# MA30085 Time series

Chris Hallsworth

February 2017

# 1 Time series as stochastic processes

## 1.1 Introduction

A time series is a collection of repeated observations of a system, made sequentially through time.
Examples occur in a variety of real life applications, ranging from economics to engineering.

- Economic and financial time series: share prices on successive days, economic indexes such as FTSE 100, export totals in successive months, average incomes in successive months, company profits in successive years etc.
- Physical time series, e.g. in meteorology, marine science and geophysics: rainfall on successive days, air temperature measured on successive hours (days or months)
- Marketing time series: sales figures in successive days or weeks, monetary receipts, advertising costs and so on.
- Demographic time series (in study of population change): population of Canada measured annually, monthly birth totals in England.
- Binary processes, a special type of time series when observations can take one of only two values: in computer science, in biology (e.g. ion channel kinetics).

## 1.2 References


- Chatfield, C. (2004). The analysis of time series. 6th Edition. Chapman & Hall
- Brockwell P.J. and Davis R.A. (1991). Time series: theory and methods. Springer-Verlag
- A first course on time series analysis (2006). Online book, available at http://statistik.mathematik.uni-wuerzburg.de/timeseries/
- Diggle, P. (1990). Time series. A biostatistical introduction.
- Harvey, A. (1989). Forecasting, structural time series models and the Kalman filter.

## 1.3 Stationary stochastic processes

### 1.3.1 Definitions and some examples

Let $T \subseteq \mathbf{R}$.

**Definition 1.1** *A stochastic process is a collection of random variables* $\{X_t = X_t(w),\, t \in T\} = \{X_t,\, t \in T\}$, *defined on a probability space* $(\Omega, \mathcal{F}, \mathsf{P})$.

In this course we consider only discrete time stochastic process, i.e., $T = \mathbf{Z}$ or $T = \mathbf{Z}_+$.

**Definition 1.2** *Given* $w \in \Omega$ *the function* $X_{\{.\}}(w)$, $w \in \Omega$ *is known as a realisation or a sample path of the process* $\{X_t(w),\, t \in \boldsymbol{Z}\}$.

**Example 1.1** *A sequence* $\{Z_t,\, t \in \boldsymbol{Z}\}$ *of i.i.d. random variables is a stochastic process.*

An i.i.d. sequence with zero mean $\mathsf{E}(Z_t) = 0$) is often called a *purely random process* or *white noise.*

**Example 1.2** *A random walk. Let* $\{Z_t,\, t \in \boldsymbol{Z}_+\}$, *be a sequence of i.i.d. random variables. A random walk is a stochastic process* $\{X_t,\, t \in \boldsymbol{Z}_+ \cup \{0\}\}$, *defined as follows*

$$X_0 = 0$$

$$X_t = X_{t-1} + Z_t,\, t \geq 1.$$

**Example 1.3** *The MA(1)-process (Moving Average process of order* 1*) is defined by the equation*

$$X_t = Z_t + \beta Z_{t-1},\, t \in \boldsymbol{Z},$$

*where* $\{Z_t,\, t \in \boldsymbol{Z}\}$, *is a sequence of i.i.d. random variables and* $\beta \in \boldsymbol{R}$.

The joint distribution function $F(x_1, \ldots, x_k)$ of a random vector $(\xi_1, \ldots, \xi_k)$ is defined as follows

$$F(x_1, \ldots, x_k) = \mathsf{P}\{\xi_1 \leq x_1, \ldots, \xi_k \leq x_k\}, \quad x_j \in \mathbf{R},\, j = 1, \ldots, k.$$

**Definition 1.3** *The finite dimensional distribution functions of a stochastic process* $\{X_t,\, t \in T\}$ *are the functions* $\{F_{t_1 \ldots t_n}(x_1, \ldots, x_n),\, t_i \in T, x_i \in \boldsymbol{R}, i = 1, \ldots, n\}$ *defined as follows*

$$F_{t_1 \ldots t_n}(x_1, \ldots, x_n) = \mathsf{P}\{X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\},$$

*i.e.,* $F_{t_1 \ldots t_n}(x_1, \ldots, x_n)$ *is a joint distribution function of* $(X_{t_1}, \ldots, X_{t_n})$.

**Definition 1.4** *A stochastic process* $X_t$, $t \in T$, *is said to be strictly stationary if for any* $t_1, \ldots, t_n \in T$ *and* $\tau$ *such that* $t_1 + \tau, \ldots, t_n + \tau \in T$ *the joint distribution function of* $(X_{t_1}, \ldots, X_{t_n})$ *is the same as the joint distribution function of* $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$.

**Example 1.4** *A sequence* $\{Z_t,\, t \in \boldsymbol{Z}\}$ *of i.i.d. random variables is a strictly stationary process.*

Recall that for a a random variable $\xi$, the $k$th moment is defined to be $\mathsf{E}(\xi^k)$, and we say that the $k$th moment exists if $\mathsf{E}(|\xi|^k) < \infty$.

**Definition 1.5** *A stochastic process $X_t$, $t \in T$, is said to be a weakly stationary or second-order stationary if its first and second moments are finite and*

$$\mathsf{E}(X_t) = const,$$

*(i.e., the process mean is a constant function) and*

$$\mathsf{Cov}(X_t, X_{t+\tau}) = \mathsf{E}((X_t - \mathsf{E}X_t)(X_{t+\tau} - \mathsf{E}X_{t+\tau})) = \gamma(\tau),$$

*(i.e., the process autocovariance function depends only on lag $\tau$) for any $t$ and $\tau \in T$, such that $t + \tau \in T$.*

**Example 1.5** *A random walk*

$$X_t = X_{t-1} + Z_t,$$

*where $\{Z_t,\, t \in \boldsymbol{Z}\}$, is i.i.d. sequence with $\mathsf{E}(Z_t) = 0$, $\mathsf{E}(Z_t^2) = \sigma^2 < \infty$, $t \in \boldsymbol{Z}$, is not a weakly stationary stochastic process, but its first difference*

$$\nabla X_t = X_t - X_{t-1},\, t \geq 1,$$

*is weakly stationary.*

*To see this, use the bilinear properties of covariance and the independence of the variables $Z_i$ to note that*

$$\mathsf{Cov}(X_t, X_{t+\tau}) = \mathsf{Cov}(\sum_{i=1}^{t} Z_i, \sum_{i=1}^{t+\tau} Z_i) = \sum_{i=1}^{t} \mathsf{Cov}(Z_i, Z_i) = \sigma^2 t,$$

*which depends on $t$. The first difference is just $Z_t$, which is clearly stationary.*

For a process with finite first and second moments, strict stationarity implies weak stationarity. But by constructing a process whose variables have first and second moments that fail to be finite, it is possible to exhibit a strictly stationary process that is not weakly stationary.

**Example 1.6** *Let $\{X_t,\, t \in \boldsymbol{Z}\}$, be i.i.d. random variables with the Cauchy distribution. This process is strictly stationary by construction, because the variables are i.i.d. but it fails to be weakly stationary because the $k$th moment of the Cauchy distribution does not exist for any $k \geq 1$.*

*(NB the preceding example was only hinted at in lectures - it is not examinable.)*

Note also that weak stationarity (as the name suggests) does not imply strict stationarity.

**Example 1.7** *Consider a sequence of independent random variables $X_t, t \in \boldsymbol{Z}_+$, such that $X_t$ is uniformly distributed on $[-1, 1]$ when $t$ is odd and normally distributed with zero mean and variance $1/3$ when $t$ is even. $X_t, t \in \boldsymbol{Z}_+$, is weakly stationary, but is not strictly stationary.*

*Remark.* In the rest of the course, by a stationary stochastic process we mean a weakly stationary stochastic process.

### 1.3.2 Autocovariance and autocorrelation functions

Two main characteristics of a weakly stationary stochastic process $\{X_t,\, t \in \mathbf{Z}\}$ are the mean $\mu = \mathsf{E}(X_t)$ (constant function of $t$) and the *autocovariance* function

$$\gamma(k) = \mathsf{Cov}(X_t, X_{t+k}),\ k \in \mathbf{Z}.$$

The process *autocorrelation* function (ac.f.)

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)},\ k \in \mathbf{Z},$$

is just its autocovariance function standardized by dividing it by the variance of the process.

**Example 1.8** *$\{Z_t,\, t \in \boldsymbol{Z}\}$, a sequence of uncorrelated random variables with $\mathsf{E}(Z_t) = 0$, $\mathsf{E}(Z_t^2) = \sigma^2 < \infty$, $t \in \boldsymbol{Z}$, is a weakly stationary stochastic process with*

$$\rho(k) = \begin{cases} 1, & k = 0, \\ 0, & k \neq 0. \end{cases}$$

*To see this, note that $\gamma(k) = \mathsf{Cov}(Z_t, Z_{t+k}) = 0$ for $k \neq 0$, because the variables $Z_t$ are uncorrelated, while $\gamma(0) = \mathsf{Cov}(Z_t, Z_t) = \sigma^2$. Normalising gives the required result.*

**Example 1.9** *The MA(1)-process is weakly stationary stochastic process with*

$$\rho(k) = \begin{cases} 0, & |k| > 1, \\ 1, & k = 0, \\ \beta/(1+\beta^2), & k = -1, 1 \end{cases}$$

*This can be seen from the defining equation of the MA(1) process as follows.*

$$\begin{aligned} \gamma(k) &= \mathsf{Cov}(X_t, X_{t+k}) = \mathsf{Cov}(Z_t + \beta Z_{t-1}, Z_{t+k} + \beta Z_{t+k-1}) \\ &= \mathsf{Cov}(Z_t, Z_{t+k}) + \beta\mathsf{Cov}(Z_t, Z_{t+k-1}) + \beta\mathsf{Cov}(Z_{t-1}, Z_{t+k}) + \beta^2\mathsf{Cov}(Z_{t-1}, Z_{t+k-1}). \end{aligned}$$

*For $|k| > 1$, the indices $t-1$, $t$, $t+k-1$, $t+k$ are all distinct, so that $\gamma(k) = 0$ because the variables $Z_t$ are independent.*

*For $k = 0$,*

$$\gamma(0) = \mathsf{Var}(Z_t) + \beta^2\mathsf{Var}(Z_{t-1}) = (1+\beta^2)\sigma^2.$$

*For $k = 1$,*

$$\gamma(1) = \beta\mathsf{Cov}(Z_t, Z_t) = \beta\sigma^2,$$

*and symmetrically, $\gamma(-1) = \beta\sigma^2$. Normalizing gives the result for $\rho(k)$.*

We see that the ac.f for an MA(1) process cuts off after the first lag. It is typically the case that $\gamma(k)$ decays exponentially to 0 as $k \to \infty$ for stationary processes.

**Theorem 1.1** *1) The autocovariance function $\gamma(t)$, $t \in \boldsymbol{Z}$, is a non-negative definite function, i.e., for any real numbers $a_i, i = 1, \ldots, n$, and any times $t_i \in \boldsymbol{Z}$, $i = 1, \ldots, n$,*

$$\sum_{i,j=1}^{n} a_i a_j \gamma(t_i - t_j) \geq 0.$$

*2) Both the autocovariance and the autocorrelation function are even functions of lag*

$$\gamma(\tau) = \gamma(-\tau),$$

$$\rho(\tau) = \rho(-\tau),$$

*$\tau \in \boldsymbol{Z}$.*

*3) $|\rho(\tau| \leq 1$, $\tau \in \boldsymbol{Z}$.*

*4) The ac.f. does not uniquely identify the underlying stochastic process.*

*Proof.*

1) Indeed

$$\sum_{i,j=1}^{n} a_i a_j \gamma(t_i - t_j) = \mathsf{Var}(a_1 X_{t_1} + \cdots + a_n X_{t_n}) \geq 0.$$

2) Check this property for the autocovariance function

$$\gamma(\tau) = \mathsf{Cov}(X_t, X_{t+\tau}) = \mathsf{Cov}(X_{t-\tau}, X_t) = \mathsf{Cov}(X_t, X_{t-\tau}) = \gamma(-\tau);$$

the same property for the autocorrelation function follows immediately.

3) Indeed

$$0 \leq \mathsf{Var}(\lambda_1 X_t + \lambda_2 X_{t+\tau}) = (\lambda_1^2 + \lambda_2^2)\gamma(0) + 2\lambda_1\lambda_2\gamma(\tau).$$

If $\lambda_1 = \lambda_2 = 1$, then $\gamma(\tau) \geq -\gamma(0)$, so that $\rho(\tau) \geq -1$. If $\lambda_1 = 1$, $\lambda_2 = -1$, then $\gamma(0) \geq \gamma(\tau)$, so that $\rho(\tau) \leq 1$. Thus $|\rho(\tau)| \leq 1$ as required. This property is also an immediate consequence of the Cauchy-Schwartz inequality.

$$|\mathsf{Cov}(X_t, X_{t+\tau})| \leq \sqrt{\mathsf{Var}(X_t)}\sqrt{\mathsf{Var}(X_{t+\tau})}.$$

4) Examples will be given later. □

### 1.3.3 The sample mean and the sample autocovariance function

Let $\{X_t,\, t \in \mathbf{Z}\}$ be a weakly stationary process with mean $\mu$ and the autocovariance function $\gamma(\cdot)$.

The sample mean

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

is used as an unbiased point estimator for $\mu$. If mean is estimated, then usually the zero mean process $Y_t = X_t - \mu$ is considered. Subtracting the mean does not change the process autocorrelation function.

Given $x_1, \ldots, x_N$ observations of a stationary process, the *sample autocovariance* is defined as follows

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}(x_t - \overline{x})(x_{t+k} - \overline{x})$$

$$\overline{x} = \frac{1}{N}\sum_{t=1}^{N} x_t$$

. $c_k$ is the usual estimator of the theoretical autocovariance coefficient $\gamma(k)$ at lag $k$. Note the following properties of the sample covariance function:

- $\mathsf{E}(c_k) \neq \gamma(k)$, i.e., it is a biased estimator.
- $\mathsf{E}(c_k) \to \gamma(k)$ as $N \to \infty$, i.e., it is an asymptotically unbiased estimator.

The sample ac.f. is defined by

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1}^{N-k}(x_t - \overline{x})(x_{t+k} - \overline{x})}{\sum_{t=1}^{N}(x_t - \overline{x})^2}$$

We often look at plots of $r_k$ as a function of time. This is known as a *correlogram*.

### 1.3.4 Linear filters

Given a time series $\{X_t,\, t \in \mathbf{Z}\}$ one can apply to it a linear operator or linear filter

$$Y_t = \sum_{k=-\infty}^{\infty} a_k X_{t-k},$$

specified by fixed (i.e. non-random) coefficients $a_k$, $k \in \mathbf{Z}$. In general this is an infinite sum, therefore its convergence in some probabilistic sense has to be justified (e.g., mean square convergence, to be discussed later).

We have already encountered several processes that were defined implicitly as linear filters, e.g. the MA(1)-process, which is obtained by applying a linear filter with two nonzero coefficients $a_0 = 1$ and $a_1 = \beta$ to a white noise process.

We can sometimes use a linear filter to transform a non-stationary time series into a stationary one.

### 1.3.5 Differencing as a linear filter

Define the first difference of the stochastic process $X_t$ at lag $d$ to be

$$\nabla_d X_t = X_t - X_{t-d}.$$

This is a linear filter with two non-zero coefficients $a_0 = 1$ and $a_d = -1$. Note that $\nabla_1 X_t$ will always be denoted $\nabla X_t$.

Further, for $j > 1$, the $j$th difference at lag $d$ is defined to be

$$\nabla_d^j X_t = \nabla_d \left( \nabla_d^{j-1} X_t \right).$$

An important property of differencing is that it preserves stationarity.

**Proposition 1.1** *If $\{X_t,\, t \in \boldsymbol{Z}\}$ is a stationary stochastic process with nonzero mean and autocovariance function $\gamma(\tau)$, $\tau \in \boldsymbol{Z}$, then its first difference at lag $d$, $\nabla_d X_t$, is a stationary stochastic process with zero mean and autocovariance function*

$$\tilde{\gamma}(k) = 2\gamma(k) - \gamma(k+d) - \gamma(k-d),\ k \in \boldsymbol{Z}.$$

*Proof.* It is clear that

$$\mathsf{E}(X_t - X_{t-d}) = 0.$$

Compute the autocovariance function of $\{Y_t = X_t - X_{t-d},\, t \in \mathbf{Z}\}$,

$$\begin{aligned}
\mathsf{Cov}(Y_t, Y_{t+k}) &= \mathsf{Cov}(X_t - X_{t-d}, X_{t+k} - X_{t+k-d}) \\
&= \mathsf{Cov}(X_t, X_{t+k}) - \mathsf{Cov}(X_{t-d}, X_{t+k}) - \mathsf{Cov}(X_t, X_{t+k-d}) + \mathsf{Cov}(X_{t-d}, X_{t+k-d}) \\
&= 2\gamma(k) - \gamma(k+d) - \gamma(k-d).
\end{aligned}$$

This depends only on $k$, hence the process is stationary (with zero mean).□

More generally, it is the case that if $X_t$ is stationary, the linear filter

$$Y_t = \sum_{k=-\infty}^{\infty} a_k X_{t-k}$$

is also stationary, so long as $\sum_{k=-\infty}^{\infty} |a_k| < \infty$.

## 1.4 Removal of trend and seasonal components

A polynomial trend can be removed by taking differences of an appropriate order. A seasonal component can also be removed, by taking differences of an appropriate lag.

**Proposition 1.2** *If* $m_t = \sum_{j=0}^{k} a_j t^j$, $t \in \boldsymbol{Z}$ *then*

$$\nabla^k m_t = k! a_k.$$

**Corollary 1.1** *If* $X_t = \sum_{j=0}^{k} a_j t^j + Y_t$, $t \in \boldsymbol{Z}$, *where* $k \geq 1$, $a_k \neq 0$ *and* $\{Y_t, t \in \boldsymbol{Z}\}$ *is a stationary process, then*

$$\nabla^k X_t = k! a_k + \nabla^k Y_t$$

*is a stationary process with mean* $k! a_k$.

This means that we can remove any polynomial trend.

Suppose we now consider a time series

$$X_t = m_t + S_t + Y_t,$$

where $\{S_t, t \in \mathbf{Z}\}$ has period $d$, i.e., $S_t = S_{t-d}$ for any $t$. Applying $\nabla_d$ gives

$$\nabla_d X_t = X_t - X_{t-d} = m_t - m_{t-d} + Y_t - Y_{t-d}$$

which gives a decomposition of the difference $\nabla_d X_t$ into a trend component $m_t - m_{t-d}$ and a stationary term $Y_t - Y_{t-d}$. If now $m_t$ is a polynomial of order $k$, $m_t - m_{t-d}$ is also a polynomial and so can be removed by taking differences of the appropriate order as above.

Another approach to removing a polynomial trend is to estimate the polynomial first, and then subtract it. e.g. suppose

$$X_t = a + bt + ct^2 + Y_t,$$

where $Y_t$ is stationary and we have observations of $X_t$ for $1 \leq t \leq N$. We can obtain estimators $(\hat{a}, \hat{b}, \hat{c})$ by looking for the minimizers of the function

$$f(u_1, u_2, u_3) = \sum_{t=1}^{N} (X_t - u_1 - u_2 t - u_3 t^2)^2$$

and then simply work with the subtracted time series

$$X_t - \hat{a} - \hat{b}t - \hat{c}t^2.$$

## 1.5 $L^2-$space and mean square convergence

*This non-examinable section contains a number of results that are useful for a rigorous understanding of the probabilistic issues underlying convergence results that we need. Proofs are generally omitted, but can be found in Chapter 2 of Brockwell and Davis.*

Consider a probability space $(\Omega, \mathcal{F}, P)$. We say that a random variable $X$ defined on $\Omega$ is square integrable if

$$\mathsf{E}(X^2) < \infty.$$

We denote by $L^2 = L^2(\Omega, \mathcal{F}, P)$ the collection of all square integrable random variables $X$ defined on $(\Omega, \mathcal{F}, P)$. Note that $L^2$ is a linear space, since if $\mathsf{E}(X^2) < \infty$ and $a \in \mathbf{R}$, then

$$\mathsf{E}((aX)^2) = a^2\mathsf{E}(X^2) < \infty,$$

so $L^2$ is closed under multiplication. Further, it is closed under addition, since if $\mathsf{E}(X^2) < \infty, \mathsf{E}(Y^2) < \infty$, then

$$\mathsf{E}((X+Y)^2) = \mathsf{E}(2X^2 + 2Y^2 - (X-Y)^2) \le 2\mathsf{E}(X^2) + 2\mathsf{E}(Y^2) < \infty.$$

*(This is the parallelogram law.)*

Two square integrable random variables $X$ and $Y$ are said to be *orthogonal* if

$$\mathsf{E}(XY) = 0.$$

A square integrable random variable $X$ is called *orthogonal* to a set of $\{Y, Z, \ldots\}$ square integrable random variables if

$$\mathsf{E}(XY) = 0,\ \mathsf{E}(XZ) = 0, \ldots,$$

i.e., if $X$ is orthogonal to any element of the set.

Note that we can also define the norm $\|X\|$ by

$$\|X\|^2 = \mathsf{E}(X^2).$$

If $X_n$, $n \ge 1$, and $X$ are square integrable random variables and

$$\|X_n - X\|^2 = \mathsf{E}(X_n - X)^2 \to 0, \quad \text{as} \quad n \to \infty,$$

then we say that the sequence $X_n$, $n \ge 1$, converges to $X$ in mean square.

**Theorem 1.2** *If $\{X_n\}$ is a Cauchy sequence, i.e., $\|X_n - X_m\| \to 0$, as $n, m \to \infty$, then there exists $X \in L^2$, such that $X_n \to X$ in mean square as $n \to \infty$.*

This theorem states that $L^2$ is *complete*, which is to say that $L^2$ is an example of a *Hilbert space*.

**Proposition 1.3** *If $X_n, Y_n \in L^2$, $n \ge 1$, $X, Y \in L^2$ and $X_n \to X$, $Y_n \to Y$ in mean square as $n \to \infty$, then*

$$\mathsf{E}(X_nY_n) \to \mathsf{E}(XY),$$

*as $n \to \infty$.*

**Lemma 1.1** *Let $Y_k$, $k \ge 1$, be a sequence of independent random variables with zero mean and $\mathsf{E}(Y_k^2) = \sigma_k^2$ such that*

$$\sum_{k=1}^{\infty} \sigma_k^2 < \infty, \tag{1}$$

*then the sequence of random variables $S_n = \sum_{k=1}^{n} Y_k$, $n \ge 1$, converges in $L^2$.*

*Proof.* By Theorem 1.2 it suffices to show that $S_n = \sum_{k=1}^{n} Y_k$, $n \geq 1$, is Cauchy sequence in $L^2$. Assuming $n > m$ we get by direct computation that

$$\mathsf{E}((S_n - S_m)^2) = \sum_{k=m+1}^{n} \sigma_k^2$$

hence $\mathsf{E}((S_n - S_m)^2) \to 0$ as $n, m \to \infty$, by assumption (1). Therefore $S_n$, $n \geq 1$, converges in $L^2$. The limit is, of course, the infinite sum $S_\infty = \sum_{k=1}^{\infty} Y_k$. Also, observe that

$$\mathsf{E}(S_\infty^2) = \sum_{k=1}^{\infty} \sigma_k^2.$$

**Theorem 1.3** *(Consistency of the sample mean.) Let $\{X_t, t \in \boldsymbol{Z}\}$ be a stationary stochastic process with mean $\mu$ and the autocovariance function $\gamma(k)$, $k \in \boldsymbol{Z}$. If $\gamma(k) \to 0$ as $k \to \infty$, then $\overline{X} \to \mu$ in mean square as $n \to \infty$.*

# 2 ARMA (Autoregressive Moving Average) processes

## 2.1 MA (moving average) processes

We have already encountered the MA(1) process, satisfying

$$X_t = Z_t + \beta Z_{t-1},$$

where the variables $Z_t$ are white noise, i.e. i.i.d with $\mathsf{E}(Z_t) = 0$.

It is often convenient to write such definitions in terms of the *backward shift* operator $B$. This acts on a time series as follows

$$BX_t = X_{t-1},$$

and powers of $B$ are defined by

$$B^k X_t = X_{t-k}.$$

In terms of this operator, the defining equation of the MA(1) process is just

$$X_t = (1 + \beta B)Z_t.$$

The moving average process of order $q$ (MA(q) process) is defined analogously:

$$X_t = Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}, \tag{2}$$

or, more compactly:

$$X_t = (1 + \beta_1 B + \beta_2 B^2 + \ldots + \beta_q B^q)Z_t,$$

where $Z_t$ is white noise with $\mathsf{Var}(Z_t) = \sigma^2 < \infty$. It will be of use to abbreviate this even further to

$$X_t = \theta(B)Z_t,$$

with $\theta(\lambda) = 1 + \beta_1\lambda + \ldots + \beta_q\lambda^q$.

It is clear that this process is weakly stationary for any $\{\beta_k\}$. Indeed, defining for convenience $\beta_0 = 1$, we compute that

$$\mathsf{E}X_t = 0$$

$$\gamma(0) = \mathsf{Var}(X_t) = \sigma^2 \sum_{i=0}^{q} \beta_i^2,$$

both of which are independent of $t$. Now considering $\gamma(k)$ for $k > 0$,

$$\gamma(k) = \mathsf{Cov}(X_t, X_{t+k}) = \mathsf{Cov}(Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q}, Z_{t+k} + \beta_1 Z_{t+k-1} + \ldots + \beta_q Z_{t+k-q}).$$

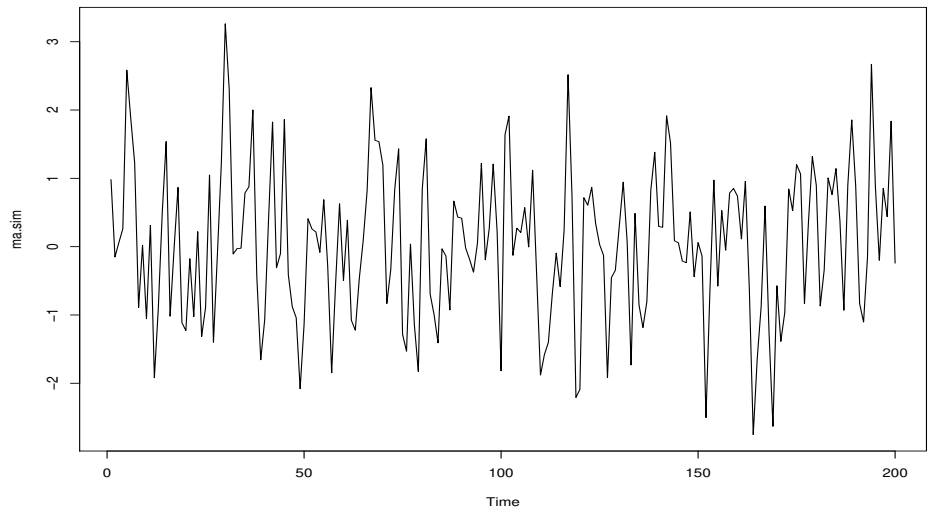


Figure 1: Plot of 200 simulated realizations of MA(1) process with $\beta = 0.5$

Note that if $k > q$, there is no overlap in the indices $t, t-1, \ldots, t-q$ and $t+k, t+k-1, \ldots, t+k-q$, so that $\gamma(k) = 0$. However, if $0 \leq \ k \leq q$, then

$$\gamma(k) = \mathsf{Cov}(X_t, X_{t+k}) = \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \mathsf{E}(Z_{t-i}^2) = \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \sigma^2.$$

Since $\gamma(k)$ is an even function of lag, this completes the calculation.

The ac.f. of the above MA(q) process can now be written as

$$\rho(k) = \begin{cases} 0 & k > q, \\ 1 & k = 0, \\ \sum_{i=0}^{q-k} \beta_i \beta_{i+k} / \sum_{i=0}^{q} \beta_i^2 & k = 1, \ldots, q, \\ \rho(-k) & k < 0. \end{cases}$$

An important property when trying to recognise an MA(q) process is that its ac.f cuts off after $q$ lags.

*(I don't think I got to the example below, but it is worth noting, and I will mention it briefly when time permits.)*

Note that it is possible to exhibit different weakly stationary MA processes with the same ac.f.
Model I: MA(1) process $X_t = Z_t + \beta Z_{t-1}$, with ac.f

$$\rho(k) = \begin{cases} 0 & |k| > 1, \\ 1 & k = 0, \\ \beta/(1+\beta^2) & k = -1, 1 \end{cases}$$

Model II: MA(1) process $X_t = Z_t + \beta^{-1} Z_{t-1}$, with ac.f.

$$\rho(k) = \begin{cases} 0 & |k| > 1, \\ 1 & k = 0, \\ \beta/(1+\beta^2) & k = -1, 1 \end{cases}$$

So, if $\beta \neq 1$, then we have two different MA(1) processes with the same ac.f. (this is an example of the last assertion of Theorem 1.1). We will consider only MA(1)-processes with $|\beta| < 1$, because they are *invertible*. Invertibility will be discussed shortly!

## 2.2 AR (autoregressive) processes

The AR(p) process satisfies the following equation

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + Z_t, \tag{3}$$

where $Z_t$ is white noise. This can be written in terms of the backward shift operator as

$$\phi(B) X_t = Z_t.$$

What conditions on $\phi$ are needed for $X_t$ as defined above to be stationary? Consider the case of the AR(1) process first, with defining equation

$$X_t = \alpha X_{t-1} + Z_t, \tag{4}$$

where $\alpha \in \mathbf{R}$ and $\{Z_t,\, t \in \mathbf{Z}\}$ is white noise.

1) Assume first that $|\alpha| < 1$. Substituting into the equation (4) $k$ times we obtain

$$\begin{aligned} X_t &= \alpha(\alpha X_{t-2} + Z_{t-1}) + Z_t \\ &= \ldots \\ &= \alpha^2(\alpha X_{t-3} + Z_{t-2}) + +\alpha Z_{t-1} + Z_t \\ &= Z_t + \alpha Z_{t-1} + \cdots + \alpha^k Z_{t-k} + \alpha^{k+1} X_{t-k-1}. \end{aligned}$$

If $\{X_t,\, t \in \mathbf{Z}\}$ is a stationary solution, then $\mathsf{E}(X_t^2)$ is constant and $\alpha^{k+1} X_{t-k-1}$ converges to zero in mean square as $k \to \infty$. This means that $X_t - \sum_{j=0}^{k} \alpha^j Z_{t-j}$ also converges to zero in mean square. Therefore, if $|\alpha| < 1$, the stochastic process

$$X_t = \sum_{j=0}^{\infty} \alpha^k Z_{t-j} \tag{5}$$

is the unique stationary solution of equation (4).

This can be expressed quite neatly in $B$-notation as follows.

$$(1 - \alpha B) X_t = Z_t$$

and so

$$\begin{aligned} X_t &= Z_t/(1-\alpha B) \\ &= (1+\alpha B+\alpha^2 B^2+\cdots)Z_t \\ &= Z_t+\alpha Z_{t-1}+\alpha^2 Z_{t-2}+\cdots \end{aligned}$$

By working directly with the representation (5), it is clear that

$$\mathsf{E}(X_t)=0,$$

$$\mathsf{Var}(X_t)=\sigma^2\sum_{k=0}^{\infty}\alpha^{2k}=\frac{\sigma^2}{(1-\alpha^2)}<\infty,$$

We can now calculate the autocovariance function $\gamma(k)$ directly from the defining equation (4). This foreshadows the approach we will take for higher order AR(p) processes. First, multiply both sides of (4) by $X_{t-k}$ and then take expectations.

$$\mathsf{E}(X_tX_{t-k})=\mathsf{E}(\alpha X_{t-1}X_{t-k})+\mathsf{E}(Z_tX_{t-k}).$$

Note that $\mathsf{E}(Z_tX_{t-k})=0$, as can be seen by considering the representation (5) for $X_{t-k}$. Now

$$\mathsf{E}(X_tX_{t-k})=\mathsf{E}(\alpha X_{t-1}X_{t-k})=\alpha\mathsf{Cov}(X_{t-1},X_{t-k})=\alpha\mathsf{Cov}(X_{t-1},X_{t-1-(k-1)}).$$

This gives

$$\mathsf{E}(X_tX_{t-k})=\alpha\mathsf{Cov}(X_{t-1},X_{t-1-(k-1)})=\alpha\mathsf{Cov}(X_0,X_{k-1})=\alpha\gamma(k-1),$$

using stationarity and the fact that $\gamma$ is an even function of $k$. This shows that $\gamma(k)=\alpha\gamma(k-1)$. Iterating this argument then gives

$$\gamma(k)=\alpha^k\gamma(0)=\frac{\alpha^k\sigma^2}{1-\alpha^2}.$$

Note that we could also obtain the same result from the representation (5) as follows. Taking $k\geq 0$,

$$\gamma(k)=\mathsf{E}(X_tX_{t+k})=\mathsf{E}\left(\left(\sum_{i=0}^{\infty}\alpha^iZ_{t-i}\right)\left(\sum_{i=0}^{\infty}\alpha^iZ_{t+k-i}\right)\right)=\sigma^2\sum_{i=0}^{\infty}\alpha^i\alpha^{k+j}=\frac{\alpha^k\sigma^2}{(1-\alpha^2)}.$$

By normalizing by the variance of the process, the ac.f. is seen to be.

$$\rho(k)=\alpha^{|k|},\qquad k\in\mathbf{Z}$$

Note that the ac.f. of an AR(1)-process decays exponentially.

2) Assume now that $|\alpha|>1$. In this case the series (5) does not converge in $L^2$, but the equation (4) can be rewritten as follows

$$X_t=-\alpha^{-1}Z_{t+1}+\alpha^{-1}X_{t+1}.$$

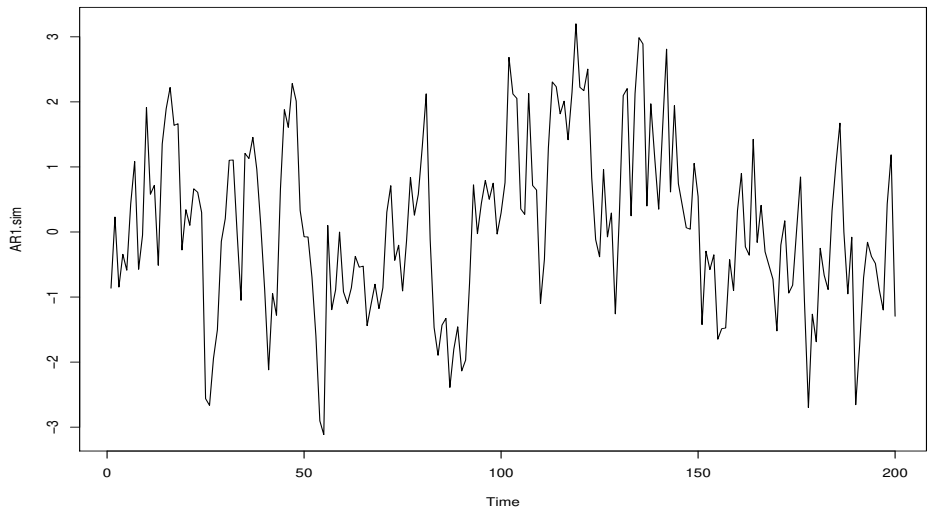


Figure 2: Simulated realization of a stationary process (AR(1) with $\alpha = 0.7$, 200 values)

This process can be repeated to gives

$$X_t = -\alpha^{-1} Z_{t+1} - \cdots - \alpha^{-k} Z_{t+k} + \alpha^{-k-1} X_{t+k+1}.$$

By the same arguments as before we obtain that

$$X_t = \sum_{k=1}^{\infty} \alpha^{-k} Z_{t+k}$$

is the unique stationary solution of (4).

We have seen that the AR(1) process with $|\alpha| < 1$: can be represented as an MA($\infty$)-process, i.e., in terms of $\mathbf{Z}_k$, $k \leq t$. Such a process is called causal or future-independent AR-process. In constrast, the AR(1) process with $|\alpha| > 1$ is future-dependent, so is regarded as unnatural and is not used in modelling stationary times series.

$|\alpha| = 1$ is a degenerate case. If, say, $\alpha = 1$, then

$$X_t = X_{t-1} + Z_t$$

is not stationary (we have already seen that the random walk is not stationary). In this case, there is no stationary solution. Higher order AR(p) processes will be discussed in a later section.

## 2.3 Definition of the general ARMA process

**Definition 2.1** *The process* $\{X_t, t \in \boldsymbol{Z}\}$, *is said to be an ARMA(p,q) process if it is weakly stationary and satisfies the following linear difference equation*

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q},\ t \in \boldsymbol{Z}$$

*where* $\{\alpha_i,\, i = 1, \ldots, p,\, \beta_j,\, j = 1, \ldots, q\}$ *are real numbers, and* $\{Z_t, t \in \boldsymbol{Z}\}$, *is a white noise process with finite variance* $\mathsf{Var}(Z_t) = \sigma^2$.

In terms of the operator $B$, the equation for an ARMA(p,q) process can be written in the form

$$\phi(B)X_t = \theta(B)Z_t$$

where $\phi(B)$ and $\theta(B)$ are polynomials of order $p, q$ respectively

$$\phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p$$

is the characteristic polynomial of the AR part, and

$$\theta(B) = 1 + \beta_1 B + \cdots + \beta_q B^q$$

is the characteristic polynomial of the MA part.

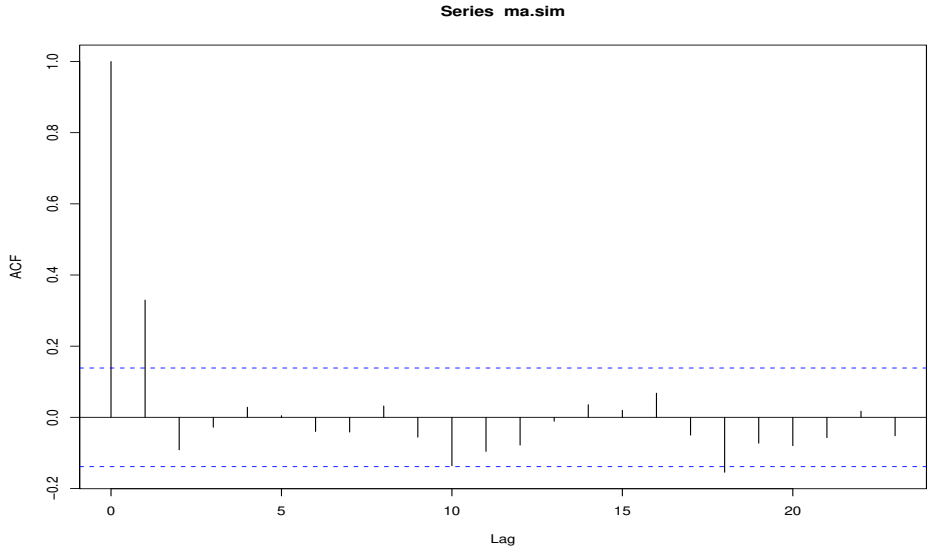


Figure 3: The sample ac.f.for the MA(1) sample plotted in Fig. 1

## 2.4 Invertibility and causality

Let $\{Z_t, t \in \mathbf{Z}\}$ be a white noise process with zero mean and variance $\sigma^2 < \infty$ and let $\{X_t, t \in \mathbf{Z}\}$ be an ARMA(p,q) process defined by the following equation

$$\phi(B)X_t = \theta(B)Z_t,\ t \in \mathbf{Z}, \tag{6}$$

where $\phi(B)$ and $\theta(B)$ are the characteristic polynomials given after (2.1).

**Definition 2.2** *An ARMA(p,q) process* $\{X_t,\, t \in \boldsymbol{Z}\}$ *defined by equation (6) is said to be invertible if there exists a sequence of constants* $\{a_i,\, i \in \boldsymbol{Z}_+\}$ *such that* $\sum |a_i| < \infty$ *and*

$$Z_t = \sum_{i=0}^{\infty} a_i X_{t-i}.$$

Invertibility means that the inverse operator

$$a(B) = \frac{\phi(B)}{\theta(B)} = \theta^{-1}(B)\phi(B),$$

exists, therefore

$$Z_t = \frac{\phi(B)}{\theta(B)} X_t = a(B)X_t.$$

Since $\phi$ and $\theta$ are polynomials with $\alpha_0 = \beta_0 = 1$ the $a(B)$ can be written as

$$a(B) = 1 - \sum_{j \geq 1} a_j B^j,$$

provided $\sum |a_j| < \infty$ (so that this formal power series expansion converges), and

$$Z_t = X_t - \sum_{j=1}^{\infty} a_j X_{t-j}.$$

As an illustration, consider the MA(1) process

$$X_t = Z_t + \beta Z_{t-1} = (1 + \beta B) Z_t.$$

Here $\phi(B) = 1$ and $\theta(B) = 1 + \beta B$. For $|\beta| < 1$, the MA(1) process is invertible and

$$Z_t = (1 + \beta B)^{-1} = X_t - \beta X_{t-1} + \beta^2 X_{t-2} - \cdots$$

Invertibility means that an AR($\infty$)-representation of the process $X_t$ is valid

$$Z_t = \frac{\phi(B)}{\theta(B)} X_t = X_t - \sum_{j=1}^{\infty} a_j X_{t-j}$$

or

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + Z_t$$

this representation is often helpful, e.g. when calculating the ac.f of a process, as we shall see later.

**Definition 2.3** *An ARMA(p,q) process defined by equation (6) is said to be causal, if there exists a sequence of constants* $\{c_i, i \in \boldsymbol{Z}_+\}$ *such that* $\sum |c_i| < \infty$ *and*

$$X_t = \sum_{i=0}^{\infty} c_i Z_{t-i}.$$

Causality means that the inverse operator

$$c(B) = \frac{\theta(B)}{\phi(B)} = \phi^{-1}(B)\theta(B)$$

exists, therefore

$$X_t = \frac{\theta(B)}{\phi(B)} Z_t = c(B) Z_t.$$

Since $\phi$ and $\theta$ are polynomials with $\alpha_0 = \beta_0 =$ the inverse operator can be written as

$$c(B) = 1 + \sum_{j=1}^{\infty} c_j Z_{t-j},$$

**Theorem 2.1** *Assume that the polynomials $\theta(\lambda)$ and $\phi(\lambda)$, $\lambda \in \boldsymbol{C}$, do not have common roots. Then the ARMA(p,q) process*

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

*1) is invertible, if and only if all roots of the characteristic polynomial $\theta(\lambda)$ (corresponding to its MA part) lie outside the unit disc $\{z \in \boldsymbol{C} : |z| \leq 1\}$, i.e., the absolute value of any root is greater than* 1.
*2) is causal, if and only if all roots of the characteristic polynomial $\phi(\lambda)$ (corresponding to its AR part) lie outside the unit disc $\{z \in \boldsymbol{C} : |z| \leq 1\}$, i.e., the absolute value of any root is greater than* 1.

Recall the earlier example of the MA(1) process $X_t = Z_t + \beta Z_{t-1}$. Its characteristic polynomial is $\theta(\lambda) = 1 + \beta\lambda$, with root $\lambda = -1/\beta$, which is a real number and lies outside the unit disc provided $|\beta| < 1$.

For the AR(1) process $X_t - \alpha X_{t-1} = Z_t$, the characteristic polynomial is $\phi(\lambda) = 1 - \alpha\lambda = 0$, with root $\lambda = 1/\alpha$, which is a real number and lies outside the unit disc provided $|\alpha| < 1$.

**Example 2.1** *Show that the ARMA(1,1) process $X_t - 0.6X_{t-1} = Z_t - 0.2Z_{t-1}$ is both invertible and causal, and find its MA representation.*

*First note that $\phi(\lambda) = 1 - 0.6\lambda$ and $\theta(\lambda) = 1 - 0.2\lambda$. This means that the root of the AR characteristic polynomial is $\lambda = \frac{10}{6} > 1$ and the root of the MA characteristic polynomial is $\lambda = 5 > 1$. These values are outside the unit disc, hence the process is both invertible and causal.*

$$X_t = \frac{\theta(B)}{\phi(B)} Z_t = \frac{(1 - 0.2B)}{(1 - 0.6B)} Z_t = (1 - 0.2B) \left( \sum_{i=0}^{\infty} 0.6^i B^i \right) Z_t = Z_t + \sum_{i=1}^{\infty} 0.4 \times 0.6^{i-1} Z_{t-i}$$

## 2.5 Computation of the ac.f. for ARMA(p,q) processes

### 2.5.1 Computation by using MA($\infty$)-representation

Suppose the ARMA(p,q) process

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

is causal. How can we compute the autocovariance function and the autocorrelation function of this process?
If we can solve

$$c(B) = \frac{\theta(B)}{\phi(B)} = \phi^{-1}(B)\theta(B),$$

i.e., compute coefficients in the expansion

$$X_t = \sum_{j=0}^{\infty} c_j Z_{t-j},$$

then we can compute

$$\gamma(k) = \mathsf{Cov}(X_t, X_{t+k}) = \sigma^2 \sum_{j=0}^{\infty} c_j c_{j+k}, \quad k > 0,$$

and the problem is theoretically solved.

However, it can be difficult to compute the coefficients $c_i$ and only in particular cases can the coefficients be computed explicitly. For instance, we can compute the coefficients for an ARMA(1,1) process. Indeed, consider

$$(1 - \alpha B)X_t = (1 + \beta B)Z_t, \tag{7}$$

where $|\alpha| < 1$ and $|\beta| < 1$.

$$X_t = \frac{1 + \beta B}{1 - \alpha B} Z_t = (1 - \alpha B)^{-1}(1 + \beta B)Z_t,$$

which gives

$$X_t = \left( \sum_{k=0}^{\infty} \alpha^k B^k \right) (1 + \beta B)Z_t = \sum_{k=0}^{\infty} \alpha^k B^k (1 + \beta B)Z_t.$$

This then simplifies to

$$Z_t + (\alpha + \beta) \sum_{k=1}^{\infty} \alpha^{k-1} Z_{t-k} \tag{8}$$

This process is a linear filter of the time series $Z_t$, and the sum converges since $|\alpha| < 1$.

This representation now allows us to calculate the ac.f of the general ARMA(1,1) process. We first set up some useful preliminary results. Multiplying (8) by $Z_t$, taking expectations and using independence of $Z_i$ and $Z_j$, when $i \neq j$ we obtain

$$\mathsf{E}(Z_t X_t) = \mathsf{E}(Z_t^2) = \sigma^2.$$

Further, multiplying (8) by $Z_{t-1}$ gives

$$\mathsf{E}(Z_{t-1}X_t) = \alpha \mathsf{E}(X_{t-1}Z_{t-1}) + \mathsf{E}(Z_t Z_{t-1}) + \beta \mathsf{E}(Z_{t-1}^2) = \sigma^2(\alpha + \beta).$$

On multiplying (7) by $X_t$ we find that

$$\gamma(0) = \alpha\gamma(1) + \sigma^2 + \beta(\alpha+\beta)\sigma^2.$$

Moreover, multiplying (7) by $X_{t-1}$ and taking expectation gives

$$\mathsf{E}(X_t X_{t-1}) = \alpha\mathsf{E}(X_{t-1}^2) + \mathsf{E}(Z_t X_{t-1}) + \beta\mathsf{E}(Z_{t-1}X_{t-1}).$$

Now $\mathsf{E}(Z_t X_{t-1}) = 0$ (by considering a similar MA representation for $X_{t-1}$), so we get the following equation

$$\gamma(1) = \alpha\gamma(0) + \beta\sigma^2.$$

We now have two equations in the two unknowns $\gamma(0)$ and $\gamma(1)$. Solving them we get

$$\gamma(0) = \sigma^2\frac{1+\beta^2+2\alpha\beta}{1-\alpha^2}$$
$$\gamma(1) = \sigma^2\frac{(1+\alpha\beta)(\alpha+\beta)}{1-\alpha^2}.$$

To obtain the equations for $\gamma(k)$, $k \geq 2$, we again multiply the equation for $X_t$ by $X_{t-k}$, $k \geq 2$ and take expectations:

$$\mathsf{E}(X_t X_{t-k}) = \alpha\mathsf{E}(X_{t-1}X_{t-k}) + \mathsf{E}(Z_t X_{t-k}) + \beta\mathsf{E}(Z_{t-1}X_{t-k}).$$

By considering the MA representation for $X_{t-k}$, it is clear that $\mathsf{E}(Z_t X_{t-k}) = 0$ and $\mathsf{E}(Z_{t-1}X_{t-k}) = 0$. Hence

$$\gamma(k) = \alpha\gamma(k-1),\ k \geq 2.$$

So, finally

$$\gamma(0) = \sigma^2\frac{1+\beta^2+2\alpha\beta}{1-\alpha^2}$$
$$\gamma(1) = \sigma^2\frac{(1+\alpha\beta)(\alpha+\beta)}{1-\alpha^2}$$
$$\gamma(k) = \alpha\gamma(k-1),\ k \geq 2,$$
$$\gamma(k) = \gamma(-k),\ \text{if } k < 0.$$

and we have the following system of equations for the ac.f.

$$\rho(0) = 1$$
$$\rho(1) = \frac{(1+\alpha\beta)(\alpha+\beta)}{1+\beta^2+2\alpha\beta}$$
$$\rho(k) = \alpha\rho(k-1),\ k \geq 2,$$
$$\rho(k) = \rho(-k),\ \text{if } k < 0.$$

The correlations of an ARMA(1,1) process decays exponentially; the same is true for any ARMA(p,q) process.

### 2.5.2 The Yule-Walker equations for an AR(2) process

Consider an AR(2)-process

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + Z_t, \;\; t \in \mathbf{Z} \tag{9}$$

and suppose that it is causal.

Multiplying both sides of the equation (9) by $X_{t-k}$, where $k \geq 0$, and taking expectations, we obtain the following equations

$$\mathsf{E}(X_t X_{t-k}) = \alpha_1 \mathsf{E}(X_{t-1} X_{t-k}) + \alpha_2 \mathsf{E}(X_{t-2} X_{t-k}) + \mathsf{E}(Z_t X_{t-k}).$$

The process $X_t$ can be expressed as an MA process of infinite order

$$X_t = Z_t + c_1 Z_{t-1} + c_2 Z_{t-2} + \cdots$$

with some coefficients $c_i$. Therefore, if $k > 0$, then $X_{t-k}$ and $Z_t$ are independent and we have that $\mathsf{E}(Z_t X_{t-k}) = 0$, hence

$$\gamma(-k) = \alpha_1 \gamma(-k+1) + \alpha_2 \gamma(-k+2), \;\; k > 0, \tag{10}$$

If $k = 0$, then the equation will be

$$\gamma(0) = \alpha_1 \gamma(1) + \alpha_2 \gamma(2) + \sigma^2.$$

Using the fact that $\gamma$ is an even function, we can rewrite these equations as follows

$$\begin{aligned}
\gamma(0) - \alpha_1 \gamma(1) - \alpha_2 \gamma(2) &= \sigma^2 \\
\gamma(1) - \alpha_1 \gamma(0) - \alpha_2 \gamma(1) &= 0 \\
\gamma(2) - \alpha_1 \gamma(1) - \alpha_2 \gamma(0) &= 0 \\
\gamma(k) - \alpha_1 \gamma(k-1) - \alpha_2 \gamma(k-2) &= 0, \; k \geq 3.
\end{aligned}$$

These are the *Yule Walker* equations for an AR(2) process.

It is clear that the first three equations form a closed system of equations for $\gamma(0), \gamma(1)$ and $\gamma(2)$, so we can obtain the form of $\gamma(0), \gamma(1)$ and $\gamma(2)$ explicitly.

If we divide all equations (10) by the process variance, i.e., by $\gamma(0)$, and use the fact that $\rho(k) = \rho(-k)$, then we get the following system of equations for the autocorrelation function $\rho(k)$, $k \in \mathbf{Z}$, of the process

$$\begin{aligned}
\rho(0) &= 1 \\
\rho(1) - \alpha_1 - \alpha_2 \rho(1) &= 0 \\
\rho(k) - \alpha_1 \rho(k-1) - \alpha_2 \rho(k-2) &= 0, \; k \geq 2.
\end{aligned}$$

For $k = 1$ we can find from the second equation that

$$\rho(1) = \frac{\alpha_1}{1 - \alpha_2}$$

therefore we get the following system of equations for the ac.f. of the AR(2) process

$$\begin{aligned}
&\rho(0) = 1\\
&\rho(1) = \frac{\alpha_1}{1-\alpha_2}\\
&\rho(k) = \alpha_1\rho(k-1) + \alpha_2\rho(k-2),\ k \geq 2,\\
&\rho(k) = \rho(-k),\ k < 0.
\end{aligned}$$

We can, in principle, compute recursively $\rho(k)$ for any $k > 0$ (and for $k < 0$ similarly), one by one.

Another way to compute $\rho(k)$, $k \geq 0$ is to note that the equations above are linear difference equations, for which there is a systematic method of solution. Some terminology from the theory of difference equations is useful here: given an AR(p) process

$$X_t - \alpha_1 X_{t-1} - \ldots - \alpha_p X_{t-p} = \theta(B)X_t = Z_t$$

with characteristic polynomial

$$\phi(\lambda) = 1 - \alpha_1\lambda - \ldots - \alpha_p\lambda^p,$$

the auxiliary polynomial is defined as

$$\lambda^p - \alpha_1\lambda^{p-1} - \ldots - \alpha_p.$$

If $\pi_i$, $i = 1, \ldots, p$ are the zeroes of the auxiliary polynomial, then $\pi_i^{-1}$, $i = 1, \ldots, p$ are the zeros of the characteristic polynomial. The AR(p)-process is causal if and only if $|\pi_i| < 1$, $i = 1, \ldots, p$.

If $\pi_i, i = 1, 2$, are the zeroes of the auxiliary polynomial for the AR(2)-process, i.e., $\pi_i, i = 1, 2$, are the roots of the following equation

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

then a general solution of Yule-Walker equations for $\rho$ is given by the following formula

$$\rho(k) = A_1\pi_1^{|k|} + A_2\pi_2^{|k|},\ k \geq 0,$$

in the case when the roots are different and by a slightly different formula, if they coincide. The constants $A_i$ are determined by the initial conditions.

For the AR(2) process, the causality-stationarity condition $|\pi_i| < 1,\ i = 1, 2$ takes the form

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

It can be shown (by considering carefully all possible choices of $\alpha_i$) that these conditions are equivalent to the following simple conditions on coefficients $\alpha_1$ and $\alpha_2$

$$\begin{aligned}
\alpha_1 + \alpha_2 &< 1\\
\alpha_1 - \alpha_2 &> -1\\
\alpha_2 &> -1
\end{aligned}$$

Assume that the coefficients $\alpha_i$ satisfy the above conditions. We will consider the different possible cases in turn.

*I. Real roots.* If $\alpha_1^2 + 4\alpha_2 > 0$, the roots $\pi_i$ are real. To find the constants $A_i$ in the formula

$$\rho(k) = A_1\pi_1^{|k|} + A_2\pi_2^{|k|}$$

we use the equations for $k = 0$ and $k = 1$. Setting $k = 0$, we get the first equation for determining $A_1$ and $A_2$

$$\rho(0) = 1 = A_1 + A_2.$$

If $k = 1$, then

$$\rho(1) = \frac{\alpha_1}{1 - \alpha_2}$$

so, we get the second equation for determining $A_1$ and $A_2$

$$\rho(1) = A_1\pi_1 + A_2\pi_2.$$

Solving the system

$$A_1 + A_2 = 1$$
$$A_1\pi_1 + A_2\pi_2 = \frac{\alpha_1}{1 - \alpha_2}$$

we find

$$A_1 = \frac{\alpha_1(1 - \alpha_2)^{-1} - \pi_2}{\pi_1 - \pi_2}$$
$$A_2 = 1 - A_1$$

*II. Coincident roots.* If $\alpha_1^2 + 4\alpha_2 = 0$, then the roots coincide, $\pi_i = \alpha_1/2$, they are real, and the solution takes the form

$$\rho(k) = (A + Bk)(\alpha_1/2)^k,\ k \geq 0,$$

Using the initial conditions $\rho(0) = 1$ and $\rho(1) = \alpha_1/(1 - \alpha_2)$ we find that

$$A = 1, \quad B = \frac{1 + \alpha_2}{1 - \alpha_2}.$$

*III. Complex roots.* If $\alpha_1^2 + 4\alpha_2 < 0$, then the roots are a complex conjugate pair

$$\pi_1 = re^{i\varphi},\ \ \pi_2 = re^{-i\varphi},$$

where $i$ is an imaginary unit, $r = \sqrt{-\alpha_2} > 0$ and $\varphi = \tan^{-1}((-\alpha_1^2 - 4\alpha_2)/\alpha_1)$, to be interpreted as lying in the range $\pi/2$ to $\pi$ if $\alpha_1$ is negative. A general (complex-valued) solution can be written then in the form

$$\rho(k) = r^k(A_1e^{ik\varphi} + A_2e^{-ik\varphi}),\ k \geq 0,$$

where $A_1, A_2 \in \mathbf{C}$ are complex numbers. We are looking for a real-valued solution, so the choice of $A_1$ and $A_2$ must give

$$\rho(k) = r^k(A\cos(k\varphi) + B\sin(k\varphi)),\ k \geq 0,$$

where $A$ and $B$ are real numbers. Using again the initial conditions $\rho(0) = 1$ and $\rho(1) = \alpha_1/(1 - \alpha_2)$ we can compute the coefficients $A$ and $B$.

**Example 2.2** *Consider the AR(2) process* $X_t = X_{t-1} - 0.5X_{t-2} + Z_t$. *This is a causal stationary process, and its ac.f. is*

$$\rho(k) = \left(\frac{1}{\sqrt{2}}\right)^k \left(\cos\left(\frac{\pi k}{4}\right) + \frac{1}{3}\sin\left(\frac{\pi k}{4}\right)\right), \; k \geq 0.$$

# 3 Integrated ARMA or ARIMA models

**Definition 3.1** *A stochastic process* $\{X_t,\, t \in \boldsymbol{Z}\}$, *is called an ARIMA(p,d,q) process if its dth difference* $W_t = (1-B)^d X_t$ *is a causal and invertible ARMA(p,q) process of order p,q, i.e.,*

$$\phi(B)W_t = \theta(B)Z_t,$$

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t.$$

Note that the ARIMA(p,d,q) process is *not a stationary* stochastic process for $d > 0$ since its characteristic polynomial is $\phi(\lambda)(1-\lambda)^d$,which has a (multiple, if $d > 1$) zero on the unit circle $\{z \in \mathbf{C} : |z| = 1\}$. There is no (causal or not) stationary solution of the equation

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t.$$

**Example 3.1** *The ARIMA(0,1,0) process*

$$(1-B)X_t = Z_t$$

$$X_t = X_{t-1} + Z_t$$

*is a random walk.*

**Example 3.2** *ARIMA(1,1,0) process*

$$(1-0.7B)(1-B)X_t = Z_t$$

$$X_t = 1.7X_{t-1} - 0.7X_{t-2} + Z_t$$

# 4 Time series prediction

One of the main goals of time series analysis is to predict the future evolution of a time series given past observations.

## 4.1 Best linear prediction in $L^2$

**Example 4.1** *Let $X_1, X_2$ and $Y$ be square integrable random variables defined on the same probability space. The problem: Find the linear combination $\widehat{Y} = b_1X_1 + b_2X_2$ that minimizes the mean squared error (m.s.e.)*

$$m.s.e. = \mathsf{E}(Y - b_1X_1 - b_2X_2)^2.$$

*Solution.I.* Minimize the function

$$f(b_1, b_2) = \mathsf{E}(Y - b_1X_1 - b_2X_2)^2 = \mathsf{E}(Y^2) + b_1^2\mathsf{E}(X_1^2) + b_2^2\mathsf{E}(X_2^2) \\ - 2b_1\mathsf{E}(YX_1) - 2b_2\mathsf{E}(YX_2) + b_1b_2\mathsf{E}(X_1X_2)$$

of two real variables by calculus.
*Solution.II.* Find a linear combination $\widehat{Y} = b_1X_1 + b_2X_2$ such that

$$\mathsf{E}((Y - \widehat{Y})X_1) = 0$$

$$\mathsf{E}((Y - \widehat{Y})X_2) = 0,$$

so, $Y - \widehat{Y}$ is orthogonal to both $X_1$ and $X_2$, and, therefore, orthogonal to any linear combination $a_1X_1 + a_2X_2$.
In both cases the coefficients minimizing the $m.s.e.$ must satisfy the equations

$$b_1\mathsf{E}(X_1^2) + b_2\mathsf{E}(X_2X_1) = \mathsf{E}(YX_1)$$

$$b_1\mathsf{E}(X_1X_2) + b_2\mathsf{E}(X_2^2) = \mathsf{E}(YX_2).$$

More generally, consider random variables $X_1, \ldots, X_n$ and $Y$ with finite second moments

$$\mathsf{E}(Y_k^2) < \infty,\ k = 1, \ldots, n,$$

$$\mathsf{E}(X^2) < \infty$$

defined on the same probability space $(\Omega, \mathcal{F}, P)$, i.e., $X_1, \ldots, X_n, Y \in L^2(\Omega, \mathcal{F}, P)$.

**Definition 4.1** *The best linear predictor $\widehat{Y}$ of $Y$ in terms of $X_1, \ldots, X_n$ is the linear combination $b_1X_1 + \cdots + b_nX_n$ such that*

$$\mathsf{E}(|Y - (b_1X_1 + \cdots + b_nX_n)|^2) = \inf_{c_1,\ldots,c_n} \mathsf{E}(|Y - (c_1X_1 + \cdots + c_nX_n)|^2).$$

**Theorem 4.1** *Let $X_1, \ldots, X_n$, and $Y$ be an arbitrary square integrable random variables defined on the same probability space $(\Omega, \mathcal{F}, P)$. If the coefficients $b_i \in \boldsymbol{R}$, $i = 1, \ldots, n$ satisfy the prediction equations*

$$b_1 \mathsf{E}(X_1 X_k) + \ldots + b_n \mathsf{E}(X_n X_k) = \mathsf{E}(Y X_k),\ k = 1, \ldots, n, \tag{11}$$

*then*

$$\widehat{Y} = b_1 X_1 + \cdots + b_n X_n$$

*is the best linear predictor of $Y$ in terms of $X_1, \ldots, X_n$.*

*Proof.* Let $b_i$, $i = 1, \ldots, n$ be a solution of the prediction equations. Consider an arbitrary linear combination of $X_1, \ldots, X_n$

$$\tilde{Y} = a_1 X_1 + \cdots + a_n X_n, \quad a_i \in \mathbf{R}, i = 1, \ldots, n.$$

Direct computation gives

$$\begin{aligned}
\mathsf{E}((Y - \tilde{Y})^2) &= \mathsf{E}((Y - \widehat{Y} + \widehat{Y} - \tilde{Y})^2) \\
&= \mathsf{E}((Y - \widehat{Y})^2) + \mathsf{E}((\widehat{Y} - \tilde{Y})^2) + 2\mathsf{E}((Y - \widehat{Y})(\widehat{Y} - \tilde{Y}) \\
&= \mathsf{E}((Y - \widehat{Y})^2) + \mathsf{E}\left(\left(\sum_{i=1}^n (b_i - a_i) X_i\right)^2\right) + 2\mathsf{E}\left((Y - \widehat{Y}) \sum_{i=1}^n (b_i - a_i) X_i\right) \\
&= \mathsf{E}((Y - \widehat{Y})^2) + \mathsf{E}\left(\left(\sum_{i=1}^n (b_i - a_i) X_i\right)^2\right) + 2 \sum_{i=1}^n (b_i - a_i) \mathsf{E}\left((Y - \widehat{Y}) X_i\right) \\
&= \mathsf{E}((Y - \widehat{Y})^2) + \mathsf{E}\left(\left(\sum_{i=1}^n (b_i - a_i) X_i\right)^2\right)
\end{aligned}$$

since

$$\mathsf{E}\left((Y - \widehat{Y}) X_i\right) = 0,\ i = 1, \ldots, n$$

by definition of $b_i \in \mathbf{R}$, $i = 1, \ldots, n$. Therefore

$$\mathsf{E}((Y - \tilde{Y})^2) = \mathsf{E}((Y - \widehat{Y})^2) + \mathsf{E}\left(\left(\sum_{i=1}^n (b_i - a_i) X_i\right)^2\right) \geq \mathsf{E}((Y - \widehat{Y})^2)$$

The theorem is proved. □

This means that the best linear predictor of $Y$ in terms of $X_1, \ldots, X_n$ is a *projection* in $L^2$ of $Y$ onto

$$Lin\{X_1, \ldots, X_n\} = \{c_1 X_1 + \cdots + c_n X_n,\ c_i \in \mathbf{R}\}, \tag{12}$$

the linear subspace generated by $X_1, \ldots, X_n$.

Notation: $\Pi(Y | X_1, \ldots, X_n)$ denotes the best linear prediction (BLP) of $Y$ in terms of $X_1, \ldots, X_n$.

The projection is a linear operator:

$$\Pi(Y + Z | X_1, \ldots, X_n) = \Pi(Y | X_1, \ldots, X_n) + \Pi(Z | X_1, \ldots, X_n).$$

**Corollary 4.1** *Let $\{X_t,\, t \in \boldsymbol{Z}\}$ be a weakly stationary stochastic process with zero mean. If the coefficients $b_i$, $i = 0, \ldots, t-1$ are solutions of the following system of equations*

$$\gamma(h+k) = \sum_{i=0}^{t-1} b_i \gamma(k-i), \quad k = 0, \ldots, t-1,$$

*or, equivalently,*

$$\rho(h+k) = \sum_{i=0}^{t-1} b_i \rho(k-i), \quad k = 0, \ldots, t-1,$$

*where $\gamma(\cdot)$ and $\rho(\cdot)$ are the autocovariance and the autocorrelation function of the process, then*

$$\widehat{X}_{t+h} = \sum_{i=0}^{t-1} b_i X_{t-i}$$

*is the best linear predictor of of $X_{t+h}$ in terms of $X_1, \ldots, X_t$.*

The equations in the corollary are the prediction equations written in this particular case in terms of the autocovariance function.

If

$$\widehat{X}_{t+h} = \sum_{i=0}^{t-1} b_i X_{t-i} = b_{t-1} X_1 + \ldots + b_0 X_t.$$

then the prediction equations take the following form

$$\mathsf{E}((X_{t+h} - \widehat{X}_{t+h}) X_{k'}) = 0,\ k' = 1, \ldots, t.$$

Or,

$$\mathsf{E}((X_{t+h} X_{k'}) = b_0 \mathsf{E}(X_t X_{k'}) + \ldots + b_{t-1} \mathsf{E}(X_1 X_{k'}),\ k' = 1, \ldots, t.$$

In terms of the autocovariance function we have

$$\gamma(t+h-k') = \gamma(t+h-k') = b_0 \gamma(t-k') + \ldots + b_{t-1} \gamma(1-k'),\ k' = 1, \ldots, t.$$

If $k' \in \{1, ..., t\}$, then $t - k' \in \{0, ..., t-1\}$. Denote $k = t - k'$ then

$$\gamma(h+k) = b_0 \gamma(k) + b_1 \gamma(k-1) + \ldots + b_{t-1} \gamma(k-(t-1)) = \sum_{i=0}^{t-1} b_i \gamma(k-i),\ k = 0, 1, ..., t-1.$$

Note that the mean-square error is defined as follows

$$m.s.e. = \mathsf{E}((X_{t+h} - \widehat{X}_{t+h})^2).$$

## 4.2 Examples of best linear prediction for ARMA-processes

### 4.2.1 AR(1)-process

**Theorem 4.2** *Consider an AR(1)-process*

$$X_t = \alpha X_{t-1} + Z_t$$

*with* $|\alpha| < 1$, *where* $\mathsf{Var}(Z_t) = \sigma^2$. *Then, for any* $t \geq 2$ *and* $h \geq 1$, $\alpha^h X_t$ *is the best linear predictor of* $X_{t+h}$ *in terms of* $X_1, \ldots, X_t$.

*Proof.* $X_t$ is a causal stationary process, since $|\alpha| < 1$. We have computed its autocovariance function as

$$\gamma(k) = \mathsf{E}(X_t X_{t+k}) = \sigma^2 \frac{\alpha^{|k|}}{1-\alpha^2}.$$

Looking at the prediction equations then gives, for $k \in \{1, \ldots t\}$,

$$\mathsf{E}\left[(X_{t+h} - \alpha^h X_t)X_k\right] = \mathsf{E}\left[X_{t+h}X_k\right] - \alpha^h \mathsf{E}\left[X_{t+h}X_k\right] = \frac{\sigma^2}{1-\alpha^2}\left(\alpha^{t+h-k} - \alpha^h \alpha^{t-k}\right) = 0.$$

This shows that $X_{t+h} - \alpha^h X_t$ is orthogonal to any $X_k, k = 1, \ldots, t$. By Theorem 4.1, $\alpha^h X_t$ is the best linear predictor of $X_{t+h}$ in terms of $X_1 \ldots, X_t$.

We could also see this more directly from the MA($\infty$) representation, by using linearity of the projection. For this process, we have shown that the MA($\infty$) representation is given by

$$X_t = \sum_{k=0}^{\infty} \alpha^k Z_{t-k}.$$

This formula can be rewritten as follows

$$X_t = Z_t + \alpha Z_{t-1} + \cdots + \alpha^{k-1} Z_{t-k+1} + \alpha^k X_{t-k},$$

for any given $k$. Therefore

$$X_{t+h} = Z_{t+h} + \alpha Z_{t+h-1} + \cdots + \alpha^{h-1} Z_{t+1} + \alpha^h X_t.$$

Using linearity of projection now gives

$$\begin{aligned}\Pi(X_{t+h}|X_1, \ldots, X_t) &= \Pi(Z_{t+h} + \alpha Z_{t+h-1} + \cdots + \alpha^{h-1} Z_{t+1} + \alpha^h X_t | X_1, \ldots, X_t) \\ &= \alpha^h \Pi(X_t|X_1, \ldots, X_t) = \alpha^h X_t,\end{aligned}$$

since $\Pi(Z_{t+h} + \alpha Z_{t+h-1} + \cdots + \alpha^{h-1} Z_{t+1}|X_1, \ldots, X_t) = 0$ by causality. □

We can compute the m.s.e. of this forecast. By definition

$$\begin{aligned}m.s.e. = \mathsf{E}((X_{t+h} - \alpha^h X_t)^2) &= \mathsf{E}((Z_{t+h} + \alpha Z_{t+h-1} + \ldots + \alpha^{h-1} Z_{t+1})^2) \\ &= \sigma^2(1 + \alpha^2 + \ldots + \alpha^{2h-2}) = \sigma^2 \frac{1-\alpha^{2h}}{1-\alpha^2}\end{aligned}$$

### 4.2.2 MA(1) process

Consider a MA(1)-process

$$X_t = Z_t + \beta Z_{t-1}.$$

where we take $\beta < 1$ for invertibility. It is clear that if $h \geq 2$, then $\widehat{X}_{t+h} = 0$. This is because

$$\mathsf{E}(X_{t+h} Z_k) = 0, \quad k = 1, \ldots, t.$$

If $h = 1$, then

$$\Pi(X_{t+1} | X_1, \ldots, X_t) = \Pi(Z_{t+1} + \beta Z_t | X_1, \ldots, X_t) = \beta \hat{Z}_t,$$

Note that $Z_t$ is not observable. This means that we need to estimate it.

If all past observations $X_k$, $k = t, t-1, ...$ were available, then by invertibility

$$Z_t = X_t - \beta X_{t-1} + \beta^2 X_{t-2} ...,$$

i.e. the exact formula for $Z_t$ are given as as a linear function of the $X$ variables, so this would give a forecast for $Z_t$ in terms of the past values of the process.

If only $X_t, X_{t-1}, ..., X_1$ are available, then by Corollary 4.1, the coefficients $b_i$, $i = 1, \ldots, t$ determining the best linear predictor, can be found by solving the system of linear equations

$$\rho(k+1) = \sum_{i=0}^{t-1} b_i \rho(k-i), \quad k = 0, \ldots, t-1.$$

Therefore the $b_i$ can be written in matrix notation as

$$\begin{pmatrix} b_0 \\ \vdots \\ b_{t-1} \end{pmatrix} = P_t^{-1} \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(t) \end{pmatrix} = P_t^{-1} \begin{pmatrix} a \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Where the $t \times t$ matrix $P_t$, which can be shown to be invertible, is given by

$$P_t = \begin{pmatrix} 1 & a & 0 & 0 & \ldots & 0 \\ a & 1 & a & 0 & \ldots & 0 \\ 0 & a & 1 & a & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & a \\ 0 & 0 & 0 & \ldots & a & 1 \end{pmatrix}$$

where $a = \beta/(1+\beta^2) = \rho(\pm 1)$.

For general ARMA(p,q) models, the solutions of the prediction equations do not have nice explicit forms. They can however be calculated using an efficient recursion (the Durbin-Levinson algorithm), which is beyond our current scope. Further details are available in Section 8.2 of Brockwell and Davis.

## 4.3 Minimum mean squared error prediction

### 4.3.1 Concept

Suppose the random variables $z_1, \ldots, z_n$ and $y$ are defined on the same probability space $(\Omega, \mathcal{F}, P)$, and have finite second moments.

**Definition 4.2** *The minimum mean squared error prediction of $y$ in terms of $z_1 \ldots z_n$ is the function $m(z_1, \ldots, z_n)$ such that*

$$\mathsf{E}\left[\{y - m\left(z_1 \ldots z_n\right)\}^2\right] = \inf_f \mathsf{E}\left[\{y - f\left(z_1 \ldots z_n\right)\}^2\right].$$

It can be shown that $m(z_1 \ldots z_n) = \mathsf{E}(y|z_1 \ldots z_n)$.

### 4.3.2 Forecast for general ARIMA(p,d,q) processes

In what follows, we'll see a recursive approach to calculating the minimum mean square error predictor of $X_{t+h}$ based on $X_1 \ldots X_t$ only.

Suppose we have a causal and invertible ARIMA(p,d,q) process

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t,$$

and only $X_k$, $k = 1 \ldots t$ are available. We want to predict $X_{t+h}$.

Let

$$\phi(B)(1-B)^d = 1 - \sum_{j=1}^{p+d} \alpha_j B^j, \quad \theta(B) = 1 + \sum_{j=1}^{q} \beta_j B^j, \quad \mathbf{X_t} = (X_t, \ldots, X_1)^T, \quad \widehat{X}_{t+h} = \mathsf{E}(X_{t+h}|\mathbf{X_t}).$$

We have

$$X_t = \sum_{j=1}^{p+d} \alpha_j X_{t-j} + Z_t + \sum_{j=1}^{q} \beta_j Z_{t-j} \tag{13}$$

and

$$\widehat{X}_{t+h} = \mathsf{E}(X_{t+h}|\mathbf{X_t}) = \sum_{j=1}^{p+d} \alpha_j \mathsf{E}(X_{t+h-j}|\mathbf{X_t}) + \sum_{j=1}^{q} \beta_j \mathsf{E}(Z_{t+h-j}|\mathbf{X_t}) \quad h \geq 1.$$

To simplify this, note that

$$\widehat{X}_{t+h-j} = \mathsf{E}(X_{t+h-j}|\mathbf{X_t}) = X_{t+h-j}, \quad j \geq h. \tag{14}$$

Further, causality gives $\widehat{Z}_{t+h-j} = 0$ for $j < h$.
This then gives

$$\widehat{X}_{t+h} = \sum_{j=1}^{h-1} \alpha_j \widehat{X}_{t+h-j} + \sum_{j=h}^{p+d} \alpha_j X_{t+h-j} + \sum_{j=h}^{q} \beta_j \widehat{Z}_{t+h-j}, \quad h \geq 1.$$

The estimators $\widehat{Z}_k$ for $k \leq t$ can be obtained through rearranging the defining equation of the process, (13) and again using (14):

$$\widehat{Z}_k = X_k - \sum_{j=1}^{p+d} \alpha_j X_{k-j} - \sum_{j=1}^{q} \beta_j \widehat{Z}_{k-j}, \tag{15}$$

with initial values set to

$$\widehat{Z}_l = 0, \quad l = 1, \ldots, \max(p+d, q).$$

**Example** Consider the ARIMA(1,2,1) process

$$(1 - 0.5B)(1 - B)^2 X_t = (1 + 0.2B) Z_t \tag{16}$$

This process is the same as

$$(1 - 2.5B + 2B^2 - 0.5B^3) X_t = (1 + 0.2B) Z_t,$$

which is

$$X_t = 2.5X_{t-1} - 2X_{t-2} + 0.5X_{t-3} + Z_t + 0.2Z_{t-1}.$$

The prediction of $X_{t+2}$ is

$$\widehat{X}_{t+2} = 2.5\widehat{X}_{t+1} - 2X_t + 0.5X_{t-1}$$

with

$$\widehat{X}_{t+1} = 2.5X_t - 2X_{t-1} + 0.5X_{t-2} + 0.2\widehat{Z}_t.$$

Notice that (16) can be written as

$$Z_t = X_t - 2.5X_{t-1} + 2X_{t-2} - 0.5X_{t-3} - 0.2Z_{t-1}.$$

$\max\{p+d, q\}$ here is 3. So, we set the initial values $\widehat{Z}_1 = \widehat{Z}_2 = \widehat{Z}_3 = 0$, and obtain the remaining $\widehat{Z}_k$ for $k \leq t$, through

$$\widehat{Z}_k = X_k - 2.5X_{k-1} + 2X_{k-2} - 0.5X_{k-3} - 0.2\widehat{Z}_{k-1}, \quad 3 < k \leq t.$$

On substituting back through, this gives a recursive forecast for $X_t$.

## 4.4 Forecast for general ARMA(p,q) processes

## 4.5 The partial ac.f.

Given two random variables $\xi$ and $\eta$, denote

$$\mathsf{Corr}(\xi, \eta) = \frac{Cov(\xi, \eta)}{\sqrt{\mathsf{Var}(\xi)}\sqrt{\mathsf{Var}(\eta)}},$$

i.e., the usual correlation coefficient.

The partial ac.f. of a zero mean stationary process $X_t$, $t \in \mathbf{Z}$ is defined by

$$a(k) = \mathsf{Corr}(X_{k+1} - \Pi(X_{k+1}|X_2, \ldots, X_k), X_1 - \Pi(X_1|X_2, \ldots, X_k)),\ k \geq 2$$

and by convention $a(1) = \rho(1)$. This is another important characteristic of a weakly stationary stochastic process.

**Example 4.2** *The partial autocorrelation function of the white noise process coincides with the autocorrelation function of the process. This can be seen by considering the prediction equations, which in this case give* $\Pi(X_{k+1}|X_2, \ldots, X_k) = 0$ *and* $\Pi(X_1|X_2, \ldots, X_k) = 0$.

An equivalent definition of the partial autocorrelation function is given below. Proving equivalence is algebraically involved, and certainly not examinable. For completeness, a derivation of the equivalence of the two definitions can be found in Brockwell and Davis, Corollary 5.2.1.

Let $b_{ki}$, $i = 1, \ldots, k$, $k \geq 1$ be the coefficients in the representation

$$\Pi(X_{k+1}|X_1, \ldots, X_k) = \sum_{i=1}^{k} b_{ki} X_{k+1-i}$$

From the prediction equations

$$\mathsf{E}((X_{k+1} - \Pi(X_{k+1}|X_1, \ldots, X_k))X_i) = 0,\ i = 1, \ldots, k$$

we obtain that the coefficients $b_{ki}$ can be found from the following system of equations

$$\sum_{i=1}^{k} b_{ki}\rho(j-i) = \rho(j),\ j = 1, \ldots, k \tag{17}$$

Then the partial ac.f. at lag $k \geq 2$ is

$$a(k) = b_{kk}.$$

## 4.6 The partial autocorrelation function for AR(p) processes

For the AR(p) process, it can be shown that

$$a(k) = 0, \quad k > p.$$

Indeed, consider for simplicity a zero mean AR(1) process

$$X_t = \alpha X_{t-1} + Z_t,$$

where $|\alpha| < 1$.
By definition

$$a(1) = \mathsf{Corr}(\alpha X_1 + Z_2, X_1) = \rho(1) = \alpha.$$

Let $k \geq 2$, then, as we know from the prediction section,

$$\Pi(X_{k+1}|X_2, \ldots, X_k) = \alpha X_k.$$

A similar direct argument using the prediction equations shows that

$$\Pi(X_1|X_2, \ldots, X_k) = \alpha X_2.$$

Therefore

$$a(k) = \mathsf{Corr}(X_{k+1} - \alpha X_k, X_1 - \alpha X_2) = \mathsf{Corr}(Z_{k+1}, X_1 - \alpha X_2) = 0, \quad k > 1,$$

because the process $X_t$ is uncorrelated with future values of the white noise process.

Similar computations can be done for an arbitrary AR(p) process. In brief, if $X_t$ is a causal AR(p) process,

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t,$$

then it can be shown that for $k > p$,

$$\Pi(X_{k+1}|X_k \ldots X_2) = \alpha_1 X_k + \ldots + \alpha_p X_{k-p+1}.$$

Then

$$\begin{aligned} a(k) &= \mathsf{Corr}(X_{k+1} - \Pi(X_{k+1}|X_k \ldots X_2), X_1 - \Pi(X_1|X_2, \ldots, X_k)) \\ &= \mathsf{Corr}(Z_{k+1}, X_1 - \Pi(X_1|X_2, \ldots, X_k)), \end{aligned}$$

which is zero since the righthand correlation argument is a linear combination of $X_1 \ldots X_k$, each of which is uncorrelated with $Z_{k+1}$.

## 4.7 Summary of acf and pacf behaviour for ARMA processes

The following table summarizes the behaviour of the autocorrelation function and partial autocorrelation functions of the different classes of process.

| | AR(p) | MA(q) | ARMA(p,q) |
|---|---|---|---|
| ACF | Tails off | Cuts off after lag $q$ | Tails off |
| pACF | Cuts off after lag $p$ | Tails off | Tails off |

Note that *tailing off* can include damped oscillatory behaviour.

## 4.8 The sample partial autocovariance function

The sample autocovariance function is a point estimator of the ac.f. of a stationary stochastic process. The *sample partial ac.f.* $a_k$, $k \in \mathbf{Z}$, is used as a point estimator of the partial ac.f. of a stochastic process and is defined as follows. First, estimate the ac.f. using the sample ac.f. and then calculate the estimates of the partial ac.f. by replacing the autocorrelations $\rho(k)$ in the equation (17) with the sample autocorrelations $r_k$, to give the system

$$\sum_{i=1}^{k} \widehat{b}_{ki} r_{j-i} = r_j, \; j = 1, \ldots, k$$

which is then solved for $\widehat{b}_{ki}$ and defining $a_k = \widehat{b}_{kk}$.

# 5 Elements of statistical inference for time series

## 5.1 Model selection, parameter estimation and verification

### 5.1.1 The Box-Jenkins methodology for model building

Suppose that we have been presented with a time series, for which we seek an adequate model (e.g. for prediction). Assume that all necessary preliminary transformations have been made and any cyclic component has been removed. If data still appear to be non-stationary and it is due to a trend, then we remove the trend by differencing. In practice, one or two differences often suffices. Then we fit an ARMA model to the stationary time series $Y_t = X_t - \mu$, where $\mathsf{E}(X_t) = \mu$.

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \ldots + \alpha_p(X_{t-p} - \mu) + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}.$$

**Overdifferencing** Note that in seeking to obtain a stationary series, it is important not to *overdifference* time series data. Though the difference of a stationary process is a stationary process, overdifferencing introduces unnecessary correlations and complicates the model. For example, suppose the time series $X_t$ is a random walk,

$$X_t = X_{t-1} + Z_t$$

then its first difference is

$$Z_t = X_t - X_{t-1}$$

the white noise process. But the second difference

$$Y_t = Z_t - Z_{t-1}$$

is a non invertible MA(1) process.

The *Box-Jenkins methodology* is an iterative model-building procedure, which consists of the following four steps.

1. Identification: decide on reasonable values for $p$, $d$ and $q$.
2. Estimation: using the values of $p$ and $q$, estimate the unknown parameters: $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$, $\mu$ and $\sigma^2$.
3. Diagnostic checking: check the model against historical data to see whether it accurately describes the underlying process that generates the series.
4. If the model doesn't fit well, repeat earlier steps using an improved model. If the model fit is adequate, begin forecasting.

### 5.1.2 Model selection/identification

The main tools in model identification are the sample ac.f. $r_k$ and the sample partial ac.f. $a_k$.

Typical patterns of $r_k$ and $a_k$ for MA processes:

- The ac.f. function of a MA(q) process cuts off after lag $q$.
- The partial ac.f. of a MA(q) process is, in general, a mixture of exponentials and damped sine waves (i.e., its asymptotic behaviour at infinity is similar to the asymptotic behaviour of the ac.f. of an AR(p) process).
- If an MA process is thought to be appropriate for a given data set, then the order of the process is usually evident from the sample ac.f.

Typical patterns pf $r_k$ and $a_k$ of AR processes:

- The ac.f. of an AR(p) process is, in general, a mixture of exponentials and damped sine waves, and is usually of little help in identifying the process order.
- The partial ac.f. function of an AR(p) process cuts off after lag $p$.
- If an AR process is thought to be appropriate for a given set of data, then the order of the process is usually evident from the sample partial ac.f.

It can be proved that for an underlying AR(p) process, the approximate sampling distribution of each $a_k$ with $k > p$ is normal with zero mean and variance $1/N$. Hence

$$\mathsf{P}\{|a_k| \leq 1.96/\sqrt{N}\} = 0.95$$

Therefore the confidence limits $\pm 1.96/\sqrt{N} \approx \pm 2/\sqrt{N}$ can be used to detect the cut off effect in the sample partial correlogram for an AR process. A similar result holds for the autocorrelation coefficients $r_k$ of a MA(q) process.

Observed coefficients that fall outside these limits are significantly different from zero at the 5% level. But note that even if a coefficient should be zero in the true underlying process, the probability of getting at least one observed coefficient outside the confidence limits increases with the number of coefficients plotted.

In practice: Consider a sample from a white noise process. If, say, the first 20 values of $r_k$ are plotted, then we can expect one significant value (at 5% level) on average. So, if just one or coefficients are significant, the size and lag of these coefficients must be taken into account when deciding if a set of data is random. A single coefficient just outside the 95% confidence limits may be ignored (consistent with being a realisation of a white noise process), but two or more values well outside the limits can be considered as an indication of significant autocorrelation (or partial autocorrelation) at the lags in question.

"Recipe" for visual inspection:

- A correlogram that decays to zero suggests that the series is stationary and one can search for an appropriate ARMA model;
- if there is a sharp cut-off in the correlogram, i.e., all $r_k$ with $k > q$ for some $q$ are within $\pm 2/\sqrt{N}$ limits, then the behaviour of the sample partial ac.f. can be neglected and an MA(q) process can be used as a possible model for the data;
- if there is a sharp cut-off in the partial correlogram, i.e., all $r_k$ with $k > p$ for some $p$ are within $\pm 2/\sqrt{N}$ limits, then an AR(p) process can be used as a possible model for the data;
- if neither correlogram nor partial correlogram cuts off then possibly a mixed ARMA(p,q) should be taken as a model. Usually, ARMA(1,1) is tested first.

For example:

- If $r_1$ is significantly different from zero but all subsequent values of $r_k$ are all close to zero, then the behaviour of the sample partial ac.f. can be neglected and an MA(1) process can be used as a possible model for the data.
- If $r_k$ appear to be decreasing exponentially and the partial correlogram cuts off at lag 1, then an AR(1) may be appropriate.

**Example 5.1** *Given a data set with* 120 *observations (of a stationary times series), the following values of the sample ac.f. was computed*

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| *Sample ac.f.* | 1 | −0.52 | −0.04 | 0.13 | −0.09 | −0.01 | 0.1 |

*Find a suitable ARMA model for the data.*

*Answer:* A 95%CI is $(-2/\sqrt{N}, 2/\sqrt{N}) = (-0.183, 0.183)$. It is easy to see that the sample ac.f. cuts at lag 2. This might indicate that an MA(1) model can be taken a possible candidate.

**Example 5.2** *Given a data set with* 120 *observations (of a stationary times series), the following values of the sample ac.f. and the sample partial ac.f. were computed*

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| *Sample ac.f.* | 1 | −0.52 | −0.04 | 0.13 | −0.09 | −0.01 | 0.1 |
| *Sample partial ac.f.* | 1 | −0.52 | −0.43 | −0.21 | −0.2 | −0.23 | −0.1 |

*Find a suitable ARMA model for the data.*

*Answer.* For both the sample ac.f. and the sample partial ac.f. a 95%CI is $(-2/\sqrt{N}, 2/\sqrt{N}) = (-0.183, 0.183)$. It is easy to see that the sample ac.f. cuts at lag 2 and that the sample partial ac.f. decays. This might indicate that an MA(1) model can be taken a possible candidate.

**Example 5.3** *Given a data set with* 100 *observations (of a stationary times series), the following values of the sample ac.f. and the sample partial ac.f. were computed*

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| *Sample ac.f.* | 1 | 0.9 | 0.81 | 0.729 | 0.657 | 0.59 | 0.532 |
| *Sample partial ac.f.* | 1 | 0.9 | 0.1 | 0.12 | 0.07 | 0.062 | 0.03 |

*Find a suitable ARMA model for the data.*

*Answer.* For both the sample ac.f. and the sample partial ac.f. a 95%CI is $(-2/\sqrt{N}, 2/\sqrt{N}) = (-0.2, 0.2)$. It is easy to see that the sample ac.f. decays exponentially ($\sim 0.9^k$) and the sample partial ac.f. cuts at lag 2. Therefore, one can try to fit AR(1) process to this data set.

**Example 5.4** *Given a data set with* 100 *observations (of a stationary times series), the following values of the sample ac.f. and the sample partial ac.f. were computed*

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| *Sample ac.f.* | 1 | 0.9 | 0.8 | 0.6 | 0.5 | 0.3 | 0.22 |
| *Sample partial ac.f.* | 1 | 0.9 | 0.5 | 0.1 | 0.03 | 0.07 | 0.04 |

*Find a suitable ARMA model for the data.*

*Answer.* For both the sample ac.f. and the sample partial ac.f. a 95%CI is $(-2/\sqrt{N}, 2/\sqrt{N}) = (-0.2, 0.2)$. It is easy to see that the sample ac.f. decays exponentially and the sample partial ac.f. cuts at lag 3. Therefore, one can try to fit AR(2) process to this data set.

## 5.2 Estimating parameters of an ARMA process

### 5.2.1 Method of moments

For reasonably large samples, we expect the sample moments to be close to their theoretical population values. This gives a method of estimating parameters of the underlying process: we equate theoretical values of the moments in terms of parameters to the observed sample values, and solve to obtain parameter estimates. The idea should be clear after seeing a few examples.

**Example** The moment estimator of the process mean of a stationary time series $X_t$ is

$$\widehat{\mu} = \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

**Example** Consider a zero-mean AR(2) process: $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + Z_t$.

The Yule-Walker equations for $\rho(1)$ and $\rho(2)$ are

$$\rho(1) = \alpha_1 + \alpha_2\rho(1)$$
$$\rho(2) = \alpha_1\rho(1) + \alpha_2$$

Replacing $\rho(1)$ and $\rho(2)$ by their sample equivalents $r_1$ and $r_2$ we get the equations

$$r_1 = \alpha_1 + \alpha_2 r_1$$
$$r_2 = \alpha_1 r_1 + \alpha_2.$$

Solving these equations we get

$$\widehat{\alpha}_1 = \frac{r_1(1-r_2)}{1-r_1^2}$$
$$\widehat{\alpha}_2 = \frac{r_2 - r_1^2}{1-r_1^2}.$$

**Example** Consider a zero-mean invertible MA(1) process: $X_t = Z_t + \beta Z_{t-1}$. We have seen already that $\rho(1) = \frac{\beta}{1+\beta^2}$, therefore, replace $\rho(1)$ by the corresponding sample coefficient

$$r_1 = \frac{\beta}{1+\beta^2}$$

and solve this quadratic equation for $\beta$.

$$\widehat{\beta} = \frac{1 \pm \sqrt{1-4r_1^2}}{2r_1}.$$

Invertibility means that we must select the negative root, to ensure $|\beta| < 1$.

### 5.2.2 Least-squares estimation for AR processes

Consider a stationary AR(p) process

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \ldots + \alpha_p(X_{t-p} - \mu) + Z_t. \tag{18}$$

Given $N$ observations $x_1, \ldots, x_N$, the parameters $\mu, \alpha_1, \ldots, \alpha_p$ may be estimated by minimizing

$$S(\mu, \alpha_1, \ldots, \alpha_p) = \sum_{t=p+1}^{N} (x_t - \mu - \alpha_1(x_{t-1} - \mu) - \ldots - \alpha_p(x_{t-p} - \mu))^2.$$

Consider in detail the case $p = 1$.

$$S(\mu, \alpha) = \sum_{t=2}^{N} (x_t - \mu - \alpha(x_{t-1} - \mu))^2.$$

The values $\widehat{\mu}$ and $\widehat{\alpha}$ that minimise the quadratic from $S(\mu, \alpha)$ can be found as solutions of the following system of equations

$$\frac{\partial S(\mu, \alpha)}{\partial \mu} = 0,$$
$$\frac{\partial S(\mu, \alpha)}{\partial \alpha} = 0$$

This gives

$$2(1 - \widehat{\alpha}) \sum_{t=2}^{N} (x_t - \widehat{\mu} - \widehat{\alpha}(x_{t-1} - \widehat{\mu})) = 0,$$
$$-2 \sum_{t=2}^{N} (x_{t-1} - \widehat{\mu})(x_t - \widehat{\mu} - \widehat{\alpha}(x_{t-1} - \widehat{\mu})) = 0.$$

Note that $\widehat{\alpha}$ cannot be equal to 1, therefore from the first equation we get that

$$\sum_{t=2}^{N} (x_t - \widehat{\mu} - \widehat{\alpha}(x_{t-1} - \widehat{\mu})) = 0.$$

Simple algebra gives

$$\widehat{\mu} = \frac{N}{N-1} \overline{x} + d_N,$$

where

$$\overline{x} = \frac{1}{N} \sum_{t=1}^{N} x_t$$

and

$$d_N = \frac{\widehat{\alpha} x_N - x_1}{(N-1)(1 - \widehat{\alpha})}.$$

For large $N$, $d_N \to 0$ and so

$$\widehat{\mu} \approx \overline{x}$$

in the sense that $\widehat{\mu}/\overline{x} \to 1$ as $N \to \infty$.
From the equation $\partial S/\partial\alpha = 0$ we get

$$\sum_{t=2}^{N}(x_{t-1} - \widehat{\mu})(x_t - \widehat{\mu}) - \widehat{\alpha}\sum_{t=1}^{N-1}(x_t - \widehat{\mu})^2 = 0.$$

From this we can find

$$\widehat{\alpha} = \frac{\sum_{t=1}^{N-1}(x_t - \widehat{\mu})(x_{t+1} - \widehat{\mu})}{\sum_{t=1}^{N}(x_t - \widehat{\mu})^2}\frac{1}{v_N}$$

where

$$v_N = 1 - \frac{(x_N - \widehat{\mu})^2}{\sum_{t=1}^{N}(x_t - \widehat{\mu})^2}$$

*(NB An error in the definition of $v_N$ from an earlier version of the notes has been corrected.)*

It can be shown that $v_N \to 1$ (in a certain sense) as $N \to \infty$. Recalling that

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{N}(x_t - \overline{x})^2}$$

and using the approximation $\widehat{\mu} \approx \overline{x}$ for large $N$, we get that asymptotically

$$\widehat{\alpha} \approx r_1$$

in the sense that $\widehat{\alpha}/r_1 \to 1$ as $N \to \infty$.

Similar computations can be done for an arbitrary AR(p) process For example, if $p = 2$, then we (approximately) recover the moment estimators computed in the previous section:

$$\widehat{\mu} \approx \overline{x},\ \ \widehat{\alpha}_1 \approx r_1(1 - r_2)/(1 - r_1^2),\ \ \widehat{\alpha}_2 \approx (r_2 - r_1^2)/(1 - r_1^2).$$

For the general AR(p) process (18), we seek the minimizer $(\widehat{\mu}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$ of

$$S(\mu, \alpha_1 \ldots \alpha_p) = \sum_{t=p+1}^{N}\{x_t - \mu - \alpha_1(x_{t-1} - \mu) - \ldots - (x_{t-p} - \mu)\}^2.$$

Let $\mathbf{Y} = (x_{p+1}, \ldots, x_N)^T$, $\zeta = \left(\mu(1 - \sum_{j=1}^{p}\alpha_j), \alpha_1, \ldots, \alpha_p\right)$ and

$$\mathbf{H} = \begin{pmatrix} 1 & x_p & x_{p-1} & \cdots & x_2 & x_1 \\ 1 & x_{p+1} & x_p & \cdots & x_3 & x_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{N-1} & x_{N-2} & \cdots & x_{N-p+1} & x_{N-p} \end{pmatrix}$$

Then $S$ can be written as a quadratic form

$$S(\mu, \alpha_1 \dots \alpha_p) = (\mathbf{Y} - \mathbf{H}\zeta)^T (\mathbf{Y} - \mathbf{H}\zeta),$$

whose minimiser is can be found by differentiation to be

$$\widehat{\zeta} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1} \mathbf{H}^T\mathbf{Y}.$$

The form of this expression should be familiar from linear models.

### 5.2.3 Least square estimation for MA processes

LSE is not so straightforward for MA processes as for AR processes. Consider, for example, a MA(1) process

$$X_t = \mu + \beta Z_{t-1} + Z_t.$$

Given observations $x_1, \dots, x_N$ we would like to write the residual sum of squares $\sum z_t^2$ in terms of observed $x_1, \dots, x_n$ and parameters $\mu$ and $\beta$, as we did in the case of the AR(1) process. This is not possible here, so the explicit least squares estimates cannot be found. Instead, the following iterative procedure is used:

- select suitable starting values for $\mu$ and $\beta$, for example $\widehat{\mu} = \overline{x} = (\sum_{k=1}^N x_k)/N$ and $\widehat{\beta}$ a solution of the moment equation

  $$r_1 = \frac{\beta}{1+\beta^2},$$

  where $r_1$ is the value of the sample ac.f. at lag 1, (one must choose the solution $|\widehat{\beta}| < 1$),

- taking $z_0 = 0$, calculate $z_1 = x_1 - \widehat{\mu}$, then $z_2 = x_2 - \widehat{\mu} - \widehat{\beta} z_1$, and so on until $z_N = x_N - \widehat{\mu} - \widehat{\beta} z_{N-1}$, and finally calculate the residual sum $\sum_{t=1}^N z_t^2$ for chosen $\widehat{\mu}$ and $\widehat{\beta}$,

- repeat the procedure for the other neighbouring values of $\mu$ and $\beta$ so that the residual sum of squares $\sum_{t=1}^N z_t^2$ is computed on a grid of points in the $(\mu, \beta)$ plane,

- determine by visual inspection of otherwise (by an iterative optimization procedure) the values of $\mu$ and $\beta$ that minimize $\sum_{t=1}^N z_t^2$. *These values are least square estimates.*

## 5.3 Maximum likelihood estimation for Gaussian ARMA(p,q) processes

Consider a causal, invertible ARMA(p,q) process $X_t$, $t \in \mathbf{Z}$,

$$X_t - \mu = \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu) + Z_t + \sum_{k=1}^{q} \beta_k Z_{t-k}, \tag{19}$$

where $Z_t$, $t \in \mathbf{Z}$, are independent random variables with zero mean and variance $\sigma^2$.

If $Z_t$, $t \in \mathbf{Z}$, are normally distributed, then the process $X_t$, $t \in \mathbf{Z}$, is said to be a *Gaussian* ARMA process. In this case, for any $t_1, \ldots, t_n \in \mathbf{Z}$ the probability distribution of the random vector $(X_{t_1}, \ldots, X_{t_n})$ is a multivariate normal distribution (see definition 5.1 below) with mean $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ and covariance matrix $\Sigma$ with entries $\Sigma_{ij} = \mathsf{Cov}(X_{t_i}, X_{t_j}) = \gamma(t_i - t_j)$.

**Definition 5.1** *The random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is said to be multivariate normal if there exist a column vector $\boldsymbol{\mu}$, a $(n \times n)$-matrix $B$ and a random vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ with independent standard normal components such that*

$$\boldsymbol{Y} = \boldsymbol{\mu} + B\boldsymbol{\eta}.$$

*The mean of $\boldsymbol{Y}$ is the vector $\boldsymbol{\mu}$ with entries $\mu_i = \mathsf{E}(Y_i)$ and the covariance matrix of $\boldsymbol{Y}$ is $\Sigma = BB^T$, with entries $\Sigma_{ij} = \mathsf{Cov}(Y_i, Y_j)$. Provided that $\det(\Sigma) > 0$, the density function of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is*

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right), \ \boldsymbol{y} = (y_1, \ldots, y_n)^T \in \boldsymbol{R}^n.$$

Taking the process (19) to be Gaussian, we would like to carry out maximum likelihood estimation of the parameters $\mu$, $\sigma^2$, $\alpha_j$ for $j = 1 \ldots p$ and $\beta_k$ for $k = 1 \ldots q$. The likelihood function is just the joint density of $(X_1 \ldots X_N)^T$, considered as a function of the parameters, for fixed observations. For general ARMA models, it is difficult to express the likelihood as an explicit function of the parameters. In the section that follows we will show how to obtain a *conditional* maximum likelihood estimator, which will be close to the full maximum likelihood estimator for sufficiently large sample sizes.

### 5.3.1 Conditional MLE

By working with $Y_t = X_t - \mu$, it is enough to develop maximum likelihood estimation for a zero-mean process. It is straightforward to incorporate $\mu$ as an additional parameter.

Define

$$\boldsymbol{Y}_t = (Y_t \ldots Y_1), \qquad \boldsymbol{\theta} = (\alpha_1 \ldots \alpha_p, \beta_1 \ \ldots \beta_q).$$

The conditional density function of $Y_t$ given $\boldsymbol{Y}_{t-1}$ is denoted $f(y_t | \boldsymbol{y}_{t-1}, \boldsymbol{\theta}, \sigma^2)$, and the density function of $\boldsymbol{Y}_p$ is $f(\boldsymbol{y}_p | \boldsymbol{\theta}, \sigma^2)$. The likelihood, which is just the joint density of $\boldsymbol{Y}_N$, can then be written as follows, by using laws of conditional probability.

$$\begin{aligned} L(\boldsymbol{Y}_N | \boldsymbol{\theta}, \sigma^2) &= f(\boldsymbol{y}_p | \boldsymbol{\theta}, \sigma^2) f(y_{p+1} \ldots y_N | \boldsymbol{y}_p, \boldsymbol{\theta}, \sigma^2) \\ &= f(\boldsymbol{y}_p | \boldsymbol{\theta}, \sigma^2) f(y_N | \boldsymbol{y}_{N-1}, \boldsymbol{\theta}, \sigma^2) f(\boldsymbol{y}_{N-1} | \boldsymbol{y}_p, \boldsymbol{\theta}, \sigma^2) \\ &= \ldots \\ &= f(\boldsymbol{y}_p | \boldsymbol{\theta}, \sigma^2) \prod_{t=p+1}^{N} f(y_t | \boldsymbol{y}_{t-1}, \boldsymbol{\theta}, \sigma^2) \end{aligned}$$

The expression

$$\prod_{t=p+1}^{N} f(y_t|\boldsymbol{y}_{t-1}, \boldsymbol{\theta}, \sigma^2),$$

which will be denoted $L^*(\boldsymbol{y}_N, \boldsymbol{\theta}, \sigma^2)$, is known as the *conditional likelihood function* of $\mu$, $\sigma^2$ and $\boldsymbol{\theta}$.

An important property of the multivariate normal distribution is that its conditional distributions are also multivariate normal (see Appendix C3 of Shumway and Stoffer for more details). Moreover, a multivariate normal distribution is determined by its mean vector and covariance matrix. This means that to determine the density $f(y_t|\boldsymbol{y}_{t-1}, \boldsymbol{\theta}, \sigma^2)$, it is enough to work out $\mathsf{E}(Y_t|\boldsymbol{Y}_{t-1})$ and $\mathsf{Var}(Y_t|\boldsymbol{Y}_{t-1})$.

We begin by making the assumption that

$$Z_1 = Z_2 = \ldots = Z_q = 0,$$

which will be reasonable in any practical situation where estimation is required.

Now, assume that $Y_1 \ldots Y_{t-1}$ are known. From the defining equation it can be seen that

$$Y_t = \sum_{k=1}^{p} \alpha_k Y_{t-k} + Z_t + \sum_{k=1}^{q} \beta_k Z_{t-k},$$

and the values of $Z_i$ for $i = q+1 \ldots t-1$ can be obtained recursively from

$$Z_i = Y_i - \sum_{k=1}^{p} \alpha_k Y_{i-k} - \sum_{k=1}^{q} \beta_k Z_{i-k},$$

as in (15) when computing forecasts for general ARMA processes. This means that the conditional expectation and variance of $Y_t$ are given in terms of known quantities as

$$\begin{aligned}\mathsf{E}(Y_t|\boldsymbol{Y}_{t-1}) &= \sum_{k=1}^{p} \alpha_k Y_{t-k} + \sum_{k=1}^{q} \beta_k Z_{t-k} \\ \mathsf{Var}(Y_t|\boldsymbol{Y}_{t-1}) &= \mathsf{Var}(Z_t|\boldsymbol{Y}_{t-1}) = \sigma^2.\end{aligned}$$

It is convenient to denote $\mathsf{E}(Y_t|\boldsymbol{Y}_{t-1})$ by $\widehat{Y}_{t|t-1}$, and define the *innovation* at $t$ to be

$$\epsilon_t(\boldsymbol{\theta}) = y_t - \widehat{Y}_{t|t-1}.$$

This means that we can write the conditional likelihood in terms of the innovations as

$$L^*(\boldsymbol{y}_N, \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-(N-p)/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=p+1}^{N} \epsilon_t^2(\boldsymbol{\theta})\right\},$$

so that the full likelihood is given by

$$L(\boldsymbol{Y}_N|\boldsymbol{\theta}, \sigma^2) = f(\boldsymbol{y}_p|\boldsymbol{\theta}, \sigma^2)(2\pi\sigma^2)^{-(N-p)/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=p+1}^{N} \epsilon_t^2(\boldsymbol{\theta})\right\}. \tag{20}$$

Maximizing (20) is a non-linear optimization problem, which is computationally expensive in general. In practice it is often enough to maximize the conditional likelihood $L^*$. To maximize $L^*$, it is enough to find $\boldsymbol{\theta}$ such that the sum of squares

$$\sum_{t=p+1}^{N} \epsilon_t^2(\boldsymbol{\theta})$$

is minimal. The conditional maximum likelihood estimator of $\sigma^2$ is then given by

$$\widehat{\sigma^2} = \frac{1}{N-p} \sum_{t=p+1}^{N} \epsilon_t^2(\widehat{\boldsymbol{\theta}}).$$

Note that for reasonably large samples, the conditional MLEs and full MLEs will be typically be very close.

## 5.4 Model verification or diagnostic checking

Once an ARIMA(p,d,q) model has been fitted to a time series $X_1 \dots X_N$, the next step is to assess how well the model fits the data. We do this by analysing the residuals $z_1, \dots, z_N$, which can be obtained from the recursive algorithm given in (15). For $W_t = (1-B)^d X_t$, the residuals are

$$\widehat{Z}_t = W_t - \sum_{j=t}^{p} \widehat{\alpha}_j W_{t-j} - \sum_{k=1}^{q} \widehat{\beta}_k \widehat{Z}_{t-k}, \qquad t > \max\{p+d, q\},$$

and $\widehat{Z}_t = 0$ for $t \leq \max\{p+d, q\}$.

For a "good" model fit, the sequence of residuals $z_1, \dots, z_N$ should behave like a realisation of a white noise process. This means that we should expect:

- the mean of the residuals should be close to zero

$$\frac{z_1 + \ldots + z_N}{N} \approx 0$$

- the spread of the residuals around the mean is constant over time
- autocorrelations between residuals are negligible, i.e.,

$$r_{z,k} = \frac{\sum_{t=1}^{N-k}((z_t - \overline{z})(z_{t+k} - \overline{z})}{\sum_{t=1}^{N}(z_t - \overline{z})^2}$$

Under assumption that the residuals are uncorrelated, approximate 95% confidence limits are $\pm 2/\sqrt{N}$. If we observe significant autocorrelations, i.e., there are values of $r_{z,k}$ which are well outside these limits, then it is worth exploring other plausible models.

### 5.4.1 The Ljung-Box statistic

The Ljung-Box statistic can be used to test whether or not the autocorrelation function of a stationary process is zero. For a sample $X_1 \dots X_t$ from a stationary process with sample autocorrelation $r_k$ at lag $k$, the Ljung-Box statistic is defined as

$$Q = t(t+2) \sum_{k=1}^{m} \frac{r_k^2}{t-k},$$

where the integer $m$ is chosen arbitrarily. Under the null hypothesis that the model fit is adequate (so that the residuals are essentially white noise), the test statistic $Q$ has an asymptotic $\chi^2$ distribution with $m - p - q$ degrees of freedom. This means that for $Q$ larger than some critical value, we reject the null hypothesis

$$H_0 : \rho(k) = 0, \qquad k \neq 0.$$

### 5.4.2 Overfitting

After specifying and fitting the model one can try to fit a more general model. As an example, suppose that we fit an AR(2) model and estimated the parameters $\mu, \alpha_1, \alpha_2$. Then repeat the estimation procedure assuming AR(3) model. If

- additional parameter $\alpha_3 \approx 0$
- $\alpha_{1,new} \approx \alpha_{1,old}$ and $\alpha_{2,new} \approx \alpha_{2,old}$,

then it is reasonable to conclude that there is no need to replace the initial model AR(2) by a more general one. You will notice this principle in use in the solutions to lab class 3. Another approach is to use the Akaike Information Criterion (AIC) - see section 2.2 of Shumway and Stoffer for more details.

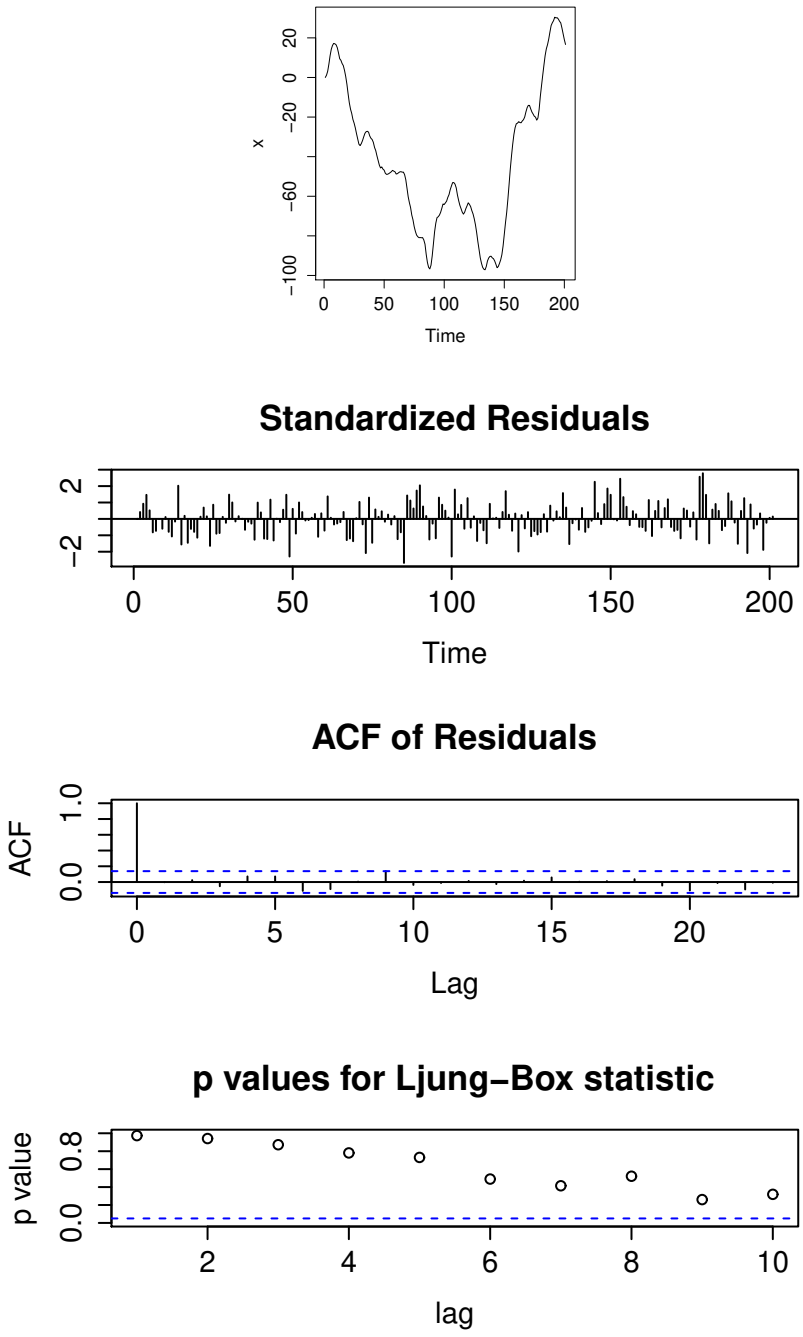


Figure 4: Good model fit for an ARIMA(2,1,2) model. No patterns in the residuals, and no significant p-values for the Ljung-Box statistic

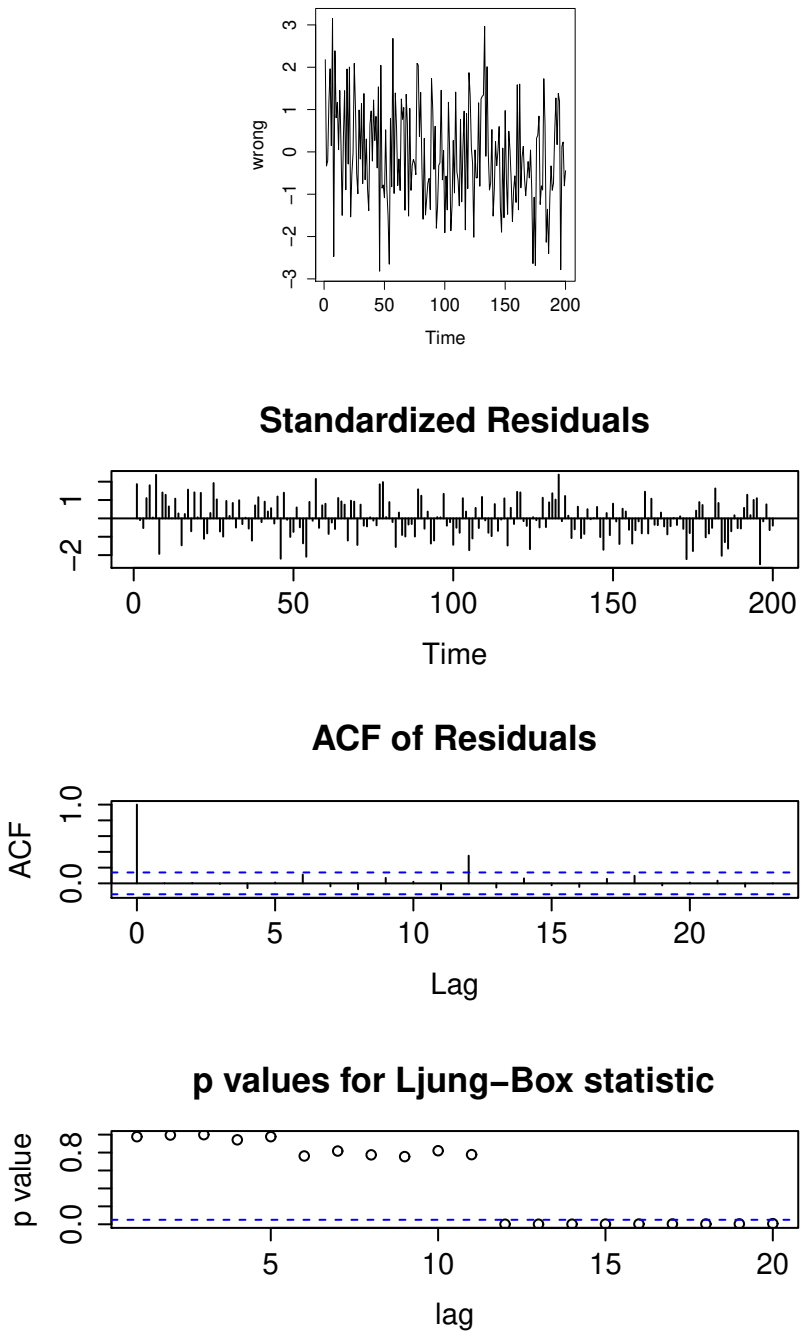


Figure 5: Fitting an AR(2) model to a dataset simulated with higher order autocorrelation. Note the significant autocorrelation in the residuals and significant p-values for the Ljung-Box test. This suggests some structure in the residuals remains unmodelled.

# 6 Spectral analysis

A time series can be considered to be a noisy observation of a curve at a set of time points. We can consider the curve as being made up of sine and cosine waves of different frequencies. (You may recall this idea from Fourier analysis). Fitting a model to a time series essentially means estimating the amplitude of the sine and cosine components at different frequencies. The *periodogram* is of use in this task.

## 6.1 The periodogram

Let $X_1 \dots X_N$ be a sample from a stationary time series, with $N = 2q+1$, an odd number. Write

$$X_t = A_0 + \sum_{i=1}^{q} A_i \cos(2\pi f_i t) + B_i \sin(2\pi f_i t) + e_t,$$

where $f_i = \frac{i}{N}$. Least squares estimates of $A_i$ and $B_i$, denoted with the corresponding lower case letters, can be obtained as follows

$$a_0 = \bar{X}, \qquad a_i = \frac{2}{N}\sum_{t=1}^{N} X_t \cos(2\pi f_i t) \qquad b_i = \frac{2}{N}\sum_{t=1}^{N} X_t \sin(2\pi f_i t), \qquad i = 1 \dots q. \tag{21}$$

This is a *saturated* model, i.e. $N$ parameters are being estimated with $N$ observations, so that we cannot obtain residuals $\widehat{e_t}$.

The *periodogram* is the set of $q$ intensity values

$$I(f_i) = \frac{N}{2}(a_i^2 + b_i^2), \qquad i = 1 \dots q.$$

Note that if instead $N = 2q$ is even, the values $a_q$ and $b_q$ have to be changed to

$$a_q = \frac{1}{N}\sum_{t=1}^{N}(-1)^t X_t, \qquad b_q = 0.$$

**Note:** If the frequency $f_i$ is indeed a component of the curve, the intensity $I(f_i)$ is expected to be relatively large.

## 6.2 The spectrum and spectral density function

Suppose $X_1 \dots X_N$ is a sample from a stationary time series with autocovariance function $\gamma(\cdot)$ and autocorrelation function $\rho(\cdot)$.

**The sample spectrum** For any frequency $0 \le f \le 0.5$, define

$$I(f) = \frac{N}{2}(a_f^2 + b_f^2),$$

where $a_f$ and $b_f$ are obtained by replacing $f_i$ by $f$ in (21). $I(f)$ is called the *sample spectrum* of $(X_t)$. It can be shown that

$$I(f) = 2\left[c_0 + 2\sum_{k=1}^{N-1} c_k \cos(2\pi f k)\right], \qquad 0 \le f \le 0.5,$$

where $c_k$ is the sample autocovariance at lag $k$.

**The power spectrum** The power spectrum is defined as

$$p(f) = \lim_{N\to\infty} \mathsf{E}\left[I(f)\right] = 2\left[\gamma(0) + 2\sum_{k=1}^{\infty} \gamma(k)\cos(2\pi f k)\right] \qquad 0 \le f \le 0.5.$$

Note that $\sum_{k=1}^{\infty} |\gamma(k)| < \infty$ is a sufficient condition for the convergence of the power spectrum. This is because $|\cos(x)| \le 1$ for real $x$, giving

$$|p(f)| \le 2\left[|\gamma(0)| + 2\sum_{k=1}^{\infty} |\gamma(k)|\right].$$

By integrating term-by-term and using the fact that $\int_0^{0.5} \cos(2\pi f k)\, df = 0$ for $k \neq 0$, it is clear that

$$\gamma(0) = \int_0^{0.5} p(f)\, df.$$

**The spectral density function** This is just a normalization of the power spectrum, which can therefore be expressed in terms of the autocorrelation function:

$$g(f) = \frac{p(f)}{\gamma(0)} = 2\left[1 + 2\sum_{k=1}^{\infty} \rho(k)\cos(2\pi f k)\right] \qquad 0 \le f \le 0.5.$$

Clearly $\int_0^{0.5} g(f)\, df = 1$.
The spectral density function shows the frequencies that dominate the variability in a time series, and guide preliminary choices of parametric models.

**Example** Let $Z_t, t = 1, 2, \ldots$ be white noise with $\mathsf{Var}(Z_t) = 1$. Consider two series
Series I: $X_t = 10 + Z_t + Z_{t-1}$.

$X_t$ has autocovariance function given by

$$\gamma(k) = \begin{cases} 2 & k = 0, \\ 1 & k = 1, \\ 0 & k \ge 2. \end{cases}$$

Its spectral density function is

$$g(f) = 2\left[1 + 2\sum_{k=1}^{\infty} \rho(k)\cos(2\pi f k)\right] = 2(1 + \cos(2\pi f)).$$

Series II: $X_t = 10 + Z_t - Z_{t-1}$.

$X_t$ has autocovariance function given by

$$\gamma(k) = \begin{cases} 2 & k = 0, \\ -1 & k = 1, \\ 0 & k \geq 2. \end{cases}$$

Analogously, its spectral density function is

$$g(f) = 2(1 - \cos(2\pi f)).$$

These two processes are dominated by different types of variation, as can be seen from the time plots, and the plots of the spectral densities.

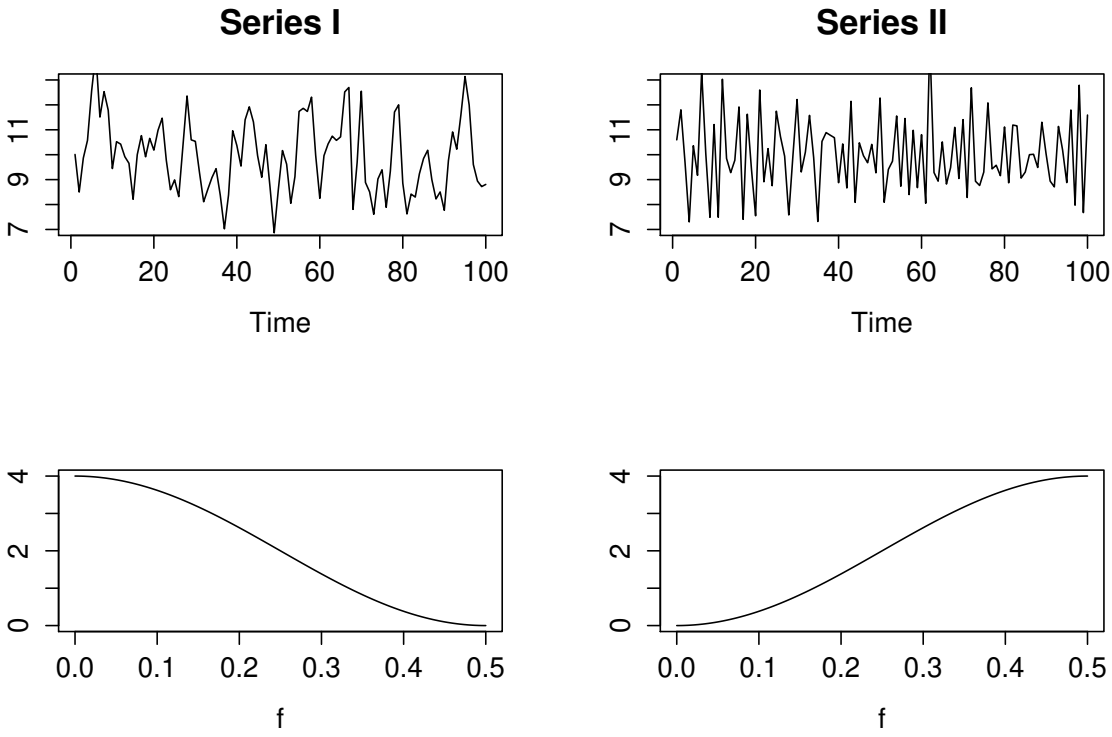


Figure 6: Spectral density functions for series I and II.

**Theorem 6.1** *If* $\sum_{k=0}^{\infty} |\gamma(k)| < \infty$, *then*

$$\gamma(k) = \int_0^{0.5} \cos(2\pi f k)\, p(f)\, df.$$

This says that the autocovariance function can be recovered if the power spectrum is known.

### 6.2.1 Spectral density of a linear filter

Let $\{Y_t,\, t \in \mathbf{Z}\}$ be a stationary process with power spectrum $p_Y(f)$. If $X_t$ is a linear filter of $Y_t$, i.e.,

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j},$$

where

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty,$$

Then $\{X_t,\, t \in \mathbf{Z}\}$ is also a stationary process with power spectrum

$$p_X(f) = |\psi(e^{-2\pi f i})|^2 p_Y(f), \quad 0 \le f \le 0.5,$$

where $i$ is the imaginary unit and

$$\psi(e^{-i2\pi f}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij2\pi f}.$$

The function $\psi(e^{-i2\pi f})$ is called the frequency response function or the transfer function of the filter. The function $|\psi(e^{-i2\pi f})|^2$ is called the power transfer function or the gain of the filter.

## 6.3 Computations of spectral density functions for some ARMA(p,q) processes

**Example 6.1 Purely random processes.** *Let $\{Z_t,\, t \in \boldsymbol{Z}\}$, be a zero mean white noise process with variance $\sigma^2$. Then*

$$\gamma(k) = \begin{cases} \sigma^2 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

*and*

$$g(f) = 2 \qquad 0 \le f \le 0.5.$$

**Example 6.2 MA(1) processes.**
*For the MA(1) process $X_t = Z_t + \beta Z_{t-1}$ we have two non-zero values of the autocovariance function: $\gamma(0) = (1+\beta^2)\sigma^2$ and $\gamma(1) = \beta\sigma^2$, hence*

$$p(f) = 2\sigma^2(1 + \beta^2 + 2\beta\cos(2\pi f))$$

$$g(f) = 2\left(1 + 2\frac{\beta\cos(2\pi f)}{1+\beta^2}\right)$$

**Example 6.3 AR(1) processes**
*For the AR(1) process $X_t = \alpha X_{t-1} + Z_t$, $|\alpha| < 1$. We can rearrange the defining equation to write*

$$Z_t = X_t - \alpha X_{t-1},$$

*so that white noise is expressed as a linear filter of the process* $X_t$*, with* $\psi_0 = 1$*,* $\psi_1 = -\alpha$ *and* $\psi_k = 0$ *for* $k \neq 0, 1$*. The transfer function is*

$$\psi(e^{-i2\pi f}) = 1 - \alpha e^{-i2\pi f},$$

*and the gain is just the magnitude of this,*

$$1 - 2\alpha\cos(2\pi f) + \alpha^2.$$

*Using the result on linear filters from above, this means that*

$$2\sigma^2 = (1 - 2\alpha\cos(2\pi f) + \alpha^2)p_X(f),$$

*so that (on normalizing)*

$$g_X(f) = \frac{2(1-\alpha^2)}{1 - 2\alpha\cos(2\pi f) + \alpha^2}.$$

Now, suppose $\{X_t,\, t \in \mathbf{Z}\}$ is an ARMA(p,q) process

$$X_t - \alpha_1 X_{t-1} - \ldots - \alpha_p X_{t-p} = Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q}$$

or,

$$\phi(B)X_t = \theta(B)Z_t$$

where

$$\phi(B) = 1 - \alpha_1 B - \ldots - \alpha_p B^p$$

and

$$\theta(B) = 1 + \beta_1 B + \ldots + \beta_q B^q.$$

If $\phi$ and $\theta$ do not have common zeroes and $\phi$ does not have zeroes on the unit circle, then

$$p_X(f) = 2\sigma^2 \frac{|\theta(e^{-i2\pi f})|^2}{|\phi(e^{-i2\pi f})|^2}, \quad 0 \leq f \leq 0.5.$$

**Example 6.4 ARMA(1,1) process** $X_t - \alpha X_{t-1} = Z_t + \beta Z_{t-1}$.

$$p_X(f) = \frac{2\sigma^2(1 + 2\beta\cos(2\pi f) + \beta^2)}{1 - 2\alpha\cos(2\pi f) + \alpha^2}$$

# 7 State-space models and the Kalman filter

## 7.1 Univariate state-space models

### 7.1.1 General form

**Definition 7.1** *A univariate state-space model is a stochastic process* $\{X_t,\, t \geq 1; \theta_t,\, t \geq 0\}$, *such that*

- $X_t \in \boldsymbol{R}$,
- $\theta_t^T = (\theta_{t,1}, \ldots, \theta_{t,k})$, *for some fixed* $k \geq 1$,

*and*

$$X_t = h^T \theta_t + n_t,\, t \geq 1, \tag{22}$$
$$\theta_t = G\theta_{t-1} + w_t,\, t \geq 1, \tag{23}$$
$$\theta_0 = \theta, \tag{24}$$

*where*

- $G = (G_{ij})$ *is a known* $(k \times k)$ *matrix,*
- $h$ *is a known* $(k \times 1)$ *column vector,* $h^T = (h_1, \ldots, h_k)$, *so*

$$h^T \theta_t = h_1 \theta_{t,1} + \ldots + h_k \theta_{t,k},$$

- $\{n_t,\, t \in \boldsymbol{Z}_+\}$ *and* $\{w_t^T = (w_{t,1}, \ldots, w_{t,k}),\, t \in \boldsymbol{Z}_+\}$, *are independent zero-mean white noise processes, with the variance* $\sigma_n^2$ *and the covariance matrix*

$$W = \mathsf{E}(w_t w_t^T) = (\mathsf{Cov}(w_{t,i}, w_{t,j}))_{i,j=1}^k \tag{25}$$

*respectively,*

*the initial value* $\theta_0$ *is uncorrelated with the noise processes (might be a constant vector).*

Terminology: $X_t$ is the *observation* at time $t$; $\theta_t$ is the *state* vector, a vector of *state variables*, a non-observable target process. Equation (22) is called the *observation* equation, equation (23) is called the *state* or *transition* equation.

Applications of the state-space models:

- Navigation
- Tracking missiles
- Extracting an object motion from video
- Computer vision applications
- Economics: forecasting economic indicators

The main problem in all these applications is prediction of unobservable state variable $\theta_t$ given observations $X_1, \ldots, X_t$. The Kalman filter is a recursive algorithm for computing the best linear predictor $\widehat{\theta}_t$ of $\theta_t$ in terms of observations $X_1, \ldots, X_t$.

### 7.1.2 The local level model

$$X_t = \theta_t + n_t \tag{26}$$
$$\theta_t = \theta_{t-1} + w_t \tag{27}$$

Here, equation (26) is the transition equation, and the state vector $\theta_t$ consists of a single variable $\theta_t \in \mathbf{R}$ which is called the local level. The unobservable local level $\theta_t$ is assumed to follow a random walk. Here $h = G = 1$. The noise process $\{n_t,\, t \in \mathbf{Z}_+\}$ and $\{w_t,\, t \in \mathbf{Z}_+\}$ are assumed to be uncorrelated with zero means and respective variances $\sigma_n^2$ and $\sigma_w^2$. If $\sigma_w^2 = 0$, then $\theta_t = \theta$ is constant and we get a constant-mean model

$$X_t = \theta + n_t \tag{28}$$

**Proposition 7.1** *The first difference $\nabla X_t$ of the local level model is a weakly stationary process with the ac.f. as the following MA(1) model*

$$Y_t = Z_t + \beta Z_{t-1}, \;\; t \geq 0,$$

*with $\beta = -1 + (\sqrt{c^2 + 4c} - c)/2$, where $c = \sigma_w^2/\sigma_n^2$.*

*Proof* This follows by direct computation.

$$\begin{aligned} \nabla X_t &= \theta_t - \theta_{t-1} + n_t - n_{t-1}. \\ &= w_t + n_t - n_{t-1}. \end{aligned}$$

We first compute the variance of $\nabla X_t$:

$$\gamma(0) = \mathsf{Var}(\nabla X_t) = \mathsf{Var}(w_t + n_t - n_{t-1}) = \sigma_w^2 + 2\sigma_n^2,$$

since the white noise terms are uncorrelated.

Now

$$\gamma(k) = \mathsf{Cov}(w_t + n_t - n_{t-1}, w_{t+k} + n_{t+k} - n_{t+k-1})$$

Note that if $k > 1$, there are no common indices in the left and right terms of the covariance, so that $\gamma(k) = 0$ for $k > 1$, as for the MA(1) process. For $k = 1$, we get

$$\gamma(1) = \mathsf{Cov}(w_t + n_t - n_{t-1}, w_{t+1} + n_{t+1} - n_t) = -\sigma_n^2.$$

This then gives

$$\rho(1) = \frac{-\sigma_n^2}{\sigma_w^2 + 2\sigma_n^2} = -\frac{1}{c+2}.$$

Since $\rho(1) = \frac{\beta}{1+\beta^2}$ for the MA(1) process, setting the above expressions equal and solving the resulting quadratic gives the stated value for $\beta$.

### 7.1.3 Linear growth model

The linear growth model is specified by these three equations

$$X_t = \mu_t + n_t \tag{29}$$
$$\mu_t = \mu_{t-1} + \beta_{t-1} + w_{1,t} \tag{30}$$
$$\beta_t = \beta_{t-1} + w_{2,t} \tag{31}$$

Equation (29) is the observation equation, two other equations are transition/state equations. The state vector $\theta_t^T = (\mu_t, \beta_t)$ has two components which are interpreted as follows: $\mu_t$ is the local level, $\beta_t$ is the local trend. Comparing these equations with the general form of the state-space model we obtain that $h^T = (1, 0)$ and

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

, which are clearly constant through time.

The components of the process $w_t^T = (w_{1,t}, w_{2,t})$ are assumed to be independent, so the covariance matrix of the process is

$$W = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

If $w_{1,t}$ and $w_{2,t}$ have zero variances, then the trend is deterministic

$$X_t = \mu_t + n_t \tag{32}$$
$$\mu_t = \mu_{t-1} + \beta = \mu_0 + \beta t. \tag{33}$$

The model is called a global linear trend model in this case. "Local linear trend" means that the trend is allowed to change.

**Proposition 7.2** *The second difference $\nabla^2 X_t$ of the linear growth model is a weakly stationary stochastic process and its ac.f. has the same structure as the ac.f. of an MA(2) model, i.e., $\rho(0) = 1,\ \rho(\pm 1) \neq 0,\ \rho(\pm 2) \neq 0$ and $\rho(k) = 0$, if $|k| > 2$.*

*Proof.* Again, this follows by direct computation with $\nabla^2 X_t$.

## 7.2 The Kalman filter

Let $\widehat{\theta}_t$ be the best linear predictor of the state variable $\theta_t$ based on observations $X_1, \ldots, X_t$. The Kalman filter is a recursive algorithm for computing $\widehat{\theta}_t$ recursively from $\widehat{\theta}_{t-1}$ and the last observation $X_t$.

$$X_t = h^T \theta_t + n_t \tag{34}$$
$$\theta_t = G\theta_{t-1} + w_t \tag{35}$$

Given $X_1, \ldots, X_t$ we want to compute

$$\widehat{\theta}_t = C_1 X_1 + \cdots + C_t X_t$$

such that the mean square error

$$\mathsf{E}((\theta_t - \widehat{\theta}_t)^T(\theta_t - \widehat{\theta}_t)) = \min_{D_1,\ldots,D_t \in \mathbf{R}^k} \mathsf{E}\left(\left(\theta_t - \sum_{i=1}^{t} D_i X_i\right)^T \left(\theta_t - \sum_{i=1}^{t} D_i X_i\right)\right) \tag{36}$$

is minimized.

The following theorem is implied by Theorem 4.1.

**Theorem 7.1** *If $\widehat{\theta}_t = C_1X_1 + \cdots + C_tX_t$ is such that*

$$\mathsf{E}((\theta_t - \widehat{\theta}_t)X_i) = 0, \quad i = 1, \ldots, t$$

*then $\widehat{\theta}_t$ is the best linear predictor of $\theta_t$ based on $X_1, \ldots, X_t$.*

Let $\boldsymbol{C}_t = (C_1, \ldots, C_t)$ and $\boldsymbol{X}_t^T = (X_1, \ldots, X_t)$. The solution to this system can be obtained as an orthogonal projection, as in Chapter 4:

$$\boldsymbol{C}_t = \mathsf{E}\left[\theta_t \boldsymbol{X}_t^T\right] \left\{\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right)\right\}^{-1}.$$

In practice, however, this representation is not an efficient way to compute $\boldsymbol{C}_t$, since computing $\left\{\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right)\right\}^{-1}$ is expensive. The Kalman filter is an algorithm that allows $\widehat{\theta}_t$ to be computed recursively from $\widehat{\theta}_{t-1}$ and the most recent observation $X_t$.

### 7.2.1 Prediction stage of the Kalman filter

At the prediction stage of the Kalman filter, a forecast $\widehat{\theta}_{t|t-1}$ of $\theta_t$ is made from the observable data up to time $t-1$.

Define

$$P_t = P_{t|t} = \mathsf{E}\left[(\theta_t - \widehat{\theta}_t)(\theta_t - \widehat{\theta}_t)^T\right]$$

and

$$P_{t|t-1} = \mathsf{E}\left[(\theta_t - \widehat{\theta}_{t|t-1})(\theta_t - \widehat{\theta}_{t|t-1})^T\right],$$

known as the error covariance matrices of $\widehat{\theta}_t$ and $\widehat{\theta}_{t|t-1}$, respectively. Assume that at time $t-1$ we know $\widehat{\theta}_{t-1}$ and the covariance matrix $P_{t-1}$ of the corresponding error $\theta_{t-1} - \widehat{\theta}_{t-1}$.

**Lemma 7.1** *$\widehat{\theta}_{t|t-1}$, the best linear predictor of $\theta_t$ based on $X_1, \ldots, X_{t-1}$, is given in terms of $\widehat{\theta}_{t-1}$ as*

$$\widehat{\theta}_{t|t-1} = G\widehat{\theta}_{t-1}, \tag{37}$$

*and the covariance matrix of the corresponding error $\theta_t - \widehat{\theta}_{t|t-1}$ is*

$$P_{t|t-1} = GP_{t-1}G^T + W \tag{38}$$

*where $W$ is the covariance matrix defined by (25).*

*Proof.* This essentially follows because of linearity of projection.

$$\begin{aligned}\widehat{\theta}_{t|t-1} &= \mathsf{E}\left[\theta_t \boldsymbol{X}_{t-1}^T\right]\left\{\mathsf{E}\left(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^T\right)\right\}^{-1}\boldsymbol{X}_{t-1} \\ &= \mathsf{E}\left[(G\theta_{t-1}+w_t)\boldsymbol{X}_{t-1}^T\right]\left\{\mathsf{E}\left(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^T\right)\right\}^{-1}\boldsymbol{X}_{t-1} \\ &= G\widehat{\theta}_{t-1}.\end{aligned}$$

Note that $\boldsymbol{X}_{t-1}^T$ is assumed known.

Now to compute the covariance matrix $P_{t|t-1}$. Denote $\eta_{t-1} = G(\theta_{t-1} - \widehat{\theta}_{t-1})$. Since $\theta_t = G\theta_{t-1} + w_t$, we obtain that

$$\begin{aligned}P_{t|t-1} &= \mathsf{E}\left[(\theta_t - G\widehat{\theta}_{t-1})(\theta_t - G\widehat{\theta}_{t-1})^T\right] \\ &= \mathsf{E}\left[(\eta_{t-1}+w_t)(\eta_{t-1}+w_t)^T\right] \\ &= \mathsf{E}(\eta_{t-1}\eta_{t-1}^T) + \mathsf{E}(w_t w_t^T) + \mathsf{E}(\eta_{t-1}w_t^T) + \mathsf{E}(w_t\eta_{t-1}^T) \\ &= \mathsf{E}(\eta_{t-1}\eta_{t-1}^T) + \mathsf{E}(w_t w_t^T) = GP_{t-1}G^T + W.\end{aligned}$$

Note that we used above that $\mathsf{E}(\eta_{t-1}w_t^T) = 0$ and $\mathsf{E}(w_t\eta_{t-1}^T) = 0$.

Equations (37) and (38) are called the *prediction* equations of the Kalman filter.

Denote by $\widehat{X}_t$ the best linear predictor of $X_t$ based on $X_1, \ldots, X_{t-1}$.

**Lemma 7.2**

$$\widehat{X}_t = h^T\widehat{\theta}_{t|t-1}.$$

*Proof.*
Again, this is essentially linearity of the projection.

$$\begin{aligned}\widehat{X}_t &= \mathsf{E}\left[X_t\boldsymbol{X}_{t-1}^T\right]\left\{\mathsf{E}\left(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^T\right)\right\}^{-1}\boldsymbol{X}_{t-1} \\ &= \mathsf{E}\left[(h^T\theta_t + n_t)\boldsymbol{X}_{t-1}^T\right]\left\{\mathsf{E}\left(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^T\right)\right\}^{-1}\boldsymbol{X}_{t-1} \\ &= h^T\mathsf{E}\left[\theta_t\boldsymbol{X}_{t-1}^T\right]\left\{\mathsf{E}\left(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^T\right)\right\}^{-1}\boldsymbol{X}_{t-1} = h^T\widehat{\theta}_{t|t-1}.\end{aligned}$$

### 7.2.2 Updating stage

When the observation at time $t$, namely, $X_t$, becomes available, it can be taken into account to modify the estimator for $\theta_t$. Let

$$e_t = X_t - \widehat{X}_t = X_t - h^T\widehat{\theta}_{t|t-1},$$

be the error of the prediction based on $X_1, \ldots, X_{t-1}$.

**Lemma 7.3** *The optimal estimator $\widehat{\theta}_t$ and its covariance matrix $P_t$ can be found by means of the following updating equations*

$$\widehat{\theta}_t = \widehat{\theta}_{t|t-1} + K_t e_t = \widehat{\theta}_{t|t-1} + K_t(X_t - h^T\widehat{\theta}_{t|t-1}) \tag{39}$$

$$P_t = P_{t|t-1} - K_t h^T P_{t|t-1} \tag{40}$$

*where*

$$K_t = \left(h^T P_{t|t-1}h + \sigma_n^2\right)^{-1} P_{t|t-1}h \tag{41}$$

*the $(k \times 1)$ matrix (vector) $K_t$ is called the Kalman gain matrix.*

*Proof.*
Let $\boldsymbol{X}_t = (X_1 \dots X_{t-1}, X_t)^T = (\boldsymbol{X}_{t-1}^T, X_t)^T$.
Because

$$\widehat{\theta_t} = \mathsf{E}\left[\theta_t \boldsymbol{X}_t^T\right] \left\{\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right)\right\}^{-1} \boldsymbol{X}_t = (\mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T), \mathsf{E}(\theta_t X_t)) \left\{\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right)\right\}^{-1} \begin{pmatrix} \boldsymbol{X}_{t-1} \\ X_t \end{pmatrix}$$

and the matrix $\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right)$ can be written as follows

$$\mathsf{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T\right) = \begin{pmatrix} \mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right) & \mathsf{E}\left(\boldsymbol{X}_{t-1} X_t\right) \\ \mathsf{E}\left(\boldsymbol{X}_{t-1} X_t\right) & \mathsf{E}(X_t^2) \end{pmatrix} = \begin{pmatrix} D_{11} & D_{12} \\ D_{12}^2 & D_{22} \end{pmatrix}.$$

The inverse of this matrix can be shown (by direct computation) to be

$$\begin{pmatrix} D_{11} & D_{12} \\ D_{12}^2 & D_{22} \end{pmatrix}^{-1} = \begin{pmatrix} D_{11}^{-1} + D_{11}^{-1} D_{12} (D_{22} - D_{12}^T D_{11}^{-1} D_{12})^{-1} D_{12}^T D_{11}^{-1} & -D_{11}^{-1} D_{12} (D_{22} - D_{12}^T D_{11}^{-1} D_{12})^{-1} \\ -(D_{22} - D_{12}^T D_{11}^{-1} D_{12})^{-1} D_{12}^T D_{11}^{-1} & (D_{22} - D_{12}^T D_{11}^{-1} D_{12})^{-1}. \end{pmatrix}$$

We can now compute that

$$D_{12}^T D_{11}^{-1} \boldsymbol{X}_{t-1} = \mathsf{E}\left[\boldsymbol{X}_{t-1}^T X_t\right] \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \boldsymbol{X}_{t-1} = \widehat{X}_t = h^T \widehat{\theta}_{t|t-1},$$

$$\begin{aligned} D_{22} - D_{12}^T D_{11}^{-1} D_{12} &= \mathsf{E}(X_t^2) - \mathsf{E}(\boldsymbol{X}_{t-1}^T X_t) \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \mathsf{E}(\boldsymbol{X}_{t-1} X_t) \\ &= \mathsf{E}(X_t - \boldsymbol{X}_{t-1}^T \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \mathsf{E}(\boldsymbol{X}_{t-1} X_t))^2 \\ &= \mathsf{E}(h^T \theta_t + n_t - h^T \widehat{\theta}_{t|t-1})^2 = \sigma_n^2 + h^T P_{t|t-1} h, \end{aligned}$$

and

$$\mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T) D_{11}^{-1} \boldsymbol{X}_{t-1} = \widehat{\theta}_{t|t-1}, \quad D_{11}^{-1} D_{12} = \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \mathsf{E}(\boldsymbol{X}_{t-1} \theta_t^T) h.$$

Hence we have

$$\widehat{\theta_t} = \widehat{\theta}_{t|t-1} + (\sigma_n^2 + h^T P_{t|t-1} h)^{-1} \left[\mathsf{E}(\theta_t \theta_t^T) - \mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T) \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \mathsf{E}(\boldsymbol{X}_{t-1} \theta_t^T)\right] h (X_t - h^T \widehat{\theta}_{t|t-1}).$$

Now consider $P_{t|t-1}$:

$$\begin{aligned} P_{t|t-1} &= \mathsf{E}\left[(\widehat{\theta}_{t|t-1} - \theta_t)(\widehat{\theta}_{t|t-1} - \theta_t)^T\right] \\ &= \mathsf{E}\left[(\mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T) \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \boldsymbol{X}_{t-1} - \theta_t)(\mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T) \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \boldsymbol{X}_{t-1} - \theta_t)^T\right] \\ &= \mathsf{E}(\theta_t \theta_t^T) - \mathsf{E}(\theta_t \boldsymbol{X}_{t-1}^T) \left\{\mathsf{E}\left(\boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^T\right)\right\}^{-1} \mathsf{E}(\boldsymbol{X}_{t-1} \theta_t^T), \end{aligned}$$

so that

$$\widehat{\theta_t} = \widehat{\theta}_{t|t-1} + (\sigma_n^2 + h^T P_{t|t-1} h)^{-1} P_{t|t-1} h (X_t - h^T \widehat{\theta}_{t|t-1}).$$

We can now compute $P_t$

$$\begin{aligned}
P_t &= \mathsf{E}\left[(\theta_t - \widehat{\theta}_t)(\theta_t - \widehat{\theta}_t)^T\right] = \mathsf{E}\left[(\theta_t - \widehat{\theta}_{t|t-1} - K_t e_t)(\theta_t - \widehat{\theta}_{t|t-1} - K_t e_t)^T\right] \\
&= \mathsf{E}\left[(\theta_t - \widehat{\theta}_{t|t-1})(\theta_t - \widehat{\theta}_{t|t-1})^T\right] - K_t \mathsf{E}\left[e_t(\theta_t - \widehat{\theta}_{t|t-1})^T\right] - \mathsf{E}\left[(\theta_t - \widehat{\theta}_{t|t-1})e_t\right] K_t^T + K_t \mathsf{E}(e_t^2) K_t^T \\
&= P_{t|t-1} - K_t h^T P_{t|t-1} - P_{t|t-1} h K_t^T + K_t (h^T P_{t|t-1} h + \sigma_n^2) K_t^T = P_{t|t-1} - K_t h^T P_{t|t-1}.
\end{aligned}$$

As required.

## 7.3 The Kalman filter for the local level model

### 7.3.1 Prediction and updating stages

Recall the observation and transition equations for the local level model.

$$\begin{aligned}
X_t &= \theta_t + n_t \\
\theta_t &= \theta_{t-1} + w_t.
\end{aligned}$$

Here, $h = G = 1$, $\{n_t,\, t \in \mathbf{Z}\}$ and $\{w_t,\, t \in \mathbf{Z}\}$ are zero mean mutually independent white noise processes:

$$\mathsf{E}(n_t) = \mathsf{E}(w_t) = 0,\ \mathsf{Var}(n_t) = \sigma_n^2,\ \mathsf{Var}(w_t) = \sigma_w^2,\ \mathsf{Cov}(w_t, n_{t'}) = 0,\ t, t' \in \mathbf{Z}_+.$$

Let $\widehat{\theta}_t$ be the BLP of $\theta$ given $X_1, \ldots, X_t$ and $P_t = \mathsf{E}((\theta_t - \widehat{\theta}_t)^2)$.

*The prediction stage:* $\widehat{\theta}_{t|t-1}$ is the BLP of $\theta_t$ given $X_1, \ldots, X_{t-1}$

$$\widehat{\theta}_{t|t-1} = G\widehat{\theta}_{t-1} = \widehat{\theta}_{t-1}$$

the variance of the corresponding error is given by

$$P_{t|t-1} = \mathsf{E}(\theta_t - \widehat{\theta}_{t|t-1})^2 = P_{t-1} + \sigma_w^2.$$

*The updating stage:*

$$\begin{aligned}
e_t &= X_t - h^T \widehat{\theta}_{t|t-1} = X_t - \widehat{\theta}_{t-1} \\
\widehat{X}_t &= h^T \widehat{\theta}_{t|t-1} = \widehat{\theta}_{t-1}.
\end{aligned}$$

We compute the Kalman gain as follows

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + \sigma_n^2} = \frac{P_{t-1} + \sigma_w^2}{P_{t-1} + \sigma_w^2 + \sigma_n^2}.$$

Now for the updating stage:

$$\begin{aligned}
\widehat{\theta}_t &= \widehat{\theta}_{t|t-1} + K_t e_t = \widehat{\theta}_{t-1} + K_t e_t \\
P_t &= P_{t|t-1} - K_t P_{t|t-1} = P_{t-1} + \sigma_w^2 - K_t (P_{t-1} + \sigma_w^2)
\end{aligned}$$

Using the explicit formula for $K_t$ in this case we can write

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + \frac{P_{t-1} + \sigma_w^2}{P_{t-1} + \sigma_w^2 + \sigma_n^2} e_t.$$

This equation can be rewritten as follows

$$\widehat{\theta}_t = (1 - K_t)\widehat{\theta}_{t-1} + K_t X_t.$$

The error covariance matrix is

$$P_t = P_{t|t-1} - K_t P_{t|t-1} = (1 - K_t)(P_{t-1} + \sigma_w^2) = \frac{(P_{t-1} + \sigma_w^2)\sigma_n^2}{P_{t-1} + \sigma_w^2 + \sigma_n^2}.$$

### 7.3.2 Long-time behaviour and steady state

It can be shown that the sequence $P_t$ converges to a certain limit as $t \to \infty$. We say that the Kalman filter converges to a steady state.
Assuming that the steady state limit exists it can be computed as follows. We have from the updating equation, that

$$\begin{aligned} P_t &= P_{t-1} + \sigma_w^2 - K_t(P_{t-1} + \sigma_w^2) \\ &= P_{t-1} + \sigma_w^2 - \frac{(P_{t-1} + \sigma_w^2)^2}{P_{t-1} + \sigma_w^2 + \sigma_n^2} \end{aligned}$$

Passing to the limit in this equation we get that the limit $P$ must be the solution of the following equation

$$P = P + \sigma_w^2 - \frac{(P + \sigma_w^2)^2}{P + \sigma_w^2 + \sigma_n^2}$$

which can be rewritten in the following quadratic form

$$P^2 + \sigma_w^2 P - \sigma_w^2 \sigma_n^2 = 0$$

This equation has two roots, $P$ is the non-negative one (as a limit of non-negative sequence)

$$P = \frac{-\sigma_w^2 + \sqrt{\sigma_w^4 + 4\sigma_w^2 \sigma_n^2}}{2}$$

Denoting $c = \sigma_w^2/\sigma_n^2$ the formula for $P$ can be rewritten as follows

$$P = \frac{\sigma_n^2}{2}(-c + \sqrt{c^2 + 4c}). \tag{42}$$

A direct computation shows that convergence of $P_t$ to the limit (42) yields that

$$K_t \to K = \frac{1}{2}\left(\sqrt{c^2 + 4c} - c\right), \quad \text{as} \quad t \to \infty.$$

### 7.3.3 State-space models of ARIMA processes

ARIMA processes have state space representations, and in general these representations are not unique. In what follows, we will give a state space representation of an AR(p) process.

Suppose we have an AR(p) process

$$Y_t = \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p} + Z_t$$

Let the state vector be

$$\theta_t^T = (Y_t, \ldots, Y_{t-p+1}),$$

and let the $p \times p$ matrix

$$G = \begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

the $p \times 1$ matrix $h^T = (1, 0, \ldots, 0)$, the white noise processes $n_t = 0$ and $w_t^T = (Z_t, 0, \ldots, 0) \in \mathbf{R}^p$. The observation variable is $X_t = Y_t \in \mathbf{R}$, so we have is a univariate model.
The observation equation is

$$X_t = h^T \theta_t,$$

and the state equation is

$$\theta_t = G\theta_{t-1} + w_t.$$

State-space representations for ARIMA processes allow us to use the general results relating to state-space models (though these are not *always* helpful).