Stationary processes

Matteo Pelagatti

October 24, 2013

1 Definitions

Let us define the main object of this book: the time series.

Definition 1 (Time Series). A time series is a sequence of observations ordered with respect to a time index \( t \), taking values in an index set \( S \). If the set \( S \) contains a finite or countable number of elements we speak of discrete-time time series and the generic observation is indicated with the symbol \( y_t \), while if \( S \) is a continuum we have a continuous-time time series, whose generic observation is represented as \( y(t) \).

Even though continuous-time time series are becoming very rare in a world dominated by digital computers\(^1\), continuous-time models are nevertheless very popular in many disciplines. Indeed, observations may be taken at approximately equispaced points in time, and in this case the discrete-time framework is the most natural, or observations may be non-equispaced and in this case continuous-time models are usually more appropriate. This book concentrates on discrete-time time series, but most of the models covered here have a continuous-time counterpart.

Since future values of real time series are generally unknown and cannot be predicted without error a quite natural mathematical model to describe their behaviour is that of a stochastic process.

\(^1\)Digital computers are finite-state machines and, thus, cannot record continuous-time time series.
Definition 2 (Stochastic Process). A stochastic process is a sequence of random variables defined on a probability space \((\Omega, \mathcal{F}, P)\) and ordered with respect to a time index \(t\), taking values in an index set \(S\).

Again, when \(S\) is numerable one speaks of discrete-time processes and denotes it as \(\{Y_t\}_{t \in S}\), when \(S\) is a continuum we have a continuous-time process and represent it as \(\{Y(t)\}_{t \in S}\) (sometimes also \(\{Y_t\}_{t \in S}\)).

By definition of random variable, for each fixed \(t\), \(Y_t\) is a function \(Y_t(\cdot)\) on the sample space \(\Omega\), while for each fixed simple event \(\omega \in \Omega\), \(Y(\omega)\) is a function on \(S\), or a realization (also sample-path) of a stochastic process.

As customary in modern time series analysis, in this book we consider a time series as a finite realisation (or sample-path) of a stochastic process.

There is a fundamental difference between classical statistical inference and time series analysis. The set-up of classical inference consists of a random variable or vector \(X\) and a random selection scheme to extract simple events, say \(\{\omega_1, \omega_2, \ldots, \omega_n\}\), from the sample space \(\Omega\). The observations, then, consist of the random variable values corresponding to the selected simple events: \(\{x_1, x_2, \ldots, x_n\}\), where \(x_i := X(\omega_i)\) for \(i = 1, 2, \ldots, n\). In time series analysis, instead, we have a stochastic process \(\{Y_t\}_{t \in S}\) and observe only one finite realisation of it through the extraction of a single event, say \(\omega_1\), from the sample space \(\Omega\): we have the time series \(\{y_1, y_2, \ldots, y_n\}\), with \(y_t := Y_t(\omega_1)\) for \(t = 1, \ldots, n\). Therefore, while in classical inference we have a sample of \(n\) observations for the random variable \(X\), in time series analysis we usually have to deal with a sample of dimension 1 with \(n\) observations coming from different time points of the process \(\{Y_t\}\). This means that, if we cannot assume some kind of time-homogeneity of the process making the single sample-path “look like” a classical sample, then we cannot make any sensible inference and prediction from a time series. In Section 2 we introduce the classes of stationary and integrated processes which are the most important time-homogeneous processes used in time series analysis.

We end up this section defining an important class of stochastic processes.

Definition 3 (Gaussian process). The process \(\{Y_t\}_{t \in S}\) is Gaussian if
for all the finite subsets \( \{t_1, t_2, \ldots, t_m\} \) of time points in \( \Omega \) the joint distribution of \( (Y_{t_1}, \ldots, Y_{t_m}) \) is normal.

## 2 Stationary processes

As we saw in Section 1, we treat a time series as a finite sample-path of a stochastic process. Unfortunately, unlike in statistical inference based on repeated random sampling, in time series analysis we have only one observation, the time series, from the data generating process. Thus, we have to base our inference on a sample of dimension one. Generally, we do have more than one observation in the sample-path but, unless we assume some kind of time-homogeneity of the data generating process, every observation \( Y_t \) in the time series is drawn from a different random variable \( Y_t \).

Most social and natural phenomena seem to evolve smoothly rather than by abrupt changes, and therefore modelling them by time-homogeneous processes is a reasonable approximation at least for a limited period of time.

The most important form of time-homogeneity used in time series analysis is stationarity, which is defined as time-invariance of the whole probability distribution of the data generating process (strict stationarity), or just of its first two moments (weak stationarity).

**Definition 4** (Strict stationarity). The process \( \{Y_t\} \) is strictly stationary if for all \( k \in \mathbb{N}, h \in \mathbb{Z}, \) and \( (t_1, t_2, \ldots, t_k) \in \mathbb{Z}^k \),

\[
(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}) \overset{d}{=} (Y_{t_1+h}, Y_{t_2+h}, \ldots, Y_{t_k+h})
\]

where \( \overset{d}{=} \) denotes equality in distribution.

**Definition 5** (Weak stationarity). The process \( \{Y_t\} \) is weakly stationary
(or covariance stationary) if, for all $h, t \in \mathbb{Z}$,

$$
E(Y_t) = \mu, \\
\text{Cov}(Y_t, Y_{t-h}) = \gamma(h),
$$

with $\gamma(0) < \infty$.

As customary in time series analysis, in the rest of the book the terms stationary and stationary will be used with the meaning of weak stationarity and weakly stationary respectively.

**Theorem 1** (Relation between strict and weak stationarity). Let $\{Y_t\}$ be a stochastic process:

1. if $\{Y_t\}$ is strictly stationary, then it is also weakly stationary if and only if $\text{Var}(Y_t) < \infty$;

2. if $\{Y_t\}$ is a Gaussian process, then strict and weak stationarity are equivalent (i.e. one form of stationarity implies the other).

**Proof.** Trivial.

Notice that the above definitions of stationarity assume that the process $\{Y_t\}$ is defined for $t \in \mathbb{Z}$ (i.e. the process originates in the infinite past and ends in the infinite future). This is a mathematical abstraction that is useful to derive some results (e.g. limit theorems), but the definitions can be easily adapted to the case of time series with $t \in \mathbb{N}$ or $t \in \{1, 2, \ldots, n\}$ by changing the domains of $t, h$ and $k$ accordingly.

The most elementary (non-trivial) stationary process is the white noise.

**Definition 6** (White noise). A stochastic process is white noise if it has zero mean, finite variance, $\sigma^2$, and covariance function

$$
\gamma(h) = \begin{cases} 
\sigma^2 & \text{for } h = 0, \\
0 & \text{for } h \neq 0.
\end{cases}
$$
As the next example clarifies, white noise processes and independent identically distributed (i.i.d.) sequences are not equivalent.

**Example 1** (White noise and i.i.d. sequences). Let \( \{X_t\} \) be a sequence of independently identically distributed (i.i.d.) random variables and \( \{Z_t\} \) be white noise.

The process \( \{X_t\} \) is strictly stationary since the joint distribution for any \( k \)-tuple of time points is the product (by independence) of the common marginal distribution (by identical distribution), say \( F(\cdot) \),

\[
Pr\{X_{t_1} \leq x_1, X_{t_2} \leq x_2, \ldots, X_{t_k} \leq x_k \} = \prod_{i=1}^{k} F(x_i),
\]

and this does not depend on \( t \). \( \{X_t\} \) is not necessarily weakly stationary since its first two moments may not exist (e.g. when \( X_t \) is Cauchy-distributed).

The process \( \{Z_t\} \) is weakly stationary since mean and covariance are time-independent, but it is not necessarily strictly stationary since its marginal and joint distributions may depend on \( t \) even when the first two moments are time-invariant.

The function \( \gamma(h) \), which characterise a weakly stationary process, is called autocovariance function and enjoys the following properties.

**Theorem 2** (Properties of the autocovariance function). Let \( \gamma(\cdot) \) the autocovariance function of a stationary process

1. (Positivity of variance) \( \gamma(0) \geq 0 \),
2. (Cauchy-Schwartz inequality) \( |\gamma(h)| \leq \gamma(0) \),
3. (Symmetry) \( \gamma(h) = \gamma(-h) \),
4. (Nonnegative definiteness) \( \sum_{i=1}^{m} \sum_{j=1}^{m} a_i \gamma(i - j) a_j \geq 0 \), \( \forall m \in \mathbb{N} \) and \( (a_1, \ldots, a_m) \in \mathbb{R}^m \).
For the proof of this theorem and in many other places in this book, we will make use of the covariance matrix of the vector of $n$ consecutive observations of a stationary process, say $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_n)\top$:

$$
\mathbf{\Gamma}_n :=
\begin{bmatrix}
\gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\
\gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\
\vdots & \vdots & \ddots & \vdots \\
\gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0)
\end{bmatrix}.
$$

As any covariance matrix, $\mathbf{\Gamma}_n$ is symmetric with respect to the main diagonal and nonnegative definite but, as the reader can easily verify, $\mathbf{\Gamma}_n$ is also symmetric with respect to the secondary diagonal. Furthermore, the element of the matrix with indexes $(i, j)$ equals the element with indexes $(i + 1, j + 1)$ (i.e. $\mathbf{\Gamma}_n$ is a Toeplitz matrix).

**Proof.** The first two properties are well-known properties of variance and covariance. The third property follows from stationarity and the symmetry of the arguments of the covariance:

$$
\gamma(h) = \text{Cov}(X_t, X_{t-h}) = \text{Cov}(X_{t+h}, X_t) = \text{Cov}(X_t, X_{t+h}) = \gamma(-h).
$$

As for the fourth property, let $\mathbf{y} := (Y_1, Y_2, \ldots, Y_m)\top$ be $m$ consecutive observations of the stationary process with autocovariance $\gamma(\cdot)$, then for any real $m$-vector of constants $\mathbf{a}$, the random variable $\mathbf{a}\top\mathbf{y}$ has variance

$$
\text{Var}(\mathbf{a}\top\mathbf{y}) = \mathbf{a}\top\mathbf{\Gamma}_m\mathbf{a} = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \gamma(i-j),
$$

which, being a variance, is nonnegative.

A stronger result asserts that any function on $\mathbb{Z}$ that satisfies the properties of Theorem 2 is the autocovariance of a stationary process, since it is always possible to build a Gaussian process with joint distributions based on such an autocovariance function.

The *autocorrelation function* (ACF) is the scale-independent version of the autocovariance function.
Definition 7 (Autocorrelation function (ACF)). If \( \{Y_t\} \) is a stationary process with autocovariance \( \gamma(\cdot) \), then its ACF is

\[
\rho(h) := \frac{\text{Cor}(Y_t, Y_{t-h})}{\gamma(0)} = \frac{\gamma(h)}{\gamma(0)}.
\]

By Theorem 2 the ACF satisfies the following properties:

1. \( \rho(0) = 1 \),
2. \( |\rho(h)| \leq 1 \),
3. \( \rho(h) = \rho(-h) \),
4. \( \sum_{i=1}^{m} \sum_{j=1}^{m} a_i \rho(i - j) a_j \geq 0, \forall m \in \mathbb{N} \) and \( (a_1, \ldots, a_m) \in \mathbb{R}^m \).

Another summary of the linear dependence of a stationary process can be obtained from the partial autocorrelation function (PACF). The PACF measures the correlation between \( Y_t \) and \( Y_{t-k} \) after their linear dependence on the intervening random variables \( Y_{t-1}, \ldots, Y_{t-h+1} \) has been removed.

Definition 8 (Partial autocorrelation function (PACF)). The partial autocorrelation function of the stationary process \( \{Y_t\} \) is the set of correlations

\[
\alpha(h) := \text{Cor}[Y_t - \mathbb{E}(Y_t|Y_{t-1:t-h+1}), Y_{t-h} - \mathbb{E}(Y_{t-h}|Y_{t-1:t-h+1})]
\]

as function of the nonnegative integer \( h \), where \( Y_{t-1:t-h+1} := (Y_{t-1}, \ldots, Y_{t-h+1})^\top \).

As from the following theorem, the PACF can be derived as linear transformation of the ACF.
Theorem 3 (Durbin-Levinson algorithm). Let \( \{Y_t\} \) be a stationary process with mean \( \mu \) and autocovariance function \( \gamma(h) \), then its PACF is given by

\[
\alpha(0) = 1, \\
\alpha(1) = \gamma(1)/\gamma(0), \\
\alpha(h) = \frac{\gamma(h) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j)}{\gamma(0) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j)}, \quad \text{for } h = 2, 3, \ldots
\]

where \( \phi_{h-1,j} \) denotes the \( j \)-th element of the vector \( \phi_{h-1} := \gamma_h^\top \Gamma_h^{-1} \) with \( \gamma_h^\top := [\gamma(1), \ldots, \gamma(h-1)] \) and \( \Gamma_h \) as in equation (1).

The coefficients \( \phi_h \) can be recursively computed as

\[
\phi_{h,h} = \alpha(h), \quad \phi_{h,j} = \phi_{h-1,j} - \alpha(h)\phi_{h-1,h-j}, \quad \text{for } j = 1, \ldots, h - 1.
\]

Furthermore, if we call \( v_{h-1} \) the denominator of \( \alpha(h) \) in equation (2), we can use the recursion \( v_0 = \gamma(0), v_h = v_{h-1}(1 - \alpha(h)^2) \) to compute it.

The first part of the theorem shows how partial autocorrelations relate to autocovariances, while the second part provide recursions to efficiently compute the PACF without the need to explicitly invert the matrices \( \Gamma_h \).

Proof. The correlation of a random variable with itself is 1, and so \( \alpha(0) = 1 \) and as no variables intervene between \( Y_t \) and \( Y_{t-1} \), \( \alpha(1) = \rho(1) = \gamma(1)/\gamma(0) \).

In order to lighten the notation, let us assume without loss of generality that \( EY_t = 0 \). First, notice that by the properties of the optimal linear predictor (1. and 2. of Theorem 2), \( E(Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}])Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}] = E(Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}])Y_{t-h} \). Thus, by definition of PACF

\[
\alpha(h) = \frac{E(Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}])Y_{t-h}}{E(Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}])Y_t} = \frac{\gamma(h) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j)}{\gamma(0) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j)},
\]

since \( \mathbb{P}[Y_t|Y_{t-1:t-h+1}] = \sum_{j=1}^{h-1} \phi_{h-1,j} Y_{t-j} \) with \( \phi_{h-1,j} \) \( j \)-th element of the vector \( \phi_{h-1}^\top = [\gamma(1), \ldots, \gamma(h)]\Gamma_h^{-1} \).
Let us concentrate on the numerator of equation (3) but for \( \alpha(h+1) \). Using the updating formula for the optimal linear predictor (Theorem ??, Property 7.) we can write the numerator of \( \alpha(h+1) \) as

\[
E(Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}]) + \mathbb{P}[Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}]|Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}])Y_{t-h} \\
= \gamma(h+1) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h+1-j) - \alpha(h) \left( \gamma(1) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h+1-j) \right),
\]

since, as it can be easily checked,

\[
\mathbb{P}[Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}]|Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}]] = \\
\frac{\mathbb{E}(Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}])^2}{\mathbb{E}(Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}])} = \\
\alpha(h)(Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}]) = \\
\alpha(h) \left( Y_{t-h} - \sum_{j=1}^{h-1} \phi_{h-1,j} Y_{t-j} \right).
\]

But, by equation (3) we have the alternative formula for the numerator of \( \alpha(h+1) \),

\[
\gamma(h+1) - \sum_{j=1}^{h-1} \phi_{h,j} \gamma(h+1-j),
\]

and equating the coefficients with the same order of autocovariance we obtain

\[
\phi_{h,h} = \alpha(h), \quad \phi_{h,j} = \phi_{h-1,j} - \alpha(h)\phi_{h-1,h-j} \quad \text{for } j = 1, \ldots, h-1.
\]

Let us denote with \( v_{h-1} \) the denominator of \( \alpha(h) \) in equation (3) and repeat the reasoning for the denominator of \( \alpha(h+1) \), which will be named
\[ v_h = \mathbb{E} \left( Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}] \right) + \]
\[ - \mathbb{P}[Y_t - \mathbb{P}[Y_t|Y_{t-1:t-h+1}]|Y_{t-h} - \mathbb{P}[Y_{t-h}|Y_{t-1:t-h+1}]] Y_t \]
\[ = v_{h-1} - \alpha(h) \left( \gamma(h) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j) \right) \]
\[ = v_{h-1} - \alpha(h)^2 \left( \gamma(0) - \sum_{j=1}^{h-1} \phi_{h-1,j} \gamma(h-j) \right) \]
\[ = v_{h-1} (1 - \alpha(h)^2). \]

Since the population mean \( \mu \), the autocovariances \( \gamma(h) \), the autocorrelations \( \rho(h) \) and the partial autocorrelations \( \alpha(h) \) are generally unknown quantities, they need to be estimated from a time series. If we do not have a specific parametric model for our time series, the natural estimators for \( \mu \) and \( \gamma(h) \) are their sample counterparts:

\[ \overline{Y_n} := \frac{1}{n} \sum_{t=1}^{n} Y_t, \]
\[ \hat{\gamma}(h) := \frac{1}{n} \sum_{t=h+1}^{n} (Y_t - \overline{Y_n})(Y_{t-h} - \overline{Y_n}). \]

Note that in the sample autocovariance the divisor is \( n \) and not \( n - h \) (or \( n - h - 1 \)) as one would expect from classical statistical inference, indeed, the latter divisor does not guarantee that the sample autocovariance function is nonnegative definite. Instead, the sample autocovariance matrix, whose generic element is the above defined \( \hat{\gamma}(i-j) \), can be expressed as the product of a matrix times its transpose and, therefore, is always nonnegative definite.
For example, define the matrix with $k$ columns,

$$
\mathbf{Y} := \begin{bmatrix}
Y_1 - \bar{Y}_n & 0 & 0 & \ldots & 0 \\
Y_2 - \bar{Y}_n & Y_1 - \bar{Y}_n & 0 & \ldots & 0 \\
Y_3 - \bar{Y}_n & Y_2 - \bar{Y}_n & Y_1 - \bar{Y}_n & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Y_n - \bar{Y}_n & Y_{n-1} - \bar{Y}_n & Y_{n-2} - \bar{Y}_n & \ldots & Y_1 - \bar{Y}_n \\
0 & Y_n - \bar{Y}_n & Y_{n-1} - \bar{Y}_n & \ldots & Y_2 - \bar{Y}_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 0 & Y_n - \bar{Y}_n
\end{bmatrix}.
$$

The autocovariance matrix containing the first $k-1$ sample autocovariances can be computed as

$$
\hat{\Gamma}_{k-1} = n^{-1} \mathbf{Y}^\top \mathbf{Y},
$$

which is always nonnegative definite.

We summarise the properties of the sample mean of a stationary process in the following theorem.

**Theorem 4** (Properties of the sample mean). Let $\{Y_t\}$ be a weakly stationary process with mean $\mu$ and autocovariance function $\gamma(h)$, then for the sample mean $\bar{Y}_n$ the following properties hold:

1. (Unbiasedness) $E\bar{Y}_n = \mu$;

2. (Variance) $\text{Var}(\bar{Y}_n) = \frac{1}{n} \sum_{h=-n+1}^{n-1} \left(1 - \left|\frac{h}{n}\right|\right) \gamma(h)$;

3. (Normality) if $\{Y_t\}$ is a Gaussian process, then $\bar{Y}_n \sim \mathcal{N}(\mu, \text{Var}(\bar{Y}_n))$;

4. (Consistency) if $\gamma(h) \to 0$ as $h \to \infty$, then $E[\bar{Y}_n - \mu]^2 \to 0$;

5. (Asymptotic variance) if $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, then $nE[\bar{Y}_n - \mu]^2 \to \sum_{h=-\infty}^{\infty} \gamma(h)$;

6. (Asymptotic normality) If $Y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ with $Z_t \sim \text{IID}(0, \sigma^2)$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$, then

$$
\sqrt{n}(\bar{Y}_n - \mu) \overset{d}{\to} \mathcal{N}
\left(0, \sum_{h=-\infty}^{\infty} \gamma(h) \right).
$$
Proof.

Unbiasedness. \( \mathbb{E} \bar{Y}_n = n^{-1} \sum_{t=1}^{n} \mathbb{E} Y_t = \mu \).

Variance.

\[
\text{Var}( \bar{Y}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(Y_i, Y_j) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma(i - j)
\]

\[
= \frac{1}{n^2} \sum_{h=-n+1}^{n-1} (n - |h|) \gamma(h) = \frac{1}{n} \sum_{h=-n+1}^{n-1} \left( 1 - \frac{|h|}{n} \right) \gamma(h).
\]

Normality. The normality of the mean follows from the assumption of joint Gaussianity of \((Y_1, \ldots, Y_n)\).

Consistency. In order to obtain the (mean-square) consistency of \(\bar{Y}_n\), the quantity \(\mathbb{E}(\bar{Y}_n - \mu) = \text{Var}(\bar{Y}_n)\) has to converge to zero. A sufficient condition for this to happen is \(\gamma(h) \to 0\) as \(h\) diverges. In fact, in this case we can always fix a small positive \(\varepsilon\) and find a positive integer \(N\) such that for all \(h > N\), \(|\gamma(h)| < \varepsilon\). Therefore, for \(n > N + 1\)

\[
\text{Var}(\bar{Y}_n) = \frac{1}{n} \sum_{h=-n+1}^{n-1} \left( 1 - \frac{|h|}{n} \right) \gamma(h) \leq \frac{1}{n} \sum_{h=-n+1}^{n-1} |\gamma(h)|
\]

\[
= \frac{1}{n} \sum_{h=-N}^{N} |\gamma(h)| + 2 \frac{1}{n} \sum_{h=N+1}^{n-1} |\gamma(h)| \leq \frac{1}{n} \sum_{h=-N}^{N} |\gamma(h)| + 2\varepsilon.
\]

As \(n\) diverges the first addend converges to zero (it is a finite quantity divided by \(n\)), while the addend can be made arbitrarily small, and so, by the very definition of limit, \(\text{Var}(\bar{Y}_n) \to 0\).

Asymptotic variance. After multiplying the variance of \(\bar{Y}_n\) times \(n\), we have the following inequalities:

\[
n \mathbb{E}|\bar{Y}_n - \mu|^2 = \sum_{h=-n+1}^{n-1} \left( 1 - \frac{|h|}{n} \right) \gamma(h) \leq \sum_{h=-n+1}^{n-1} |\gamma(h)|.
\]
Therefore, a sufficient condition for the asymptotic variance to converge as $n \to \infty$ is that $\sum_{h=-n+1}^{n-1} |\gamma(h)|$ converges. Furthermore, by Cesàro theorem

$$\lim_{n \to \infty} \sum_{n=-\infty}^{n-1} \left( 1 - \frac{|h|}{n} \right) \gamma(h) = \lim_{n \to \infty} \sum_{h=-\infty}^{n-1} \gamma(h).$$

**Asymptotic normality.** This is the central limit theorem for linear processes, for the proof refer to Brockwell and Davis (1991, Sec. 7.3), for instance \(\square\)

Of course, if \(\{Y_t\}\) is a Gaussian process, also the sample mean is Gaussian. If the process is not Gaussian, the distribution of the sample mean can be approximated by a normal only if some central limit theorem (CLT) for dependent processes applies; in the Theorem we provide one for linear processes, but there are alternative CLT under weaker conditions (in particular under mixing conditions).

For the sample autocorrelation $\hat{\rho}(h) := \hat{\gamma}(h)/\hat{\gamma}(0)$ we provide the following result without proof, which is rather lengthy and cumbersome and can be found in Brockwell and Davis (2002, Sec. 7.3).

**Theorem 5 (Asymptotic distribution of the sample ACF).** Let \(\{Y_t\}\) be the stationary process,

$$Y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad \text{IID}(0, \sigma^2).$$

If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and either one of the following conditions hold

- $\mathbb{E} Z_t^4 < \infty$,
- $\sum_{j=-\infty}^{\infty} \psi_j^2 |j| < \infty$;

then for each $h \in \{1, 2, \ldots\}$

$$\sqrt{n} \begin{bmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(h) \end{bmatrix} - \begin{bmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{bmatrix} \overset{d}{\to} \mathcal{N}(0, V),$$

with the generic \((i, j)\)-th element of the covariance matrix $V$ being $v_{ij} = \sum_{k=1}^{\infty} [\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)][\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)]$. 

13
A corollary to this theorem is that, when a process is IID(0, σ²) the sample autocorrelations at different lags are asymptotically independent and √n̂ρ(h) converges in distribution to a standard normal. This result is used in the following portmanteau test statistics for the null hypothesis $Y_t \sim$ IID(0, σ²):

**Box-Pierce** $Q_{BP}(h) = n \sum_{k=1}^{h} \hat{\rho}(k)^2$;

**Ljung-Box** $Q_{LB}(h) = n(n + 2) \sum_{k=1}^{h} \hat{\rho}(k)^2/(n - k)$.

**Corollary 6** (Portmanteau tests). Under the hypothesis $Y_t \sim$ IID(0, σ²), the test statistics $Q_{BP}$ and $Q_{LB}$ converge in distribution to a chi-square with $h$ degrees of freedom.

Proof. Since $Q_{BP}(h)$ is the sum of $h$ asymptotically independent standard normal random variables, it converges in distribution to a chi-square with $h$ degrees of freedom. The statistic $Q_{LB}(h)$ is asymptotically equivalent to $Q_{BP}(h)$, in fact, for each $h$:

$$\sqrt{\frac{n(n + 2)}{n - h}} \hat{\rho}(h) - \sqrt{n} \hat{\rho}(h) = \sqrt{n} \hat{\rho}(h) \left( \sqrt{\frac{n + 2}{n - h}} - 1 \right)$$

which converges in probability to zero as $n$ diverges. 

The Ljung-Box statistic is more popular then the Box-Pierce since it approximates the asymptotic distribution better in small samples.

We conclude this section on stationary processes with a celebrated result used also as a justification for the use of the class of ARMA models as approximation to any stationary processes.

Before presenting the result we need the concept of (linearly) deterministic processes.

**Definition 9** (Deterministic process). The stationary process $\{V_t\}$ is (linearly) deterministic if

$$\lim_{k \to \infty} \mathbb{E}\left( V_t - \mathbb{E}[V_t|V_{t-1}, V_{t-2}, \ldots, V_{t-k}] \right)^2 = 0.$$
In other words, a stationary process is deterministic if it can be predicted without error by a linear function of its (possibly) infinite past.

**Example 2** (Two deterministic processes). Let

\[ W_t = W, \quad V_t = X \cos(\lambda t) + Y \sin(\lambda t), \]

where \(W, X\) and \(Y\) are random variables with zero means and finite variances; furthermore, \(\text{Var}(X) = \text{Var}(Y) = \sigma^2\) and \(\text{Cov}(X, Y) = 0\). The two processes are stationary, in fact their mean is zero for all \(t\) and their autocovariance functions are

\[
\begin{align*}
\mathbb{E}W_t W_{t-h} &= \mathbb{E}W^2, \\
\mathbb{E}V_t V_{t-h} &= \sigma^2 \left[ \cos(\lambda t) \cos(\lambda(t-h)) + \sin(\lambda t) \sin(\lambda(t-h)) \right], \\
&= \sigma^2 \cos(\lambda h),
\end{align*}
\]

which are invariant with respect to \(t\).

Their linear predictions based on the past are

\[
\begin{align*}
\mathbb{P}[W_t | W_{t-1}] &= W, \\
\mathbb{P}[V_t | V_{t-1}, V_{t-2}] &= 2 \cos(\lambda)V_{t-1} - V_{t-2},
\end{align*}
\]

(the reader is invited to derive the latter formula by computing the optimal linear prediction and applying trigonometric identities).

For the first process it is evident that the prediction \((W)\) is identical to the outcome \((W)\). In the second case, we need to show that the prediction is equal to the process outcome:

\[
\mathbb{P}[V_t | V_{t-1}, V_{t-2}] = \cos(\lambda)V_{t-1} + \cos(2\lambda)V_{t-2}
\]

\[
= 2 \cos(\lambda) [X \cos(\lambda t - \lambda) + Y \sin(\lambda t - \lambda)] - [X \cos(\lambda t - 2\lambda) + Y \sin(\lambda t - 2\lambda)]
\]

\[
= X [2 \cos(\lambda) \cos(\lambda t - \lambda) - \cos(\lambda t - 2\lambda)] + Y [2 \cos(\lambda) \sin(\lambda t - \lambda) - \sin(\lambda t - 2\lambda)]
\]

\[
= X \cos(\lambda t) + Y \sin(\lambda t),
\]

where the last line is obtained by applying well-known trigonometric identities.
Theorem 7 (Wold decomposition). Let $Y_t$ be a stationary process, then

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t$$

where

1. $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,
2. $Z_t$ is white noise,
3. $\text{Cov}(Z_t, V_s) = 0$ for all $t$ and $s$,
4. $V_t$ is (linearly) deterministic,
5. $\lim_{k \to \infty} \mathbb{E}(Z_t - \mathbb{P}[Z_t|Y_t, Y_{t-1}, \ldots, Y_{t-k}])^2 = 0$;
6. $\lim_{k \to \infty} \mathbb{E}(V_t - \mathbb{P}[V_t|Y_s, Y_{s-1}, \ldots, Y_{s-k}])^2 = 0$ for all $t$ and $s$.

The message of this theorem is that every stationary process can be seen as the sum of two orthogonal components: one, the deterministic, is perfectly predictable using a linear function of the past of the process (point 6.), the other, the purely non-deterministic, is expressible as a (possibly) infinite linear combination of past and present observations of a white noise process $Z_t$. This process, is the prediction error of $Y_t$ based on its (possibly) infinite past,

$$Z_t = Y_t - \mathbb{P}[Y_t|Y_{t-1}, Y_{t-2}, \ldots],$$

and is generally termed innovation of the process $Y_t$. Thus, the coefficients $\psi_j$ are the projection coefficients of $Y_t$ on its past innovations $Z_{t-j}$:

$$\psi_j = \mathbb{E}[Y_t Z_{t-j}]/\mathbb{E}[Z_{t-j}^2].$$

As for the deterministic component, point 6. implies that it can be predicted without error also using the infinite past of $\{Y_t\}$ and not only using its own past $\{V_s\}_{s<t}$, as by the definition of deterministic process. In a single time series, the deterministic component cannot be discriminated from a deterministic mean and/or periodic function of time, such as a deterministic
seasonal component, so the main focus of time series modelling is on the purely non-deterministic component. The class of ARMA models, that will be introduced in Section ??, provide a parsimonious approximation for purely non-deterministic stationary processes when the sequence of coefficients $\psi_j$ approaches zero not too slowly.

In the rest of the book, we will consider only stationary processes that are purely non-deterministic. In particular, we will assume that a stationary process has always a representation as causal linear process.

**Definition 10 (Linear process).** The process $\{Y_t\}$ is a linear process if it has the representation

$$Y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad \forall t \in \mathbb{Z},$$

where $Z_t \sim \text{WN}(0, \sigma^2)$, $\mu$ is a constant mean, and $\{\psi_j\}$ is a sequence of constants such that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

If $\psi_j = 0$ for all $j < 0$, then $\{Y_t\}$ is a causal linear process.

The absolute summability condition in the is slightly stronger than the square summability property resulting from the Wold decomposition\(^2\), but, as we have already seen in Theorems 4 and 5, it brings some advantages.

The autocovariance of the linear process is

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h},$$

with the summation starting from $j = 0$ and $h \geq 0$ for the causal case. One advantage of the absolute summability condition is that it implies the abso-
lute summability of the autocovariance function$^3$ that, according to Theorem 4, is sufficient for the consistency of the sample mean and the existence of its asymptotic variance.

References


$^3$To see why (all summation indexes range from $-\infty$ to $\infty$):

$$\sigma^2 \sum_h \left| \sum_j \psi_j \psi_{j+h} \right| \leq \sigma^2 \sum_j |\psi_j| \sum_{h} |\psi_{j+h}| \leq \sigma^2 \sum_j |\psi_j| \sum_i |\psi_i| = \sigma^2 \left( \sum_j |\psi_j| \right)^2.$$