0 \Delta X_t + (\alpha - 1) \left[Y_{t-1} + \frac{\beta_0 + \beta_1}{\alpha - 1} X_{t-1} \right] + \varepsilon_t.$$

Finally, by changing the sign inside the brackets:

$$\Delta Y_t = \beta_0 \Delta X_t + (\alpha - 1) \left[Y_{t-1} - \frac{\beta_0 + \beta_1}{1 - \alpha} X_{t-1} \right] + \varepsilon_t. \tag{8.22}$$

The model in (8.22) is an equivalent transformation of (8.21) and is called the *error correction mechanism (ECM)* form.

Why consider the ECM form (8.22) instead of the original ADL form (8.21)?

The main motivation is the following.
As discussed in §8.1, if $Y_t \sim I(1)$ and $X_t \sim I(1)$, then estimating the ADL form(8.21) does not guarantee that significant OLS estimates are not the result of a spurious regression.

Under the same assumptions, the ECM form (8.22) reduces the risk of spurious regression because the regression is performed with first-differenced variables, which are stationary. However, in the ECM specification, other issues must be addressed.
The process

$$u_{t-1} = Y_{t-1} - \frac{\beta_0 + \beta_1}{1-\alpha} X_{t-1},$$

defined by the term in square brackets, is a new regressor. Equation (8.22) can be rewritten as:

$$\Delta Y_t = \beta_0 \Delta X_t + (\alpha - 1) u_{t-1} + \varepsilon_t.$$

Remark 8.1. The u_{t-1} process is a linear combination of Y_t and X_t. From an integration standpoint, is $u_{t-1} \sim I(1)$ or $u_{t-1} \sim I(0)$?
For the ECM to represent a *balanced regression*, it is necessary that $u_{t-1} \sim I(0)$.
This condition ensures that the steady-state relation introduced in *Chapter 5* corresponds to a valid stochastic equilibrium.
How do we verify the integration order of u_{t-1}?

Remark 8.2. The ratio

$$k = \frac{\beta_0 + \beta_1}{1-\alpha}$$

is the long-run coefficient when the system (8.21) is in equilibrium (steady state), with constant inputs and outputs X and Y respectively. In the absence of shocks, substituting the equilibrium values into (8.21) yields:

$$Y = \alpha Y + \beta_0 X + \beta_1 X,$$

from which:

$$Y = \frac{\beta_0 + \beta_1}{1-\alpha} X = kX. \tag{8.23}$$

Using the definition $u_{t-1} = Y_{t-1} - kX_{t-1}$, we can write:

$$Y_{t-1} = kX_{t-1} + u_{t-1},$$

and for any t:

$$Y_t = kX_t + u_t. \tag{8.24}$$

Equation (8.24) is the static regression of Y_t on X_t, with u_t as the error term. If the system is stable and the equilibrium relation holds, then from (8.23), the steady-state values satisfy Y - kX = 0. From (8.24) we have $Y_t - kX_t = u_t$, and thus u_t must have unconditional mean zero. If also $E(u_t|X_t) = 0$, then u_t can be interpreted as the deviation of Y_t from the equilibrium value $kX_t = E(Y_t|X_t)$.

Remark 8.3. The meaning of k is completely different from that of β_0, even though both multiply X_t. The coefficient estimated via static regression should be interpreted as k, the long-run coefficient of X_t, and **not** as β_0, which would be biased due to omission of Y_{t-1} and X_{t-1}.

Remark 8.4. The component u_{t-1} is unobservable because the coefficient $(\beta_0+\beta_1)/(1-\alpha)$ is unknown. If we estimate k by OLS from (8.24), the residuals $\hat{u}_t$ may be autocorrelated, violating standard regression assumptions. This is evident in the equivalent form of (8.21):

$$\begin{cases} Y_t = \beta_0 X_t + u_t, \\ u_t = \alpha Y_{t-1} + \beta_1 X_{t-1} + \varepsilon_t. \end{cases} \tag{8.25}$$

If the DGP follows (8.25) and we use static regression, we omit α and β_1, so u_t retains information that often leads to autocorrelation in $\hat{u}_t$. However, if $X_t \sim I(1)$ and $Y_t \sim I(1)$ are cointegrated, then $\hat{k}$ from (8.24) is superconsistent, converging quickly to k. In this case, kX_t can be interpreted as the expected equilibrium value. This is one of the results of the *Engle–Granger Theorem* in §8.9, which proposes a two-step estimation: (1) estimate (8.24) and check that $\hat{u}_t$ is $I(0)$; (2) estimate the ECM (8.22) replacing u_{t-1} with $\hat{u}_{t-1}$. Monte Carlo studies show that in small samples the bias in the estimated cointegrating relation may be substantial, unless R^2 is close to unity without artificial inclusion of regressors[108].

Remark 8.5. In general, the OLS estimator of the cointegrating parameter has a non-standard distribution, so inference based on its t-statistic may be misleading.
A favourable case occurs when X_t is *strictly exogenous* in (8.24), i.e.

$$E(\Delta X_{t-j} u_t) = 0, \quad -\infty < j < \infty.$$

In this case, the estimator's distribution is *asymptotically mixed normal*, and standard inference is asymptotically valid. Given that residuals are typically autocorrelated, robust standard errors can be obtained using a *heteroskedasticity and autocorrelation consistent (HAC) estimator*[109]. If (8.24) is interpreted as the *long-run relationship* between Y_t and X_t, the ECM (8.22) represents the *short-run relationship* between ΔY_t and ΔX_t, with short-run dynamics also depending on deviations of Y_t from its expected path. These deviations contribute to explaining changes in Y_t via the *adjustment parameter* $(\alpha - 1)$, which measures the proportion of disequilibrium corrected each period.
The stability condition $|\alpha| < 1$ implies $-2 < (\alpha - 1) < 0$. Thus, as the system approaches instability, the proportion of disequilibrium affecting ΔY_t remains large, even when deviations from the equilibrium path are substantial.

[108] See Banerjee et al. (1993), *op. cit.*, §7.4.
[109] See Davidson (2013).

8.7 ECM Model as a Transformation of the *ADL(p,q)*

In the previous section, we considered the transformation of the *ADL(1,1)* model with only one exogenous variable into its equivalent ECM form. We now extend the transformation to an *ADL(p,q)* model with more general dynamics.
Consider the model:

$$\begin{cases} \alpha(L)Y_t = \beta(L)X_t + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, \sigma_\varepsilon^2), \\ \alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p, \\ \beta(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q \end{cases} \tag{8.26}$$

To obtain the equivalent ECM form, it is convenient to use the following polynomial decomposition (*Beveridge–Nelson decomposition*)[110]:

$$\begin{aligned} \alpha(L) &= \alpha(1)L + \alpha^*(L)\Delta, \\ \beta(L) &= \beta(1)L + \beta^*(L)\Delta, \end{aligned} \tag{8.27}$$

where:

$$\begin{aligned} \Delta &= 1 - L, \\ \alpha^*(L) &= \alpha_0^* - \alpha_1^* L - \alpha_2^* L^2 - \cdots - \alpha_{p-1}^* L^{p-1}, \\ \beta^*(L) &= \beta_0^* + \beta_1^* L + \beta_2^* L^2 + \cdots + \beta_{q-1}^* L^{q-1}. \end{aligned} \tag{8.28}$$

The relationships between the coefficients of $\alpha^*(L)$ and $\beta^*(L)$ and those of $\alpha(L)$ and $\beta(L)$ are given in *Table 8.3*.

$\alpha_0^* = 1,$	$\beta_0^* = \beta_0,$
$\alpha_1^* = -\sum_{j=1}^{p-1} \alpha_{j+1},$	$\beta_1^* = -\sum_{j=1}^{q-1} \beta_{j+1},$
$\vdots$	$\vdots$
$\alpha_k^* = -\sum_{j=1}^{p-k} \alpha_{j+k},$	$\beta_k^* = -\sum_{j=1}^{q-k} \beta_{j+k},$
$\vdots$	$\vdots$
$\alpha_{p-1}^* = -\alpha_p$	$\beta_{q-1}^* = -\beta_q$

Table 8.3: α_k^* and β_k^* in terms of α_k and β_k coefficients.

The Beveridge–Nelson decomposition avoids lengthy algebraic manipulations when transforming an ADL model into ECM form. For example, consider $\alpha(L)$ of order $p = 3$:

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \alpha_3 L^3. \tag{8.29}$$

[110] See Beveridge and Nelson (1981).

By repeatedly adding and subtracting equal terms on the right-hand side, we obtain:

$$\begin{aligned}
\alpha(L) &= 1-\alpha_1 L-\alpha_2 L^2-\alpha_3 L^3+\alpha_3 L^2-\alpha_3 L^2 \\
&= 1-\alpha_1 L-(\alpha_2+\alpha_3)L^2+\alpha_3 L^2(1-L)+(\alpha_2+\alpha_3)L-(\alpha_2+\alpha_3)L \\
&= 1-(\alpha_1+\alpha_2+\alpha_3)L+(\alpha_2+\alpha_3)L(1-L)+\alpha_3 L^2(1-L)+L-L \\
&= \alpha(1)L+(1-L)+(\alpha_2+\alpha_3)L(1-L)+\alpha_3 L^2(1-L) \\
&= \alpha(1)L+\Delta+(\alpha_2+\alpha_3)L\Delta+\alpha_3 L^2\Delta \\
&= \alpha(1)L+\alpha^*(L)\Delta.
\end{aligned} \tag{8.30}$$

Thus:

$$\alpha^*(L)=\alpha_0^*-\alpha_1^* L-\alpha_2^* L^2, \quad \alpha_0^*=1,\ \alpha_1^*=-(\alpha_2+\alpha_3),\ \alpha_2^*=-\alpha_3.$$

These values are easily obtained using the Beveridge–Nelson decomposition.
In general, replacing (8.27) into (8.26) yields:

$$\alpha(1)LY_t+\alpha^*(L)\Delta Y_t=\beta(1)LX_t+\beta^*(L)\Delta X_t+\varepsilon_t,$$

and therefore the ECM form:

$$\alpha^*(L)\Delta Y_t=\beta^*(L)\Delta X_t-\alpha(1)\left[Y_{t-1}-kX_{t-1}\right]+\varepsilon_t, \tag{8.31}$$

with the long-run coefficient:

$$k=\frac{\beta(1)}{\alpha(1)}=\frac{\beta_0+\beta_1+\cdots+\beta_q}{1-\alpha_1-\cdots-\alpha_p}. \tag{8.32}$$

Expanding $\alpha^*(L)$ in (8.31) gives:

$$\begin{aligned}
\Delta Y_t &= \alpha_1^*\Delta Y_{t-1}+\alpha_2^*\Delta Y_{t-2}+\cdots+\alpha_{p-1}^*\Delta Y_{t-p+1} \\
&+ \beta_0^*\Delta X_t+\beta_1^*\Delta X_{t-1}+\cdots+\beta_{q-1}^*\Delta X_{t-q+1} \\
&- \alpha(1)\left[Y_{t-1}-kX_{t-1}\right]+\varepsilon_t.
\end{aligned} \tag{8.33}$$

As in the *ADL(1,1)* case:

(1) The ECM representation in (8.33) yields a balanced regression only if $u_t=Y_t-kX_t$ is $I(0)$.

(2) *Table 8.3* provides the α_k^* values given α_k. The inverse relationship is also useful: given α_k^* and $\alpha(1)$, one can recover α_k for the original ADL(p,q) form. This allows moving from the OLS estimates of the ECM parameters to those of the ADL(p,q) coefficients.

With some algebraic manipulation (starting from 8.29), the inverse relationships are given in *Table 8.4*.

$$\begin{cases} \alpha_0 = 1 \\ \alpha_1 = 1 - \alpha(1) + \alpha_1^* \\ \alpha_j = \alpha_j^* - \alpha_{j-1}^*, \quad j = 2, \ldots, p-1 \\ \alpha_p = -\alpha_{p-1}^* \end{cases} \qquad \begin{cases} \beta_0 = \beta_0^* \\ \beta_1 = \alpha(1)k + \beta_1^* - \beta_0^* \\ \beta_j = \beta_j^* - \beta_{j-1}^*, \quad j = 2, \ldots, q-1 \\ \beta_q = -\beta_{q-1}^* \end{cases} \tag{8.34}$$

Table 8.4: α_k and β_k in terms of the coefficients α_k^* and β_k^*

8.8 ECM Model with *m* Exogenous Variables

In this section, the ECM model is generalized considering a number $m > 1$ of exogenous variables. The *ADL(p,q)* model is given in this case by:

$$\alpha(L)Y_t = \beta_1(L)X_{1t} + \cdots + \beta_m(L)X_{mt} + \varepsilon_t \,. \tag{8.35}$$

Using the polynomial decomposition we obtain:

$$\begin{array}{rcl} \alpha^*(L)\Delta Y_t & = & \beta_1^*(L)\Delta X_{1t} + \cdots + \beta_m^*(L)\Delta X_{mt} \\ & - & \alpha(1)\left[Y_{t-1} - k_1 X_{1,t-1} - \cdots - k_m X_{m,t-1}\right] + \varepsilon_t \end{array} \,, \tag{8.36}$$

with long-run coefficients $k_1 = \dfrac{\beta_1(1)}{\alpha(1)}, \cdots, k_m = \dfrac{\beta_m(1)}{\alpha(1)}$.
Explicitly, the (8.36) can be rewritten as follows:

$$\begin{array}{rcl} \Delta Y_t & = & \sum_{j=1}^{p-1} \alpha_j^* \Delta Y_{t-j} \; + \; \sum_{j=0}^{q-1} \beta_{1j}^* \Delta X_{1,t-j} \; + \; \cdots \; + \; \sum_{j=0}^{q-1} \beta_{mj}^* \Delta X_{m,t-j} \\ & & - \alpha(1)[Y_{t-1} - k_1 X_{1,t-1} - \cdots - k_m X_{m,t-1}] + \varepsilon_t \end{array} \,. \tag{8.37}$$

In the model (8.37) we assume, for simplicity, that all the polynomials are of the same degree $q - 1$.

8.9 Engle-Granger Representation Theorem

The Superconsistent Theorem of J. H. Stock mentioned in §8.5 and the representation in ECM form of the cointegrated processes led Engle and Granger to formulate the following theorem::

Theorem 8.1. *(Engle-Granger, 1987)*[111]

[111] See Engle and Granger (1987). A sketch-proof of Engle-Granger Theorem (Bivariate Case) is found in Banerjee et al. (1993), p. 159.

The two-step estimator of a single equation of an error-correction system with one co-integrating vector, obtained by taking the estimate $\hat{k}$ of k from the static regression in place of the true value for estimation of the error-correction form at a second stage, will have the same limiting distribution as the maximum-likelihood estimator using the true value of k. Least-squares standard errors in the second stage will provide consistent estimates of the true standard errors.

The result of the theorem allows us to implement the following estimation procedure.

1) For each of the variables in the specification of the model, it is necessary to verify whether they are $I(0)$, $I(1)$, or $I(2)$ with an appropriate test (for example the ADF test). The specification of the model depends on the determination of this preliminary analysis.

2) If it happens that all the variables are $I(1)$, then the first step of the E-G procedure suggests estimating (with the OLS method) the coefficients of the static regression:

$$Y_t = c + k_1 X_{1t} + k_2 X_{2t} + \cdots + k_m X_{mt} + u_t \tag{8.38}$$

In this way, we obtain:

$$Y_t = \hat{c} + \hat{k}_1 X_{1t} + \hat{k}_2 X_{2t} + \cdots + \hat{k}_m X_{mt} + \hat{u}_t \tag{8.39}$$

It is essential to test whether the residuals $\hat{u}_t$ can be thought of as the realization of the $I(0)$ process (e.g., using an ADF test)..

If the residuals are $I(0)$, then the $\{Y_t, X_{1t}, \cdots, X_{mt}\}$ variables are cointegrated and the regression obtained in (8.39) avoids the possibility of being a spurious regression. In addition, the ECM model is consistent in the sense that all the variables that enter the equation (8.37) are $I(0)$.

All the variables in the first difference are $I(0)$, and the residual variable, which is used to replace the unobservable regressor $\hat{u}_{t-1}$ is also $I(0)$. Therefore, we can proceed to the second step of the E-G procedure which provides for the estimate of the parameters of the equation (8.37) with the OLS method.

The model specification becomes:

$$\Delta Y_t = \sum_{j=1}^{p-1} \alpha_j^* \Delta Y_{t-j} + \sum_{j=0}^{q-1} \beta_{1j}^* \Delta X_{1t-j} + \cdots + \sum_{j=0}^{q-1} \beta_{mj}^* \Delta X_{mt-j} + \gamma \hat{u}_{t-1} + \varepsilon_t \,, \tag{8.40}$$

where $\gamma = -\alpha(1)$ represents the *error correction coefficient*. For the stability of the model, this coefficient must be inside the interval[112] $(-2, 0)$.

[112] The stability of the model requires that $\sum_{j=1}^{p-1} \alpha_j < 1$ as a necessary (but not sufficient) condition. This is the same condition already seen in §2.5.1 regarding the stationarity of a stochastic process.

3) If the test on the $\hat{u}_{t-1}$ residuals rejects the hypothesis of $I(0)$ behavior, then the specification of the model in ECM form no longer makes sense.

   Suppose we do not want to give up the analysis of the variables in levels. In that case, we can try to change the model specification, for example, inserting further explanatory variables in the regression.

   Alternatively, the analysis of the variables in levels must be renounced and the relationship between the dependent and the explanatory variables with a simple *model in the first differences* can be studied, that is:

$$\Delta Y_t = \sum_{j=1}^{p-1} \alpha_j \Delta Y_{t-j} + \sum_{j=0}^{q-1} \beta_{1j} \Delta X_{1t-j} + \cdots + \sum_{j=0}^{q-1} \beta_{mj} \Delta X_{mt-j} + \varepsilon_t \tag{8.41}$$

8.10 Forecasting with ECM models

In this section, we consider forecasting with ECM models through a simulation example. Consider the following $ADL(1,1)$ and the corresponding ECM models:

$$\begin{aligned} ADL(1,1): &\quad Y_t = c + \alpha Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t \\ ECM: &\quad \Delta Y_t = \beta_0 \Delta X_t + (\alpha - 1)\left[Y_{t-1} - k X_{t-1}\right] + \varepsilon_t \\ &\quad k = \frac{\beta_0 + \beta_1}{1 - \alpha} \end{aligned} \tag{8.42}$$

with $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.
The advantage of a simulation exercise lies in the knowledge of the theoretical parameter values of regression coefficients and the variance of the errors. Holding the realization of X_t fixed and exogenous in (8.42), these theoretical values can be changed to see their effects on estimations and forecasts.
Furthermore, since the Y_t variable is generated conditionally on the $X_t \sim I(1)$ variable using the model $ADL(1,1)$, there is no need to verify the cointegration between X_t and Y_t. The Y_t variable inherits a unit root from the X_t variable and the two processes are cointegrated.
The estimation applied here is Engle–Granger's two-step procedure.
In the first step, we estimate the static equation:

$$Y_t = d + kX_t + u_t\,, \tag{8.43}$$

by OLS. The $\hat{k}$ is the long-run coefficient and the estimation is superconsistent. The $\hat{u}_t$ residuals constitute the variable to be included, lagged one period, in the second-step dynamic equation:

$$\Delta Y_t = \beta_0 \Delta X_t + \alpha^* \hat{u}_{t-1} + \varepsilon_t, \quad \alpha^* = \alpha - 1\,. \tag{8.44}$$

Some software (for example EViews) allows the use of the explicit difference that defines $\hat{u}_t$, that is:

$$\Delta Y_t = \beta_0 \Delta X_t + \alpha^* \left[Y_{t-1} - \hat{k} X_{t-1} \right] + \varepsilon_t \,. \tag{8.45}$$

Both specifications (8.44) and (8.45) give the same estimation results. This is not true for forecasting. Therefore, it is convenient to refer to (8.44) as dynamic regression with *Implicit Error Correction Mechanism (I-ECM)* and to (8.45) as dynamic regression with *Explicit Error Correction Mechanism (E-ECM)*.
We distinguish *in-sample forecasts* from *out-of-sample forecasts*.
In-sample forecasting uses the first subset of available sample data (*training set*) for estimation. The last part of the available sample data (*validation set*) is used to compare actual and forecast values. The validation set refers to a virtual future, which is a useful device to test the forecast capabilities. In-sample forecasting also has the advantage of using the actual "future" values of the exogenous variables as if they had been foreseen without any prediction error (*perfect foresight*).
On the other hand, out-of-sample forecasting uses all available sample data without partition into sub-periods and does not implement any comparison. In this case, the values of the exogenous variables must also be provided, usually with statistical models in the class of the $ARIMA(p, d, q)$ models.
For out-of-sample forecasting, errors concerning the endogenous variable contain more components, such as:

1. Model specification errors;
2. Errors in estimating the regression coefficients;
3. Errors in estimating the variance of the regression error;
4. The three previous errors referring to the forecast of the exogenous variables.

A further distinction concerns *one-step* (or *static*) and *multi-step* (or *dynamic*) forecasts. In the one-step forecast, only one period is foreseen and in the subsequent period we update the forecast by taking into account the actual value of the endogenous variable (and of the exogenous variables if we are considering the out-of-sample forecast). In the one-step forecast, the $\widehat{\Delta Y}_t$ differences have the following decomposition:

$$\widehat{\Delta Y}_t = \hat{Y}_t - Y_{t-1}, \quad t = 2, 3, \ldots \tag{8.46}$$

from which $\hat{Y}_2 = Y_1 + \widehat{\Delta Y}_2$, $\hat{Y}_3 = Y_2 + \widehat{\Delta Y}_3$, and so on.
In a multi-step forecast, at each step we use the forecast value of the previous period. In the case of out-of-sample forecasting, even for the exogenous variables the replacement takes place at each step with the values forecasted for the whole future period instead of

actual values. In the multi-step forecast, the $\widehat{\Delta Y}_t$ differences have the following decomposition:

$$\widehat{\Delta Y}_t = \tilde{Y}_t - \tilde{Y}_{t-1}, \quad t = 2, 3, \ldots \tag{8.47}$$

from which $\tilde{Y}_2 = \tilde{Y}_1 + \widehat{\Delta Y}_2$, $\tilde{Y}_3 = \tilde{Y}_2 + \widehat{\Delta Y}_3$, and so on, with initial value $\tilde{Y}_1 = Y_1$.
Note that $\hat{Y}_2 = \tilde{Y}_2$, while in general $\hat{Y}_t \neq \tilde{Y}_t$ for $t > 2$.
This section refers only to the in-sample forecast.
For the model in ECM form, it is possible to calculate the one-step and multi-step forecasts of both the Y_t levels and the ΔY_t differences.
The multi-step forecast is not possible if the regression does not contain the lagged endogenous variable.
For example, regarding the model (8.44), OLS estimation can be written[113]:

$$\widehat{\Delta Y}_t = \hat{c} + \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* \hat{u}_{t-1}, \quad t = 2, \ldots, T \tag{8.48}$$

If the forecast concerns the variable in differences, i.e. ΔY_t, then only the one-step forecast is possible because in (8.48) the lagged variable of ΔY_t does not appear.
The forecast values of the differences for the h time horizons are given by:

$$\widehat{\Delta Y}_{T+1} = \hat{c} + \hat{\beta}_0 \Delta X_{T+1} + \hat{\alpha}^* \hat{u}_T, \quad \text{(one-step only)} \tag{8.49}$$

On the other hand, if the forecast concerns the variable Y_t in levels, then both one-step and multi-step forecasts can be computed. This is possible because the differences obtained from decomposition (8.46) in the first case, or (8.47) in the second, can be used.
To highlight the difference between level forecasts of Y_t, let $\tilde{Y}_t$ denote the I-ECM forecast and $\hat{Y}_t$ the E-ECM forecast. Consider the following.
For the *I-ECM* forecast, we refer to specification (8.48), so the level forecasts are obtained from:

$$\widehat{\Delta Y}_t = \tilde{Y}_t - \tilde{Y}_{t-1} = \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* \hat{u}_{t-1} + \hat{c}, \quad \hat{\alpha}^* = \hat{\alpha} - 1\,. \tag{8.50}$$

Solving for $\tilde{Y}_t$ gives:

$$\begin{aligned} \tilde{Y}_t &= \tilde{Y}_{t-1} + \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* \hat{u}_{t-1} + \hat{c} \\ &= \tilde{Y}_{t-1} + \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* Y_{t-1} - \hat{\alpha}^* \hat{k} X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c} \\ &= \tilde{Y}_{t-1} + \hat{\alpha}^* Y_{t-1} + \hat{\beta}_0 X_t + \hat{\beta}_1 X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c}\,, \end{aligned} \tag{8.51}$$

where $\hat{\alpha}^* \hat{k} = (\hat{\alpha} - 1) \dfrac{\hat{\beta}_0 + \hat{\beta}_1}{1 - \hat{\alpha}} = -(\hat{\beta}_0 + \hat{\beta}_1)$, and $\hat{d}$ is the estimated intercept of the long-run relationship.

[113] Unlike the model (8.44), in the model (8.48) the estimated constant is inserted to obtain OLS residuals with zero sum.

For the *E-ECM* forecast, we use specification (8.45), in which the deviation from the long-run equilibrium is included directly in the dynamic regression (all variables lagged one period). Here, Y_{t-1} is replaced by its forecast $\hat{Y}_{t-1}$, yielding:

$$\widehat{\Delta Y}_t = \hat{Y}_t - \hat{Y}_{t-1} = \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* [\hat{Y}_{t-1} - \hat{k} X_{t-1} - \hat{d}] + \hat{c}, \quad \hat{\alpha}^* = \hat{\alpha} - 1. \tag{8.52}$$

Solving for $\hat{Y}_t$ gives:

$$\begin{aligned} \hat{Y}_t &= \hat{Y}_{t-1} + \hat{\beta}_0 \Delta X_t + \hat{\alpha}^* \hat{Y}_{t-1} - \hat{\alpha}^* \hat{k} X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c} \\ &= \hat{\alpha} \hat{Y}_{t-1} + \hat{\beta}_0 X_t + \hat{\beta}_1 X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c}\,. \end{aligned} \tag{8.53}$$

The difference between the *I-ECM* and *E-ECM* forecasts is:

$$\begin{aligned} \tilde{Y}_t - \hat{Y}_t &= \tilde{Y}_{t-1} + \hat{\alpha}^* Y_{t-1} + \hat{\beta}_0 X_t + \hat{\beta}_1 X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c} \\ &\quad - \hat{\alpha} \hat{Y}_{t-1} - \hat{\beta}_0 X_t - \hat{\beta}_1 X_{t-1} + \hat{\alpha}^* \hat{d} + \hat{c} \\ &= \tilde{Y}_{t-1} + \hat{\alpha}^* Y_{t-1} - \hat{\alpha} \hat{Y}_{t-1} \\ &= \tilde{Y}_{t-1} + (\hat{\alpha} - 1) Y_{t-1} - \hat{\alpha} \hat{Y}_{t-1} \\ &= \hat{\alpha} (Y_{t-1} - \hat{Y}_{t-1}) - (Y_{t-1} - \tilde{Y}_{t-1})\,. \end{aligned} \tag{8.54}$$

Thus, the two forecasts coincide only if $\hat{\alpha}$ approaches unity, i.e., when the error correction mechanism becomes completely ineffective. The characteristics of the two forecasts differ significantly. In the *E-ECM* forecast, replacing the lagged dependent variable with its predicted value ensures the effectiveness of the error correction mechanism, keeping the forecast trajectory close to the long-run equilibrium curve — the curve defined by the static regression.
Writing both recursions in a compact form clarifies the difference:

$$\begin{aligned} \tilde{Y}_t &= \tilde{Y}_{t-1} + \hat{\alpha}^* Y_{t-1} + r_t, \\ \hat{Y}_t &= \hat{\alpha} \hat{Y}_{t-1} + r_t, \end{aligned} \tag{8.55}$$

where $r_t = \hat{\beta}_0 X_t + \hat{\beta}_1 X_{t-1} - \hat{\alpha}^* \hat{d} + \hat{c}$.
By iterative substitution, the I-ECM forecast accumulates past values of the dependent variable:

$$\tilde{Y}_t = \tilde{Y}_0 + \hat{\alpha}^* \sum_{j=0}^{t-1} Y_j + \sum_{j=1}^{t} r_j\,. \tag{8.56}$$

where $\tilde{Y}_0$ is an arbitrary initial value.
Consequently, $\tilde{Y}_t$ behaves like a unit-root type accumulation and does not necessarily converge to the long-run equilibrium curve.

The E-ECM forecast admits the compact representation

$$(1-\hat{\alpha}L)\hat{Y}_t = r_t, \qquad \text{hence} \qquad \hat{Y}_t = \frac{r_t}{1-\hat{\alpha}L}. \tag{8.57}$$

If $|\hat{\alpha}| < 1$ the inverse can be expanded as the convergent power series

$$\hat{Y}_t = \sum_{j=0}^{\infty} \hat{\alpha}^j \, r_{t-j},$$

which highlights the stabilizing autoregressive effect of the E-ECM recursion.
Since the processes X_t and Y_t are cointegrated, the equilibrium property also applies to the steady-state forecast $\hat{Y}$::

$$\begin{aligned} \hat{Y} &= \frac{\hat{\beta}_0 + \hat{\beta}_1}{1-\hat{\alpha}} X + \hat{d} + \frac{\hat{c}}{1-\hat{\alpha}} \\ &= \hat{k}\, X + \hat{d} + \frac{\hat{c}}{1-\hat{\alpha}}\,. \end{aligned} \tag{8.58}$$

Thus, the *E-ECM* forecast shares the same steady-state solution as the Y_t process, up to a constant term whose relevance increases only if $\hat{\alpha}$ is close to one.
In conclusion, while the stochastic process Y_t fluctuates around the long-run curve, the error-correction mechanism ensures that this curve acts as an attracting reference path. In contrast, E-ECM forecasts are explicitly driven back toward the long-run equilibrium. The distinction between I-ECM and E-ECM clarifies why, in practice, the E-ECM specification is generally preferred for forecasting purposes, as it ensures convergence toward the long-run equilibrium.

Example 8.5. *(Simulated example)*
We consider a numerically simulated example with 200 observations. For the in-sample forecast, the estimation sub-sample is set to $T = 150$, while the forecast horizon covers periods 151 to 200.
First, we generate the exogenous variable $X_t \sim I(1)$ using the following $AR(4)$ model:

$$X_t = 0.85X_{t-1} + 0.35X_{t-2} + 0.05X_{t-3} - 0.25X_{t-4} + \eta_t, \quad \eta_t \sim WN(0,\, 16)\,. \tag{8.59}$$

Simulation of the X_t series requires assigning four initial values. For simplicity, we set them all to zero. In the $AR(4)$ model, no constant term is included. Adding a non-zero constant would allow the simulation of a deterministic trend component.
The Y_t process is generated as a function of X_t using the following *ADL(1,1)* model:

$$Y_t = 0.7Y_{t-1} + 2.5X_t - 3.6X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,\, 81)\,, \tag{8.60}$$

where the error process ε_t is stochastically independent of the η_s process for all t, s.

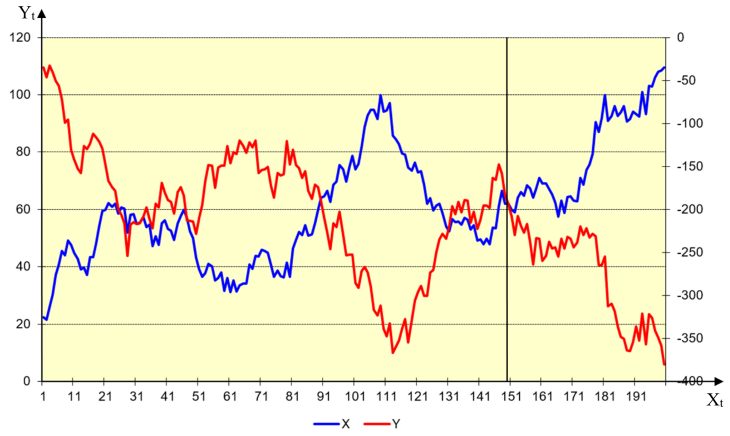


Figure 8.7: Simulation of cointegrated realizations from model (8.60). The vertical line separates the estimation sample from the forecast period.

We generate 200 observations from (8.60). The estimation sample includes the first $T = 150$ data points, while the last 50 are used to compare actual and forecasted data.[114]
The theoretical static regression is:

$$\begin{aligned} Y_t &= kX_t + u_t, \\ &= -3.667X_t + u_t, \quad u_t \sim I(0)\,. \end{aligned} \tag{8.61}$$

First step: we estimate the static regression by OLS. The results are reported in *Table 8.5.*
The conditions in *Remark 8.5* are satisfied, so the standard normal approximation is asymptotically valid. We use the HAC estimator to obtain consistent standard errors for the coefficients.
As expected, the intercept is not statistically significant.[115] Regarding the null hypothesis $H_0 : k = 0$, the estimate $\hat{k}$ is highly significant. For the null $H_0 : k = -3.667$, the result is not significant, with $\Pr(-0.327 < z < 0.327) = 0.256$, where $z \sim N(0, 1)$.
The Durbin–Watson statistic is 0.474, indicating at least first-order autocorrelation, as confirmed in *Figure 8.8.*
Since the bivariate vector (X_t, Y_t) is cointegrated by construction, testing the residuals for unit roots is unnecessary. Nevertheless, the DF test statistic equals -4.525, supporting

[114] The generated series initially contains 250 data points. The first 50 are discarded to reduce the impact of the arbitrary initial values, although unit-root processes never completely lose the memory of their initial state.
[115] Given the parameters in the *ADL(1,1)* process, the constant should not appear in the static model. However, in OLS estimation, including the intercept ensures that the residuals sum to zero.

Dependent Variable: Y
Method: Least Squares
Included observations: 150
HAC standard errors & covariance (Bartlett kernel, Newey-West fixed bandwidth = 5.0000)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.875126	21.00684	0.041659	0.9668
X	-3.546401	0.367842	-9.641107	0.0000

R-squared	0.694978	Mean dependent var	-195.5082
Adjusted R-squared	0.692917	S.D. dependent var	70.11766
S.E. of regression	38.85576	Akaike info criterion	10.17083
Sum squared resid	223446.0	Schwarz criterion	10.21098
Log likelihood	-760.8125	Hannan-Quinn criter.	10.18714
F-statistic	337.2107	Durbin-Watson stat	0.473736
Prob(F-statistic)	0.000000	Wald F-statistic	92.95094
Prob(Wald F-statistic)	0.000000		

Table 8.5: OLS estimates for the static regression (8.61).

the alternative hypothesis of stationarity[116].
The estimated long-run curve is displayed in *Figure 8.9*.

Second step: The theoretical form is:

$$\Delta Y_t = 2.5\,\Delta X_t - 0.3\,[Y_{t-1} + 3.667\,X_{t-1}] + \varepsilon_t\,. \tag{8.62}$$

The positive sign within the square brackets is due to the negative long-run coefficient.
The dynamic regression for model (8.62) is reported in *Table 8.6*.
Thus, the short-run regression in *I-ECM* form of model (8.62) is:
(t statistics in parentheses)

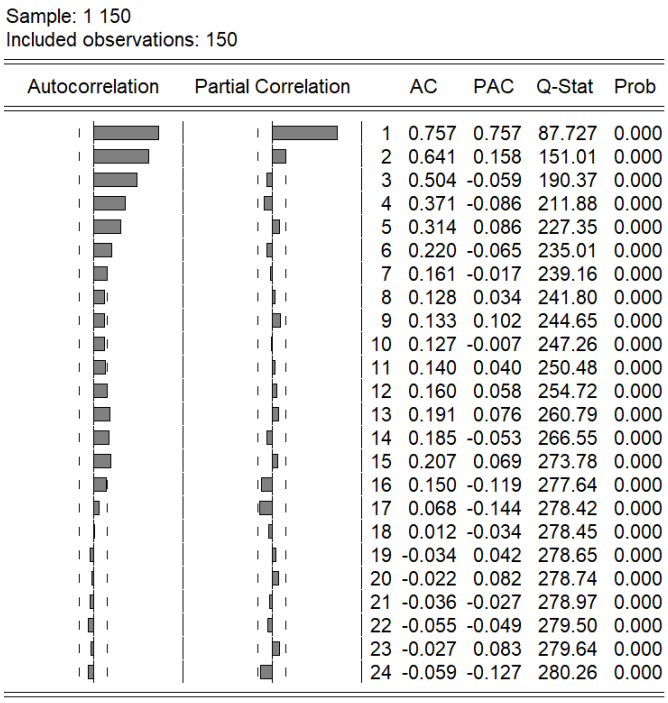
$$\widehat{\Delta Y}_t = \underset{(-2.56)}{-1.799} + \underset{(14.06)}{2.564}\ \Delta X_t - \underset{(-16.55)}{0.302}\ \hat{u}_{t-1}.$$

$$The equivalent E\text{-}ECM formis^{117} : \widehat{\Delta Y}_t = \underset{(-2.56)}{-1.799} + \underset{(14.06)}{2.564}\ \Delta X_t - \underset{(-16.55)}{0.302}\left[Y_{t-1} + \underset{(-9.64)}{3.546}\ X_{t-1}\right]. \tag{8.63}$$

The positive sign inside the square brackets is due to the negative long-run coefficient k.
As previously noted, both *I-ECM* and *E-ECM* provide identical estimation results.
The Durbin–Watson statistic is 2.05, indicating no first-order autocorrelation. *Figure 8.10* shows a correlogram consistent with a white noise process.

[116] The Dickey–Fuller statistic reported here refers to the specification with a constant (Case 2 in Table 8.1). Comparing the value -4.525 with the simulated critical values reported in Table 8.2 for that case, the statistic is more negative than the usual 5% and 1% critical values (approximately -2.9 and -3.5, respectively). Therefore, the null hypothesis of a unit root is rejected at conventional significance levels.

[117] Note that -3.667 is the theoretical long-run coefficient used to generate the data, whereas -3.546 is the OLS estimate obtained from the sample (*Table 8.5*).


Sample: 1 150
Included observations: 150

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.757	0.757	87.727	0.000
		2	0.641	0.158	151.01	0.000
		3	0.504	-0.059	190.37	0.000
		4	0.371	-0.086	211.88	0.000
		5	0.314	0.086	227.35	0.000
		6	0.220	-0.065	235.01	0.000
		7	0.161	-0.017	239.16	0.000
		8	0.128	0.034	241.80	0.000
		9	0.133	0.102	244.65	0.000
		10	0.127	-0.007	247.26	0.000
		11	0.140	0.040	250.48	0.000
		12	0.160	0.058	254.72	0.000
		13	0.191	0.076	260.79	0.000
		14	0.185	-0.053	266.55	0.000
		15	0.207	0.069	273.78	0.000
		16	0.150	-0.119	277.64	0.000
		17	0.068	-0.144	278.42	0.000
		18	0.012	-0.034	278.45	0.000
		19	-0.034	0.042	278.65	0.000
		20	-0.022	0.082	278.74	0.000
		21	-0.036	-0.027	278.97	0.000
		22	-0.055	-0.049	279.50	0.000
		23	-0.027	0.083	279.64	0.000
		24	-0.059	-0.127	280.26	0.000

Figure 8.8: Correlogram of residuals from the static regression.

Residuals are also normally distributed, as indicated by the Jarque–Bera test in *Figure 8.11*.
The short-run curve is shown in *Figure 8.12*.
The regressions (8.61), (8.62), and (8.63) are then used to produce different forecasts. Their comparison is illustrated in *Figure 8.13*.
A broader comparison, from $T = 2$ to $T = 200$, is given in *Figure 8.14*. The vertical line separates the estimation sample from the forecasting period.
Undoubtedly, the best forecast is provided by regression (8.62) or equivalently (8.63), corresponding to the one-step dynamic prediction procedure. In this case, the actual value of Y_t at each time step acts as an anchor, preventing the forecast from deviating significantly.
The comparison between the *I-ECM* and *E-ECM* forecasts is noteworthy. While the latter closely follows the path of the actual values, the former appears entirely disconnected from the actual curve.
We also compare each forecast curve with the actual values using the *Theil's U inequality coefficient*. As is well known, the closer this coefficient is to zero, the nearer the forecasts are to the actual values. The coefficients are shown in *Table 8.7*.

To explore the behavior of the curves along the entire time span, forecasts are computed from $T = 2$ to $T = 200$. The representation is shown in *Figure 8.14*.

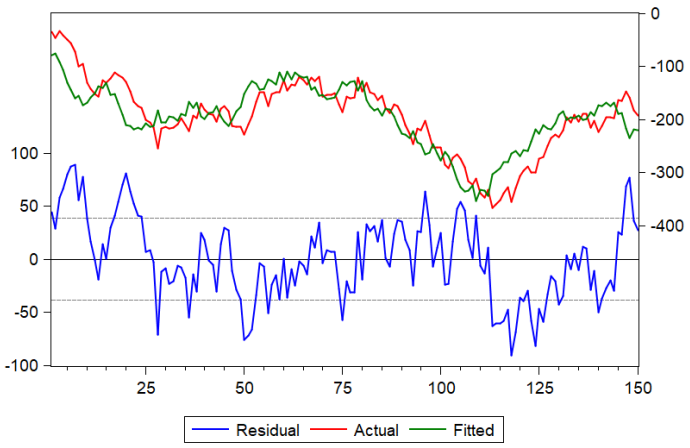


Figure 8.9: Estimated long-run curve from the static regression in *Table 8.5*.

The predictive path is thus divided into two segments. In the first, from $T = 2$ to $T = 150$, we compare both static and dynamic interpolations, where the *I-ECM* curve amplifies its deviations from the actual path. In the second, from $T = 151$ to $T = 200$, we present the in-sample forecast, which coincides exactly with the representation in *Figure 8.13*.
The Theil's U inequality coefficients, reported in *Table 8.8*, again confirm the superiority of the one-step forecasting procedure. Furthermore, all curves except *I-ECM* track the actual path.

An important point concerns the dynamic *I-ECM* interpolation.
This curve, despite diverging considerably from the actual series within the sample period ($t = 1, \ldots, 150$), converges towards the actual values at the end of the sample, satisfying $\widehat{Y}_{150} = Y_{150}$. By contrast, in the out-of-sample period ($t = 151, \ldots, 200$), a substantial and permanent divergence emerges.
The reason for this convergence at $T = 150$ is linked to the properties of OLS estimation. The relationship between the actual and fitted values in the sample period can be written as:

$$\Delta Y_t = \widehat{\Delta Y}_t + \hat{\varepsilon}_t,$$

where $\widehat{\Delta Y}_t$ is the fitted curve and $\hat{\varepsilon}_t$ the residuals. Summing over the sample period gives:

$$\sum \Delta Y_t = \sum \widehat{\Delta Y}_t + \sum \hat{\varepsilon}_t.$$

By the OLS property, when an intercept is included in the regression, we have:

$$\sum \hat{\varepsilon}_t = 0,$$

Dependent Variable: D(Y)
Method: Least Squares
Sample (adjusted): 2 150
Included observations: 149 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.798995	0.703211	-2.558259	0.0115
D(X)	2.564240	0.182410	14.05753	0.0000
ECM(-1)	-0.301841	0.018237	-16.55071	0.0000

R-squared	0.746110	Mean dependent var	-1.059504
Adjusted R-squared	0.742632	S.D. dependent var	16.87868
S.E. of regression	8.562799	Akaike info criterion	7.152660
Sum squared resid	10704.94	Schwarz criterion	7.213142
Log likelihood	-529.8732	Hannan-Quinn criter.	7.177233
F-statistic	214.5260	Durbin-Watson stat	2.050036
Prob(F-statistic)	0.000000		

Table 8.6: Dynamic regression for model (8.62).

Long-run	**I-ECM**	**E-ECM**	**One-step**
0.00040012	0.00097633	0.00015031	0.00011153

Table 8.7: Theil's U inequality coefficients for different forecasting curves.

and therefore:

$$\sum \Delta Y_t = \sum \widehat{\Delta Y}_t.$$

Expressed in terms of mean values:

$$\overline{\Delta Y_t} = \overline{\widehat{\Delta Y}_t}.$$

This equality constitutes a constraint that the fitted curve must satisfy in the sample period, regardless of how large its deviations from the actual series path are. In other words, $\widehat{\Delta Y}_t$ values are free to differ from ΔY_t except for one value that ensures the equality of the means (or sums). It can be said that $\widehat{\Delta Y}_t$ loses one degree of freedom due to this constraint.

8.11 Introduction to Multivariate Cointegration

Starting from §8.8, we introduce a more compact representation of the ECM form. Given the vectors:

$$\begin{cases} \mathbf{z}'_t &= \begin{bmatrix} Y_t & X_{1t} & \cdots & X_{mt} \end{bmatrix} \\ \pi'(L) &= \begin{bmatrix} \alpha^*(L) & -\beta_1^*(L) & \cdots & -\beta_m^*(L) \end{bmatrix} \\ \pi'(1) &= \begin{bmatrix} \alpha(1) & -\beta_1(1) & \cdots & -\beta_m(1) \end{bmatrix}, \end{cases}$$

Sample: 1 150
Included observations: 149
Q-statistic probabilities adjusted for 2 dynamic regressors

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob*
		1	-0.027	-0.027	0.1113	0.739
		2	-0.005	-0.005	0.1147	0.944
		3	0.139	0.139	3.0994	0.377
		4	0.020	0.028	3.1624	0.531
		5	0.088	0.092	4.3706	0.497
		6	-0.031	-0.046	4.5210	0.607
		7	0.021	0.013	4.5891	0.710
		8	0.122	0.099	6.9574	0.541
		9	0.083	0.100	8.0596	0.528
		10	0.061	0.062	8.6558	0.565
		11	0.070	0.055	9.4534	0.580
		12	0.099	0.076	11.077	0.522
		13	-0.062	-0.094	11.720	0.551
		14	-0.019	-0.050	11.779	0.624
		15	0.083	0.050	12.943	0.607
		16	0.060	0.069	13.548	0.632
		17	0.124	0.125	16.190	0.510
		18	-0.024	-0.031	16.286	0.573
		19	0.020	-0.030	16.357	0.633
		20	0.171	0.097	21.482	0.369
		21	0.071	0.089	22.366	0.379
		22	-0.050	-0.045	22.816	0.412
		23	0.008	-0.026	22.829	0.471
		24	-0.037	-0.092	23.072	0.516

*Probabilities may not be valid for this equation specification.

Figure 8.10: Correlogram of residuals from regression model (8.62).

L-Run	I-ECM	E-ECM	One-step
0.00070869	0.00129481	0.00023567	0.00016388

Table 8.8: Theil's U inequality coefficients for forecasting curves from $T = 2$.

where $\mathbf{z}_t \sim I(1)$, we can rewrite model (8.35) as:

$$\pi'(L)\,\Delta\mathbf{z}_t = -\pi'(1)\,\mathbf{z}_{t-1} + \varepsilon_t\,. \tag{8.64}$$

Expression (8.64) is perfectly equivalent to (8.36); it simply provides a more compact way of writing the ECM model. Within the vector $\mathbf{z}_t$ we can still distinguish between the endogenous variable Y_t and the exogenous variables $\{X_{jt},\, j = 1, \ldots, m\}$.
If we drop the hypothesis of exogeneity for the X variables, then representation (8.64)—based on a single equation—must be generalized to a system of $n = m + 1$ equations, adding m further equations, one for each now-endogenous X.
In this case, a suggestion from the econometric literature is to consider a VAR(p) representation:

$$\underbrace{\mathbf{\Pi}(L)}_{n\times n}\;\underbrace{\mathbf{z}_t}_{n\times 1} = \boldsymbol{\varepsilon}_t\,, \tag{8.65}$$

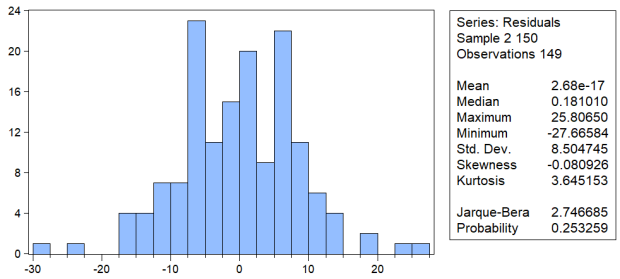


Figure 8.11: Empirical distribution of residuals from regression model (8.62).

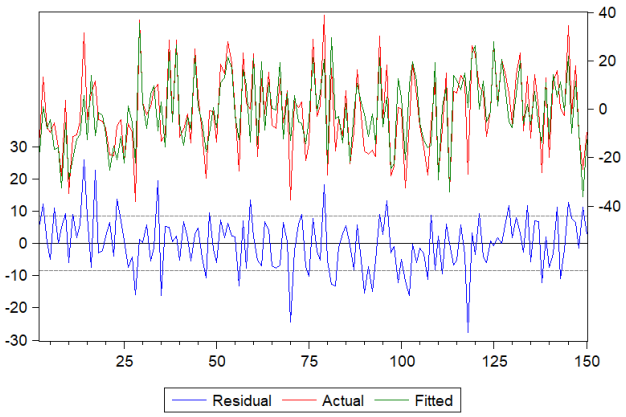


Figure 8.12: Short-run curve estimated from the dynamic regression.

where

$$\mathbf{\Pi}(L) = \mathbf{I} - \mathbf{\Pi}_1 L - \mathbf{\Pi}_2 L^2 - \cdots - \mathbf{\Pi}_p L^p \,.$$

If we assume that $\mathbf{z}_t \sim I(1)$, then some roots of $|\mathbf{\Pi}(L)| = 0$ lie on or outside the unit circle. Following Banerjee et al. (1993)[118], the VAR in levels can be rewritten in error-correction form as:

$$\begin{aligned} \Delta \mathbf{z}_t = \mathbf{\Pi}_1^* \Delta \mathbf{z}_{t-1} + \mathbf{\Pi}_2^* \Delta \mathbf{z}_{t-2} + \cdots + \mathbf{\Pi}_{p-1}^* \Delta \mathbf{z}_{t-p+1} \\ + \mathbf{\Pi}\, \mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t \end{aligned}$$

$$\begin{cases} \mathbf{\Pi}_i^* = -\sum_{j=i+1}^{p} \mathbf{\Pi}_j \,, \quad i = 1, \ldots, p-1 \\ \mathbf{\Pi} = \sum_{j=1}^{p} \mathbf{\Pi}_j - \mathbf{I} \end{cases} \tag{8.66}$$

[118] See Banerjee et al. (1993), p.147 ff.

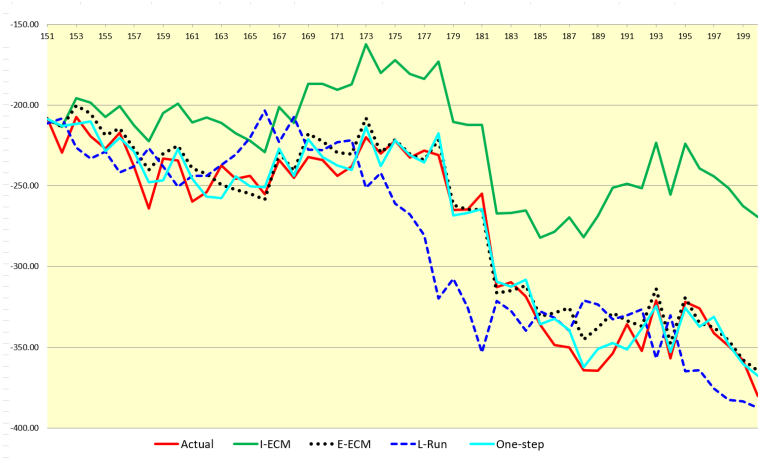


Figure 8.13: Graphical performance of different forecasting procedures.

Note that $\mathbf{\Pi} = -\mathbf{\Pi}(1)$. If we define the polynomial matrix

$$\mathbf{\Pi}^*(L) = \mathbf{I} - \mathbf{\Pi}_1^* L - \mathbf{\Pi}_2^* L^2 - \cdots - \mathbf{\Pi}_{p-1}^* L^{p-1} \,,$$

expression (8.66) can be rewritten as:

$$\mathbf{\Pi}^*(L)\, \Delta \mathbf{z}_t = -\mathbf{\Pi}(1)\, L\, \mathbf{z}_t + \boldsymbol{\varepsilon}_t \,. \tag{8.67}$$

From (8.67) and (8.65) we can write:

$$\big[\, \mathbf{\Pi}^*(L)\, \Delta + \mathbf{\Pi}(1)\, L \,\big] \mathbf{z}_t = \mathbf{\Pi}(L)\, \mathbf{z}_t = \boldsymbol{\varepsilon}_t \,, \tag{8.68}$$

showing that the square bracket contains a decomposition of $\mathbf{\Pi}(L)$ similar to the Beveridge–Nelson decomposition in the univariate case.
The matrix $\mathbf{\Pi}(1)$ plays a crucial role in cointegration analysis for systems of equations. Each element of $-\mathbf{\Pi}(1)\, \mathbf{z}_{t-1}$ is a linear combination of $I(1)$ processes. Row by row, we can check whether these linear combinations are $I(0)$ or $I(1)$.
If all are $I(1)$, then (8.67) is unbalanced: on the LHS we have $I(0)$ combinations while on the RHS they are $I(1)$. In such a case, applying decomposition (8.68) yields a meaningless representation.
The decomposition becomes relevant if some elements of $-\mathbf{\Pi}(1)\, \mathbf{z}_{t-1}$ are $I(0)$—that is, if there are cointegration relationships among the processes.
If $\mathbf{z}_t$ has dimension $n = 2$, there can be only one cointegration relationship; for $n > 2$, there may be multiple cointegrating vectors, i.e., multiple equilibrium relationships among the endogenous variables.
If the number of equilibrium relationships is $r < n$, then r is exactly the number of linearly independent cointegrating vectors. Engle and Granger (1987) show that in this

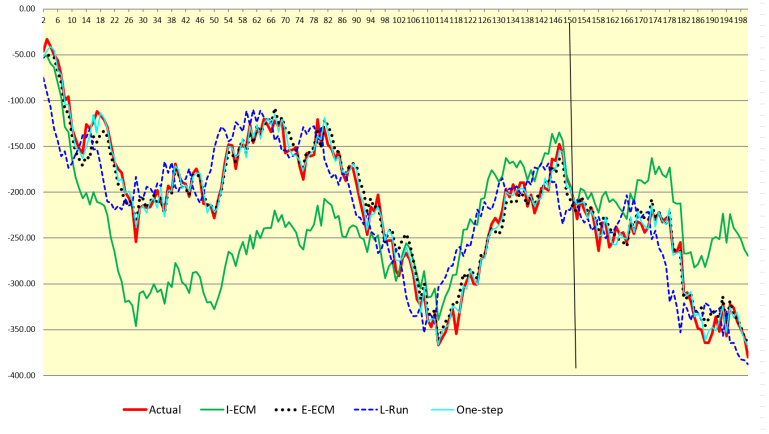


Figure 8.14: Comparison of curves over the full sample ($T = 2$ to $T = 200$). Vertical line separates estimation and forecast periods.

case $\text{rank}\big(\mathbf{\Pi}(1)\big) = r < n$, so that $\mathbf{\Pi}(1)$ can be factored as:

$$-\mathbf{\Pi}(1)_{n\times n} = \underbrace{\boldsymbol{\alpha}}_{n\times r}\underbrace{\boldsymbol{\beta}'}_{r\times n}, \tag{8.69}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(n \times r)$ matrices of rank r (the *cointegration rank*). The rows of $\boldsymbol{\beta}'$ form a basis for the r *cointegrating vectors*, while the elements of $\boldsymbol{\alpha}$, called the *matrix of adjustment coefficients*, are the factor loadings in the *Vector Error Correction Model* (VECM):

$$\mathbf{\Pi}^*(L)\,\Delta\mathbf{z}_t = \boldsymbol{\alpha}\,\boldsymbol{\beta}'\,\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t\,, \tag{8.70}$$

where $\mathbf{\Pi}^*(L)$ measures the *transitory effects*.
We consider two particular cases:

1) If $\text{rank}[\mathbf{\Pi}(1)] = 0$, then $\mathbf{\Pi}(1) = \mathbf{0}$, implying $\mathbf{z}_t \sim I(1)$ and not cointegrated. In this case, model (8.67) reduces to:

$$\mathbf{\Pi}^*(L)\,\Delta\mathbf{z}_t = \boldsymbol{\varepsilon}_t\,, \tag{8.71}$$

meaning that stability must be assessed on the processes in first differences only.

2) If $\text{rank}[\mathbf{\Pi}(1)] = n$, then $\mathbf{z}_t$ cannot be $I(1)$ but is stationary, i.e. $\mathbf{z}_t \sim I(0)$. In this case, $\mathbf{\Pi}(1)$ is invertible and we can write:

$$\mathbf{\Pi}(1)^{-1}\,\mathbf{\Pi}^*(L)\,\Delta\mathbf{z}_t = -\,\mathbf{z}_{t-1} + \mathbf{\Pi}(1)^{-1}\,\boldsymbol{\varepsilon}_t\,.$$

Since the LHS is $I(0)$, the RHS must also be $I(0)$.

Notice that the factorization in (8.69) is not unique. For any non-singular $r \times r$ matrix $\boldsymbol{Q}$:

$$\boldsymbol{\alpha}\,\boldsymbol{\beta}' = (\boldsymbol{\alpha}\,\boldsymbol{Q})\,(\boldsymbol{Q}^{-1}\,\boldsymbol{\beta}') = \boldsymbol{\alpha}^*\,\boldsymbol{\beta}^{*'}\,, \tag{8.72}$$

which creates an *identification problem*. Uniqueness requires imposing restrictions on the VECM representation.

Example 8.6. Consider the bivariate VAR(1) model for

$$\mathbf{z}_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix}\,,$$

so that:

$$\boldsymbol{\Pi}(L)\,\mathbf{z}_t = \boldsymbol{\varepsilon}_t\,, \quad (\mathbf{I} - \boldsymbol{\Pi}_1 L)\,\mathbf{z}_t = \boldsymbol{\varepsilon}_t\,.$$

Assume $\mathbf{z}_t \sim I(1)$ and cointegrated. The VECM form is:

$$\Delta\mathbf{z}_t = -\boldsymbol{\Pi}(1)\,\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t\,,$$

or, equivalently:

$$\Delta\mathbf{z}_t = (\boldsymbol{\Pi}_1 - \mathbf{I})\,\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t\,.$$

If $\mathbf{z}_t$ is cointegrated with one cointegrating vector, then $\text{rank}[\boldsymbol{\Pi}(1)] = 1$, and:

$$-\boldsymbol{\Pi}(1) = \boldsymbol{\alpha}\,\boldsymbol{\beta}' = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}\,.$$

Since the factorization is not unique, we impose $\beta_1 = 1$ and $\beta_2 = -\beta$, giving:

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ -\beta \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{\beta}'\,\mathbf{z}_t = Y_t - \beta X_t \sim I(0)\,.$$

The normalization suggests the long-run equilibrium:

$$Y_t = \beta X_t + u_t\,,$$

and the VECM form:

$$\begin{cases} \Delta Y_t = \alpha_1\,(Y_{t-1} - \beta X_{t-1}) + \varepsilon_{1t} \\ \Delta X_t = \alpha_2\,(Y_{t-1} - \beta X_{t-1}) + \varepsilon_{2t} \end{cases}$$

Stability requires $u_t \sim I(0)$, which imposes constraints on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since $u_t = \boldsymbol{\beta}'\,\mathbf{z}_t$ and:

$$\boldsymbol{\beta}'\,\mathbf{z}_t = \boldsymbol{\beta}'\,\boldsymbol{\Pi}_1\,\mathbf{z}_{t-1} + \boldsymbol{\beta}'\,\boldsymbol{\varepsilon}_t\,,$$

substituting $\mathbf{\Pi}_1 = \mathbf{I} - \mathbf{\Pi}(1) = \mathbf{I} + \boldsymbol{\alpha}\,\boldsymbol{\beta}'$ yields:

$$\begin{aligned} \boldsymbol{\beta}'\mathbf{z}_t &= \boldsymbol{\beta}'(\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\beta}')\mathbf{z}_{t-1} + \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\beta}'\mathbf{z}_{t-1} + \boldsymbol{\beta}'\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{z}_{t-1} + \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t \quad , \\ &= (1 + \boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\beta}'\mathbf{z}_{t-1} + \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t \end{aligned}$$

or

$$\begin{aligned} & u_t = \delta u_{t-1} + \nu_t \\ & \begin{cases} \delta = (1 + \boldsymbol{\beta}'\boldsymbol{\alpha}) \\ \nu_t = \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t \end{cases} \end{aligned}$$

The stability condition $|\delta| < 1$ implies:

$$|\, 1 + (\alpha_1 - \beta\alpha_2)\,| < 1\,.$$

If $\beta > 0$, this is equivalent to:

$$-2 < \alpha_1 < 0 \quad \text{and} \quad \frac{\alpha_1}{\beta} < \alpha_2 < \frac{\alpha_1 + 2}{\beta}\,.$$

Alternatively, to use a VAR(p), we may start with a Moving Average representation:

$$\Delta\mathbf{z}_t = \mathbf{C}(L)\,\boldsymbol{\varepsilon}_t\,, \tag{8.73}$$

where $\mathbf{C}(L)$ is a polynomial matrix satisfying:

$$\sum_{j=1}^{\infty} j\,\|\mathbf{C}_j\| < \infty\,, \quad \mathbf{C}(0) = \mathbf{I}_n\,,$$

and $\boldsymbol{\varepsilon}_t$ is multivariate white noise.
The matrix $\mathbf{C}(L)$ can be decomposed similarly to the univariate case in (8.30). For $q = 2$:

$$\mathbf{C}(L) = \mathbf{I}_n + \mathbf{C}_1 L + \mathbf{C}_2 L^2\,. \tag{8.74}$$

Adding and subtracting $\mathbf{C}_2 L$:

$$\mathbf{C}(L) = \mathbf{I}_n + (\mathbf{C}_1 + \mathbf{C}_2)L - \mathbf{C}_2 L(1 - L)\,. \tag{8.75}$$

Adding and subtracting $\mathbf{C}_1 + \mathbf{C}_2$:

$$\begin{aligned} \mathbf{C}(L) &= \mathbf{I}_n + (\mathbf{C}_1 + \mathbf{C}_2)L - \mathbf{C}_2 L(1 - L) + (\mathbf{C}_1 + \mathbf{C}_2) - (\mathbf{C}_1 + \mathbf{C}_2) \\ &= \mathbf{I}_n + \mathbf{C}_1 + \mathbf{C}_2 - (\mathbf{C}_1 + \mathbf{C}_2)(1 - L) - \mathbf{C}_2 L(1 - L) \\ &= \mathbf{C}(1) + \mathbf{C}_0^*(1 - L) + \mathbf{C}_1^* L(1 - L) \\ &= \mathbf{C}(1) + \mathbf{C}^*(L)\Delta\,, \end{aligned} \tag{8.76}$$

where:

$$\mathbf{C}^*(L) = \mathbf{C}_0^* + \mathbf{C}_1^* L$$
$$\begin{cases} \mathbf{C}_0^* = -(\mathbf{C}_1 + \mathbf{C}_2) = \mathbf{I}_n - \mathbf{C}(1) \\ \mathbf{C}_1^* = -\mathbf{C}_2 \end{cases}$$

In general[119], for $q \to \infty$:

$$\mathbf{C}(L) = \mathbf{C}(1) + \mathbf{C}^*(L)\,\Delta\,, \tag{8.77}$$

with:

$$\begin{cases} \mathbf{C}_0^* = \mathbf{I}_n - \mathbf{C}(1) \\ \mathbf{C}_i^* = -\sum_{j>i} \mathbf{C}_j\,, \quad i \geq 1 \end{cases} \tag{8.78}$$

Stock and Watson (1988) apply this decomposition assuming $\Delta\mathbf{z}_t$ has a Wold representation:

$$\Delta\mathbf{z}_t = \mathbf{C}(L)\,\boldsymbol{\varepsilon}_t\,, \tag{8.79}$$

so that:

$$\Delta\mathbf{z}_t = \mathbf{C}(1)\,\boldsymbol{\varepsilon}_t + \mathbf{C}^*(L)\,\Delta\boldsymbol{\varepsilon}_t\,. \tag{8.80}$$

Integrating both sides (under the assumption that $\mathbf{z}_t \sim I(1)$) gives the *trend–cycle decomposition*:

$$\mathbf{z}_t = \mathbf{C}(1)\sum_{j=0}^{\infty}\boldsymbol{\varepsilon}_{t-j} + \mathbf{C}^*(L)\,\boldsymbol{\varepsilon}_t = \boldsymbol{\tau}_t + \mathbf{c}_t\,, \tag{8.81}$$

where $\boldsymbol{\tau}_t$ is the stochastic trend and $\mathbf{c}_t$ is the transitory (cyclical) part.
According to Vahid and Engle (1993): *"If* $\mathbf{C}(1)$ *has full rank, the trend is a linear combination of* n *random walks and no linear combination of* $\mathbf{z}$ *is stationary. If* $\text{rank}[\mathbf{C}(1)] = r < n$, $\mathbf{C}(1)$ *can be factored into two matrices of rank* r*, and the trend reduces to* r *random walks rather than* n*."*
In the framework of Stock and Watson (1988), if $\text{rank}[\mathbf{C}(1)] = r < n$, then $\mathbf{z}_t$ has r *common trends* (random walks) and possibly $(n-r)$ *common cycles*.
Further discussion is provided in Vahid and Engle (1993).

8.12 Johansen's Methodology for Modelling Cointegration

In this section, we follow the Johansen procedure to model cointegration and to test the number of cointegrating vectors.
Johansen's procedure can be summarised as follows:

1. Specification and estimation of the unrestricted VAR(p) model (8.65).

[119] See Vahid and Engle (1993).

Recalling expression (8.67), we can rewrite it as:

$$\begin{aligned}
\mathbf{\Pi}^*(L)\Delta\mathbf{z}_t &= -\mathbf{\Pi}(1)L\mathbf{z}_t + \boldsymbol{\varepsilon}_t \\
\Delta\mathbf{z}_t &= \mathbf{\Pi}_1^*\Delta\mathbf{z}_{t-1} + \cdots + \mathbf{\Pi}_{p-1}^*\Delta\mathbf{z}_{t-p+1} - \mathbf{\Pi}(1)\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t \\
(1-L)\mathbf{z}_t &= \mathbf{\Pi}_1^*\Delta\mathbf{z}_{t-1} + \cdots + \mathbf{\Pi}_{p-1}^*\Delta\mathbf{z}_{t-p+1} - \mathbf{\Pi}(1)\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t \\
\mathbf{z}_t &= \mathbf{\Pi}_1^*\Delta\mathbf{z}_{t-1} + \cdots + \mathbf{\Pi}_{p-1}^*\Delta\mathbf{z}_{t-p+1} + \big[\mathbf{I} - \mathbf{\Pi}(1)\big]\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t \,.
\end{aligned} \tag{8.82}$$

The $\mathbf{\Pi}_j$ and $\mathbf{\Pi}_j^*$ matrices are linked by the equalities:

$$\begin{cases}
\mathbf{\Pi}_1 = \mathbf{I} - \mathbf{\Pi}(1) + \mathbf{\Pi}_1^* \\
\mathbf{\Pi}_j = \mathbf{\Pi}_j^* - \mathbf{\Pi}_{j-1}^* \,, \quad j = 2, \ldots, p-1 \\
\mathbf{\Pi}_p = -\mathbf{\Pi}_{p-1}^* \,,
\end{cases} \tag{8.83}$$

which represent the multivariate version of the left panel of *Table 8.4* in the univariate case.
Using OLS, the equalities in (8.83) ensure that the fitted values obtained from (8.67) and (8.82) coincide.
We can now discuss the properties of the estimators.
Following Hamilton[120], $\hat{\mathbf{\Pi}}_j^*$ converge to $\mathbf{\Pi}_j^*$ at rate $T^{1/2}$, and moreover $T^{1/2}(\hat{\mathbf{\Pi}}_j^* - \mathbf{\Pi}_j^*)$ is asymptotically normal. Similarly, $\hat{\mathbf{\Pi}}_j$ for $j = 2, \ldots, p-1$ converge to $\mathbf{\Pi}_j$ at rate $T^{1/2}$, and $T^{1/2}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)$ is asymptotically normal as well, since they are linear combinations of the $\hat{\mathbf{\Pi}}_j^*$ parameters.
A problem arises with the estimation of the $\mathbf{\Pi}(1)$ parameter. In the presence of unit roots, superconsistency ensures that $\hat{\mathbf{\Pi}}(1)$ converges to $\mathbf{\Pi}(1)$ at rate T. Its asymptotic distribution is non-normal, but this faster rate of convergence implies that $\hat{\mathbf{\Pi}}_1 = \mathbf{I} - \hat{\mathbf{\Pi}}(1) + \hat{\mathbf{\Pi}}_1^*$ also converges at the $T^{1/2}$ rate to an asymptotically normal distribution, because the coefficients with the slower convergence rate dominate.
Consequently, the presence of unit roots does not prevent the use of conventional t- and F-tests in VAR(p) levels as long as the restrictions do not involve $\mathbf{I} - \mathbf{\Pi}(1) = \mathbf{\Pi}_1 + \cdots + \mathbf{\Pi}_p$.
Determining the order p of the VAR model is crucial for the next step.

2. Estimation of VECM and tests of the cointegrating rank.

Without going into detail, we refer to the concise explanation of Hamilton[121], which summarises the maximum likelihood (ML) procedure proposed by Johansen[122].
The maintained hypothesis is that the vector $\mathbf{z}_t$ follows a VAR(p) in levels, which admits an error correction representation as in (8.66). More precisely:

[120] See Hamilton (1994), Ch. 18.2.
[121] See Hamilton (1994), Ch. 20.2, pp. 635 ff.
[122] Johansen (1991).

$$\Delta \mathbf{z}_t = \boldsymbol{\mu} + \boldsymbol{\Pi}_1^* \Delta \mathbf{z}_{t-1} + \cdots + \boldsymbol{\Pi}_{p-1}^* \Delta \mathbf{z}_{t-p+1} + \boldsymbol{\Pi}_0 \mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t$$

$$\begin{cases} \boldsymbol{\Pi}_0 = \boldsymbol{\Pi}(1) = -\boldsymbol{\alpha}\boldsymbol{\beta}' \\ \boldsymbol{\varepsilon}_t \sim \text{NID}(\mathbf{0},\, \boldsymbol{\Omega}) \end{cases} \tag{8.84}$$

In the first step, we consider the multivariate regression:

$$\Delta \mathbf{z}_t = \hat{\boldsymbol{\delta}} + \sum_{j=1}^{p-1} \hat{\boldsymbol{\Gamma}}_j \, \Delta \mathbf{z}_{t-j} + \hat{\mathbf{u}}_t \,, \tag{8.85}$$

where $\hat{\boldsymbol{\Gamma}}_j$ denotes an $(n \times n)$ matrix of OLS coefficient estimates and $\hat{\mathbf{u}}_t$ the $(n \times 1)$ vector of OLS residuals.
The second step is to estimate the multivariate regression:

$$\mathbf{z}_{t-1} = \hat{\boldsymbol{\theta}} + \sum_{j=1}^{p-1} \hat{\boldsymbol{\Theta}}_j \, \Delta \mathbf{z}_{t-j} + \hat{\mathbf{v}}_t \,, \tag{8.86}$$

where $\hat{\mathbf{v}}_t$ are the residuals.
From regressions (8.85) and (8.86), we compute the covariance matrices:

$$\hat{\boldsymbol{\Sigma}}_{vv} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{v}}_t \hat{\mathbf{v}}_t' \tag{8.87}$$

$$\hat{\boldsymbol{\Sigma}}_{uu} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_t' \tag{8.88}$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{uv} &= \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \hat{\mathbf{u}}_t \hat{\mathbf{v}}_t' \\ \hat{\boldsymbol{\Sigma}}_{vu} &= \hat{\boldsymbol{\Sigma}}_{uv}' \,. \end{aligned} \tag{8.89}$$

From these covariance matrices, we extract the ordered eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_n$ of:

$$\hat{\boldsymbol{\Sigma}}_{vv}^{-1} \hat{\boldsymbol{\Sigma}}_{vu} \hat{\boldsymbol{\Sigma}}_{uu}^{-1} \hat{\boldsymbol{\Sigma}}_{uv} \,. \tag{8.90}$$

Associated with the r largest eigenvalues are the $(n \times 1)$ eigenvectors $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_r$.
Johansen suggested normalizing each eigenvector so that $\hat{\boldsymbol{\beta}}_i' \hat{\boldsymbol{\Sigma}}_{vv} \hat{\boldsymbol{\beta}}_i = 1, \;\; i = 1, \cdots, r$.
The ML estimate of the $(n \times r)$ matrix $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = \left[\hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \cdots \; \hat{\boldsymbol{\beta}}_r \right], \tag{8.91}$$

and the ML estimate of the $(n \times r)$ matrix $\boldsymbol{\alpha}$ is:

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}_{uv} \hat{\boldsymbol{\beta}} \,. \tag{8.92}$$

Other ML estimates are given by[123]:

$$\begin{aligned} \hat{\mathbf{\Pi}}_0 &= \hat{\mathbf{\Sigma}}_{uv}\,\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \\ \hat{\mathbf{\Pi}}_i^* &= \hat{\mathbf{\Gamma}}_i - \hat{\mathbf{\Pi}}_0\hat{\mathbf{\Theta}}_i \quad . \\ \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\delta}} - \hat{\mathbf{\Pi}}_0\hat{\boldsymbol{\theta}} \end{aligned} \tag{8.93}$$

Hamilton justifies the first step in terms of *auxiliary regressions* involving a *concentrated likelihood function*[124]. The second step is motivated by the concept of *canonical correlation*[125].

The maximum value of the log-likelihood function subject to the constraint of r cointegrating relations is:

$$\begin{aligned} \mathcal{L}(r) = \; & - \frac{Tn}{2}\log(2\pi) - \frac{Tn}{2} \\ & - \frac{T}{2}\log|\hat{\mathbf{\Sigma}}_{uu}| - \frac{T}{2}\textstyle\sum_{i=1}^{r}\log(1-\hat{\lambda}_i)\,. \end{aligned} \tag{8.94}$$

If the rank r is known, ML estimates can be obtained from (8.94). If instead r is unknown, inference is based on the log-likelihood in the unrestricted case:

$$\begin{aligned} \mathcal{L}(n) = \; & - \frac{Tn}{2}\log(2\pi) - \frac{Tn}{2} \\ & - \tfrac{T}{2}\log|\hat{\mathbf{\Sigma}}_{uu}| - \frac{T}{2}\textstyle\sum_{i=1}^{n}\log(1-\hat{\lambda}_i)\,. \end{aligned} \tag{8.95}$$

Both (8.94) and (8.95) provide the essential ingredients for testing whether the rank is $r < n$ through a likelihood ratio test, namely H_0: rank $= r$ against H_A: rank $= n$, which is carried out by:

$$\mathcal{L}(n) - \mathcal{L}(r) = -\frac{T}{2}\sum_{i=r+1}^{n}\log(1-\hat{\lambda}_i)\,. \tag{8.96}$$

If the hypothesis involved only $I(0)$ variables, we would expect twice the log-likelihood ratio, that is

$$2\left[\mathcal{L}(n) - \mathcal{L}(r)\right] = -T\sum_{i=r+1}^{n}\log(1-\hat{\lambda}_i)\,. \tag{8.97}$$

This is known as the *trace statistic*, and testing proceeds sequentially to check whether the rank r is equal to zero, equal to one, up to $r = n-1$. The sequential procedure is:

---

[123] By definition, $\mathbf{\Pi}_0 = -\boldsymbol{\alpha}\boldsymbol{\beta}'$. Since the ML estimator of the adjustment coefficients is $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{\Sigma}}_{uv}\,\hat{\boldsymbol{\beta}}$, it follows that $\hat{\mathbf{\Pi}}_0 = \hat{\mathbf{\Sigma}}_{uv}\,\hat{\boldsymbol{\beta}}\,\hat{\boldsymbol{\beta}}'$. The positive sign is therefore consistent: the estimator incorporates the direction of adjustment through $\hat{\boldsymbol{\alpha}}$, while the theoretical restriction introduces the negative sign in the definition of $\mathbf{\Pi}_0$.

[124] See Gourieroux and Monfort (1995), §7.2.4, p. 170.

[125] See Mardia, Kent, and Bibby (1979), Ch. 10, p. 281 ff..

1. $H_0 : r = 0$ vs. $H_1 : r > 0$
   If H_0 is not rejected, then $r = 0$ (no cointegration). If rejected, continue to $r = 1$.

2. $H_0 : r \leq 1$ vs. $H_1 : r > 1$
   If H_0 is not rejected, then $r = 1$ (only one cointegration relation). If rejected, continue sequentially up to $r = n - 1$.

Formally, the nested hypotheses are:

$$H(0) \subset \cdots \subset H(r) \subset \cdots \subset H(n)\,, \tag{8.98}$$

where:

- $r = 0$: no $I(0)$ linear combinations of $\mathbf{z}_t$ exist, so the model must be built only on $\Delta\mathbf{z}_t$;
- $r = n$: all variables in $\mathbf{z}_t$ are stationary $I(0)$;
- $0 < r < n$: up to r cointegration relations $\boldsymbol{\beta}'\mathbf{z}_t$ exist.

Johansen[126] also proposed the *maximal eigenvalue test* for testing:

$$H_0 : r \quad \text{vs.} \quad H_1 : r + 1\,, \tag{8.99}$$

given by:

$$LR_{\max}(r) = -T \log(1 - \hat{\lambda}_{r+1})\,. \tag{8.100}$$

It is well known that both the trace test and the maximal eigenvalue test do not follow standard χ^2 asymptotic distributions under H_0, but instead non-standard ones (see Johansen (1995)).
The econometric literature has discussed their finite-sample power. In particular, the asymptotic distributions of these tests depend on whether a deterministic trend is included in the data-generating process.
Since researchers are often uncertain about the presence of a linear trend, Harvey et al.[127] suggest that applying both versions of the test (with and without a deterministic trend) is a sound strategy.

[126] See Johansen (1995).
[127] D. I. Harvey, Leybourne, and A. M. R. Taylor (2009).

Appendix 8.A *(Wiener Process (Brownian Motion))*

The Wiener process[128], also known as Brownian motion[129], is a continuous-time stochastic process defined on the interval $[0, 1]$ (and more generally on $[0, \infty)$). It can be viewed as the limit of a discrete-time random walk[130].
In the literature, several constructions of the Wiener process have been proposed[131]. Here we follow the result due to Donsker[132]. His theorem, known as the *invariance principle* (or *functional central limit theorem*), shows that a discrete-time random walk defined on $[0, T]$ converges in distribution to a Wiener process when properly rescaled to the continuous interval.
Let X_T be a random walk defined as:

$$X_T = \sum_{j=1}^{T} \varepsilon_j \,, \tag{8.A1}$$

where $\{\varepsilon_j,\, j = 1, \ldots, T\}$ is an i.i.d. process with $\Pr(\varepsilon_j = \pm 1) = \frac{1}{2}$.
A direct calculation gives $E[\varepsilon_1] = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$, and $Var(\varepsilon_1) = E(\varepsilon_1^2) = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1$.
Hence, $E(X_T) = 0$ and

$$Var(X_T) = Var\left(\sum_{j=1}^{T} \varepsilon_j\right) = \sum_{j=1}^{T} Var(\varepsilon_j) = T,$$

since the covariances are zero.
By the Central Limit Theorem:

$$\frac{X_T}{\sqrt{T}} \xrightarrow[T\to\infty]{\mathcal{D}} \mathcal{N}(0, 1) \,, \tag{8.A2}$$

that is, as $T \to \infty$, the normalized partial sum converges in distribution to the standard normal law.
The discrete interval $[0, T]$ can be mapped onto the fixed interval $[0, 1]$. For instance, if $T = 1000$, we define the discrete variable $r = t/T \in [0, 1]$, $t = 0, 1, \ldots, 1000$. This divides the unit interval into $T + 1$ parts: $0, 1/T, 2/T, \ldots, 1$. Hence, r takes 1001 values in $[0, 1]$ and represents the time points at which the rescaled process $X_T/\sqrt{T}$ is evaluated on the real line.

[128] The name refers to the American mathematician Norbert Wiener (1894–1964).
[129] From the Scottish botanist Robert Brown (1773–1858).
[130] A good introductory book on the Wiener process is Mikosch (1998).
[131] See, for example, Schilling and Partzsch (2012).
[132] From the American mathematician Monroe David Donsker (1924–1991).

Defining $\lfloor rT \rfloor$ as the integer part of rT, the process

$$R_T(r) = \frac{X_{\lfloor rT \rfloor}}{\sqrt{T}}$$

is therefore a *step function*, constant between successive jump points.
To make this construction more explicit, consider the following numerical example.
Suppose $T = 200$, while the unit interval $[0, 1]$ is represented on the same fixed fine grid of 1001 equally spaced points introduced above. Since the process $X_T/\sqrt{T}$ has only 201 distinct values, it must be extended to this finer grid by keeping each value constant until the next jump.
For example, if $r = 0.009$, then $rT = 1.8$ and $\lfloor rT \rfloor = 1$, while at $r = 0.01$ we have $rT = 2$ and $\lfloor rT \rfloor = 2$. Consequently,

$$R_{200}(0.009) = \frac{X_1}{\sqrt{200}}, \quad R_{200}(0.01) = \frac{X_2}{\sqrt{200}},$$

with $X_1 = \varepsilon_1$ and $X_2 = \varepsilon_1 + \varepsilon_2$. The values of X_1 and X_2 depend on the underlying ± 1 sequence.
Thus, for all $r \in [0.005, 0.009]$, the process $R_{200}(r)$ remains constant at $X_1/\sqrt{200}$, then jumps at $r = 0.01$ to $X_2/\sqrt{200}$, and remains constant until $r = 0.014$. A new jump occurs at $r = 0.015$, where $\lfloor rT \rfloor = 3$.
Figure 8.A1 shows realizations of the process $R_T(r)$ for $r \in [0, 1]$ when the unit interval is divided into 1001 parts, for $T = 50$, 200, and 1000.

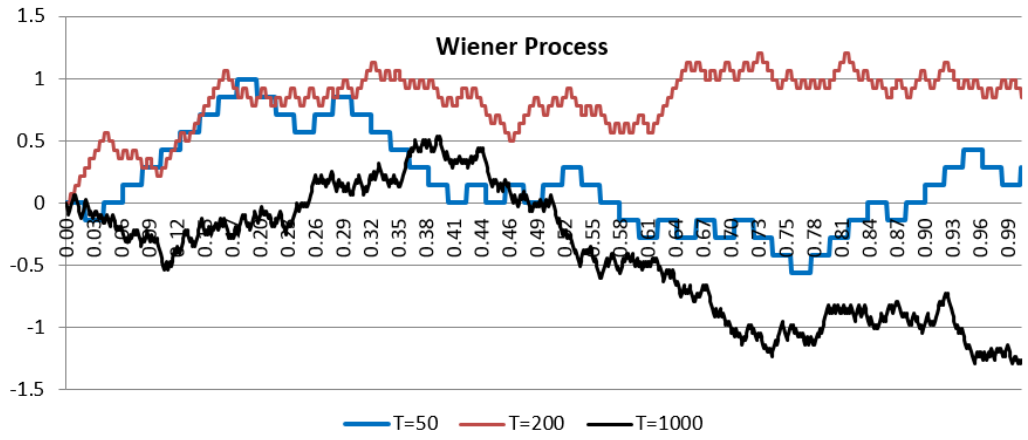


Figure 8.A1: Realizations of the step process $R_T(r)$ on $[0, 1]$ for different values of T

As T increases, the steps become narrower, and it is natural to ask what happens in the limit: *does the discontinuous step function converge to a continuous one?* This is the motivation for the following theorem:

Theorem 8.A1. *(Donsker) Suppose ε_t is an i.i.d. sequence with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = 1$, and define the partial sums $X_t = \sum_{j=1}^{t} \varepsilon_j$ and the rescaled process $R_T(r) = T^{-1/2} X_{\lfloor rT \rfloor}$ for $r \in [0, 1]$. Then, as $T \to \infty$,*

$$R_T(\cdot) \xrightarrow{d} W(\cdot),$$

where W is a Wiener process[133].

Proof. See Billingsley (1999), p. 91. □

Properties of the Wiener process

Intuitively, the Wiener process can be regarded as a continuous-time version of the scaled random walk on [0, 1]. Although continuous, its sample paths fluctuate erratically over any subinterval.

From its construction as the limit of a random walk, the Wiener process has the following important properties. Before listing them, we recall two definitions:

Definition 8.A1. *(Stochastic process with independent increments)*
Given an index set $T \subset \mathbb{R}$, a process $Z = \{Z_t,\, t \in T\}$ has independent increments if, for any $t_1 < t_2 < \cdots < t_n$, the random variables

$$Z_{t_2} - Z_{t_1}, \;\; Z_{t_3} - Z_{t_2}, \;\; \ldots, \;\; Z_{t_n} - Z_{t_{n-1}}$$

are mutually independent.

Definition 8.A2. ***(Stochastic process with stationary increments)***
A process $Z = \{Z_t,\, t \in T\}$ has stationary increments if

$$Z_t - Z_s \;\; \stackrel{\mathcal{D}}{=} \;\; Z_{t+h} - Z_{s+h},$$

for all $s, t \in T$ and shifts h such that $t + h$, $s + h \in T$. The symbol $\stackrel{\mathcal{D}}{=}$ means equality in distribution.

Note that a process with stationary increments is not necessarily stationary.

Following Mikosch, the convergence in distribution of $R_T(t)$ to W_t has two complementary aspects:

1) The finite-dimensional distributions of $R_T(t)$ converge to those of W_t. That is,

$$\Pr(R_T(t_1) \le x_1, \ldots, R_T(t_n) \le x_n) \;\to\; \Pr(W_{t_1} \le x_1, \ldots, W_{t_n} \le x_n)\,,$$

for all $t_j \in [0,\, 1]$, $x_j \in \mathbb{R}$, $j = 1, \ldots, n$ and any $n \ge 1$.

[133] The notation $R_T(\cdot) \xrightarrow{d} W(\cdot)$ indicates convergence in distribution of stochastic processes, not merely convergence at each fixed $r \in [0, 1]$.

> *"But the convergence of the finite-dimensional distributions is not sufficient for the convergence in distribution of stochastic processes. Finite-dimensional distribution convergence determines the Gaussian limit distribution for every choice of finitely many fixed instants of time t_j, but stochastic processes are infinite-dimensional objects, and therefore unexpected events may happen. For example, the sample paths of the converging $R_T(t)$ processes may fluctuate very wildly with increasing T, in contrast to the limiting process of Brownian motion which has continuous sample path"*[134].

To prevent such irregular behavior,

2) a *tightness* (or *stochastic compactness*) condition must also be satisfied.

It can be shown that the partial sum processes $R_T(t)$ are indeed tight[135].

In summary, the properties of the Wiener process are:

a) *Independence*

$W_t - W_s$ is independent of the past $\{W_\tau : \tau \leq s\}$[136] and has the same distribution as W_{t-s}.

b) *Independent and stationary increments*

For $0 \leq s < t$, the increment $W_t - W_s$ is independent of the past, that is, of all values $\{W_\tau : \tau \leq s\}$.

c) *Gaussianity*

W_t is a Gaussian process with mean $E(W_t) = 0$ and covariance $E(W_t W_s) = \min(t, s)$.

d) *Continuity*

W_t is a continuous function of t with probability 1 (almost sure continuity). In other words, the points of discontinuity have probability zero.

Some comments on these properties are in order.
Independence and independent increments follow from the fact that

$$X_{t+h} - X_t = \sum_{j=t+1}^{t+h} \varepsilon_j, \qquad 0 \leq t \leq t+h,$$

[134] Mikosch (1998), p. 47.
[135] For a clear explanation of tightness and stochastic compactness, see Billingsley (1999), §7, p. 80 ff.
[136] Formally, independence is understood with respect to the σ-algebra generated by $\{W_\tau : \tau \leq s\}$, which represents the information available up to time s.

is independent of X_t and is distributed as $X_{t+h} - X_t = X_h$, since $\{\varepsilon_j\}$ is an i.i.d. process. Consequently, taking $W_0 = 0$ and using Gaussianity (explained below), it can be shown that

$$W_{t+h} - W_t \overset{\mathcal{D}}{=} W_h \,, \tag{8.A3}$$

that is, the increment follows $N(0, h)$. In this way, the variance of the increment is proportional to the length of the interval $[t,\, t+h]$: the larger the interval, the larger the fluctuations of the Wiener process. It is important to note, however, that (8.A3) does not imply that the trajectory generated by $W_{t+h} - W_t$ is identical to that generated by W_h. Gaussianity follows from the fact that if $t \in [0, 1]$ is fixed, then

$$R_T(t) = \frac{X_{\lfloor tT \rfloor}}{\sqrt{T}} = \frac{X_{\lfloor tT \rfloor}}{\sqrt{\lfloor tT \rfloor}} \cdot \frac{\sqrt{\lfloor tT \rfloor}}{\sqrt{T}} \xrightarrow[T\to\infty]{\mathcal{D}} Z\sqrt{t} \overset{\mathcal{D}}{=} N(0, t),$$

where $Z \sim N(0, 1)$.
Since $R_T(t) \xrightarrow{d} W_t$ and tightness (stochastic compactness condition) holds, we also have:

$$E(W_t) = 0, \qquad Var(W_t) = t.$$

The covariance is obtained as follows. Suppose $0 \le s \le t$, then

$$\begin{aligned} Cov(W_t,\, W_s) &= E(W_t W_s) = E\left[(W_s + (W_t - W_s))W_s\right] \\ &= E\left[W_s^2 + (W_t - W_s)W_s\right] \\ &= s + E\big[(W_t - W_s)W_s\big] \\ &= s \,, \end{aligned}$$

since W_s and the increment $(W_t - W_s)$ are independent.
If $0 \le t \le s$, then clearly $Cov(W_t, W_s) = t$. In general, therefore,

$$Cov(W_t, W_s) = \min(t, s).$$

In summary, a stochastic process with these properties can be considered a Wiener process. This motivates the following definition:

Definition 8.A3. ***(Wiener Process or Brownian Motion)***
A stochastic process $\{W_t,\, t \in [0, \infty)\}$ *is called a* Wiener process *or* Brownian motion *if the following properties hold:*

1. *Its starting point is zero:* $W_0 = 0$.
2. *It has independent and stationary increments.*

3. *For each* $t > 0$, $W_t \sim N(0, t)$.

4. *It has continuous realizations (sample paths), i.e., the probability of jumps is zero.*

Since a Gaussian stochastic process is fully characterized by its mean and covariance function, an alternative definition of the Wiener process is:

Definition 8.A4. *The Wiener process (or Brownian motion) is a Gaussian stochastic process with*

$$E(W_t) = 0, \qquad Cov(W_t, W_s) = \min(t, s).$$

Further properties of the Wiener process, such as the non-differentiability of its trajectories and their infinite variation, can be found in the introductory text by Mikosch (1998). These striking properties highlight the irregularity of Brownian paths despite their continuity.
The Wiener process is also linked to concepts that are central in stochastic calculus for finance, such as *martingales* (of which the Wiener process is an example) and the *Itô formula*. The discussion of these topics, however, lies beyond the scope of this Appendix.

9 Dynamic Systems in State-Space Form

9.1 Introduction

As discussed in Chapter 5, the dynamic properties of stochastic models allow them to be regarded as systems. Univariate processes belonging to the ARMA(p, q) class can also be interpreted in this way. Consider, for example, the simplest case of an *AR(1)* model:

$$X_t = \alpha X_{t-1} + \varepsilon_t \,. \tag{9.1}$$

Interpreting ε_t as the system input (and α as known), the process X_t represents the system output determined by that input. This is the interpretation assigned to (9.1) when simulating a realisation of X_t based on random draws of ε_t. Such a generation mechanism has a different meaning from the usual econometric interpretation. In econometrics, the realisation of the stochastic process X_t is assumed to be observed, and inference is carried out to determine the characteristics of the system that generates it: the class of models to which it belongs and, within that class, the specific parameter values. Moreover, since the input component ε_t is unobservable, inference is based solely on the system output, under certain assumptions on the process generating the input.
For the determination of the values of X_t in the input–output interpretation of (9.1), an arbitrary initial value X_0 is assigned and the system is solved by iterated substitutions:

$$\begin{aligned} X_1 &= \alpha X_0 + \varepsilon_1 \\ X_2 &= \alpha X_1 + \varepsilon_2 = \alpha(\alpha X_0 + \varepsilon_1) + \varepsilon_2 = \alpha^2 X_0 + \alpha\varepsilon_1 + \varepsilon_2 \\ &\vdots \\ X_t &= \alpha^t X_0 + \sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j} \,. \end{aligned} \tag{9.2}$$

From a mathematical point of view, (9.2) is the solution of the *finite stochastic difference equation* (9.1). The first term, $\alpha^t X_0$, is the *general solution of the homogeneous part*, while the second term, $\sum_{j=0}^{t-1} \alpha^j \varepsilon_{t-j}$, is the *particular solution of the inhomogeneous part*. If the input ε_t is assigned, the solution of the finite difference equation identifies a family of output processes, each depending on the choice of the initial value X_0. If the condition $|\alpha| < 1$ holds, then $\lim_{j\to\infty} \alpha^j = 0$ and the process loses memory of its initial value. Under this condition, the effect of the initial value vanishes asymptotically and the solution is stable; in the AR(1) case this coincides with covariance stationarity.

Consider now the multivariate case:

$$\underset{(k\times 1)}{\mathbf{x}_t} = \underset{(k\times k)}{\mathbf{A}}\, \mathbf{x}_{t-1} + \underset{(k\times 1)}{\boldsymbol{\varepsilon}_t} \,. \tag{9.3}$$

where $\boldsymbol{\varepsilon}_t$ is a vector of state disturbances (or system shocks), assumed to be a zero-mean white noise process.

Given an initial vector $\mathbf{x}_0$ of arbitrary values, the solution of the system is:

$$\mathbf{x}_t = \mathbf{A}^t \mathbf{x}_0 + \sum_{j=0}^{t-1} \mathbf{A}^j \boldsymbol{\varepsilon}_{t-j} \,. \tag{9.4}$$

The stability of the solution is less immediate in this case. Assume that $\mathbf{A}$ is *diagonalizable*, which holds if all its eigenvalues $\{\lambda_j;\, j = 1, \ldots, k\}$ are distinct[137]. If $\mathbf{A}$ is diagonalizable, it can be written as:

$$\mathbf{AC} = \mathbf{C}\boldsymbol{\Lambda} \,, \tag{9.5}$$

where:

$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is the $k \times k$ diagonal matrix containing the eigenvalues;

$\mathbf{C}$ is the matrix of eigenvectors of $\mathbf{A}$.

Since the eigenvectors are linearly independent, $\mathbf{C}$ is invertible. From (9.5):

$$\boldsymbol{\Lambda} = \mathbf{C}^{-1}\mathbf{AC} \,, \tag{9.6}$$

or, equivalently,

$$\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^{-1} \,. \tag{9.7}$$

Relations (9.6) and (9.7) are referred to as *similarity transformations*. An important property is:

$$\mathbf{A}^2 = (\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^{-1})^2 = \mathbf{C}\boldsymbol{\Lambda}^2\mathbf{C}^{-1} \,. \tag{9.8}$$

By iteration:

$$\mathbf{A}^t = \mathbf{C}\boldsymbol{\Lambda}^t\mathbf{C}^{-1} \,. \tag{9.9}$$

Substituting (9.9) into (9.4) yields:

$$\mathbf{x}_t = \mathbf{C}\boldsymbol{\Lambda}^t\mathbf{C}^{-1}\mathbf{x}_0 + \sum_{j=0}^{t-1} \mathbf{C}\boldsymbol{\Lambda}^j\mathbf{C}^{-1}\boldsymbol{\varepsilon}_{t-j} \,. \tag{9.10}$$

The process $\mathbf{x}_t$ loses memory of its initial value $\mathbf{x}_0$ (asymptotic stability) if and only if all eigenvalues satisfy[138] $|\lambda_j| < 1$, $j = 1, \ldots, k$.

[137] Distinct eigenvalues are only a sufficient condition. A matrix may be diagonalizable even with repeated eigenvalues, provided it has a full set of k linearly independent eigenvectors; equivalently, the geometric multiplicity of each eigenvalue must equal its algebraic multiplicity. In particular, when $\mathrm{rank}(\mathbf{A} - \lambda_j\mathbf{I}) = k - 1$ the eigenvalue λ_j has geometric multiplicity one.

[138] Equivalently, the spectral radius $\rho(\mathbf{A}) = \max_j |\lambda_j|$ must satisfy $\rho(\mathbf{A}) < 1$.

Consider now the case of an *AR(2)* process:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t \,. \tag{9.11}$$

The interpretation of (9.11) as the solution of a finite-difference equation is less straightforward than in the *AR(1)* case. We can proceed by iterative substitutions, noting that it is necessary to start from two initial conditions, denoted X_0 and X_{-1}. Thus, we obtain:

$$\begin{aligned} X_1 &= \alpha_1 X_0 + \alpha_2 X_{-1} + \varepsilon_1 \\ X_2 &= \alpha_1 X_1 + \alpha_2 X_0 + \varepsilon_2 = (\alpha_1^2 + \alpha_2) X_0 + \alpha_1 \alpha_2 X_{-1} + \alpha_1 \varepsilon_1 + \varepsilon_2 \\ X_3 &= \alpha_1 X_2 + \alpha_2 X_1 + \varepsilon_3 = (\alpha_1^3 + 2\alpha_1 \alpha_2) X_0 + (\alpha_2^2 + \alpha_1^2 \alpha_2) X_{-1} \\ &\quad + (\alpha_1^2 + \alpha_2)\varepsilon_1 + \alpha_1 \varepsilon_2 + \varepsilon_3 \\ &\vdots \end{aligned} \tag{9.12}$$

The continuation of the substitutions is omitted, but it is evident that this procedure generates an increasing number of additive terms involving the coefficients of the initial conditions and the input shocks, as well as their powers. This makes the final solution far more complex than in the AR(1) case.
The complexity of the final solution becomes unmanageable in the case of AR(p) processes with high order p. For this reason, it is preferable to adopt an alternative formalization of AR(p) models that reduces them to the case of single-lag models. This formalization is obtained by redefining the output variables of the system and is motivated by the multivariate formulation in (9.3)–(9.10), which provides the general stability criterion for first-order vector systems.
For simplicity, consider again the *AR(2)* process. We define two new variables:

$$\begin{cases} z_{1t} = X_t \\ z_{2t} = X_{t-1} \,, \end{cases} \tag{9.13}$$

so that $z_{2,\,t-1} = X_{t-2}$.
With these substitutions, the difference equation (9.11) can be rewritten as:

$$z_{1t} = \alpha_1 z_{1,\,t-1} + \alpha_2 z_{2,\,t-1} + \varepsilon_t \,. \tag{9.14}$$

The expression (9.14) shows that we have returned to a relationship with only a one-period lag.
The price to be paid for this simplification is the introduction of a new variable z_{2t}; however, since

$$z_{2t} = z_{1,\,t-1} \,, \tag{9.15}$$

both (9.14) and (9.15) can be considered jointly in the following system:

$$\begin{cases} z_{1t} = \alpha_1 z_{1,t-1} + \alpha_2 z_{2,t-1} + \varepsilon_t \\ z_{2t} = z_{1,t-1}\,, \end{cases} \tag{9.16}$$

or, in matrix form:

$$\begin{bmatrix} z_{1t} \\ \\ z_{2t} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \varepsilon_t\,. \tag{9.17}$$

Defining

$$\mathbf{z}_t = \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

we can rewrite the system (9.17) as

$$\mathbf{z}_t = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{G}\varepsilon_t\,. \tag{9.18}$$

Equation (9.18) is interpreted as a first-order stochastic difference equation.
The last important step in this formal transformation of the second-order stochastic equation (9.11) into the first-order form (9.18) concerns making explicit the link between the scalar process X_t and the new vector process $\mathbf{z}_t$. This link is given by:

$$X_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \mathbf{H}\mathbf{z}_t\,. \tag{9.19}$$

In conclusion, the dynamic system (9.11) can be equivalently represented by the following system of equations:

$$\begin{cases} X_t &= \mathbf{H}\mathbf{z}_t \\ \mathbf{z}_t &= \mathbf{F}\mathbf{z}_{t-1} + \mathbf{G}\varepsilon_t\,, \end{cases} \tag{9.20}$$

which is known as the *state-space representation* of the system. Here $\mathbf{z}_t$ is called the *state of the system*, and its components are referred to as *state variables*.
The first equation in (9.20) is generally called the *measurement (or output) equation*, while the second one is the *transition (or state) equation*. The matrix $\mathbf{F}$ is called the *transition matrix*.

Remark 9.1. The state–space representation of a stochastic dynamic system is not unique. Different formulations may arise from alternative definitions of the state vector, all leading to dynamically equivalent systems.

In particular, the formulation adopted by Harvey[139] corresponds to a different choice of state variables. While in (9.18) the state vector is defined as

$$\mathbf{z}_t = \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix},$$

Harvey defines the state vector by interchanging the roles of the lagged components, which leads to the representation

$$\begin{cases} X_t &= \mathbf{H}\mathbf{z}_t \\ \mathbf{z}_t &= \mathbf{F}'\mathbf{z}_{t-1} + \mathbf{G}\varepsilon_t \,, \end{cases} \tag{9.21}$$

where the transition matrix $\mathbf{F}'$ is numerically equal to the transpose of the matrix $\mathbf{F}$ in (9.18).
In explicit form, this yields

$$\begin{cases} X_t &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_t \\ \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} &= \begin{bmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{bmatrix} \begin{bmatrix} z_{1,\,t-1} \\ z_{2,\,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \varepsilon_t \,. \end{cases} \tag{9.22}$$

The appearance of the transposed transition matrix should not be interpreted as a transposition of the system dynamics. It simply reflects a different linear transformation of the state vector. Both formulations share the same characteristic polynomial, eigenvalues, and stability properties, and therefore describe the same underlying stochastic process.

Remark 9.2. The iterative substitutions highlighted in (9.12) can be performed more easily in the state-space form. Denoting the initial state as

$$\mathbf{z}_0 = \begin{bmatrix} z_{10} \\ z_{20} \end{bmatrix},$$

and using (9.4), we have

$$\mathbf{z}_t = \mathbf{F}^t \mathbf{z}_0 + \sum_{j=0}^{t-1} \mathbf{F}^j \mathbf{G}\, \varepsilon_{t-j} \,. \tag{9.23}$$

At step $t = 3$ this becomes

$$\mathbf{z}_3 = \mathbf{F}^3 \mathbf{z}_0 + \sum_{j=0}^{2} \mathbf{F}^j \mathbf{G}\, \varepsilon_{t-j},$$

that is,

$$\mathbf{z}_3 = \mathbf{F}^3 \mathbf{z}_0 + \mathbf{F}^2 \mathbf{G}\, \varepsilon_{t-2} + \mathbf{F}\mathbf{G}\, \varepsilon_{t-1} + \mathbf{G}\, \varepsilon_t,$$

[139] A. C. Harvey (1993), p. 84.

with:

$$\mathbf{F}^2 = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \alpha_1^2 + \alpha_2 & \alpha_1\alpha_2 \\ \alpha_1 & \alpha_2 \end{bmatrix},$$

$$\mathbf{F}^3 = \begin{bmatrix} \alpha_1^2 + \alpha_2 & \alpha_1\alpha_2 \\ \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \alpha_1^3 + 2\alpha_1\alpha_2 & \alpha_1^2\alpha_2 + \alpha_2^2 \\ \alpha_1^2 + \alpha_2 & \alpha_1\alpha_2 \end{bmatrix},$$

$$\mathbf{FG} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ 1 \end{bmatrix},$$

$$\mathbf{F}^2\mathbf{G} = \begin{bmatrix} \alpha_1^2 + \alpha_2 & \alpha_1\alpha_2 \\ \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha_1^2 + \alpha_2 \\ \alpha_1 \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} z_{13} \\ z_{23} \end{bmatrix} = \begin{bmatrix} \alpha_1^3 + 2\alpha_1\alpha_2 & \alpha_1^2\alpha_2 + \alpha_2^2 \\ \alpha_1^2 + \alpha_2 & \alpha_1\alpha_2 \end{bmatrix} \begin{bmatrix} z_{10} \\ z_{20} \end{bmatrix} + \begin{bmatrix} \alpha_1^2 + \alpha_2 \\ \alpha_1 \end{bmatrix} \varepsilon_1 + \begin{bmatrix} \alpha_1 \\ 1 \end{bmatrix} \varepsilon_2 + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \varepsilon_3.$$

Using the first equation of (9.22) and renaming the initial conditions $z_{10} = X_0$ and $z_{20} = X_{-1}$, we obtain:

$$\begin{aligned} X_3 &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_3 \\ &= z_{13} \\ &= (\alpha_1^3 + 2\alpha_1\alpha_2)X_0 + (\alpha_1^2\alpha_2 + \alpha_2^2)X_{-1} + (\alpha_1^2 + \alpha_2)\varepsilon_1 + \alpha_1\varepsilon_2 + \varepsilon_3, \end{aligned}$$

which coincides with (9.12).

Remark 9.3. The dimension of the state coincides with the number of initial conditions required for the uniqueness of the solution of the finite difference equation. The initial state of the dynamic system coincides exactly with the set of initial conditions.

Remark 9.4. The system dynamics summarised in the state dynamics represent a first-order autoregressive (or Markovian) vector process.

9.2 Formalization of the State-Space System

From a formal point of view, the state-space system is based on two definitions.

Definition 9.1. *(State Variables)*
The state variables of a system consist of a minimum *set of parameters that completely summarise the system's status in the following sense.*
If at any initial time $t_0 \in \mathcal{T}$ the values of the state variables $\{z_{1t_0}, \ldots, z_{kt_0}\}$ are known, then the output X_{t_1}, where t_1 denotes the final time, and the values $\{z_{1t_1}, \ldots, z_{kt_1}\}$ can be uniquely *determined for any time $t_1 \in \mathcal{T}$, $t_1 > t_0$, provided that $\{\varepsilon_{t_0}, \ldots, \varepsilon_{t_1}\}$ is known.*

Definition 9.2. *(State-Space System)*
The state-space system at any time $t_0 \in \mathcal{T}$ is a set of the minimum number of parameters that allows a unique output segment $\{X_{t_0}, \ldots, X_t\}$ to be associated with each input segment $\{\varepsilon_{t_0}, \ldots, \varepsilon_t\}$ for every $t_0 \in \mathcal{T}$ and for all $t > t_0$, $t \in \mathcal{T}$.

The two definitions can be interpreted in terms of spaces to which the input, output, and state belong, indicated respectively with $\mathcal{E}$, $\mathcal{X}$, $\mathcal{Z}$.
At each t, let $\varepsilon_t \in \mathcal{E}$, $X_t \in \mathcal{X}$, and the vector state

$$\mathbf{z}_t = \begin{bmatrix} z_{1t}, & \cdots, & z_{kt} \end{bmatrix}' \in \mathcal{Z},$$

then given:

1) $(t_0, t_1) \in \mathcal{T}$,

2) $\{\varepsilon_{t_0}, \ldots, \varepsilon_{t_1}\} = \{\varepsilon_t\}_{t_0}^{t_1} \in \mathcal{E}$,

3) $\mathbf{z}_{t_0} \in \mathcal{Z}$,

we consider the transformation $\mathbf{f}$ that maps the elements

$$(t_0, t_1, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_1})$$

from the Cartesian product $\mathcal{T} \times \mathcal{T} \times \mathcal{Z} \times \mathcal{E}$ into a unique element in $\mathcal{Z}$, that is:

$$\mathbf{z}_{t_1} = f\left(t_0, t_1, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_1}\right). \tag{9.24}$$

In addition, a second transformation $\mathbf{h}$ must be given, mapping the elements

$$(t_1, \mathbf{z}_{t_1}, \varepsilon_{t_1})$$

from the Cartesian product $\mathcal{T} \times \mathcal{Z} \times \mathcal{E}$ into a unique element in $\mathcal{X}$, that is:

$$X_{t_1} = h\left(t_1, \mathbf{z}_{t_1}, \varepsilon_{t_1}\right). \tag{9.25}$$

Therefore, a causal dynamic system can be represented by the triad of $\mathcal{T} \times \mathcal{Z} \times \mathcal{E}$ spaces with associated $\mathbf{f}$ and $\mathbf{h}$ functions that satisfy the following properties:

1) $\mathbf{h}$ represents an *instantaneous, or memoryless, transformation.*

2) $\mathbf{f}$ represents a *nonanticipative* (for this reason also called *causal*) transformation such that:
$$f\left(t_0, t_0, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_0}\right) = \mathbf{z}_{t_0}, \quad \forall t_0 \in \mathcal{T}. \tag{9.26}$$

Property (9.26) indicates that $\mathbf{f}$ becomes the identity transformation when its time arguments coincide. This property has full meaning if $\mathcal{T}$ is continuous, while it is trivially satisfied if it is discrete.
Moreover, if two input trajectories $\{\varepsilon_t\}_{t_0}^{t_1}$ and $\{\eta_t\}_{t_0}^{t_1}$ in $\mathcal{E}$ coincide for all $t \in [t_0, t_1]$, then:

$$f\left(t_0, t_1, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_1}\right) = f\left(t_0, t_1, \mathbf{z}_{t_0}, \{\eta_t\}_{t_0}^{t_1}\right). \tag{9.27}$$

This property is known as the *state transition property* and indicates that $\mathbf{z}_{t_1}$ does not depend on inputs prior to t_0 except for the past already summarised in $\mathbf{z}_{t_0}$, and does not depend on inputs after t_1. In other words, the state at time t_1 depends only on the initial state $\mathbf{z}_{t_0}$ and on the input trajectory restricted to the interval $[t_0, t_1]$, thus expressing the causal nature of the transformation.
If $(t_0, t_1, t_2) \in \mathcal{T}$ and $t_0 < t_1 < t_2$, then:

$$\begin{aligned} \mathbf{z}_{t_2} &= f\left(t_0, t_2, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_2}\right) \\ &= f\left(t_1, t_2, \mathbf{z}_{t_1}, \{\varepsilon_t\}_{t_1}^{t_2}\right) \\ &= f\Big(t_1, t_2, f\left(t_0, t_1, \mathbf{z}_{t_0}, \{\varepsilon_t\}_{t_0}^{t_1}\right), \{\varepsilon_t\}_{t_1}^{t_2}\Big). \end{aligned} \tag{9.28}$$

This is the *semigroup property*, which states that it is indifferent whether $\mathbf{z}_{t_2}$ is obtained directly from $\mathbf{z}_{t_0}$ and $\{\varepsilon_t\}_{t_0}^{t_2}$, or indirectly by first computing $\mathbf{z}_{t_1}$ from $\mathbf{z}_{t_0}$ and $\{\varepsilon_t\}_{t_0}^{t_1}$, and then using it with $\{\varepsilon_t\}_{t_1}^{t_2}$.

In the case of linear systems, $\mathbf{f}$ and $\mathbf{h}$ are linear transformations. In the general case where there is a vector $\mathbf{x}_t$ of stochastic processes of order N, the model with state variables takes the form:

$$\begin{cases} \underset{(N\times 1)}{\mathbf{x}_t} = \underset{(N\times m)}{\mathbf{H}_t}\ \mathbf{z}_t + \underset{(N\times n)}{\mathbf{S}_t}\ \boldsymbol{\varepsilon}_t \\ \underset{(m\times 1)}{\mathbf{z}_t} = \underset{(m\times m)}{\mathbf{F}_t}\ \mathbf{z}_{t-1} + \underset{(m\times g)}{\mathbf{G}_t}\ \boldsymbol{\eta}_t \end{cases}, \qquad t = 1, 2, \ldots, T. \tag{9.29}$$

The system (9.29) generalises previous representations in the state space by allowing matrices in the state and measurement equations to vary over time.

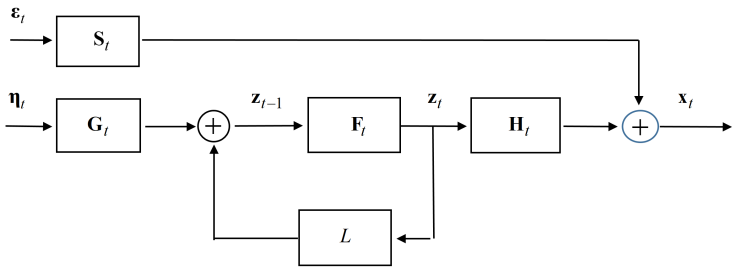


Figure 9.1: Block diagram of the dynamic state-space system (9.29).

The model (9.29) satisfies the stated properties of $\mathbf{f}$ and $\mathbf{h}$. Using the iterative replacement procedure, we obtain:

$$\begin{aligned}
\mathbf{z}_1 &= \mathbf{F}_1\mathbf{z}_0 + \mathbf{G}_1\boldsymbol{\eta}_1, \\
\mathbf{z}_2 &= \mathbf{F}_2\mathbf{F}_1\mathbf{z}_0 + \mathbf{F}_2\mathbf{G}_1\boldsymbol{\eta}_1 + \mathbf{G}_2\boldsymbol{\eta}_2, \\
\mathbf{z}_3 &= \mathbf{F}_3\mathbf{F}_2\mathbf{F}_1\mathbf{z}_0 + \mathbf{F}_3\mathbf{F}_2\mathbf{G}_1\boldsymbol{\eta}_1 + \mathbf{F}_3\mathbf{G}_2\boldsymbol{\eta}_2 + \mathbf{G}_3\boldsymbol{\eta}_3, \\
&\vdots \\
\mathbf{z}_t &= \textstyle\prod_{j=1}^{t}\mathbf{F}_j\mathbf{z}_0 + \sum_{k=1}^{t}\left(\prod_{j=k+1}^{t}\mathbf{F}_j\right)\mathbf{G}_k\boldsymbol{\eta}_k \\
&= \boldsymbol{\Phi}_{0,t}\mathbf{z}_0 + \textstyle\sum_{k=1}^{t}\boldsymbol{\Phi}_{k,t}\mathbf{G}_k\boldsymbol{\eta}_k,
\end{aligned} \tag{9.30}$$

where $\boldsymbol{\Phi}_{k,t} = \prod_{j=k+1}^{t}\mathbf{F}_j$, for $k < t$ and $\boldsymbol{\Phi}_{0,t} = \prod_{j=1}^{t}\mathbf{F}_j$.
Assuming the existence of the inverse matrices $\mathbf{F}_j^{-1}$ for $k < j < t$, we can define by convention:

$$\boldsymbol{\Phi}_{t,k} = \boldsymbol{\Phi}_{k,t}^{-1} = \left(\prod_{j=k+1}^{t}\mathbf{F}_j\right)^{-1} = \prod_{j=1}^{t-k}\mathbf{F}_{t-j+1}^{-1}, \quad k < t. \tag{9.31}$$

Furthermore, for $k = t$ we conventionally assume:

$$\boldsymbol{\Phi}_{k,k} = \boldsymbol{\Phi}_{t,t} = \mathbf{I}. \tag{9.32}$$

Thus, the matrix $\boldsymbol{\Phi}_{k,t}$ is the transition matrix that satisfies all the properties listed previously:

1) $\boldsymbol{\Phi}_{t,t} = \mathbf{I}$.

2) $\boldsymbol{\Phi}_{t,k}\boldsymbol{\Phi}_{k,s} = \boldsymbol{\Phi}_{t,s}, \quad \forall s \le k \le t.$
   Equivalently, $\boldsymbol{\Phi}_{t,s}^{-1} = (\boldsymbol{\Phi}_{t,k}\boldsymbol{\Phi}_{k,s})^{-1} = \boldsymbol{\Phi}_{k,s}^{-1}\boldsymbol{\Phi}_{t,k}^{-1} = \boldsymbol{\Phi}_{s,k}\boldsymbol{\Phi}_{k,t} = \boldsymbol{\Phi}_{s,t}$.
   Property 2 also holds in the alternative direction: $\boldsymbol{\Phi}_{s,t} = \boldsymbol{\Phi}_{s,k}\boldsymbol{\Phi}_{k,t}, \quad \forall s \le k \le t.$

3) $\boldsymbol{\Phi}_{t,k} = \boldsymbol{\Phi}_{k,t}^{-1}$, whenever $\mathbf{F}_j^{-1}$ exist for all $k < j \le t$.

The second property is the *semigroup property*. The third property is called the *temporal inversion property* and allows us to use the first two properties for any temporal direction. In fact, using the convention introduced above we can write:

$$\boldsymbol{\Phi}_{t,0}\boldsymbol{\Phi}_{0,t} = \boldsymbol{\Phi}_{t,t} = \mathbf{I}\,.$$

The inversion property for the $\mathbf{F}_j$ matrices is not always required, as will be seen later.

9.2.1 *ARMA* Process in State-Space Form

A univariate or multivariate $ARMA(p, q)$ model can always be represented in the state-space form. For example, the model:

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_m X_{t-m} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_{m-1} \varepsilon_{t-m+1}, \quad m = \max(p,\, q+1), \tag{9.33}$$

has a representation, in the form suggested by A. Harvey, obtained by defining a state vector of size m as follows:

$$\begin{cases} \mathbf{z}_t = \begin{bmatrix} \alpha_1 & \vdots & \\ \alpha_2 & \vdots & \mathbf{I}_{m-1} \\ \vdots & \vdots & \\ \cdots & \vdots & \cdots \\ \alpha_m & \vdots & \mathbf{0} \end{bmatrix} \mathbf{z}_{t-1} + \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{m-1} \end{bmatrix} \varepsilon_t \\ X_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \mathbf{z}_t \end{cases} . \tag{9.34}$$

This form can be applied even to moving-average models. For example, for the $MA(1)$ model we have:

$$\begin{cases} \mathbf{z}_t &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{z}_{t-1} + \begin{bmatrix} 1 \\ \beta \end{bmatrix} \varepsilon_t \\ X_t &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_t \end{cases} ,$$

that is,

$$\begin{cases} z_{1t} &= z_{2,t-1} + \varepsilon_t, \\ z_{2t} &= \beta \varepsilon_t, \\ X_t &= z_{1t}, \end{cases}$$

In the state-space representation of the $MA(1)$ model, the transition matrix $\mathbf{F} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is constant and not invertible. The non-invertibility of the transition matrix constitutes the mathematical "cost" incurred for a process that has a dual representation in terms of an autoregressive process of infinite order, while in the state-space form it is represented as an $AR(1)$ model.

9.3 Properties of State-Space Representation

This section discusses the properties of dynamic systems in state-space form that are important in the model-building procedure[140].

[140] See the synthesis presented in the volume Casals et al. (2016). See also Antsaklis and Michel (2006), Ch. 3.

Definition 9.3. *(Stability)*

A discrete-time state-space model is stable *if all eigenvalues λ_i of the transition matrix $\mathbf{F}$ satisfy $|\lambda_i| < 1$, that is, if all roots of $\det(\mathbf{F} - \lambda\mathbf{I}) = 0$ lie inside the unit circle. Equivalently, all roots of the characteristic equation $\det(\mathbf{I} - \mu\mathbf{F}) = 0$ lie outside the unit circle, since the roots satisfy $\mu = \frac{1}{\lambda_i}$, where λ_i are the eigenvalues of $\mathbf{F}$.*

If inputs are held constant (and disturbances are set to zero), stability implies that the state vector converges asymptotically to a steady state.

The stability property is widely treated in the theory of dynamic systems and in *control theory* with reference to linear and nonlinear, continuous and discrete systems. In this regard, for example, the definitions of *equilibrium point of the system; stability in Lyapunov's sense; global and local asymptotic stability; bounded and unbounded stability*, etc., are important but will not be treated here[141].

Definition 9.4. *(Observability)*

The state-space model is observable if, given the information on inputs and outputs, the initial state can be determined uniquely.

To make this aspect evident, consider the model (9.29), where, for simplicity, we take $\mathbf{S}_t \equiv \mathbf{0}$ and the output $\mathbf{x}_t \in \mathbb{R}^N$. If any $k < t$ is fixed, we have:

$$\begin{aligned}
&\mathbf{x}_0 = \mathbf{H}_0\mathbf{z}_0 \\
&\mathbf{x}_1 = \mathbf{H}_1\mathbf{z}_1 = \mathbf{H}_1\mathbf{F}_1\mathbf{z}_0 + \mathbf{H}_1\mathbf{G}_1\boldsymbol{\eta}_1 \\
&\mathbf{x}_2 = \mathbf{H}_2\mathbf{z}_2 = \mathbf{H}_2\mathbf{F}_2\mathbf{F}_1\mathbf{z}_0 + \mathbf{H}_2\mathbf{F}_2\mathbf{G}_1\boldsymbol{\eta}_1 + \mathbf{H}_2\mathbf{G}_2\boldsymbol{\eta}_2 \\
&\vdots \\
&\mathbf{x}_{k-1} = \underset{(N\times m)}{\mathbf{H}_{k-1}}\ \underset{(m\times m)}{\boldsymbol{\Phi}_{0,\,k-1}}\,\mathbf{z}_0 + \sum_{j=1}^{k-1}\mathbf{H}_{k-1}\boldsymbol{\Phi}_{j,\,k-1}\mathbf{G}_j\boldsymbol{\eta}_j
\end{aligned} \quad . \tag{9.35}$$

[141] For a thorough discussion of the concept of stability and its multiple definitions, see for example Michel and Hou (2008).

Hence:

$$\begin{aligned}\underset{(kN\times 1)}{\mathbf{x}} &= \begin{bmatrix}\mathbf{x}_0\\ \vdots \\ \mathbf{x}_{k-1}\end{bmatrix}\\ &= \underbrace{\begin{bmatrix}\mathbf{H}_0\\ \mathbf{H}_1\boldsymbol{\Phi}_{0,1}\\ \vdots \\ \mathbf{H}_{k-1}\boldsymbol{\Phi}_{0,k-1}\end{bmatrix}}_{(kN\times m)}\mathbf{z}_0\\ &\quad + \underbrace{\begin{bmatrix}0 & 0 & \cdots & 0\\ \mathbf{h}_{11} & 0 & \cdots & 0\\ \mathbf{h}_{21} & \mathbf{h}_{22} & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{h}_{k-1,1} & \mathbf{h}_{k-1,2} & \cdots & \mathbf{h}_{k-1,k-1}\end{bmatrix}}_{[kN\times(k-1)g]}\underbrace{\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\\ \vdots \\ \boldsymbol{\eta}_{k-1}\end{bmatrix}}_{[(k-1)g\times 1]},\end{aligned} \tag{9.36}$$

where

$$\underset{(N\times g)}{\mathbf{h}_{ij}} = \underset{(N\times m)}{\mathbf{H}_i}\ \underset{(m\times m)}{\boldsymbol{\Phi}_{j,\,i}}\ \underset{(m\times g)}{\mathbf{G}_j}\ .$$

Expression (9.36) shows that, given the input sequence $\{\boldsymbol{\eta}_t\}_1^{k-1}$, the output vector depends only on the initial state $\mathbf{z}_0$.

The observability property concerns the possibility of determining $\mathbf{z}_0$ from the system when the input $\{\boldsymbol{\eta}_t\}_1^{k-1}$ and the output $\{\mathbf{x}_t\}_1^{k-1}$ are known.

If we define:

$$\begin{cases}\underset{[kN\times m]}{\mathbf{R}(0,\,k-1)} = \begin{bmatrix}\mathbf{H}_0\\ \mathbf{H}_1\boldsymbol{\Phi}_{0,\,1}\\ \vdots \\ \mathbf{H}_k\boldsymbol{\Phi}_{0,\,k-1}\end{bmatrix}\\ \underset{[kN\times(k-1)g]}{\mathbf{T}} = \begin{bmatrix}0 & 0 & \cdots & 0\\ \mathbf{h}_{11} & 0 & \cdots & 0\\ \mathbf{h}_{21} & \mathbf{h}_{22} & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{h}_{k-1,1} & \mathbf{h}_{k-1,2} & \cdots & \mathbf{h}_{k-1,k-1}\end{bmatrix}\end{cases}, \quad \underset{[(k-1)g\times 1]}{\boldsymbol{\eta}} = \begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\\ \vdots \\ \boldsymbol{\eta}_{k-1}\end{bmatrix},$$

then (9.36) can be written compactly as:

$$\mathbf{x} = \mathbf{R}(0,\,k-1)\mathbf{z}_0 + \mathbf{T}\boldsymbol{\eta}. \tag{9.37}$$

Since $\boldsymbol{\eta}$ is known, we do not lose generality by taking $\boldsymbol{\eta} = \mathbf{0}$. Therefore, expression (9.37) reduces to:

$$\mathbf{x} = \mathbf{R}(0,\, k-1)\mathbf{z}_0\,. \tag{9.38}$$

The $\mathbf{T}$ matrix becomes irrelevant for the definition of state observability; only the $\mathbf{R}(0,\, k-1)$ matrix remains important. Therefore, the observability property is referred to the $\mathbf{R}(0,\, k-1)$ matrix rather than to the system (9.29).

Proposition 9.1. *(Observability)*
The state-space system (or equivalently, the pair $(\mathbf{F}, \mathbf{H})$*) is observable if and only if:*

$$\operatorname{rank}\{\mathbf{R}(0,\, k-1)\} = m\,. \tag{9.39}$$

where $\mathbf{R}(0,\, k-1)$ *is the associated observability matrix.*

Consequently, we have:

$$\mathbf{R}(0,\, k-1)'\mathbf{x} = \mathbf{R}(0,\, k-1)'\mathbf{R}(0,\, k-1)\mathbf{z}_0,$$

from which:

$$\mathbf{z}_0 = \left[\mathbf{R}(0,\, k-1)'\mathbf{R}(0,\, k-1)\right]^{-1}\mathbf{R}(0,\, k-1)'\mathbf{x}.$$

Application to invariant systems
For invariant systems, we have:

$$\boldsymbol{\Phi}_{0,\,t} = \prod_{j=1}^{t}\mathbf{F} = \mathbf{F}^t, \qquad \boldsymbol{\Phi}_{k,\,t} = \prod_{j=k+1}^{t}\mathbf{F} = \mathbf{F}^{\,t-k},$$

hence[142]:

$$\mathbf{R}(0,\, k-1)' = \begin{bmatrix}\mathbf{H}' & \mathbf{F}'\mathbf{H}' & \cdots & (\mathbf{F}')^{k-1}\mathbf{H}'\end{bmatrix}.$$

In the case of invariant systems, the $\mathbf{T}$ matrix is the triangular *Toeplitz matrix*:

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \mathbf{h}_1 & 0 & \cdots & 0 \\ \mathbf{h}_2 & \mathbf{h}_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{h}_{k-1} & \mathbf{h}_{k-2} & \cdots & \mathbf{h}_1 \end{bmatrix}.$$

A matrix is said to be *Toeplitz* if each element (i, j) depends only on the difference $(i-j)$. Consequently, its entries are constant along each diagonal, and the matrix is completely specified by the elements in its first column.

[142] See also Aoki (1990), p. 39.

Definition 9.5. *(Constructibility)*
Constructibility requires that it be possible to determine $\mathbf{z}_k$ *(instead of* $\mathbf{z}_0$*) from the system, when both the input* $\{\boldsymbol{\eta}_t\}_1^{k-1}$ *and the output* $\{\mathbf{x}_t\}_0^{k-1}$ *are known.*

Constructibility is a concept very close to that of observability.
Given the system (9.29), fixing a value $p < k$, in general we can write:

$$\begin{aligned}
\mathbf{z}_k &= \mathbf{F}_k\mathbf{z}_{k-1} + \mathbf{G}_k\boldsymbol{\eta}_k \\
&= \mathbf{F}_k\mathbf{F}_{k-1}\mathbf{z}_{k-2} + \mathbf{F}_k\mathbf{G}_{k-1}\boldsymbol{\eta}_{k-1} + \mathbf{G}_k\boldsymbol{\eta}_k \\
&= \mathbf{F}_k\mathbf{F}_{k-1}\mathbf{F}_{k-2}\mathbf{z}_{k-3} + \mathbf{F}_k\mathbf{F}_{k-1}\mathbf{G}_{k-2}\boldsymbol{\eta}_{k-2} + \mathbf{F}_k\mathbf{G}_{k-1}\boldsymbol{\eta}_{k-1} + \mathbf{G}_k\boldsymbol{\eta}_k \quad . \qquad (9.40) \\
&\vdots \\
&= \boldsymbol{\Phi}_{k-p,\,k}\mathbf{z}_{k-p} + \textstyle\sum_{j=0}^{p-1} \boldsymbol{\Phi}_{k-j,\,k}\mathbf{G}_{k-j}\boldsymbol{\eta}_{k-j}
\end{aligned}$$

Assuming that the inverse matrix $\mathbf{F}_j^{-1}$ exists for $k - p < j \leq k$, then multiplying both sides of (9.40) by the matrix $\boldsymbol{\Phi}_{k,\,k-p}$ we obtain:

$$\boldsymbol{\Phi}_{k,\,k-p}\mathbf{z}_k = \mathbf{z}_{k-p} + \sum_{j=0}^{p-1} \boldsymbol{\Phi}_{k,\,k-p}\boldsymbol{\Phi}_{k-j,\,k}\mathbf{G}_{k-j}\boldsymbol{\eta}_{k-j}. \qquad (9.41)$$

Using equality (9.41), the following system is obtained for $p = 1, 2, \ldots, k$:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{k-1} \\ \vdots \\ \mathbf{x}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{k-1}\boldsymbol{\Phi}_{k,\,k-1} \\ \vdots \\ \mathbf{H}_1\boldsymbol{\Phi}_{k,\,1} \\ \mathbf{H}_0\boldsymbol{\Phi}_{k,\,0} \end{bmatrix} \mathbf{z}_k + \mathbf{L} \begin{bmatrix} \boldsymbol{\eta}_k \\ \boldsymbol{\eta}_{k-1} \\ \vdots \\ \boldsymbol{\eta}_1 \end{bmatrix}, \qquad (9.42)$$

with $\mathbf{L} = \{\mathbf{l}_{ij}\}$ and $\mathbf{l}_{ij} = -\boldsymbol{\Phi}_{k,\,k-i}\boldsymbol{\Phi}_{k-j+1,\,k}$.
Also in this case we can define:

$$\underset{[(k+1)N\times m]}{\mathbf{R}(k,\,0)} = \begin{bmatrix} \mathbf{H}_{k-1}\boldsymbol{\Phi}_{k,\,k-1} \\ \vdots \\ \mathbf{H}_1\boldsymbol{\Phi}_{k,\,1} \\ \mathbf{H}_0\boldsymbol{\Phi}_{k,\,0} \end{bmatrix},$$

so that we rewrite (9.42) as follows:

$$\mathbf{x} = \mathbf{R}(k-1,\,0)\mathbf{z}_k + \mathbf{L}\boldsymbol{\eta}, \qquad (9.43)$$

where

$$\boldsymbol{\eta}' = \begin{bmatrix} \boldsymbol{\eta}_k' & \cdots & \boldsymbol{\eta}_1' \end{bmatrix}.$$

As for constructibility, similarly to observability, the conditions relating to the $\mathbf{R}(k-1,\,0)$ matrix are relevant; the following proposition states this:

Proposition 9.2. *(Constructibility)*
The state-space system is constructible if and only if

$$\operatorname{rank}\{\mathbf{R}(k-1,\,0)\} = m,$$

where $\mathbf{R}(k-1,\,0)$ *denotes the associated constructibility matrix.*

Consequently, assuming for simplicity $\boldsymbol{\eta} = \mathbf{0}$ and multiplying both sides of expression (9.43) by $\mathbf{R}(k-1,\,0)'$, we obtain

$$\mathbf{R}(k-1,\,0)'\mathbf{x} = \mathbf{R}(k-1,\,0)'\mathbf{R}(k-1,\,0)\mathbf{z}_k,$$

from which

$$\mathbf{z}_k = \left[\mathbf{R}(k-1,\,0)'\mathbf{R}(k-1,\,0)\right]^{-1}\mathbf{R}(k-1,\,0)'\mathbf{x}\,. \tag{9.44}$$

Application to invariant systems
In the case of invariant systems, assuming that $\mathbf{F}$ is nonsingular, the constructibility matrix can be written as:

$$\mathbf{R}(k-1,\,0)' = \left[\,(\mathbf{F}')^{-1}\mathbf{H}'\;\;(\mathbf{F}')^{-2}\mathbf{H}'\;\;\cdots\;\;(\mathbf{F}')^{-k}\mathbf{H}'\,\right]. \tag{9.45}$$

Some remarks are important:

1) Constructibility does not imply observability, as the following example shows.

   Example 9.1. The following matrices are defined[143]:

$$\mathbf{H} = \begin{bmatrix}1 & 1\end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.$$

   Then the observability matrix is singular:

$$\mathbf{R}(0,\,1) = \begin{bmatrix}\mathbf{H}\\ \mathbf{HF}\end{bmatrix} = \begin{bmatrix}1 & 1\\ 2 & 2\end{bmatrix}.$$

   In terms of the system, we have:

$$\begin{bmatrix}z_{11}\\ z_{21}\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}z_{10}\\ z_{20}\end{bmatrix},$$

   and

$$X_0 = z_{10} + z_{20},$$
$$X_1 = z_{10} + z_{20} + z_{10} + z_{20} = 2(z_{10} + z_{20}).$$

[143] This *Example* is suggested by Kailath (1980), p.98.

Only the sum $z_{10}+z_{20}$ is observable, not z_{10} and z_{20} individually. For constructibility, the matrix $\mathbf{R}(1,\,0)$ cannot be calculated because $\mathbf{F}^{-1}$ does not exist; however, we can write

$$\begin{bmatrix} z_{10} \\ z_{20} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z_{1,-1} \\ z_{2,-1} \end{bmatrix} = \begin{bmatrix} z_{1,-1} + z_{2,-1} \\ z_{1,-1} + z_{2,-1} \end{bmatrix}.$$

If X_{-1} is available, since $X_{-1} = z_{1,-1} + z_{2,-1}$ we obtain

$$z_{10} = X_{-1}, \qquad z_{20} = X_{-1}.$$

2) If $\mathbf{F}^{-1}$ exists in invariant systems, then constructibility and observability are equivalent. Indeed, with $\boldsymbol{\eta}_{k-j} \equiv 0$ $(0 \leq j < p)$, the relations derived from (9.41) reduce to

$$\mathbf{z}_k = \mathbf{F}^p \mathbf{z}_{k-p} \quad \Longleftrightarrow \quad \mathbf{F}^{-p}\mathbf{z}_k = \mathbf{z}_{k-p}.$$

3) Observability implies constructibility.

4) The concept of observability can be translated into econometric terms as *parametric identification.*

Definition 9.6. *(Reachability)*
The reachability property describes the possibility of reaching any state from the origin in finite time by a suitable choice of the input sequence.

Reachability is often referred to as controllability from the origin. The state of the system is reachable if, by appropriately choosing the input $\boldsymbol{\eta}_t$, any state vector can be reached starting from the null state condition $\mathbf{z}_0 = \mathbf{0}$.
Considering expression (9.40), if we set $p = k$, then:

$$\mathbf{z}_k = \boldsymbol{\Phi}_{0,\,k}\mathbf{z}_0 + \sum_{j=0}^{k-1} \boldsymbol{\Phi}_{k-j,\,k}\mathbf{G}_{k-j}\boldsymbol{\eta}_{k-j}.$$

In matrix form (with $\mathbf{z}_0$ known), this becomes:

$$\mathbf{z}_k - \boldsymbol{\Phi}_{0,\,k}\mathbf{z}_0 = \begin{bmatrix} \mathbf{G}_k \\ \boldsymbol{\Phi}_{k-1,\,k}\mathbf{G}_{k-1} \\ \vdots \\ \boldsymbol{\Phi}_{1,\,k}\mathbf{G}_1 \end{bmatrix}' \begin{bmatrix} \boldsymbol{\eta}_k \\ \boldsymbol{\eta}_{k-1} \\ \vdots \\ \boldsymbol{\eta}_1 \end{bmatrix}.$$

To analyze reachability, we consider the effect of inputs starting from the origin ($\mathbf{z}_0 = \mathbf{0}$); equivalently, we focus on the forced component of the state, defined as $\mathbf{z}_k^* = \mathbf{z}_k - \boldsymbol{\Phi}_{0,\,k}\mathbf{z}_0$.

Defining

$$\mathbf{Q}(k,\,0) = \big[\ \mathbf{G}_k \ \ \mathbf{\Phi}_{k-1,\,k}\mathbf{G}_{k-1} \ \ \cdots \ \ \mathbf{\Phi}_{1,\,k}\mathbf{G}_1\ \big],$$
$$\mathbf{z}_k^* = \mathbf{z}_k - \mathbf{\Phi}_{0,\,k}\mathbf{z}_0,$$
$$\boldsymbol{\eta}_k^* = \Big[\boldsymbol{\eta}_k' \ \ \boldsymbol{\eta}_{k-1}' \ \ \cdots \ \ \boldsymbol{\eta}_1'\Big]',$$

we obtain:

$$\mathbf{z}_k^* = \mathbf{Q}(k,\,0)\,\boldsymbol{\eta}_k^*\,. \tag{9.46}$$

To allow the transition from $\mathbf{0}$ to $\mathbf{z}_k^*$ it is necessary to determine the sequence of inputs; hence the following proposition is relevant.

Proposition 9.3. *(Reachability)*
The state-space system (or equivalently, the pair $(\mathbf{F}, \mathbf{G})$*) is reachable if and only if*

$$\operatorname{rank}\big\{\mathbf{Q}(k,\,0)\big\} = m\,. \tag{9.47}$$

As a consequence of Proposition 9.3, we can write:

$$\boldsymbol{\eta}_k^* = \mathbf{Q}(k,\,0)'\big[\mathbf{Q}(k,\,0)\mathbf{Q}(k,\,0)'\big]^{-1}\mathbf{z}_k^*\,, \tag{9.48}$$

where

$$\mathbf{Q}(k,\,0)'\big[\mathbf{Q}(k,\,0)\mathbf{Q}(k,\,0)'\big]^{-1}$$

is the *right inverse* of $\mathbf{Q}(k,\,0)$.[144]

Application to invariant systems
In the case of invariant systems,

$$\mathbf{Q}(k,\,0) = \big[\ \mathbf{G} \ \ \mathbf{FG} \ \ \cdots \ \ \mathbf{F}^{k-1}\mathbf{G}\ \big]\,. \tag{9.49}$$

Definition 9.7. *(Controllability)*
The system is controllable if any initial state can be driven to the origin in finite time by a suitable choice of the input sequence.

If $\mathbf{F}$ is nonsingular, controllability can be equivalently characterized through the time-reversed dynamics, leading to a controllability matrix involving powers of $\mathbf{F}^{-1}$ (see Figure 9.2).
In the case of invariant systems, if $\mathbf{F}$ is nonsingular, we have:

$$\mathbf{F}^{-k}\mathbf{Q}(k,\,0) = \big[\ \mathbf{F}^{-k}\mathbf{G} \ \ \mathbf{F}^{-k+1}\mathbf{G} \ \ \cdots \ \ \mathbf{F}^{-1}\mathbf{G}\ \big]. \tag{9.50}$$

[144] Recall that if a matrix $\mathbf{A}$ is $m \times n$ with $m < n$ and rank m, the pseudo-inverse $\mathbf{A}_g = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}$ is a right inverse since $\mathbf{A}\mathbf{A}_g = \mathbf{I}$.

Remark 9.5. Reachability (controllability from the origin) implies controllability (to the origin), whereas the converse is not true in general. If $\mathbf{F}$ is nonsingular, the two properties are equivalent.

Example 9.2. The following matrices are defined[145]:

$$\mathbf{G} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

Consequently,

$$\begin{bmatrix} \mathbf{G} & \mathbf{FG} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

is singular, so an arbitrary state (e.g. $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$) cannot be reached from the zero state regardless of the input.
On the other hand, any initial state can be returned to zero. Indeed, let

$$z_{11} = \alpha, \quad z_{21} = \beta, \quad \eta_0 = -(\alpha + \beta).$$

Then

$$\mathbf{z}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-\alpha - \beta) = \mathbf{0}.$$

If $\mathbf{F}$ is nonsingular, reachability and controllability are equivalent.

Following Kailath[146], *Figure 9.2* gives a useful summary of the above properties.

[145] See Kailath (1980), p. 96.
[146] Kailath (1980), Figure 2.3-5. p.100.

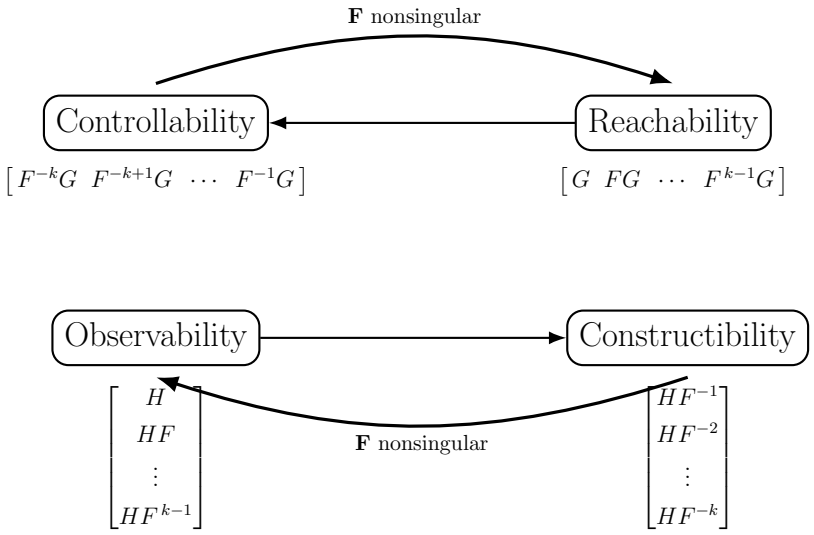


Figure 9.2: Relations among controllability, reachability, observability, and constructibility in linear discrete-time state-space systems.

Example 9.3. *[Observability and Reachability in a Trend–Cycle Model]* Consider a simple unobserved-components representation for a macroeconomic time series y_t (e.g., log real GDP), decomposed into a permanent component (trend) and a transitory component (cycle):

$$z_t = \begin{bmatrix} \tau_t \\ c_t \end{bmatrix}, \qquad y_t = Hz_t, \qquad H = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

The state equation is

$$z_t = Fz_{t-1} + G\eta_t, \qquad F = \begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

where

$$\eta_t = \begin{bmatrix} \varepsilon_t^\tau \\ \varepsilon_t^c \end{bmatrix}$$

collects a permanent shock ε_t^τ and a transitory shock ε_t^c.
This model corresponds to the standard unobserved-components representation commonly used in macroeconometrics[147].
The permanent component τ_t follows a random walk, while the cyclical component c_t follows a stationary AR(1) process with parameter ρ. The econometric question is whether the latent components are uniquely identifiable from the observed series and whether the structural shocks are sufficiently rich to generate the whole state space.

[147] See A. C. Harvey (1989).

Observability. The observability matrix is

$$\mathbf{R}(0,\,1) = \begin{bmatrix} H \\ HF \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & \rho \end{bmatrix}.$$

If $\rho \neq 1$, the determinant of $\mathbf{R}(0,\,1)$ is $\rho - 1 \neq 0$, so the matrix has full rank. Therefore, the state vector $(\tau_t, c_t)'$ can be uniquely recovered from the observed data (given the parameters). In econometric terms, the trend and the cycle are separately identifiable components.
If instead $\rho = 1$, the observability matrix becomes

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

which has rank one. In this case only the sum $\tau_t + c_t$ is observable, and the decomposition into permanent and transitory components is not uniquely identified. This illustrates how lack of observability translates into lack of identification of latent components.

Reachability. The reachability matrix for the invariant system is

$$\mathbf{Q}(2,\,0) = \begin{bmatrix} G & FG \end{bmatrix}.$$

Since $G = I_2$, the reachability matrix has full rank. Both components of the state vector are directly affected by structural shocks. Therefore, starting from the origin, any state vector can be generated by an appropriate sequence of structural shocks.
In econometric terms, the model allows shocks to affect all components of the system: the permanent and transitory components are both driven by independent innovations. The model allows for sufficiently rich dynamics to reproduce any combination of trend and cycle.
Suppose instead that the trend is assumed deterministic, so that

$$G = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The reachability matrix becomes

$$\mathbf{Q}(2,\,0) = \begin{bmatrix} 0 & 0 \\ 1 & \rho \end{bmatrix},$$

which has rank one. In this case, the permanent component cannot be generated by shocks starting from the origin. The model restricts the dynamics by excluding permanent innovations.

Econometrically, this corresponds to imposing the absence of permanent shocks. Such a restriction may be theoretically motivated, but it reduces the dynamic richness of the system. In this case, structural shocks do not affect all components of the state vector, and some directions of motion are excluded by construction.
This example illustrates the econometric interpretation of the structural properties discussed in this section. Observability is closely related to identification of latent components, while reachability concerns the ability of structural shocks to generate movements in all components of the state vector.
In empirical macroeconomic modelling, lack of observability leads to non-identification, whereas lack of reachability corresponds to restrictive assumptions on the role of structural shocks in the system.

Definition 9.8. *(Minimality)*
The state of a system is said to be minimal *if there is no alternative representation that produces the same output* $\mathbf{X}_t$ *with fewer states.*

It can be demonstrated that a model is minimal if and only if it is both observable and controllable.
To illustrate the importance of minimality, consider the following simple example.

Example 9.4. Consider the non-stochastic finite difference equation:

$$\Delta^2 y_t = a, \quad a \in \mathbb{R},\ a \text{ constant parameter.} \tag{9.51}$$

This equation could represent the acceleration of penetration of a new product in the market. Once the parameter a is fixed, the sequence y_t of the new product can be determined. A solution can be found using the *inverse truncated (or partial) sum operator* defined[148] as:

$$\Delta_d^{-1} = \sum_{k=0}^{d} L^k\,, \tag{9.52}$$

where d represents a time instant linked to the initial values of the operator argument. Its introduction is due to the fact that for finite time series, as in *Example 9.4*, $d \nrightarrow \infty$. Some properties of the operator Δ_d^{-1} are reported below:

1) $\Delta_d^{-1} c = (d+1)c, \quad c \in \mathbb{R},\ c \text{ constant}$

[148] Recall that $\Delta = 1 - L$, where L is the lag operator. In general, we can refer to the *inverse indefinite summation operator* Δ^{-1} for discrete processes, which is the equivalent of integration for continuous processes. If $\Delta Y_t = y_t$ then $\Delta^{-1} y_t = Y_t$. The application of the inverse operator to the time series requires the definition of an *arbitrary initial value*, e.g. y_{t-d-1}, so that $\Delta^{-1}\Delta y_t = y_t - y_{t-d-1}$.

$$\text{2)} \quad \begin{aligned} \Delta_d^{-1}\Delta &= \sum_{k=0}^{d} L^k - \sum_{k=0}^{d} L^{k+1} = 1 + \sum_{k=1}^{d} L^k - \left(\sum_{k=0}^{d-1} L^{k+1} + L^{d+1} \right) \\ &= 1 + \sum_{k=1}^{d} L^k - \sum_{s=1}^{d} L^s - L^{d+1} \\ &= 1 - L^{d+1} \end{aligned}$$

These properties are used to obtain the solution for y_t through inverse iterative operations in two steps.

Step 1: The Δ_d^{-1} operator is applied to the starting equation (9.51), obtaining:

$$\begin{aligned} \Delta_d^{-1}\Delta^2 y_t &= \Delta_d^{-1} a, \\ \Delta_d^{-1}\Delta(\Delta y_t) &= (d+1)a, \\ (1 - L^{d+1})\Delta y_t &= (d+1)a, \\ \Delta y_t - \Delta y_{t-d-1} &= (d+1)a\,, \end{aligned}$$

that is:

$$\Delta y_t = \Delta y_{t-d-1} + (d+1)a\,. \tag{9.53}$$

When an arbitrary value of Δy_{t-d-1} is assigned, the system output in terms of Δy_t for all t is uniquely determined. However, we want to find the solution in terms of y_t, so we proceed to the second step.

Step 2: The Δ_d^{-1} operator is applied again to (9.53), obtaining:

$$\begin{aligned} \Delta_d^{-1}\Delta y_t &= \Delta_d^{-1}\left[\Delta y_{t-d-1} + (d+1)a\right] \\ (1 - L^{d+1})y_t &= (d+1)\left[\Delta y_{t-d-1} + (d+1)a\right] \\ y_t - y_{t-d-1} &= (d+1)\Delta y_{t-d-1} + (d+1)^2 a\,, \end{aligned}$$

from which:

$$y_t = y_{t-d-1} + (d+1)\Delta y_{t-d-1} + (d+1)^2 a\,. \tag{9.54}$$

The solution in (9.54) shows the need to assign a new arbitrary initial value y_{t-d-1}. This value, together with Δy_{t-d-1} previously set, uniquely determines the system solution in terms of y_t for all t.
The previous derivation highlights that two independent initial conditions are required to determine the trajectory of y_t uniquely.
The y_t variable cannot represent a state of the system because the assignment of the initial y_{t-d-1} value is not sufficient to uniquely determine y_t. It is necessary to assign an initial value also to Δy_{t-d-1}. Not even the set $\{\Delta^2 y_t,\, \Delta y_t,\, y_t\}$ can constitute the state of

the system since it is not the minimal set for the determination of y_t, the variable $\Delta^2 y_t$ being superfluous.
In fact, a representation in the state space can be given by defining:

$$\mathbf{z}_t = \begin{bmatrix} y_t \\ \Delta y_t \end{bmatrix},$$

for which:

$$\begin{cases} \mathbf{z}_t = \begin{bmatrix} 1 & d+1 \\ 0 & 1 \end{bmatrix} \mathbf{z}_{t-d-1} + \begin{bmatrix} (d+1)^2 \\ d+1 \end{bmatrix} a, \\ y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_t. \end{cases}$$

This state-space representation synthesizes both equations (9.53) and (9.54), and shows that the minimal state vector is given by $\mathbf{z}_t = (y_t, \Delta y_t)'$, which contains the essential information without redundancy.

9.4 Kalman Filter

The State-space model (9.29) is the basis of a recursive algorithm known as the *Kalman filter (KF).*
For completeness, we restate model (9.29) in probabilistic terms as follows:

$$\begin{cases} \underset{(k\times 1)}{\mathbf{x}_t} = \underset{(k\times m)}{\mathbf{H}_t}\ \mathbf{z}_t + \underset{(k\times n)}{\mathbf{S}_t}\ \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim NID(\mathbf{0}, \mathbf{R}) \\ \mathbf{z}_t = \underset{(m\times m)}{\mathbf{F}_t}\ \mathbf{z}_{t-1} + \underset{(m\times g)}{\mathbf{G}_t}\ \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim NID(\mathbf{0}, \mathbf{Q}) \end{cases} \qquad t = 1, 2, \ldots, T \tag{9.55}$$

where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_s$ are stochastically independent for all t, s.
For simplicity, when deriving the Kalman filter we do not time-index the matrices $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{S}$. The resulting formulas remain valid in the time-varying case by allowing these matrices to depend on t.
The distribution of the initial state is assumed to be known: $\mathbf{z}_0 \sim N(\bar{\mathbf{z}}_0, \mathbf{P}_0)$.
Starting from this initial state, the aim is to derive a recursive algorithm valid for each time t, conditional on the information available at $t-1$. The information set up to time $t-1$ is denoted by $\mathbf{Z}_{t-1}$.

Prediction step. The *one-step-ahead state predictor* is defined as:

$$\begin{aligned} \mathbf{z}_{t|t-1} &= E[\mathbf{z}_t \,|\, \mathbf{Z}_{t-1}] \\ &= \mathbf{F} E[\mathbf{z}_{t-1} \,|\, \mathbf{Z}_{t-1}] + \mathbf{G} E[\boldsymbol{\eta}_t \,|\, \mathbf{Z}_{t-1}] \\ &= \mathbf{F}\mathbf{z}_{t-1|t-1}. \end{aligned} \tag{9.56}$$

Remark 9.6. Since $\mathbf{z}_{t-1} \subset \mathbf{Z}_{t-1}$, we have $E[\mathbf{z}_{t-1} \,|\, \mathbf{Z}_{t-1}] = \mathbf{z}_{t-1}$. For convenience, we denote $\mathbf{z}_{t-1|t-1} = E[\mathbf{z}_{t-1} \,|\, \mathbf{Z}_{t-1}]$.

The relation between the state and the system output error is described by the vector

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{z}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} .$$

Its conditional mean is

$$E\left[\mathbf{y}_t | \mathbf{Z}_{t-1}\right] = \begin{bmatrix} E\left[\mathbf{z}_t | \mathbf{Z}_{t-1}\right] \\ E\left[\boldsymbol{\varepsilon}_t | \mathbf{Z}_{t-1}\right] \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{t|t-1} \\ \mathbf{0} \end{bmatrix} . \tag{9.57}$$

since $E[\boldsymbol{\varepsilon}_t \,|\, \mathbf{Z}_{t-1}] = \mathbf{0}$ by independence of the errors.
The corresponding conditional variance is

$$\begin{aligned} Var\left[\mathbf{y}_t | \mathbf{Z}_{t-1}\right] &= E\left\{ \begin{bmatrix} \mathbf{z}_t - \mathbf{z}_{t|t-1} \\ \boldsymbol{\varepsilon}_t \end{bmatrix} \begin{bmatrix} (\mathbf{z}_t - \mathbf{z}_{t|t-1})' & \boldsymbol{\varepsilon'}_t \end{bmatrix} | \mathbf{Z}_{t-1} \right\} \\ &= \begin{bmatrix} E(\mathbf{z}_t - \mathbf{z}_{t|t-1})(\mathbf{z}_t - \mathbf{z}_{t|t-1})' | \mathbf{Z}_{t-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{t|t-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{aligned} . \tag{9.58}$$

where $\mathbf{P}_{t|t-1}$ is the variance matrix of the state prediction error and stochastic independence between model errors implies that

$$E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon'}_t | \mathbf{Z}_{t-1}) = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon'}_t) = \mathbf{R} .$$

The variance of the state prediction error can be expressed in terms of the variance updated at the previous step $t-1$. Indeed, we can write:

$$\begin{aligned} \mathbf{z}_t - \mathbf{z}_{t|t-1} &= \mathbf{z}_t - \mathbf{F}\mathbf{z}_{t-1|t-1} \\ &= \mathbf{F}\mathbf{z}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t - \mathbf{F}\mathbf{z}_{t-1|t-1} \\ &= \mathbf{F}\big(\mathbf{z}_{t-1} - \mathbf{z}_{t-1|t-1}\big) + \mathbf{G}\boldsymbol{\eta}_t . \end{aligned} \tag{9.59}$$

which leads to

$$\begin{aligned} Var\left[(\mathbf{z}_t - \mathbf{z}_{t|t-1}) | \mathbf{Z}_{t-1}\right] &= Var\left\{\left[\mathbf{F}(\mathbf{z}_{t-1} - \mathbf{z}_{t-1|t-1}) + \mathbf{G}\boldsymbol{\eta}_t\right] | \mathbf{Z}_{t-1}\right\} \\ &= \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{G}\mathbf{Q}\mathbf{G}' \end{aligned} .$$

Hence, the prediction error variance recursion is

$$\mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{G}\mathbf{Q}\mathbf{G}' , \tag{9.60}$$

known as the *Riccati recursion*.

Relations (9.56) and (9.60) represent the first two formulas of the Kalman filter. They express, respectively, the prediction of the state at time $t-1$ for the following time t and the prediction error variance. To complete the formulation of the prediction component of the filter, we consider the further transformation that relates the state and the output at the same time t:

$$\mathbf{w}_t = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{z}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{H}\mathbf{z}_t + \boldsymbol{\varepsilon}_t \end{bmatrix} = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{x}_t \end{bmatrix}. \tag{9.61}$$

For the vector $\mathbf{w}_t$ we compute the mean and variance conditional on the information set $\mathbf{Z}_{t-1}$, obtaining:

$$\mathbf{w}_{t|t-1} = E[\mathbf{w}_t \mid \mathbf{Z}_{t-1}] = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{z}_{t|t-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{t|t-1} \\ \mathbf{H}\mathbf{z}_{t|t-1} \end{bmatrix}, \tag{9.62}$$

and

$$\begin{aligned} Var[\mathbf{w}_t \mid \mathbf{Z}_{t-1}] &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{H} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{t|t-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{H}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{H}' \\ \mathbf{H}\mathbf{P}_{t|t-1} & \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R} \end{bmatrix}. \end{aligned} \tag{9.63}$$

Update step. At time t, the new observation $\mathbf{x}_t$ becomes available and allows us to update the distribution of the state variable conditional on the new information. We define:

$$\mathbf{z}_{t|t} = E[\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1}], \qquad \mathbf{P}_{t|t} = Var[\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1}].$$

The determination of these quantities is based on the standard result for the conditional distribution of a multivariate normal random vector (*Proposition 9.4*).

Proposition 9.4. *(Conditional normal distribution).*
Let

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\alpha \\ \boldsymbol{\mu}_\beta \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\beta} \\ \boldsymbol{\Sigma}_{\beta\alpha} & \boldsymbol{\Sigma}_{\beta\beta} \end{bmatrix} \right), \quad \textit{with } \boldsymbol{\Sigma}_{\beta\beta} \textit{ nonsingular.}$$

Then

$$\boldsymbol{\alpha} \mid \boldsymbol{\beta} \sim \mathcal{N}\Big(\boldsymbol{\mu}_\alpha + \boldsymbol{\Sigma}_{\alpha\beta}\boldsymbol{\Sigma}_{\beta\beta}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta),\ \boldsymbol{\Sigma}_{\alpha\alpha} - \boldsymbol{\Sigma}_{\alpha\beta}\boldsymbol{\Sigma}_{\beta\beta}^{-1}\boldsymbol{\Sigma}_{\beta\alpha} \Big).$$

Moreover, defining

$$\mathbf{u} = \boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha - \boldsymbol{\Sigma}_{\alpha\beta}\boldsymbol{\Sigma}_{\beta\beta}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta),$$

we have the orthogonality $\mathrm{Cov}(\mathbf{u}, \boldsymbol{\beta}) = \mathbf{0}$.

Applying this result with $\boldsymbol{\alpha} = \mathbf{z}_t$ and $\boldsymbol{\beta} = \mathbf{x}_t$, we obtain:

$$\begin{aligned} &\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1} \sim N\left(\mathbf{z}_{t|t},\, \mathbf{P}_{t|t}\right) \\ &\mathbf{z}_{t|t} = \mathbf{z}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}'\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R}\right)^{-1}\left(\mathbf{x}_t - \mathbf{H}\mathbf{z}_{t|t-1}\right) \\ &\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}'\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R}\right)^{-1}\mathbf{H}\mathbf{P}_{t|t-1} \end{aligned} \quad . \tag{9.64}$$

The Kalman filter is therefore characterized by formulas (9.56), (9.60), and (9.64). In this way, the Kalman recursions can be viewed as repeated applications of the conditional normal update.
To highlight some important features of the filter, let us introduce the following definitions:

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}'\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R}\right)^{-1} \tag{9.65}$$

$$\nu_t = \mathbf{x}_t - \mathbf{H}\mathbf{z}_{t|t-1} \tag{9.66}$$

Expression (9.65) defines the *Kalman gain* K_t, while (9.66) defines the *innovation*, i.e. the one-step-ahead prediction error of the observation equation. The innovation measures the discrepancy between the actual observation $\mathbf{x}_t$ and its prediction $\mathbf{H}_t\mathbf{z}_{t|t-1}$ and is the quantity used to update the state estimate.

Using these definitions, the update equations in (9.64) can be rewritten in compact form:

$$\begin{aligned} \mathbf{z}_{t|t} &= \mathbf{z}_{t|t-1} + \mathbf{K}_t\nu_t \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{H}\mathbf{P}_{t|t-1} \end{aligned} \quad . \tag{9.67}$$

Summary of the algorithm. In conclusion, the Kalman filter algorithm is summarized by the following set of recursive formulas:

$$\begin{aligned} 1)\quad \mathbf{z}_{t|t-1} &= \mathbf{F}\mathbf{z}_{t-1|t-1} \\ 2)\quad \mathbf{P}_{t|t-1} &= \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{G}\mathbf{Q}\mathbf{G}' \\ 3)\quad \mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{H}'\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R}\right)^{-1} \\ 4)\quad \nu_t &= \mathbf{x}_t - \mathbf{H}\mathbf{z}_{t|t-1} \\ 5)\quad \mathbf{z}_{t|t} &= \mathbf{z}_{t|t-1} + \mathbf{K}_t\nu_t \\ 6)\quad \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{H}\mathbf{P}_{t|t-1} \end{aligned} \quad . \tag{9.68}$$

9.5 OLS Method versus Kalman Filter Method

Considering a model with constant parameters, we compare the Kalman filter—which processes data sequentially—with the OLS method, which is a batch estimation approach in which all observations are processed jointly.
We refer to the regression model:

$$y_t = \sum_{i=1}^{k} \beta_i x_{it} + u_t, \qquad x_{1t} = 1 \ \ \forall t, \qquad u_t \sim NID(0,\, \sigma^2)\,. \tag{9.69}$$

This regression model can be rewritten in the state-space form (9.55). To this end, define[149]:

$$\begin{aligned}
&\mathbf{x}_t = \big[y_t\big],\\
&\mathbf{z}_t = \big[\beta_{1t},\, \beta_{2t},\, \ldots,\, \beta_{kt}\big]',\\
&\mathbf{H}_t = \big[1,\, x_{2t},\, \ldots,\, x_{kt}\big],\\
&\mathbf{S}_t = \mathbf{I},\\
&\mathbf{F}_t = \mathbf{I},\\
&\mathbf{G}_t = \mathbf{0},\\
&\boldsymbol{\varepsilon}_t = \big[u_t\big],\\
&\mathbf{R} = \big[\sigma^2\big],\\
&\mathbf{Q} = \mathbf{0}.
\end{aligned} \tag{9.70}$$

Note that $\mathbf{F}_t = \mathbf{I}$ imposes the constraint $\mathbf{z}_t = \mathbf{z}_{t-1}$, i.e. the constancy of the regression coefficients. To apply the Kalman filter, we start from an arbitrary prior for the initial state, $\mathbf{z}_0 \sim N(\bar{\mathbf{z}}_0,\, \mathbf{P}_0)$. The algorithm becomes:

$$\begin{aligned}
&1)\ \ \mathbf{z}_{1|0} = \mathbf{z}_{0|0} = \bar{\mathbf{z}}_0,\\
&2)\ \ \mathbf{P}_{1|0} = \mathbf{P}_{0|0} = \mathbf{P}_0,\\
&3)\ \ \mathbf{v}_1 = \mathbf{x}_1 - \mathbf{H}_1\mathbf{z}_{1|0} = \mathbf{x}_1 - \mathbf{H}_1\bar{\mathbf{z}}_0,\\
&4)\ \ c_{1|0} = \mathbf{H}_1\mathbf{P}_{1|0}\mathbf{H}_1' + \sigma^2 = \mathbf{H}_1\mathbf{P}_0\mathbf{H}_1' + \sigma^2,\\
&5)\ \ \mathbf{K}_1 = \mathbf{P}_{1|0}\mathbf{H}_1'\left(\mathbf{H}_1\mathbf{P}_{1|0}\mathbf{H}_1' + \sigma^2\right)^{-1} = \mathbf{P}_0\mathbf{H}_1'\left(\mathbf{H}_1\mathbf{P}_0\mathbf{H}_1' + \sigma^2\right)^{-1},\\
&6)\ \ \mathbf{z}_{1|1} = \mathbf{z}_{1|0} + \mathbf{K}_1\mathbf{v}_1,\\
&7)\ \ \mathbf{P}_{1|1} = \mathbf{P}_{1|0} - \mathbf{K}_1\mathbf{H}_1\mathbf{P}_{1|0} = \left(\mathbf{I} - \mathbf{K}_1\mathbf{H}_1\right)\mathbf{P}_0.
\end{aligned} \tag{9.71}$$

[149] Here $\mathbf{x}_t$ denotes the (possibly vector-valued) output in the state-space form; in this example it is scalar, hence $\mathbf{x}_t = [y_t]$. The regressors in the original regression are denoted by x_{it}.

At step $t = T$, steps 6) and 7) become:

$$\begin{aligned} &6^*) \;\; \mathbf{z}_{T|T} = \mathbf{z}_{T|T-1} + \mathbf{K}_T \mathbf{v}_T, \\ &7^*) \;\; \mathbf{P}_{T|T} = (\mathbf{I} - \mathbf{K}_T \mathbf{H}_T)\, \mathbf{P}_{T|T-1}. \end{aligned} \tag{9.72}$$

Are the results obtained in 6) and 7*) comparable with those produced by OLS?*
To answer this question, it is useful to express 6*) and 7*) in a non-recursive form, in terms of the initial values $\bar{\mathbf{z}}_0$ and $\mathbf{P}_0$.
Equivalently, we consider the posterior distribution of the parameter vector $\boldsymbol{\beta}$ given all observations $y_1, y_2, \ldots, y_T$.
Suppose that $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$ is the prior (initial) distribution of the parameters.

How does the distribution change after observing y_1?

We can proceed by observing that: y_1 with $\mathbf{x'}_1\boldsymbol{\beta} \sim N(\mathbf{x'}_1\boldsymbol{\beta}_0, \mathbf{x'}_1\boldsymbol{\Sigma}\mathbf{x}_1)$. The unconditional value of y_1 is obtained using the *law of iterated expectations* and is:

$$E(y_1) = E_{\mathbf{x'}_1\boldsymbol{\beta}} E_{y_1|_{\mathbf{x'}_1\boldsymbol{\beta}}}(\mathbf{x'}_1\boldsymbol{\beta} + \varepsilon_1) = E_{\mathbf{x'}_1\boldsymbol{\beta}}(\mathbf{x'}_1\boldsymbol{\beta}) = \mathbf{x'}_1\boldsymbol{\beta}_0\,.$$

Similarly, the unconditional variance of y_1 is:

$$\begin{aligned} Var(y_1) &= Ey_1^2 - (Ey_1)^2 = E_{\mathbf{x'}_1\boldsymbol{\beta}} E_{y_1|_{\mathbf{x'}_1\boldsymbol{\beta}}}(y_1^2) - \mathbf{x'}_1\boldsymbol{\beta}_0\boldsymbol{\beta'}_0\mathbf{x}_1 \\ &= \mathbf{x'}_1\boldsymbol{\Sigma}\mathbf{x}_1 + \sigma^2 + \mathbf{x'}_1\boldsymbol{\beta}_0\boldsymbol{\beta'}_0\mathbf{x}_1 - \mathbf{x'}_1\boldsymbol{\beta}_0\boldsymbol{\beta'}_0\mathbf{x}_1 = \mathbf{x'}_1\boldsymbol{\Sigma}\mathbf{x}_1 + \sigma^2 \end{aligned}$$

The calculation of the $Cov(y_1, \boldsymbol{\beta})$ is given by:

$$\begin{aligned} Cov(y_1, \boldsymbol{\beta}) &= E(\boldsymbol{\beta} y'_1) - E(\boldsymbol{\beta})E(y_1)' \\ &= E_{\mathbf{x'}_1\boldsymbol{\beta}} E_{y_1|_{\mathbf{x'}_1\boldsymbol{\beta}}}(\boldsymbol{\beta} y'_1) - E_{\mathbf{x'}_1\boldsymbol{\beta}} E_{y_1|_{\mathbf{x'}_1\boldsymbol{\beta}}}(\boldsymbol{\beta}) E_{\mathbf{x'}_1\boldsymbol{\beta}} E_{y_1|_{\mathbf{x'}_1\boldsymbol{\beta}}}(y_1)' \\ &= E_{\mathbf{x'}_1\boldsymbol{\beta}}(\boldsymbol{\beta}\boldsymbol{\beta'}\mathbf{x}_1) - E_{\mathbf{x'}_1\boldsymbol{\beta}}(\boldsymbol{\beta}) E_{\mathbf{x'}_1\boldsymbol{\beta}}(\boldsymbol{\beta'}\mathbf{x}_1) \\ &= (\boldsymbol{\Sigma} + \boldsymbol{\beta}_0\boldsymbol{\beta'}_0)\mathbf{x}_1 - \boldsymbol{\beta}_0\boldsymbol{\beta'}_0\mathbf{x}_1 \\ &= \boldsymbol{\Sigma}\mathbf{x}_1 \end{aligned} \quad .$$

Therefore, the joint distribution of the regression coefficients and the first observation is:

$$\begin{bmatrix} \boldsymbol{\beta} \\ y_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta}_0 \\ \mathbf{x}'_1\boldsymbol{\beta}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{x}_1 \\ \mathbf{x}'_1\boldsymbol{\Sigma} & \mathbf{x}'_1\boldsymbol{\Sigma}\mathbf{x}_1 + \sigma^2 \end{bmatrix} \right). \tag{9.73}$$

Using *Proposition 9.4* we obtain:

$$\begin{aligned} &\boldsymbol{\beta} \mid y_1 \sim N\big(\boldsymbol{\mu}_{\boldsymbol{\beta}|y_1},\, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|y_1}\big)\,, \\ &\boldsymbol{\mu}_{\boldsymbol{\beta}|y_1} = \boldsymbol{\beta}_0 + \boldsymbol{\Sigma}\mathbf{x}_1 \left(\mathbf{x}'_1\boldsymbol{\Sigma}\mathbf{x}_1 + \sigma^2\right)^{-1} (y_1 - \mathbf{x}'_1\boldsymbol{\beta}_0)\,, \\ &\boldsymbol{\Sigma}_{\boldsymbol{\beta}|y_1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{x}_1 \left(\mathbf{x}'_1\boldsymbol{\Sigma}\mathbf{x}_1 + \sigma^2\right)^{-1} \mathbf{x}'_1\boldsymbol{\Sigma}. \end{aligned} \tag{9.74}$$

By taking $\bar{\mathbf{z}}_0 = \boldsymbol{\beta}_0$ and $\mathbf{P}_0 = \boldsymbol{\Sigma}$, formulation (9.74) coincides with steps 6) and 7) in (9.71). Extending this procedure to all observations, the joint distribution of $\boldsymbol{\beta}$ and $\mathbf{y} = (y_1, \ldots, y_T)'$ is:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta}_0 \\ \mathbf{X}\boldsymbol{\beta}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{X}' \\ \mathbf{X}\boldsymbol{\Sigma} & \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I} \end{bmatrix} \right). \tag{9.75}$$

Hence:

$$\begin{aligned} &\boldsymbol{\beta} \mid \mathbf{y} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}},\, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}}\right), \\ &\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} = \boldsymbol{\beta}_0 + \boldsymbol{\Sigma}\mathbf{X}' \left(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I}\right)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0), \\ &\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{X}' \left(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I}\right)^{-1} \mathbf{X}\boldsymbol{\Sigma}. \end{aligned} \tag{9.76}$$

Now consider the OLS estimator and its variance matrix:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ols} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \\ \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{9.77}$$

To compare (9.77) with (9.76), it is useful to reformulate the latter through appropriate matrix transformations based on the following *Proposition 9.5*.

Proposition 9.5. *Let* $\mathbf{P}$, $\mathbf{R}$, *and* $\mathbf{M}$ *be* $n \times n$, $m \times m$, *and* $m \times m$ *matrices, respectively. If both* $\mathbf{P}$ *and* $\mathbf{R}$ *are positive definite, then the following equalities hold*[150]*:*

$$\mathbf{I} - \mathbf{P}\mathbf{M}'(\mathbf{M}\mathbf{P}\mathbf{M}' + \mathbf{R})^{-1}\mathbf{M} = (\mathbf{I} + \mathbf{P}\mathbf{M}'\mathbf{R}^{-1}\mathbf{M})^{-1}, \tag{9.78}$$

$$\mathbf{P}\mathbf{M}'(\mathbf{M}\mathbf{P}\mathbf{M}' + \mathbf{R})^{-1} = (\mathbf{I} + \mathbf{P}\mathbf{M}'\mathbf{R}^{-1}\mathbf{M})^{-1}\mathbf{P}\mathbf{M}'\mathbf{R}^{-1}, \tag{9.79}$$

$$\mathbf{P} - \mathbf{P}\mathbf{M}'(\mathbf{M}\mathbf{P}\mathbf{M}' + \mathbf{R})^{-1}\mathbf{M}\mathbf{P} = (\mathbf{P}^{-1} + \mathbf{M}'\mathbf{R}^{-1}\mathbf{M})^{-1}. \tag{9.80}$$

By setting $\mathbf{P} = \boldsymbol{\Sigma}$, $\mathbf{M} = \mathbf{X}$, and $\mathbf{R} = \sigma^2\mathbf{I}$, these equalities become:

$$\mathbf{I} - \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{X} = (\mathbf{I} + \sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}, \tag{9.81}$$

$$\boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1} = (\mathbf{I} + \sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Sigma}\mathbf{X}'\sigma^{-2}, \tag{9.82}$$

$$\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}. \tag{9.83}$$

Hence:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} &= \boldsymbol{\beta}_0 + \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) \\ &= \boldsymbol{\beta}_0 + \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{y} - \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\beta}_0 \\ &= \left[\mathbf{I} - \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{X}\right]\boldsymbol{\beta}_0 + \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{y} \end{aligned} \tag{9.84}$$

[150] See Jazwinski (1970), p. 262.

Using equality (9.81) for the matrix multiplying $\boldsymbol{\beta}_0$ and equality (9.82) for the matrix multiplying $\mathbf{y}$, we obtain:

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} &= (\mathbf{I}+\sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}\beta_0 + (\mathbf{I}+\sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Sigma}\mathbf{X}'\sigma^{-2}\mathbf{y} \\
&= (\mathbf{I}+\sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\beta}_0 + \sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{y}) \\
&= (\mathbf{I}+\sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_0 + \sigma^{-2}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{y}) \\
&= (\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{X}'\mathbf{y}) \\
&= (\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}'\mathbf{y})
\end{aligned} \quad . \tag{9.85}$$

Furthermore, by using equality (9.83) we obtain:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}' + \sigma^2\mathbf{I})^{-1}\mathbf{X}\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}\frac{\sigma^2}{\sigma^2} \\
&= \sigma^2(\sigma^2\boldsymbol{\Sigma}^{-1}+\mathbf{X}'\mathbf{X})^{-1}
\end{aligned} \quad . \tag{9.86}$$

The matrix transformations yield the following relevant results:

1) An equivalent form of the conditional mean, which can be interpreted as a weighted average. The weights reflect the precision of the observed estimate $\hat{\boldsymbol{\beta}}_{ols}$ and the degree of prior belief in the initial parameter $\boldsymbol{\beta}_0$. From (9.85) we obtain:

$$\begin{aligned}
&(\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}'\mathbf{y}\big) \\
&= (\boldsymbol{\Sigma}^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}'\mathbf{X}\,\hat{\boldsymbol{\beta}}_{ols}\big)\,,
\end{aligned} \tag{9.87}$$

   where the weights associated with $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{ols}$ are $\boldsymbol{\Sigma}^{-1}$ and $\sigma^{-2}\mathbf{X}'\mathbf{X}$, respectively.

   The former reflects the degree of prior belief in the value of $\boldsymbol{\beta}_0$, while the latter represents the precision of $\hat{\boldsymbol{\beta}}_{ols}$ as a synthetic measure of the unknown parameter $\boldsymbol{\beta}$. By substituting the expression of $\hat{\boldsymbol{\beta}}_{ols}$ into (9.87), one exactly recovers the last equality of (9.85).

2) A second result follows from (9.86). In the positive definite ordering sense, it is evident that

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} < \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

   since:

$$(\sigma^2\boldsymbol{\Sigma}^{-1}+\mathbf{X}'\mathbf{X})^{-1} < (\mathbf{X}'\mathbf{X})^{-1}. \tag{9.88}$$

   However, this inequality does not imply that the Kalman filter is more efficient in the usual (frequentist) sense, because the posterior mean $\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}}$ is generally a biased estimator of $\boldsymbol{\beta}$:

$$E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}}\big) \neq \boldsymbol{\beta}. \tag{9.89}$$

   In contrast, the OLS estimator is unbiased.

Equality between $\boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}}$ and $\hat{\boldsymbol{\beta}}_{ols}$ is achieved when $\boldsymbol{\Sigma}^{-1} = \mathbf{0}$. This condition is intuitive: assigning very large values to the prior variance matrix $\boldsymbol{\Sigma}$ is equivalent to having no prior preference for the initial values $\boldsymbol{\beta}_0$, thereby allowing the Kalman filter to "learn" exclusively from the data.

Glossary of Acronyms and Abbreviations

ADF Augmented Dickey–Fuller

ADL Autoregressive Distributed Lag

AR Autoregressive

ARMA Autoregressive Moving Average

ARIMA Autoregressive Integrated Moving Average

DF Dickey-Fuller

DS Difference-Stationary

ECM Error Correction Mechanism

GLS Generalized Least Squares

i.i.d. Independent and Identically Distributed

I(0) Integrated of order zero

I(1) Integrated of order one

L Lag

MA Moving Average

NID Normally and Independently Distributed

OLS Ordinary Least Squares

PACF Partial Autocorrelation Function

PP Phillips-Perron

RW Random Walk

TS Trend-Stationary

VAR Vector Autoregression

VARMA Vector Autoregressive Moving Average

VECM Vector Error Correction Mechanism

WN White Noise

References

Anderson, Theodore W. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley & Sons.

Antsaklis, Panos J. and Anthony N. Michel (2006). *Linear Systems*. Boston, MA: Birkhäuser. ISBN: 978-0-8176-4434-5. DOI: 10.1007/0-8176-4459-5.

Aoki, Masanao (1990). *State Space Modeling of Time Series*. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-540-52870-8. DOI: 10.1007/978-3-642-75883-6.

Banerjee, Anindya et al. (1993). *Co-integration, Error Correction, and the Econometric Analysis of Non-Stationary Data*. Advanced Texts in Econometrics. Oxford: Oxford University Press.

Bartlett, Maurice Stevenson (1946). "On the theoretical specification of sampling properties of auto-correlated time series". In: *Journal of the Royal Statistical Society*. B 8, pp. 27–41.

Beveridge, Stephen and Charles Nelson (1981). "A New Approach to the Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to the Measurement of Business Cycle". In: *Journal of Monetary Economics* 2, pp. 151–174.

Billingsley, Patrick (1999). *Convergence of Probability Measures*. 2nd. New York: John Wiley & Sons.

Brockwell, Peter J. and Richard A. Davis (2016). *Introduction to Time Series and Forecasting*. 3rd. New York: Springer.

Casals, J. et al. (2016). *State-Space Methods for Time Series Analysis: Theory, Applications and Software*. Boca Raton, Florida: Chapman and Hall/CRC.

Chatfield, Chris (2000). *Time-Series Forecasting*. London: Chapman & Hall.

Davidson, James (2013). "Cointegrazione e correzione degli errori". In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Ed. by Nigar Hashim -zade and Michael A. Thornton. Edward Elgar Publishing. Chap. 7.

De Biasio, Francesco, Giorgio Baldin, and S. Vignudelli (2020). "Revisiting Vertical Land Motion and Sea Level Trends in the Northeastern Adriatic Sea Using Satellite Altimetry and Tide Gauge Data". In: *Journal of Marine Science and Engineering* 8.11, 949. DOI: 10.3390/jmse8110949. URL: https://doi.org/10.3390/jmse8110949.

Dhrymes, Phoebus J. (1974). *Econometrics: Statistical Foundations and Applications*. Second printing. New York: Springer-Verlag.

Dickey, David A. and Wayne A. Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root". In: *Econometrica* 49, pp. 1057–1072.

Engle, Robert F. and Clive W. J. Granger (1987). "Co-integration and Error Correction: Representation, Estimation and Testing". In: *Econometrica* 55, pp. 251–276.

Fama, Eugene F. (Sept. 1965). "Random Walks in Stock Market Prices". In: *Financial Analysts Journal* 21.5, pp. 55–59.

Feller, William (1968). *An Introduction to Probability Theory and Its Applications*. 3rd. New York: John Wiley & Sons.

Fuller, Wayne A. (1995). *Introduction to Statistical Time Series*. New York: Wiley & Sons.

Gnedenko, Boris V. (1969). *The Theory of Probability*. Moscow: Mir Publishers.

Gooijer, Jan De et al. (1985). "Methods for Determining the Order of an Autoregressive-Moving Average Process: A Survey". In: *International Statistical Review/Revue Internationale De Statistique* 53.3, pp. 301–329.

Gourieroux, Christian and Alain Monfort (1995). *Statistics and Econometric Models*. Vol. 1. Cambridge, UK: Cambridge University Press.

Graham, Ronald L., Donald E. Knuth, and Oren Patashnik (1990). *Concrete Mathematics*. Sixth printing, with corrections, October 1990. Reading, MA: Addison-Wesley.

Granger, Clive W. J. (1969). "Prediction with a Generalized Cost of the Error Function". In: *Operational Research Quarterly* 20, pp. 199–207.

— (1981). "Some properties of time series data and their use in econometric model specification". In: *Journal of Econometrics* 16.1, pp. 121–130.

Granger, Clive W. J. and Paul Newbold (1974). "Spurious regressions in econometrics". In: *Journal of Econometrics* 2.2, pp. 111–120.

— (1977). *Forecasting Economic Time Series*. New York: Academic Press.

Hamilton, James D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Harvey, Andrew C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

— (1993). *Time Series*. 2nd. Cambridge, Massachusetts: The MIT Press.

Harvey, David I., Stephen J. Leybourne, and A. M. Robert Taylor (2009). "Unit root testing in practice: dealing with uncertainty over the trend and initial condition". In: *Econometric Theory* 25.3, pp. 587–636. DOI: 10.1017/S0266466608090353.

Hyndman, Rob J. and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd. Melbourne, Australia: OTexts. URL: https://otexts.com/fpp3/.

Jazwinski, Andrew H. (1970). *Stochastic Processes and Filtering Theory*. Appendix 7B. Cambridge: Academic Press.

Johansen, Søren (1991). "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models". In: *Econometrica* 59, pp. 1551–1580.

— (1995). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Kailath, Thomas (1980). *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall. ISBN: 978-0135369616.

Karlin, Samuel and Howard M. Taylor (1975). *A First Course in Stochastic Processes*. 2nd. London: Academic Press.

Kemp, Gordon C.R. (1997). "Linear Combinations of Stationary Processes—Solution". In: *Econometric Theory* 13.6, pp. 897–898. DOI: 10.1017/S026646660000640X.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability*. Trans. by N. Morrison. 2nd English ed. With an added bibliography by A. T. Bharucha-Reid. New York: Chelsea Publishing Co.

Lee, Yuk-wing (1960). *Statistical Theory of Communication*. New York: John Wiley & Sons.

Lo, Andrew W. and A. Craig MacKinlay (2002). *A Non-Random Walk Down Wall Street*. The first edition (not Paperback) is from 1999. New Jersey: Princeton University Press.

Malkiel, Burton G. (2003). *A Random Walk Down Wall Street*. New York: W. W. Norton & Company.

Mardia, Kantilal V., John T. Kent, and John M. Bibby (1979). *Multivariate analysis*. London: Academic Press.

Michel, Anthony N. and Ling Hou (2008). *Stability of Dynamical Systems: Continuous, Discontinuous, and Discrete Systems*. Boston, MA: Birkhäuser. ISBN: 978-0-8176-4649-3. DOI: 10.1007/978-0-8176-4649-3.

Mikosch, Thomas (1998). *Elementary Stochastic Calculus - with Finance in View*. New Jersey: World Scientific Publishing Co. Pte. Ltd.

Newey, Whitney K. and Kenneth D. West (1987). "A Simple, Positive Semi-definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix". In: *Econometrica* 55.3, pp. 703–708.

Pearson, Karl (1905). "The Problem of the Random Walk". In: *Nature* 72, p. 294.

Phillips, Peter C. B. (1986). "Understanding Spurious Regressions in Econometrics". In: *Journal of Econometrics* 33, pp. 311–340.

Proietti, Tommaso (2006). "Exact Beveridge–Nelson decomposition for I(1) and I(2) processes". In: *Journal of Economic Dynamics and Control* 30.12, pp. 2227–2240.

Sartore, Domenico (1975). "Analisi statistica comparata degli andamenti del livello marino medio mensile a Venezia e a Porto Corsini (1947-1971)". In: *Rendiconti del Comitato Veneto per il Potenziamento degli Studi Economici e per la Programmazione* X, pp. 120–207. ISSN: 1591-9811.

Schilling, Rudolf L. and Lutz Partzsch (2012). *Brownian Motion: An Introduction to Stochastic Processes*. Berlin/Boston: Gruyter GmbH & Co. KG.

Stock, James H. (1987). "Asymptotic Properties of Least-Squares Estimators of Cointegrating Vectors". In: *Econometrica* 55, pp. 1035–1056.

Stock, James H. and Mark W. Watson (1988). "Testing for common trends". In: *Journal of the American Statistical Association* 83, pp. 1097–1107.

Stralkowski, C. M., S. M. Wu, and R. E. DeVor (1970). "Charts for the Interpretation and Estimation of the Second Order Autoregressive Model". In: *Technometrics* 12.3, pp. 669–685. DOI: 10.2307/1267211.

Tomasin, Alberto (2005). *Il software Polifemo per l'analisi delle maree*. Nota Tecnica 202. Venezia: CNR-ISMAR.

Vahid, Farshid and Robert F. Engle (1993). "Common Trends and Common Cycles". In: *Journal of Applied Econometrics* 8.4, pp. 341–360.

Verbeek, Marno (2017). *A Guide to Modern Econometrics*. 5th. Chichester: John Wiley & Sons.

Wold, Herman (1953). *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist & Wiksell.

Yule, G. Udny (1926). "Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series". In: *Journal of the Royal Statistical Society* 89.1, pp. 1–63.

Index

This digital version is archived by Zenodo (CERN) as part of its Open Access repository.
Deposit date: **February 2026**
DOI: **https://doi.org/10.5281/zenodo.17572969**
ISBN: **979-12-243-1114-0**

For comments, suggestions, or notices of any typographical errors, please contact the author at: sartore@unive.it