F70TS: Time Series

Contents

1. Introduction, ACF, Stationarity and Operators

A **time series** (TS) is a set of observations of a particular quantity made one after another in time. These data are usually equally-spaced in time (daily, monthly, quarterly, annually,...), but this is not always the case.

The fields of application of modern time-series analysis are numerous, encompassing finance, economics, geography, demography, management, medicine, meteorology etc. See some examples of time series data in Figures 1–4.

The most important thing to note is that these observations are **NOT** just an i.i.d. sample – they will be **related** to one another by a dependence structure. It is this relationship of "serial dependence" that is of interest.

Formally, we have the following definition:

Definition 1.1 *A **Time Series** is a stochastic process defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *That means, for a given index set* $\mathbb{T}$*, a time series* $X = \{X_t\}_{t\in\mathbb{T}}$ *is a collection of random variables defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$.

Remarks:

- In the majority of cases we have $\mathbb{T} = \{0, 1, \ldots, \infty\}$, or $\mathbb{T} = \{0, 1, \ldots, n\}$, or $\mathbb{T} = \{-\infty, \ldots, -1, 0, 1, \ldots, \infty\}$. This means $X = \{X_t\}_{t\in\mathbb{T}}$ is a stochastic process in discrete time, that is, $X = \{X_t\}_{t\in\mathbb{T}}$ is a set of random variables X_t in **time order**.

- The series

$$x_1, x_2, \ldots, x_n = \{x_t\}_{t=1}^n$$

  is called an observation (or observed series) of the time series $X_1, \ldots, X_n = \{X_t\}_{t=1}^n$. The term "time series" is used both for the sequence of random variables $\{X_t\}$ and for the sequence of observations $\{x_t\}$.

- "Time" could be a more general coordinate system. For example, the length along a river, road, pipeline, or power line, or a vector-valued coordinate system (e.g., spatial data analysis). However, in this course we will focus on temporal series analysis.

- We mainly focus on univariate series $X_t \in \mathbb{R}$ and discuss multivariate series at the end of the course.

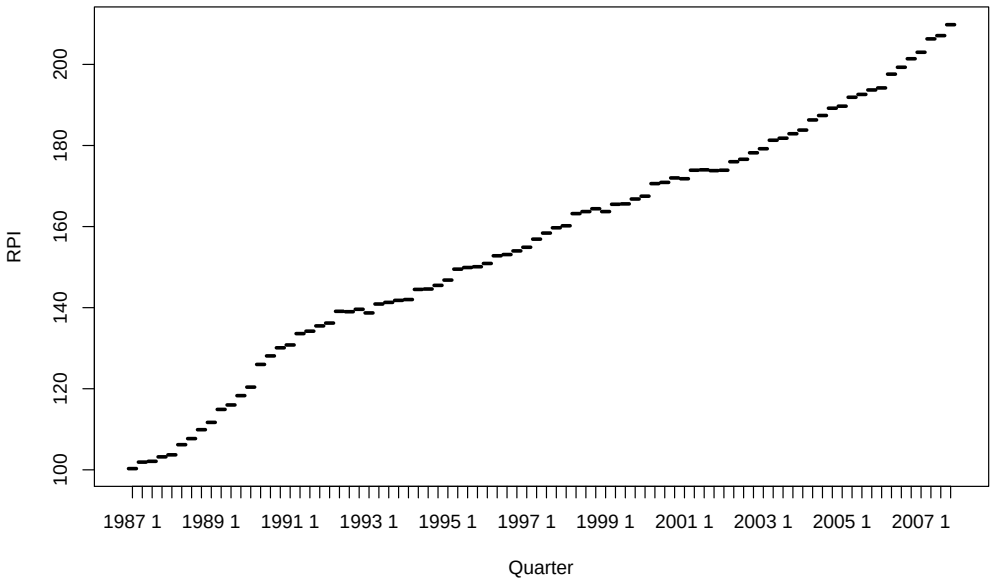


Figure 1: Retail Price Index (UK)

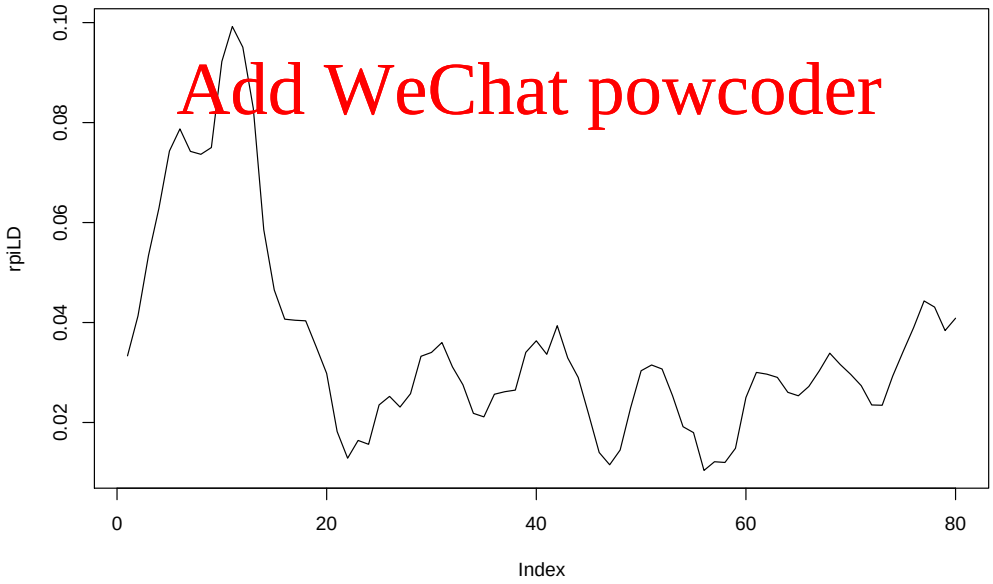


Figure 2: Annual force of inflation (annual log-difference of RPI)

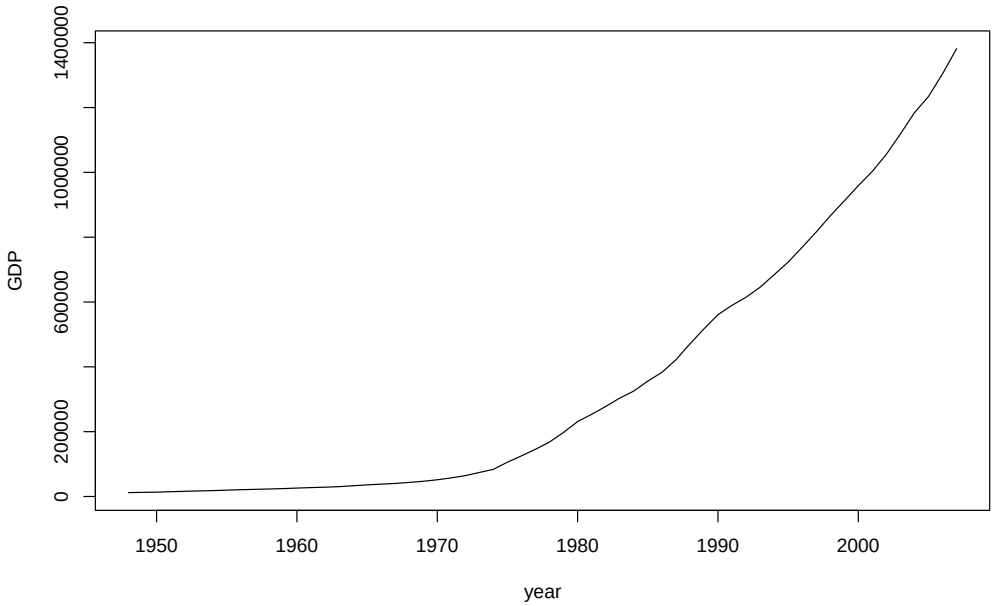


Figure 3: UK annual GDP at current prices (in Million GBP)

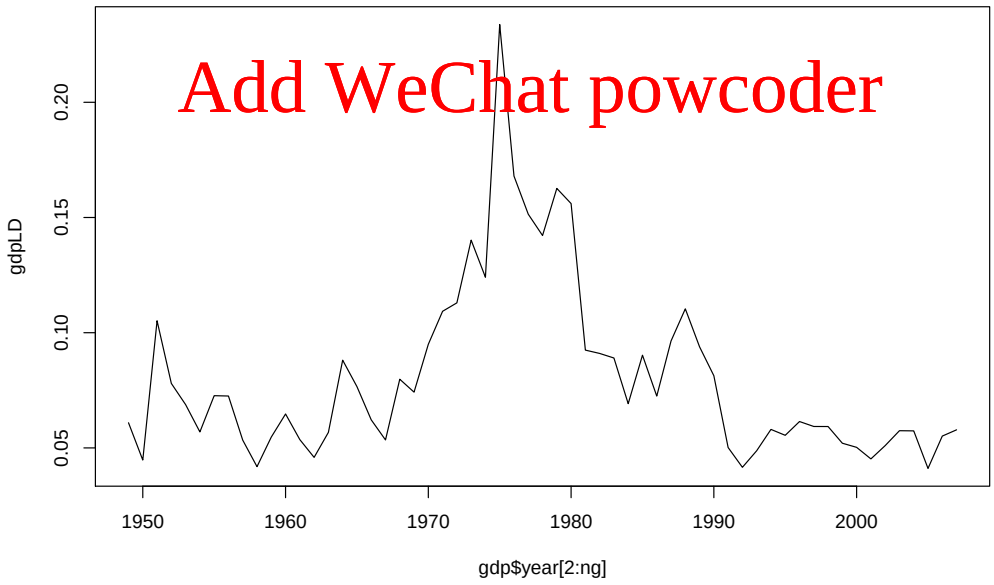


Figure 4: Log-difference of UK annual GDP at current prices

1.1. Objectives of time series analysis

The objectives of time series analysis can be summarised as follows:

1. **Description and analysis**

   a) Plot the series. A data plot shows up the basic properties of the series and may indicate appropriate methods of analysis. It may, for example, suggest that a transformation of the data is appropriate (e.g. if standard deviation increases with mean, a log transformation may stabilise the variance). It may also show up the presence of *outliers*; i.e., points which do not fit the overall data pattern.

   b) Calculation of simple descriptive summary measures, such as mean, variance, and autocorrelation plots (see Autocorrelation Function).

   c) Modelling: a mathematical model may be proposed and fitted. This is useful, for example, for forecasting future values of the series. Time series models usually describe the observed series $\{x_t\}$ by combining three terms:

   - Seasonal model: deterministic periodic patterns (e.g., a restaurant that has more business during the end of the week and the weekend, this would produce a seasonality of 7 days). Some series include several periodic patterns (e.g., increased business during summer months, or during the main sports season, etc.)
   - Trend model: underlying deterministic upwards or downwards trend of the observed quantities (e.g., demand for certain commodities might be growing, or shrinking).
   - Stochastic model: random fluctuations in the time series that are not due to trend or seasonality.

   The seasonality and trend components of a time series are deterministic, and can therefore be estimated and removed from the observed series to isolate the stochastic component of the series. Analysing this stochastic component is significantly more challenging.

2. **Forecasting:** Given an observed series we predict future values, with uncertainty quantification on the delivered prediction.

3. **Control:** In a controlled physical process, analysing an appropriate time series allows adjusting the input variables in order to stabilise the process. (e.g., the temperature control system controlling an oven, the automatic pilot system in an autonomous vehicle).

4. **Explanation:** In a multivariate context, we may use the variation in one time series to "explain" the behaviour of another time series (e.g., the relationship between the price of petrol and the price of oil, or between demand for certain products and the weather).

1.2. Autocorrelation Function and Stationarity

Our main tool for understanding the dependence structure within time series data is the autocovariance function, defined as follows:

Definition 1.2 *The covariance between two elements X_t and X_s ($t, s \in \mathbb{T}$) of a time series X, is called the autocovariance function of X, denoted by*

$$\gamma(t,s) = Cov(X_t, X_s) = E[(X_t - E(X_t))(X_s - E(X_s))].$$

In particular, $\gamma(t,t) = \text{Var}(X_t)$.

Roughly speaking, a time series is said to be *stationary* if its properties do not change over time. There are many different ways to make this idea mathematically precise. One common way is the property of *weak stationarity*, which is defined based on autocovariances:

Definition 1.3 *A time series is said to be **weakly stationary** if*

(i) $E(X_t^2) < \infty$ *for all* $t \in \mathbb{T}$,

(ii) $E(X_t) = \mu$ *for all* $t \in \mathbb{T}$, *and*

(iii) $\gamma(r,s) = \gamma(r+t, s+t)$ *for all* $r, s, t \in \mathbb{T}$.

We note the following properties of a weakly stationary time series:

- the mean and variance are constant.
- the autocovariances $\gamma(r,s) = \text{Cov}(X_r, X_s)$ only depend on the time lag $k = s - r$. It is hence a one-dimensional function

$$\gamma(k) = \text{Cov}(X_t, X_{t+k}) \text{ for } k = \ldots -1, 0, 1, \ldots \text{ and } \forall\, t \in \mathbb{T},$$

  called the autocovariance function of the process X_t.

We may also define a normalized version of the autocovariance function, called the *autocorrelation function*:

Definition 1.4 *The autocorrelation function (ACF) is the standardised autocovariance function of a (weakly) stationary process X_t:*

$$\rho(k) \;=\; \frac{\gamma(k)}{\gamma(0)} = \frac{Cov(X_t, X_{t+k})}{Var(X_t)} = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)Var(X_{t+k})}}.$$

Example: White Noise
This is the simplest stationary series. The elements X_t for all $t \in \mathbb{T}$ are i.i.d. with mean 0 and variance σ^2. So, $\gamma(t,s) = 0$ for $t \neq s$, and $\gamma(t,t) = \sigma^2$. The White Noise process is usually

denoted by ϵ (or $\epsilon(t)$). It has no "structure" but is used as a fundamental building block for models with interesting and complex structure (see later sections).

Properties of the autocovariance function of a (weakly) stationary process (similar properties hold for the ACF):

- $\gamma(0) = \text{Var}(X_t) = \sigma_X^2 \geq 0$ and $\rho(0) \equiv 1$.
- $|\gamma(k)| \leq \gamma(0)$ for all $k \in \mathbb{T}$,
- $\gamma(-k) = \gamma(k)$ for all $k \in \mathbb{T}$.

An alternative definition of stationarity that is sometimes useful is strong stationarity:

Definition 1.5 *A time series is said to be* ***strongly (or strictly) stationary****, if*

$$P(X_{t_1} < x_1, ..., X_{t_k} < x_k) = P(X_{t_1+t} < x_1, ..., X_{t_k+t} < x_k)$$
$$\forall\, k = 1, 2, ...;\ t, t_1, ..., t_k \in \mathbb{T} \textit{ and } x_1, ..., x_k \in \mathbb{R}.$$

Strong stationarity is a property of the whole distribution of a process. Weak stationarity is just defined based on the first two moments of the process.

It is clear that strong stationarity does not follow from weak stationarity. However, note that, in general, weak stationarity also does not follow from strong stationarity either, because the second moments (or even the mean) of a strongly stationary process may not exist.

If the variance of a strongly stationary process exists, then it is also weakly stationary.

Note that if the process is Gaussian, i.e. the distribution of X_t for any $t \in \mathbb{T}$ is normal, then the two stationary properties are equivalent to each other.

Examples:

1. $X_t = \epsilon_t$, $t = 0, 1, ...$ with i.i.d. Cauchy ϵ_t is strongly, but not weakly, stationary (since the variance is not finite here).

2. Let ϵ_t be i.i.d. with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) < \infty$, then the *moving average* process (of order 1) defined by
$$X_t = \psi\epsilon_{t-1} + \epsilon_t,\ |\psi| < 1$$
and the *autoregressive* process (of order 1) defined by
$$X_t = \varphi X_{t-1} + \epsilon_t,\ |\varphi| < 1$$
are both strongly and weakly stationary. See later on for more detail.

3. Let ϵ_t be i.i.d. $N(0, 1)$ random variables. Then the process
$$X_t = \begin{cases} \epsilon_t & \text{for } t = 1, 3, ... \\ (\epsilon_t^2 - 1)/\sqrt{2} & \text{for } t = 2, 4, ..., \end{cases}$$
is weakly, but not strongly, stationary.

4. The Random Walk process $X_t = X_{t-1} + \epsilon_t$ with i.i.d. ϵ_t is neither weakly nor strongly stationary. To see this, check that the variance of X_t depends on t.

1.3. Operators

A time series model is often defined by using mathematical operators and some base stochastic process (such as a white noise process).

Operators allow transforming one time series into another, and can be applied to the random variables X_t or the observations x_t.

We will focus on the following widely used operators:

Backward shift: $BX_t := X_{t-1}$, and $B^r X_t := X_{t-r}$

Difference: $\Delta X_t = X_t - X_{t-1}$

Seasonal difference: $D_r X_t = X_t - X_{t-r} = (1 - B^r)X_t$

Forward Shift: $FX_t = X_{t+1}$, and $F^r X_t = X_{t+r}$

Identity: $1X_t = X_t$

Summation: $SX_t = X_t + X_{t-1} + X_{t-2} + \ldots$

Operators can be handled algebraically and admit inverse operators. For example:

$$\Delta x_t := x_t - x_{t-1} = x_t - Bx_t = (1 - B)x_t,$$

$\implies \Delta \equiv 1 - B.$

Also,

$$\begin{aligned} S(1 - B)x_t = S(x_t - x_{t-1}) &= Sx_t - Sx_{t-1} \\ &= (x_t + x_{t-1} + x_{t-2} + \ldots) \\ &\qquad - (x_{t-1} + x_{t-2} + x_{t-3} + \ldots) \\ &= x_t \end{aligned}$$

$\implies S(1 - B) \equiv 1 \qquad$ i.e. $S \equiv (1 - B)^{-1}$.

$S\Delta \equiv 1 \qquad$ i.e. $S \equiv \Delta^{-1}$.

Alternatively,

$$\begin{aligned} Sx_t &= x_t + x_{t-1} + x_{t-2} + \ldots \\ &= (1 + B + B^2 + \ldots)x_t = (1 - B)^{-1}x_t. \end{aligned}$$

The Moving Average Operator

Moving averages are a useful way of building models for stationary time series data, especially by applying a moving average operator to a white noise process.

Notation:

$$T = [w_{-k}, \ldots, w_{-1}, w_0, w_1, \ldots, w_k]$$

is called a **moving average (MA) operator**. We write $m_t = Tx_t$.

T is defined by:

$$Tx_t = \sum_{i=-k}^{k} w_i x_{t+i} = w_{-k}x_{t-k} + \cdots + w_0 x_t + \cdots + w_k x_{t+k}$$

1.4. Univariate Linear Processes

A very wide class of time series models is that of linear processes, including the well known AR (autoregressive), MA (moving average), ARMA (autoregressive moving average) processes with i.i.d. innovations.

The simplest process is the *White Noise* (WN) model we have already met, defined as follows:

Definition 1.6 *A White Noise process is an i.i.d. series ϵ_t with $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma_\epsilon^2$.*

A WN process is strongly and weakly stationary with

$$\gamma(k) = \begin{cases} \sigma_\epsilon^2, & k = 0, \\ 0, & k \neq 0 \end{cases} \quad \text{and} \quad \rho(k) = \begin{cases} 1, & k = 0, \\ 0, & k \neq 0. \end{cases} \tag{1}$$

Observe that infinite variance i.i.d. series such as i.i.d. Cauchy series, are not included in our definition.

A time series is said to be linear, if it can be represented as a linear (possibly infinite) sum (called a *linear filter*) of ϵ_t, where $\{\epsilon_t\}$ is a WN process. See the MA(∞) process defined later.

Examples of Time Series Models Based on White Noise;

1. First order moving average process - MA(1):

   $$X_t = 0.8\epsilon_{t-1} + \epsilon_t = (1 + 0.8B)\epsilon_t,$$

   ϵ_t are, for example, i.i.d. $N(0, 1)$ random variables.

2. Second order moving average process - MA(2):

   $$X_t = 0.5\epsilon_{t-1} + 0.3\epsilon_{t-2} + \epsilon_t = (1 + 0.5B + 0.3B^2)\epsilon_t,$$

   ϵ_t are as above.
   Using the MA-operator: $X_t = T\epsilon_t$ with $T = [0.3, 0.5, 1, 0, 0]$

3. First order autoregressive process - AR(1):

$$X_t = 0.8X_{t-1} + \epsilon_t \iff (1 - 0.8B)X_t = \epsilon_t,$$

ϵ_t as above.

4. Random Walk:

$$X_t = X_{t-1} + \epsilon_t \iff \Delta X_t = \epsilon_t, \quad t = 0, 1, ...,$$

ϵ_t as above.

$$X_t = S\epsilon_t$$

2. Moving Average Processes

Definition 2.1 *Let ϵ_t be a white noise process. The moving average (MA) process of order q with parameter vector $(\psi_1, \ldots, \psi_q)$ is given by*

$$MA(q): \quad X_t = \psi_1 \epsilon_{t-1} + \ldots + \psi_q \epsilon_{t-q} + \epsilon_t = \sum_{i=0}^{q} \psi_i \epsilon_{t-i}, \tag{2}$$

where $\psi_q \neq 0$ and $\psi_0 = 1$.

Remarks:

- $E(X_t) = 0$ for an MA(q) process
- Using the backshift operator B we have

$$\begin{aligned} X_t &= \psi_0 B^0 \epsilon_t + \psi_1 B \epsilon_t + \ldots + \psi_q B^q \epsilon_t \\ &= \left[\psi_0 B^0 + \psi_1 B + \ldots + \psi_q B^q\right] \epsilon_t = \psi(B)\epsilon_t \end{aligned} \tag{3}$$

where $\psi(B) = \sum_{i=0}^{q} \psi_i B^i$.

Examples:

(i) An MA(1) process: $X_t = 0.8\epsilon_{t-1} + \epsilon_t$.

(ii) Another MA(1) process: $X_t = -0.9\epsilon_{t-1} + \epsilon_t$.

(iii) An MA(2) process: $X_t = 0.7\epsilon_{t-1} + 0.4\epsilon_{t-2} + \epsilon_t$. See Figure 5 for plots of two sets of data simulated from this process.

(iv) Another MA(2) process: $X_t = 0.6\epsilon_{t-1} - 0.3\epsilon_{t-2} + \epsilon_t$.

2.1. Properties of MA Processes

The MA process defined above does not involve ϵ_i with $i > t$, i.e. does not involve information about the future. A process which does not involve future information is called *causal*.

An MA model can also be defined with coefficients $\psi_i \neq 0$ for $i < 0$ (see, e.g., the MA(∞) process given later). In this case the process is not causal.

A finite order MA process is always stationary.

We can obtain the autocovariances $\gamma(k)$ and the autocorrelations $\rho(k)$ of an MA process easily by using the fact $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$, $\text{Cov}(\epsilon_t, \epsilon_{t\pm k}) = 0$ for all t and $k \neq 0$. For example, for the

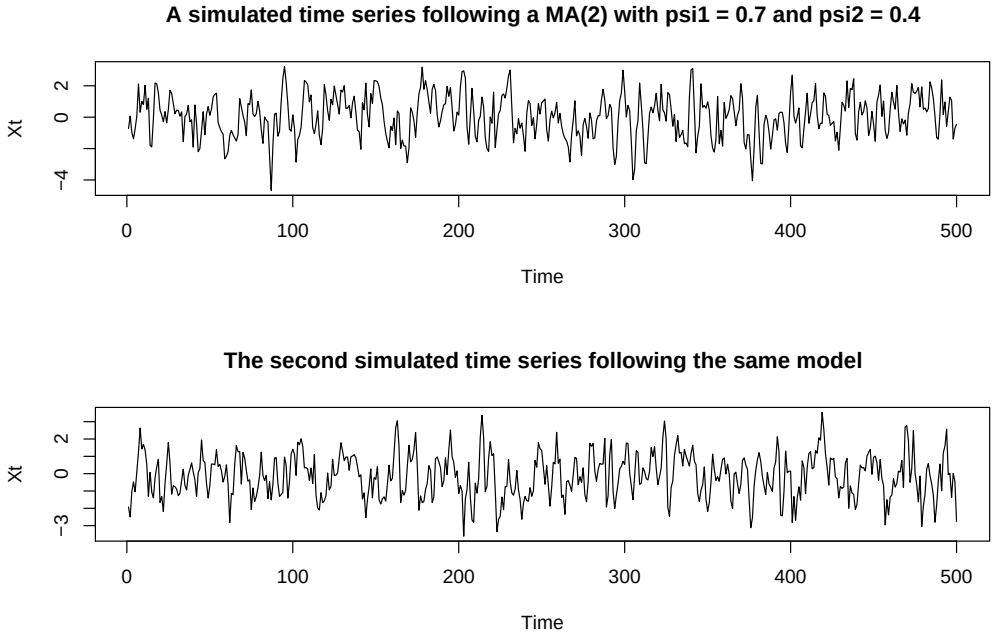


Figure 5: Two simulated realisations following the MA(2) Model (iii) given above.

MA(1) process $X_t = 0.8\epsilon_{t-1} + \epsilon_t$ we have

$$\begin{aligned}
\text{Var}(X_t) &= \text{Var}(0.8\epsilon_{t-1} + \epsilon_t) = 0.8^2\text{Var}(\epsilon_t) + \text{Var}(\epsilon_t) = 1.64\sigma_\epsilon^2, \\
\gamma(\pm 1) &= \text{Cov}(X_t, X_{t+1}) = \text{Cov}(0.8\epsilon_{t-1} + \epsilon_t, 0.8\epsilon_t + \epsilon_{t+1}) = 0.8\sigma_\epsilon^2, \\
\gamma(\pm k) &= 0 \text{ for } |k| > 1, \text{ and hence} \\
\rho(0) &= 1, \\
\rho(\pm 1) &= \frac{0.8\sigma_\epsilon^2}{1.64\sigma_\epsilon^2} = 0.488, \text{ and} \\
\rho(\pm k) &= 0 \text{ for } |k| > 1.
\end{aligned}$$

Let $\gamma_\epsilon(k) = \text{Cov}(\epsilon_t, \epsilon_{t+k})$. Then $\gamma_\epsilon(0) = \sigma_\epsilon^2$ and $\gamma_\epsilon(k) = 0$ for $k \neq 0$.

For an MA(q) process we obtain

a) The autocovariances:

$$\begin{aligned}
\gamma(k) &= \sum_{i=0}^{q}\sum_{j=0}^{q}\psi_i\psi_j\gamma_\epsilon(k+i-j) \\
&= \begin{cases} \sigma_\epsilon^2\left[\sum_{i=0}^{q-k}\psi_i\psi_{i+k}\right], & 0 \le k \le q, \\ 0, & k > q, \\ \gamma(-k), & k < 0. \end{cases}
\end{aligned} \tag{4}$$

b) In particular, $\text{Var}(X_t) = \sigma^2 = \sigma_\epsilon^2 \left[\sum_{i=0}^{q} \psi_i^2\right]$.

c) The autocorrelations:

$$\rho(k) = \begin{cases} \left[\sum_{i=0}^{q-k} \psi_i \psi_{i+k}\right] \Big/ \left[\sum_{i=0}^{q} \psi_i^2\right], & 0 \le k \le q, \\ 0, & k > q, \\ \rho(-k), & k < 0. \end{cases} \tag{5}$$

We see that $\gamma(k)$ and $\rho(k)$ of an MA(q) process are zero for $|k| > q$.

Examples:

1. For the MA(1) process $X_t = -0.9\epsilon_{t-1} + \epsilon_t$ we have

   a) $\gamma(0) = \text{Var}(X_t) = (\psi_0^2 + \psi_1^2)\sigma_\epsilon^2 = 1.81\sigma_\epsilon^2$,

   b) $\gamma(\pm 1) = \psi_1 \sigma_\epsilon^2 = -0.9\sigma_\epsilon^2$, $\rho(\pm 1) = -0.497$ and

   c) $\gamma(\pm k) = \rho(\pm k) = 0$ for $|k| > 1$.

2. For the MA(2) process $X_t = 0.7\epsilon_{t-1} + 0.4\epsilon_{t-2} + \epsilon_t$ we have

   a) $\gamma(0) = \text{Var}(X_t) = (1 + \psi_1^2 + \psi_2^2)\sigma_\epsilon^2 = 1.65\sigma_\epsilon^2$,

   b) $\gamma(\pm 1) = (\psi_0\psi_1 + \psi_1\psi_2)\sigma_\epsilon^2 = 0.98\sigma_\epsilon^2$, $\rho(\pm 1) = 0.594$,

   c) $\gamma(\pm 2) = \psi_0\psi_2\sigma_\epsilon^2 = 0.4\sigma_\epsilon^2$, $\rho(\pm 2) = 0.242$ and

   d) $\gamma(\pm k) = \rho(\pm k) = 0$ for $|k| > 2$.

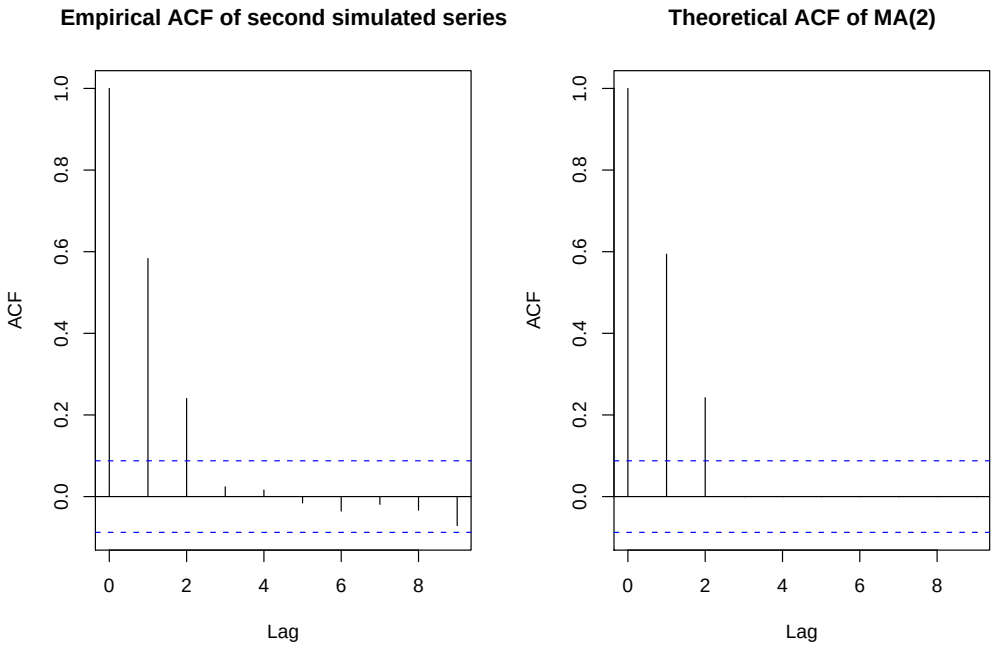


Figure 6: The ACFs (estimated and theoretical) for the MA(2) process. The estimated ACF is based on the second realisation shown in Figure 5.

3. Autoregressive Processes

Autoregressive (AR) processes are another important class of time series models that are widely used to analyse data.

Definition 3.1 *A pth order* autoregressive *process AR(p) is defined by*

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \epsilon_t, \tag{6}$$

where $\phi_p \neq 0$.

Equivalently, we have

$$\epsilon_t = X_t - \sum_{i=1}^{p} \phi_i X_{t-i}, \tag{7}$$

Observe that these processes are closely related to regression analysis, with regressors given by shifted (past) instances of the same time series.

Setting $\phi_0 = 1$, we have

$$\epsilon_t = \phi_0 B^0 X_t - \phi_1 B X_t - ... - \phi_p B^p X_t = \phi(B) X_t, \tag{8}$$

where $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$.

Examples: Given a WN process ϵ_t, we can define

1. An AR(1) process: $X_t = 0.8X_{t-1} + \epsilon_t$,
2. An AR(2) process: $X_t = 1.1X_{t-1} - 0.3X_{t-2} + \epsilon_t$.

Replacing B in $\phi(B)$ with a variable z we obtain a pth order polynomial $\phi(z)$.

Definition 3.2 *The polynomial $\phi(x)$ is called the* characteristic polynomial *of an AR(p) model. Similarly, $\psi(z)$ is the characteristic polynomial of an MA(q) model.*

Setting $\phi(z) = 0$ we obtain the *characteristic equation* of an AR(p) model:

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p = 0. \tag{9}$$

This equation has exactly p roots z_1, ..., z_p (possibly multiple or complex).

Note that $\phi(z)$ and $\{z_1, ..., z_p\}$ determine each other. Hence, the correlation structure and some other important properties of X_t are determined by z_1, ..., z_p.

To show this we introduce first the important concept of the unit circle in time series analysis.

The *unit circle* in the complex plane is the set of all complex numbers with norm one:

$$z = a + ib \text{ such that } |z| = 1,$$

where $i = \sqrt{-1}$ denotes the imaginary unit and $|z| = \sqrt{a^2 + b^2}$ is the Euclidean norm of z.

We have seen that an MA process with summable coefficients are stationary. But this is not true for an AR model. For instance, $X_t = X_{t-1} + \epsilon_t$, the random walk, is non-stationary.

Theorem 3.3 *An AR(p) process X_t is* causal and stationary*, iff (if and only if) all of the roots of $\phi(z)$ lie outside the unit circle, i.e., iff*

$$|z_k| > 1, \forall\, 1 \leq k \leq p.$$

Proof. See, for example, Theorem 3.1.1 of Brockwell and Davis (1991). ◇

These results are based on the assumption that the process starts at $t = -\infty$. For an AR process starting at $t = 0$ or $t = 1$, these results only hold asymptotically, i.e., X_t will converge (almost surely) to another $\tilde{X}_t$ as $t \to \infty$, where $\tilde{X}_t$ is defined following the same AR model but starting at $t = -\infty$.

If we can find a factorised version of $\phi(z)$, then it is easy to check whether the process is causal stationary. For the AR(3) process

$$X_t = 1.8X_{t-1} - 1.07X_{t-2} + 0.21X_{t-3} + \epsilon_t$$

we have

$$\phi(z) = (1 - 0.5z)(1 - 0.6z)(1 - 0.7z)$$

with

$$z_1 = 2,\ z_2 = 10/6 \text{ and } z_3 = 10/7.$$

These are all outside the unit circle and X is therefore stationary.

If the conditions of Theorem 3.3 are fulfilled, we have

$$X_t = \phi^{-1}(B)\epsilon_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \tag{10}$$

which is a causal stationary MA(∞) process with

$$\sum_{i=0}^{\infty} |\psi_i| < \infty.$$

Generally, $\psi_i \neq 0$ for all i, because the reciprocal $\phi^{-1}(z)$ of a finite order polynomial $\phi(z)$ is an infinite order polynomial. We study MA(∞) process in the following chapter.

3.1. AR(1) Processes

The simplest AR process is the AR(1) process

$$X_t = \phi X_{t-1} + \epsilon_t. \tag{11}$$

For this process the condition of Theorem 3.3 reduces to $|\phi| < 1$, and we have seen that $\psi_i = \phi^i$ and

$$X_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}. \tag{12}$$

We obtain

$$\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi^i = 1/(1 - \phi) < \infty.$$

The ACF of an AR(1) model can be calculated using its MA(∞) representation.

$$\begin{aligned}
\gamma(k) &= E(X_t X_{t+k}) \\
&= E\left\{ \left(\sum_{i=0}^{\infty} \phi^i \epsilon_{t-i} \right) \left(\sum_{j=0}^{\infty} \phi^j \epsilon_{t+k-j} \right) \right\} \\
&= \sigma_\epsilon^2 \phi^{|k|} \sum_{i=0}^{\infty} \phi^{2i} = \sigma_\epsilon^2 \frac{\phi^{|k|}}{1 - \phi^2},
\end{aligned}$$

and

$$\rho(k) \quad = \quad \gamma(k)/\gamma(0) = \phi^{|k|},$$

for $k = 0, \pm 1, \pm 2, \ldots$. Note that $\gamma(k) \neq 0$ for any k. For $\phi > 0$, $\rho(k)$ are always positive and decrease monotonically. For $\phi < 0$, $|\rho(k)|$ decrease monotonically but with alternating signs.

These results can also be obtained recursively. Note that

$$E(X_t) = 0,\ \gamma(k) = E[X_t X_{t+k}] \text{ and } E[X_{t-k}\epsilon_t] = \begin{cases} \sigma_\epsilon^2 & \text{for} \quad k = 0 \\ 0 & \text{for} \quad k \neq 0 \end{cases}$$

Multiply both sides of (11) with X_{t-k} and take expectations:

For $k = 0$

$$\begin{aligned} E(X_t X_t) &= E(X_t \phi X_{t-1}) + E(X_t \epsilon_t) \\ \gamma(0) &= \phi\gamma(1) + \sigma_\epsilon^2. \end{aligned} \tag{13}$$

For $k = 1$

$$\begin{aligned} E(X_{t-1} X_t) &= E(X_{t-1} \phi X_{t-1}) + E(X_{t-1}\epsilon_t) \\ \gamma(1) &= \phi\gamma(0) \end{aligned} \tag{14}$$

For $k \geq 2$

$$\begin{aligned} E(X_{t-k} X_t) &= E(X_{t-k} \phi X_{t-1}) + E(X_{t-k}\epsilon_t) \\ \gamma(k) &= \phi\gamma(k-1) \end{aligned} \tag{15}$$

By solving this system of equations system we obtain the same results as given above, i.e.,

$$\gamma(0) = \frac{1}{1-\phi^2}\sigma_\epsilon^2,\ \gamma(1) = \frac{\phi}{1-\phi^2}\sigma_\epsilon^2,\ \gamma(2) = \frac{\phi^2}{1-\phi^2}\sigma_\epsilon^2, ...,$$

$$\rho(0) \equiv 1,\ \rho(1) = \phi,\ \rho(2) = \phi^2, ...$$

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\phi\gamma(k-1)}{\gamma(0)} = \phi\rho(k-1) = \phi^k$$

Example $\rho(k)$ for AR(1) processes with $\phi = 0.8$ and $\phi = -0.8$, respectively. Figure 7 displays two realisations following each of these two AR(1) models with i.i.d. $N(0,1)$ ϵ_t. The realisations following AR models with $\phi > 0$ and $\phi < 0$ look quite different. The former is dominated by positive and the latter by negative correlations.

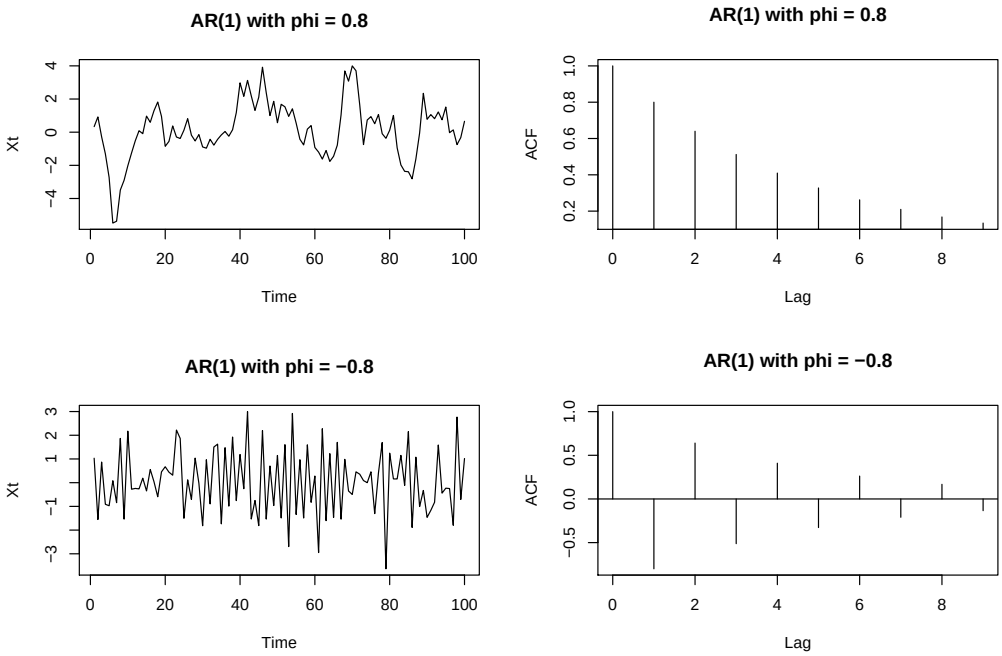


Figure 7: Two AR(1) processes and their autocorrelations

3.2. AR(2) Processes

Some properties of the more complex AR(2) processes will be discussed in this section.

For a given AR(2) model we have to first check whether it is causal stationary. In some special cases this can be done by means of a factorisation of $\phi(z)$.

Examples:

1) For the AR(2) process $X_t = 1.1X_{t-1} - 0.3X_{t-2} + \epsilon_t$ we have

$$\phi(z) = 1 - 1.1z + 0.3z^2 = (1 - 0.5z)(1 - 0.6z)$$

with the roots $z_1 = 2$ and $z_2 = 1\frac{2}{3}$. This process is causal stationary.

2) For the AR(2) process $X_t = -1.5X_{t-1} + X_{t-2} + \epsilon_t$ we have

$$\phi(z) = 1 + 1.5z - z^2 = (1 - 0.5z)(1 + 2z)$$

with the roots $z_1 = 2$ and $z_2 = -\frac{1}{2}$. This process is not causal stationary (however, there exist a non-causal stationary solution of X_t).

Usually we have to check whether the conditions of Theorem 3.3 are fulfilled or not. For an AR(2) model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$$

these conditions are equivalent to all of the following simple conditions on the coefficients holding:

(i) $\phi_1 + \phi_2 < 1$,

(ii) $\phi_2 - \phi_1 < 1$,

(iii) $-1 < \phi_2 < 1$.

See Appendix A for a proof of this.

Using these conditions it is easy to check whether an AR(2) model is causal stationary or not. An AR(2) model is stationary if conditions (i)–(iii) all hold. It is not stationary if one (or more) of these conditions does not hold.

Examples:

1) $X_t = 0.556X_{t-1} + 0.322X_{t-2} + \epsilon_t$ is causal stationary, because
$\phi_1 + \phi_2 = 0.878 < 1$, $\phi_2 - \phi_1 = -0.234 < 1$ and $-1 < \phi_2 = 0.322 < 1$.

2) $X_t = 0.424X_{t-1} - 1.122X_{t-2} + \epsilon_t$ is not causal stationary, because $\phi_2 = -1.122 < -1$.

The above stationary conditions define a triangle in the ϕ_1-ϕ_2 plane, which can be divided into four regions (see Figure 8):

Region A $\phi_1 > 0$ and $\phi_1^2 + 4\phi_2 > 0$ (two real roots z_1 and z_2)

Region B $\phi_1 < 0$ and $\phi_1^2 + 4\phi_2 > 0$ (two real roots z_1 and z_2)

Region C $\phi_1 < 0$ and $\phi_1^2 + 4\phi_2 < 0$ (two conjugate complex roots z_1 and z_2)

Region D $\phi_1 > 0$ and $\phi_1^2 + 4\phi_2 < 0$ (two conjugate complex roots z_1 and z_2)

The ACFs of an AR(2) process with coefficients in different regions have different functional forms.

The ACF of an AR(2) model

We have seen explicit formulas for the ACFs of MA processes and AR(1) processes. The ACF of an AR(p) process with $p > 1$ has to be calculated in a recursive way.

Consider an AR(2) process

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \epsilon_t$$

Multiply both sides with X_{t-k} and take expectations. We obtain, for $k = 0, 1$ and 2 respectively,

$$\begin{array}{lllll} \gamma(0) & -\phi_1\gamma(1) & -\phi_2\gamma(2) & = & \sigma_\epsilon^2, \\ \gamma(1) & -\phi_1\gamma(0) & -\phi_2\gamma(1) & = & 0, \\ \gamma(2) & -\phi_1\gamma(1) & -\phi_2\gamma(0) & = & 0, \end{array} \quad \text{or} \quad \begin{array}{lllll} \gamma(0) & -\phi_1\gamma(1) & -\phi_2\gamma(2) & = & \sigma_\epsilon^2, \\ -\phi_1\gamma(0) & +(1-\phi_2)\gamma(1) & +0\gamma(2) & = & 0, \\ -\phi_2\gamma(0) & -\phi_1\gamma(1) & +\gamma(2) & = & 0. \end{array} \tag{16}$$

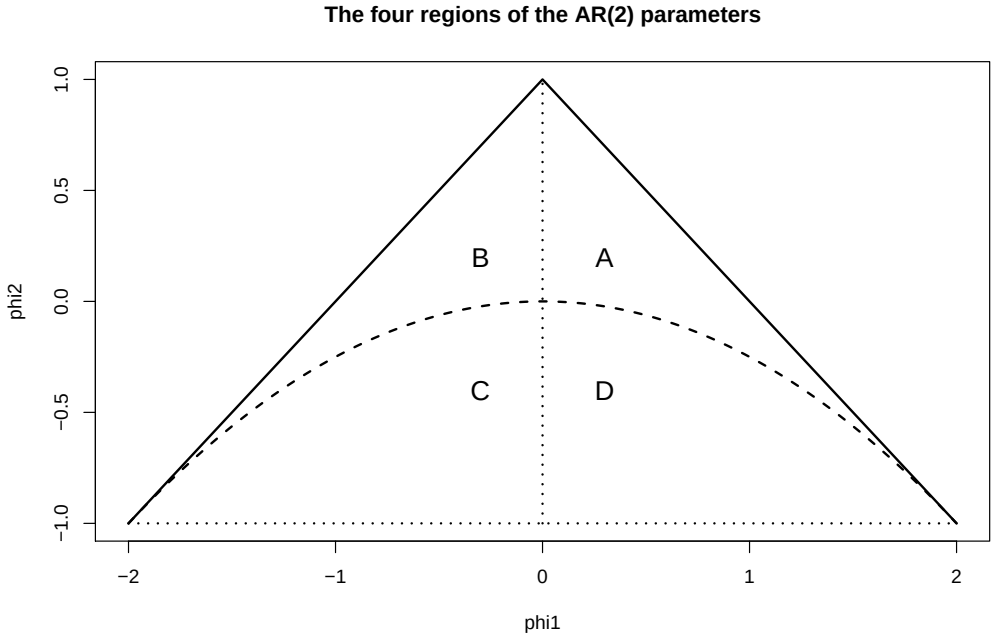


Figure 8: AR(2) stationarity triangle

Solving these equations, we have

$$\begin{aligned}\gamma(0) &= \frac{(1-\phi_2)\sigma_\epsilon^2}{(1+\phi_2)[(1-\phi_2)^2-\phi_1^2]},\\ \gamma(1) &= \frac{\phi_1\sigma_\epsilon^2}{(1+\phi_2)[(1-\phi_2)^2-\phi_1^2]},\\ \gamma(2) &= \frac{[\phi_1^2+\phi_2(1-\phi_2)]\sigma_\epsilon^2}{(1+\phi_2)[(1-\phi_2)^2-\phi_1^2]},\end{aligned} \tag{17}$$

and, for $k \geq 2$,

$$\gamma(k) = \phi_1\gamma(k-1) + \phi_2\gamma(k-2). \tag{18}$$

This recursive formula and the initial solutions given in (17) allow us to calculate the autocovariances $\gamma(k)$ of an AR(2) process for any finite k.

This idea can be generalized to a common AR(p) model, for which $p+1$ initial values have to be solved.

For $\rho(k)$ we have $\rho(0) \equiv 1$. For $k = 1$, dividing the second equation of (16) by $\gamma(0)$, we have

$$\rho(1) - \phi_1 - \phi_2\rho(1) = 0. \tag{19}$$

Hence

$$\rho(1) = \frac{\phi_1}{1-\phi_2}. \tag{20}$$

Analogously, we obtain recursive formulas for $\rho(k)$ for $k \geq 2$:

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2). \tag{21}$$

We see, for an AR(2) model, the calculation of $\rho(k)$ is a little bit easier than that of $\gamma(k)$.

Example. Calculate $\rho(k)$, $k = 0, 1, ..., 50$, of the AR(2) process $X_t = 0.5X_{t-1} + 0.4X_{t-2} + \epsilon_t$.

Solution: Following the above formulas we have
$\rho(0) = 1$, $\rho(1) = \phi_1/(1-\phi_2) = 0.833$, $\rho(2) = \phi_1\rho(1) + \phi_2\rho(0) = 0.817$,
$\rho(3) = \phi_1\rho(2) + \phi_2\rho(1) = 0.742$, $\rho(4) = 0.698$, $\rho(5) = 0.645$, $\rho(6) = 0.602$, ... ,
$\rho(48) = 0.029$, $\rho(49) = 0.027$ and $\rho(50) = 0.025$.

Similarly, we can obtain recursive formulas for computing the coefficients α_i in the MA(∞) representation of an AR process given in (10). But this will not be discussed further here. What we need to know are only some properties of α_i, such as absolute summability.

The ACF for the four different regions (see Figure 9):

1. $X_t = 0.5X_{t-1} + 0.4X_{t-2} + \epsilon_t$ with ϕ_1 and ϕ_2 in area A;
2. $X_t = -0.5X_{t-1} + 0.4X_{t-2} + \epsilon_t$ with ϕ_1 and ϕ_2 in area B;
3. $X_t = -1.8X_{t-1} - 0.9X_{t-2} + \epsilon_t$ with ϕ_1 and ϕ_2 in area C;
4. $X_t = 1.8X_{t-1} - 0.9X_{t-2} + \epsilon_t$ with ϕ_1 and ϕ_2 in area D.

We see that the ACF of an AR(2) process with coefficients in region A looks like that of an AR(1) with positive coefficient and that of an AR(2) process with coefficients in region B looks like that of an AR(1) with negative coefficient.

If the roots of $\phi(z)$ are complex, then $\rho(k)$ appears to be like damped sine waves. If the coefficients are in region C, the sign of $\rho(k)$ changes quite frequently. If the coefficients are in region D, the sign of $\rho(k)$ keeps the same in a half period, as by a sine function.

See Figure 10 for plots of the corresponding time series data.

3.3. Partial Autocorrelations

If we want to fit an AR(p) model to our data, we have to choose a proper order of the model. By means of the ACF, it is not easy to determine which p should be used. The partial autocorrelation function (PACF) was traditionally introduced as a tool for selecting the order of an AR model, because, whereas an AR(p) model has an ACF which is infinite in extent, its partial autocorrelations are only non-zero until lag $k = p$, similarly to the autocorrelations of an MA model.

The partial autocorrelation of a stationary process at lag k, denoted by $\alpha(k)$ is the correlation between X_t and X_{t-k} conditionally on $\{X_{t-1}, ..., X_{t-k+1}\}$. In other words, the par-

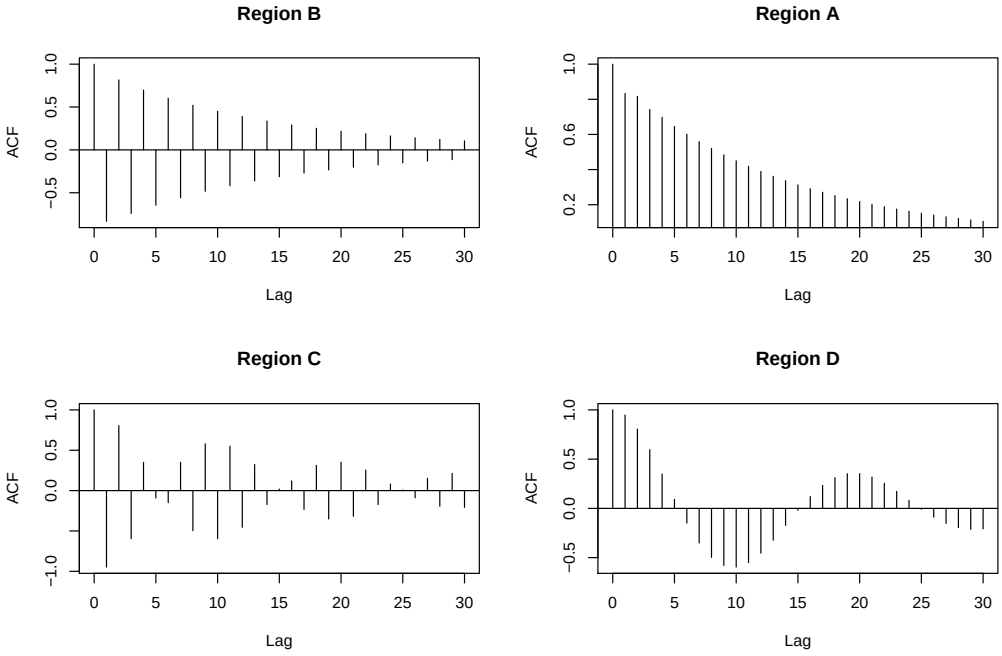


Figure 9: Autocorrelations for four different AR(2) processes

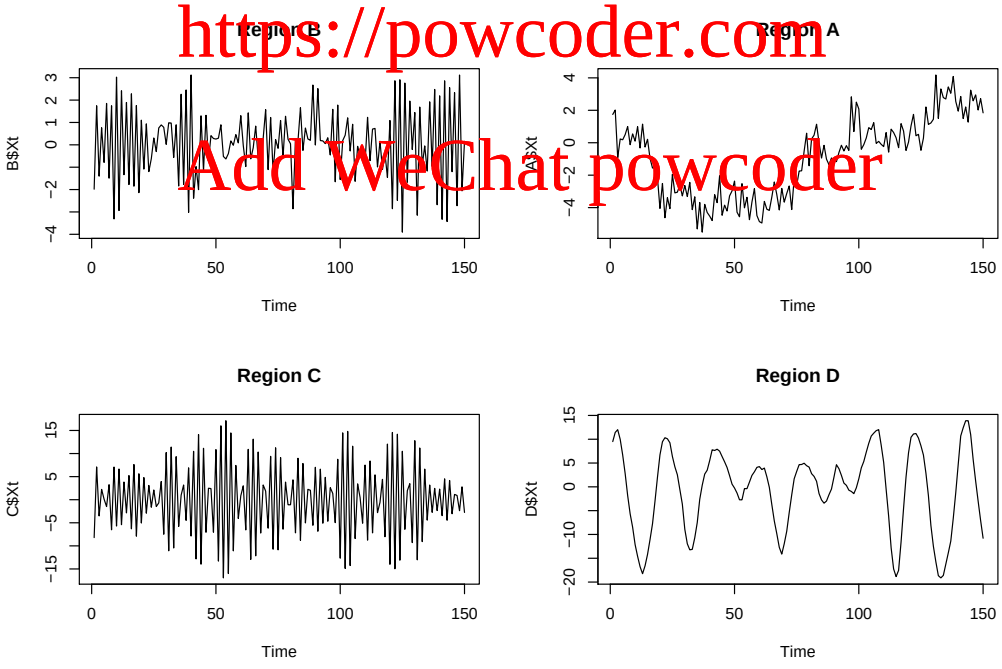


Figure 10: Simulations of four different AR(2) processes

tial autocorrelation $\alpha(k)$ is the correlation between the residuals of the regression of X_t on $\{X_{t-1}, ..., X_{t-k+1}\}$ and the residuals of the regression of X_{t-k} on $\{X_{t-1}, ..., X_{t-k+1}\}$.

Assume that we are given a stationary process with autocorrelations $\rho(k)$, $k = ..., -1, 0, 1, ...$ such that $\rho(k) \to 0$ as $k \to \infty$. The formulas for its PACF are very complex. However, it is easy to calculate the sample PACFs from your data using software, e.g., R.

The PACF for an AR(2) process:

1. Define $\alpha(0) = 1$ and $\alpha(1) = \rho(1)$ (no observations coming in between X_t and X_{t-1}).
2. For $k = 2$ we have

$$\alpha(2) = \frac{\begin{vmatrix} 1 & \rho(1) \\ \rho(1) & \rho(2) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{vmatrix}} = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)},$$

For an AR(p) process it can be shown that $\alpha(p) = \phi_p$ and $\alpha(k) = 0$ for all $k > p$. The result $\alpha(k) = 0$ for all $k > p$ is due to the fact that an AR(p) model is a pth order Markov process. For $k > p$, the influence of X_{t-k} on X_t is totally included in $X_{t-1}, ..., X_{t-p}$.

Example 1: For the AR(1) process: $X_t = \phi X_{t-1} + \epsilon_t$ $|\phi| < 1$ we have $\alpha(0) = 1$, $\alpha(1) = \phi$ and $\alpha(k) = 0$ for $k > 1$.

Example 2: For a causal AR(2) process $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$ we have $\alpha(0) = 1$, $\alpha(\pm 1) = \rho_1 = \phi_1/(1 - \phi_2)$, $\alpha(2) = \phi_2$ and $\alpha(k) = 0$ for $k > 2$ (following the above results). Note in particular that $-1 < \alpha(2) = \phi_2 < 1$, because it is some kind of correlation coefficient.

The PACFs of MA processes are very complex and are in general nonzero for all lags, like the ACF of AR models. For the simplest invertible MA(1) process $X_t = \epsilon_t - \psi\epsilon_{t-1} + \epsilon_t$ with $|\psi| < 1$, it can be shown that, after lengthy calculation,

$$\alpha(k) = -\frac{(-\psi)^k(1 - \psi^2)}{1 - \psi^{2(k+1)}}. \tag{22}$$

See Box and Jenkins (1976). For $\psi < 0$, $\alpha(k) < 0 \; \forall \; k \neq 0$. For $\psi > 0$, $\alpha(k)$ has alternating signs.

We see that the ACF of a MA(q) process is zero for lag $|k| > q$. And the ACF of an AR(p) process is nonzero for all lags. In contrast to this, the PACF of a MA(q) process is nonzero for all lags. And the PACF of an AR(p) process is zero for lag $|k| > p$.

4. The MA(∞) and AR(∞) Processes

4.1. The MA(∞) Processes

In time series analysis we often need to consider infinite order MA processes, denoted by MA(∞), which is an generalisation of the MA(q) processes described above. This will be motivated in the following by the simple AR(1) (first order autoregressive) process.

Example: The AR(1) process $X_t = \phi X_{t-1} + \epsilon_t$ with $|\phi| < 1$ has an MA(∞) representation

$$X_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}. \tag{23}$$

Proof: $X_t = \phi X_{t-1} + \epsilon_t = \phi(\phi X_{t-2} + \epsilon_{t-1}) + \epsilon_t = ... = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}.$ ◇

A general MA(∞) process is defined through a *linear filter* (MA-filter) of ϵ_t:

$$X_t = \sum_{i=-\infty}^{\infty} \psi_i \epsilon_{t-i}. \tag{24}$$

The MA(∞) process is causal if $\psi_i = 0 \; \forall \; i < 0$. Observe that the MA($\infty$) representation of an AR(1) process is causal. Hereafter, we will mainly consider causal MA(∞) processes

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{25}$$

Without loss of generality, we will often put $\psi_0 = 1$.

The MA(∞) process is stationary if ψ_i, $i = 0, 1, ...$, are squared summable, i.e.

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty. \tag{26}$$

Provided that (26) holds, by extending the results on the ACF of an MA(q) process, we have

1. $\gamma(0) = \text{Var}(X_t) = \left(\sum_{i=0}^{\infty} \psi_i^2 \right) \sigma_\epsilon^2 < \infty;$
2. $\gamma(k) = \text{Cov}(X_t, X_{t+k}) = \left(\sum_{i=0}^{\infty} \psi_i \psi_{i+|k|} \right) \sigma_\epsilon^2, k = 0, \pm 1, ...$
3. $\rho(k) = \left(\sum_{i=0}^{\infty} \psi_i \psi_{i+|k|} \right) \Big/ \left(\sum_{i=0}^{\infty} \psi_i^2 \right), k = 0, \pm 1, ...$

Many practically relevant autoregressive processes admit the causal MA(∞) representations defined in (25). The above results can hence be used to calculate (or approximate) the ACFs of autoregressive processes.

A stronger condition on the coefficients ψ_i of an MA(∞) process is absolute summability:

$$\sum_{i=0}^{\infty} |\psi_i| < \infty. \tag{27}$$

It is easy to show that

$$\sum_{i=0}^{\infty} |\psi_i| < \infty \Longrightarrow \sum_{i=0}^{\infty} \psi_i^2 < \infty,$$

but not vice versa. A series that converges faster than $i^{-1/2}$ is squared summable, and that which converges faster than i^{-1} is absolutely summable (see examples below).

Examples: Let $\alpha_0 = 1$.

1. $\psi_i = i^{-1/2}$ for $i = 1, 2, ... \Longrightarrow \sum_{i=0}^{\infty} \psi_i^2 = \infty, \sum_{i=0}^{\infty} |\psi_i| = \infty.$
2. $\psi_i = i^{-1}$ for $i = 1, 2, ... \Longrightarrow \sum_{i=0}^{\infty} \psi_i^2 < \infty, \sum_{i=0}^{\infty} |\psi_i| = \infty.$
3. $\psi_i = i^{-3/2}$ for $i = 1, 2, ... \Longrightarrow \sum_{i=0}^{\infty} \psi_i^2 < \infty, \sum_{i=0}^{\infty} |\psi_i| < \infty.$

Summability conditions play an important role in the theory of time series analysis, which can be seen from the following simple lemma.

Lemma 4.1 *For an MA(∞) process defined by (25), we have*

$$\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty, \text{ if } \sum_{i=0}^{\infty} |\psi_i| < \infty \tag{28}$$

and

$$\sum_{k=-\infty}^{\infty} \gamma^2(k) < \infty, \text{ if } \sum_{i=0}^{\infty} \psi_i^2 < \infty. \tag{29}$$

This also holds for the general MA(∞) given in (24).

The proof of Lemma 4.1 is left as an (optional) exercise.

The coefficients in (23) are obviously absolutely summable, i.e. AR(1) with $|\phi| < 1$ is stationarity with absolutely summable $\gamma(k)$.

A stationary process with absolutely summable $\gamma(k)$ is said to have *short memory*. This means that a stationary AR(1) process has short memory.

4.2. AR(∞) Processes

The AR(∞) process is given by

$$X_t = \sum_{i=1}^{\infty} \psi_i X_{t-i} + \epsilon_t \tag{30}$$

or, equivalently,

$$\epsilon_t = \sum_{i=0}^{\infty} \psi_i X_{t-i} \tag{31}$$

with $\psi_0 = 1$. Usually, it is assumed that $\sum_{i=0}^{\infty} |\psi_i| < \infty$.

Example: An MA(1) process $X_t = \psi\epsilon_{t-1} + \epsilon_t$ with $|\psi| < 1$ has the following AR(∞) representation:

$$\epsilon_t = \sum_{i=0}^{\infty} (-\psi)^i X_{t-i} \tag{32}$$

with absolutely summable coefficients $\phi_i = (-\psi)^i$, $i = 0, 1, \ldots$.

Note that X_t are observable but ϵ_t are often unobservable. Hence, the above AR(∞) model of the innovations ϵ_t is very useful in theory and practice because now it is possible to estimate ϵ_t from the data. The question is whether such an AR(∞) model is well defined for a given MA process. The answer is that this is only possible if the MA process is invertible.

4.3. Invertibility of MA(q) Processes

Invertibility: Any process X_t is said to be invertible if it can be represented as an AR(∞) process with (absolutely) summable coefficients.

Note that any AR process with summable coefficients is invertible.

Example: The RW $X_t = X_{t-1} + \epsilon_t$ is non-stationary but invertible.

There are however some simple MA models which are not invertible.

Example: The MA(1) process $X_t = \epsilon_{t-1} + \epsilon_t$ is stationary but not invertible.

Hence, we need to discuss when an MA process is invertible. For this we have the following theorem, which is closely related to Theorem 3.3 on the causal stationarity of an AR process. The characteristic equation of an MA(q) model is

$$\psi(z) = 1 + \psi_1 z + \ldots + \psi_q z^q = 0. \tag{33}$$

Again, this equation has q roots z_1, ..., z_q. And $\psi(z)$ and $\{z_1, \ldots, z_q\}$ determine each other. Hence, the correlation structure of X_t is determined by z_1, ..., z_q.

Theorem 4.2 *An MA(q) process X_t is* invertible*, iff all of the roots of $\psi(z)$ lie outside the unit circle, i.e., iff*

$$|z_i| > 1, \forall\, 1 \leq i \leq q.$$

Proof. See Brockwell and Davis (1991). ◇

The conditions $|z_i| > 1$ in Theorem 4.2 imply that $\psi^{-1}(B)$ is well defined with non-negative powers and absolutely summable coefficients. Now we have

$$\epsilon_t = \psi^{-1}(B)X_t = \sum_{i=0}^{\infty} \psi_i X_{t-i},$$

which is a causal stationary AR(∞) process with $\sum_{i=0}^{\infty} |\phi_i| < \infty$.

We see there is a dual relationship between the MA and AR processes.

- A causal invertible MA process can be represented as a causal stationary AR(∞).
- A causal stationary AR process can be represented as a causal, stationary (and also invertible) MA(∞).

Analogously to the causal/stationary conditions for AR(1) and AR(2), we have the following invertible conditions for MA(1) and MA(2) processes:

1) An MA(1) process $X_t = \psi\epsilon_{t-1} + \epsilon_t$ is invertible iff $|\psi| < 1$.

2) An MA(2) process $X_t = \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \epsilon_t$ is invertible iff all three of the following conditions hold:

(i) $\psi_1 + \psi_2 > -1$;

(ii) $\psi_1 - \psi_2 < 1$;

(iii) $-1 < \psi_2 < 1$.

Examples:

1) $X_t = 1.5\epsilon_{t-1} + 0.75\epsilon_{t-2} + \epsilon_t$ is invertible, because
$\psi_1 + \psi_2 = 2.25 > -1$, $\psi_1 - \psi_2 = 0.75 < 1$, and $-1 < \psi_2 = 0.75 < 1$.

2) $X_t = 0.75\epsilon_{t-1} - 0.5\epsilon_{t-2} + \epsilon_t$ is not invertible, because $\psi_1 - \psi_2 = 1.25 > 1$.

5. The ARMA and ARIMA Processes

5.1. ARMA Processes

Definition 5.1 *An Autoregressive Moving Average process of order* (p, q) *(ARMA(p, q)) is defined by*

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \psi_1 \epsilon_{t-1} + ... + \psi_q \epsilon_{t-q} + \epsilon_t. \tag{34}$$

An ARMA model combines an AR and an MA models. Equivalently, (34) can be represented in the following way:

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \psi_1 \epsilon_{t-1} + ... + \psi_q \epsilon_{t-q} + \epsilon_t, \tag{35}$$

That is,

$$\phi(B) X_t = \psi(B) \epsilon_t, \tag{36}$$

where

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

is the characteristic polynomial of the AR part and

$$\psi(z) = 1 + \psi_1 z + ... + \psi_q z^q$$

the characteristic polynomial of the MA part.

An AR(p) model is an ARMA(p, 0) model with $\psi(z) \equiv 1$, and an MA(q) model is an ARMA(0, q) model with $\phi(z) \equiv 1$.

Examples: Given a WN process ϵ_t, we can define

1. An ARMA(1, 1) process:

$$X_t = 0.8 X_{t-1} + 0.6 \epsilon_{t-1} + \epsilon_t$$

2. An ARMA(2,2) process:

$$X_t = 0.7 X_{t-1} + 0.1 X_{t-2} + 0.8 \epsilon_{t-1} + 0.16 \epsilon_{t-2} + \epsilon_t$$

The following theorem is one of the most important theorems in time series analysis. For an ARMA(p, q) process we have

Theorem 5.2 *Assume that* $\phi(z)$ *and* $\psi(z)$ *have no common factors. Then the ARMA(p, q) process is*

a) causal (stationary) iff all roots of $\phi(z)$ *lie outside the unit circle,*

b) invertible iff all roots of $\psi(z)$ *lie outside the unit circle,*

c) causal (stationary) and invertible iff all roots of $\phi(z)$ and $\psi(z)$ lie outside the unit circle.

Theorem 5.2 combines Theorems 3.3 and 4.2.

The causal stationarity and invertibility conditions of an ARMA model do not depend on each other.

By combining the conditions given above we can check whether an ARMA(p, q), for $p = 0, 1, 2$ and $q = 0, 1, 2$, process is causal stationary and/or invertible or not.

Remark: $\phi(z)$ and $\psi(z)$ have a common factor if there exists a function $f(z)$ such that

$$\phi(z) = f(z)\tilde{\phi}(z) \text{ and } \psi(z) = f(z)\tilde{\psi}(z)$$

If a common factor exists then $\tilde{\phi}(z)$ and $\tilde{\psi}(z)$ instead of $\phi(z)$ and $\psi(z)$ should be used in Theorem 5.2.

Examples:

1) $X_t - 0.6X_{t-1} - 0.3X_{t-2} = 1.5\epsilon_{t-1} + 0.75\epsilon_{t-2} + \epsilon_t$ is both causal stationary and invertible.

2) $X_t - 0.6X_{t-1} - 0.3X_{t-2} = 0.75\epsilon_{t-1} - 0.5\epsilon_{t-2} + \epsilon_t$ is causal stationary but not invertible.

3) $X_t - 0.6X_{t-1} - 0.5X_{t-2} = 1.5\epsilon_{t-1} + 0.75\epsilon_{t-2} + \epsilon_t$ is invertible but not causal stationary.

4) $X_t - 0.6X_{t-1} - 0.5X_{t-2} = 0.75\epsilon_{t-1} - 0.5\epsilon_{t-2} + \epsilon_t$ is neither causal stationary nor invertible.

Remark: Under the assumptions of Theorem 5.2 c), an ARMA(p,q) process has on the one hand the MA(∞) representation

$$X_t = \sum_{i=0}^{\infty} \alpha_i \epsilon_{t-i}, \tag{37}$$

where $\alpha(z) = \phi(z)^{-1}\psi(z)$ with $\sum |\alpha_i| < \infty$, and on the other hand the AR(∞) representation

$$\epsilon_t = \sum_{i=0}^{\infty} \beta_i X_{t-i}, \tag{38}$$

where $\beta(z) = \phi(z)\psi(z)^{-1}$, with $\sum |\beta_i| < \infty$.

The fact that α_i or β_i are absolutely summable follows since the convolution of two absolutely summable sequences is absolutely summable.

Under the assumptions of Theorem 5.2 c), the autocovariances of an ARMA(p,q) process are always absolutely summable, i.e.,

$$\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty.$$

The mean of MA, AR or ARMA processes may be non-zero:

Definition 5.3 *An ARMA(p, q) with mean μ is defined by*

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + ... + \phi_p(X_{t-p} - \mu) + \psi_1\epsilon_{t-1} + ... + \psi_q\epsilon_{t-q} + \epsilon_t. \quad (39)$$

Note that, if the mean is known, we can simply assume that $\mu = 0$ as before. If the mean is unknown, it is not difficult to estimate μ from the data and to remove it. Hence, ARMA processes with unknown mean have similar properties as those given above.

The ACF

We rewrite the ARMA(p, q) process as

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \psi_1\epsilon_{t-1} + ... + \psi_q\epsilon_{t-q} + \epsilon_t. \quad (40)$$

Multiplying both sides of (40) by X_{t-k} and taking the expectations, we obtain

$$\gamma(k) - \phi_1\gamma(k-1) - ... - \phi_p\gamma(k-p) = \sigma_\epsilon^2 \sum_{k \le j \le q} \psi_j\psi_{j-k} \quad (41)$$

for $0 \le k < \max(p, q+1)$, and

$$\gamma(k) - \phi_1\gamma(k-1) - ... - \phi_p\gamma(k-p) = 0 \quad (42)$$

for $k \ge \max(p, q+1)$. The recursive equation (42) is the same as for an AR(p).

5.2. ARIMA Processes

The class of ARMA processes is the most important class of stationary time series. However, in practice, in particular in finance and insurance, most time series observed are non-stationary. One important reason for non-stationarity is the effect due to the integration of two stationary observations. Non-stationary processes in this sense are hence called integrated ones. In the sequel, the ARMA processes will be extended to well-known linear non-stationary integrated processes.

Processes X_t whose dth differenced series Y_t are ARMA processes, where $d = 0, 1, ...$, are called ARIMA (autoregressive integrated moving average) processes, denoted by ARIMA(p, d, q).

Definition 5.4 *(The ARIMA(p, d, q) process) If d is a non-negative integer, then $\{X_t\}$ is said to be an ARIMA(p, d, q) process if $Y_t := (1-B)^d X_t$ is a causal stationary ARMA process.*

This definition means that X_t satisfies the difference equation

$$\phi(B)(1-B)^d X_t = \psi(B)\epsilon_t \quad \text{or} \quad \phi(B)Y_t = \psi(B)\epsilon_t \text{ with } Y_t = (1-B)^d X_t, \quad (43)$$

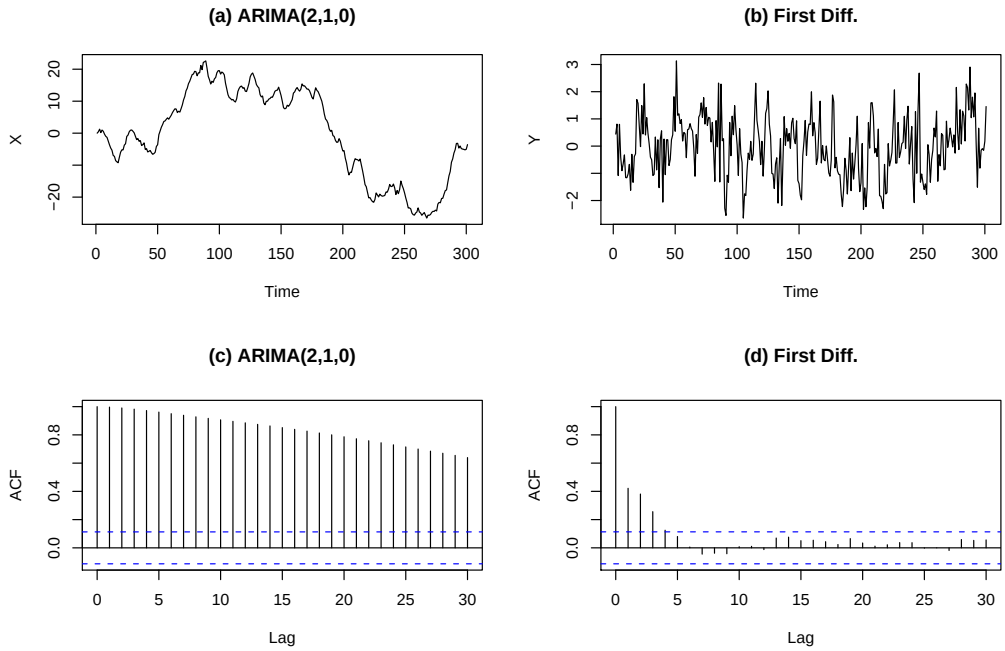


Figure 11: An ARIMA(2,1,0) process, its first difference, and their correlograms

where $d = 0, 1, ...$, $\phi(z)$ and $\psi(z)$ are characteristic polynomials of the AR and MA parts, respectively, and $\{\epsilon_t\}$ is white noise.

If $d = 0$ we will simply have an ARMA process. The process is stationary if and only if $d = 0$. This means that an ARIMA model with $d \geq 1$ is non-stationary.

If a time series follows an ARIMA(p, 1, q) model, then the series of the first differences, i.e. $Y_t = \Delta X_t = X_t - X_{t-1}$, will follow an ARMA($p$, q) model and is stationary. Similarly, if a time series follows an ARIMA(p, 2, q) model, then the *second* differences follow an ARMA(p, q) model and are stationary.

Suppose we are given observations $x_1, x_2, ..., x_n$ of a time series X_t. The graph of the sample autocorrelations $\hat{\rho}(k)$, called the *correlogram*, can be empirically used to check whether X_t is non-stationary or stationary.

In Figure 11, we have a simulation from the following ARIMA(2, 1, 0) model displayed in part (a):

$$(1 - B)X_t = Y_t \qquad \text{with} \qquad Y_t = 0.4Y_{t-1} + 0.2Y_{t-2} + \epsilon_t,$$

where ϵ_t are i.i.d. $N(0, 1)$ random variables and where $t = 0, 1, ..., 300$.

The differenced series $y_t = x_t - x_{t-1}$, $t = 1, 2, ..., 300$, is shown in part (b).

Note that the original series $\{x_i\}$ is of length 301, but the differenced series $\{y_i\}$ is of length 300. One observation is lost by taking first order differences.

The estimated ACFs (the correlograms) of x_t and y_t are shown respectively in part (c) and (d).

The two dashed lines in a correlogram are the so-called $\pm 2/\sqrt{n}$ confidence bounds, which will be explained later.

Random Walk process

We now consider the random walk process, which is a simple example of an ARIMA process that has applications in financial time series analysis.

For practical and theoretical reason, a random walk is often assumed to start at the time point $t = 0$ with known $X_0 = x_0$. Without loss of generality it is often assumed $X_0 = 0$.

Definition 5.5 *(Random Walk) A random walk is the stochastic process defined by*

$$X_t = \begin{cases} X_0, & t = 0, \\ X_{t-1} + \epsilon_t, & t > 0, \end{cases} \tag{44}$$

where $\{\epsilon_t\}$ is a white noise with mean zero and variance σ_ϵ^2.

A random walk is indeed an ARIMA(0, 1, 0) model starting from $t = 1$. Obviously, we have

$$X_t = \sum_{i=1}^{t} \epsilon_i + X_0. \tag{45}$$

$E(X_t) = X_0$ and $\mathrm{Var}(X_t) = t\sigma_\epsilon^2$, (not stationary), causal.

Due to the very big variance, a realisation of a random walk often shows a (stochastic) trend, which is however purely random.

Two realisations of length $n = 301$ following the same random walk model with i.i.d. $N(0, 1)$ innovations are shown in Figure 12 together with their sample ACFs.

A random walk is also invertible, because we have

$$\epsilon_t = X_t - X_{t-1}, \tag{46}$$

for $t > 0$, where ϵ_t are i.i.d.

A time series in practice may also have a non-stochastic trend together with a stochastic one. If there is a simple linear trend in a random walk, then it can be modelled by a random walk with drift defined by

$$X_t = \begin{cases} X_0, & t = 0, \\ X_{t-1} + \mu + \epsilon_t, & t > 0, \end{cases} \tag{47}$$

where $\{\epsilon_t\}$ is white noise with mean zero and variance σ_ϵ^2, and $\mu \neq 0$ is an unknown constant. Now, we have

$$X_t = t\mu + \sum_{i=1}^{t} \epsilon_i + X_0 \tag{48}$$

with mean $E(X_t) = t\mu + X_0$, which forms a linear non-stochastic trend in such a time series.

Moreover, it holds that $\epsilon_t = X_t - X_{t-1} - \mu$.

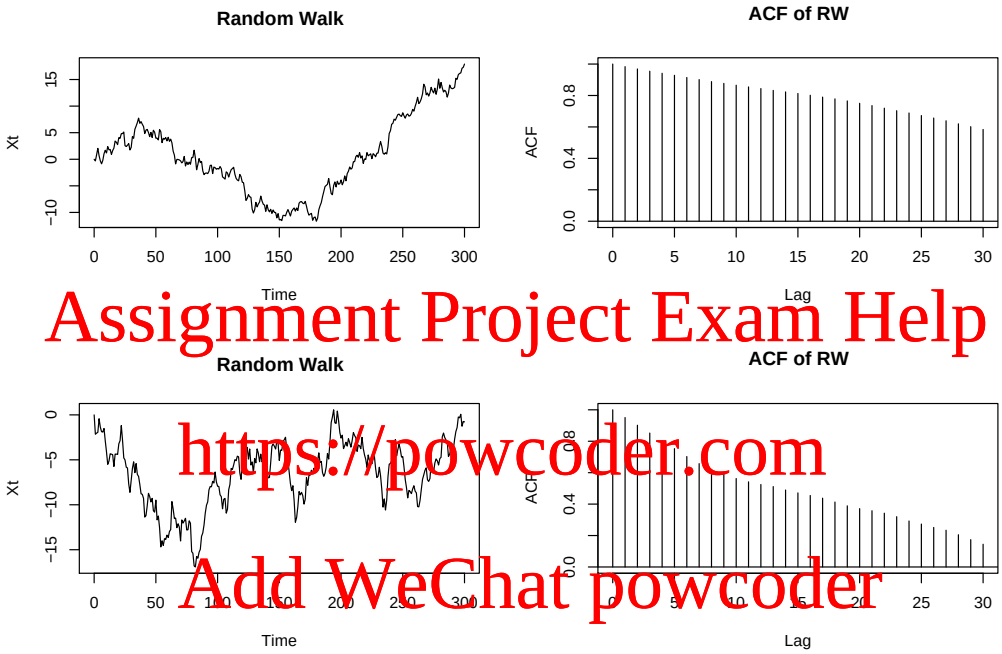


Figure 12: Two realisations of a random walk

6. Forecasting for Linear Processes

6.1. Box-Jenkins Forecasting

We want to forecast the value of a future value of the time series X_{t+k} given the observed values of the series $x_t, x_{t-1}, \ldots, x_1$, for some lead time $k > 0$.

There are many different approaches to forecasting. We concentrate on the Box-Jenkins method to forecasting, which is based on minimising mean square forecast error under fitted ARIMA models. For simplicity, we present the method with ARMA models, the extension to ARIMA models is conceptionally straightforward but requires more calculations.

Consider the MA(∞) representation of an ARMA(p, q) model

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \quad \text{with } \psi_0 = 1$$

So

$$\begin{aligned} X_{t+k} &= \sum_{j=0}^{\infty} \psi_j \epsilon_{t+k-j} \\ &= \epsilon_{t+k} + \psi_1 \epsilon_{t+k-1} + \psi_2 \epsilon_{t+k-2} + \ldots + \psi_{k-1} \epsilon_{t+1} + \psi_k \epsilon_t + \psi_{k+1} \epsilon_{t-1} + \ldots \end{aligned}$$

We focus on linear forecasting techniques that construct estimates of X_{t+k} by taking a linear combination of $x_t, x_{t-1}, x_{t-2}, \ldots$ or, equivalently, of $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots$ (of course, the values of $\epsilon_t, \epsilon_{t-1}, \ldots$ are unknown in practice, so estimates $\hat{\epsilon}_t, \hat{\epsilon}_{t-1}, \ldots$ are used instead). Again, the extension to non-linear forecasts is conceptually simple, but computationally more demanding.

More precisely, we formulate the forecast of X_{t+k} as

$$X_t(k) = \sum_{j=k}^{\infty} a_j \epsilon_{t+k-j} = a_k \epsilon_t + a_{k+1} \epsilon_{t-1} + \ldots$$

and use the *mean square error* (MSE) of the forecast

$$E[\{X_{t+k} - X_t(k)\}^2]$$

as measure of forecasting accuracy.

We aim to find the coefficients $\{a_j\}$ that produce the forecast $X_t(k)$ with optimal (smallest) forecasting MSE. By substituting $X_t(k)$ in $E[\{X_{t+k} - X_t(k)\}^2]$ we obtain

$$\text{MSE} = E[\{(\epsilon_{t+k} + \psi_1 \epsilon_{t+k-1} + \psi_2 \epsilon_{t+k-2} + \ldots) - (a_k \epsilon_t + a_{k+1} \epsilon_{t-1} + \ldots)\}^2],$$

which is equivalent to

$$\begin{aligned} E[\{\epsilon_{t+k} + \psi_1 \epsilon_{t+k-1} + \cdots + \psi_{k-1} \epsilon_{t+1} + (\psi_k - a_k)\epsilon_t + (\psi_{k+1} - a_{k+1})\epsilon_{t-1} + \ldots\}^2] \\ = \left\{ \sum_{j=0}^{k-1} \psi_j^2 + \sum_{j=k}^{\infty} (\psi_j - a_j)^2 \right\} \sigma_\epsilon^2 \end{aligned}$$

This is a quadratic function w.r.t. the coefficients $\{a_j\}$ that can be minimised analytically. The optimal coefficients are given by $a_j = \psi_j \;\; j = k, k+1, \ldots$.

We conclude that the minimum MSE forecast is:

$$X_t(k) = \psi_k \epsilon_t + \psi_{k+1}\epsilon_{t-1} + \cdots = \sum_{j=k}^{\infty} \psi_j \epsilon_{t+k-j}$$

which corresponds to the conditional expectation of X_{t+k} given the observed data; that is, the optimal forecast $X_t(k) = \mathrm{E}(X_{t+k}|X_t = x_t, X_{t-1} = x_{t-1}, \ldots)$. For AR($p$) processes

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

which suggests the more natural and equivalent representation

$$X_t(k) = \phi_1 X_t(k-1) + \phi_2 X_t(k-2) + \ldots + \phi_p X_t(k-p)$$

that is constructed recursively from $X_t(1) = \phi_1 x_t + \phi_2 x_{t-1} + \ldots + \phi_p x_{t-p+1}$, then $X_t(1) = \phi_1 X_t(1) + \phi_2 x_t + \ldots + \phi_p x_{t-p+2}$, which does not explicitly involve estimates of the unobserved process $\{\epsilon_t\}$. More precisely, for an AR(p), the point prediction for X_{n+k} and given observations x_1, ..., x_n is

$$\hat{x}_{n+k} = \phi_1(\hat{x}_{n+k-1}|x_n, ..., x_1) + \cdots + \phi_p(\hat{x}_{n+k-p}|x_n, ..., x_1), \tag{49}$$

where $(\hat{x}_{n+k-i}|x_n, ..., x_1)$, $i = 1, ..., p$, are either the observations x_{n+k-i}, if $k - i \leq 0$ or the prediction $\hat{x}_{n+k-i}$ obtained before, if $k - i > 0$.

The prediction procedure is as follows:

1. The first step optimal linear prediction for $k = 1$ is

$$\hat{x}_{n+1} \;\; = \;\; \phi_1 x_n + ... + \phi_p x_{n+1-p}.$$

2. The second step optimal linear prediction for $k = 2$ is

$$\hat{x}_{n+2} \;\; = \;\; \phi_1 \hat{x}_{n+1} + \phi_2 x_n + ... + \phi_p x_{n+2-p}.$$

3. The k-step optimal linear prediction for $k > p$ is

$$\hat{x}_{n+k} \;\; = \;\; \phi_1 \hat{x}_{n+k-1} + ... + \phi_p x_{n+k-p}.$$

Remark: Point predictions for a given an AR(p) model depend only on the last p observations (see examples below).

Remark: The prediction $\hat{X}_{n+k}$ tends to the process mean as k increases. This indicates that the process has finite memory and that the information from observations from the past decreases as we move into the future.

k	1	2	3	4	5	6	7	8	9	10
$\hat{x}_{n+k}$	-0.840	-0.588	-0.412	-0.288	-0.202	-0.141	-0.099	-0.069	-0.048	-0.034

Table 1: The first 10 point predictions for our AR(1) model with $x_n = -1.2$

k	1	2	3	4	5	6	7	8	9	10
$\hat{x}_{n+k}$	0.630	0.528	0.506	0.462	0.429	0.396	0.366	0.338	0.313	0.289

Table 2: The first 10 point predictions for our AR(2) model with $x_{n-1} = 1.1$, $x_{n-1} = 0.5$.

Example: From a time series x_1, ..., x_n we obtained the following AR(1) model:

$$X_t = 0.7X_{t-1} + \epsilon_t,$$

where ϵ_t are iid WN with mean zero. The last observation is $x_n = -1.2$. Calculate the first 10 point predictions.

Solution: $\hat{x}_{n+1} = \phi x_n = 0.7 * (-1.2) = -0.84$, $\hat{x}_{n+k} = \phi \hat{x}_{n+k-1} = \phi^k x_n$ for $k \geq 2$. All predictions are listed in Table 1.

Example: From a time series x_1, ..., x_n we obtained the following AR(2) model:

$$X_t = 0.6X_{t-1} + 0.3X_{t-2} + \epsilon_t,$$

where ϵ_t are iid WN with mean zero. The last two observation are $x_{n-1} = 1.1$ and $x_n = 0.5$. Calculate the first 10 point predictions.

Solution: $\hat{x}_{n+1} = \phi_1 x_n + \phi_2 x_{n-1} = 0.6 * 0.5 + 0.3 * 1.1 = 0.630$, $\hat{x}_{n+2} = \phi_1 \hat{x}_{n+1} + \phi_2 x_n = 0.6 * 630 + 0.3 * 0.5 = 0.528$, $\hat{x}_{n+k} = \phi_1 \hat{x}_{n+k-1} + \phi_2 \hat{x}_{n+k-2}$ for $k \geq 3$. All predictions are listed in Table 2.

If estimates of $\{\epsilon_t\}$ are required, e.g., for a general ARMA(p, q) model, we can adopt an AR(∞) representation, or construct estimates $\hat{\epsilon}_t$ recursively as follows:

$$\hat{\epsilon}_1 = x_1,$$

$$\hat{\epsilon}_2 = x_2 - \phi_1 x_1 - \psi_1 \hat{\epsilon}_1,$$

$$\dots$$

$$\hat{\epsilon}_t = x_t - \phi_1 x_{t-1} \dots - \phi_p x_{t-p} - \psi_1 \hat{\epsilon}_{t-1} \dots - \psi_q \hat{\epsilon}_{t-q},$$

by assuming that $\epsilon_t = X_t = 0$ for $t < 0$.

These estimates can then be used in $X_t(k)$ to produce forecasts.

6.2. Forecasting intervals and error

Many time series analysis applications require quantifying the uncertainty in the forecasts and reporting forecasting intervals. Again, we focus on ARMA models and note that the generalisation to other linear models is conceptually straightforward but requires more calculations.

Recall that a causal stationary ARMA(p, q) process can be represented in MA(∞) form:

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_i.$$

In a manner akin to the previous section where we studied point prediction of $\hat{X}_{n+k}$ based on the conditional mean of a future observation X_{n+k}, here we analyse the uncertainty in our forecasts by analysing the conditional variance of X_{n+k} given $x_1, ..., x_n$:

$$\text{Var}(X_{n+k}|X_n, ..., X_1) = \sigma_\epsilon^2 \sum_{i=0}^{k-1} \psi_i^2.$$

Observe that $\psi_0 = 1$ and $\text{Var}(X_{n+k}) = \sigma_\epsilon^2 \left(\sum_{i=0}^{\infty} \psi_i^2 \right)$. We have

1. $\text{Var}(X_{n+k}|X_n, ..., X_1)$ increases as k increases.
2. $\text{Var}(X_{n+1}|X_n, ..., X_1) = \sigma_\epsilon^2$.
3. $\sigma_\epsilon^2 \leq \text{Var}(X_{n+k}|X_n, ..., X_1) \leq \sigma_\epsilon^2 \sum_{i=0}^{\infty} \psi_i^2 = \text{Var}(X_{n+k})$ for $k \geq 2$.
4. $\lim_{k \to \infty} \text{Var}(X_{n+k}|X_n, ..., X_1) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \psi_i^2 = \text{Var}(X_{n+k})$.

The approximate 95% forecasting interval (FI) for X_{n+k} is

$$X_{n+k} \in \left(\hat{x}_{n+k} - 2\sigma_\epsilon \sqrt{\sum_{i=0}^{k-1} \psi_i^2}, \quad \hat{x}_{n+k} + 2\sigma_\epsilon \sqrt{\sum_{i=0}^{k-1} \psi_i^2} \right).$$

Remark: The FI here is for one observation, which is not the same as the FI for the sample mean $\bar{x}$. Note that $\lim_{k \to \infty} \hat{X}_{n+k} = 0 = \mu_X$ and $\lim_{k \to \infty} \text{Var}(X_{n+k}|X_n, ..., X_1) = \text{Var}(X_{n+k})$. This means that the larger k is, the less information about a future observation X_{n+k} is contained in the past observations.

Example: Figure 13 shows a realisation of length 150 following the AR(1) model

$$X_t = 0.8X_{t-1} + \epsilon_t,$$

where ϵ_t are i.i.d. $N(0, 1)$ random variables. Note that we have now $\psi_i = 0.8^i$, $i = 0, 1, ...$ The first 100 observations are shown in solid lines, whereas the last 50 (see the first row of the table) in points, which are assumed to be unknown future values.

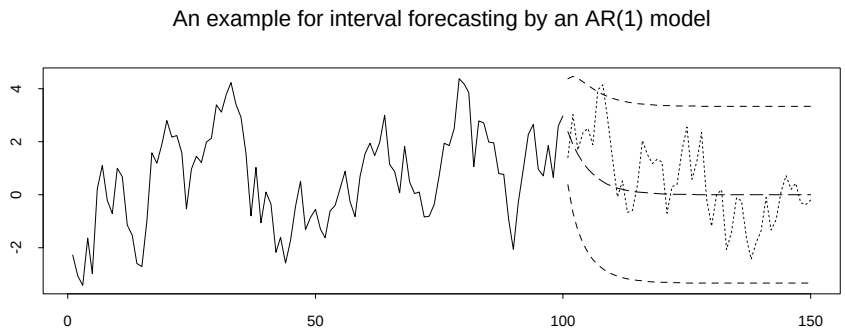


Figure 13: A realisation of an AR(1) model and the corresponding forecast intervals

k	1	2	3	4	5	6	7	8	9	10
x_{n+k}	1.403	3.032	1.687	2.337	2.509	1.888	3.956	4.148	2.917	1.375
$\hat{x}_{n+k}$	2.373	1.898	1.519	1.215	0.972	0.778	0.622	0.498	0.398	0.318
lower bound	0.373	-0.663	-1.345	-1.826	-2.177	-2.439	-2.637	-2.788	-2.905	-2.996
upper bound	4.373	4.460	4.382	4.256	4.121	3.994	3.881	3.784	3.701	3.633

Table 3: The first 10 point predictions and the corresponding 95%-forecasting intervals

The point predictions for $n = 100$ and $k = 1, 2, ..., 50$, and the 95%-forecasting intervals are shown in long- resp. short-dashed lines, where $x_{100} = 2.966$. The first 10 point predictions and the corresponding forecasting intervals are listed in Table 3.

To analyse the forecast error, we note that $X_t(k)$ coincides with the MA(∞) representation of X_{t+k} with the ϵ's between $t+1$ and $t+k$ set to zero, as this is their expected value. Consequently,

$$X_{t+k} = (\epsilon_{t+k} + \psi_1\epsilon_{t+k-1} + \psi_2\epsilon_{t+k-2} + \ldots + \psi_{k-1}\epsilon_{t+1}) + (\psi_k\epsilon_t + \psi_{k+1}\epsilon_{t-1} + \ldots)$$

$$= \underbrace{\sum_{j=0}^{k-1} \psi_j\epsilon_{t+k-j}}_{\downarrow} + \underbrace{\sum_{j=k}^{\infty} \psi_j\epsilon_{t+k-j}}_{\downarrow} \quad (*)$$

the forecast error $R_t(k)$ + the forecast $X_t(k)$

$R_t(1) = \epsilon_{t+1}$, i.e., the one-step-ahead forecast error is simply the next noise term. Hence

$$\epsilon_{t+1} = X_{t+1} - X_t(1).$$

Note this useful fact!

$$\begin{aligned} R_t(2) &= \epsilon_{t+2} + \psi_1\epsilon_{t+1} \\ R_t(3) &= \epsilon_{t+3} + \psi_1\epsilon_{t+2} + \psi_2\epsilon_{t+1} \end{aligned}$$

etc. $E[R_t(k)] = 0$, so the forecast is *unbiased*. The variance of the forecast error is

$$V(k) = \text{Var}(R_t(k)) = \left(\sum_{j=0}^{k-1} \psi_j^2\right) \sigma_\epsilon^2.$$

This gives the *standard error* of the forecast (the "*standard error of prediction*") as

$$s.e.(k) = \sqrt{V(k)} = \left(\sum_{j=0}^{k-1} \psi_j^2\right)^{1/2} \widehat{\sigma}_\epsilon$$

Assuming normality of the ϵ's, we can derive forecast intervals that contain X_{t+k} with 95% probability

$$X_t(k) \pm 2\, s.e.(k)$$

(the 95% probability is calculated under the assumption that the coefficients $\{\phi_j\}$ are perfectly known, which is not the case in practice).

Notice that:

1.

$$\begin{aligned} V(1) &= \sigma_\epsilon^2 \\ V(2) &= (1+\psi_1^2)\sigma_\epsilon^2 \\ V(3) &= (1+\psi_1^2+\psi_2^2)\sigma_\epsilon^2 \end{aligned}$$

   etc. Clearly $V(k)$ *increases* with k: this matches our intuition.

2. For *stationary* ARMA processes

$$V(k) \to \left(\sum_{j=0}^{\infty} \psi_j^2\right) \sigma_\epsilon^2 = \sigma_X^2 \quad \text{as } k \to \infty.$$

   So for large k, $V(k) \approx \sigma_X^2$.

3. The above results can be generalised to *non-stationary* processes i.e. ARIMA(p, d, q) with $d = 1, 2, \ldots$. In that case, the coefficients ψ_j's do not tend to zero, and so $V(k)$ diverges as k increases.

4. One-step-ahead forecast errors are *independent* (being single ϵ terms). However for other lead times, the errors are *correlated* (generally positively). The forecast errors tend to be of the same sign and so the forecasts tend to be all too high or all too low.

7. Estimation for Univariate Linear Processes

We will now discuss the estimation of

- the mean $E(X_t) = \mu$,
- the autocovariances $\gamma(k) = E[(X_t - \mu)(X_{t+k} - \mu)]$, and
- the parameters,

under the assumption that X_t are an ARIMA(p, d, q) process.

The data will be one realisation $x_1, ..., x_n$ of X_t, which will be called a time series.

Furthermore, the selection of the unknown model and the application of the estimated model for forecasting will also be discussed.

We will mainly consider the so-called *large sample properties* of an estimator, based on the assumption that we have a relatively long time series.

7.1. Estimation of μ

The expected value $\mu = E(X_t)$ can be estimated by the sample mean

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{50}$$

If $X_t = \epsilon_t$ are i.i.d. with $E(X_t) = \mu$ and $\text{Var}(X_t) = \sigma_X^2$, we have $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{1}{n}\sigma_X^2$ so that

$$\lim_{n\to\infty} \mathrm{E}\left[(\hat{\mu} - \mu)^2\right] = 0. \tag{51}$$

For an estimator $\hat{\theta}$ of an unknown parameter θ, the quantity

$$\mathrm{E}\left[[\hat{\theta} - \theta]^2\right] = \left(\mathrm{E}\left[\hat{\theta}\right] - \theta\right)^2 + \text{Var}(\hat{\theta})$$

is called the **mean squared error (MSE)** of $\hat{\theta}$.

If MSE($\hat{\theta}$)$\to$ 0 as $n \to \infty$, then $\hat{\theta}$ is said to be **consistent**, denoted by $\hat{\theta} \to \theta$, as $n \to \infty$. In the i.i.d. case $\bar{X}$ is consistent.

7.2. Properties of $\bar{X}$

If X_t is stationary, then $\bar{X}$ defined above is clearly unbiased, i.e., $E(\bar{X}) = \mu$. Hence $E[\bar{X} - \mu]^2 = \text{Var}(\bar{X})$.

In this section we will give some important properties of $\bar{X}$ as an estimator of μ. Proofs are given in Appendix B.

Theorem 7.1 *Let $\{X_t; t = 1, 2, ...\}$ be a time series satisfying*

$$\lim_{t\to\infty} E(X_t) = \mu,$$

$$\lim_{t\to\infty} Cov(\bar{X}, X_t) = 0,$$

where $\bar{X}$ is as defined before. Then

$$\lim_{t\to\infty} E[(\bar{X} - \mu)^2] = 0.$$

The conditions of Theorem 7.1 mean that $E(X_t)$ is asymptotically a constant, and any single observation does not dominate the covariance. Note that stationarity is not required.

Theorem 7.2 *Let $\{X_t\}$ be a stationary time series with mean μ and autocovariances $\gamma(k)$ such that $\gamma(k) \to 0$ as $k \to \infty$. Then $\bar{X}$ is a consistent estimator of the mean μ.*

Example: For any causal stationary ARMA process X_t we have $\gamma(k) \to 0$. Hence, $\bar{X} \to \mu$ as $n \to \infty$.

Example: Let Z be a Bernoulli random variable with distribution $P(Z = 1) = P(Z) = 0 = 0.5$. Define

$$X_t = \text{sign}(Z - 0.5) = \begin{cases} 1, & \text{for } Z = 1, \\ -1, & \text{for } Z = 0, \end{cases} \qquad t = 1, 2, ...$$

It is easy to show that $\{X_t\}$ is a *stationary* process with zero mean and $\gamma(k) \equiv 1$ for all k, i.e. $\gamma(k) \not\to 0$ as $k \to \infty$. For this process we have either $\bar{X} \equiv 1$ (for $z = 1$) or $\bar{X} \equiv -1$ (for $z = 0$). None of them is equal to or converges to $\mu = 0$.

Theorem 7.3 *Assume that $\{X_t\}$ is a stationary time series with absolutely summable autocovariances $\gamma(k)$. Then $\bar{X} \to \mu$, as $n \to \infty$. Furthermore,*

$$\lim_{n\to\infty} n\, Var(\bar{X}) = \sum_{k=-\infty}^{\infty} \gamma(k).$$

The asymptotic variance of $\bar{X}$ is larger than that of the sample mean of i.i.d. random variables Y_t with the same variance as X_t, if $\sum_{k=-\infty}^{\infty} \gamma(k) > \gamma(0)$, and smaller if $\sum_{k=-\infty}^{\infty} \gamma(k) < \gamma(0)$.

For a causal stationary ARMA(p, q) process, the above result reduces to

$$\text{Var}(\bar{X}) \approx \frac{\sigma_\epsilon^2}{n} \left(\frac{\sum_{j=0}^{q} \psi_j}{1 - \sum_{i=1}^{p} \phi_i} \right)^2. \tag{52}$$

Example: Let $x_1, x_2, ..., x_{400}$ be an observed time series following the theoretical model

$$X_t - \mu = 0.5(X_{t-1} - \mu) + 0.3(X_{t-2} - \mu) + \epsilon_t,$$

where ϵ_t are i.i.d. $N(0, \sigma_\epsilon^2)$ random variables. Calculate the asymptotic variance of the sample mean $\bar{x}$. Furthermore, assume that y_t, $t = 1, 2, ..., 400$, are i.i.d. random variables with $\text{Var}(Y_t) = \text{Var}(X_t)$ and unknown mean. Compare $\text{Var}(\bar{y})$ with the asymptotic variance of $\bar{x}$, where $\bar{y}$ is the sample mean of y_t.

Solution: We have

$$\text{Var}(\bar{x}) \doteq \frac{\sigma_\epsilon^2}{400}(1 - 0.5 - 0.3)^{-2} = \frac{25}{400}\sigma_\epsilon^2 = 0.0625\sigma_\epsilon^2.$$

Using the formula for $\gamma(0)$ of an AR(2) model, we obtain

$$\begin{aligned} \text{Var}(Y_t) &= \text{Var}(X_t) = \gamma(0) = \frac{(1-\phi_2)\sigma_\epsilon^2}{(1+\phi_2)[(1-\phi_2)^2 - \phi_1^2]} \\ &= \frac{(1-0.3)\sigma_\epsilon^2}{(1+0.3)[(1-0.3)^2 - 0.5^2]} = \frac{0.7}{1.3 * 0.24}\sigma_\epsilon^2 = 2.244\sigma_\epsilon^2. \end{aligned}$$

We have $\text{Var}(\bar{y}) = \text{Var}(Y_t)/400 = 0.0056\sigma_\epsilon^2$, i.e., asymptotically, $\text{Var}(\bar{x}) \geq 10\text{Var}(\bar{y})$.

The following theorem provides a CLT for a general linear process, which is a linear filter of an i.i.d. white noise ϵ_t with $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$.

Theorem 7.4 *Let X_t be a causal stationary process with MA(∞) representation*

$$X_t = \sum_{j=0}^{\infty} \alpha_j \epsilon_{t-j}$$

with $\sum_{j=0}^{\infty} |\alpha_j| < \infty$, $\sum_{j=0}^{\infty} \alpha_j \neq 0$, and where the ϵ_t are i.i.d. random variables with $E(\epsilon_t) = 0$ and Var$(\epsilon_t) = \sigma_\epsilon^2$. Then

$$\sqrt{n}\bar{X} \to_D N(0, V), \tag{53}$$

where

$$V = \sum_{k=-\infty}^{\infty} \gamma(k) = \left(\sum_{j=0}^{\infty} \alpha_j\right)^2 \sigma_\epsilon^2.$$

The sign $\to_D$ means *convergence in distribution.*

Note that Theorem 7.4 holds for all ARMA processes.

We can give a confidence interval for μ of an ARMA model.

Since $\sqrt{n}\bar{X}$ tends to a normal distribution, $\bar{X}$ is called $\sqrt{n}$ convergent.

In the general case with $E(X_t) = \mu \neq 0$, we have

$$\sqrt{n}(\bar{X} - \mu) \to_D N(0, V), \tag{54}$$

where V is the same as in (53).

The assumption $\sum_{j=0}^{\infty} \alpha_j \neq 0$ is necessary.

Example: Let $X_t = \epsilon_t - \epsilon_{t-1}$. Now we have $\sum_{j=0}^{\infty} \alpha_j = 1 - 1 = 0$ and hence the results of Theorem 7.4 do not hold for this X_t.

Example: Continue the example about the variances of $\bar{x}$ and $\bar{y}$. Assume that there $\sigma_\epsilon^2 = 1$, we have the standard deviations of $\bar{x}$ and $\bar{y}$ are about 0.25 and 0.075, respectively. For simplicity, we can use the standard normal quantile $Z_{0.025} = 1.96 \doteq 2$ to calculate the approximate 95% confidence interval. Now an approximate 95% confidence interval, e.g. for μ_X, is simply $\bar{x} \pm 2 \times \text{SD}_{\bar{X}}$. Assume that we obtained $\bar{x} = 10.5$ and $\bar{y} = 15.15$ from the data. Then the approximate 95% confidence intervals are $\mu_X \in [10, 11]$ and $\mu_Y \in [15, 15.30]$. The length of the former is more than three times of that of the latter.

Following the CLT $\sqrt{n}(\bar{X} - \mu) \to_D N(0, V)$ we have, asymptotically,

$$\frac{\bar{X} - \mu}{\sqrt{\frac{V}{n}}} \sim N(0, 1),$$

where $\sqrt{V/n}$ is the asymptotic standard deviation of $\bar{X}$. For an ARMA model we have simply

$$V = \sigma_\epsilon^2 \left(\frac{\sum_{i=0}^{q} \psi_j}{1 - \sum_{i=1}^{p} \phi_i} \right)^2.$$

This means, for any (upper) normal quantile $Z_{\alpha/2}$ we have, with about $(1 - \alpha)$ cover probability,

$$-Z_{\alpha/2}\sqrt{\frac{V}{n}} \leq \bar{X} - \mu \leq Z_{\alpha/2}\sqrt{\frac{V}{n}},$$

or equivalently,

$$\mu \in \left[\bar{X} - Z_{\alpha/2}\sqrt{\frac{V}{n}}, \bar{X} + Z_{\alpha/2}\sqrt{\frac{V}{n}} \right].$$

Example: Let $x_1, x_2, ..., x_{900}$ be an observed time series following the theoretical model

$$X_t - \mu_X = -0.8\epsilon_{t-1} + 0.3\epsilon_{t-2} + \epsilon_t,$$

where ϵ_t are i.i.d. $N(0, \sigma_\epsilon^2)$ random variables with $\sigma_\epsilon^2 = 9$. Furthermore, assume that y_t, $t = 1, 2, ..., 900$, are i.i.d. random variables with $\text{Var}(Y_t) = \text{Var}(X_t)$ and unknown mean.

Assume we have $\bar{x} = 35.25$ and $\bar{y} = 25.75$. Calculate $\text{Var}(\bar{x})$ asymptotically. Calculate $\text{Var}(\bar{y})$ and compare it with $\text{Var}(\bar{x})$. And then construct the approximate 95% confidence intervals of μ_X and μ_Y, and compare them with each other.

Solution: X_t is MA(2). We have

$$\text{Var}(\bar{x}) \doteq \frac{\sigma_\epsilon^2}{n}(1 - 0.8 + 0.3)^2 = \frac{0.25 \cdot 9}{900} = 0.0025.$$

and $\gamma(0) = (1 + \psi_1^2 + \psi_2^2)\sigma_\epsilon^2 = 1.73\sigma_\epsilon^2$ and hence $\text{Var}(\bar{y}) = \text{Var}(Y_t)/n = \gamma(0)/900 = 0.0173$.

Asymptotically, $\text{Var}(\bar{x}) \doteq 0.1445\text{Var}(\bar{y})$ and $\text{SD}_{\bar{X}} = 0.05$, $\text{SD}_{\bar{Y}} = 0.1315$.

The approximate 95% confidence intervals are:

$$\mu_X \in [\bar{x} - 2\text{SD}_{\bar{X}}, \bar{x} + 2\text{SD}_{\bar{X}}] = [35.15, 35.35]$$

and

$$\mu_Y \in [\bar{y} - 2\text{SD}_{\bar{Y}}, \bar{y} + 2\text{SD}_{\bar{Y}}] = [25.487, 26.013].$$

In this example, the asymptotic variance of $\bar{x}$ is about 14.5% of that of $\bar{y}$ for an i.i.d. random variable with the same variance. Hence, the asymptotic standard deviation of $\bar{x}$ is also much smaller than that of $\bar{y}$. Consequently, the confidence interval of μ_X is much shorter than that of μ_Y.

This example provides a case where the estimation of the unknown mean in dependent data is more accurate than the estimation in independent data.

7.3. Estimation of the Autocorrelation Function

Given a stationary process with mean μ and autocovariances $\gamma(k)$, two reasonable estimators of $\gamma(k)$ are (sample size n):

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X}) \tag{55}$$

and

$$\tilde{\gamma}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X}) \tag{56}$$

for $k = 0, 1, ..., n-1$. For $k = -(n-1), ..., -1$ we define $\hat{\gamma}(k) = \hat{\gamma}(-k)$ and $\tilde{\gamma}(k) = \tilde{\gamma}(-k)$. Note that $\hat{\gamma}(0) = \tilde{\gamma}(0)$.

Remark: If μ is known, we can use μ instead of $\bar{X}$ in the above definitions. Now the error caused by $\bar{X}$ is avoided. However, it can be shown that, for fixed k, the asymptotic properties of $\hat{\gamma}(k)$ or $\tilde{\gamma}(k)$ are the same for cases with known or unknown μ.

The autocorrelations can be estimated by

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} \tag{57}$$

or

$$\tilde{\rho}(k) = \frac{\tilde{\gamma}(k)}{\tilde{\gamma}(0)}. \tag{58}$$

For $k = 0$ we have $\hat{\rho}(0) = \tilde{\rho}(0) \equiv 1$. Hence, for estimating $\rho(k)$ we only need to discuss the properties of these estimators with $k \neq 0$.

Which of these two estimators should be used?

It might appear that $\tilde{\gamma}(k)$ is more appropriate than $\hat{\gamma}(k)$ because we just have $n - k$ product terms in the sum. However,

- MSE($\hat{\gamma}(k)$) <MSE($\tilde{\gamma}(k)$) (in general)
- $\tilde{\gamma}(k)$ is not necessarily positive semi-definite but $\hat{\gamma}(k)$ is always positive semi-definite.

For these reasons, $\hat{\gamma}(k)$ is often used in a statistical software.

In addition, $\hat{\gamma}(k)$ also has a clear disadvantage, i.e.,

$$\sum_{k=-(n-1)}^{n-1} \hat{\gamma}(k) = 0$$

for any data set, no matter what the process is. This is a restriction introduced by the definition and is not a property of the underlying process. Hence, only $\hat{\gamma}(k)$ with relatively small lags should be estimated and used. Also the estimator $\tilde{\gamma}(k)$ should only be calculated for relatively small ks. A rule of thumb is that the maximal lag used, say m, should be negligible compared to n, e.g., $m = \sqrt{n}$.

Remark: If k is negligible compared to n, then the difference between $\hat{\gamma}(k)$ and $\tilde{\gamma}(k)$ is small.

Moreover, for fixed k, $\tilde{\gamma}(k)$ and $\tilde{\rho}(k)$ have the same asymptotic properties as $\hat{\gamma}(k)$ and $\hat{\rho}(k)$ respectively.

Theorem 7.5 *Let the time series* $\{X_t\}$ *be defined by*

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \alpha_j \epsilon_{t-j},$$

where the sequence $\{\alpha_j\}$ *is absolutely summable and* ϵ_t *are i.i.d.* $(0, \sigma_\epsilon^2)$ *innovations with* $E(\epsilon_t^4) = \eta\sigma_\epsilon^4$. *Then, for fixed* $k \geq h \geq 0$

$$E\{\hat{\gamma}(k) - \gamma(k)\} = -\frac{|k|}{n}\gamma(k) - \frac{n - |k|}{n} Var(\bar{X}) + O(n^{-2}) \tag{59}$$

and

$$\begin{aligned}\lim_{n\to\infty}\frac{n^2}{n-k}&Cov\{\hat{\gamma}(k),\hat{\gamma}(h)\}\\ &=(\eta-3)\gamma(k)\gamma(h)\\ &\quad+\sum_{j=-\infty}^{\infty}[\gamma(j)\gamma(j-k+h)+\gamma(j+h)\gamma(j-k)]\end{aligned}\tag{60}$$

Remark: $O(n^{-2})$ denotes a series of the order n^{-2}. For two series a_i and b_i, we say a_i is of the same order as b_i, denoted by $a_i = O(b_i)$, if $\lim_{i\to\infty} a_i/b_i$ tends to some non-zero constant. And if $\lim_{i\to\infty} a_i/b_i \to 0$, we say that a_i is of a smaller order than b_i, denoted by $a_i = o(b_i)$.

Remark What we can learn from the above results are: For estimating μ, the existence of the variance is often required. And for estimating $\gamma(k)$, the existence of the fourth order moments is often required. Note that, if ϵ_t are independent $N(0,\sigma_\epsilon^2)$ random variables, then $(\eta-3)\gamma(k)\gamma(h)=0$, since $\eta=3$ for a normal distribution.

Theorem 7.6 *Under the assumptions of Theorem 7.5 and for fixed $i \geq j > 0$, the following results hold:*

$$E\{\hat{\rho}(i)\} = 1 - \frac{i}{n} + \frac{i}{n}\rho(i) + O(n^{-2})$$

and

$$Cov\{\hat{\rho}(i),\hat{\rho}(j)\} = w_{ij}/n,\tag{61}$$

where w_{ij} are some constants depending on $\rho(k)$.

Define $\rho(k) = (\rho(1),...,\rho(k))'$ and $\hat{\rho}(k) = (\hat{\rho}(1),...,\hat{\rho}(k))'$. The following theorem shows that $\hat{\rho}(h)$, $h = 1, 2, ..., k$, for fixed $k \geq 1$ are asymptotically normal.

Theorem 7.7 *Under the assumptions of Theorem 7.5 and for fixed $k \geq 1$, we have*

$$\hat{\rho} \to_D N(\rho, n^{-1}W),$$

where $W = (w_{ij})$ with w_{ij} defined above.

Example: Let $X_t = \{\epsilon_t\}$ be white noise with $E(\epsilon_t) = 0$ and $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$ ($\rho(k) = 0$ for $k \neq 0$)

$$w_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.}\end{cases}$$

$\hat{\rho}(1), \hat{\rho}(2), ..., \hat{\rho}(k)$, are asymptotically independent with mean zero and variance $\frac{1}{n}$. Therefore, 95% of the sample autocorrelations should lie between the bounds

$$\pm\frac{1.96}{\sqrt{n}} \doteq \pm\frac{2}{\sqrt{n}}.$$

These are the so-called $\pm 2/\sqrt{n}$ confidence bands given on a correlogram in R for the sample ACF, which shows empirically, whether the underlying process could be white noise or not.

Remark: Note that this results is obtained under the i.i.d. assumption. It should be noticed that, if more than 5% of the $\hat{\rho}(k)$ lie outside these bounds, then we can say that the underlying process is possibly not an independent white noise process. However, if more than 95% of $\hat{\rho}(k)$ lie between these two bounds, we cannot say that the process is i.i.d., because the second order properties of an i.i.d. process and an uncorrelated white noise process are the same.

8. Estimation of the ARMA Model

In this section we will discuss the estimation of the unknown parameters of an ARMA model.

At first it is assumed that the orders p and q are known. For an ARMA(p, q) model with

$$\phi(B)X_t = \psi(B)\epsilon_t \tag{62}$$

the unknown parameters are $\theta = (\sigma_\epsilon^2; \phi_1, ..., \phi_p; \psi_1, ..., \psi_q)$. More advanced analysis can be performed to check whether the normal distribution assumption is appropriate or not.

8.1. Estimation for an AR(1) Model

$$X_t = \phi X_{t-1} + \epsilon_t, \tag{63}$$

where $|\phi| < 1$ and ϵ_t is a white noise process.

We have two unknown parameters σ_ϵ^2 and ϕ.

Note that $\phi = \rho(1)$. Hence, an estimator of $\rho(1)$ will also be an estimator of ϕ.

Given observations $x_1, ..., x_n$, the simplest estimator of ϕ is

$$\hat{\phi} = \hat{\rho}(1) = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{64}$$

If we assume that $E(X_t) = 0$, the use of $\bar{x}$ is not necessary and $\hat{\phi}$ is given by

$$\hat{\phi} = \hat{\rho}(1) = \frac{\sum_{i=1}^{n-1} x_i x_{i+1}}{\sum_{i=1}^{n} x_i^2} \tag{65}$$

The properties of $\hat{\phi}$ are the same as discussed in the last section. In particular, $|\hat{\phi}| \leq 1$ (why?), and under additional conditions it is $\sqrt{n}$-consistent and asymptotically normally distributed.

Example: AR(1) model with zero mean: From a time series $x_1, ..., x_{400}$ we have obtained

$$\sum_{i=1}^{400} x_i^2 = 1012.74 \text{ and } \sum_{i=1}^{399} x_i x_{i+1} = 798.15.$$

Then we have $\hat{\phi} = \hat{\rho}(1) = 798.15/1012.74 = 0.788$. The fitted model is

$$X_t = 0.788 X_{t-1} + \epsilon_t.$$

This example is indeed calculated using a realisation following the AR(1) model given by $X_t = 0.8X_{t-1} + \epsilon_t$, where ϵ_t are i.i.d. $N(0, 1)$ random variables.

The estimation of an AR(1) model in the case with unknown mean μ_X is similar. Now we should first calculate $\bar{x}$ (from the sum of all x_t, $\sum_{t=1}^{n} x_t$), and then calculate the two sums $\sum_{t=1}^{n}(x_t - \bar{x})^2$ and $\sum_{t=1}^{n}(x_t - \bar{x})(x_{t+1} - \bar{x})$. If information about these three sums are given, then it is enough for all further calculations.

Example: Assume that we have a time series $x_1, ..., x_{900}$ following an AR(1) model with unknown mean. From the data we obtained

$$\sum_{t=1}^{900} x_t = 27254.45 \ , \sum_{t=1}^{900}(x_t - \bar{x})^2 = 13347.46$$

and

$$\sum_{t=1}^{899}(x_t - \bar{x})(x_{t+1} - \bar{x}) = 8385.93.$$

Then

$$\bar{x} = 27254.45/900 = 30.28$$

and

$$\hat{\phi} = \hat{\rho}(1) = \sum_{t=1}^{899}(x_t - \bar{x})(x_{t+1} - \bar{x}) / \sum_{t=1}^{900}(x_t - \bar{x})^2 = 8385.93/13347.46 = 0.6283.$$

The fitted model is

$$X_t - 30.28 = 0.6283(X_{t-1} - 30.28) + \epsilon_t.$$

For an AR(1) model, the unknown innovation variance σ_ϵ^2 can also be estimated from the information given in the above examples. Let $Y_t = X_t - \mu$. Y_t is an AR(1) process with zero mean, unknown parameter ϕ and $\text{Var}(Y_t) = \text{Var}(X_t)$. We know Y_t has the MA(∞) representation

$$Y_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i},$$

from which we have

$$\text{Var}(X_t) = \text{Var}(Y_t) = \gamma(0) = \left(\sum_{i=0}^{\infty} \phi^{2i}\right) \sigma_\epsilon^2 = 1/(1-\phi^2)\sigma_\epsilon^2.$$

By definition, we have

$$\hat{\gamma}(0) = \frac{1}{n}\sum_{t=1}^{n} x_t^2.$$

The unknown innovation variance σ_ϵ^2 can hence be estimated from $\hat{\phi}$ and $\hat{\gamma}(0)$,

$$\hat{\sigma}_\epsilon^2 = \hat{\gamma}(0)(1 - \hat{\phi}^2).$$

Example: In the first example above, where it is assumed that $\mu = 0$, we have simply

$$\hat{\gamma}(0) = \frac{1}{n}\sum_{t=1}^{n} x_t^2 = 1012.74/400 = 2.532$$

and

$$\hat{\sigma}_\epsilon^2 = (1 - \hat{\phi}^2)\hat{\gamma}(0) = 2.532(1 - 0.788^2) = 0.960.$$

Example: In the second example above, we have

$$\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 = 13347.46/900 = 14.831$$

and

$$\hat{\sigma}_\epsilon^2 = (1 - \hat{\phi}^2)\hat{\gamma}(0) = 14.831(1 - 0.6283^2) = 8.976.$$

Furthermore, we have $\text{Var}(\bar{x}) \doteq \frac{8.976}{900}(1 - 0.6283)^{-2} = 0.0722$. And $\text{SD}_{\bar{x}} \doteq 0.27$. The approximate 95%-CI of μ is $\mu \in [30.28 - 2 * 0.27, 30.28 + 2 * 0.27] = [29.74, 30.82]$.

Remark: Denote by $Y_t = X_t - \mu$ and $X_{1t} = X_{t-1}$. Then we have, for $t = 2, ..., n$,

$$Y_t = \phi X_{1t} + \epsilon_t$$

We have

$$E[Y_t|(X_{1t} = x_{t-1})] = \phi x_{t-1}.$$

That is we obtain a linear regression function $f(x_{1t}) = a + bx_{1t}$ with $a = 0$ and $b = \phi$, which is called a *regression through the origin* because of the fact $f(0) = 0$. The proposed $\hat{\phi}$ here is indeed an approximate least squares estimator of ϕ in the above regression model. This idea can be extended to general AR(p) models.

8.2. The Yule-Walker Estimators of an AR(p) Model

For an AR(p) Y with unknown mean μ we have

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \epsilon_t. \tag{66}$$

Let $X_t = Y_t - \mu$, We obtain

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t. \tag{67}$$

Remark: This is a linear regression model (through the origin). Although this is not a usual regression model, because the *regressors* X_{t-1}, ..., X_{t-p} depend strongly on each other, the unknown parameters ϕ_1, ..., ϕ_p can still be estimated using the method of *least squares* as in standard regression analysis. A solution based on such a least squares method will lead to the same result as the following **Yule-Walker estimation**.

Consider again the equations for γ of an AR(p):

$$\begin{array}{rcrcrcrcrcl}
\gamma(0) & - & \phi_1\gamma(1) & - & \phi_2\gamma(2) & - & \ldots & - & \phi_p\gamma(p) & = & \sigma_\epsilon^2, \\
 & & & & \text{and} & & & & & & \\
\gamma(1) & - & \phi_1\gamma(0) & - & \phi_2\gamma(1) & - & \ldots & - & \phi_p\gamma(p-1) & = & 0, \\
\gamma(2) & - & \phi_1\gamma(1) & - & \phi_2\gamma(0) & - & \ldots & - & \phi_p\gamma(p-2) & = & 0, \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\gamma(p) & - & \phi_1\gamma(p-1) & - & \phi_2\gamma(p-2) & - & \ldots & - & \phi_p\gamma(0) & = & 0,
\end{array}$$

The last p equations can be rewritten as

$$\underbrace{\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}}_{\gamma} = \underbrace{\begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \ldots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \ldots & \gamma(p-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \ldots & \gamma(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma(p-1) & \gamma(p-2) & \gamma(p-3) & \ldots & \gamma(0) \end{pmatrix}}_{\Gamma} \underbrace{\begin{pmatrix} \phi(1) \\ \phi(2) \\ \phi(3) \\ \vdots \\ \phi(p) \end{pmatrix}}_{\phi} \tag{68}$$

$\gamma = \Gamma\phi \Longleftrightarrow \phi = \Gamma^{-1}\gamma$

provided the inverse of Γ exists.

We now replace γ and Γ by their estimates $\hat{\gamma}$ and $\hat{\Gamma}$ to obtain our estimates for ϕ:

$$\hat{\phi} = \hat{\Gamma}^{-1}\hat{\gamma} \tag{69}$$

provided the inverse of $\hat{\Gamma}$ exists. Generally, $\hat{\Gamma}^{-1}$ exists because $\hat{\Gamma}(k)$ is positive semi-definite. Results based on (68) are called **Yule-Walker Estimators**.

Remark: If all $\hat{\gamma}(k)$ are replaced by $\hat{\rho}(k)$, the solutions will not change (why?).

Once we have estimates for γ and ϕ we obtain an estimate for σ_ϵ^2 from

$$\gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2) - \ldots - \phi_p\gamma(p) = \sigma_\epsilon^2$$

or from the explicit formulas for σ_ϵ^2 for AR(1) and AR(2) processes.

Two special cases:

1. For AR(1) with $p = 1$, there is only one equation in (69), i.e.

   $\hat{\phi}_1 = [\hat{\gamma}(0)]^{-1}\hat{\gamma}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \hat{\rho}(1)$, as proposed earlier.

2. For AR(2) with $p = 2$, by replacing all $\hat{\gamma}(k)$ with $\hat{\rho}(k)$, we have

$$\begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} = \begin{pmatrix} 1 & \hat{\rho}(1) \\ \hat{\rho}(1) & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \end{pmatrix}.$$

The solutions are

$$\hat{\phi}_1 = \hat{\rho}(1)\left(\frac{1-\hat{\rho}(2)}{1-\hat{\rho}^2(1)}\right) \text{ and } \hat{\phi}_2 = 1 - \left(\frac{1-\hat{\rho}(2)}{1-\hat{\rho}^2(1)}\right).$$

Example: From a time series with $n = 400$, $x_1, ..., x_{400}$, we have $\bar{x} = 10.06$, $\hat{\gamma}(0) = 1.6378$, $\hat{\gamma}(1) = 0.9176$ and $\hat{\gamma}(2) = 0.1969$. Assume that the theoretical model is an AR(2) with unknown mean μ. Estimate and write down the model. Find a 95%-CI for μ.

Solution: We have $\hat{\rho}(1) = 0.9176/1.6378 = 0.5603$ and $\hat{\rho}(2) = 0.1969/1.6378 = 0.1202$.

$$\hat{\phi}_1 = \hat{\rho}(1)\left(\frac{1-\hat{\rho}(2)}{1-\hat{\rho}^2(1)}\right) = 0.5603\frac{1-0.1202}{1-0.5603^2} = 0.7185$$

and

$$\hat{\phi}_2 = 1 - \frac{1-\hat{\rho}(2)}{1-\hat{\rho}^2(1)} = 1 - \frac{1-0.1202}{1-0.5603^2} = -0.2824.$$

The estimated AR(2) model is

$$X_t - 10.06 = 0.7185(X_{t-1} - 10.06) - 0.2824(X_{t-2} - 10.06) + \epsilon_t.$$

Now, insert $\hat{\phi}_1$ and $\hat{\phi}_2$ into the formula

$$\gamma(0) = \frac{(1-\phi_2)\sigma_\epsilon^2}{(1+\phi_2)[(1-\phi_2)^2 - \phi_1^2]}$$

we have $\hat{\gamma}(0) = 1.5839\sigma_\epsilon^2 \Longrightarrow \sigma_\epsilon^2 = 1.6378/1.5839 = 1.034$.

Furthermore, we have

$$\text{Var}(\bar{x}) \doteq \frac{\sigma_\epsilon^2}{400}(1-\hat{\phi}_1-\hat{\phi}_2)^{-2} = \frac{1.034}{400}(1-0.7185+0.2824)^{-2} = 0.0081.$$

Hence, $\text{SD}_{\bar{x}} \doteq 0.09$. The approximate 95% CI for μ is $\mu \in [10.06-2*0.09, 10.06+2*0.09] = [9.88, 10.24]$.

Remark: The estimation of an AR(p) model is not difficult, because it is related to a regression problem. The first $p+1$ estimates of the autocovariances $\hat{\gamma}(k)$, $k = 0, 1, ..., p$, (together with n) contain all information we need to estimate an AR(p) model using the above proposal.

8.3. Least Squares Estimators of an ARMA Model

We now describe the basic approach to estimate the parameters of an ARMA(p, q). Assume that $E(X_t) = 0$. For an AR(p) we have

$$\epsilon_t = X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p}.$$

The least squares estimators $\hat{\phi}_1$, ..., $\hat{\phi}_p$ are obtained by minimising the squared sum

$$Q := \sum_{t=p+1}^{n} (\hat{\epsilon}_t)^2 = \sum_{t=p+1}^{n} (X_t - \hat{\phi}_1 X_{t-1} - ... - \hat{\phi}_p X_{t-p})^2.$$

This approach does not apply if the model has an MA component. In that case we have

$$\epsilon_t = X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} - \psi_1 \epsilon_{t-1} - ... - \psi_q \epsilon_{t-q}, \tag{70}$$

which involves unobservable ϵ-values from the past. However, an approximate recursion can be used to carry out a least squares estimation of a general ARMA model.

The parameters $\phi_1, ..., \phi_p$ and $\psi_1, ..., \psi_q$, $\hat{\epsilon}_t$ can be approximated as follows:

1. Calculate $\hat{\mu} = \bar{x}$ and define $y_t = x_t - \bar{x}$, the centralised data again.
2. Put $\hat{\epsilon}_t = 0$ and $y_t = 0$ for $t \leq 0$, because we do not have information about them.
3. Set

$$\begin{aligned} \hat{\epsilon}_1 &= y_1 \\ \hat{\epsilon}_2 &= y_2 - \phi_1 y_1 - \psi_1 \hat{\epsilon}_1 \\ &\vdots \\ \hat{\epsilon}_t &= y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p} - \psi_1 \hat{\epsilon}_{t-1} - ... - \psi_q \hat{\epsilon}_{t-q} \end{aligned}$$

for $t > \max(p, q)$.

The effect due to the above approximation is negligible for large n.

Let

$$Q(\phi_1, ..., \phi_p; \psi_1, ..., \psi_q) = \sum_{t=1}^{n} \hat{\epsilon}_t^2. \tag{71}$$

One approach to estimate $\phi_1, ..., \phi_p$ and $\psi_1, ..., \psi_q$ is to minimise Q over all values of $\phi_1, ..., \phi_p$ and $\psi_1, ..., \psi_q$ subject to a causality constraint (i.e., the causal stationary regions of these parameters). These are the so-called approximate least squares estimators of an ARMA model. Note that, an estimated ARMA model will always be causal stationary, because parameters outside the causal stationary regions will not be considered. This is a restriction of the estimation method, which does not mean that the underlying model should certainly be a causal stationary ARMA model. It is possible to perform a hypothesis test to assess the goodness-of-fit of the ARMA model, but this is beyond the scope of this course.

Remark: Another commonly used method for estimating an ARMA model is the maximum-likelihood (ML) method, which will not be discussed in our lecture. When the distribution of X_t is normal, then the least squares method and the ML method are asymptotically equivalent.

8.4. Selection of an ARMA(p, q) Model

Assume that the underlying process X_t is an ARMA(p_0, q_0) model with unknown (p_0, q_0). The question is how the unknown parameters (p_0, q_0) can be estimated. This is the so-called model selection problem, which also occurs in other areas of statistics. As mentioned previously, the partial ACF can be used for selecting an AR model. A more general approach useful for ARMA models is to select the orders p_0 and q_0 by optimising some statistical information criteria that balances model-fit-to-data and model complexity.

Note that, for given (p, q), an ARMA(p, q) model is estimated by minimising the least squares criterion

$$Q = \sum_{t=1}^{n} \hat{\epsilon}_t^2. \tag{72}$$

However, the minimisation of the criterion Q cannot be directly used for selecting p and q because this would lead to model overfitting (i.e., $p \to \infty$ and $q \to \infty$ as n increases). For model selection it is hence necessary to penalise model complexity.

For given (p, q), denote the minimised value of Q as $Q(p, q)$. Then define

$$\hat{\sigma}_{p,q}^2 = \frac{Q(p, q)}{n}. \tag{73}$$

Note that, Q is the squared sum of $\hat{\epsilon}_t$. Hence, $\hat{\sigma}_{p,q}^2$ is in fact an estimator of σ_ϵ^2. Two *information criteria* based on $\hat{\sigma}_{p,q}^2$ that are widely used in the literature are

1. the AIC (Akaike's Information Criterion, Akaike, 1978)

$$\text{AIC}(p, q) = \ln \hat{\sigma}_{p,q}^2 + 2\frac{p+q}{n}, \tag{74}$$

2. the BIC (Bayesian Information Criterion, Akaike, 1977, Schwarz, 1978)

$$\text{BIC}(p, q) = \ln \hat{\sigma}_{p,q}^2 + \ln(n)\frac{(p+q)}{n}. \tag{75}$$

The penalty term in the BIC is larger than that in the AIC. Hence, the orders selected following the BIC are not larger than those selected following the AIC.

The main difference between these two methods is that the estimates $\hat{p}$ and $\hat{q}$ related to BIC are consistent estimators of p_0 and q_0 (i.e. they tends to p_0 and q_0 respectively as $n \to \infty$), while those related to AIC are not consistent. For this reason we prefer to use BIC.

Simulated examples: Three realisations are generated following the AR(4) model

$$X_t = 0.4X_{t-1} + 0.2X_{t-2} + 0.0X_{t-3} - 0.3X_{t-4} + \epsilon_t$$

with i.i.d. $N(0, 1)$ ϵ_t. Table 4 shows the estimated BIC's for $p = 0, 1, ..., 8$, where $\hat{p}$ with minimal BIC is indicated with a small star.

The estimated models obtained from these three time series using $\hat{p}$ as the AR order are:

p	0	1	2	3	4	5	6	7	8
1. S.	0.446	0.208	0.223	0.224	0.137*	0.158	0.179	0.188	0.201
2. S.	0.233	0.115	0.119	0.133	0.091*	0.122	0.153	0.181	0.207
3. S.	0.548	0.272	0.272	0.293	0.183*	0.211	0.237	0.268	0.297

Table 4: BICs and $\hat{p}$ for the three simulated time series

1. $X_t = 0.407X_{t-1} + 0.254X_{t-2} - 0.027X_{t-3} - 0.323X_{t-4} + \epsilon_t$
2. $X_t = 0.326X_{t-1} + 0.111X_{t-2} - 0.029X_{t-3} - 0.264X_{t-4} + \epsilon_t$
3. $X_t = 0.462X_{t-1} + 0.142X_{t-2} + 0.088X_{t-3} - 0.358X_{t-4} + \epsilon_t$.

8.5. Selection of d in an ARIMA(p, d,q) Model

Assume that now the time series X_t follows an ARIMA(p, d, q) with $d \in \{0, 1\}$. We have also to determine whether X_t is integrated (with $d = 1$) or not (with $d = 0$). This can be solved in the following way.

1. Set $d = 0$. Estimate ARMA models from the original data $x_1, ..., x_n$ and select $\hat{p}_0$, $\hat{q}_0$ using BIC. We obtain the minimised BIC in this case, denoted by BIC(0).
2. Set $d = 1$. Let $y_t = \Delta x_t = x_t - x_{t-1}$ for $t = 2, ..., n$. Estimate ARMA models from the $y_2, ..., y_n$ and select p_1, q_1 using BIC. We obtain the minimised BIC in this case, denoted by BIC(1).
3. Compare the two values of BIC. Select $d = 0$ if BIC(0)<BIC(1) and use the ARIMA($\hat{p}_0$, 0, $\hat{q}_0$) (i.e. the ARMA($\hat{p}_0$, $\hat{q}_0$)) model. Otherwise select $d = 1$ and use the ARIMA($\hat{p}_1$, 1, $\hat{q}_1$) model.

8.6. Examples of Model Selection

The following two examples present some time series data and its differences, together with relevant sample ACFs and partial ACFs. In each example, we consider the question of how to choose appropriate values of p, d and q such that the original time series data may be modelled as an ARIMA(p, d, q) process.

Example 1

Consider the plots in Figure 14, which show a time series, its first two differences and the sample ACF of each of these.

The time series plot of the original data clearly shows a trend, and the corresponding sample

ACF tails off relatively slowly. Hence, the original time series is not stationary. Arguably, there is little trend present in the time series plot of the first differenced data, but the corresponding ACF still tails off slowly, indicating that these first differences are not stationary either. However, when we difference again, the second difference of our time series shows no obvious trend in the time series plot, and the ACF no longer decays slowly. We conclude that these second differences are stationary. Since we had to difference twice to reach stationary data, a model with $d = 2$ would be suitable here.

Looking at the ACF of the second differenced data, it appears to cut off after lag 3, indicating that an MA(3) model would be good for these second differences (i.e., $p = 0$ and $q = 3$). This is the same as an ARIMA(0,2,3) model for the original time series.

Example 2

Consider the plots in Figure 15, which show a time series, its first difference and the sample ACF and partial ACF of each of these.

The time series plot and ACF of the original data clearly indicate it is not stationary. Using the time series plot and the ACF, the differenced data do appear to be stationary. Hence $d = 1$ seems appropriate.

The ACF of the differenced data seems to cut off at lag 5, which would suggest $p = 0$ and $q = 5$ (i.e., an MA(5) model for the first differenced data). On the other hand, the partial ACF of the differenced data seems to cut off at lag 2. This would suggest an alternative model with $p = 2$ and $q = 0$ (i.e, an AR(2) model for the first differenced data). On the grounds that fewer parameters are generally preferable in a statistical model, we would begin by choosing the second of these options, an ARIMA(2,1,0) model for the original data.

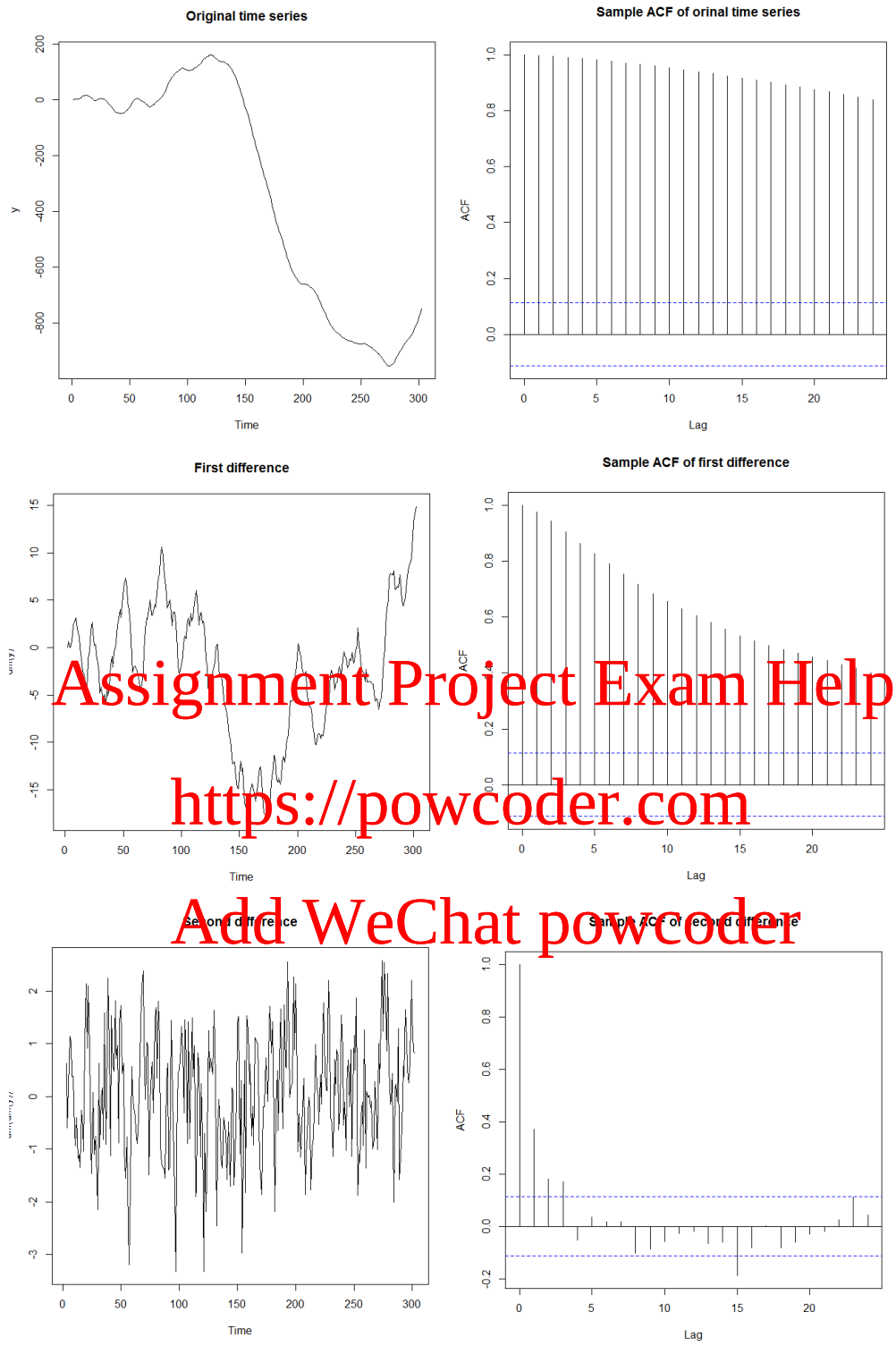


Figure 14: Some time series data, its first two differences, and the corresponding sample ACFs

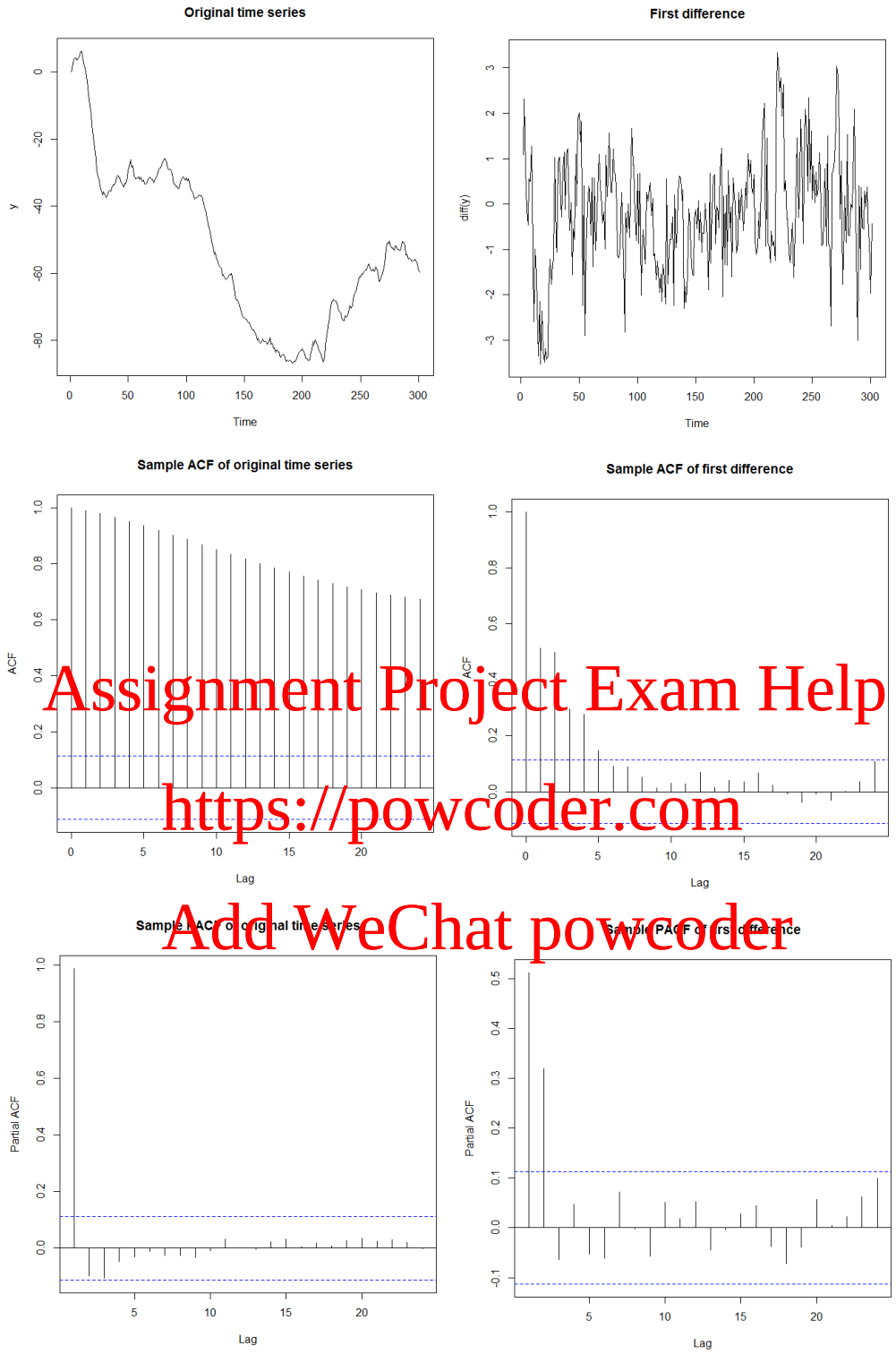


Figure 15: Some time series data and its first difference, and the sample ACFs and partial ACFs

9. Vector Time Series

See: Brockwell, P. J. and Davis, R. A.: Introduction to Time Series and Forecasting, Springer, 2002

Definition 9.1 *A multivariate time series (or vector time series) is a sequence of m-dimensional random vectors*

$$X(t) = (X_1(t), \ldots, X_m(t)), \qquad t \in \mathbb{T}$$

for a given index set $\mathbb{T}$.

Observe that X_i with $i = \{1, \ldots, m\}$ is a univariate time series.

Example: Let X_1 be the inflation rate in the UK measured by the Retail Price index (RPI), and X_2 the inflation rate in the UK measured by the Consumer Price index (CPI). See Figures 16 and 17.

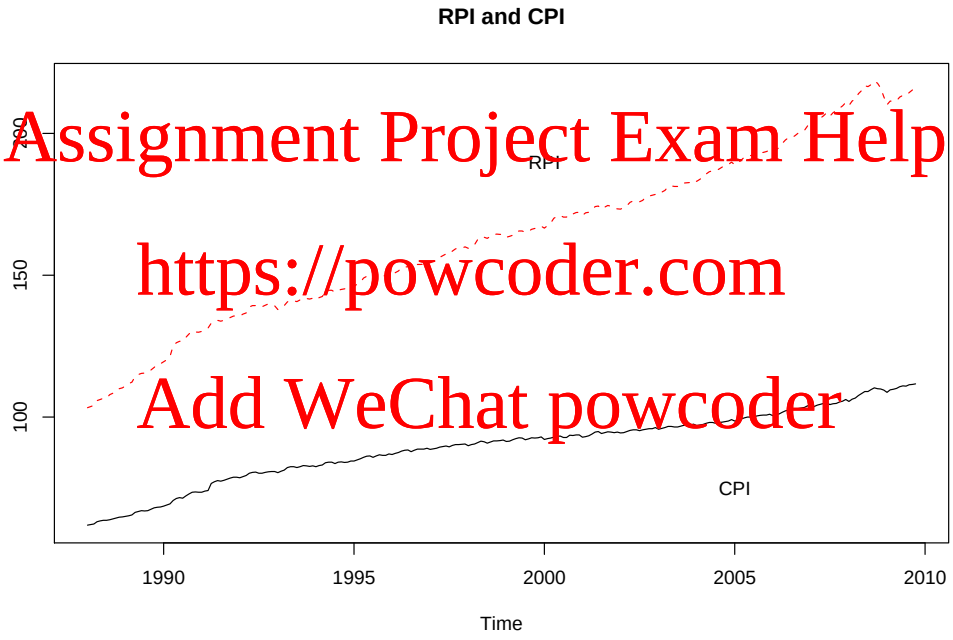


Figure 16: RPI and CPI.

We can now define the **second order properties** of X:

- the mean of $X(t)$:

$$\mu(t) = \begin{pmatrix} \mu_1(t) \\ \vdots \\ \mu_m(t) \end{pmatrix} = \begin{pmatrix} \mathrm{E}\left[X_1(t)\right] \\ \vdots \\ \mathrm{E}\left[X_m(t)\right] \end{pmatrix} = \mathrm{E}\left[X(t)\right]$$

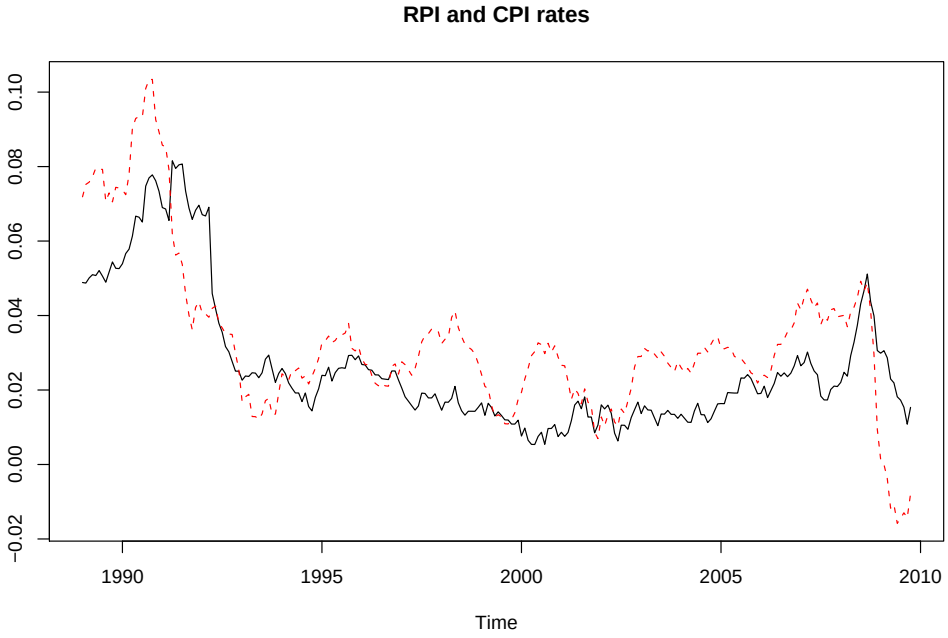


Figure 17: RPI and CPI rates

- and the autocovariance matrix:

$$\begin{aligned}\Gamma(t,s) &= \operatorname{Cov}(X(t),X(s)) \\ &= \mathrm{E}\left[\Big(X(t)-\mu(t)\Big)\Big(X(s)-\mu(s)\Big)'\right].\end{aligned}$$

$\Gamma(t,s)$ is a matrix of the form

$$\Gamma(t,s) = \begin{pmatrix} \gamma_{1,1}(t,s) & \dots & \gamma_{1,m}(t,s) \\ \vdots & & \vdots \\ \gamma_{m,1}(t,s) & \dots & \gamma_{m,m}(t,s) \end{pmatrix}$$

with

$$\gamma_{i,j}(t,s) = \operatorname{Cov}(X_i(t), X_j(s))$$

for $i,j = 1,\dots,m$.

Therefore, the diagonal elements of $\Gamma(t,s)$ are the autocovariance functions of the single dimensional time series X_i.

9.1. Stationary Multivariate Time Series

As in the one-dimensional case we call X **weakly stationary** if

i) $E(X_i(t)^2) < \infty$ for all $i \in 1, \ldots, m$, $t \in \mathbb{T}$,

ii) $EX(t) = \mu \in \mathbb{R}^m$ for all $t \in \mathbb{T}$, and

iii) $\Gamma(r, s) = \Gamma(r + t, s + t)$ for all $r, s, t \in \mathbb{T}$.

Note that μ is an m-dimensional vector.

For a weakly stationary time series we use the notation

$$\Gamma(h) = \Gamma(t + h, t)$$

and call it the **covariance matrix at lag** h.

The element $\gamma_{i,i}(h)$ of $\Gamma(h)$ is the autocovariance function of the stationary TS X_i, and $\gamma_{i,j}(h)$ is called the cross-covariance function of X_i and X_j.

We can also define the **correlation matrix at lag** h:

$$R(h) = \begin{pmatrix} \rho_{1,1}(h) & \ldots & \rho_{1,m}(h) \\ \vdots & & \vdots \\ \rho_{m,1}(h) & \ldots & \rho_{m,m}(h) \end{pmatrix} \text{ with } \rho_{i,j}(h) = \frac{\gamma_{i,j}(h)}{\sqrt{\gamma_{i,i}(0)\gamma_{j,j}(0)}}$$

where $\rho_{i,j}(h)$ is the correlation between $X_i(t)$ and $X_j(0)$.

Properties of Γ and R:

1. $\gamma_{i,j}(h) \neq \gamma_{j,i}(h)$ (in general)
2. $\Gamma(h) = \Gamma'(-h) \neq \Gamma(-h)$
3. $\rho_{i,i}(0) = 1$
4. $\rho_{i,j}(0) \neq 1$ (in general)

9.2. Multivariate ARMA Processes

We call a vector TS Z an m-dimensional **White Noise** process with covariance matrix Σ if Z is a sequence of i.i.d. random vectors, $\mathrm{E}[Z] = (0, \ldots, 0)' \in \mathbb{R}^m$ and $\Gamma(0) = \Sigma$.

Note that some authors (for example Brockwell & Davis) call Z a white noise process if Z is stationary with zero mean and covariance

$$\Gamma(h) = \begin{cases} \Sigma & \text{for } h = 0 \\ 0 & \text{otherwise} \end{cases}$$

This would be a sequence of uncorrelated identically distributed random variables.

Similarly to the definition of a one-dimensional ARMA process, the multivariate ARMA process is defined as follows:

Definition 9.2 *A multivariate Autoregressive Moving Average process of order (p, q), the ARMA(p, q) model, is defined by*

$$X(t) - \Phi_1 X(t-1) - ... - \Phi_p X(t-p) = \Psi_1 Z(t-1) + ... + \Psi_q Z(t-q) + Z(t)$$

where $\Phi_1, \ldots, \Phi_p$ and $\Psi_1, \ldots, \Psi_q$ are $m \times m$ matrices.

Causality and invertibility are also easily generalised from one-dimensional to multivariate time series (see for example Brockwell & Davis).

Of particular importance and widely used for modelling econometric data are vector autoregressive processes VAR(p), that is, $q = 0$ (no moving average term).

9.3. VAR(p)

Example: Consider the following two-dimensional VAR(1) process.

$$\begin{aligned} X(t) &= \Phi X(t-1) + Z(t) \\ \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} &= \begin{pmatrix} 0.5 & 0.2 \\ -0.5 & 0.2 \end{pmatrix} \begin{pmatrix} X_1(t-1) \\ X_2(t-1) \end{pmatrix} + \begin{pmatrix} Z_1(t) \\ Z_2(t) \end{pmatrix} \\ X_1(t) &= 0.5X_1(t-1) + 0.2X_2(t-1) + Z_1(t) \\ X_2(t) &= -0.5X_1(t-1) + 0.2X_2(t-1) + Z_2(t) \end{aligned}$$

Note that any VAR(p) process can be written as a VAR(1) process. Consider the m-dimensional VAR(p)-process

$$X(t) = \Phi_1 X(t-1) + ... + \Phi_p X(t-p) + Z(t).$$

$$\text{Define: } \xi(t) = \begin{pmatrix} X(t) \\ X(t-1) \\ \vdots \\ X(t-p+1) \end{pmatrix} \text{ and } V(t) = \begin{pmatrix} Z(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$F = \begin{pmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \cdots & \Phi_{p-1} & \Phi_p \\ I & 0 & 0 & \cdots & 0 & 0 \\ 0 & I & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I & 0 \end{pmatrix}$$

where I is the $m \times m$ identity matrix.

We then obtain

$$\xi(t) = F\xi(t-1) + V(t)$$

where $\xi(t), V(t) \in \mathbb{R}^{mp}$ and F is a $mp \times mp$ matrix.

The covariance function Γ of a VAR(1) process can be found recursively, similarly to the one-dimensional case.

$$\begin{aligned} X(t) &= \Phi X(t-1) + Z(t) \\ X(t)X(t-h)' &= \Phi X(t-1)X(t-h)' + Z(t)X(t-h)' \\ \mathrm{E}\left[X(t)X(t-h)'\right] &= \Phi \mathrm{E}\left[X(t-1)X(t-h)'\right] + \mathrm{E}\left[Z(t)X(t-h)'\right] \end{aligned}$$

In particular:

$$\begin{aligned} \Gamma(0) &= \Phi\Gamma(-1) + \mathrm{E}\left[Z(t)X(t)'\right] \\ &= \Phi\Gamma(1)' + \mathrm{E}\left[Z(t)Z(t)'\right] \\ &= \Phi\Gamma(1)' + \Sigma \\ \Gamma(1) &= \Phi\Gamma(0) \\ \Gamma(h) &= \Phi\Gamma(h-1) \end{aligned}$$

It follows that:

$$\Phi = \Gamma(1)\Gamma(0)^{-1} \text{ if } \Gamma(0)^{-1} \text{ exists.}$$

This relationship can be used for estimating the coefficient matrix Φ.

Once Φ has been estimated, we can further estimate Σ by using

$$\Gamma(0) - \Phi\Gamma(1)' = \Sigma.$$

10. Introduction to ARCH and GARCH

The purpose of this final section is to introduce two non-linear time series models, namely ARCH (*autoregressive conditional heterocedastic*) (Engle, 1982, joint winner of Nobel Prize in Economics 2003) and GARCH (*generalised* ARCH, Bollerslev, 1986), which are widely used for modelling financial time series.

10.1. Properties of Financial Time Series

Financial time series have some special properties. Some of these are not consistent with linear models such as ARMA or ARIMA. In particular

- Financial time series are generally integrated, i.e., they are not stationary but stationary when differenced. Hence, we should either study the difference series or some other transforms of them. Commonly used transformations of a financial time series X_t are the difference series $Y_t = \Delta X_t$, the series of the log-returns

$$Y_t = \Delta(\log X_t)$$

or that of the absolute returns

$$Y_t = \Delta X_t / X_{t-1}.$$

- The conditional mean of a financial time series is often independent of the past value of X, but the conditional variance is not.

- Financial returns are often uncorrelated but not independent because they exhibit non-linear dependences. For example, the quadratic returns are often clearly correlated. Due to this fact, such processes are called non-linear processes.

- The conditional distribution of financial returns may be normal but the joint distribution is usually not normal.
- Financial returns often exhibit heavy tails, i.e., the moments $\text{Var}(Y_t) < \infty$ but $E(Y_t^4) = \infty$.

Some of the above properties are illustrated in Figure 18 which shows the time series of the daily log-returns of the FTSE100 stock market index from Jan 1985 to Feb 2010. There are some clustered observations with large variances, which illustrate the so-called GARCH-effect. The ACFs of the returns and squared returns (Figure 19) show that the former seem to be uncorrelated but the latter are clearly correlated to each other. This also indicates that the return are not independent.

In 1982 Robert Engle introduced the very important ARCH model for modelling the above special properties of financial time series.

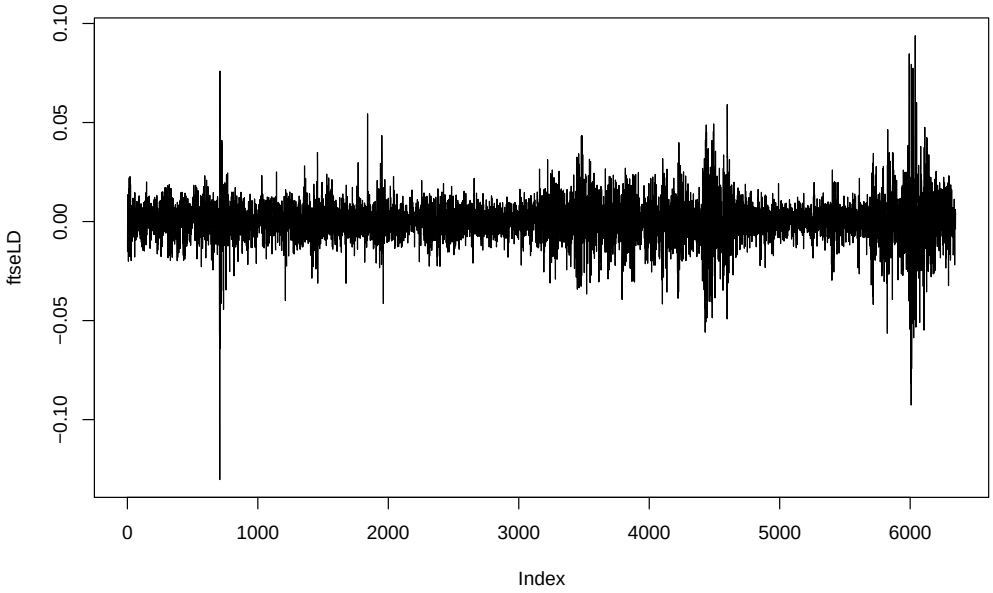


Figure 18: Daily returns of the FTSE100.

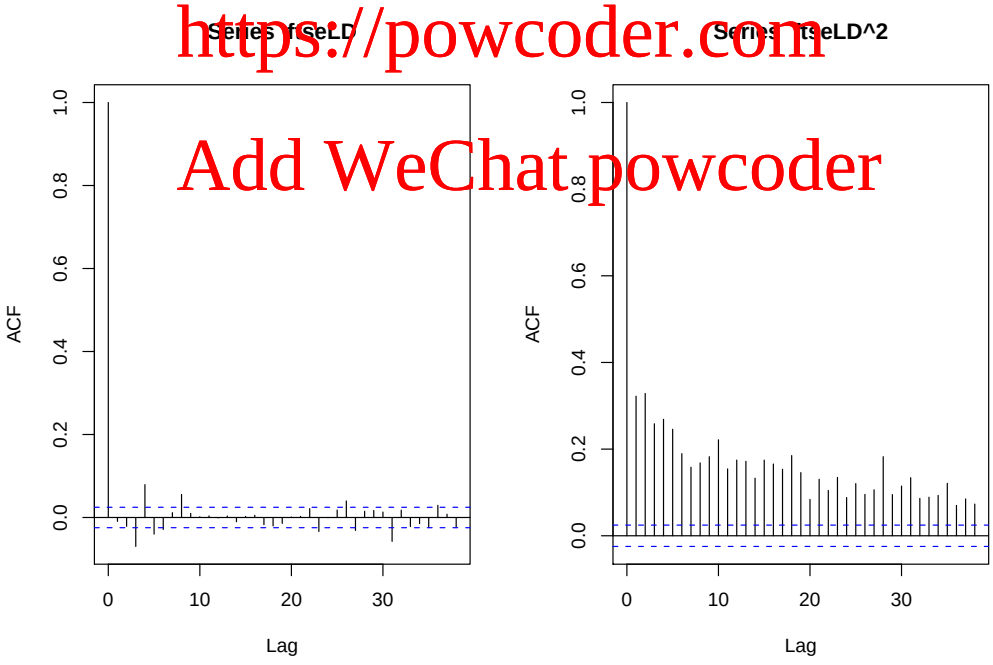


Figure 19: ACF of the daily returns of the FTSE100, and ACF of the squared daily returns.

10.2. ARCH Models

Conditional expectation of AR(p):

$$\mathrm{E}\left[X_t \,|\, X_{t-1}, \ldots, X_{t-p}\right] = \phi_1 X_{t-1} + \ldots + \phi_1 X_{t-p}$$

Conditional variance of AR(p):

$$\begin{aligned}
&\mathrm{E}\left[\left(X_t - \mathrm{E}\left[X_t \,|\, X_{t-1}, \ldots, X_{t-p}\right]\right)^2 \,|\, X_{t-1}, \ldots, X_{t-p}\right] \\
&= \mathrm{E}\left[\left(X_t - \phi_1 X_{t-1}, \ldots, \phi_1 X_{t-p}\right)^2 \,|\, X_{t-1}, \ldots, X_{t-p}\right] \\
&= \mathrm{E}\left[\epsilon^2\right] = \sigma_\epsilon^2
\end{aligned}$$

This means that the conditional mean depends on the past observations, but the conditional variance does not.

ARCH models have been introduced as an alternative. These models are about the error term ϵ_t. The particular form of the linear term is not important, and is therefore ignored in the following definition.

An ARCH model of order p (ARCH(p)) is defined by

$$Z_t = \sqrt{h_t}\epsilon_t \text{ with } h_t = \alpha_0 + \alpha_1 Z_{t-1}^2 + \ldots + \alpha_p Z_{t-p}^2$$

where ϵ is a white noise process with $\mathrm{E}\left[\epsilon_t^2\right] = 1$, and $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, .., p$. We often make the assumption that $\epsilon_t \sim N(0, 1)$.

Here $h_t > 0$ is the conditional variance of Z_t, which depends on some past observations. This is what *conditional heteroscedasticity* means. The larger the squared past observations are, the larger the conditional variance.

$Z(t)$ can now replace the white noise term in linear models or can be used directly to model a time series.

We obtain:

$$\mathrm{E}\left[Z_t | Z_{t-1}, \ldots\right] = \mathrm{E}\left[\sqrt{h_t}\epsilon_t | Z_{t-1}, \ldots\right] = \mathrm{E}\left[\sqrt{h_t} | Z_{t-1}, \ldots\right] \mathrm{E}\left[\epsilon_t\right] = 0$$

since ϵ_t is independent of h_t. This is a common property of many financial return series (the so-called Martingale property). Note, that $\mathrm{E}\left[Z(t)\right] = 0$ too.

We obtain for the covariance function of Z:

$$\begin{aligned}
\mathrm{Cov}(Z_t, Z_{t+k}) &= \mathrm{E}\left[Z_t Z_{t+k}\right] \\
&= \mathrm{E}\left[\sqrt{h_t}\epsilon_t\sqrt{h_{t+k}}\epsilon_{t+k}\right] \\
&= \mathrm{E}\left[\sqrt{h_t}\epsilon_t\sqrt{h_{t+k}}\right] \mathrm{E}\left[\epsilon_{t+k}\right] \\
&= 0
\end{aligned}$$

since ϵ_{t+k}, for $k > 0$, is independent of all other three terms and has mean zero. This is another important property of financial returns, that is, they are often uncorrelated with each other.

However, it is clear that Z_t and Z_{t+k} are not independent, because the variance of Z_{t+k} depends on the past of Z, and therefore on Z_t. An ARCH model is an example of uncorrelated random variables which are not independent.

The process is called a non-linear process because there are non-linear correlations between the observations (and no linear correlation). In particular, here Z_t^2 and Z_{t+k}^2 are correlated.

Furthermore, note that the conditional distribution of an ARCH model is normal if $\epsilon_t \sim N(0,1)$. Hence, the conditional moments of any order exist.

However, this does not mean that the *unconditional* variance of an ARCH model exists.

If $\text{Var}(Z_t) < \infty$, we have $\text{Var}(Z_t) = E(Z_t^2)$. Assume further that the process started from the infinite past. It can be shown that an ARCH process with finite variance is weakly (and strongly) stationary. We then obtain

$$\begin{aligned} \text{Var}(Z_t) = \text{E}\left[Z_t^2\right] &= \text{E}\left[h_t \epsilon_t^2\right] \\ &= \text{E}\left[\alpha_0 + \alpha_1 Z_{t-1}^2 + ... + \alpha_p Z_{t-p}^2\right] \\ &= \alpha_0 + \alpha_1 \text{Var}(Z_t) + ... + \alpha_p \text{Var}(Z_t) \end{aligned}$$

since Z is stationary. We solve this equation for $\text{Var}(Z_t)$ and obtain

$$\text{Var}(Z_t) = \frac{\alpha_0}{1 - \alpha_1 - ... - \alpha_p}. \tag{76}$$

This leads to the condition

$$0 < \text{Var}(Z_t) < \infty \iff \sum_{i=1}^{p} \alpha_i < 1.$$

If the variance exists, its quantity depends linearly on α_0, which is generally called the *scale parameter*.

Finally, note that an ARCH may have finite variance but infinite fourth order moments, which as mentioned previously is an important property of financial returns.

10.3. GARCH Models

ARCH models were generalised by Bollerslev (1986) to GARCH(p, q) models, defined by

$$Z_t = \epsilon_t \sqrt{h_t}$$

with $\epsilon_t \sim N(0,1)$ i.i.d. and

$$\begin{aligned} h_t &= \alpha_0 + \alpha_1 Z_{t-1}^2 + ... + \alpha_p Z_{t-p}^2 \\ &\quad + \beta_1 h_{t-1} + ... + \beta_q h_{t-q}, \end{aligned} \tag{77}$$

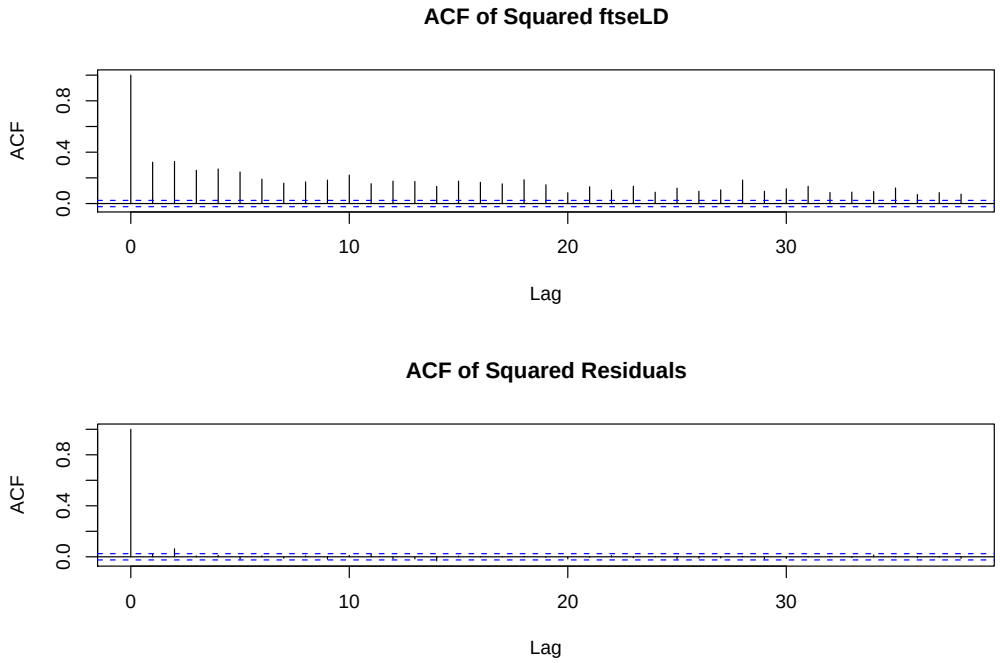


Figure 20: ACF of the squared residuals of a GARCH(1,1) model fitted to the FTSE100 returns.

where $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, p$, $\beta_j \geq 0$ for $j = 1, \ldots, q$.

A GARCH model has similar properties as those for an ARCH model, with some key differences that make the GARCH model more practically relevant.

In particular, in GARCH models the variance of Z_t depends on Z_{t-1}^2 and on the variance of Z_{t-1}, while it only depends on Z_{t-1}^2 in ARCH-models.

Figure 20 shows the acf of the squared residuals of a GARCH(1,1) model fitted to the daily returns of the FTSE100.

11. Further Reading

- Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis*. Holden-Day.
- Brockwell, P.J. and R.A. Davis (1991). *Time Series: Theory and Methods* (2nd Ed.). Springer.
- Chatfield, C. and Xing, H. (2019). *The Analysis of Tine Series: An Introduction with R* (7th Ed). CRC Press.
- Cowpertwait, P.S.P. and Metcalf, A.V. (2009). *Introductory Time Series with R*. Springer.
- Diggle, P.J. (1990). *Time Series – A Biological Introduction*. Oxford University Press.
- Fuller, W.A. (1996). *Introduction to Time Series Analysis*. John Wiley.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

A. Stationarity Triangle for an AR(2) Process

In this appendix we prove that the stationarity conditions (i)–(iii) for an AR(2) process given in Section 3.2 are equivalent to the conditions in Theorem 3.3.

We have our AR(2) process $\phi(B)X_t = \varepsilon_t$, where $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$.

One method of proving the stationarity region for this process uses the quadratic formula to give expressions for the roots of $\phi(z)$ in terms of ϕ_1 and ϕ_2; analysis of these roots gives conditions under which they both have modulus larger than 1.

Here we present an alternative proof, representing the characteristic polynomial ϕ in a different way. We then have to relate this new representation back to the original representation in terms of ϕ_1 and ϕ_2.

We write $\phi(z) = (1 + a_1 z)(1 + a_2 z) = 1 + (a_1 + a_2)z + a_1 a_2 z^2$, so that

$$a_1 + a_2 = -\phi_1 \,, \qquad a_1 a_2 = -\phi_2 \,. \tag{78}$$

(by comparing coefficients)

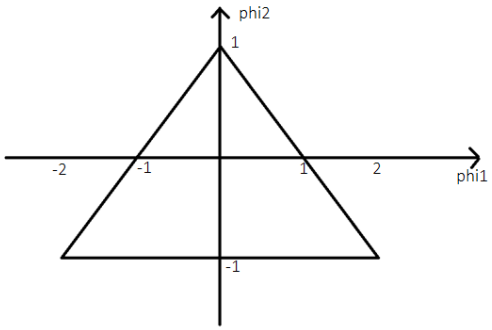
The solutions to $(1+a_1 z)(1+a_2 z) = 0$ are $z = -1/a_1$ and $z = -1/a_2$. Hence, for stationarity we need $|a_i| < 1$ for $i = 1, 2$. This corresponds to a square in the (a_1, a_2) plane, with corners at $(-1, 1)$, $(1, 1)$, $(1, -1)$ and $(-1, -1)$.

We map this square into the (ϕ_1, ϕ_2) plane using (78), by considering where the corner points of this square map to:

$$(a_1, a_2) = (-1, 1) \text{ maps to } (\phi_1, \phi_2) = (0, 1) \,.$$
$$(a_1, a_2) = (1, 1) \text{ maps to } (\phi_1, \phi_2) = (-2, -1) \,.$$

$$(a_1, a_2) = (1, -1) \text{ maps to } (\phi_1, \phi_2) = (0, 1) \,.$$
$$(a_1, a_2) = (-1, -1) \text{ maps to } (\phi_1, \phi_2) = (2, -1) \,.$$

This gives us that our square maps to the following triangular region in the (ϕ_1, ϕ_2) plane:



Since the processes represented by points inside our square in the (a_1, a_2) plane are precisely those which are stationary, so are the processes represented by points inside this triangle. It is easy to check that this triangle is defined by the required inequalities:

$$\begin{aligned} \phi_1 + \phi_2 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ |\phi_2| &< 1\,. \end{aligned}$$

B. Proofs of Results in Section 7.2

The following lemmas will be used for some of the proofs in this section.

Lemma B.1 *Let $\{a_n\}$ be a sequence of real numbers and $a \in \mathbb{R}$.*

$$\lim_{n\to\infty} a_n = a \implies \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i = a$$

Proof: By assumption, given $\epsilon > 0$, we may choose an N such that $|a_n - a| < \frac{1}{2}\epsilon$ for all $n > N$. For $n > N$, we have

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} a_i - a\right| &\leq \frac{1}{n}\sum_{i=1}^{N} |a_i - a| + \frac{1}{n}\sum_{i=N+1}^{n} |a_i - a| \\ &\leq \frac{1}{n}\sum_{i=1}^{N} |a_i - a| + \frac{1}{2}\epsilon. \end{aligned}$$

We can choose n large enough so that the first term is smaller than $\frac{1}{2}\epsilon$. The result follows. $\diamond$

Lemma B.2 *(Kronecker's lemma) If the sequence $\{a_j\}$ is such that*

$$\lim_{n\to\infty} \sum_{j=0}^{n} |a_j| < \infty,$$

then

$$\lim_{n\to\infty} \sum_{j=0}^{n} \frac{j}{n}|a_j| = 0.$$

Proof: Set, for example, $N = n^{1/3}$. Then we have

$$\sum_{j=0}^{n} \frac{j}{n}|a_j| = \sum_{j=0}^{N} \frac{j}{n}|a_j| + \sum_{j=N+1}^{n} \frac{j}{n}|a_j|,$$

where the second term tends to zero by assumption, because $N \to \infty$, and

$$\sum_{j=0}^{N} \frac{j}{n}|a_j| < \max(|a_j|)n^{-1/3} \to 0.$$

too, as $n \to \infty$. The result holds. ◇

B.1. Proof of Theorem 7.1

We have that

$$E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) + \left(\frac{1}{n}\sum_{t=1}^{n} E(X_t) - \mu\right)^2,$$

where the second term on the right hand side converges to zero by Lemma B.1. Furthermore,

$$\begin{aligned}
\text{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{j=1}^{n}\text{Cov}(X_t, X_j) \\
&= \frac{2}{n^2}\sum_{t=1}^{n}\sum_{j=1}^{t}\text{Cov}(X_t, X_j) - \frac{1}{n^2}\sum_{t=1}^{n}\text{Var}(X_t) \\
&= \frac{2}{n^2}\sum_{t=1}^{n} t\,\text{Cov}(\bar{X}_t, X_t) - \frac{1}{n^2}\sum_{t=1}^{n}\text{Var}(X_t) \\
&\leq \frac{2}{n}\sum_{t=1}^{n}|\text{Cov}(\bar{X}_t, X_t)|,
\end{aligned}$$

where $\bar{X}_t$ is the sample mean of the first t observations. By Lemma B.1 the last sum converges to zero.

B.2. Proof of Theorem 7.2

Note that for a stationary process we have $E(\bar{X}) = \mu$. Hence we only need to show

$$\lim_{n\to\infty} \text{Var}(\bar{X}) = 0.$$

$$\begin{aligned}
\text{Var}(\bar{X}) &= \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{i=1}^{n}\text{Cov}\left(X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\text{Cov}\left(X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right)\right|
\end{aligned}$$

$$\left|\text{Cov}\left(X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right)\right| = \frac{1}{n}\sum_{k=1}^{n}|\text{Cov}(X_i, X_k)| = \frac{1}{n}\sum_{k=1}^{n}|\gamma(k-i)|$$
$$\leq \frac{2}{n}\sum_{k=0}^{n}|\gamma(k)| =: a_n \to 0.$$

We obtain

$$\text{Var}(\bar{X}) \leq \frac{1}{n}\sum_{i=1}^{n} a_n = a_n \to 0$$

.

B.3. Proof of Theorem 7.3

We only need to show the second statement, for which we have

$$n\text{Var}(\bar{X}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{t=1}^{n}\gamma(t-j)$$
$$= \frac{1}{n}\sum_{k=-(n-1)}^{n-1}(n-|k|)\gamma(k)$$
$$= \sum_{k=-(n-1)}^{n-1}\gamma(k) - 2\sum_{k=1}^{n-1}\frac{k}{n}\gamma(k).$$

By Lemma B.2 the second term of the last equation tends to zero. The theorem is proved.