Chapter 3

ARMA Processes

3.1 Introduction

In this chapter we discuss *autoregressive moving average* processes, which play a crucial role in specifying time series models for applications. They are defined as the solutions of *stochastic difference equations* with constant coefficients and therefore possess a linear structure.

Definition 3.1.1 (ARMA processes) *(a) A weakly stationary process* $(X_t\colon t \in \mathbb{Z})$ *is called an autoregressive moving average time series of order* (p, q)*, abbreviated by ARMA*(p, q)*, if it satisfies the difference equations*

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \qquad t \in \mathbb{Z}, \tag{3.1.1}$$

where $\phi_1, \ldots, \phi_p$ *and* $\theta_1, \ldots, \theta_q$ *are real constants,* $\phi_p \neq 0 \neq \theta_q$*, and* $(Z_t\colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$*.*

(b) A weakly stationary stochastic process $(X_t\colon t \in \mathbb{Z})$ *is called an ARMA*(p, q) *time series with mean* μ *if the process* $(X_t - \mu\colon t \in \mathbb{Z})$ *satisfies the equation system (3.1.1).*

A more concise representation of (3.1.1) can be obtained with the use of the backshift operator B. To this end, we define the *autoregressive polynomial* and the *moving average polynomial* by

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p, \qquad z \in \mathbb{C},$$

and

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q, \qquad z \in \mathbb{C},$$

respectively, where $\mathbb{C}$ denotes the set of complex numbers. Inserting the backshift operator into these polynomials, the equations in (3.1.1) become

$$\phi(B)X_t = \theta(B)Z_t, \qquad t \in \mathbb{Z}. \tag{3.1.2}$$

Example 3.1.1 Figure 3.1 displays realizations of three different autoregressive moving average time series based on independent, standard normally distributed $(Z_t\colon t \in \mathbb{Z})$. The left panel is an ARMA(2,2) process with parameter specifications $\phi_1 = .2$, $\phi_2 = -.3$, $\theta_1 = -.5$ and $\theta_2 = .3$. The middle plot is obtained from an ARMA(1,4) process with

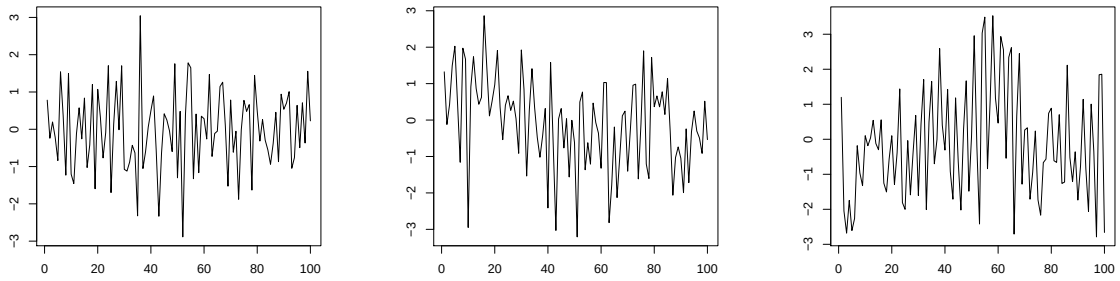

Figure 3.1: Realizations of three autoregressive moving average processes.

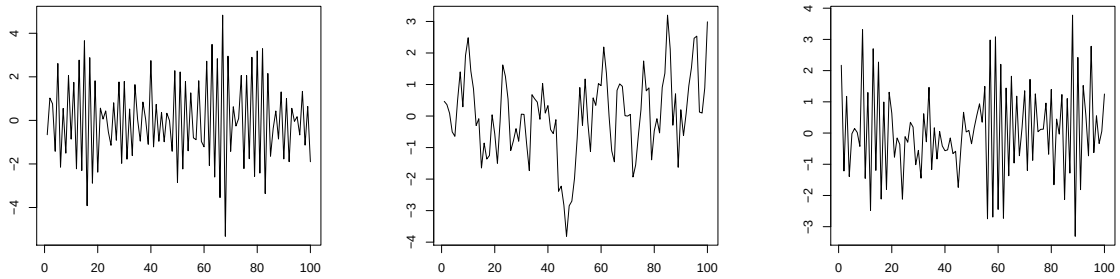

Figure 3.2: Realizations of three autoregressive processes.

parameters $\phi_1 = .3$, $\theta_1 = -.2$, $\theta_2 = -.3$, $\theta_3 = .5$, and $\theta_4 = .2$, while the right plot is from an ARMA(4,1) with parameters $\phi_1 = -.2$, $\phi_2 = -.3$, $\phi_3 = .5$ and $\phi_4 = .2$ and $\theta_1 = .6$. The plots indicate that ARMA models can provide a flexible tool for modeling diverse residual sequences. We shall find out in the next section that all three realizations here come from (strictly) stationary processes.

Some special cases which we cover in the following two examples have particular relevance in time series analysis.

Example 3.1.2 (AR processes) If the moving average polynomial in (3.1.2) is equal to one, that is, if $\theta(z) \equiv 1$, then the resulting $(X_t \colon t \in \mathbb{Z})$ is referred to as *autoregressive process of order p, AR(p)*. These time series interpret the value of the current variable X_t as a linear combination of p previous variables $X_{t-1}, \ldots, X_{t-p}$ plus an additional distortion by the white noise Z_t. Figure 3.2 displays two AR(1) processes with respective parameters $\phi_1 = -.9$ (left) and $\phi_1 = .8$ (middle) as well as an AR(2) process with parameters $\phi_1 = -.5$ and $\phi_2 = .3$.

Example 3.1.3 (MA processes) If the autoregressive polynomial in (3.1.2) is equal to one, that is, if $\phi(z) \equiv 1$, then the resulting $(X_t \colon t \in \mathbb{Z})$ is referred to as *moving average process of order q, MA(q)*. Here the present variable X_t is obtained as superposition of q white noise terms $Z_t, \ldots, Z_{t-q}$. Figure 3.3 shows two MA(1) processes with respective parameters $\theta_1 = .5$ (left) and $\theta_1 = -.8$ (middle). The right plot is observed from an MA(2) process with parameters $\theta_1 = -.5$ and $\theta_2 = .3$.

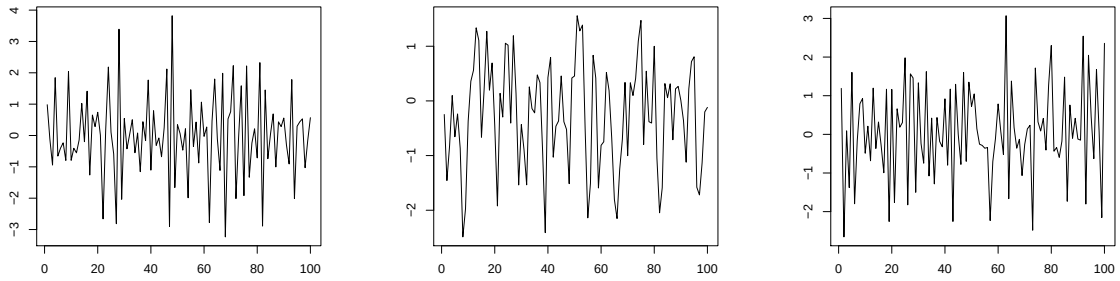

Figure 3.3: Realizations of three moving average processes.

For the analysis upcoming in the next chapters, we now introduce moving average processes of infinite order ($q = \infty$). They are an important tool for determining stationary solutions to the difference equations (3.1.1).

Definition 3.1.2 (Linear processes) *A stochastic process $(X_t\colon t \in \mathbb{Z})$ is called linear process or MA(∞) time series if there is a sequence $(\psi_j\colon j \in \mathbb{N}_0)$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \qquad t \in \mathbb{Z}, \tag{3.1.3}$$

where $(Z_t\colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$.

Moving average time series of any order q are special cases of linear processes. Just pick $\psi_j = \theta_j$ for $j = 1, \ldots, q$ and set $\psi_j = 0$ if $j > q$. It is common to introduce the power series

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j, \qquad z \in \mathbb{C},$$

to express a linear process in terms of the backshift operator. We can now rewrite display (3.1.3) in the form

$$X_t = \psi(B) Z_t, \qquad t \in \mathbb{Z}.$$

With the definitions of this section at hand, we shall investigate properties of ARMA processes such as stationarity and invertibility in the next section. We close the current section giving meaning to the notation $X_t = \psi(B)Z_t$. Note that we are possibly dealing with an *infinite* sum of random variables and that therefore issues of convergence have to be addressed. This is done in the following proposition.

Proposition 3.1.1 *If $(Y_t\colon t \in \mathbb{Z})$ is any stochastic process such that $\sup_t E[|Y_t|] < \infty$ and if $\sum_j |\psi_j| < \infty$, then the series*

$$\psi(B) Y_t = \sum_j \psi_j Y_{t-j}$$

converges absolutely with probability one. If in addition $\sup_t E[Y_t^2] < \infty$, then the series also converges in mean square to the same limit.

Proof. See Proposition 3.1.1 in Brockwell and Davis (1991). □

3.2 Causality and Invertibility

While a moving average process of order q will always be stationary without conditions on the coefficients $\theta_1, \ldots, \theta_q$, some deeper thoughts are required in the case of AR(p) and ARMA(p, q) processes. For simplicity, we start by investigating the autoregressive process of order one, which is given by the equations $X_t = \phi X_{t-1} + Z_t$ (writing $\phi = \phi_1$). Repeated iterations yield that

$$X_t = \phi X_{t-1} + Z_t = \phi^2 X_{t-2} + Z_t + \phi Z_{t-1} = \ldots = \phi^N X_{t-N} + \sum_{j=0}^{N-1} \phi^j Z_{t-j}.$$

This leads to the candidate

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

for the weakly stationary solution, provided that $|\phi| < 1$. To arrive at this conclusion, observe that $E[X_t^2]$ is a finite constant if we are dealing with a weakly stationary solution. Therefore,

$$E\left[\left(X_t - \sum_{j=0}^{N-1} \phi^j Z_{t-j}\right)^2\right] = \phi^{2N} E[X_{t-N}^2] \to 0 \qquad (N \to \infty),$$

which proves mean-square convergence and, by Proposition 3.1.1, convergence with probability one. The calculations just performed indicate moreover, that an autoregressive process of order one can be represented as linear process with coefficients $\psi_j = \phi^j$.

Example 3.2.1 (Mean and ACVF of an AR(1) process) Since we have identified an autoregressive process of order one as an example of a linear process, we can easily determine its expected value as

$$E[X_t] = \sum_{j=0}^{\infty} \phi^j E[Z_{t-j}] = 0 \qquad t \in \mathbb{Z}.$$

For the ACVF, we obtain that

$$\begin{aligned}
\gamma(h) &= \mathrm{Cov}(X_{t+h}, X_t) \\
&= E\left[\sum_{j=0}^{\infty} \phi^j Z_{t+h-j} \sum_{k=0}^{\infty} \phi^k Z_{t-k}\right] \\
&= \sigma^2 \sum_{k=0}^{\infty} \phi^{k+h} \phi^k = \sigma^2 \phi^h \sum_{k=0}^{\infty} \phi^{2k} = \frac{\sigma^2 \phi^h}{1-\phi^2},
\end{aligned}$$

where $h \geq 0$. This determines the ACVF for all h using that $\gamma(-h) = \gamma(h)$. It is also immediate that the ACF satisfies $\rho(h) = \phi^h$.

Example 3.2.2 (Nonstationary AR(1) processes) In Example 1.2.3 we have introduced the random walk as a nonstationary time series. It is also an AR(1) process with parameter $\phi = 1$. In general, autoregressive processes of order one with coefficients $|\phi| > 1$ are called *explosive* for they do not admit a weakly stationary solution that could be expressed as a linear process. However, one may proceed as follows. Rewrite the defining equations of an AR(1) process as

$$X_t = -\phi^{-1} Z_{t+1} + \phi^{-1} X_{t+1}, \qquad t \in \mathbb{Z}.$$

Apply now the same iterations as before to arrive at

$$X_t = \phi^{-N} X_{t+N} - \sum_{j=1}^{N} \phi^{-j} Z_{t+j}, \qquad t \in \mathbb{Z}.$$

Note that in the weakly stationary case, the present observation has been described in terms of past innovations. The representation in the last equation however contains only future observations with time lags larger than the present time t. From a statistical point of view this does not make much sense, even though by identical arguments as above we may obtain

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}, \qquad t \in \mathbb{Z},$$

as the weakly stationary solution in the explosive case.

The result of the previous example leads to the notion of causality which means that the process $(X_t \colon t \in \mathbb{Z})$ has a representation in terms of the white noise $(Z_s \colon s \leq t)$ and that is hence independent of the future as given by $(Z_s \colon s > t)$. We give the definition for the general ARMA case.

Definition 3.2.1 (Causality) *An ARMA(p, q) process given by (3.1.1) is causal if there is a sequence $(\psi_j \colon j \in \mathbb{N}_0)$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \qquad t \in \mathbb{Z}.$$

Theorem 3.2.1 *Let $(X_t \colon t \in \mathbb{Z})$ be an ARMA(p, q) process such that the polynomials $\phi(z)$ and $\theta(z)$ have no common zeroes. Then $(X_t \colon t \in \mathbb{Z})$ is causal if and only if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. The coefficients $(\psi_j \colon j \in \mathbb{N}_0)$ are determined by the power series expansion*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \qquad |z| \leq 1.$$

Proof. Let $(X_t \colon t \in \mathbb{Z})$ be causal. Then, there is a sequence $(\psi_j \colon j \in \mathbb{N}_0)$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \psi(B) Z_t, \qquad t \in \mathbb{Z}. \tag{3.2.1}$$

Replacing X_t in equation (3.1.2) by the expression on the right-hand side of display (3.2.1), we obtain

$$\theta(B)Z_t = \phi(B)X_t = \phi(B)\psi(B)Z_t \qquad t \in \mathbb{Z}.$$

Denoting $\eta(z) = \phi(z)\psi(z) = \sum_{j=0}^{\infty} \eta_j z^j$, the previous equation can be rewritten as

$$\sum_{j=0}^{q} \theta_j Z_{t-j} = \sum_{j=0}^{\infty} \eta_j Z_{t-j}, \qquad t \in \mathbb{Z}$$

and we conclude that $\theta_j = \eta_j$ for $k = 0, 1, \ldots, q$ and $\eta_j = 0$ if $j > q$. It follows consequently that $\theta(z) = \eta(z) = \phi(z)\psi(z)$. Recall that, by assumption, $\phi(z)$ and $\theta(z)$ have no common zeroes. Hence,

$$\psi(z) = \frac{\theta(z)}{\phi(z)}, \qquad z \in \mathbb{C},\ |z| \leq 1$$

and, because of $|\psi(z)| < \infty$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$, $\phi(z) \neq 0$ for these z. This proves the first part of the theorem.

As for the second part, assume that $\phi(z) \neq 0$ for all $|z| \leq 1$. Since a power series is continuous, there is an $\epsilon > 0$ such that

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \xi_j z^j = \xi(z), \qquad |z| < 1 + \epsilon.$$

This series is convergent and thus $\xi_j(1+\epsilon/2)^j \to 0$ as $j \to \infty$. It follows that we can pick a constant K such that $|\xi_j| \leq K(1+\epsilon/2)^{-j}$ for all $j \in \mathbb{N}_0$. Now, estimate

$$\sum_{j=0}^{\infty} |\xi_j| \leq K \sum_{j=0}^{\infty} \left(1 + \frac{\epsilon}{2}\right)^{-j} < \infty.$$

Note that by construction $\xi(z)\phi(z) \equiv 1$ for all $|z| \leq 1$. With the defining ARMA equations in the form of (3.1.2) we can therefore write

$$X_t = \xi(B)\phi(B)X_t = \xi(B)\theta(B)Z_t, \qquad t \in \mathbb{Z}.$$

Setting $\psi(z) = \xi(z)\theta(z)$, the proof is complete. □

A concept closely related to causality is *invertibility*. We motivate this notion with the following example that studies properties of a moving average time series of order 1.

Example 3.2.3 Let $(X_t : t \in \mathbb{Z})$ be an MA(1) process with parameter $\theta = \theta_1$. It is an easy exercise to compute the ACVF and the ACF (for $h \geq 0$ only) as

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2, & h = 0, \\ \theta\sigma^2, & h = 1 \\ 0 & h > 1, \end{cases} \qquad \rho(h) = \begin{cases} 1 & h = 0. \\ \theta(1+\theta^2)^{-1}, & h = 1. \\ 0 & h > 1. \end{cases}$$

These results lead to the conclusion that $\rho(h)$ does not change if the parameter θ is replaced with θ^{-1}. Moreover, there exist pairs (θ, σ^2) that lead to the same ACVF, for

example $(5, 1)$ and $(1/5, 25)$. Consequently, we arrive at the fact that the two MA(1) models

$$X_t = Z_t + \frac{1}{5} Z_{t-1}, \qquad t \in \mathbb{Z}, \qquad (Z_t \colon t \in \mathbb{Z}) \sim \text{iid } \mathcal{N}(0, 25),$$

and

$$X_t = \tilde{Z}_t + 5\tilde{Z}_{t-1}, \qquad t \in \mathbb{Z}, \qquad (\tilde{Z}_t \colon t \in \mathbb{Z}) \sim \text{iid } \mathcal{N}(0, 1),$$

are indistinguishable because we only observe X_t but not the noise variables Z_t and $\tilde{Z}_t$.

For convenience, the statistician will pick the model which satisfies the invertibility criterion which is to be defined next. It specifies that the noise sequence can be represented as a linear process in the observations.

Definition 3.2.2 (Invertibility) *An ARMA(p, q) process given by (3.1.1) is invertible if there is a sequence* $(\pi_j \colon j \in \mathbb{N}_0)$ *such that* $\sum_{j=0}^{\infty} |\pi_j| < \infty$ *and*

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \qquad t \in \mathbb{Z}.$$

Theorem 3.2.2 *Let* $(X_t \colon t \in \mathbb{Z})$ *be an ARMA(p, q) process such that the polynomials* $\phi(z)$ *and* $\theta(z)$ *have no common zeroes. Then* $(X_t \colon t \in \mathbb{Z})$ *is invertible if and only if* $\theta(z) \neq 0$ *for all* $z \in \mathbb{C}$ *with* $|z| \leq 1$*. The coefficients* $(\pi_j \colon j \in \mathbb{N}_0)$ *are determined by the power series expansion*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \qquad |z| \leq 1.$$

Proof. Similar to the proof of Theorem 3.2.1. □.

From now on we assume that all ARMA sequences specified in the sequel are causal and invertible ones if not stated otherwise explicitly. Two examples highlight the usefulness of the theory we have established in the section. The first one deals with parameter redundancy and the calculation of the causality and invertibility sequences $(\psi_j \colon j \in \mathbb{N}_0)$ and $(\pi_j \colon j \in \mathbb{N}_0)$, the second one with the causal solution of the AR(2) equations.

Example 3.2.4 (Parameter redundancy) Consider the ARMA equations

$$X_t = .4X_{t-1} + .21X_{t-2} + Z_t + .6Z_{t-1} + .09Z_{t-2},$$

which seem to generate an ARMA(2,2) sequence. However, the autoregressive and moving average polynomials have a common zero:

$$\tilde{\phi}(z) = 1 - .4z - .21z^2 = (1 - .7z)(1 + .3z),$$

$$\tilde{\theta}(z) = 1 + .6z + .09z^2 = (1 + .3z)^2.$$

Therefore, we can reset the ARMA equations to a sequence of order (1,1) and obtain

$$X_t = .7X_{t-1} + Z_t + .3Z_{t-1}.$$

Now, the corresponding polynomials have no common roots. Note that the roots of $\phi(z) = 1 - .7z$ and $\theta(z) = 1 + .3z$ are $10/7 > 1$ and $-10/3 < -1$, respectively. Thus Theorems 3.2.1 and 3.2.2 imply that causal and invertible solutions exist. In the following, we are going to calculate the corresponding coefficients in the expansions

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \qquad \text{and} \qquad Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \qquad t \in \mathbb{Z}.$$

We start with the causality sequence $(\psi_j \colon j \in \mathbb{N}_0)$. Writing, for $|z| \leq 1$,

$$\sum_{j=0}^{\infty} \psi_j z^j = \psi(z) = \frac{\theta(z)}{\phi(z)} = \frac{1 + .3z}{1 - .7z} = (1 + .3z) \sum_{j=0}^{\infty} (.7z)^j,$$

it can be obtained from a comparison of coefficients that

$$\psi_0 = 1 \qquad \text{and} \qquad \psi_j = (.7 + .3)(.7)^{j-1} = (.7)^{j-1}, \qquad j \in \mathbb{N}.$$

Similarly one computes the invertibility coefficients $(\pi_j \colon j \in \mathbb{N}_0)$ from from the equation

$$\sum_{j=0}^{\infty} \pi_j z^j = \pi(z) = \frac{\phi(z)}{\theta(z)} = \frac{1 - .7z}{1 + .3z} = (1 - .7z) \sum_{j=0}^{\infty} (-.3z)^j$$

$(|z| \leq 1)$ as

$$\pi_0 = 1 \qquad \text{and} \qquad \pi_j = (-1)^j (.3 + .7)(.3)^{j-1} = (-1)^j (.3)^{j-1}.$$

Together, the previous calculations yield to the explicit representations

$$X_t = Z_t + \sum_{j=1}^{\infty} (.7)^{j-1} Z_{t-j} \qquad \text{and} \qquad Z_t = X_t + \sum_{j=1}^{\infty} (-1)^j (.3)^{j-1} X_{t-j}.$$

Example 3.2.5 (Causality for AR(2) processes) Consider the difference equations for an AR(2) process, $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$, where $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$. To obtain a causal solution, the roots of the autoregressive polynomial

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2$$

must lie outside the complex unit circle. Setting $\phi(z) = 0$ and solving the resulting quadratic equation, one arrives at the equivalent formulation

$$\left| \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \right| > 1.$$

Homework Problem 5 contains the statement that the causality region of possible parameter choices ϕ_1 and ϕ_2 is given by the triangle determined by the three inequalities $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$ and $|\phi_2| < 1$.

In the remainder of this section, we provide a way to determine the weights $(\psi_j \colon j \geq 1)$ for a causal ARMA(p, q) process given by $\phi(B)X_t = \theta(B)Z_t$, where $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. Since $\psi(z) = \theta(z)/\phi(z)$ for these z, the weight ψ_j can be computed by matching the corresponding coefficients in the equation $\psi(z)\phi(z) = \theta(z)$, that is,

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \ldots)(1 - \phi_1 z - \ldots - \phi_p z^p) = 1 + \theta_1 z + \ldots + \theta_q z^q.$$

Recursively solving for $\psi_0, \psi_1, \psi_2, \ldots$ gives

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 - \phi_1 \psi_0 &= \theta_1, \\ \psi_2 - \phi_1 \psi_1 - \phi_2 \psi_0 &= \theta_2, \end{aligned}$$

and so on as long as $j < \max\{p, q+1\}$. The general solution can be stated as

$$\psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, \qquad 0 \leq j < \max\{p, q+1\}, \tag{3.2.2}$$

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = 0, \qquad j \geq \max\{p, q+1\}, \tag{3.2.3}$$

if we define $\phi_j = 0$ if $j > p$ and $\theta_j = 0$ if $j > q$. To obtain the coefficients ψ_j one therefore has to solve the homogeneous linear difference equation (3.2.3) subject to the initial conditions specified by (3.2.2). For more on this subject, see Section 3.6 of Brockwell and Davis (1991) and Section 3.3 of Shumway and Stoffer (2006).

3.3 The ACVF and PACF of a causal ARMA Process

In this section, we study first three methods to compute the ACVF of a causal ARMA process of arbitrary order (p, q). In the second part, we introduce the *partial autocorrelation function (PACF)* to further assess the dependence structure of stationary processes in general and causal ARMA processes in particular. To start with, let us compute the ACVF of an moving average process of order q.

Example 3.3.1 (The ACVF of an MA(q) process) Let $(X_t \colon t \in \mathbb{Z})$ be an MA(q) process specified by the polynomial $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$. Then, letting $\theta_0 = 0$, it holds that

$$E[X_t] = \sum_{j=0}^{q} \theta_j E[Z_{t-j}] = 0.$$

To compute the ACVF, suppose that $h \geq 0$ and write

$$\begin{aligned} \gamma(h) = \mathrm{Cov}(X_{t+h}, X_t) &= E[X_{t+h} X_t] \\ &= E\left[\left(\sum_{j=0}^{q} \theta_j Z_{t+h-j} \right) \left(\sum_{k=0}^{q} \theta_k Z_{t-k} \right) \right] \end{aligned}$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k E[Z_{t+h-j} Z_{t-k}]$$

$$= \begin{cases} \sigma^2 \sum_{k=0}^{q-h} \theta_{k+h} \theta_k, & 0 \le h \le q. \\ 0, & h > q. \end{cases}$$

The result can easily be extended to linear processes and hence to all causal ARMA sequences.

The second example hints at another possible strategy to compute the ACVF for a causal ARMA process.

Example 3.3.2 (The ACVF of an ARMA(1,1) process) Let $(X_t \colon t \in \mathbb{Z})$ be given by the causal ARMA(1,1) equations

$$X_t = \phi X_{t-1} + \theta Z_{t-1} + Z_t, \qquad |\phi| < 1.$$

For $h \ge 2$, it holds that

$$\gamma(h) = E[X_{t+h} X_t] = E[(\phi X_{t+h-1} + \theta Z_{t+h-1} + Z_{t+h}) X_t] = \phi E[X_{t+h-1} X_t] = \phi \gamma(h-1).$$

This suggests that the ACVF can be computed by solving for $\gamma(h)$ recursively after determining a few initial values first.

Next, we are going to introduce the three methods to evaluate the ACVF for causal ARMA processes. The first relies on the equations for the weights $(\psi_j \colon j \in \mathbb{N}_0)$ derived in displays (3.2.2) and (3.2.3), the second establishes a similar set of difference equations directly for the ACVF, while the third approach is based on the recursions introduced in Example 3.3.2.

Method 1 (Using the ψ-weights) If $(X_t \colon t \in \mathbb{Z})$ is a causal ARMA process, it can be represented as $X_t = \psi(B) Z_t$, that is, as a linear process. The covariance function has the form (this can be established as in Example 3.3.1 for the MA(q) process)

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}, \qquad h \in \mathbb{N}_0.$$

Since the coefficients $(\psi_j \colon j \in \mathbb{N}_0)$ can be computed as in (3.2.2) and (3.2.3), $\gamma(h)$ follows readily.

Example 3.3.3 Let $(X_t \colon t \in \mathbb{Z})$ be given by

$$X_t = X_{t-1} - \tfrac{1}{4} X_{t-2} + Z_t + Z_{t-1}, \qquad t \in \mathbb{Z}. \tag{3.3.1}$$

Since $p = 2$ and $q = 1$, (3.2.2) gives the initial conditions $\psi_0 = 1$ and $\psi_1 = 2$, while (3.2.3) yields the homogeneous difference equation $\psi_j - \psi_{j-1} + \frac{1}{4} \psi_{j-2} = 0$ $(j \ge 2)$. This system has the general solution

$$\psi_j = (\alpha_0 + \alpha_1 j) 2^{-j}, \qquad j \ge 0.$$

Since $\psi_0 = 1$ and $\psi_1 = 2$, this leads to the values $\alpha_0 = 1$ and $\alpha_1 = 3$ and consequently to $\psi_j = (1 + 3j)2^{-j}$. Thus, for $h \geq 0$,

$$\begin{aligned}
\gamma(h) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \\
&= \sigma^2 \sum_{j=0}^{\infty} (1 + 3j)(1 + 3j + 3h) 2^{-2j-h} \\
&= \frac{\sigma^2}{2^h} \sum_{j=0}^{\infty} \left[(3h + 1)4^{-j} + 3(3h + 2) j 4^{-j} + 9j^2 4^{-j} \right] \\
&= \frac{\sigma^2}{2^h} \left[\frac{4}{3}(3h + 1) + \frac{12}{9}(3h + 2) + \frac{180}{27} \right] \\
&= \frac{\sigma^2}{2^h} \left[\frac{32}{3} + 8h \right]
\end{aligned}$$

where various geometric sum arguments have been used to solve the series.

Method 2 (A difference equation for γ) Take the ARMA(p, q) equations $\phi(B)X_t = \theta(B)Z_t$ and multiply them first with X_{t-h} and take expectations thereafter. This leads to the equations

$$E\left[\{\phi(B)X_t\}X_{t-h}\right] = E\left[\{\theta(B)Z_t\}X_{t-h}\right].$$

The left-hand side of this equality can be further evaluated as

$$E[\{X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}\}X_{t-h}] = \gamma(h) - \phi_1\gamma(h-1) - \ldots - \phi_p\gamma(h-p), \quad (3.3.2)$$

whereas the right-hand side, using causality, leads to

$$E\left[\sum_{j=0}^{q} \theta_j Z_{t-j} \sum_{k=0}^{\infty} \psi_k Z_{t-h-k}\right] = \begin{cases} \sigma^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}, & h < \max\{p, q+1\}. \\ 0, & h \geq \max\{p, q+1\}. \end{cases} \quad (3.3.3)$$

using $\theta_0 = 1$. The general solution to these equations can be found in Section 3.6 of Brockwell and Davis (1991). Here, we continue with the computation of the ACVF of the ARMA process from Example 3.3.3 with the help of Method 2.

Example 3.3.4 Let $(X_t \colon t \in \mathbb{Z})$ be given by the equations (3.3.1). Then, (3.3.2) and (3.3.3) become

$$\gamma(h) - \gamma(h-1) + \tfrac{1}{4}\gamma(h-2) = 0, \qquad h \geq 2,$$

subject to the initial conditions

$$\begin{aligned}
\gamma(0) - \gamma(1) + \tfrac{1}{4}\gamma(2) &= 3\sigma^2, \\
\gamma(1) - \gamma(0) + \tfrac{1}{4}\gamma(1) &= \sigma^2,
\end{aligned}$$

which has the general solution $\gamma(h) = (\beta_0 + \beta_1 h)2^{-h}$, where $h \geq 0$. Now $\psi_0 = 1$ and $\psi_1 = \theta_1 + \phi_1 = 2$, so that one can solve the initial conditions to obtain $\beta_0 = 32\sigma^2/3$ and $\beta_1 = 8\sigma^2$. Finally,

$$\gamma(h) = \frac{\sigma^2}{2^h}\left[\frac{32}{3} + 8h\right],$$

which is, of course, the same result as in Example 3.3.3.

Method 3 (Using the recursion) The third approach suggests to first solve $p + 1$ initial equations given by (3.3.2) and (the first one of) (3.3.3) to obtain the values for $\gamma(0), \gamma(1), \ldots, \gamma(p)$. Then use the iterations to build up $\gamma(h)$ for larger h.

Example 3.3.5 For the ARMA process specified by (3.3.1) the first $p+1 = 3$ equations are

$$\begin{aligned}
\gamma(0) - \gamma(1) + \tfrac{1}{4}\gamma(2) &= 3\sigma^2, \\
\gamma(1) - \gamma(0) + \tfrac{1}{4}\gamma(1) &= \sigma^2, \\
\gamma(2) - \gamma(1) + \tfrac{1}{4}\gamma(0) &= 0
\end{aligned}$$

which have the solutions

$$\gamma(0) = \frac{32\sigma^2}{3}, \qquad \gamma(1) = \frac{28\sigma^2}{3}, \qquad \text{and} \qquad \gamma(2) = \frac{20\sigma^2}{3}.$$

Then, one utilizes the formula $\gamma(h) = \gamma(h-1) - \frac{1}{4}\gamma(h-2)$ to calculate the $\gamma(h)$ for $h \geq 3$.

In the second part of the section, we first introduce the partial autocorrelation function of a weakly stationary process and use it thereafter to assess in further detail dependence structures of ARMA sequences. As motivation, we discuss the following example.

Example 3.3.6 Let $(X_t \colon t \in \mathbb{Z})$ be a causal AR(1) process with parameter $|\phi| < 1$. It holds that

$$\gamma(2) = \mathrm{Cov}(X_2, X_0) = \mathrm{Cov}(\phi^2 X_0 + \phi Z_1 + Z_2, X_0) = \phi^2\gamma(0) \neq 0.$$

To break the linear dependence between X_0 and X_2, subtract ϕX_1 from both variables. Calcualting the resulting covariance yields

$$\mathrm{Cov}(X_2 - \phi X_1, X_0 - \phi X_1) = \mathrm{Cov}(Z_2, X_0 - \phi X_1) = 0$$

due to the causality of the AR(1) process. Note that ϕX_1 coincides with the projection of X_2 onto the closed linear subspace $\mathcal{M} = \overline{\mathrm{sp}}(X_1)$.

The previous example motivates the following general definition.

Definition 3.3.1 (Partial autocorrelation function) *Let $(X_t \colon t \in \mathbb{Z})$ be a weakly stationary stochastic process with zero mean. Then, we call the sequence $(\phi_{hh} \colon h \in \mathbb{N})$ given by*

$$\begin{aligned}
\phi_{11} &= \rho(1) = \mathrm{Corr}(X_1, X_0), \\
\phi_{hh} &= \mathrm{Corr}(X_h - X_h^{h-1}, X_0 - X_0^{h-1}), \qquad h \geq 2,
\end{aligned}$$

the partial autocorrelation function (PACF) of $(X_t\colon t \in \mathbb{Z})$. *Therein,*

$$\begin{aligned} X_h^{h-1} &= P_{\mathcal{M}} X_h \\ &= \beta_1 X_{h-1} + \beta_2 X_{h-2} + \ldots + \beta_{h-1} X_1 \\ X_0^{h-1} &= P_{\mathcal{M}} X_0 \\ &= \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{h-1} X_{h-1}, \end{aligned}$$

where $P_{\mathcal{M}}$ *is the projection onto the closed linear subspace* $\mathcal{M} = \overline{\mathrm{sp}}(X_1, \ldots, X_{h-1})$.

The projection theorem guarantees the existence of the regression representation with coefficients $\beta_1, \ldots, \beta_{h-1}$. The symmetry of the ACVF yields the relation between X_h^{h-1} and X_0^{h-1}. Notice also that there is no intercept coefficient β_0 in the regression parameters, since it is assumed that $E[X_t] = 0$. We demonstrate how to calculate the regression parameters in the case of an AR(1) process.

Example 3.3.7 (PACF of an AR(1) process) If $(X_t\colon t \in \mathbb{Z})$ is a causal AR(1) process, then we have that $\phi_{11} = \rho(1) = \phi$. To calculate ϕ_{22}, we first calculate $X_2^1 = \beta X_1$, that is β. This coefficient is determined by minimizing the mean-squared error between X_2 and βX_1:

$$\|X_2 - \beta X_1\|^2 = E[X_2 - \beta X_1]^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0)$$

which is minimized by $\beta = \rho(1) = \phi$. (This follows easily by taking the derivative and setting it to zero.) Therefore $X_2^1 = \phi X_1$. Similarly, one computes $X_0^1 = \phi X_1$ and it follows from Example 3.3.6 that $\phi_{22} = 0$. Indeed all lags $h \geq 2$ of the PACF are zero.

More generally, let us briefly consider a causal AR(p) process given by $\phi(B)X_t = Z_t$ with $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$. Then, for $h > p$,

$$X_h^{h-1} = P_{\mathcal{M}} X_h = \sum_{j=1}^{p} \phi_j X_{h-j}$$

and consequently

$$\phi_{hh} = \mathrm{Corr}(X_h - X_h^{h-1}, X_0 - X_0^{h-1}) = \mathrm{Corr}(Z_h, X_0 - X_0^{h-1}) = 0$$

if $h > p$ by causality. Observe, however, that ϕ_{hh} is not necessarily zero if $h \leq p$. On the other hand, for an invertible MA(q) process, we can write $Z_t = \pi(B)X_t$ or, equivalently,

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

which shows that the PACF of an MA(q) process will be nonzero for all lags, since for a "perfect" regression one would have to use all past variables $(X_s\colon s < t)$ instead of only the quantity X_t^{t-1} given in Definition 3.3.1.

In summary, the PACF reverses the behavior of the ACVF for autoregressive and moving average processes. While the latter have an ACVF that vanishes after lag q and a PACF that is nonzero (though decaying) for all lags, AR processes have an ACVF that is nonzero (though decaying) for all lags but a PACF that vanishes after lag p.

ACVF (ACF) and PACF hence provide useful tools in assessing the dependence of given ARMA processes. If the estimated ACVF (the estimated PACF) is essentially zero after some time lag, then the underlying time series can be conveniently modeled with an MA (AR) process—and no general ARMA sequence has to be fitted.

3.4 Forecasting

Suppose that we have observed the variables $X_1, \ldots, X_n$ of a weakly stationary time series $(X_t \colon t \in \mathbb{Z})$ in $L^2(\Omega, \mathcal{A}, P)$ and that our goal is to predict or forecast the future values of $X_{n+1}, X_{n+2}, \ldots$ based on this information. We shall focus here on so-called *one-step best linear predictors (BLP)*. These are, by definition, the projections

$$\hat{X}_{n+1} = P_{\mathcal{K}_n} X_{n+1} = \phi_{n0} + \phi_{n1} X_n + \ldots + \phi_{nn} X_1, \tag{3.4.1}$$

where $\mathcal{K}_n = \overline{\mathrm{sp}}(1, X_1, \ldots, X_n)$, that is, linear combinations of the observations $X_1, \ldots, X_n$ that minimize the mean-squared error

$$\|X_{n+1} - \hat{X}_{n+1}\|^2 = E\big[(X_{n+1} - \hat{X}_{n+1})^2\big]$$

Straightforward generalizations yield definitions for the m-step best linear predictors $\hat{X}_{n+m}$ of X_{n+m} for arbitrary $m \in \mathbb{N}$ in the same fashion. Here

$$\hat{X}_{n+m} = P_{\mathcal{K}_n} X_{n+m} = \phi_{n0}^{(m)} + \phi_{n1}^{(m)} X_n + \ldots + \phi_{nn}^{(m)} X_1,$$

where the coefficients can be found minimizing

$$\|X_{n+m} - \hat{X}_{n+m}\|^2 = E\big[(X_{n+m} - \hat{X}_{n+m})^2\big]$$

Restating the projection theorem, the following result will be the starting point for our considerations.

Theorem 3.4.1 (Best linear prediction) *Let $(X_t \colon t \in \mathbb{Z})$ be a weakly stationary stochastic process of which we observe $X_1, \ldots, X_n$. Then, the one-step BLP $\hat{X}_{n+1}$ of X_{n+1} is determined by the equations*

$$E\left[(X_{n+1} - \hat{X}_{n+1}) X_{n+1-j}\right] = 0$$

for all $j = 1, \ldots, n+1$, where $X_0 = 1$.

The equations specified in Theorem 3.4.1 can be used to calculate the coefficients $\phi_{n0}, \ldots, \phi_{nn}$ in (3.4.1). We can focus on mean zero processes $(X_t \colon t \in \mathbb{Z})$ and thus set $\phi_{n0} = 0$ as the following calculations show. Assume that $E[X_t] = \mu$ for all $t \in \mathbb{Z}$. Then,

Theorem 3.4.1 gives that $E[\hat{X}_{n+1}] = E[X_{n+1}] = \mu$ (using the equation with $j = n+1$). Consequently, it holds that

$$\mu = E[\hat{X}_{n+1}] = E\left[\phi_{n0} + \sum_{\ell=1}^{n} \phi_{n\ell} X_{n+1-\ell}\right] = \phi_{n0} + \sum_{\ell=1}^{n} \phi_{n\ell}\mu.$$

Using now that $\phi_{n0} = \mu(1 - \phi_{n1} - \ldots - \phi_{nn})$, equation (3.4.1) can be rewritten as

$$\hat{Y}_{n+1} = \phi_{n1} Y_n + \ldots + \phi_{nn} Y_1,$$

where $\hat{Y}_{n+1} = \hat{X}_{n+1} - \mu$ has mean zero. This means that instead of projecting on $\mathcal{K}_n$ in the general case, we can focus on projecting the mean zero variable Y_{n+1} onto the space $\mathcal{H}_n$ and $\hat{Y}_{n+1} = P_{\mathcal{H}_n} Y_{n+1}$. Similar reasoning applies to the m-step prediction case.

The following is a repetition of Chapter 2 material. With the ACVF γ of $(X_t \colon t \in \mathbb{Z})$, the equations in Theorem 3.4.1 can be expressed as

$$\sum_{\ell=1}^{n} \phi_{n\ell}\gamma(j-\ell) = \gamma(j), \qquad j = 1, \ldots, n. \tag{3.4.2}$$

Note that due to the convention $\phi_{n0} = 0$, the last equation in Theorem 3.4.1 (for which $j = n+1$) is omitted. More conveniently, this is restated in matrix notation. To this end, let $\boldsymbol{\Gamma}_n = (\gamma(j-\ell))_{j,\ell=1,\ldots,n}$, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})'$ and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))'$, where $'$ denotes the transpose. With these notations, (3.4.2) becomes

$$\boldsymbol{\Gamma}_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n \qquad \Longleftrightarrow \qquad \boldsymbol{\phi}_n = \boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n, \tag{3.4.3}$$

provided that $\boldsymbol{\Gamma}_n$ is nonsingular. The determination of the coefficients $\phi_{n\ell}$ has thus been reduced to solving a linear equation system and depends only on second-order properties of $(X_t)_{t\in\mathbb{Z}}$ which are given by the ACVF γ.

Let $\boldsymbol{X}_n = (X_n, X_{n-1}, \ldots, X_1)'$. Then, $\hat{X}_{n+1} = \boldsymbol{\phi}_n' \boldsymbol{X}_n$. To assess the quality of the prediction, one computes the mean-squared error with the help of (3.4.3) as follows:

$$\begin{aligned}
v_n &= E\left[(X_{n+1} - \hat{X}_{n+1})^2\right] \\
&= E\left[(X_{n+1} - \boldsymbol{\phi}_n' \boldsymbol{X}_n)^2\right] \\
&= E\left[(X_{n+1} - \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n)^2\right] \\
&= E\left[X_{n+1}^2 - 2\boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n X_{n+1} + \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{X}_n \boldsymbol{X}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n\right] \\
&= \gamma(0) - 2\boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n + \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\Gamma}_n \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n \\
&= \gamma(0) - \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n.
\end{aligned} \tag{3.4.4}$$

As an initial example, we explain the prediction procedure for an autoregressive process of order 2.

Example 3.4.1 (Prediction of an AR(2) Process) Let $(X_t\colon t \in \mathbb{Z})$ be the causal AR(2) process $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Suppose that you have only an observation of X_1 to forecast the value of X_2. In this simplified case, the single prediction equation (3.4.2) is

$$\phi_{11}\gamma(0) = \gamma(1),$$

so that $\phi_{11} = \rho(1)$ and $\hat{X}_{1+1} = \rho(1)X_1$. In the next step, assume that we have observed values of X_1 and X_2 at hand to forecast the value of X_3. Then, one similarly obtains from (3.4.2) that the predictor can be computed from

$$\begin{aligned}\hat{X}_{2+1} &= \phi_{21}X_2 + \phi_{22}X_1 = \boldsymbol{\phi}_2'\boldsymbol{X}_2 = (\boldsymbol{\Gamma}_2^{-1}\boldsymbol{\gamma}_2)'\boldsymbol{X}_2 \\ &= (\gamma(1),\gamma(2)) \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} X_2 \\ X_1 \end{pmatrix}.\end{aligned}$$

However, applying the arguments leading to the definition of the PACF in Section 3.3, one finds that

$$\begin{aligned}E\left[\{X_3 - (\phi_1 X_2 + \phi_2 X_1)\}X_1\right] &= E[Z_3 X_1] = 0, \\ E\left[\{X_3 - (\phi_1 X_2 + \phi_2 X_1)\}X_2\right] &= E[Z_3 X_2] = 0.\end{aligned}$$

Hence, $\hat{X}_{2+1} = \phi_1 X_2 + \phi_2 X_1$ and even $\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1}$ for all $n \geq 2$, exploiting the particular autoregressive structure. Since similar results can be proved for general causal AR(p) processes, the one-step predictors have the form

$$\hat{X}_{n+1} = \phi_1 X_n + \ldots + \phi_p X_{n-p+1}$$

whenever the number of observed variables n is at least p.

The major drawback of this approach is immediately apparent from the previous example: For larger sample sizes n, the prediction procedure requires the calculation of the inverse matrix $\boldsymbol{\Gamma}_n^{-1}$ which is computationally expensive. In the remainder of this section, we introduce two recursive prediction methods that bypass the inversion altogether. They are known as *Durbin-Levinson algorithm* and *innovations algorithm.* Finally, we deal with predictors based on the *infinite past* which are, in several cases, easily applicable for the class of causal and invertible ARMA processes.

Method 1 (The Durbin-Levinson algorithm) If $(X_t\colon t \in \mathbb{Z})$ is a zero mean weakly stationary process with ACVF γ such that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then the coefficients $\phi_{n\ell}$ in (3.4.2) and the mean squared errors v_n in (3.4.4) satisfy the recursions

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)}, \qquad v_0 = \gamma(0),$$

and, for $n \geq 1$,

$$\phi_{nn} = \frac{1}{v_{n-1}}\left(\gamma(n) - \sum_{\ell=1}^{n-1}\phi_{n-1,\ell}\gamma(n-\ell)\right),$$

$$\begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{nn} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}$$

and

$$v_n = v_{n-1}(1 - \phi_{nn}^2).$$

Proof. Let $\mathcal{G}_1 = \overline{\text{sp}}(X_2, \ldots, X_n)$ and $\mathcal{G}_2 = \overline{\text{sp}}(X_1 - P_{\mathcal{G}_1} X_1)$. Then, $\mathcal{G}_1$ and $\mathcal{G}_2$ are orthogonal and for any $Y \in L^2$, we have the unique representation $P_{\mathcal{H}_n} Y = P_{\mathcal{G}_1} Y + P_{\mathcal{G}_2} Y$. In particular,

$$\hat{X}_{n+1} = P_{\mathcal{H}_n} X_{n+1} = P_{\mathcal{G}_1} X_{n+1} + P_{\mathcal{G}_2} X_{n+1} = P_{\mathcal{G}_1} X_{n+1} + \alpha(X_1 - P_{\mathcal{G}_1} X_1),$$

where $\alpha = \langle X_{n+1}, X_1 - P_{\mathcal{G}_1} X_1 \rangle / \|X_1 - P_{\mathcal{G}_1} X_1\|^2$. Using the prediction equations, stationarity and the symmetry of the covariance matrix Γ_n, we obtain the representations

$$P_{\mathcal{G}_1} X_{n+1} = \sum_{\ell=1}^{n-1} \phi_{n-1,\ell} X_{n+1-\ell},$$

$$P_{\mathcal{G}_1} X_1 = \sum_{\ell=1}^{n-1} \phi_{n-1,\ell} X_{\ell+1} = \sum_{\ell=1}^{n-1} \phi_{n-1,n-\ell} X_{n+1-\ell}$$

as well as

$$\|X_1 - P_{\mathcal{G}_1} X_1\|^2 = \|X_{n+1} - P_{\mathcal{G}_1} X_{n+1}\|^2 = \|X_n - \hat{X}_n\|^2 = v_{n-1}.$$

Combining the previous steps gives that

$$\hat{X}_{n+1} = \alpha X_1 + \sum_{\ell=1}^{n-1} \big(\phi_{n-1,\ell} - \alpha \phi_{n-1,n-\ell}\big) X_{n+1-\ell},$$

with

$$\alpha = \frac{1}{v_{n-1}} \Big(\langle X_{n+1}, X_1 \rangle - \sum_{\ell=1}^{n-1} \phi_{n-1,\ell} \langle X_{n+1}, X_{\ell+1} \rangle \Big)$$

$$= \frac{1}{v_{n-1}} \Big(\gamma(n) - \sum_{\ell=1}^{n-1} \phi_{n-1,\ell} \gamma(n-\ell) \Big)$$

On the other hand, we know that the representation $\hat{X}_{n+1} = \phi_{n1} X_n + \ldots + \phi_{nn} X_1$ is unique (Γ_n is invertible by assumption on the ACVF γ). Comparing coefficients implies therefore that $\phi_{nn} = \alpha$ and $\phi_{n\ell} = \phi_{n-1,\ell} - \alpha \phi_{n-1,n-\ell} = \phi_{n-1,\ell} - \phi_{nn} \phi_{n-1,n-\ell}$ for $\ell = 1, \ldots, n-1$. It remains to verify that $v_n = v_{n-1}(1 - \phi_{nn}^2)$. Note that

$$\begin{aligned} v_n &= \|X_{n+1} - \hat{X}_{n+1}\|^2 \\ &= \|X_{n+1} - P_{\mathcal{G}_1} X_{n+1} - P_{\mathcal{G}_2} X_{n+1}\|^2 \\ &= \|X_{n+1} - P_{\mathcal{G}_1} X_{n+1}\|^2 + \|P_{\mathcal{G}_2} X_{n+1}\|^2 - 2\langle X_{n+1} - P_{\mathcal{G}_1} X_{n+1}, P_{\mathcal{G}_2} X_{n+1} \rangle \\ &= v_{n-1} + \alpha v_{n-1} - 2\alpha \langle X_{n+1}, X_1 - P_{\mathcal{G}_1} X_1 \rangle \\ &= v_{n-1}(1 - \phi_{nn}^2). \end{aligned}$$

Therein, we have used that $P_{\mathcal{G}_2} X_{n+1} = \alpha(X_1 - P_{\mathcal{G}_1} X_1)$ and the definition of α to obtain the fourth and fifth equality sign. This completes the proof. □

It can be shown that under the assumptions made on the process $(X_t\colon t \in \mathbb{Z})$, it holds indeed that ϕ_{nn} is equal to the value of the PACF of $(X_t\colon t \in \mathbb{Z})$ at lag n. The result is formulated as Corollary 5.2.1 in Brockwell and Davis (1991). We highlight this fact in an example.

Example 3.4.2 (The PACF of an AR(2) process) Let $(X_t\colon t \in \mathbb{Z})$ be a causal AR(2) process. Then, $\rho(1) = \phi_1/(1-\phi_2)$ and all other values can be computed recursively from

$$\rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \qquad h \geq 2.$$

Note that the ACVF γ satisfies a difference equation with the same coefficients, which is easily seen by multiplying the latter equation with $\gamma(0)$. Applying the Durbin-Levinson algorithm gives first that

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)} = \rho(1) \qquad \text{and} \qquad v_1 = v_0(1-\phi_{11}^2) = \gamma(0)(1-\rho(1)^2).$$

Ignoring the recursion for the error terms v_n in the following, the next $\phi_{n\ell}$ values are obtained as

$$\begin{aligned}
\phi_{22} &= \frac{1}{v_1}\left[\gamma(2) - \phi_{11}\gamma(1)\right] = \frac{1}{1-\rho(1)^2}\left[\rho(2) - \rho(1)^2\right] \\
&= \frac{\phi_1^2(1-\phi_2)^{-1} + \phi_2 - [\phi_1(1-\phi_2)^{-1}]^2}{1 - [\phi_1(1-\phi_2)^{-1}]^2} = \phi_2,
\end{aligned}$$

$$\phi_{21} = \phi_{11} - \phi_{22}\phi_{11} = \rho(1)(1-\phi_2) = \phi_1,$$

$$\phi_{33} = \frac{1}{v_2}\left[\gamma(3) - \phi_{21}\gamma(2) - \phi_{22}\gamma(1)\right] = \frac{1}{v_2}\left[\gamma(3) - \phi_1\gamma(2) - \phi_2\gamma(2)\right] = 0.$$

Now, referring to the remarks after Example 3.3.7, no further computations are necessary to determine the PACF because $\phi_{nn} = 0$ for all $n > p = 2$.

Method 2 (The innovations algorithm) In contrast to the Durbin-Levinson algorithm, this method can also be applied to nonstationary processes. It should thus, in general, be preferred over Method 1. The nonstationary version will prove in particular useful in finding a simplified version for predictors of causal and invertible ARMA(p,q) processes. The innovations algorithm gets its name from the fact that one directly uses the form of the prediction equations in Theorem 3.4.1 which are stated in terms of the *innovations* $(X_{t+1} - \hat{X}_{t+1}\colon t \in \mathbb{Z})$. Observe that the sequence consists of uncorrelated random variables.

Let $(X_t\colon t \in \mathbb{Z})$ be a centered stochastic process with autocovariances $\gamma(i,j) = E[X_i X_j]$ such that the matrix $(\gamma(i,j))_{i,j=1,\ldots,n}$ is regular for all $n \in \mathbb{N}$. Then, the one-step predictors $\hat{X}_{n+1}$ can be calculated from the recursions

$$\hat{X}_{0+1} = 0, \qquad v_0 = \gamma(1,1)$$

and, for $n \geq 1$,

$$\hat{X}_{n+1} = \sum_{\ell=1}^{n} \theta_{n\ell}(X_{n+1-\ell} - \hat{X}_{n+1-\ell}),$$

$$v_n = \gamma(0) - \sum_{\ell=0}^{n-1} \theta_{n,n-\ell}^2 v_\ell,$$

where the coefficients are obtained from the equations

$$\theta_{n,n-\ell} = \frac{1}{v_\ell}\left(\gamma(n+1,\ell+1) - \sum_{i=0}^{\ell-1} \theta_{\ell,\ell-i}\theta_{n,n-i}v_i\right), \qquad \ell = 0, 1, \ldots, n-1.$$

Proof. Note first that $X_i - \hat{X}_i \in \mathcal{H}_{j-1}$ if $i < j$ and that $X_j - \hat{X}_j$ is orthogonal to $\mathcal{H}_{j-1}$ by definition of $\hat{X}_j$. This shows that $\{X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n\}$ is an orthogonal set. Using the definition of $\hat{X}_{n+1}$, we therefore obtain

$$\langle \hat{X}_{n+1}, X_{\ell+1} - \hat{X}_{\ell+1}\rangle = \sum_{j=1}^{n} \theta_{nj}\langle X_{n+1-j} - \hat{X}_{n+1-j}, X_{\ell+1} - \hat{X}_{\ell+1}\rangle = \theta_{n,n-\ell}v_\ell \qquad (3.4.5)$$

Consequently,

$$\begin{aligned}
\theta_{n,n-\ell} &= \frac{1}{v_\ell}\langle \hat{X}_{n+1}, X_{\ell+1} - \hat{X}_{\ell+1}\rangle \\
&= \frac{1}{v_\ell}\langle X_{n+1}, X_{\ell+1} - \hat{X}_{\ell+1}\rangle \\
&= \frac{1}{v_\ell}\Big\langle X_{n+1}, X_{\ell+1} - \sum_{i=1}^{\ell} \theta_{\ell i}(X_{\ell+1-i} - \hat{X}_{\ell+1-i})\Big\rangle \\
&= \frac{1}{v_\ell}\Big\langle X_{n+1}, X_{\ell+1} - \sum_{i=0}^{\ell-1} \theta_{\ell,\ell-i}(X_{i+1} - \hat{X}_{i+1})\Big\rangle \\
&= \frac{1}{v_\ell}\Big(\langle X_{n+1}, X_{\ell+1}\rangle - \sum_{i=0}^{\ell-1} \theta_{\ell,\ell-i}\langle X_{n+1}, X_{i+1} - \hat{X}_{i+1}\rangle\Big) \\
&= \frac{1}{v_\ell}\Big(\gamma(n+1,\ell+1) - \sum_{i=0}^{\ell-1} \theta_{\ell,\ell-i}\theta_{n,n-i}v_i\Big).
\end{aligned}$$

Here, the first equality sign is obtained noticing that $\hat{X}_{n+1} = X_{n+1} + (\hat{X}_{n+1} - X_{n+1})$ with $\hat{X}_{n+1} - X_{n+1}$ being orthogonal to $X_{\ell+1} - \hat{X}_{\ell+1}$. The second equality follows upon using the innovations expansion for $X_{\ell+1}$, while the last equality follows from (3.4.5). The recursion for v_n can be proved similarly and is omitted. □

As example we show how the innovations algorithm is applied to a moving average time series of order 1.

Example 3.4.3 (Prediction of an MA(1) Process) Let $(X_t\colon t \in \mathbb{Z})$ be the MA(1) process $X_t = Z_t + \theta Z_{t-1}$. Note that

$$\gamma(i,i) = \gamma_X(0) = (1+\theta^2)\sigma^2, \qquad \gamma(i,i+1) = \gamma_X(1) = \theta\sigma^2, \qquad i \in \mathbb{N},$$

and $\gamma(i,j) = 0$ if $|i-j| \geq 2$. Using the innovations algorithm, we can compute the one-step predictor from the values

$$\theta_{n1} = \frac{\theta\sigma^2}{v_{n-1}}, \qquad \theta_{n\ell} = 0 \quad (\ell = 2,\dots,n),$$

and

$$\begin{aligned} v_0 &= (1+\theta^2)\sigma^2, \\ v_n &= (1+\theta^2-\theta\theta_{n1})\sigma^2 \end{aligned}$$

as

$$\hat{X}_{n+1} = \frac{\theta\sigma^2}{v_{n-1}}(X_n - \hat{X}_n).$$

Example 3.4.4 (Innovations Algorithm for ARMA(p,q)) A major simplification can be achieved in the prediction of a causal and invertible ARMA(p,q) process $(X_t\colon t \in \mathbb{Z})$ if the innovations algorithm is not applied directly but to the transformed series $(W_t\colon t \in \mathbb{N})$ given by

$$W_t = \begin{cases} \sigma^{-1}X_t, & t = 1,\dots,m, \\ \sigma^{-1}\phi(B)X_t, & t = m+1, m+2, \dots, \end{cases}$$

where $m = \max\{p,q\}$. It can be seen that $\mathcal{H}_n = \overline{\mathrm{sp}}(W_1,\dots,W_n)$ and one can explore the connection between the predictors $\hat{X}_{n+1} = P_{\mathcal{H}_n}X_{n+1}$ and $\hat{X}_{n+1} = P_{\mathcal{H}_n}W_{n+1}$. It is clear from its definition that $(W_t\colon t \in \mathbb{N})$ is nonstationary. Its covariances can be expressed in terms of the ACVF γ_X as

$$\gamma(i,j) = \begin{cases} \sigma^{-2}\gamma_X(|i-j|), & 1 \leq i,j \leq m, \\ \sigma^{-2}\big[\gamma_X(|i-j|) - \sum_{\ell=1}^{p}\phi_\ell\gamma_X(\ell-|i-j|)\big], & \min\{i,j\} \leq m < \max\{i,j\} \leq 2m, \\ \sum_{\ell=0}^{q}\theta_\ell\theta_{\ell+|i-j|}, & \min\{i,j\} > m, \\ 0, & \text{else}, \end{cases}$$

where $\theta_0 = 1$ and $\theta_\ell = 0$ if $\ell > q$. An application of the innovations algorithm gives

$$\hat{W}_{n+1} = \begin{cases} \sum_{\ell=1}^{n}\theta_{n\ell}(W_{n+1-\ell} - \hat{W}_{n+1-\ell}), & 1 \leq n < m, \\ \sum_{\ell=1}^{q}\theta_{n\ell}(W_{n+1-\ell} - \hat{W}_{n+1-\ell}), & n \geq m, \end{cases}$$

with corresponding prediction error $r_n = E[(W_{n+1} - \hat{W}_{n+1})^2] = \sigma^{-2}v_n$. This is because the form of $\gamma(i,j)$ implies that $\theta_{n\ell} = 0$ if $\ell > q$ and $n \geq m$. The predictors $\hat{W}_t$ can be related to the predictors $\hat{X}_t$ as

$$\hat{W}_t = P_{\mathcal{H}_{t-1}}W_t = \begin{cases} \sigma^{-1}\hat{X}_t, & t = 1,\dots,m. \\ \sigma^{-1}(\hat{X}_t - \sum_{\ell=1}^{p}\phi_\ell X_{t-\ell}), & t = m+1, m+2, \dots. \end{cases}$$

Therefore $X_t - \hat{X}_t = \sigma(W_t - \hat{W}_t)$ for all $t \in \mathbb{N}$ and thus

$$\hat{X}_{n+1} = \begin{cases} \sum_{\ell=1}^{n} \theta_{n\ell}(X_{n+1-\ell} - \hat{X}_{n+1-\ell}), & 1 \le n < m, \\ \sum_{\ell=1}^{p} \phi_\ell X_{n+1-\ell} + \sum_{\ell=1}^{q} \theta_{n\ell}(X_{n+1-\ell} - \hat{X}_{n+1-\ell}), & n \ge m, \end{cases}$$

The prediction error is $v_n = \sigma^2 r_n$.

Method 3 (Prediction based on the infinite past) Suppose that we are analyzing a causal and invertible ARMA(p, q) process. Assume further that we have the (unrealistic) ability to store the history of the process and that we can thus access all past variables $(X_t \colon t \le n)$. Define now $\mathcal{M}_n = \overline{\text{sp}}(X_t \colon t \le n)$ and

$$\tilde{X}_{n+m} = P_{\mathcal{M}_n} X_{n+m} = E[X_{n+m} | X_n, X_{n-1}, \ldots],$$

as the m-step ahead predictor based on the infinite past. It can be shown that, for large sample sizes n, the difference between the values of $\hat{X}_{n+m}$ and $\tilde{X}_{n+m}$ vanishes at an exponential rate. Exploiting causality and invertibility of the ARMA process, one can transform the predictor $\tilde{X}_{n+m}$ so that it is in a computationally more feasible form. To do so, note that by causality

$$\begin{aligned} \tilde{X}_{n+m} &= E[X_{n+m} | X_n, X_{n-1}, \ldots] \\ &= E\left[\sum_{j=0}^{\infty} \psi_j Z_{n+m-j} \Big| X_n, X_{n-1}, \ldots \right] \\ &= \sum_{j=m}^{\infty} \psi_j Z_{n+m-j} \end{aligned} \tag{3.4.6}$$

because $E[Z_t | X_n, X_{n-1}, \ldots]$ equals zero if $t > n$ and equals Z_t if $t \le n$ (due to invertibility!). The representation in (3.4.6) can be used to compute the mean squared prediction error $\tilde{P}_{n+m}$. Using causality, we obtain that

$$\tilde{P}_{n+m} = E[(X_{n+m} - \tilde{X}_{n+m})^2] = E\left[\Big(\sum_{j=0}^{m-1} \psi_j Z_{n+m-j} \Big)^2 \right] = \sigma^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.4.7}$$

On the other hand, (3.4.6) does not allow to directly calculate the forecasts because $\tilde{X}_{n+m}$ is given in terms of the noise variables Z_{n+m-j}. Instead we will use invertibility. Observe first that

$$E[X_{n+m-j} | X_n, X_{n-1}, \ldots] = \begin{cases} \tilde{X}_{n+m-j}, & j < m. \\ X_{n+m-j}, & j \ge m. \end{cases}$$

By invertibility (the "0 =" part follows again from causality),

$$\begin{aligned} 0 &= E[Z_{n+m} | X_n, X_{n-1}, \ldots] \\ &= E\left[\sum_{j=0}^{\infty} \pi_j X_{n+m-j} \Big| X_n, X_{n-1}, \ldots \right] \end{aligned}$$

$$= \sum_{j=0}^{\infty} \pi_j E[X_{n+m-j} | X_n, X_{n-1}, \ldots].$$

Combining the previous two statements, we arrive at

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}. \tag{3.4.8}$$

The equations can now be solved recursively for $m = 1, 2, \ldots$ Note, however, that for any $m \geq 1$ the sequence $(X_{n+m+t} - \tilde{X}_{n+m+t} \colon t \in \mathbb{Z})$ does not consist of uncorrelated random variables. In fact, if $h \in \mathbb{N}_0$, it holds that

$$\begin{aligned} &E[(X_{n+m} - \tilde{X}_{n+m})(X_{n+m+h} - \tilde{X}_{n+m+h})] \\ &\quad = E\left[\sum_{j=0}^{m-1} \psi_j Z_{n+m-j} \sum_{i=0}^{m+h-1} \psi_i Z_{n+m+h-i}\right] \\ &\quad = \sigma^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+h}. \end{aligned}$$

Finally, for practical purposes we need to truncate the given forecast for n large. This is accomplished by setting

$$\sum_{j=n+m}^{\infty} \pi_j X_{n+m-j} = 0.$$

The resulting equations (see (3.4.8) for comparison) yield recursively the truncated m-step predictors X^*_{n+m}:

$$X^*_{n+m} = -\sum_{j=1}^{m-1} \pi_j X^*_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j X_{n+m-j}. \tag{3.4.9}$$

3.5 Parameter Estimation

Let $(X_t \colon t \in \mathbb{Z})$ be a causal and invertible ARMA(p, q) process with known orders p and q, possibly with mean μ. We are in this section concerned with estimation procedures for the unknown parameter vector

$$\boldsymbol{\beta} = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'. \tag{3.5.1}$$

To simplify the estimation procedure, we assume that we can work with data that has been adjusted by subtraction of the mean and we can restrict the discussion to zero mean ARMA models.

In the following, we shall introduce three methods of estimation. The method of moments works best in case of pure AR processes, while it does not lead to efficient estimation procedures for general ARMA processes. For the latter, more efficient estimators are

provided by the maximum likelihood and least squares methods which will be discussed subsequently.

Method 1 (Method of Moments) Since this method is only efficient in their case, we restrict the presentation here to AR(p) processes

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$. The parameter vector $\boldsymbol{\beta}$ consequently reduces to $(\boldsymbol{\phi}, \sigma^2)'$ with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and can be estimated using the *Yule-Walker equations*

$$\boldsymbol{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \qquad \text{and} \qquad \sigma^2 = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_p,$$

where $\boldsymbol{\Gamma}_p = (\boldsymbol{\gamma}(k-j))_{k,j=1,\ldots,p}$ and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$. Observe that the equations are obtained by the same arguments applied to derive the Durbin-Levinson algorithm in the previous section. The method of moments suggests to replace every quantity in the Yule-Walker equations with their estimated counterparts, which yields the *Yule-Walker estimators*

$$\widehat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p = \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p \tag{3.5.2}$$

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p' \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p = \hat{\gamma}(0) \left[1 - \hat{\boldsymbol{\rho}}_p' \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p \right]. \tag{3.5.3}$$

Therein, $\hat{\boldsymbol{R}}_p = \hat{\gamma}(0)^{-1} \hat{\boldsymbol{\Gamma}}_p$ and $\hat{\boldsymbol{\rho}}_p = \hat{\gamma}(0)^{-1} \hat{\boldsymbol{\gamma}}_p$ with $\hat{\gamma}(h)$ defined as in (1.2.1). Using $\hat{\gamma}(h)$ as estimator for the ACVF at lag h, we implicitly obtain a dependence on the sample size n. This dependence is suppressed in the notation used here. The following theorem contains the limit behavior of the Yule-Walker estimators as n tends to infinity.

Theorem 3.5.1 *If $(X_t \colon t \in \mathbb{Z})$ is a causal AR(p) process with $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{IID}(0, \sigma^2)$, then*

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma}_p^{-1}) \qquad \textit{and} \qquad \hat{\sigma}^2 \xrightarrow{P} \sigma^2$$

as $n \to \infty$, where $\to^P$ indicates convergence in probability.

Proof. Putting the observations $X_1, \ldots, X_n$ into a vector $\mathbf{Y}$ and using the autoregression equations yields the linear model

$$\mathbf{Y} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_0 & X_{-1} & \ldots & X_{1-p} \\ X_1 & X_0 & \ldots & X_{2-p} \\ \vdots & \vdots & & \vdots \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = X\boldsymbol{\phi} + \boldsymbol{Z}.$$

The foregoing motivates the (infeasible) estimator $\boldsymbol{\phi}^* = (X'X)^{-1} X'\boldsymbol{Y}$. In the following we will show that

$$\sqrt{n}(\boldsymbol{\phi}^* - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2 \Gamma_p^{-1}) \tag{3.5.4}$$

and

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*) = o_P(1), \tag{3.5.5}$$

from which the first assertion of the theorem readily follows.

To prove (3.5.4) observe that

$$\sqrt{n}(\boldsymbol{\phi}^* - \boldsymbol{\phi}) = \sqrt{n}\big[(X'X)^{-1}X'\boldsymbol{Y} - \boldsymbol{\phi}\big] = n(X'X)^{-1}\frac{1}{\sqrt{n}}X'\boldsymbol{Z}.$$

For $t \in \mathbb{N}$ set $\mathbf{U}_t = (X_{t-1}, \ldots, X_{t-p})'Z_t$. Then,

$$\frac{1}{\sqrt{n}}X'\boldsymbol{Z} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbf{U}_t,$$

where $E[\mathbf{U}_t] = \mathbf{0}$, $E[\mathbf{U}_t\mathbf{U}_t'] = \sigma^2\Gamma_p$ and $E[\mathbf{U}_t\mathbf{U}_{t+h}'] = 0_{p\times p}$ for all $h \neq 0$. Recall the causal representation $\sum_{j=1}^{\infty}\psi_j Z_{t-j}$ and define its truncated version

$$X_t^{(m)} = \sum_{j=1}^{m}\psi_j Z_{t-j}, \qquad t \in \mathbb{Z},$$

as well as $\mathbf{U}_t^{(m)} = (X_{t-1}^{(m)}, \ldots, X_{t-p}^{(m)})'Z_t$. Let $\boldsymbol{\lambda} \in \mathbb{R}^p$. Then

$$\boldsymbol{\lambda}'\mathbf{U}_t^{(m)} = \sum_{i=1}^{p}\lambda_i X_{t-i}Z_t$$

and $E[\boldsymbol{\lambda}'\mathbf{U}_t^{(m)}] = 0$, $E[(\boldsymbol{\lambda}'\mathbf{U}_t^{(m)})^2] = \sigma^2\boldsymbol{\lambda}'\Gamma_p^{(m)}\boldsymbol{\lambda}$ and $E[(\boldsymbol{\lambda}'\mathbf{U}_t^{(m)})\boldsymbol{\lambda}'\mathbf{U}_{t+h}^{(m)})] = 0$ for $h \neq 0$, where $\Gamma_p^{(m)}$ is the variance-covariance matrix of the vector $(X_{t-1}^{(m)}, \ldots, X_{t-p}^{(m)})'$. Therefore, $(\boldsymbol{\lambda}'\mathbf{U}_t^{(m)} : t \in \mathbb{Z})$ is a strictly stationary, $(m+p)$-dependent white noise sequence. The central limit theorem for such sequences (see Theorem 6.4.2 in Brockwell and Davis, 1991) implies that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{\lambda}'\mathbf{U}_t^{(m)} \xrightarrow{\mathcal{D}} \boldsymbol{\lambda}'\mathbf{V}^{(m)},$$

where $\mathbf{V}^{(m)} \sim N(\mathbf{0}, \sigma^2\Gamma_p^{(m)})$. Now $\sigma^2\Gamma_p^{(m)} \to \sigma^2\Gamma_p$ and so we have that $\boldsymbol{\lambda}'\mathbf{V}^{(m)}$ converges in distribution to $\boldsymbol{\lambda}'\mathbf{V}$, where $\mathbf{V} \sim N(\mathbf{0}, \sigma^2\Gamma_p)$. Since $n^{-1}E[(\boldsymbol{\lambda}'\sum_{t=1}^{n}(\mathbf{U}_t - \mathbf{U}_t^{(m)}))^2] \to 0$ and since $X_t^{(m)}$ converges to X_t in mean square, (3.5.4) follows from the Cramér-Wold device and the fact that $n(X'X)^{-1}$ converges to Γ_p^{-1} in probability.

It remains to prove (3.5.5). To this end, write

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*) &= \sqrt{n}\big(\hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p - (X'X)^{-1}X'\boldsymbol{Y}\big) \\
&= \sqrt{n}\hat{\Gamma}_p^{-1}\big(\hat{\boldsymbol{\gamma}}_p - \tfrac{1}{n}X'\boldsymbol{Y}\big) + \sqrt{n}\big(\hat{\Gamma}_p^{-1} - n(X'X)^{-1}\big)\tfrac{1}{n}X'\boldsymbol{Y} \\
&= R_1 + R_2.
\end{aligned}$$

The ith coordinate of $\hat{\Gamma}_p R_1 = \sqrt{n}(\hat{\boldsymbol{\gamma}}_p - \frac{1}{n}X'\boldsymbol{Y})$ is given by

$$\frac{1}{\sqrt{n}}\left(\sum_{k=1}^{n-i}(X_k - \bar{X}_n)(X_{k+i} - \bar{X}_n) - \sum_{k=1-i}^{n-i}X_k X_{k+i}\right)$$

$$= \frac{1}{\sqrt{n}} \sum_{k=1-i}^{0} X_k X_{k+i} + \sqrt{n}\bar{X}_n \left(\Big(1 - \frac{i}{n}\Big)\bar{X}_n - \frac{1}{n}\sum_{k=1}^{n-1}(X_k + X_{k+1}) \right)$$

which converges to zero in probability. Since $\hat{\Gamma}_p \to \Gamma_p$ in probability, we obtain that $R_1 = o_P(1)$. To show that $R_2 = o_P(1)$ note that it suffices to show that $\sqrt{n}\|\hat{\Gamma}_p^{-1} - n(X'X)^{-1}\| = o_P(1)$. But this is done as follows:

$$\begin{aligned}\sqrt{n}\|\hat{\Gamma}_p^{-1} - n(X'X)^{-1}\| &= \sqrt{n}\|\hat{\Gamma}_p^{-1}(\tfrac{1}{n}X'X - \hat{\Gamma}_p)n(X'X)^{-1}\| \\ &\leq \sqrt{n}\|\hat{\Gamma}_p^{-1}\|\|\tfrac{1}{n}X'X - \hat{\Gamma}_p\|\|n(X'X)^{-1}\|,\end{aligned}$$

where $\sqrt{n}\|\frac{1}{n}X'X - \hat{\Gamma}_p\| = o_P(1)$ follows from the last display above. The other norms are both bounded in probability. Combining these results gives (3.5.5).

Since both $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}$ and $\hat{\boldsymbol{\phi}} \to \boldsymbol{\phi}$ in probability, we must have $\hat{\sigma}^2 \to \sigma^2$ in probability also. This completes the proof of the theorem. □

Since equations (3.5.2) and (3.5.3) have the same structure as the corresponding equations (3.4.3) and (3.4.4), the Durbin-Levinson algorithm can be used to solve recursively for the estimators $\widehat{\boldsymbol{\phi}}_h = (\widehat{\phi}_{h1}, \ldots, \widehat{\phi}_{hh})$. Moreover, since ϕ_{hh} is equal to the value of the PACF of $(X_t)_{t\in\mathbb{Z}}$ at lag h, the estimator $\widehat{\phi}_{hh}$ can be used as its proxy. Since we already know that, in the case of AR(p) processes, $\phi_{hh} = 0$ if $h > p$, Theorem 3.5.1 implies immediately the following corollary.

Corollary 3.5.1 *If $(X_t\colon t \in \mathbb{Z})$ is a causal AR(p) process, then*

$$\sqrt{n}\widehat{\phi}_{hh} \xrightarrow{\mathcal{D}} Z \qquad (n \to \infty)$$

for all $h > p$, where Z stands for a standard normal random variable.

Example 3.5.1 (Yule-Walker estimates for AR(2) processes) Suppose that we have observed $n = 144$ values of the autoregressive process $X_t = 1.5X_{t-1} - .75X_{t-2} + Z_t$, where $(Z_t\colon t \in \mathbb{Z})$ is a sequence of independent standard normal variates. Assume further that $\hat{\gamma}(0) = 8.434$, $\hat{\rho}(1) = 0.834$ and $\hat{\rho}(2) = 0.476$ have been calculated from the data. The Yule-Walker estimators for the parameters are then given by

$$\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} = \begin{pmatrix} 1.000 & 0.834 \\ 0.834 & 1.000 \end{pmatrix}^{-1} \begin{pmatrix} 0.834 \\ 0.476 \end{pmatrix} = \begin{pmatrix} 1.439 \\ -0.725 \end{pmatrix}$$

and

$$\hat{\sigma}^2 = 8.434 \left[1 - (0.834, 0.476) \begin{pmatrix} 1.439 \\ -0.725 \end{pmatrix} \right] = 1.215.$$

To construct asymptotic confidence intervals using Theorem 3.5.1, the unknown limiting covariance matrix $\sigma^2\boldsymbol{\Gamma}_p^{-1}$ needs to be estimated. This can be done using the estimator

$$\frac{\hat{\sigma}^2\hat{\boldsymbol{\Gamma}}_p^{-1}}{n} = \frac{1}{144}\frac{1.215}{8.434} \begin{pmatrix} 1.000 & 0.834 \\ 0.834 & 1.000 \end{pmatrix}^{-1} = \begin{pmatrix} 0.057^2 & -0.003 \\ -0.003 & 0.057^2 \end{pmatrix}.$$

Then, the $1-\alpha$ level confidence interval for the parameters ϕ_1 and ϕ_2 are computed as

$$1.439 \pm 0.057 z_{1-\alpha/2} \qquad \text{and} \qquad -0.725 \pm 0.057 z_{1-\alpha/2},$$

respectively, where $z_{1-\alpha/2}$ is the corresponding normal quantile.

Example 3.5.2 Consider the invertible MA(1) process $X_t = Z_t + \theta Z_{t-1}$, where $|\theta| < 1$. Using invertibility, each X_t has an infinite autoregressive representation

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

that is nonlinear in the unknown parameter θ to be estimated. The method of moments is here based on solving

$$\hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{1+\hat{\theta}^2}.$$

for $\hat{\theta}$. The foregoing quadratic equation has the two solutions

$$\hat{\theta} = \frac{1 \pm \sqrt{1-4\hat{\rho}(1)^2}}{2\hat{\rho}(1)},$$

of which we pick the invertible one. Note moreover, that $|\hat{\rho}(1)|$ is not necessarily less or equal to $1/2$ which is required for the existence of real solutions. (The theoretical value $|\rho(1)|$, however, is always less than $1/2$ for any MA(1) process, as an easy computation shows). Hence, θ can not always be estimated from given data samples.

Method 2 (Maximum Likelihood Estimation) In a first step, we rewrite the likelihood of a possibly correlated Gaussian process with the help of the innovations algorithm. Let $(X_t \colon t \in \mathbb{Z})$ be a centered Gaussian process with ACVF $\gamma(i,j) = E[X_i X_j]$. Assume we have observed $X_1, \ldots, X_n$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$ and assume that $\Gamma_n = E[\boldsymbol{X}_n \boldsymbol{X}_n']$ is regular. The Gaussian likelihood is then given by

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp\big(-\tfrac{1}{2} \boldsymbol{X}_n' \Gamma_n^{-1/2} \boldsymbol{X}_n\big).$$

In the following, we will simplify this likelihood with the use of the innovations algorithm, in particular we shall avoid the inversion of Γ_n. Let $\hat{\boldsymbol{X}}_n = (\hat{X}_1, \ldots, \hat{X}_n)'$, where $\hat{X}_1 = 0$ and

$$\hat{X}_{i+1} = P_{\mathcal{H}_i} X_{i+1} = E[X_i | X_{i-1}, \ldots, X_1] = \sum_{\ell=1}^{i} \theta_{i\ell} (X_{i+1-\ell} - \hat{X}_{i+1-\ell})$$

for $i = 1, \ldots, n-1$, where the coefficients $\theta_{i\ell}$, $\ell = 1, \ldots, i$ and $i = 1, 2, \ldots$, are supplied by the innovations algorithm. This leads to the equation system

$$\hat{\boldsymbol{X}}_n = \begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \\ \hat{X}_3 \\ \vdots \\ \hat{X}_n \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ \theta_{11} & 0 & 0 & \ldots & 0 \\ \theta_{22} & \theta_{21} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \ldots & 0 \end{pmatrix} \begin{pmatrix} X_1 - \hat{X}_1 \\ X_2 - \hat{X}_2 \\ X_3 - \hat{X}_3 \\ \vdots \\ X_n - \hat{X}_n \end{pmatrix}$$

$$= T(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n) = (C - I)(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n),$$

so that $C = T + I$ where I is the identity matrix. This implies that $\boldsymbol{X}_n = C(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)$. Notice that $E[(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)'] = \operatorname{diag}(v_0, \ldots, v_{n-1}) = D$. Therefore

$$\Gamma_n = E[\boldsymbol{X}_n \boldsymbol{X}_n'] = E[C(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)'C'] = CDC'$$

and

$$\begin{aligned}
\boldsymbol{X}_n \Gamma_n^{-1} \boldsymbol{X}_n &= (\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)' C' (C')^{-1} D^{-1} C^{-1} C (\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n) \\
&= (\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)' D^{-1} (\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n) \\
&= \sum_{\ell=1}^{n} \frac{(X_\ell - \hat{X}_\ell)^2}{v_{\ell-1}}.
\end{aligned}$$

Furthermore $\det \Gamma_n = (\det C)^2 \det D = v_0 \cdot \ldots \cdot v_{n-1}$. This yields the reformulation

$$L(\Gamma_n) = (2\pi)^{-n/2} (v_0 \cdot \ldots \cdot v_{n-1})^{-1/2} \exp\left(-\frac{1}{2} \sum_{\ell=1}^{n} \frac{(X_\ell - \hat{X}_\ell)^2}{v_{\ell-1}} \right).$$

of the above likelihood. In many cases the likelihood can be expressed in terms of a finite-dimensional parameter vector $\boldsymbol{\beta}$, for example in the case of causal ARMA(p, q) processes discussed in the next paragraph. Then, the likelihood can be simplified further. If the underlying process $(X_t \colon t \in \mathbb{Z})$ is non-Gaussian, $L(\Gamma_n)$ is viewed as a quasi-likelihood function.

Let now $(X_t \colon t \in \mathbb{Z})$ be a causal ARMA(p, q) process and assume further that the moving average polynomial satisfies $\theta(z) \neq 0$ for $|z| < 1$ (which is slightly less restrictive than assuming invertibility). The latter can always be achieved by adjusting the MA coefficients and the innovations variance accordingly. The innovations algorithm gives

$$\hat{X}_{i+1} = \sum_{\ell=1}^{i} \theta_{i\ell} (X_{i+1-\ell} - \hat{X}_{i+1-\ell}), \qquad 1 \leq i < \max\{p, q\},$$

$$\hat{X}_{i+1} = \sum_{\ell=1}^{p} \phi_\ell X_{i+1-\ell} + \sum_{\ell=1}^{q} \theta_{i\ell} (X_{i+1-\ell} - \hat{X}_{i+1-\ell}), \qquad i \geq \max\{p, q\},$$

with prediction error

$$v_i = \sigma^2 r_i.$$

Importantly, in these last expressions we have that the coefficients $\theta_{i\ell}$ and the r_i are independent of σ^2. Recall that in the ARMA case $\boldsymbol{\beta} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)'$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. Consequently,

$$L(\boldsymbol{\beta}) = (2\pi\sigma^2)^{-n/2} (r_0 \cdot \ldots \cdot r_{n-1})^{-1/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} \frac{(X_\ell - \hat{X}_\ell)^2}{r_{\ell-1}} \right). \tag{3.5.6}$$

Taking the partial derivative of $\ln L(\boldsymbol{\beta})$ with respect to the variable σ^2 reveals that the MLE for σ^2 can be calculated from

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{n}, \qquad S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = \sum_{\ell=1}^{n} \frac{(X_\ell - \hat{X}_\ell)^2}{r_{\ell-1}}.$$

Therein, $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ denote the MLEs of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ obtained from minimizing the *profile likelihood* or *reduced likelihood*

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}) = \ln\left(\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n}\right) + \frac{1}{n}\sum_{\ell=1}^{n} \ln(r_{\ell-1}).$$

which can also be computed using the innovations algorithm. The maximization of the likelihood depends heavily on the quality of initial estimates, some of which are discussed at the end of this section. For numerical methods, such as the *Newton-Raphson* and *scoring algorithms*, see Section 3.6 in Shumway and Stoffer (2006).

The limit distribution of the MLE procedure is given as the following theorem. Its proof can be found in Section 8.8 of Brockwell and Davis (1991).

Theorem 3.5.2 *Let $(X_t\colon t \in \mathbb{Z})$ be a causal and invertible ARMA(p,q) process defined with $(Z_t\colon t \in \mathbb{Z}) \sim \mathrm{IID}(0, \sigma^2)$. Consider the MLE $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})'$ of $\boldsymbol{\beta}^* = (\boldsymbol{\phi}, \boldsymbol{\theta})'$ obtained from minimizing the profile likelihood. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} N(0, \sigma^2 \boldsymbol{\Gamma}_{p,q}^{-1}) \qquad (n \to \infty).$$

The result is optimal. For $p \geq 1$ and $q \geq 1$, the covariance matrix

$$\boldsymbol{\Gamma}_{p,q} = \begin{pmatrix} E[\mathbf{U}_t \mathbf{U}_t'] & E[\mathbf{U}_t \mathbf{V}_t'] \\ E[\mathbf{V}_t \mathbf{U}_t'] & E[\mathbf{V}_t \mathbf{V}_t'] \end{pmatrix}$$

is in block form, $\mathbf{U}_t = (U_t, \ldots, U_{t+1-p})'$, $\mathbf{V}_t = (V_t, \ldots, V_{t+1-q})'$ and $(U_t\colon t \in \mathbb{Z})$ and $(V_t\colon t \in \mathbb{Z})$ are the solutions of the AR equations $\phi(B)U_t = \sigma^{-1}Z_t$ and $\theta(B)V_t = \sigma^{-1}Z_t$, respectively.

Method 3 (Least Squares Estimation) An alternative to the method of moments and the MLE is provided by the least squares estimation (LSE). For causal and invertible ARMA(p,q) processes, it is based on minimizing the weighted sum of squares

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{\ell=1}^{n} \frac{(X_\ell - \hat{X}_\ell)^2}{r_{\ell-1}} \tag{3.5.7}$$

with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Assuming that $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\theta}}$ denote these LSEs, the LSE for σ^2 is computed as

$$\tilde{\sigma}^2 = \frac{S(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\theta}})}{n - p - q}.$$

The least squares procedure has the same asymptotics as the MLE.

Theorem 3.5.3 *The result of Theorem 3.5.2 holds also if $\hat{\boldsymbol{\beta}}$ is replaced with $\tilde{\boldsymbol{\beta}}$.*

Note that the LS estimation has been done for both causal and invertible ARMA processes. If noninvertible processes were admitted, then one can typically not minimize the weighted sum of squares in (3.5.7) for finite values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Initial Estimators. In the remainder of this section, we briefly discuss three ways to obtain initial estimators which can be used to initialize the numerical optimization of both MLE and LSE. They are respectively based on the Durbin-Levinson algorithm, the innovations algorithm and the causal representation of ARMA processes.

(a) Fitting autoregressive processes. Let $(X_t\colon t \in \mathbb{Z})$ be a zero-mean (not necessarily autoregressive) process. Observe $X_1, \ldots, X_n$. If for some $m \geq 1$, $\hat{\gamma}(0) > 0$ (so that $\hat{\Gamma}_m^{-1}$ exists) one can fit the AR(m) process

$$X_t - \hat{\phi}_{m1}X_{t-1} - \ldots - \hat{\phi}_{mm}X_{t-m} = Z_t, \qquad (Z_t) \sim \mathrm{WN}(0, \hat{v}_m),$$

to the data, where $\hat{\boldsymbol{\phi}}_m = \hat{R}_m^{-1}\hat{\boldsymbol{\rho}}_m$ and $\hat{v}_m = \hat{\gamma}(0)(1 - \hat{\boldsymbol{\rho}}_m'\hat{R}_m^{-1}\hat{\boldsymbol{\rho}}_m)$ are calculated with the Durbin-Levinson algorithm for the sample correlations. If an AR model is appropriate for the data, then $\hat{\phi}_{mm} \approx 0$ for m large enough. Since the distribution of $\hat{\phi}_{mm}$ is known, we can select p by requiring that

$$|\hat{\phi}_{mm}| < \frac{1.96}{\sqrt{n}} \qquad \text{for all } m > p.$$

Now, the coefficients for the resulting AR(p) process can be estimated by Yule-Walker.

(b) Fitting moving average processes. In analogy to (a) and starting under the same set of assumptions, one can successively fit moving average models

$$X_t = Z_t + \hat{\theta}_{m1}Z_{t-1} + \ldots + \hat{\theta}_{mm}Z_{t-m}, \qquad (Z_t) \sim \mathrm{WN}(0, \hat{v}_m),$$

where the coefficients $\hat{\theta}_{m\ell}$ and the mean-squared prediction error $\hat{v}_m$ can be calculated from the innovations algorithm. If an MA model is appropriate, then $\hat{\theta}_{m\ell} \to \theta_\ell$ in probability (Homework Problem 7). Even in the general case of causal time series, the limit for $\hat{\theta}_{m\ell}$ can be determined, but even for a causal ARMA(p, q) model, the limit is generally not θ_ℓ (see below). On the practical side this means that one should monitor the estimators $\theta_{m\ell}$ for increasing m and wait until the values stabilize (not necessarily around zero) to select the order q.

(c) Fitting autoregressive moving average processes. If $(X_t\colon t \in \mathbb{Z})$ is a causal ARMA(p, q) process it admits the representation $X_t = \psi(B)Z_t$, $t \in \mathbb{Z}$, for which the coefficients can be determined as $\psi_0 = 1$ and

$$\psi_j = \theta_j + \sum_{\ell=1}^{\min\{j,p\}} \phi_\ell \psi_{j-\ell}, \qquad j \in \mathbb{N}, \tag{3.5.8}$$

with the understanding that $\theta_j = 0$ for $j > q$ and $\phi_\ell = 0$ for $\ell > p$. Under a set of suitable assumptions, $\hat{\theta}_{m\ell} \to \psi_\ell$ in probability. This suggests to compute $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{m,p+q}$ from the

innovations algorithm and then to solve (3.5.8) with ψ_j replaced by $\hat{\theta}_{m\ell}$, that is,

$$\hat{\theta}_{mj} = \theta_j + \sum_{\ell=1}^{\min\{j,p\}} \phi_\ell \hat{\theta}_{m,j-\ell}, \qquad j = 1, \ldots, p+q.$$

In particular, one uses first equations $q+1, \ldots, p$ in the linear system

$$\begin{pmatrix} \hat{\theta}_{m,q+1} \\ \vdots \\ \hat{\theta}_{m,q+p} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_{m,q} & \ldots & \hat{\theta}_{m,q+1-p} \\ \vdots & & \vdots \\ \hat{\theta}_{m,p+q+1} & \ldots & \hat{\theta}_{m,q} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}$$

to obtain $\hat{\phi}_1, \ldots, \hat{\phi}_p$, which may not be causal. Then one computes $\hat{\theta}_j$ from

$$\hat{\theta}_{mj} = \theta_j + \sum_{\ell=1}^{\min\{j,p\}} \hat{\phi}_\ell \hat{\theta}_{m,j-\ell}, \qquad j = 1, \ldots, q.$$

The innovations variance is, as before, given by $\hat{\sigma}^2 = \hat{v}_m$.

3.6 Integrated ARMA Models

In Section 1.3, we have seen how the difference operator $\nabla = 1 - B$ can be applied to generate stationary sequences. A similar concept leads to the class of *autoregressive-integrated moving average processes* which are studied in this section.

Definition 3.6.1 (ARIMA processes) *A process $(X_t : t \in \mathbb{Z})$ is called autoregressive-integrated moving average of order (p, d, q), abbreviated by ARIMA(p, d, q), if it satisfies the difference equations*

$$\phi^*(B)X_t \equiv \phi(B)(1-B)^d X_t = \theta(B)Z_t, \qquad t \in \mathbb{Z}, \tag{3.6.1}$$

where $(Z_t : t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$.

Remark 3.6.1 *Note that $\phi^*(z)$ has a zero of order d at $z = 1$. Consequently, an ARIMA(p, d, q) process is stationary if and only if $d = 0$, in which case it is ARMA in the sense of Definition 3.1.1. For any $d \geq 1$, (3.6.1) will still be satisfied if X_t is replaced with $X_t + \mu_t$, where μ_t is a polynomial of degree $d-1$. Therefore, ARIMA processes can potentially play an important role in representing data with trend.*

The next lemma shows that, for $d \geq 1$, the defining equations (3.6.1) determine only the second order properties of the process $(\nabla^d X_t : t \in \mathbb{Z})$ but not those of $(X_t : t \in \mathbb{Z})$ itself.

Lemma 3.6.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be an ARIMA(p, d, q) process. Then, the difference equations (3.6.1) are also satisfied by the process $(W_t : t \in \mathbb{Z})$ given by*

$$W_t = X_t + A_0 + A_1 t + \ldots + A_{d-1} t^{d-1},$$

where $A_0, \ldots, A_{d-1}$ are arbitrary random variables.

Proof. Clearly $(1-B)^d(W_t - X_t) = 0$ and therefore

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

which completes the proof. □

Example 3.6.1 (ARIMA(0,1,1) processes) Let the process $(X_t\colon t \in \mathbb{Z})$ be given by the ARIMA(0,1,1) equations $(1-B)X_t = \theta(B)Z_t$, where $\theta(z) = 1 - \theta z$ and $|\theta| < 1$. In this case, X_t admits the invertible representation

$$X_t = \sum_{j=1}^{\infty}(1-\theta)\theta^{j-1}X_{t-j} + Z_t.$$

From the foregoing equation it follows immediately that the one-step ahead predictor based on the infinite past is given by

$$\tilde{X}_{n+1} = \sum_{j=1}^{\infty}(1-\theta)\theta^{j-1}X_{n+1-j} = (1-\theta)X_n + \theta\tilde{X}_n,$$

where the latter equality sign follows after some basic algebra. Finally, the truncated predictor is determined by the equations

$$X^*_{n+1} = (1-\theta)X_n + \theta X^*_n, \qquad n \in \mathbb{N}.$$

See Homework Problem 13 for more on ARIMA(0,1,1) processes.

The intuitive approach in a concrete data analysis is to apply the operator ∇ repeatedly to the observations in the hope of finding a d such that $(\nabla^d X_t\colon t \in \mathbb{Z})$ has rapidly decaying ACF and PACF and fitting an ARMA sequence to these differences. Proceeding in this way is oftentimes successful but some care is necessary as the following examples show.

Example 3.6.2 (ARIMA(1,1,0) processes) Figure 3.4 displays in its right panel 200 simulated values of the ARIMA(1,1,0) process $(X_t\colon t \in \mathbb{Z})$ defined by the equations

$$(1-.8B)(1-B)X_t = Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t\colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$. The scale has been adjusted so that $X_0 = 0$. The sample ACF given in the middle panel of Figure 3.4 shows the typical slow decay characteristic for ARIMA processes. The PACF in the right panel of the same figure is essentially zero with the exception of the first two (three) values. The left panel of Figure 3.5 shows the differenced observations which do not display any significant deviations from stationarity. An inspection of the sample ACF and PACF of the differences indicates that fitting an AR(1) process will lead to a satisfactory result. Using LSE for the autoregressive parameter ϕ and the variance σ^2 returns the values $\hat{\phi} = .86$ and $\hat{\sigma}^2 = 1.032$. The fitted model thus becomes

$$(1-.86B)(1-B)X_t = Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t\colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, 1.032)$. This is in proximity of the true model.

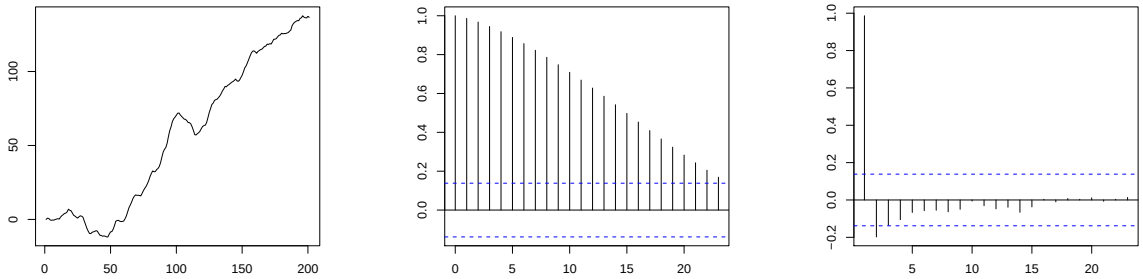

Figure 3.4: The ARIMA(1,1,0) process from Example 3.6.2, its ACF and PACF.

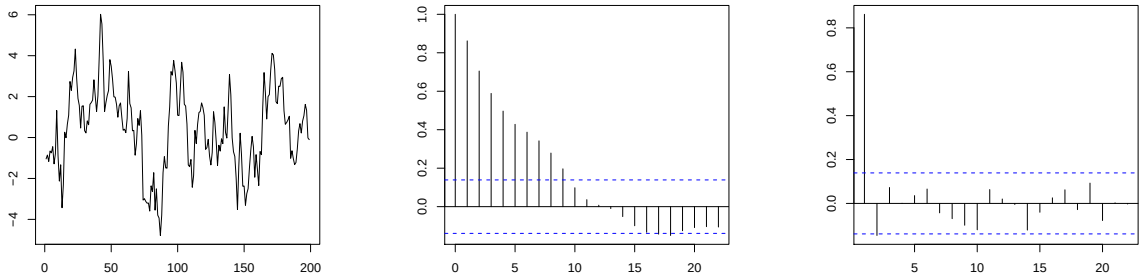

Figure 3.5: The differences of the ARIMA(1,1,0) process from Example 3.6.2, its ACF and PACF.

If, on the other hand, one tries to fit an AR(2) model via LSE directly to the observations, as suggested by the PACF in Figure 3.4, one would estimate the corresponding parameters as

$$\hat{\phi}_1 = 1.8672, \qquad \hat{\phi}_2 = -.8721 \qquad \text{and} \qquad \sigma^2 = 1.011,$$

resulting in the fitted model

$$(1 - 1.8672B + .8721B^2)X_t = Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, 1.011)$. The autoregressive polynomial $\phi(z) = 1 - \hat{\phi}_1 z - \hat{\phi}_2 z^2$ has complex roots close to the unit circle. Figure 3.6 shows the time series plot of 200 observations obtained from a simulated AR(2) process with the given estimated parameters together with the ACF and PACF. Detecting apparent differences between the respective ACFs and PACFs in this figure and Figure 3.4 is rather difficult. Notice that the Yule-Walker estimation would give the estimates $\hat{\phi}_1 = 1.1815$ and $\hat{\phi}_2 = -.1978$ which are nowhere near the actual values. The reason for the deviation lies in the fact that the covariance matrix $\mathbf{\Gamma}_2$ needed to solve the Durbin-Levinson algorithm is almost singular. Estimation results are therefore distorted.

The nonstationarities allowed for in the ARIMA framework are special, since the autoregressive polynomial ϕ^* has a zero of order d in $z = 1$. That is of course only one particular point on the unit circle. The following example shows what can be done if an AR process has roots that are close to the unit circle but not necessarily in the vicinity of $z = 1$.

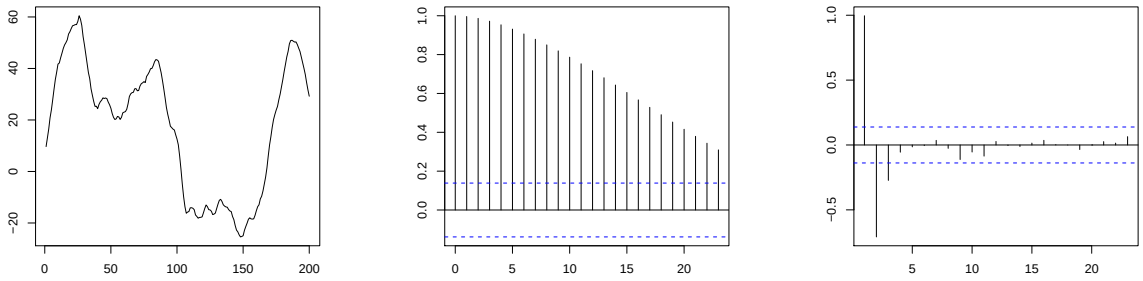

Figure 3.6: The AR(2) process from Example 3.6.2, its ACF and PACF.

Example 3.6.3 The processes discussed in this example show that suitable modifications of the difference operator can be applied in certain cases to improve fitting procedures. Figure 3.7 displays 200 observations of the AR(1) process

$$X_t = -.99X_{t-1} + Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0, \sigma^2)$. The corresponding ACF displays a cyclical behavior and is slowly decaying, while the PACF is essentially zero after lag 1. Taking into account that the cycles have period 2, it is more convenient to apply the operator $1 - B^2$ (instead of ∇), which returns a time series that appears stationary (see Figure 3.7, upper right panel). The reason for this lies in the fact that the process is "close" to the periodic process $X_t^* = -X_{t-1}^* + Z_t^*$ which has a unit root in $z = e^{i\pi} = -1$.

In a similar vein, if one studies the AR(2) process

$$X_t = X_{t-1} - .99X_{t-2} + Z_t, \qquad t \in \mathbb{Z},$$

one can use that it is close to the periodic process $X_t^* = X_{t-1}^* - X_{t-2}^* + Z_t^*$ which has the autoregressive polynomial unit roots $e^{\pm\pi/3}$. Its ACF will consequently display cycles with period 6 and one can recommend the use of $1 - B^6$ to obtain a more stationary process.

3.7 Model Selection

In this section, a rough guide for going about the data analysis will be provided. It consists of several parts, most of which have been discussed previously. The main focus is on the order selection of p and q in the case that these parameters are in fact unknown.

Step 1. Plot the data and check whether or not the variability remains reasonably stable throughout the observation period. If that is not the case, use preliminary transformations to stabilize the variance. One popular class is given by the *Box-Cox transformations* (Box and Cox, 1964)

$$f_\lambda(U_t) = \begin{cases} \lambda^{-1}(U_t^\lambda - 1), & U_t \geq 0,\ \lambda > 0. \\ \ln U_t & U_t > 0,\ \lambda = 0. \end{cases}$$

In practice f_0 or $f_{1/2}$ are often adequate choices. (Recall, for instance, the Australian wine sales data of Example 1.4.1.)

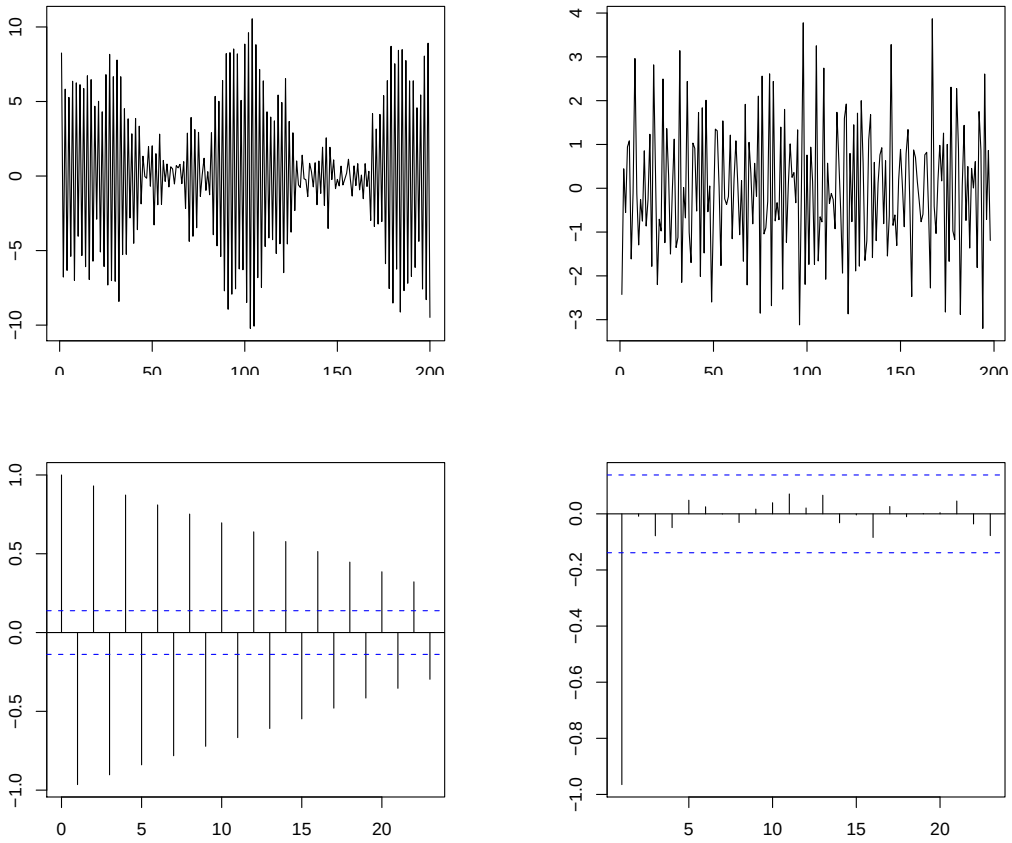

Figure 3.7: The AR(1) process from Example 3.6.3 (upper left), its ACF and PACF (lower panel) and the differences obtained after an application of $1 - B^2$ (upper right).

Step 2. Remove, if present, trend and seasonal components from the data. Chapter 1 introduced a number of tools to do so, based on the classical decomposition of a time series

$$X_t = m_t + s_t + Y_t$$

into a trend, a seasonality and a residual component. Note that differencing works also without the specific representation in the last display. If the data appears stationary, move on to the next step. Else apply, for example, another set of difference operations.

Step 3. Suppose now that Steps 1 and 2 have provided us with observations that are well described by a stationary sequence $(Y_t\colon t \in \mathbb{Z})$. The goal is then to find the most appropriate ARMA(p, q) model to describe the process. In the unlikely case that p and q can be assumed known, utilize the estimation procedures of Section 3.5 directly. Otherwise, choose them according to one of the following criteria.

(a) The standard criterion that is typically implemented in software packages is a modification of *Akaike's information criterion*, see Akaike (1969), which was given by Hurvich and Tsai (1989). In this paper, it is suggested that the ARMA model parameters be chosen that they minimize the objective function

$$\mathrm{AIC}_C(\boldsymbol{\phi}, \boldsymbol{\theta}, p, q) = -2 \ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, S(\boldsymbol{\phi}, \boldsymbol{\theta})/n) + \frac{2(p+q+1)n}{n-p-q-2}. \tag{3.7.1}$$

Here, $L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ denotes the Gaussian likelihood defined in (3.5.6) and $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ is the weighted sum of squares in (3.5.7). It can be seen from the definition that the AIC_C does not attempt to minimize the log-likelihood function directly. The introduction of the penalty term on the right-hand side of (3.7.1) reduces the risk of overfitting.

(b) For pure autoregressive processes, Akaike (1969) introduced a criterion that is based on a minimization of the *final prediction error*. Here, the order p is chosen as the minimizer of the objective function

$$\mathrm{FPE} = \hat{\sigma}^2 \frac{n+p}{n-p},$$

where $\hat{\sigma}^2$ denotes the MLE of the unknown noise variance σ^2. For more on this topic and other procedures that help fit a model, we refer here to Section 9.3 of Brockwell and Davis (1991).

Step 4. The last step in the analysis is concerned with *diagnostic checking* by applying the goodness of fit tests of Section 1.5.

3.8 Summary

The class of autoregressive moving average processes has been introduced to model stationary stochastic processes. We have examined theoretical properties such as causality and invertibility, which depend on the zeroes of the autoregressive and moving average polynomials, respectively.

We have learned how the causal representation of an ARMA process can be utilized to compute its covariance function which contains all information about the dependence structure.

Assuming known parameter values, several forecasting procedures have been discussed. The Durbin-Levinson algorithm works well for pure AR processes, while the innovations algorithm is particularly useful for pure MA processes. Predictions using an infinite past work well for causal and invertible ARMA processes. For practical purposes, however, a truncated version is more relevant.

Since the exact parameter values are in general unknown, we have introduced various estimation procedures. The Yule-Walker procedure is only optimal in the AR case but provides useful initial estimates that can be used for the numerical derivation of maximum likelihood or least squares estimates.

We have shown how integrated ARIMA processes can be used to describe nonstationary processes. The nonstationary behavior is of a particular kind, assuming a multiple root of the autoregressive polynomial at 1.

Finally, we have provided a framework that may potentially be useful when facing the problem of analyzing a data set in practice.