3

ARIMA Models

3.1 Introduction

In Chapters 1 and 2, we introduced autocorrelation and cross-correlation functions (ACFs and CCFs) as tools for clarifying relations that may occur within and between time series at various lags. In addition, we explained how to build linear models based on classical regression theory for exploiting the associations indicated by large values of the ACF or CCF. The time domain, or regression, methods of this chapter are appropriate when we are dealing with possibly nonstationary, shorter time series; these series are the rule rather than the exception in many applications. In addition, if the emphasis is on forecasting future values, then the problem is easily treated as a regression problem. This chapter develops a number of regression techniques for time series that are all related to classical ordinary and weighted or correlated least squares.

Classical regression is often insufficient for explaining all of the interesting dynamics of a time series. For example, the ACF of the residuals of the simple linear regression fit to the global temperature data (see Example 2.4 of Chapter 2) reveals additional structure in the data that the regression did not capture. Instead, the introduction of correlation as a phenomenon that may be generated through lagged linear relations leads to proposing the autoregressive (AR) and autoregressive moving average (ARMA) models. Adding nonstationary models to the mix leads to the autoregressive integrated moving average (ARIMA) model popularized in the landmark work by Box and Jenkins (1970). The Box–Jenkins method for identifying a plausible ARIMA model is given in this chapter along with techniques for parameter estimation and forecasting for these models. A partial theoretical justification of the use of ARMA models is discussed in Appendix B, §B.4.

3.2 Autoregressive Moving Average Models

The classical regression model of Chapter 2 was developed for the static case, namely, we only allow the dependent variable to be influenced by current values of the independent variables. In the time series case, it is desirable to allow the dependent variable to be influenced by the past values of the independent variables and possibly by its own past values. If the present can be plausibly modeled in terms of only the past values of the independent inputs, we have the enticing prospect that forecasting will be possible.

Introduction to Autoregressive Models

Autoregressive models are based on the idea that the current value of the series, x_t, can be explained as a function of p past values, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value. As a typical case, recall Example 1.10 in which data were generated using the model

$$x_t = x_{t-1} - .90x_{t-2} + w_t,$$

where w_t is white Gaussian noise with $\sigma_w^2 = 1$. We have now assumed the current value is a particular *linear* function of past values. The regularity that persists in Figure 1.9 gives an indication that forecasting for such a model might be a distinct possibility, say, through some version such as

$$x_{n+1}^n = x_n - .90x_{n-1},$$

where the quantity on the left-hand side denotes the forecast at the next period $n+1$ based on the observed data, $x_1, x_2, \ldots, x_n$. We will make this notion more precise in our discussion of forecasting (§3.5).

The extent to which it might be possible to forecast a real data series from its own past values can be assessed by looking at the autocorrelation function and the lagged scatterplot matrices discussed in Chapter 2. For example, the lagged scatterplot matrix for the Southern Oscillation Index (SOI), shown in Figure 2.7, gives a distinct indication that lags 1 and 2, for example, are linearly associated with the current value. The ACF shown in Figure 1.14 shows relatively large positive values at lags 1, 2, 12, 24, and 36 and large negative values at 18, 30, and 42. We note also the possible relation between the SOI and Recruitment series indicated in the scatterplot matrix shown in Figure 2.8. We will indicate in later sections on transfer function and vector AR modeling how to handle the dependence on values taken by other series.

The preceding discussion motivates the following definition.

Definition 3.1 *An* **autoregressive model** *of order p, abbreviated* **AR(p)**, *is of the form*

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{3.1}$$

where x_t is stationary, and $\phi_1, \phi_2, \ldots, \phi_p$ are constants ($\phi_p \neq 0$). Although it is not necessary yet, we assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2, unless otherwise stated. The mean of x_t in (3.1) is zero. If the mean, μ, of x_t is not zero, replace x_t by $x_t - \mu$ in (3.1),

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t,$$

or write

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{3.2}$$

where $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

We note that (3.2) is similar to the regression model of §2.2, and hence the term auto (or self) regression. Some technical difficulties, however, develop from applying that model because the regressors, $x_{t-1}, \ldots, x_{t-p}$, are random components, whereas $\boldsymbol{z}_t$ was assumed to be fixed. A useful form follows by using the backshift operator (2.33) to write the AR(p) model, (3.1), as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_t = w_t, \tag{3.3}$$

or even more concisely as

$$\phi(B) x_t = w_t. \tag{3.4}$$

The properties of $\phi(B)$ are important in solving (3.4) for x_t. This leads to the following definition.

Definition 3.2 *The* **autoregressive operator** *is defined to be*

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p. \tag{3.5}$$

We initiate the investigation of AR models by considering the first-order model, AR(1), given by $x_t = \phi x_{t-1} + w_t$. Iterating backwards k times, we get

$$\begin{aligned} x_t &= \phi x_{t-1} + w_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t \\ &= \phi^2 x_{t-2} + \phi w_{t-1} + w_t \\ &\vdots \\ &= \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j}. \end{aligned}$$

This method suggests that, by continuing to iterate backward, and provided that $|\phi| < 1$ and x_t is stationary, we can represent an AR(1) model as a linear process given by[1]

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{3.6}$$

[1] Note that $\lim_{k\to\infty} E\left(x_t - \sum_{j=0}^{k-1} \phi^j w_{t-j}\right)^2 = \lim_{k\to\infty} \phi^{2k} E\left(x_{t-k}^2\right) = 0$, so (3.6) exists in the mean square sense (see Appendix A for a definition).

The AR(1) process defined by (3.6) is stationary with mean

$$E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0,$$

and autocovariance function,

$$\begin{aligned}\gamma(h) &= \mathrm{cov}(x_{t+h}, x_t) = E\left[\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j}\right)\left(\sum_{k=0}^{\infty} \phi^k w_{t-k}\right)\right] \\ &= E\left[\left(w_{t+h} + \cdots + \phi^h w_t + \phi^{h+1} w_{t-1} + \cdots\right)\left(w_t + \phi w_{t-1} + \cdots\right)\right] \\ &= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j}\phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1-\phi^2}, \quad h \geq 0.\end{aligned} \tag{3.7}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only exhibit the autocovariance function for $h \geq 0$. From (3.7), the ACF of an AR(1) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \quad h \geq 0, \tag{3.8}$$

and $\rho(h)$ satisfies the recursion

$$\rho(h) = \phi\,\rho(h-1), \quad h = 1, 2, \ldots . \tag{3.9}$$

We will discuss the ACF of a general AR(p) model in §3.4.

Example 3.1 The Sample Path of an AR(1) Process

Figure 3.1 shows a time plot of two AR(1) processes, one with $\phi = .9$ and one with $\phi = -.9$; in both cases, $\sigma_w^2 = 1$. In the first case, $\rho(h) = .9^h$, for $h \geq 0$, so observations close together in time are positively correlated with each other. This result means that observations at contiguous time points will tend to be close in value to each other; this fact shows up in the top of Figure 3.1 as a very smooth sample path for x_t. Now, contrast this with the case in which $\phi = -.9$, so that $\rho(h) = (-.9)^h$, for $h \geq 0$. This result means that observations at contiguous time points are negatively correlated but observations two time points apart are positively correlated. This fact shows up in the bottom of Figure 3.1, where, for example, if an observation, x_t, is positive, the next observation, x_{t+1}, is typically negative, and the next observation, x_{t+2}, is typically positive. Thus, in this case, the sample path is very choppy.

The following R code can be used to obtain a figure similar to Figure 3.1:

```
par(mfrow=c(2,1))
plot(arima.sim(list(order=c(1,0,0), ar=.9), n=100), ylab="x",
     main=(expression(AR(1)~~~phi==+.9)))
plot(arima.sim(list(order=c(1,0,0), ar=-.9), n=100), ylab="x",
     main=(expression(AR(1)~~~phi==-.9)))
```

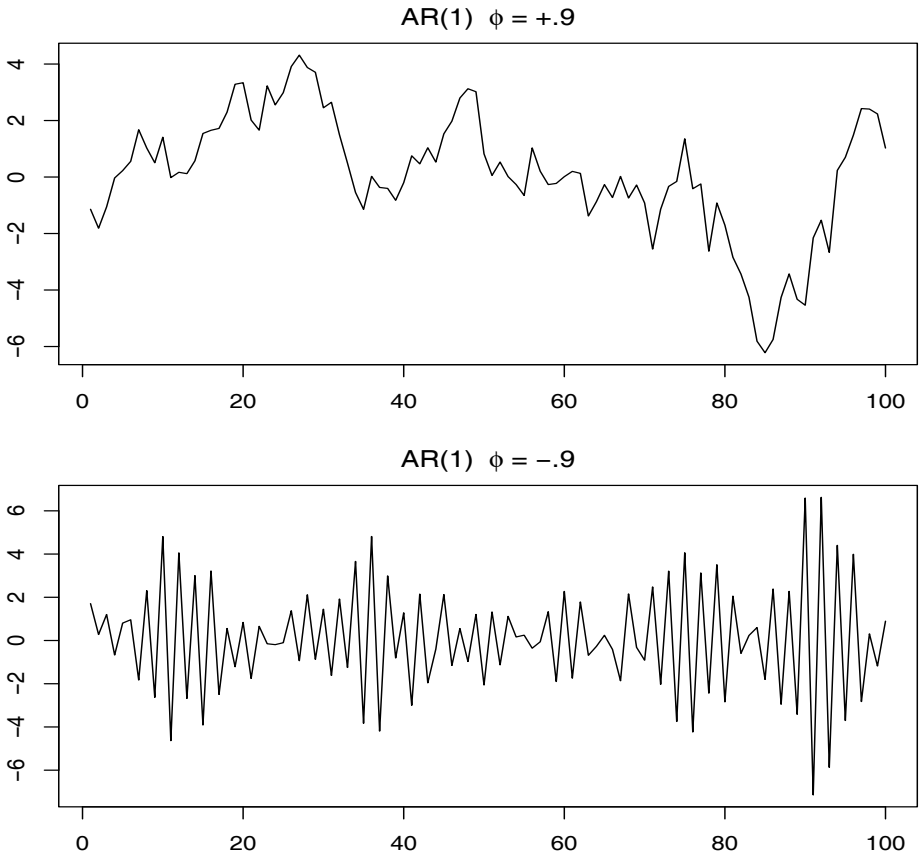

Fig. 3.1. Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).

Example 3.2 Explosive AR Models and Causality

In Example 1.18, it was discovered that the random walk $x_t = x_{t-1} + w_t$ is not stationary. We might wonder whether there is a stationary AR(1) process with $|\phi| > 1$. Such processes are called explosive because the values of the time series quickly become large in magnitude. Clearly, because $|\phi|^j$ increases without bound as $j \to \infty$, $\sum_{j=0}^{k-1} \phi^j w_{t-j}$ will not converge (in mean square) as $k \to \infty$, so the intuition used to get (3.6) will not work directly. We can, however, modify that argument to obtain a stationary model as follows. Write $x_{t+1} = \phi x_t + w_{t+1}$, in which case,

$$\begin{aligned} x_t &= \phi^{-1} x_{t+1} - \phi^{-1} w_{t+1} = \phi^{-1}\left(\phi^{-1} x_{t+2} - \phi^{-1} w_{t+2}\right) - \phi^{-1} w_{t+1} \\ &\;\;\vdots \\ &= \phi^{-k} x_{t+k} - \sum_{j=1}^{k-1} \phi^{-j} w_{t+j}, \end{aligned} \tag{3.10}$$

by iterating forward k steps. Because $|\phi|^{-1} < 1$, this result suggests the stationary future dependent AR(1) model

$$x_t = -\sum_{j=1}^{\infty} \phi^{-j} w_{t+j}. \tag{3.11}$$

The reader can verify that this is stationary and of the AR(1) form $x_t = \phi x_{t-1} + w_t$. Unfortunately, this model is useless because it requires us to know the future to be able to predict the future. When a process does not depend on the future, such as the AR(1) when $|\phi| < 1$, we will say the process is causal. In the explosive case of this example, the process is stationary, but it is also future dependent, and not causal.

Example 3.3 Every Explosion Has a Cause

Excluding explosive models from consideration is not a problem because the models have causal counterparts. For example, if

$$x_t = \phi x_{t-1} + w_t \quad \text{with} \quad |\phi| > 1$$

and $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$, then using (3.11), $\{x_t\}$ is a non-causal stationary Gaussian process with $E(x_t) = 0$ and

$$\begin{aligned}\gamma_x(h) = \mathrm{cov}(x_{t+h}, x_t) &= \mathrm{cov}\left(-\sum_{j=1}^{\infty} \phi^{-j} w_{t+h+j}, -\sum_{k=1}^{\infty} \phi^{-k} w_{t+k}\right) \\ &= \sigma_w^2 \phi^{-2}\, \phi^{-h}/(1-\phi^{-2}).\end{aligned}$$

Thus, using (3.7), the causal process defined by

$$y_t = \phi^{-1} y_{t-1} + v_t$$

where $v_t \sim$ iid $\mathrm{N}(0, \sigma_w^2 \phi^{-2})$ is stochastically equal to the x_t process (i.e., all finite distributions of the processes are the same). For example, if $x_t = 2x_{t-1} + w_t$ with $\sigma_w^2 = 1$, then $y_t = \frac{1}{2} y_{t-1} + v_t$ with $\sigma_v^2 = 1/4$ is an equivalent causal process (see Problem 3.3). This concept generalizes to higher orders, but it is easier to show using Chapter 4 techniques; see Example 4.7.

The technique of iterating backward to get an idea of the stationary solution of AR models works well when $p = 1$, but not for larger orders. A general technique is that of matching coefficients. Consider the AR(1) model in operator form

$$\phi(B) x_t = w_t, \tag{3.12}$$

where $\phi(B) = 1 - \phi B$, and $|\phi| < 1$. Also, write the model in equation (3.6) using operator form as

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B) w_t, \tag{3.13}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\psi_j = \phi^j$. Suppose we did not know that $\psi_j = \phi^j$. We could substitute $\psi(B)w_t$ from (3.13) for x_t in (3.12) to obtain

$$\phi(B)\psi(B)w_t = w_t. \tag{3.14}$$

The coefficients of B on the left-hand side of (3.14) must be equal to those on right-hand side of (3.14), which means

$$(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_j B^j + \cdots) = 1. \tag{3.15}$$

Reorganizing the coefficients in (3.15),

$$1 + (\psi_1 - \phi)B + (\psi_2 - \psi_1\phi)B^2 + \cdots + (\psi_j - \psi_{j-1}\phi)B^j + \cdots = 1,$$

we see that for each $j = 1, 2, \ldots$, the coefficient of B^j on the left must be zero because it is zero on the right. The coefficient of B on the left is $(\psi_1 - \phi)$, and equating this to zero, $\psi_1 - \phi = 0$, leads to $\psi_1 = \phi$. Continuing, the coefficient of B^2 is $(\psi_2 - \psi_1\phi)$, so $\psi_2 = \phi^2$. In general,

$$\psi_j = \psi_{j-1}\phi,$$

with $\psi_0 = 1$, which leads to the solution $\psi_j = \phi^j$.

Another way to think about the operations we just performed is to consider the AR(1) model in operator form, $\phi(B)x_t = w_t$. Now multiply both sides by $\phi^{-1}(B)$ (assuming the inverse operator exists) to get

$$\phi^{-1}(B)\phi(B)x_t = \phi^{-1}(B)w_t,$$

or

$$x_t = \phi^{-1}(B)w_t.$$

We know already that

$$\phi^{-1}(B) = 1 + \phi B + \phi^2 B^2 + \cdots + \phi^j B^j + \cdots,$$

that is, $\phi^{-1}(B)$ is $\psi(B)$ in (3.13). Thus, we notice that working with operators is like working with polynomials. That is, consider the polynomial $\phi(z) = 1 - \phi z$, where z is a complex number and $|\phi| < 1$. Then,

$$\phi^{-1}(z) = \frac{1}{(1 - \phi z)} = 1 + \phi z + \phi^2 z^2 + \cdots + \phi^j z^j + \cdots, \quad |z| \leq 1,$$

and the coefficients of B^j in $\phi^{-1}(B)$ are the same as the coefficients of z^j in $\phi^{-1}(z)$. In other words, we may treat the backshift operator, B, as a complex number, z. These results will be generalized in our discussion of ARMA models. We will find the polynomials corresponding to the operators useful in exploring the general properties of ARMA models.

Introduction to Moving Average Models

As an alternative to the autoregressive representation in which the x_t on the left-hand side of the equation are assumed to be combined linearly, the moving average model of order q, abbreviated as MA(q), assumes the white noise w_t on the right-hand side of the defining equation are combined linearly to form the observed data.

Definition 3.3 *The* **moving average model** *of order q, or* **MA(q)** *model, is defined to be*

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{3.16}$$

where there are q lags in the moving average and $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q \neq 0$) are parameters.[2] *Although it is not necessary yet, we assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2, unless otherwise stated.*

The system is the same as the infinite moving average defined as the linear process (3.13), where $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \ldots, q$, and $\psi_j = 0$ for other values. We may also write the MA(q) process in the equivalent form

$$x_t = \theta(B) w_t, \tag{3.17}$$

using the following definition.

Definition 3.4 *The* **moving average operator** *is*

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \tag{3.18}$$

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$; details of this result are provided in §3.4.

Example 3.4 The MA(1) Process

Consider the MA(1) model $x_t = w_t + \theta w_{t-1}$. Then, $E(x_t) = 0$,

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0, \\ \theta\sigma_w^2 & h = 1, \\ 0 & h > 1, \end{cases}$$

and the ACF is

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

Note $|\rho(1)| \leq 1/2$ for all values of θ (Problem 3.1). Also, x_t is correlated with x_{t-1}, but not with $x_{t-2}, x_{t-3}, \ldots$. Contrast this with the case of the AR(1)

[2] Some texts and software packages write the MA model with negative coefficients; that is, $x_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \cdots - \theta_q w_{t-q}$.

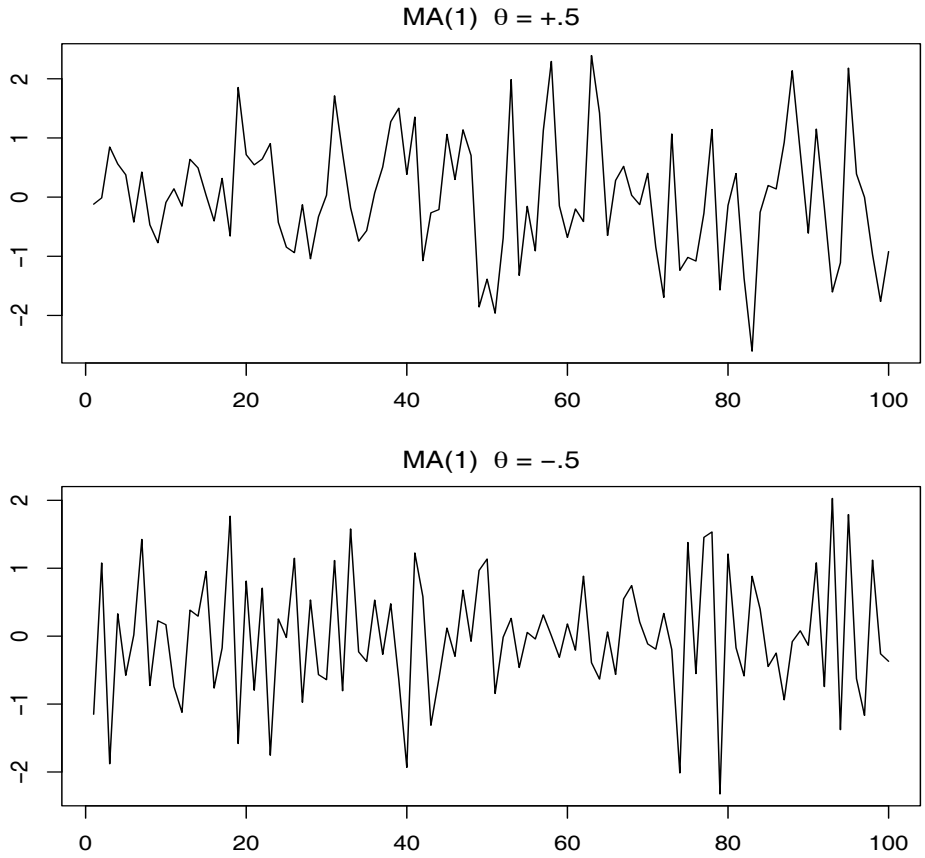

Fig. 3.2. Simulated MA(1) models: $\theta = .5$ (top); $\theta = -.5$ (bottom).

model in which the correlation between x_t and x_{t-k} is never zero. When $\theta = .5$, for example, x_t and x_{t-1} are positively correlated, and $\rho(1) = .4$. When $\theta = -.5$, x_t and x_{t-1} are negatively correlated, $\rho(1) = -.4$. Figure 3.2 shows a time plot of these two processes with $\sigma_w^2 = 1$. The series in where $\theta = .5$ is smoother than the series where $\theta = -.5$.

A figure similar to Figure 3.2 can be created in R as follows:

```
par(mfrow = c(2,1))
plot(arima.sim(list(order=c(0,0,1), ma=.5), n=100), ylab="x",
    main=(expression(MA(1)~~~theta==+.5)))
plot(arima.sim(list(order=c(0,0,1), ma=-.5), n=100), ylab="x",
    main=(expression(MA(1)~~~theta==-.5)))
```

Example 3.5 Non-uniqueness of MA Models and Invertibility

Using Example 3.4, we note that for an MA(1) model, $\rho(h)$ is the same for θ and $\frac{1}{\theta}$; try 5 and $\frac{1}{5}$, for example. In addition, the pair $\sigma_w^2 = 1$ and $\theta = 5$ yield the same autocovariance function as the pair $\sigma_w^2 = 25$ and $\theta = 1/5$, namely,

$$\gamma(h) = \begin{cases} 26 & h = 0, \\ 5 & h = 1, \\ 0 & h > 1. \end{cases}$$

Thus, the MA(1) processes

$$x_t = w_t + \tfrac{1}{5} w_{t-1}, \quad w_t \sim \text{iid } \text{N}(0, 25)$$

and

$$y_t = v_t + 5v_{t-1}, \quad v_t \sim \text{iid } \text{N}(0, 1)$$

are the same because of normality (i.e., all finite distributions are the same). We can only observe the time series, x_t or y_t, and not the noise, w_t or v_t, so we cannot distinguish between the models. Hence, we will have to choose only one of them. For convenience, by mimicking the criterion of causality for AR models, we will choose the model with an infinite AR representation. Such a process is called an invertible process.

To discover which model is the invertible model, we can reverse the roles of x_t and w_t (because we are mimicking the AR case) and write the MA(1) model as $w_t = -\theta w_{t-1} + x_t$. Following the steps that led to (3.6), if $|\theta| < 1$, then $w_t = \sum_{j=0}^{\infty} (-\theta)^j x_{t-j}$, which is the desired infinite AR representation of the model. Hence, given a choice, we will choose the model with $\sigma_w^2 = 25$ and $\theta = 1/5$ because it is invertible.

As in the AR case, the polynomial, $\theta(z)$, corresponding to the moving average operators, $\theta(B)$, will be useful in exploring general properties of MA processes. For example, following the steps of equations (3.12)–(3.15), we can write the MA(1) model as $x_t = \theta(B)w_t$, where $\theta(B) = 1 + \theta B$. If $|\theta| < 1$, then we can write the model as $\pi(B)x_t = w_t$, where $\pi(B) = \theta^{-1}(B)$. Let $\theta(z) = 1 + \theta z$, for $|z| \le 1$, then $\pi(z) = \theta^{-1}(z) = 1/(1 + \theta z) = \sum_{j=0}^{\infty} (-\theta)^j z^j$, and we determine that $\pi(B) = \sum_{j=0}^{\infty} (-\theta)^j B^j$.

Autoregressive Moving Average Models

We now proceed with the general development of autoregressive, moving average, and mixed autoregressive moving average (ARMA), models for stationary time series.

Definition 3.5 *A time series* $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ *is* **ARMA**(p, q) *if it is stationary and*

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \tag{3.19}$$

with $\phi_p \neq 0$, $\theta_q \neq 0$, *and* $\sigma_w^2 > 0$. *The parameters* p *and* q *are called the autoregressive and the moving average orders, respectively. If* x_t *has a nonzero mean* μ, *we set* $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ *and write the model as*

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}. \tag{3.20}$$

Although it is not necessary yet, we assume that w_t is a Gaussian white noise series with mean zero and variance σ_w^2, unless otherwise stated.

As previously noted, when $q = 0$, the model is called an autoregressive model of order p, AR(p), and when $p = 0$, the model is called a moving average model of order q, MA(q). To aid in the investigation of ARMA models, it will be useful to write them using the AR operator, (3.5), and the MA operator, (3.18). In particular, the ARMA(p, q) model in (3.19) can then be written in concise form as

$$\phi(B)x_t = \theta(B)w_t. \tag{3.21}$$

Before we discuss the conditions under which (3.19) is causal and invertible, we point out a potential problem with the ARMA model.

Example 3.6 Parameter Redundancy

Consider a white noise process $x_t = w_t$. Equivalently, we can write this as $.5x_{t-1} = .5w_{t-1}$ by shifting back one unit of time and multiplying by .5. Now, subtract the two representations to obtain

$$x_t - .5x_{t-1} = w_t - .5w_{t-1},$$

or

$$x_t = .5x_{t-1} - .5w_{t-1} + w_t, \tag{3.22}$$

which looks like an ARMA(1, 1) model. Of course, x_t is still white noise; nothing has changed in this regard [i.e., $x_t = w_t$ is the solution to (3.22)], but we have hidden the fact that x_t is white noise because of the parameter redundancy or over-parameterization. Write the parameter redundant model in operator form as $\phi(B)x_t = \theta(B)w_t$, or

$$(1 - .5B)x_t = (1 - .5B)w_t.$$

Apply the operator $\phi(B)^{-1} = (1 - .5B)^{-1}$ to both sides to obtain

$$x_t = (1 - .5B)^{-1}(1 - .5B)x_t = (1 - .5B)^{-1}(1 - .5B)w_t = w_t,$$

which is the original model. We can easily detect the problem of over-parameterization with the use of the operators or their associated polynomials. That is, write the AR polynomial $\phi(z) = (1 - .5z)$, the MA polynomial $\theta(z) = (1 - .5z)$, and note that both polynomials have a common factor, namely $(1 - .5z)$. This common factor immediately identifies the parameter redundancy. Discarding the common factor in each leaves $\phi(z) = 1$ and $\theta(z) = 1$, from which we conclude $\phi(B) = 1$ and $\theta(B) = 1$, and we deduce that the model is actually white noise. The consideration of parameter redundancy will be crucial when we discuss estimation for general ARMA models. As this example points out, we might fit an ARMA(1, 1) model to white noise data and find that the parameter estimates are significant. If we were unaware of parameter redundancy, we might claim the data are correlated when in fact they are not (Problem 3.20).

Example 3.2, Example 3.5, and Example 3.6 point to a number of problems with the general definition of ARMA(p, q) models, as given by (3.19), or, equivalently, by (3.21). To summarize, we have seen the following problems:

(i) parameter redundant models,
(ii) stationary AR models that depend on the future, and
(iii) MA models that are not unique.

To overcome these problems, we will require some additional restrictions on the model parameters. First, we make the following definitions.

Definition 3.6 *The* **AR and MA polynomials** *are defined as*

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \phi_p \neq 0, \tag{3.23}$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \quad \theta_q \neq 0, \tag{3.24}$$

respectively, where z is a complex number.

To address the first problem, we will henceforth refer to an ARMA(p, q) model to mean that it is in its simplest form. That is, in addition to the original definition given in equation (3.19), we will also require that $\phi(z)$ and $\theta(z)$ have no common factors. So, the process, $x_t = .5x_{t-1} - .5w_{t-1} + w_t$, discussed in Example 3.6 is not referred to as an ARMA$(1, 1)$ process because, in its reduced form, x_t is white noise.

To address the problem of future-dependent models, we formally introduce the concept of causality.

Definition 3.7 *An ARMA(p, q) model is said to be* **causal**, *if the time series $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ can be written as a one-sided linear process:*

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B) w_t, \tag{3.25}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$, and $\sum_{j=0}^{\infty} |\psi_j| < \infty$; we set $\psi_0 = 1$.

In Example 3.2, the AR(1) process, $x_t = \phi x_{t-1} + w_t$, is causal only when $|\phi| < 1$. Equivalently, the process is causal only when the root of $\phi(z) = 1 - \phi z$ is bigger than one in absolute value. That is, the root, say, z_0, of $\phi(z)$ is $z_0 = 1/\phi$ (because $\phi(z_0) = 0$) and $|z_0| > 1$ because $|\phi| < 1$. In general, we have the following property.

Property 3.1 Causality of an ARMA(p, q) Process

An ARMA(p, q) model is causal if and only if $\phi(z) \neq 0$ for $|z| \leq 1$. The coefficients of the linear process given in (3.25) can be determined by solving

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

Another way to phrase Property 3.1 is that *an ARMA process is causal only when the roots of $\phi(z)$ lie outside the unit circle*; that is, $\phi(z) = 0$ only when $|z| > 1$. Finally, to address the problem of uniqueness discussed in Example 3.5, we choose the model that allows an infinite autoregressive representation.

Definition 3.8 *An ARMA(p, q) model is said to be* **invertible**, *if the time series $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ can be written as*

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t, \tag{3.26}$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$, and $\sum_{j=0}^{\infty} |\pi_j| < \infty$; we set $\pi_0 = 1$.

Analogous to Property 3.1, we have the following property.

Property 3.2 Invertibility of an ARMA(p, q) Process

An ARMA(p, q) model is invertible if and only if $\theta(z) \neq 0$ for $|z| \leq 1$. The coefficients π_j of $\pi(B)$ given in (3.26) can be determined by solving

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

Another way to phrase Property 3.2 is that *an ARMA process is invertible only when the roots of $\theta(z)$ lie outside the unit circle*; that is, $\theta(z) = 0$ only when $|z| > 1$. The proof of Property 3.1 is given in Appendix B (the proof of Property 3.2 is similar and, hence, is not provided). The following examples illustrate these concepts.

Example 3.7 Parameter Redundancy, Causality, Invertibility

Consider the process

$$x_t = .4x_{t-1} + .45x_{t-2} + w_t + w_{t-1} + .25w_{t-2},$$

or, in operator form,

$$(1 - .4B - .45B^2)x_t = (1 + B + .25B^2)w_t.$$

At first, x_t appears to be an ARMA$(2, 2)$ process. But, the associated polynomials

$$\phi(z) = 1 - .4z - .45z^2 = (1 + .5z)(1 - .9z)$$
$$\theta(z) = (1 + z + .25z^2) = (1 + .5z)^2$$

have a common factor that can be canceled. After cancellation, the polynomials become $\phi(z) = (1 - .9z)$ and $\theta(z) = (1 + .5z)$, so the model is an ARMA$(1, 1)$ model, $(1 - .9B)x_t = (1 + .5B)w_t$, or

$$x_t = .9x_{t-1} + .5w_{t-1} + w_t. \tag{3.27}$$

The model is causal because $\phi(z) = (1 - .9z) = 0$ when $z = 10/9$, which is outside the unit circle. The model is also invertible because the root of $\theta(z) = (1 + .5z)$ is $z = -2$, which is outside the unit circle.

To write the model as a linear process, we can obtain the ψ-weights using Property 3.1, $\phi(z)\psi(z) = \theta(z)$, or

$$(1 - .9z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots) = (1 + .5z).$$

Matching coefficients we get $\psi_0 = 1$, $\psi_1 = .5 + .9 = 1.4$, and $\psi_j = .9\psi_{j-1}$ for $j > 1$. Thus, $\psi_j = 1.4(.9)^{j-1}$ for $j \geq 1$ and (3.27) can be written as

$$x_t = w_t + 1.4 \sum_{j=1}^{\infty} .9^{j-1} w_{t-j}.$$

Similarly, the invertible representation using Property 3.2 is

$$x_t = 1.4 \sum_{j=1}^{\infty} (-.5)^{j-1} x_{t-j} + w_t.$$

Example 3.8 Causal Conditions for an AR(2) Process

For an AR(1) model, $(1-\phi B)x_t = w_t$, to be causal, the root of $\phi(z) = 1-\phi z$ must lie outside of the unit circle. In this case, the root (or zero) occurs at $z_0 = 1/\phi$ [i.e., $\phi(z_0) = 0$], so it is easy to go from the causal requirement on the root, $|1/\phi| > 1$, to a requirement on the parameter, $|\phi| < 1$. It is not so easy to establish this relationship for higher order models.

For example, the AR(2) model, $(1 - \phi_1 B - \phi_2 B^2)x_t = w_t$, is causal when the two roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ lie outside of the unit circle. Using the quadratic formula, this requirement can be written as

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

The roots of $\phi(z)$ may be real and distinct, real and equal, or a complex conjugate pair. If we denote those roots by z_1 and z_2, we can write $\phi(z) = (1-z_1^{-1}z)(1-z_2^{-1}z)$; note that $\phi(z_1) = \phi(z_2) = 0$. The model can be written in operator form as $(1-z_1^{-1}B)(1-z_2^{-1}B)x_t = w_t$. From this representation, it follows that $\phi_1 = (z_1^{-1} + z_2^{-1})$ and $\phi_2 = -(z_1 z_2)^{-1}$. This relationship and the fact that $|z_1| > 1$ and $|z_2| > 1$ can be used to establish the following equivalent condition for causality:

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \text{and} \quad |\phi_2| < 1. \tag{3.28}$$

This causality condition specifies a triangular region in the parameter space; see Figure 3.3 We leave the details of the equivalence to the reader (Problem 3.5).

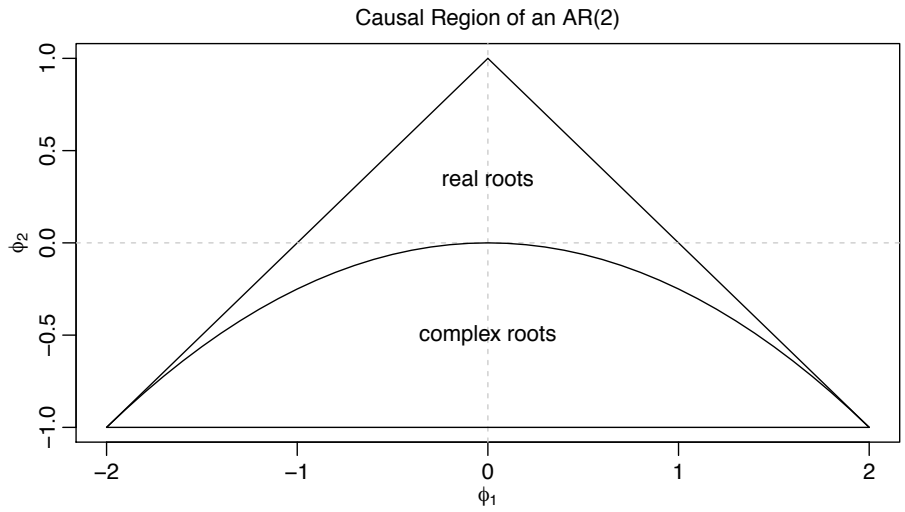

Fig. 3.3. Causal region for an AR(2) in terms of the parameters.

3.3 Difference Equations

The study of the behavior of ARMA processes and their ACFs is greatly enhanced by a basic knowledge of difference equations, simply because they are difference equations. This topic is also useful in the study of time domain models and stochastic processes in general. We will give a brief and heuristic account of the topic along with some examples of the usefulness of the theory. For details, the reader is referred to Mickens (1990).

Suppose we have a sequence of numbers $u_0, u_1, u_2, \ldots$ such that

$$u_n - \alpha u_{n-1} = 0, \quad \alpha \neq 0, \quad n = 1, 2, \ldots \ . \tag{3.29}$$

For example, recall (3.9) in which we showed that the ACF of an AR(1) process is a sequence, $\rho(h)$, satisfying

$$\rho(h) - \phi\rho(h-1) = 0, \quad h = 1, 2, \ldots \ .$$

Equation (3.29) represents a homogeneous difference equation of order 1. To solve the equation, we write:

$$\begin{aligned} u_1 &= \alpha u_0 \\ u_2 &= \alpha u_1 = \alpha^2 u_0 \\ &\vdots \\ u_n &= \alpha u_{n-1} = \alpha^n u_0. \end{aligned}$$

Given an initial condition $u_0 = c$, we may solve (3.29), namely, $u_n = \alpha^n c$.

In operator notation, (3.29) can be written as $(1 - \alpha B)u_n = 0$. The polynomial associated with (3.29) is $\alpha(z) = 1 - \alpha z$, and the root, say, z_0, of this

polynomial is $z_0 = 1/\alpha$; that is $\alpha(z_0) = 0$. We know a solution (in fact, *the* solution) to (3.29), with initial condition $u_0 = c$, is

$$u_n = \alpha^n c = \left(z_0^{-1}\right)^n c. \tag{3.30}$$

That is, the solution to the difference equation (3.29) depends only on the initial condition and the inverse of the root to the associated polynomial $\alpha(z)$.

Now suppose that the sequence satisfies

$$u_n - \alpha_1 u_{n-1} - \alpha_2 u_{n-2} = 0, \quad \alpha_2 \neq 0, \quad n = 2, 3, \ldots \tag{3.31}$$

This equation is a homogeneous difference equation of order 2. The corresponding polynomial is

$$\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2,$$

which has two roots, say, z_1 and z_2; that is, $\alpha(z_1) = \alpha(z_2) = 0$. We will consider two cases. First suppose $z_1 \neq z_2$. Then the general solution to (3.31) is

$$u_n = c_1 z_1^{-n} + c_2 z_2^{-n}, \tag{3.32}$$

where c_1 and c_2 depend on the initial conditions. The claim that is *a* solution can be verified by direct substitution of (3.32) into (3.31):

$$\begin{aligned}
&\left(c_1 z_1^{-n} + c_2 z_2^{-n}\right) - \alpha_1\left(c_1 z_1^{-(n-1)} + c_2 z_2^{-(n-1)}\right) - \alpha_2\left(c_1 z_1^{-(n-2)} + c_2 z_2^{-(n-2)}\right) \\
&\quad = c_1 z_1^{-n}\left(1 - \alpha_1 z_1 - \alpha_2 z_1^2\right) + c_2 z_2^{-n}\left(1 - \alpha_1 z_2 - \alpha_2 z_2^2\right) \\
&\quad = c_1 z_1^{-n}\alpha(z_1) + c_2 z_2^{-n}\alpha(z_2) = 0.
\end{aligned}$$

Given two initial conditions u_0 and u_1, we may solve for c_1 and c_2:

$$u_0 = c_1 + c_2 \quad \text{and} \quad u_1 = c_1 z_1^{-1} + c_2 z_2^{-1},$$

where z_1 and z_2 can be solved for in terms of α_1 and α_2 using the quadratic formula, for example.

When the roots are equal, $z_1 = z_2$ $(= z_0)$, a general solution to (3.31) is

$$u_n = z_0^{-n}(c_1 + c_2 n). \tag{3.33}$$

This claim can also be verified by direct substitution of (3.33) into (3.31):

$$\begin{aligned}
&z_0^{-n}(c_1 + c_2 n) - \alpha_1\left(z_0^{-(n-1)}[c_1 + c_2(n-1)]\right) - \alpha_2\left(z_0^{-(n-2)}[c_1 + c_2(n-2)]\right) \\
&\quad = z_0^{-n}(c_1 + c_2 n)\left(1 - \alpha_1 z_0 - \alpha_2 z_0^2\right) + c_2 z_0^{-n+1}\left(\alpha_1 + 2\alpha_2 z_0\right) \\
&\quad = c_2 z_0^{-n+1}\left(\alpha_1 + 2\alpha_2 z_0\right).
\end{aligned}$$

To show that $(\alpha_1 + 2\alpha_2 z_0) = 0$, write $1 - \alpha_1 z - \alpha_2 z^2 = (1 - z_0^{-1} z)^2$, and take derivatives with respect to z on both sides of the equation to obtain $(\alpha_1 + 2\alpha_2 z) = 2z_0^{-1}(1 - z_0^{-1} z)$. Thus, $(\alpha_1 + 2\alpha_2 z_0) = 2z_0^{-1}(1 - z_0^{-1} z_0) = 0$,

as was to be shown. Finally, given two initial conditions, u_0 and u_1, we can solve for c_1 and c_2:

$$u_0 = c_1 \quad \text{and} \quad u_1 = (c_1 + c_2)z_0^{-1}.$$

It can also be shown that these solutions are unique.

To summarize these results, in the case of distinct roots, the solution to the homogeneous difference equation of degree two was

$$\begin{aligned} u_n = z_1^{-n} \times (\text{a polynomial in } n \text{ of degree } m_1 - 1) \\ + \ z_2^{-n} \times (\text{a polynomial in } n \text{ of degree } m_2 - 1), \end{aligned} \tag{3.34}$$

where m_1 is the multiplicity of the root z_1 and m_2 is the multiplicity of the root z_2. In this example, of course, $m_1 = m_2 = 1$, and we called the polynomials of degree zero c_1 and c_2, respectively. In the case of the repeated root, the solution was

$$u_n = z_0^{-n} \times (\text{a polynomial in } n \text{ of degree } m_0 - 1), \tag{3.35}$$

where m_0 is the multiplicity of the root z_0; that is, $m_0 = 2$. In this case, we wrote the polynomial of degree one as $c_1 + c_2 n$. In both cases, we solved for c_1 and c_2 given two initial conditions, u_0 and u_1.

Example 3.9 The ACF of an AR(2) Process

Suppose $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ is a causal AR(2) process. Multiply each side of the model by x_{t-h} for $h > 0$, and take expectation:

$$E(x_t x_{t-h}) = \phi_1 E(x_{t-1} x_{t-h}) + \phi_2 E(x_{t-2} x_{t-h}) + E(w_t x_{t-h}).$$

The result is

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2), \quad h = 1, 2, \ldots . \tag{3.36}$$

In (3.36), we used the fact that $E(x_t) = 0$ and for $h > 0$,

$$E(w_t x_{t-h}) = E\Big(w_t \sum_{j=0}^{\infty} \psi_j w_{t-h-j}\Big) = 0.$$

Divide (3.36) through by $\gamma(0)$ to obtain the difference equation for the ACF of the process:

$$\rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \quad h = 1, 2, \ldots . \tag{3.37}$$

The initial conditions are $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1 - \phi_2)$, which is obtained by evaluating (3.37) for $h = 1$ and noting that $\rho(1) = \rho(-1)$.

Using the results for the homogeneous difference equation of order two, let z_1 and z_2 be the roots of the associated polynomial, $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$. Because the model is causal, we know the roots are outside the unit circle: $|z_1| > 1$ and $|z_2| > 1$. Now, consider the solution for three cases:

(i) When z_1 and z_2 are real and distinct, then

$$\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h},$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.

(ii) When $z_1 = z_2\,(= z_0)$ are real and equal, then

$$\rho(h) = z_0^{-h}(c_1 + c_2 h),$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.

(iii) When $z_1 = \bar{z}_2$ are a complex conjugate pair, then $c_2 = \bar{c}_1$ (because $\rho(h)$ is real), and

$$\rho(h) = c_1 z_1^{-h} + \bar{c}_1 \bar{z}_1^{-h}.$$

Write c_1 and z_1 in polar coordinates, for example, $z_1 = |z_1|e^{i\theta}$, where θ is the angle whose tangent is the ratio of the imaginary part and the real part of z_1 (sometimes called $\arg(z_1)$; the range of θ is $[-\pi, \pi]$). Then, using the fact that $e^{i\alpha} + e^{-i\alpha} = 2\cos(\alpha)$, the solution has the form

$$\rho(h) = a|z_1|^{-h}\cos(h\theta + b),$$

where a and b are determined by the initial conditions. Again, $\rho(h)$ dampens to zero exponentially fast as $h \to \infty$, but it does so in a sinusoidal fashion. The implication of this result is shown in the next example.

Example 3.10 An AR(2) with Complex Roots

Figure 3.4 shows $n = 144$ observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

with $\sigma_w^2 = 1$, and with complex roots chosen so the process exhibits pseudo-cyclic behavior at the rate of one cycle every 12 time points. The autoregressive polynomial for this model is $\phi(z) = 1 - 1.5z + .75z^2$. The roots of $\phi(z)$ are $1 \pm i/\sqrt{3}$, and $\theta = \tan^{-1}(1/\sqrt{3}) = 2\pi/12$ radians per unit time. To convert the angle to cycles per unit time, divide by 2π to get $1/12$ cycles per unit time. The ACF for this model is shown in §3.4, Figure 3.5.

To calculate the roots of the polynomial and solve for arg in R:

```
z = c(1,-1.5,.75)      # coefficients of the polynomial
(a = polyroot(z)[1])   # print one root: 1+0.57735i = 1 + i/sqrt(3)
arg = Arg(a)/(2*pi)    # arg in cycles/pt
1/arg                  # = 12,  the pseudo period
```

To reproduce Figure 3.4:

```
set.seed(90210)
ar2 = arima.sim(list(order=c(2,0,0), ar=c(1.5,-.75)), n = 144)
plot(1:144/12, ar2, type="l", xlab="Time (one unit = 12 points)")
abline(v=0:12, lty="dotted", lwd=2)
```

To calculate and display the ACF for this model:

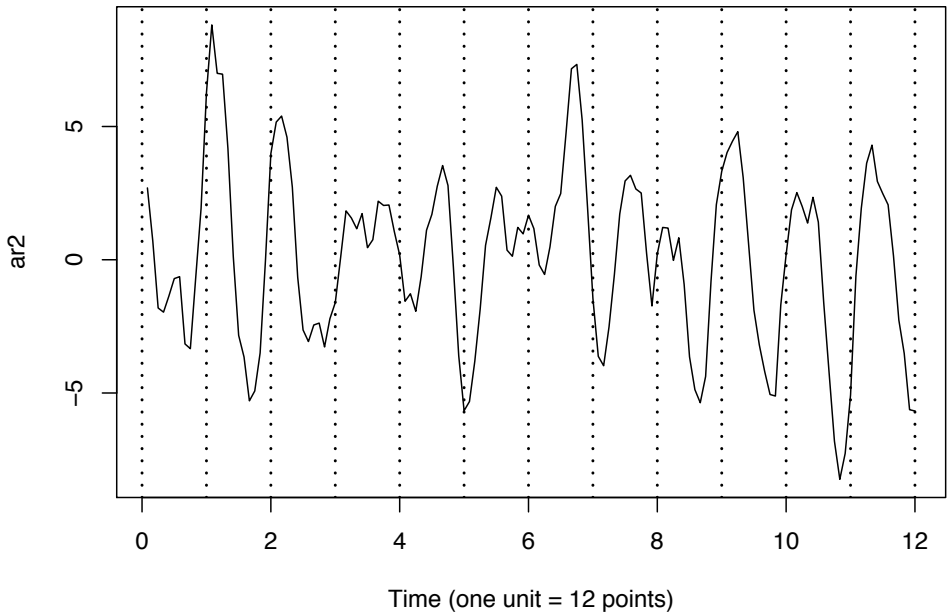

Fig. 3.4. Simulated AR(2) model, $n = 144$ with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 50)
plot(ACF, type="h", xlab="lag")
abline(h=0)
```

We now exhibit the solution for the general homogeneous difference equation of order p:

$$u_n - \alpha_1 u_{n-1} - \cdots - \alpha_p u_{n-p} = 0, \quad \alpha_p \neq 0, \quad n = p, p+1, \ldots . \tag{3.38}$$

The associated polynomial is

$$\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p.$$

Suppose $\alpha(z)$ has r distinct roots, z_1 with multiplicity m_1, z_2 with multiplicity m_2, ..., and z_r with multiplicity m_r, such that $m_1 + m_2 + \cdots + m_r = p$. The general solution to the difference equation (3.38) is

$$u_n = z_1^{-n} P_1(n) + z_2^{-n} P_2(n) + \cdots + z_r^{-n} P_r(n), \tag{3.39}$$

where $P_j(n)$, for $j = 1, 2, \ldots, r$, is a polynomial in n, of degree $m_j - 1$. Given p initial conditions $u_0, \ldots, u_{p-1}$, we can solve for the $P_j(n)$ explicitly.

Example 3.11 The ψ-weights for an ARMA Model

For a causal ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, where the zeros of $\phi(z)$ are outside the unit circle, recall that we may write

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

where the ψ-weights are determined using Property 3.1.

For the pure MA(q) model, $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \ldots, q$, and $\psi_j = 0$, otherwise. For the general case of ARMA(p, q) models, the task of solving for the ψ-weights is much more complicated, as was demonstrated in Example 3.7. The use of the theory of homogeneous difference equations can help here. To solve for the ψ-weights in general, we must match the coefficients in $\phi(z)\psi(z) = \theta(z)$:

$$(1 - \phi_1 z - \phi_2 z^2 - \cdots)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots) = (1 + \theta_1 z + \theta_2 z^2 + \cdots).$$

The first few values are

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 - \phi_1\psi_0 &= \theta_1 \\ \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 &= \theta_2 \\ \psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3\psi_0 &= \theta_3 \\ &\vdots \end{aligned}$$

where we would take $\phi_j = 0$ for $j > p$, and $\theta_j = 0$ for $j > q$. The ψ-weights satisfy the homogeneous difference equation given by

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = 0, \quad j \geq \max(p, q+1), \tag{3.40}$$

with initial conditions

$$\psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, \quad 0 \leq j < \max(p, q+1). \tag{3.41}$$

The general solution depends on the roots of the AR polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, as seen from (3.40). The specific solution will, of course, depend on the initial conditions.

Consider the ARMA process given in (3.27), $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Because $\max(p, q+1) = 2$, using (3.41), we have $\psi_0 = 1$ and $\psi_1 = .9 + .5 = 1.4$. By (3.40), for $j = 2, 3, \ldots$, the ψ-weights satisfy $\psi_j - .9\psi_{j-1} = 0$. The general solution is $\psi_j = c\,.9^j$. To find the specific solution, use the initial condition $\psi_1 = 1.4$, so $1.4 = .9c$ or $c = 1.4/.9$. Finally, $\psi_j = 1.4(.9)^{j-1}$, for $j \geq 1$, as we saw in Example 3.7.

To view, for example, the first 50 ψ-weights in R, use:

```
ARMAtoMA(ar=.9, ma=.5, 50)          # for a list
plot(ARMAtoMA(ar=.9, ma=.5, 50))    # for a graph
```

3.4 Autocorrelation and Partial Autocorrelation

We begin by exhibiting the ACF of an MA(q) process, $x_t = \theta(B)w_t$, where $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$. Because x_t is a finite linear combination of white noise terms, the process is stationary with mean

$$E(x_t) = \sum_{j=0}^{q} \theta_j E(w_{t-j}) = 0,$$

where we have written $\theta_0 = 1$, and with autocovariance function

$$\begin{aligned}\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) &= \operatorname{cov}\Big(\sum_{j=0}^{q} \theta_j w_{t+h-j}, \sum_{k=0}^{q} \theta_k w_{t-k}\Big) \\ &= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \le h \le q \\ 0 & h > q. \end{cases}\end{aligned} \tag{3.42}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only display the values for $h \ge 0$. The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (3.42) by $\gamma(0)$ yields the ACF of an MA(q):

$$\rho(h) = \begin{cases} \dfrac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \theta_1^2 + \cdots + \theta_q^2} & 1 \le h \le q \\ 0 & h > q. \end{cases} \tag{3.43}$$

For a causal ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, where the zeros of $\phi(z)$ are outside the unit circle, write

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}. \tag{3.44}$$

It follows immediately that $E(x_t) = 0$. Also, the autocovariance function of x_t can be written as

$$\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}, \quad h \ge 0. \tag{3.45}$$

We could then use (3.40) and (3.41) to solve for the ψ-weights. In turn, we could solve for $\gamma(h)$, and the ACF $\rho(h) = \gamma(h)/\gamma(0)$. As in Example 3.9, it is also possible to obtain a homogeneous difference equation directly in terms of $\gamma(h)$. First, we write

$$\begin{aligned}\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) &= \operatorname{cov}\Big(\sum_{j=1}^{p} \phi_j x_{t+h-j} + \sum_{j=0}^{q} \theta_j w_{t+h-j}, x_t\Big) \\ &= \sum_{j=1}^{p} \phi_j \gamma(h-j) + \sigma_w^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}, \quad h \ge 0,\end{aligned} \tag{3.46}$$

where we have used the fact that, for $h \ge 0$,

$$\text{cov}(w_{t+h-j},\, x_t) = \text{cov}\Big(w_{t+h-j},\, \sum_{k=0}^{\infty} \psi_k w_{t-k}\Big) = \psi_{j-h}\sigma_w^2.$$

From (3.46), we can write a general homogeneous equation for the ACF of a causal ARMA process:

$$\gamma(h) - \phi_1\gamma(h-1) - \cdots - \phi_p\gamma(h-p) = 0, \quad h \geq \max(p, q+1), \tag{3.47}$$

with initial conditions

$$\gamma(h) - \sum_{j=1}^{p} \phi_j\gamma(h-j) = \sigma_w^2 \sum_{j=h}^{q} \theta_j\psi_{j-h}, \quad 0 \leq h < \max(p, q+1). \tag{3.48}$$

Dividing (3.47) and (3.48) through by $\gamma(0)$ will allow us to solve for the ACF, $\rho(h) = \gamma(h)/\gamma(0)$.

Example 3.12 The ACF of an AR(p)

In Example 3.9 we considered the case where $p = 2$. For the general case, it follows immediately from (3.47) that

$$\rho(h) - \phi_1\rho(h-1) - \cdots - \phi_p\rho(h-p) = 0, \quad h \geq p. \tag{3.49}$$

Let $z_1, \ldots, z_r$ denote the roots of $\phi(z)$, each with multiplicity $m_1, \ldots, m_r$, respectively, where $m_1 + \cdots + m_r = p$. Then, from (3.39), the general solution is

$$\rho(h) = z_1^{-h}P_1(h) + z_2^{-h}P_2(h) + \cdots + z_r^{-h}P_r(h), \quad h \geq p, \tag{3.50}$$

where $P_j(h)$ is a polynomial in h of degree $m_j - 1$.

Recall that for a causal model, all of the roots are outside the unit circle, $|z_i| > 1$, for $i = 1, \ldots, r$. If all the roots are real, then $\rho(h)$ dampens exponentially fast to zero as $h \to \infty$. If some of the roots are complex, then they will be in conjugate pairs and $\rho(h)$ will dampen, in a sinusoidal fashion, exponentially fast to zero as $h \to \infty$. In the case of complex roots, the time series will appear to be cyclic in nature. This, of course, is also true for ARMA models in which the AR part has complex roots.

Example 3.13 The ACF of an ARMA$(1,1)$

Consider the ARMA$(1,1)$ process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$, where $|\phi| < 1$. Based on (3.47), the autocovariance function satisfies

$$\gamma(h) - \phi\gamma(h-1) = 0, \quad h = 2, 3, \ldots,$$

and it follows from (3.29)–(3.30) that the general solution is

$$\gamma(h) = c\,\phi^h, \quad h = 1, 2, \ldots\ . \tag{3.51}$$

To obtain the initial conditions, we use (3.48):

$$\gamma(0) = \phi\gamma(1) + \sigma_w^2[1 + \theta\phi + \theta^2] \quad \text{and} \quad \gamma(1) = \phi\gamma(0) + \sigma_w^2\theta.$$

Solving for $\gamma(0)$ and $\gamma(1)$, we obtain:

$$\gamma(0) = \sigma_w^2 \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2} \quad \text{and} \quad \gamma(1) = \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}.$$

To solve for c, note that from (3.51), $\gamma(1) = c\,\phi$ or $c = \gamma(1)/\phi$. Hence, the specific solution for $h \geq 1$ is

$$\gamma(h) = \frac{\gamma(1)}{\phi}\,\phi^h = \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}\,\phi^{h-1}.$$

Finally, dividing through by $\gamma(0)$ yields the ACF

$$\rho(h) = \frac{(1 + \theta\phi)(\phi + \theta)}{1 + 2\theta\phi + \theta^2}\,\phi^{h-1}, \quad h \geq 1. \tag{3.52}$$

Notice that the general pattern of $\rho(h)$ in (3.52) is not different from that of an AR(1) given in (3.8). Hence, it is unlikely that we will be able to tell the difference between an ARMA(1,1) and an AR(1) based solely on an ACF estimated from a sample. This consideration will lead us to the partial autocorrelation function.

The Partial Autocorrelation Function (PACF)

We have seen in (3.43), for MA(q) models, the ACF will be zero for lags greater than q. Moreover, because $\theta_q \neq 0$, the ACF will not be zero at lag q. Thus, the ACF provides a considerable amount of information about the order of the dependence when the process is a moving average process. If the process, however, is ARMA or AR, the ACF alone tells us little about the orders of dependence. Hence, it is worthwhile pursuing a function that will behave like the ACF of MA models, but for AR models, namely, the partial autocorrelation function (PACF).

To motivate the idea, consider a causal AR(1) model, $x_t = \phi x_{t-1} + w_t$. Then,

$$\begin{aligned}\gamma_x(2) = \operatorname{cov}(x_t, x_{t-2}) &= \operatorname{cov}(\phi x_{t-1} + w_t, x_{t-2}) \\ &= \operatorname{cov}(\phi^2 x_{t-2} + \phi w_{t-1} + w_t, x_{t-2}) = \phi^2\gamma_x(0).\end{aligned}$$

This result follows from causality because x_{t-2} involves $\{w_{t-2}, w_{t-3}, \ldots\}$, which are all uncorrelated with w_t and w_{t-1}. The correlation between x_t and x_{t-2} is not zero, as it would be for an MA(1), because x_t is dependent on x_{t-2} through x_{t-1}. Suppose we break this chain of dependence by removing (or partial out) the effect x_{t-1}. That is, we consider the correlation between $x_t - \phi x_{t-1}$ and $x_{t-2} - \phi x_{t-1}$, because it is the correlation between x_t and x_{t-2} with the linear dependence of each on x_{t-1} removed. In this way, we have broken the dependence chain between x_t and x_{t-2}. In fact,

$$\text{cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \text{cov}(w_t, x_{t-2} - \phi x_{t-1}) = 0.$$

Hence, the tool we need is partial autocorrelation, which is the correlation between x_s and x_t with the linear effect of everything "in the middle" removed.

To formally define the PACF for mean-zero stationary time series, let $\widehat{x}_{t+h}$, for $h \geq 2$, denote the regression[3] of x_{t+h} on $\{x_{t+h-1}, x_{t+h-2}, \ldots, x_{t+1}\}$, which we write as

$$\widehat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1}. \tag{3.53}$$

No intercept term is needed in (3.53) because the mean of x_t is zero (otherwise, replace x_t by $x_t - \mu_x$ in this discussion). In addition, let $\widehat{x}_t$ denote the regression of x_t on $\{x_{t+1}, x_{t+2}, \ldots, x_{t+h-1}\}$, then

$$\widehat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1}. \tag{3.54}$$

Because of stationarity, the coefficients, $\beta_1, \ldots, \beta_{h-1}$ are the same in (3.53) and (3.54); we will explain this result in the next section.

Definition 3.9 *The* **partial autocorrelation function (PACF)** *of a stationary process, x_t, denoted ϕ_{hh}, for $h = 1, 2, \ldots$, is*

$$\phi_{11} = \text{corr}(x_{t+1}, x_t) = \rho(1) \tag{3.55}$$

and

$$\phi_{hh} = \text{corr}(x_{t+h} - \widehat{x}_{t+h},\, x_t - \widehat{x}_t), \quad h \geq 2. \tag{3.56}$$

Both $(x_{t+h} - \widehat{x}_{t+h})$ and $(x_t - \widehat{x}_t)$ are uncorrelated with $\{x_{t+1}, \ldots, x_{t+h-1}\}$. The PACF, ϕ_{hh}, is the correlation between x_{t+h} and x_t with the linear dependence of $\{x_{t+1}, \ldots, x_{t+h-1}\}$ on each, removed. If the process x_t is Gaussian, then $\phi_{hh} = \text{corr}(x_{t+h}, x_t \mid x_{t+1}, \ldots, x_{t+h-1})$, that is, ϕ_{hh} is the correlation coefficient between x_{t+h} and x_t in the bivariate distribution of (x_{t+h}, x_t) conditional on $\{x_{t+1}, \ldots, x_{t+h-1}\}$.

Example 3.14 The PACF of an AR(1)

Consider the PACF of the AR(1) process given by $x_t = \phi x_{t-1} + w_t$, with $|\phi| < 1$. By definition, $\phi_{11} = \rho(1) = \phi$. To calculate ϕ_{22}, consider the regression of x_{t+2} on x_{t+1}, say, $\widehat{x}_{t+2} = \beta x_{t+1}$. We choose β to minimize

$$E(x_{t+2} - \widehat{x}_{t+2})^2 = E(x_{t+2} - \beta x_{t+1})^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0).$$

Taking derivatives with respect to β and setting the result equal to zero, we have $\beta = \gamma(1)/\gamma(0) = \rho(1) = \phi$. Next, consider the regression of x_t on x_{t+1}, say $\widehat{x}_t = \beta x_{t+1}$. We choose β to minimize

[3] The term regression here refers to regression in the population sense. That is, $\widehat{x}_{t+h}$ is the linear combination of $\{x_{t+h-1}, x_{t+h-2}, \ldots, x_{t+1}\}$ that minimizes the mean squared error $E(x_{t+h} - \sum_{j=1}^{h-1} \alpha_j x_{t+j})^2$.

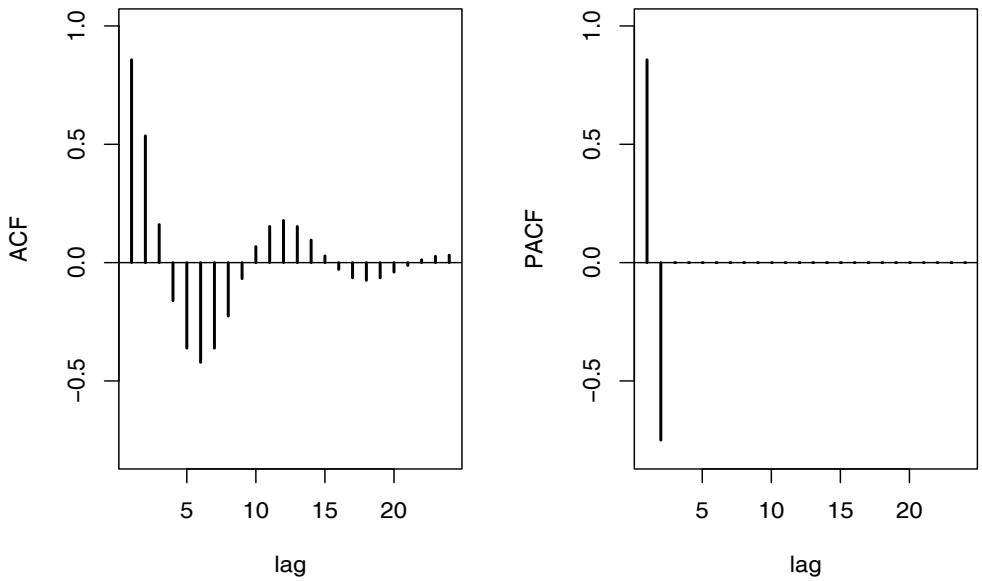

Fig. 3.5. The ACF and PACF of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

$$E(x_t - \widehat{x}_t)^2 = E(x_t - \beta x_{t+1})^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0).$$

This is the same equation as before, so $\beta = \phi$. Hence,

$$\phi_{22} = \text{corr}(x_{t+2} - \widehat{x}_{t+2}, x_t - \widehat{x}_t) = \text{corr}(x_{t+2} - \phi x_{t+1}, x_t - \phi x_{t+1})$$
$$= \text{corr}(w_{t+2}, x_t - \phi x_{t+1}) = 0$$

by causality. Thus, $\phi_{22} = 0$. In the next example, we will see that in this case, $\phi_{hh} = 0$ for all $h > 1$.

Example 3.15 The PACF of an AR(p)

The model implies $x_{t+h} = \sum_{j=1}^{p} \phi_j x_{t+h-j} + w_{t+h}$, where the roots of $\phi(z)$ are outside the unit circle. When $h > p$, the regression of x_{t+h} on $\{x_{t+1}, \ldots, x_{t+h-1}\}$, is

$$\widehat{x}_{t+h} = \sum_{j=1}^{p} \phi_j x_{t+h-j}.$$

We have not proved this obvious result yet, but we will prove it in the next section. Thus, when $h > p$,

$$\phi_{hh} = \text{corr}(x_{t+h} - \widehat{x}_{t+h},\, x_t - \widehat{x}_t) = \text{corr}(w_{t+h},\, x_t - \widehat{x}_t) = 0,$$

because, by causality, $x_t - \widehat{x}_t$ depends only on $\{w_{t+h-1}, w_{t+h-2}, \ldots\}$; recall equation (3.54). When $h \leq p$, ϕ_{pp} is not zero, and $\phi_{11}, \ldots, \phi_{p-1,p-1}$ are not necessarily zero. We will see later that, in fact, $\phi_{pp} = \phi_p$. Figure 3.5 shows the ACF and the PACF of the AR(2) model presented in Example 3.10.

To reproduce Figure 3.5 in R, use the following commands:

Table 3.1. Behavior of the ACF and PACF for ARMA Models

	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

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24)[-1]
PACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24, pacf=TRUE)
par(mfrow=c(1,2))
plot(ACF, type="h", xlab="lag", ylim=c(-.8,1));  abline(h=0)
plot(PACF, type="h", xlab="lag", ylim=c(-.8,1));  abline(h=0)
```

Example 3.16 The PACF of an Invertible MA(q)

For an invertible MA(q), we can write $x_t = -\sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t$. Moreover, no finite representation exists. From this result, it should be apparent that the PACF will never cut off, as in the case of an AR(p).

For an MA(1), $x_t = w_t + \theta w_{t-1}$, with $|\theta| < 1$, calculations similar to Example 3.14 will yield $\phi_{22} = -\theta^2/(1 + \theta^2 + \theta^4)$. For the MA(1) in general, we can show that

$$\phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}, \quad h \geq 1.$$

In the next section, we will discuss methods of calculating the PACF. The PACF for MA models behaves much like the ACF for AR models. Also, the PACF for AR models behaves much like the ACF for MA models. Because an invertible ARMA model has an infinite AR representation, the PACF will not cut off. We may summarize these results in Table 3.1.

Example 3.17 Preliminary Analysis of the Recruitment Series

We consider the problem of modeling the Recruitment series shown in Figure 1.5. There are 453 months of observed recruitment ranging over the years 1950-1987. The ACF and the PACF given in Figure 3.6 are consistent with the behavior of an AR(2). The ACF has cycles corresponding roughly to a 12-month period, and the PACF has large values for $h = 1, 2$ and then is essentially zero for higher order lags. Based on Table 3.1, these results suggest that a second-order ($p = 2$) autoregressive model might provide a good fit. Although we will discuss estimation in detail in §3.6, we ran a regression (see §2.2) using the data triplets $\{(x; z_1, z_2) : (x_3; x_2, x_1), (x_4; x_3, x_2), \ldots, (x_{453}; x_{452}, x_{451})\}$ to fit a model of the form

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

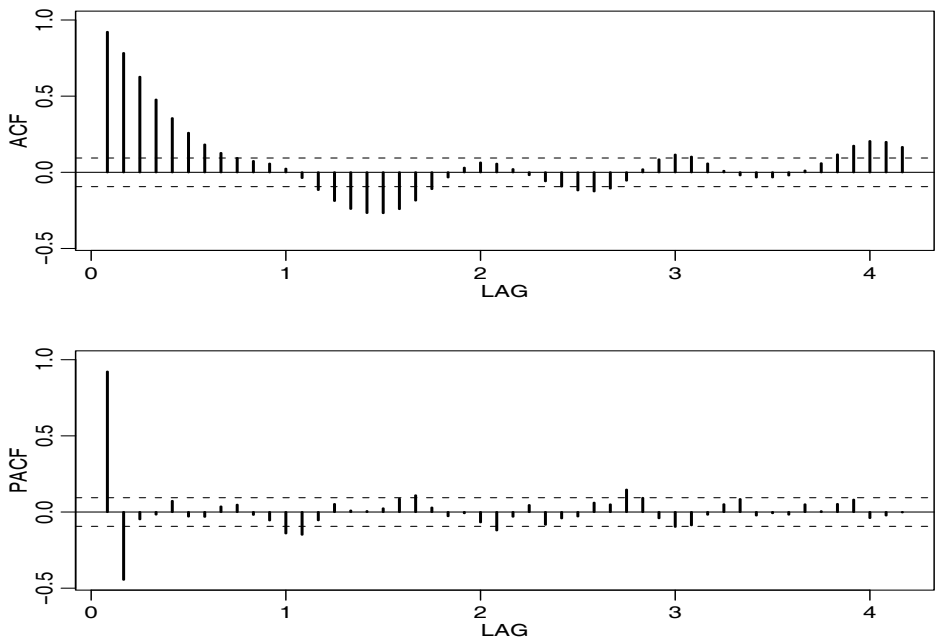

Fig. 3.6. ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).

for $t = 3, 4, \ldots, 453$. The values of the estimates were $\widehat{\phi}_0 = 6.74_{(1.11)}$, $\widehat{\phi}_1 = 1.35_{(.04)}$, $\widehat{\phi}_2 = -.46_{(.04)}$, and $\widehat{\sigma}_w^2 = 89.72$, where the estimated standard errors are in parentheses.

The following R code can be used for this analysis. We use the script `acf2` to print and plot the ACF and PACF; see Appendix R for details.

```
acf2(rec, 48)      # will produce values and a graphic
(regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE))
regr$asy.se.coef  # standard errors of the estimates
```

3.5 Forecasting

In forecasting, the goal is to predict future values of a time series, x_{n+m}, $m = 1, 2, \ldots$, based on the data collected to the present, $\boldsymbol{x} = \{x_n, x_{n-1}, \ldots, x_1\}$. Throughout this section, we will assume x_t is stationary and the model parameters are known. The problem of forecasting when the model parameters are unknown will be discussed in the next section; also, see Problem 3.26. The minimum mean square error predictor of x_{n+m} is

$$x_{n+m}^n = E(x_{n+m} \mid \boldsymbol{x}) \tag{3.57}$$

because the conditional expectation minimizes the mean square error

$$E\left[x_{n+m} - g(\boldsymbol{x})\right]^2, \tag{3.58}$$

where $g(\boldsymbol{x})$ is a function of the observations $\boldsymbol{x}$; see Problem 3.14.

First, we will restrict attention to predictors that are linear functions of the data, that is, predictors of the form

$$x_{n+m}^n = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k, \tag{3.59}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_n$ are real numbers. Linear predictors of the form (3.59) that minimize the mean square prediction error (3.58) are called best linear predictors (BLPs). As we shall see, linear prediction depends only on the second-order moments of the process, which are easy to estimate from the data. Much of the material in this section is enhanced by the theoretical material presented in Appendix B. For example, Theorem B.3 states that if the process is Gaussian, minimum mean square error predictors and best linear predictors are the same. The following property, which is based on the Projection Theorem, Theorem B.1 of Appendix B, is a key result.

Property 3.3 Best Linear Prediction for Stationary Processes

Given data $x_1, \ldots, x_n$, the best linear predictor, $x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$, of x_{n+m}, for $m \geq 1$, is found by solving

$$E\left[\left(x_{n+m} - x_{n+m}^n\right) x_k\right] = 0, \quad k = 0, 1, \ldots, n, \tag{3.60}$$

where $x_0 = 1$, for $\alpha_0, \alpha_1, \ldots \alpha_n$.

The equations specified in (3.60) are called the prediction equations, and they are used to solve for the coefficients $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$. If $E(x_t) = \mu$, the first equation ($k = 0$) of (3.60) implies

$$E(x_{n+m}^n) = E(x_{n+m}) = \mu.$$

Thus, taking expectation in (3.59), we have

$$\mu = \alpha_0 + \sum_{k=1}^{n} \alpha_k \mu \qquad \text{or} \qquad \alpha_0 = \mu\Big(1 - \sum_{k=1}^{n} \alpha_k\Big).$$

Hence, the form of the BLP is

$$x_{n+m}^n = \mu + \sum_{k=1}^{n} \alpha_k (x_k - \mu).$$

Thus, until we discuss estimation, there is no loss of generality in considering the case that $\mu = 0$, in which case, $\alpha_0 = 0$.

First, consider one-step-ahead prediction. That is, given $\{x_1, \ldots, x_n\}$, we wish to forecast the value of the time series at the next time point, x_{n+1}. The BLP of x_{n+1} is of the form

$$x_{n+1}^n = \phi_{n1}x_n + \phi_{n2}x_{n-1} + \cdots + \phi_{nn}x_1, \tag{3.61}$$

where, for purposes that will become clear shortly, we have written α_k in (3.59), as $\phi_{n,n+1-k}$ in (3.61), for $k = 1, \ldots, n$. Using Property 3.3, the coefficients $\{\phi_{n1}, \phi_{n2}, \ldots, \phi_{nn}\}$ satisfy

$$E\Big[\Big(x_{n+1} - \sum_{j=1}^{n} \phi_{nj}x_{n+1-j}\Big)x_{n+1-k}\Big] = 0, \quad k = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} \phi_{nj}\gamma(k-j) = \gamma(k), \quad k = 1, \ldots, n. \tag{3.62}$$

The prediction equations (3.62) can be written in matrix notation as

$$\Gamma_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n, \tag{3.63}$$

where $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^n$ is an $n \times n$ matrix, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})'$ is an $n \times 1$ vector, and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))'$ is an $n \times 1$ vector.

The matrix Γ_n is nonnegative definite. If Γ_n is singular, there are many solutions to (3.63), but, by the Projection Theorem (Theorem B.1), x_{n+1}^n is unique. If Γ_n is nonsingular, the elements of $\boldsymbol{\phi}_n$ are unique, and are given by

$$\boldsymbol{\phi}_n = \Gamma_n^{-1}\boldsymbol{\gamma}_n. \tag{3.64}$$

For ARMA models, the fact that $\sigma_w^2 > 0$ and $\gamma(h) \to 0$ as $h \to \infty$ is enough to ensure that Γ_n is positive definite (Problem 3.12). It is sometimes convenient to write the one-step-ahead forecast in vector notation

$$x_{n+1}^n = \boldsymbol{\phi}_n'\boldsymbol{x}, \tag{3.65}$$

where $\boldsymbol{x} = (x_n, x_{n-1}, \ldots, x_1)'$.

The mean square one-step-ahead prediction error is

$$P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n. \tag{3.66}$$

To verify (3.66) using (3.64) and (3.65),

$$\begin{aligned} E(x_{n+1} - x_{n+1}^n)^2 &= E(x_{n+1} - \boldsymbol{\phi}_n'\boldsymbol{x})^2 = E(x_{n+1} - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x})^2 \\ &= E(x_{n+1}^2 - 2\boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x}x_{n+1} + \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x}\boldsymbol{x}'\Gamma_n^{-1}\boldsymbol{\gamma}_n) \\ &= \gamma(0) - 2\boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n + \boldsymbol{\gamma}_n'\Gamma_n^{-1}\Gamma_n\Gamma_n^{-1}\boldsymbol{\gamma}_n \\ &= \gamma(0) - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n. \end{aligned}$$

Example 3.18 Prediction for an AR(2)

Suppose we have a causal AR(2) process $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$, and one observation x_1. Then, using equation (3.64), the one-step-ahead prediction of x_2 based on x_1 is

$$x_2^1 = \phi_{11} x_1 = \frac{\gamma(1)}{\gamma(0)} x_1 = \rho(1) x_1.$$

Now, suppose we want the one-step-ahead prediction of x_3 based on two observations x_1 and x_2; i.e., $x_3^2 = \phi_{21} x_2 + \phi_{22} x_1$. We could use (3.62)

$$\begin{aligned} \phi_{21}\gamma(0) + \phi_{22}\gamma(1) &= \gamma(1) \\ \phi_{21}\gamma(1) + \phi_{22}\gamma(0) &= \gamma(2) \end{aligned}$$

to solve for ϕ_{21} and ϕ_{22}, or use the matrix form in (3.64) and solve

$$\begin{pmatrix} \phi_{21} \\ \phi_{22} \end{pmatrix} = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} \gamma(1) \\ \gamma(2) \end{pmatrix},$$

but, it should be apparent from the model that $x_3^2 = \phi_1 x_2 + \phi_2 x_1$. Because $\phi_1 x_2 + \phi_2 x_1$ satisfies the prediction equations (3.60),

$$E\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_1\} = E(w_3 x_1) = 0,$$

$$E\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_2\} = E(w_3 x_2) = 0,$$

it follows that, indeed, $x_3^2 = \phi_1 x_2 + \phi_2 x_1$, and by the uniqueness of the coefficients in this case, that $\phi_{21} = \phi_1$ and $\phi_{22} = \phi_2$. Continuing in this way, it is easy to verify that, for $n \geq 2$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1}.$$

That is, $\phi_{n1} = \phi_1, \phi_{n2} = \phi_2$, and $\phi_{nj} = 0$, for $j = 3, 4, \ldots, n$.

From Example 3.18, it should be clear (Problem 3.40) that, if the time series is a causal AR(p) process, then, for $n \geq p$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \phi_p x_{n-p+1}. \tag{3.67}$$

For ARMA models in general, the prediction equations will not be as simple as the pure AR case. In addition, for n large, the use of (3.64) is prohibitive because it requires the inversion of a large matrix. There are, however, iterative solutions that do not require any matrix inversion. In particular, we mention the recursive solution due to Levinson (1947) and Durbin (1960).

Property 3.4 The Durbin–Levinson Algorithm

Equations (3.64) *and* (3.66) *can be solved iteratively as follows:*

$$\phi_{00} = 0, \quad P_1^0 = \gamma(0). \tag{3.68}$$

For $n \geq 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi_{n-1,k}\ \rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k}\ \rho(k)}, \quad P_{n+1}^{n} = P_{n}^{n-1}(1 - \phi_{nn}^2), \tag{3.69}$$

where, for $n \geq 2$,

$$\phi_{nk} = \phi_{n-1,k} - \phi_{nn}\phi_{n-1,n-k}, \quad k = 1, 2, \ldots, n-1. \tag{3.70}$$

The proof of Property 3.4 is left as an exercise; see Problem 3.13.

Example 3.19 Using the Durbin–Levinson Algorithm

To use the algorithm, start with $\phi_{00} = 0$, $P_1^0 = \gamma(0)$. Then, for $n = 1$,

$$\phi_{11} = \rho(1), \quad P_2^1 = \gamma(0)[1 - \phi_{11}^2].$$

For $n = 2$,

$$\phi_{22} = \frac{\rho(2) - \phi_{11}\ \rho(1)}{1 - \phi_{11}\ \rho(1)},\ \phi_{21} = \phi_{11} - \phi_{22}\phi_{11},$$
$$P_3^2 = P_2^1[1 - \phi_{22}^2] = \gamma(0)[1 - \phi_{11}^2][1 - \phi_{22}^2].$$

For $n = 3$,

$$\phi_{33} = \frac{\rho(3) - \phi_{21}\ \rho(2) - \phi_{22}\ \rho(1)}{1 - \phi_{21}\ \rho(1) - \phi_{22}\ \rho(2)},$$
$$\phi_{32} = \phi_{22} - \phi_{33}\phi_{21},\ \phi_{31} = \phi_{21} - \phi_{33}\phi_{22},$$
$$P_4^3 = P_3^2[1 - \phi_{33}^2] = \gamma(0)[1 - \phi_{11}^2][1 - \phi_{22}^2][1 - \phi_{33}^2],$$

and so on. Note that, in general, the standard error of the one-step-ahead forecast is the square root of

$$P_{n+1}^{n} = \gamma(0) \prod_{j=1}^{n} [1 - \phi_{jj}^2]. \tag{3.71}$$

An important consequence of the Durbin–Levinson algorithm is (see Problem 3.13) as follows.

Property 3.5 Iterative Solution for the PACF

The PACF of a stationary process x_t, *can be obtained iteratively via* (3.69) *as* ϕ_{nn}, *for* $n = 1, 2, \ldots$.

Using Property 3.5 and putting $n = p$ in (3.61) and (3.67), it follows that for an AR(p) model,

$$\begin{aligned} x_{p+1}^{p} &= \phi_{p1}\, x_p + \phi_{p2}\, x_{p-1} + \cdots + \phi_{pp}\, x_1 \\ &= \phi_1\, x_p + \phi_2\, x_{p-1} + \cdots + \phi_p\, x_1. \end{aligned} \tag{3.72}$$

Result (3.72) shows that for an AR(p) model, the partial autocorrelation coefficient at lag p, ϕ_{pp}, is also the last coefficient in the model, ϕ_p, as was claimed in Example 3.15.

Example 3.20 The PACF of an AR(2)

We will use the results of Example 3.19 and Property 3.5 to calculate the first three values, ϕ_{11}, ϕ_{22}, ϕ_{33}, of the PACF. Recall from Example 3.9 that $\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0$ for $h \geq 1$. When $h = 1, 2, 3$, we have $\rho(1) = \phi_1/(1-\phi_2)$, $\rho(2) = \phi_1\rho(1) + \phi_2$, $\rho(3) - \phi_1\rho(2) - \phi_2\rho(1) = 0$. Thus,

$$\phi_{11} = \rho(1) = \frac{\phi_1}{1-\phi_2}$$

$$\phi_{22} = \frac{\rho(2)-\rho(1)^2}{1-\rho(1)^2} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right)+\phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1-\left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2$$

$$\phi_{21} = \rho(1)[1-\phi_2] = \phi_1$$

$$\phi_{33} = \frac{\rho(3)-\phi_1\rho(2)-\phi_2\rho(1)}{1-\phi_1\rho(1)-\phi_2\rho(2)} = 0.$$

Notice that, as shown in (3.72), $\phi_{22} = \phi_2$ for an AR(2) model.

So far, we have concentrated on one-step-ahead prediction, but Property 3.3 allows us to calculate the BLP of x_{n+m} for any $m \geq 1$. Given data, $\{x_1, \ldots, x_n\}$, the m-step-ahead predictor is

$$x^n_{n+m} = \phi^{(m)}_{n1}x_n + \phi^{(m)}_{n2}x_{n-1} + \cdots + \phi^{(m)}_{nn}x_1, \tag{3.73}$$

where $\{\phi^{(m)}_{n1}, \phi^{(m)}_{n2}, \ldots, \phi^{(m)}_{nn}\}$ satisfy the prediction equations,

$$\sum_{j=1}^{n} \phi^{(m)}_{nj} E(x_{n+1-j}x_{n+1-k}) = E(x_{n+m}x_{n+1-k}), \quad k = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} \phi^{(m)}_{nj}\ \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \ldots, n. \tag{3.74}$$

The prediction equations can again be written in matrix notation as

$$\Gamma_n \boldsymbol{\phi}^{(m)}_n = \boldsymbol{\gamma}^{(m)}_n, \tag{3.75}$$

where $\boldsymbol{\gamma}^{(m)}_n = (\gamma(m), \ldots, \gamma(m+n-1))'$, and $\boldsymbol{\phi}^{(m)}_n = (\phi^{(m)}_{n1}, \ldots, \phi^{(m)}_{nn})'$ are $n \times 1$ vectors.

The mean square m-step-ahead prediction error is

$$P^n_{n+m} = E\left(x_{n+m} - x^n_{n+m}\right)^2 = \gamma(0) - \boldsymbol{\gamma}^{(m)\prime}_n \Gamma^{-1}_n \boldsymbol{\gamma}^{(m)}_n. \tag{3.76}$$

Another useful algorithm for calculating forecasts was given by Brockwell and Davis (1991, Chapter 5). This algorithm follows directly from applying the projection theorem (Theorem B.1) to the innovations, $x_t - x^{t-1}_t$, for $t = 1, \ldots, n$, using the fact that the innovations $x_t - x^{t-1}_t$ and $x_s - x^{s-1}_s$ are uncorrelated for $s \neq t$ (see Problem 3.41). We present the case in which x_t is a mean-zero stationary time series.

Property 3.6 The Innovations Algorithm

The one-step-ahead predictors, x_{t+1}^t, and their mean-squared errors, P_{t+1}^t, can be calculated iteratively as

$$x_1^0 = 0, \quad P_1^0 = \gamma(0)$$

$$x_{t+1}^t = \sum_{j=1}^{t} \theta_{tj}(x_{t+1-j} - x_{t+1-j}^{t-j}), \quad t = 1, 2, \ldots \tag{3.77}$$

$$P_{t+1}^t = \gamma(0) - \sum_{j=0}^{t-1} \theta_{t,t-j}^2 P_{j+1}^j \quad t = 1, 2, \ldots, \tag{3.78}$$

where, for $j = 0, 1, \ldots, t-1$,

$$\theta_{t,t-j} = \left(\gamma(t-j) - \sum_{k=0}^{j-1} \theta_{j,j-k}\theta_{t,t-k}P_{k+1}^k\right) / P_{j+1}^j. \tag{3.79}$$

Given data $x_1, \ldots, x_n$, the innovations algorithm can be calculated successively for $t = 1$, then $t = 2$ and so on, in which case the calculation of x_{n+1}^n and P_{n+1}^n is made at the final step $t = n$. The m-step-ahead predictor and its mean-square error based on the innovations algorithm (Problem 3.41) are given by

$$x_{n+m}^n = \sum_{j=m}^{n+m-1} \theta_{n+m-1,j}(x_{n+m-j} - x_{n+m-j}^{n+m-j-1}), \tag{3.80}$$

$$P_{n+m}^n = \gamma(0) - \sum_{j=m}^{n+m-1} \theta_{n+m-1,j}^2 P_{n+m-j}^{n+m-j-1}, \tag{3.81}$$

where the $\theta_{n+m-1,j}$ are obtained by continued iteration of (3.79).

Example 3.21 Prediction for an MA(1)

The innovations algorithm lends itself well to prediction for moving average processes. Consider an MA(1) model, $x_t = w_t + \theta w_{t-1}$. Recall that $\gamma(0) = (1+\theta^2)\sigma_w^2$, $\gamma(1) = \theta\sigma_w^2$, and $\gamma(h) = 0$ for $h > 1$. Then, using Property 3.6, we have

$$\theta_{n1} = \theta\sigma_w^2/P_n^{n-1}$$
$$\theta_{nj} = 0, \quad j = 2, \ldots, n$$
$$P_1^0 = (1+\theta^2)\sigma_w^2$$
$$P_{n+1}^n = (1 + \theta^2 - \theta\theta_{n1})\sigma_w^2.$$

Finally, from (3.77), the one-step-ahead predictor is

$$x_{n+1}^n = \theta\left(x_n - x_n^{n-1}\right)\sigma_w^2/P_n^{n-1}.$$

Forecasting ARMA Processes

The general prediction equations (3.60) provide little insight into forecasting for ARMA models in general. There are a number of different ways to express these forecasts, and each aids in understanding the special structure of ARMA prediction. Throughout, we assume x_t is a causal and invertible ARMA(p, q) process, $\phi(B)x_t = \theta(B)w_t$, where $w_t \sim$ iid N$(0, \sigma_w^2)$. In the non-zero mean case, $E(x_t) = \mu_x$, simply replace x_t with $x_t - \mu_x$ in the model. First, we consider two types of forecasts. We write x_{n+m}^n to mean the minimum mean square error predictor of x_{n+m} based on the data $\{x_n, \ldots, x_1\}$, that is,

$$x_{n+m}^n = E(x_{n+m} \mid x_n, \ldots, x_1).$$

For ARMA models, it is easier to calculate the predictor of x_{n+m}, assuming we have the complete history of the process $\{x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots\}$. We will denote the predictor of x_{n+m} *based on the infinite past* as

$$\widetilde{x}_{n+m} = E(x_{n+m} \mid x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots).$$

In general, x_{n+m}^n and $\widetilde{x}_{n+m}$ are not the same, but the idea here is that, for large samples, $\widetilde{x}_{n+m}$ will provide a good approximation to x_{n+m}^n.

Now, write x_{n+m} in its causal and invertible forms:

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}, \quad \psi_0 = 1 \tag{3.82}$$

$$w_{n+m} = \sum_{j=0}^{\infty} \pi_j x_{n+m-j}, \quad \pi_0 = 1. \tag{3.83}$$

Then, taking conditional expectations in (3.82), we have

$$\widetilde{x}_{n+m} = \sum_{j=0}^{\infty} \psi_j \widetilde{w}_{n+m-j} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}, \tag{3.84}$$

because, by causality and invertibility,

$$\widetilde{w}_t = E(w_t \mid x_n, x_{n-1}, \ldots, x_0, x_{-1}, \ldots) = \begin{cases} 0 & t > n \\ w_t & t \le n. \end{cases}$$

Similarly, taking conditional expectations in (3.83), we have

$$0 = \widetilde{x}_{n+m} + \sum_{j=1}^{\infty} \pi_j \widetilde{x}_{n+m-j},$$

or

$$\widetilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, \tag{3.85}$$

using the fact $E(x_t \mid x_n, x_{n-1}, \ldots, x_0, x_{-1}, \ldots) = x_t$, for $t \leq n$. Prediction is accomplished recursively using (3.85), starting with the one-step-ahead predictor, $m = 1$, and then continuing for $m = 2, 3, \ldots$. Using (3.84), we can write

$$x_{n+m} - \widetilde{x}_{n+m} = \sum_{j=0}^{m-1} \psi_j w_{n+m-j},$$

so the mean-square prediction error can be written as

$$P_{n+m}^{n} = E(x_{n+m} - \widetilde{x}_{n+m})^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.86}$$

Also, we note, for a fixed sample size, n, the prediction errors are correlated. That is, for $k \geq 1$,

$$E\{(x_{n+m} - \widetilde{x}_{n+m})(x_{n+m+k} - \widetilde{x}_{n+m+k})\} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}. \tag{3.87}$$

Example 3.22 Long-Range Forecasts

Consider forecasting an ARMA process with mean μ_x. Replacing x_{n+m} with $x_{n+m} - \mu_x$ in (3.82), and taking conditional expectation as is in (3.84), we deduce that the m-step-ahead forecast can be written as

$$\widetilde{x}_{n+m} = \mu_x + \sum_{j=m}^{\infty} \psi_j w_{n+m-j}. \tag{3.88}$$

Noting that the ψ-weights dampen to zero exponentially fast, it is clear that

$$\widetilde{x}_{n+m} \to \mu_x \tag{3.89}$$

exponentially fast (in the mean square sense) as $m \to \infty$. Moreover, by (3.86), the mean square prediction error

$$P_{n+m}^{n} \to \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_x(0) = \sigma_x^2, \tag{3.90}$$

exponentially fast as $m \to \infty$; recall (3.45).

It should be clear from (3.89) and (3.90) that ARMA forecasts quickly settle to the mean with a constant prediction error as the forecast horizon, m, grows. This effect can be seen in Figure 3.7 on page 119 where the Recruitment series is forecast for 24 months; see Example 3.24.

When n is small, the general prediction equations (3.60) can be used easily. When n is large, we would use (3.85) by truncating, because we do not observe $x_0, x_{-1}, x_{-2}, \ldots$, and only the data $x_1, x_2, \ldots, x_n$ are available. In this case, we can truncate (3.85) by setting $\sum_{j=n+m}^{\infty} \pi_j x_{n+m-j} = 0$. The truncated predictor is then written as

$$\widetilde{x}_{n+m}^n = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}_{n+m-j}^n - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}, \tag{3.91}$$

which is also calculated recursively, $m = 1, 2, \ldots$. The mean square prediction error, in this case, is approximated using (3.86).

For AR(p) models, and when $n > p$, equation (3.67) yields the exact predictor, x_{n+m}^n, of x_{n+m}, and there is no need for approximations. That is, for $n > p$, $\widetilde{x}_{n+m}^n = \widetilde{x}_{n+m} = x_{n+m}^n$. Also, in this case, the one-step-ahead prediction error is $E(x_{n+1} - x_{n+1}^n)^2 = \sigma_w^2$. For pure MA($q$) or ARMA($p, q$) models, truncated prediction has a fairly nice form.

Property 3.7 Truncated Prediction for ARMA

For ARMA(p, q) models, the truncated predictors for $m = 1, 2, \ldots,$ are

$$\widetilde{x}_{n+m}^n = \phi_1 \widetilde{x}_{n+m-1}^n + \cdots + \phi_p \widetilde{x}_{n+m-p}^n + \theta_1 \widetilde{w}_{n+m-1}^n + \cdots + \theta_q \widetilde{w}_{n+m-q}^n, \tag{3.92}$$

where $\widetilde{x}_t^n = x_t$ for $1 \le t \le n$ and $\widetilde{x}_t^n = 0$ for $t \le 0$. The truncated prediction errors are given by: $\widetilde{w}_t^n = 0$ for $t \le 0$ or $t > n$, and

$$\widetilde{w}_t^n = \phi(B)\widetilde{x}_t^n - \theta_1 \widetilde{w}_{t-1}^n - \cdots - \theta_q \widetilde{w}_{t-q}^n$$

for $1 \le t \le n$.

Example 3.23 Forecasting an ARMA$(1, 1)$ Series

Given data $x_1, \ldots, x_n$, for forecasting purposes, write the model as

$$x_{n+1} = \phi x_n + w_{n+1} + \theta w_n.$$

Then, based on (3.92), the one-step-ahead truncated forecast is

$$\widetilde{x}_{n+1}^n = \phi x_n + 0 + \theta \widetilde{w}_n^n.$$

For $m \ge 2$, we have

$$\widetilde{x}_{n+m}^n = \phi \widetilde{x}_{n+m-1}^n,$$

which can be calculated recursively, $m = 2, 3, \ldots$.

To calculate $\widetilde{w}_n^n$, which is needed to initialize the successive forecasts, the model can be written as $w_t = x_t - \phi x_{t-1} - \theta w_{t-1}$ for $t = 1, \ldots, n$. For truncated forecasting using (3.92), put $\widetilde{w}_0^n = 0$, $x_0 = 0$, and then iterate the errors forward in time

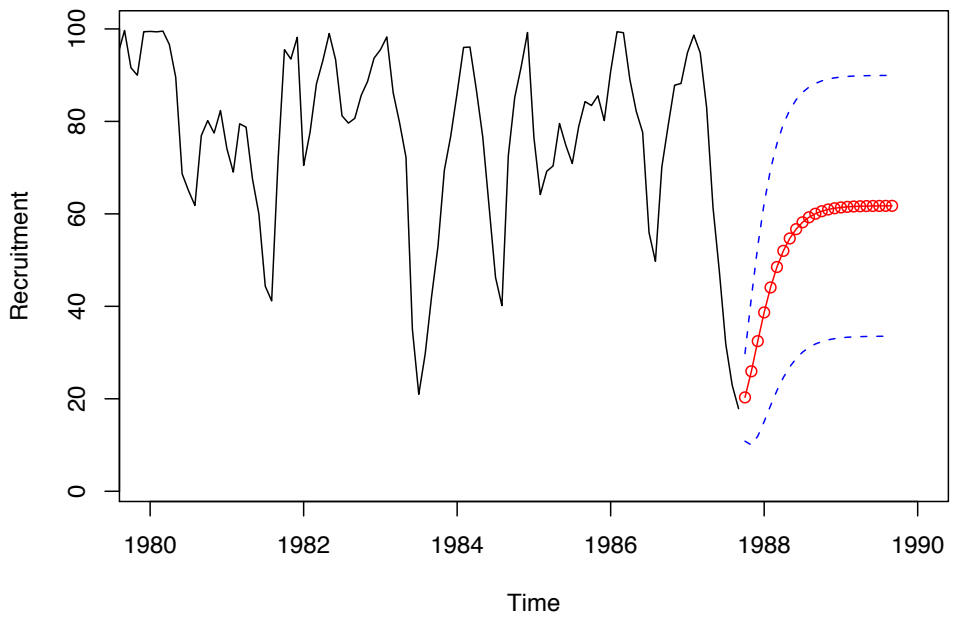

Fig. 3.7. Twenty-four month forecasts for the Recruitment series. The actual data shown are from about January 1980 to September 1987, and then the forecasts plus and minus one standard error are displayed.

$$\widetilde{w}_t^n = x_t - \phi x_{t-1} - \theta \widetilde{w}_{t-1}^n, \quad t = 1, \ldots, n.$$

The approximate forecast variance is computed from (3.86) using the ψ-weights determined as in Example 3.11. In particular, the ψ-weights satisfy $\psi_j = (\phi + \theta)\phi^{j-1}$, for $j \geq 1$. This result gives

$$P_{n+m}^n = \sigma_w^2 \left[1 + (\phi + \theta)^2 \sum_{j=1}^{m-1} \phi^{2(j-1)} \right] = \sigma_w^2 \left[1 + \frac{(\phi + \theta)^2 (1 - \phi^{2(m-1)})}{(1 - \phi^2)} \right].$$

To assess the precision of the forecasts, prediction intervals are typically calculated along with the forecasts. In general, $(1-\alpha)$ prediction intervals are of the form

$$x_{n+m}^n \pm c_{\frac{\alpha}{2}} \sqrt{P_{n+m}^n}, \tag{3.93}$$

where $c_{\alpha/2}$ is chosen to get the desired degree of confidence. For example, if the process is Gaussian, then choosing $c_{\alpha/2} = 2$ will yield an approximate 95% prediction interval for x_{n+m}. If we are interested in establishing prediction intervals over more than one time period, then $c_{\alpha/2}$ should be adjusted appropriately, for example, by using Bonferroni's inequality [see (4.55) in Chapter 4 or Johnson and Wichern, 1992, Chapter 5].

Example 3.24 Forecasting the Recruitment Series

Using the parameter estimates as the actual parameter values, Figure 3.7 shows the result of forecasting the Recruitment series given in Example 3.17

over a 24-month horizon, $m = 1, 2, \ldots, 24$. The actual forecasts are calculated as

$$x^n_{n+m} = 6.74 + 1.35x^n_{n+m-1} - .46x^n_{n+m-2}$$

for $n = 453$ and $m = 1, 2, \ldots, 12$. Recall that $x^s_t = x_t$ when $t \leq s$. The forecasts errors P^n_{n+m} are calculated using (3.86). Recall that $\widehat{\sigma}^2_w = 89.72$, and using (3.40) from Example 3.11, we have $\psi_j = 1.35\psi_{j-1} - .46\psi_{j-2}$ for $j \geq 2$, where $\psi_0 = 1$ and $\psi_1 = 1.35$. Thus, for $n = 453$,

$$\begin{aligned} P^n_{n+1} &= 89.72, \\ P^n_{n+2} &= 89.72(1 + 1.35^2), \\ P^n_{n+3} &= 89.72(1 + 1.35^2 + [1.35^2 - .46]^2), \end{aligned}$$

and so on.

Note how the forecast levels off quickly and the prediction intervals are wide, even though in this case the forecast limits are only based on one standard error; that is, $x^n_{n+m} \pm \sqrt{P^n_{n+m}}$.

To reproduce the analysis and Figure 3.7, use the following commands:

```
regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE)
fore = predict(regr, n.ahead=24)
ts.plot(rec, fore$pred, col=1:2, xlim=c(1980,1990),
    ylab="Recruitment")
lines(fore$pred, type="p", col=2)
lines(fore$pred+fore$se, lty="dashed", col=4)
lines(fore$pred-fore$se, lty="dashed", col=4)
```

We complete this section with a brief discussion of backcasting. In backcasting, we want to predict x_{1-m}, for $m = 1, 2, \ldots$, based on the data $\{x_1, \ldots, x_n\}$. Write the backcast as

$$x^n_{1-m} = \sum_{j=1}^{n} \alpha_j x_j. \tag{3.94}$$

Analogous to (3.74), the prediction equations (assuming $\mu_x = 0$) are

$$\sum_{j=1}^{n} \alpha_j E(x_j x_k) = E(x_{1-m} x_k), \quad k = 1, \ldots, n, \tag{3.95}$$

or

$$\sum_{j=1}^{n} \alpha_j \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \ldots, n. \tag{3.96}$$

These equations are precisely the prediction equations for forward prediction. That is, $\alpha_j \equiv \phi^{(m)}_{nj}$, for $j = 1, \ldots, n$, where the $\phi^{(m)}_{nj}$ are given by (3.75). Finally, the backcasts are given by

$$x^n_{1-m} = \phi^{(m)}_{n1} x_1 + \cdots + \phi^{(m)}_{nn} x_n, \quad m = 1, 2, \ldots. \tag{3.97}$$

Example 3.25 Backcasting an ARMA$(1,1)$
Consider an ARMA$(1,1)$ process, $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$; we will call this the *forward model*. We have just seen that best linear prediction backward in time is the same as best linear prediction forward in time for stationary models. Because we are assuming ARMA models are Gaussian, we also have that minimum mean square error prediction backward in time is the same as forward in time for ARMA models.[4] Thus, the process can equivalently be generated by the *backward model*,

$$x_t = \phi x_{t+1} + \theta v_{t+1} + v_t,$$

where $\{v_t\}$ is a Gaussian white noise process with variance σ_w^2. We may write $x_t = \sum_{j=0}^{\infty} \psi_j v_{t+j}$, where $\psi_0 = 1$; this means that x_t is uncorrelated with $\{v_{t-1}, v_{t-2}, \ldots\}$, in analogy to the forward model.

Given data $\{x_1, \ldots, x_n\}$, truncate $v_n^n = E(v_n \,|\, x_1, \ldots, x_n)$ to zero and then iterate backward. That is, put $\widetilde{v}_n^n = 0$, as an initial approximation, and then generate the errors backward

$$\widetilde{v}_t^n = x_t - \phi x_{t+1} - \theta \widetilde{v}_{t+1}^n, \quad t = (n-1), (n-2), \ldots, 1.$$

Then,

$$\widetilde{x}_0^n = \phi x_1 + \theta \widetilde{v}_1^n + \widetilde{v}_0^n = \phi x_1 + \theta \widetilde{v}_1^n,$$

because $\widetilde{v}_t^n = 0$ for $t \leq 0$. Continuing, the general truncated backcasts are given by

$$\widetilde{x}_{1-m}^n = \phi \widetilde{x}_{2-m}^n, \quad m = 2, 3, \ldots \ .$$

3.6 Estimation

Throughout this section, we assume we have n observations, $x_1, \ldots, x_n$, from a causal and invertible Gaussian ARMA(p,q) process in which, initially, the order parameters, p and q, are known. Our goal is to estimate the parameters, $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, and σ_w^2. We will discuss the problem of determining p and q later in this section.

We begin with method of moments estimators. The idea behind these estimators is that of equating population moments to sample moments and then solving for the parameters in terms of the sample moments. We immediately see that, if $E(x_t) = \mu$, then the method of moments estimator of μ is the sample average, $\bar{x}$. Thus, while discussing method of moments, we will assume $\mu = 0$. Although the method of moments can produce good estimators, they can sometimes lead to suboptimal estimators. We first consider the case

[4] In the stationary Gaussian case, (a) the distribution of $\{x_{n+1}, x_n, \ldots, x_1\}$ is the same as (b) the distribution of $\{x_0, x_1 \ldots, x_n\}$. In forecasting we use (a) to obtain $E(x_{n+1} \,|\, x_n, \ldots, x_1)$; in backcasting we use (b) to obtain $E(x_0 \,|\, x_1, \ldots, x_n)$. Because (a) and (b) are the same, the two problems are equivalent.

in which the method leads to optimal (efficient) estimators, that is, AR(p) models.

When the process is AR(p),

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t,$$

the first $p+1$ equations of (3.47) and (3.48) lead to the following:

Definition 3.10 *The* **Yule–Walker equations** *are given by*

$$\gamma(h) = \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p), \quad h = 1, 2, \ldots, p, \tag{3.98}$$

$$\sigma_w^2 = \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p). \tag{3.99}$$

In matrix notation, the Yule–Walker equations are

$$\Gamma_p\boldsymbol{\phi} = \boldsymbol{\gamma}_p, \quad \sigma_w^2 = \gamma(0) - \boldsymbol{\phi}'\boldsymbol{\gamma}_p, \tag{3.100}$$

where $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p\times p$ matrix, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ is a $p\times 1$ vector, and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.100) by $\widehat{\gamma}(h)$ [see equation (1.34)] and solve

$$\widehat{\boldsymbol{\phi}} = \widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0) - \widehat{\boldsymbol{\gamma}}_p'\widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p. \tag{3.101}$$

These estimators are typically called the Yule–Walker estimators. For calculation purposes, it is sometimes more convenient to work with the sample ACF. By factoring $\widehat{\gamma}(0)$ in (3.101), we can write the Yule–Walker estimates as

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0)\left[1 - \widehat{\boldsymbol{\rho}}_p'\widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p\right], \tag{3.102}$$

where $\widehat{R}_p = \{\widehat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\widehat{\boldsymbol{\rho}}_p = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p))'$ is a $p \times 1$ vector.

For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\widehat{\sigma}_w^2$ is close to the true value of σ_w^2. We state these results in Property 3.8; for details, see Appendix B, §B.3.

Property 3.8 Large Sample Results for Yule–Walker Estimators

The asymptotic ($n \to \infty$) behavior of the Yule–Walker estimators in the case of causal AR(p) processes is as follows:

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_w^2\Gamma_p^{-1}\right), \qquad \widehat{\sigma}_w^2 \xrightarrow{p} \sigma_w^2. \tag{3.103}$$

The Durbin–Levinson algorithm, (3.68)-(3.70), can be used to calculate $\widehat{\boldsymbol{\phi}}$ without inverting $\widehat{\Gamma}_p$ or $\widehat{R}_p$, by replacing $\gamma(h)$ by $\widehat{\gamma}(h)$ in the algorithm. In running the algorithm, we will iteratively calculate the $h \times 1$ vector, $\widehat{\boldsymbol{\phi}}_h = (\widehat{\phi}_{h1}, \ldots, \widehat{\phi}_{hh})'$, for $h = 1, 2, \ldots$. Thus, in addition to obtaining the desired forecasts, the Durbin–Levinson algorithm yields $\widehat{\phi}_{hh}$, the sample PACF. Using (3.103), we can show the following property.

Property 3.9 Large Sample Distribution of the PACF

For a causal AR(p) process, asymptotically ($n \to \infty$),

$$\sqrt{n}\,\widehat{\phi}_{hh} \xrightarrow{d} N(0,1)\,, \quad \text{for} \quad h > p. \tag{3.104}$$

Example 3.26 Yule–Walker Estimation for an AR(2) Process

The data shown in Figure 3.4 were $n = 144$ simulated observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

where $w_t \sim$ iid N$(0,1)$. For these data, $\widehat{\gamma}(0) = 8.903$, $\widehat{\rho}(1) = .849$, and $\widehat{\rho}(2) = .519$. Thus,

$$\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} = \begin{bmatrix} 1 & .849 \\ .849 & 1 \end{bmatrix}^{-1} \begin{pmatrix} .849 \\ .519 \end{pmatrix} = \begin{pmatrix} 1.463 \\ -.723 \end{pmatrix}$$

and

$$\widehat{\sigma}_w^2 = 8.903 \left[1 - (.849, .519) \begin{pmatrix} 1.463 \\ -.723 \end{pmatrix}\right] = 1.187.$$

By Property 3.8, the asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\phi}}$,

$$\frac{1}{144}\frac{1.187}{8.903}\begin{bmatrix} 1 & .849 \\ .849 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} .058^2 & -.003 \\ -.003 & .058^2 \end{bmatrix},$$

can be used to get confidence regions for, or make inferences about $\widehat{\boldsymbol{\phi}}$ and its components. For example, an approximate 95% confidence interval for ϕ_2 is $-.723 \pm 2(.058)$, or $(-.838, -.608)$, which contains the true value of $\phi_2 = -.75$.

For these data, the first three sample partial autocorrelations are $\widehat{\phi}_{11} = \widehat{\rho}(1) = .849$, $\widehat{\phi}_{22} = \widehat{\phi}_2 = -.721$, and $\widehat{\phi}_{33} = -.085$. According to Property 3.9, the asymptotic standard error of $\widehat{\phi}_{33}$ is $1/\sqrt{144} = .083$, and the observed value, $-.085$, is about only one standard deviation from $\phi_{33} = 0$.

Example 3.27 Yule–Walker Estimation of the Recruitment Series

In Example 3.17 we fit an AR(2) model to the recruitment series using regression. Below are the results of fitting the same model using Yule-Walker estimation in R, which are nearly identical to the values in Example 3.17.

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean  # = 62.26 (mean estimate)
rec.yw$ar      # =  1.33, -.44  (parameter estimates)
sqrt(diag(rec.yw$asy.var.coef))  # = .04, .04  (standard errors)
rec.yw$var.pred  # = 94.80 (error variance estimate)
```

To obtain the 24 month ahead predictions and their standard errors, and then plot the results as in Example 3.24, use the R commands:

```
rec.pr = predict(rec.yw, n.ahead=24)
U = rec.pr$pred + rec.pr$se
L = rec.pr$pred - rec.pr$se
minx = min(rec,L); maxx = max(rec,U)
ts.plot(rec, rec.pr$pred, xlim=c(1980,1990), ylim=c(minx,maxx))
lines(rec.pr$pred, col="red", type="o")
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
```

In the case of AR(p) models, the Yule–Walker estimators given in (3.102) are optimal in the sense that the asymptotic distribution, (3.103), is the best asymptotic normal distribution. This is because, given initial conditions, AR(p) models are linear models, and the Yule–Walker estimators are essentially least squares estimators. If we use method of moments for MA or ARMA models, we will not get optimal estimators because such processes are nonlinear in the parameters.

Example 3.28 Method of Moments Estimation for an MA(1)

Consider the time series

$$x_t = w_t + \theta w_{t-1},$$

where $|\theta| < 1$. The model can then be written as

$$x_t = \sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + w_t,$$

which is nonlinear in θ. The first two population autocovariances are $\gamma(0) = \sigma_w^2(1+\theta^2)$ and $\gamma(1) = \sigma_w^2\theta$, so the estimate of θ is found by solving:

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

Two solutions exist, so we would pick the invertible one. If $|\widehat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real, otherwise, a real solution does not exist. Even though $|\rho(1)| < \frac{1}{2}$ for an invertible MA(1), it may happen that $|\widehat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\widehat{\rho}(1) = .507$ when the true value is $\rho(1) = .9/(1+.9^2) = .497$.

```
set.seed(2)
ma1 = arima.sim(list(order = c(0,0,1), ma = 0.9), n = 50)
acf(ma1, plot=F)[1]  # = .507 (lag 1 sample ACF)
```

When $|\widehat{\rho}(1)| < \frac{1}{2}$, the invertible estimate is

$$\widehat{\theta} = \frac{1 - \sqrt{1 - 4\widehat{\rho}(1)^2}}{2\widehat{\rho}(1)}.$$

It can be shown that[5]

[5] The result follows from Theorem A.7 given in Appendix A and the delta method. See the proof of Theorem A.7 for details on the delta method.

$$\widehat{\theta} \sim \text{AN}\left(\theta, \frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{n(1-\theta^2)^2}\right);$$

AN is read *asymptotically normal* and is defined in Definition A.5, page 515, of Appendix A. The maximum likelihood estimator (which we discuss next) of θ, in this case, has an asymptotic variance of $(1-\theta^2)/n$. When $\theta = .5$, for example, the ratio of the asymptotic variance of the method of moments estimator to the maximum likelihood estimator of θ is about 3.5. That is, for large samples, the variance of the method of moments estimator is about 3.5 times larger than the variance of the MLE of θ when $\theta = .5$.

Maximum Likelihood and Least Squares Estimation

To fix ideas, we first focus on the causal AR(1) case. Let

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{3.105}$$

where $|\phi| < 1$ and $w_t \sim \text{iid } \text{N}(0, \sigma_w^2)$. Given data $x_1, x_2, \ldots, x_n$, we seek the likelihood

$$L(\mu, \phi, \sigma_w^2) = f\left(x_1, x_2, \ldots, x_n \mid \mu, \phi, \sigma_w^2\right).$$

In the case of an AR(1), we may write the likelihood as

$$L(\mu, \phi, \sigma_w^2) = f(x_1) f(x_2 \mid x_1) \cdots f(x_n \mid x_{n-1}),$$

where we have dropped the parameters in the densities, $f(\cdot)$, to ease the notation. Because $x_t \mid x_{t-1} \sim \text{N}\left(\mu + \phi(x_{t-1} - \mu), \sigma_w^2\right)$, we have

$$f(x_t \mid x_{t-1}) = f_w[(x_t - \mu) - \phi(x_{t-1} - \mu)],$$

where $f_w(\cdot)$ is the density of w_t, that is, the normal density with mean zero and variance σ_w^2. We may then write the likelihood as

$$L(\mu, \phi, \sigma_w) = f(x_1) \prod_{t=2}^{n} f_w\left[(x_t - \mu) - \phi(x_{t-1} - \mu)\right].$$

To find $f(x_1)$, we can use the causal representation

$$x_1 = \mu + \sum_{j=0}^{\infty} \phi^j w_{1-j}$$

to see that x_1 is normal, with mean μ and variance $\sigma_w^2/(1-\phi^2)$. Finally, for an AR(1), the likelihood is

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-\frac{S(\mu, \phi)}{2\sigma_w^2}\right], \tag{3.106}$$

where

$$S(\mu,\phi) = (1-\phi^2)(x_1-\mu)^2 + \sum_{t=2}^{n}\left[(x_t-\mu)-\phi(x_{t-1}-\mu)\right]^2. \tag{3.107}$$

Typically, $S(\mu,\phi)$ is called the unconditional sum of squares. We could have also considered the estimation of μ and ϕ using unconditional least squares, that is, estimation by minimizing $S(\mu,\phi)$.

Taking the partial derivative of the log of (3.106) with respect to σ_w^2 and setting the result equal to zero, we see that for any given values of μ and ϕ in the parameter space, $\sigma_w^2 = n^{-1}S(\mu,\phi)$ maximizes the likelihood. Thus, the maximum likelihood estimate of σ_w^2 is

$$\widehat{\sigma}_w^2 = n^{-1}S(\widehat{\mu},\widehat{\phi}), \tag{3.108}$$

where $\widehat{\mu}$ and $\widehat{\phi}$ are the MLEs of μ and ϕ, respectively. If we replace n in (3.108) by $n-2$, we would obtain the unconditional least squares estimate of σ_w^2.

If, in (3.106), we take logs, replace σ_w^2 by $\widehat{\sigma}_w^2$, and ignore constants, $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the criterion function

$$l(\mu,\phi) = \log\left[n^{-1}S(\mu,\phi)\right] - n^{-1}\log(1-\phi^2); \tag{3.109}$$

that is, $l(\mu,\phi) \propto -2\log L(\mu,\phi,\widehat{\sigma}_w^2)$.[6] Because (3.107) and (3.109) are complicated functions of the parameters, the minimization of $l(\mu,\phi)$ or $S(\mu,\phi)$ is accomplished numerically. In the case of AR models, we have the advantage that, conditional on initial values, they are linear models. That is, we can drop the term in the likelihood that causes the nonlinearity. Conditioning on x_1, the conditional likelihood becomes

$$\begin{aligned} L(\mu,\phi,\sigma_w^2 \mid x_1) &= \prod_{t=2}^{n} f_w\left[(x_t-\mu)-\phi(x_{t-1}-\mu)\right] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2}\exp\left[-\frac{S_c(\mu,\phi)}{2\sigma_w^2}\right], \end{aligned} \tag{3.110}$$

where the conditional sum of squares is

$$S_c(\mu,\phi) = \sum_{t=2}^{n}\left[(x_t-\mu)-\phi(x_{t-1}-\mu)\right]^2. \tag{3.111}$$

The conditional MLE of σ_w^2 is

$$\widehat{\sigma}_w^2 = S_c(\widehat{\mu},\widehat{\phi})/(n-1), \tag{3.112}$$

and $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the conditional sum of squares, $S_c(\mu,\phi)$. Letting $\alpha = \mu(1-\phi)$, the conditional sum of squares can be written as

[6] The criterion function is sometimes called the profile or concentrated likelihood.

$$S_c(\mu, \phi) = \sum_{t=2}^{n} [x_t - (\alpha + \phi x_{t-1})]^2 . \tag{3.113}$$

The problem is now the linear regression problem stated in §2.2. Following the results from least squares estimation, we have $\widehat{\alpha} = \bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}$, where $\bar{x}_{(1)} = (n-1)^{-1}\sum_{t=1}^{n-1} x_t$, and $\bar{x}_{(2)} = (n-1)^{-1}\sum_{t=2}^{n} x_t$, and the conditional estimates are then

$$\widehat{\mu} = \frac{\bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}}{1 - \widehat{\phi}} \tag{3.114}$$

$$\widehat{\phi} = \frac{\sum_{t=2}^{n}(x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^{n}(x_{t-1} - \bar{x}_{(1)})^2}. \tag{3.115}$$

From (3.114) and (3.115), we see that $\widehat{\mu} \approx \bar{x}$ and $\widehat{\phi} \approx \widehat{\rho}(1)$. That is, the Yule–Walker estimators and the conditional least squares estimators are approximately the same. The only difference is the inclusion or exclusion of terms involving the endpoints, x_1 and x_n. We can also adjust the estimate of σ_w^2 in (3.112) to be equivalent to the least squares estimator, that is, divide $S_c(\widehat{\mu}, \widehat{\phi})$ by $(n-3)$ instead of $(n-1)$ in (3.112).

For general AR(p) models, maximum likelihood estimation, unconditional least squares, and conditional least squares follow analogously to the AR(1) example. For general ARMA models, it is difficult to write the likelihood as an explicit function of the parameters. Instead, it is advantageous to write the likelihood in terms of the innovations, or one-step-ahead prediction errors, $x_t - x_t^{t-1}$. This will also be useful in Chapter 6 when we study state-space models.

For a normal ARMA(p, q) model, let $\boldsymbol{\beta} = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ be the $(p+q+1)$-dimensional vector of the model parameters. The likelihood can be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = \prod_{t=1}^{n} f(x_t \mid x_{t-1}, \ldots, x_1).$$

The conditional distribution of x_t given $x_{t-1}, \ldots, x_1$ is Gaussian with mean x_t^{t-1} and variance P_t^{t-1}. Recall from (3.71) that $P_t^{t-1} = \gamma(0)\prod_{j=1}^{t-1}(1 - \phi_{jj}^2)$. For ARMA models, $\gamma(0) = \sigma_w^2\sum_{j=0}^{\infty}\psi_j^2$, in which case we may write

$$P_t^{t-1} = \sigma_w^2 \left\{ \left[\sum_{j=0}^{\infty}\psi_j^2\right] \left[\prod_{j=1}^{t-1}(1 - \phi_{jj}^2)\right] \right\} \stackrel{\text{def}}{=} \sigma_w^2\, r_t,$$

where r_t is the term in the braces. Note that the r_t terms are functions only of the regression parameters and that they may be computed recursively as $r_{t+1} = (1 - \phi_{tt}^2)r_t$ with initial condition $r_1 = \sum_{j=0}^{\infty}\psi_j^2$. The likelihood of the data can now be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}\,[r_1(\boldsymbol{\beta})r_2(\boldsymbol{\beta})\cdots r_n(\boldsymbol{\beta})]^{-1/2}\exp\left[-\frac{S(\boldsymbol{\beta})}{2\sigma_w^2}\right], \tag{3.116}$$

where

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left[\frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t(\boldsymbol{\beta})} \right]. \tag{3.117}$$

Both x_t^{t-1} and r_t are functions of $\boldsymbol{\beta}$ alone, and we make that fact explicit in (3.116)-(3.117). Given values for $\boldsymbol{\beta}$ and σ_w^2, the likelihood may be evaluated using the techniques of §3.5. Maximum likelihood estimation would now proceed by maximizing (3.116) with respect to $\boldsymbol{\beta}$ and σ_w^2. As in the AR(1) example, we have

$$\widehat{\sigma}_w^2 = n^{-1} S(\widehat{\boldsymbol{\beta}}), \tag{3.118}$$

where $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes the concentrated likelihood

$$l(\boldsymbol{\beta}) = \log \left[n^{-1} S(\boldsymbol{\beta}) \right] + n^{-1} \sum_{t=1}^{n} \log r_t(\boldsymbol{\beta}). \tag{3.119}$$

For the AR(1) model (3.105) discussed previously, recall that $x_1^0 = \mu$ and $x_t^{t-1} = \mu + \phi(x_{t-1} - \mu)$, for $t = 2, \ldots, n$. Also, using the fact that $\phi_{11} = \phi$ and $\phi_{hh} = 0$ for $h > 1$, we have $r_1 = \sum_{j=0}^{\infty} \phi^{2j} = (1 - \phi^2)^{-1}$, $r_2 = (1 - \phi^2)^{-1}(1 - \phi^2) = 1$, and in general, $r_t = 1$ for $t = 2, \ldots, n$. Hence, the likelihood presented in (3.106) is identical to the innovations form of the likelihood given by (3.116). Moreover, the generic $S(\boldsymbol{\beta})$ in (3.117) is $S(\mu, \phi)$ given in (3.107) and the generic $l(\boldsymbol{\beta})$ in (3.119) is $l(\mu, \phi)$ in (3.109).

Unconditional least squares would be performed by minimizing (3.117) with respect to $\boldsymbol{\beta}$. Conditional least squares estimation would involve minimizing (3.117) with respect to $\boldsymbol{\beta}$ but where, to ease the computational burden, the predictions and their errors are obtained by conditioning on initial values of the data. In general, numerical optimization routines are used to obtain the actual estimates and their standard errors.

Example 3.29 The Newton–Raphson and Scoring Algorithms

Two common numerical optimization routines for accomplishing maximum likelihood estimation are Newton–Raphson and scoring. We will give a brief account of the mathematical ideas here. The actual implementation of these algorithms is much more complicated than our discussion might imply. For details, the reader is referred to any of the *Numerical Recipes* books, for example, Press et al. (1993).

Let $l(\boldsymbol{\beta})$ be a criterion function of k parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ that we wish to minimize with respect to $\boldsymbol{\beta}$. For example, consider the likelihood function given by (3.109) or by (3.119). Suppose $l(\widehat{\boldsymbol{\beta}})$ is the extremum that we are interested in finding, and $\widehat{\boldsymbol{\beta}}$ is found by solving $\partial l(\boldsymbol{\beta})/\partial \beta_j = 0$, for $j = 1, \ldots, k$. Let $l^{(1)}(\boldsymbol{\beta})$ denote the $k \times 1$ vector of partials

$$l^{(1)}(\boldsymbol{\beta}) = \left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} \right)'.$$

Note, $l^{(1)}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, the $k \times 1$ zero vector. Let $l^{(2)}(\boldsymbol{\beta})$ denote the $k \times k$ matrix of second-order partials

$$l^{(2)}(\boldsymbol{\beta}) = \left\{ -\frac{\partial l^2(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right\}_{i,j=1}^{k},$$

and assume $l^{(2)}(\boldsymbol{\beta})$ is nonsingular. Let $\boldsymbol{\beta}_{(0)}$ be an initial estimator of $\boldsymbol{\beta}$. Then, using a Taylor expansion, we have the following approximation:

$$\mathbf{0} = l^{(1)}(\widehat{\boldsymbol{\beta}}) \approx l^{(1)}(\boldsymbol{\beta}_{(0)}) - l^{(2)}(\boldsymbol{\beta}_{(0)}) \left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(0)} \right].$$

Setting the right-hand side equal to zero and solving for $\widehat{\boldsymbol{\beta}}$ [call the solution $\boldsymbol{\beta}_{(1)}$], we get

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \left[l^{(2)}(\boldsymbol{\beta}_{(0)}) \right]^{-1} l^{(1)}(\boldsymbol{\beta}_{(0)}).$$

The Newton–Raphson algorithm proceeds by iterating this result, replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ to get $\boldsymbol{\beta}_{(2)}$, and so on, until convergence. Under a set of appropriate conditions, the sequence of estimators, $\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \ldots$, will converge to $\widehat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$.

For maximum likelihood estimation, the criterion function used is $l(\boldsymbol{\beta})$ given by (3.119); $l^{(1)}(\boldsymbol{\beta})$ is called the score vector, and $l^{(2)}(\boldsymbol{\beta})$ is called the Hessian. In the method of scoring, we replace $l^{(2)}(\boldsymbol{\beta})$ by $E[l^{(2)}(\boldsymbol{\beta})]$, the information matrix. Under appropriate conditions, the inverse of the information matrix is the asymptotic variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$. This is sometimes approximated by the inverse of the Hessian at $\widehat{\boldsymbol{\beta}}$. If the derivatives are difficult to obtain, it is possible to use quasi-maximum likelihood estimation where numerical techniques are used to approximate the derivatives.

Example 3.30 MLE for the Recruitment Series

So far, we have fit an AR(2) model to the Recruitment series using ordinary least squares (Example 3.17) and using Yule–Walker (Example 3.27). The following is an R session used to fit an AR(2) model via maximum likelihood estimation to the Recruitment series; these results can be compared to the results in Example 3.17 and Example 3.27.

```
rec.mle = ar.mle(rec, order=2)
rec.mle$x.mean   # 62.26
rec.mle$ar       # 1.35, -.46
sqrt(diag(rec.mle$asy.var.coef))  # .04, .04
rec.mle$var.pred     # 89.34
```

We now discuss least squares for ARMA(p, q) models via Gauss–Newton. For general and complete details of the Gauss–Newton procedure, the reader is referred to Fuller (1996). As before, write $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$, and

for the ease of discussion, we will put $\mu = 0$. We write the model in terms of the errors

$$w_t(\boldsymbol{\beta}) = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}(\boldsymbol{\beta}), \tag{3.120}$$

emphasizing the dependence of the errors on the parameters.

For conditional least squares, we approximate the residual sum of squares by conditioning on $x_1, \ldots, x_p$ (if $p > 0$) and $w_p = w_{p-1} = w_{p-2} = \cdots = w_{1-q} = 0$ (if $q > 0$), in which case, given $\boldsymbol{\beta}$, we may evaluate (3.120) for $t = p+1, p+2, \ldots, n$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} w_t^2(\boldsymbol{\beta}). \tag{3.121}$$

Minimizing $S_c(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the conditional least squares estimates. If $q = 0$, the problem is linear regression and no iterative technique is needed to minimize $S_c(\phi_1, \ldots, \phi_p)$. If $q > 0$, the problem becomes nonlinear regression and we will have to rely on numerical optimization.

When n is large, conditioning on a few initial values will have little influence on the final parameter estimates. In the case of small to moderate sample sizes, one may wish to rely on unconditional least squares. The unconditional least squares problem is to choose $\boldsymbol{\beta}$ to minimize the unconditional sum of squares, which we have generically denoted by $S(\boldsymbol{\beta})$ in this section. The unconditional sum of squares can be written in various ways, and one useful form in the case of ARMA(p, q) models is derived in Box et al. (1994, Appendix A7.3). They showed (see Problem 3.19) the unconditional sum of squares can be written as

$$S(\boldsymbol{\beta}) = \sum_{t=-\infty}^{n} \widehat{w}_t^2(\boldsymbol{\beta}), \tag{3.122}$$

where $\widehat{w}_t(\boldsymbol{\beta}) = E(w_t \,|\, x_1, \ldots, x_n)$. When $t \leq 0$, the $\widehat{w}_t(\boldsymbol{\beta})$ are obtained by backcasting. As a practical matter, we approximate $S(\boldsymbol{\beta})$ by starting the sum at $t = -M + 1$, where M is chosen large enough to guarantee $\sum_{t=-\infty}^{-M} \widehat{w}_t^2(\boldsymbol{\beta}) \approx 0$. In the case of unconditional least squares estimation, a numerical optimization technique is needed even when $q = 0$.

To employ Gauss–Newton, let $\boldsymbol{\beta}_{(0)} = (\phi_1^{(0)}, \ldots, \phi_p^{(0)}, \theta_1^{(0)}, \ldots, \theta_q^{(0)})'$ be an initial estimate of $\boldsymbol{\beta}$. For example, we could obtain $\boldsymbol{\beta}_{(0)}$ by method of moments. The first-order Taylor expansion of $w_t(\boldsymbol{\beta})$ is

$$w_t(\boldsymbol{\beta}) \approx w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}), \tag{3.123}$$

where

$$\boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) = \left(-\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_1}, \ldots, -\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_{p+q}}\right)', \quad t = 1, \ldots, n.$$

The linear approximation of $S_c(\boldsymbol{\beta})$ is

$$Q(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} \left[w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) \right]^2 \tag{3.124}$$

and this is the quantity that we will minimize. For approximate unconditional least squares, we would start the sum in (3.124) at $t = -M + 1$, for a large value of M, and work with the backcasted values.

Using the results of ordinary least squares (§2.2), we know

$$\widehat{(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)})} = \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) \boldsymbol{z}_t'(\boldsymbol{\beta}_{(0)}) \right)^{-1} \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) w_t(\boldsymbol{\beta}_{(0)}) \right) \tag{3.125}$$

minimizes $Q(\boldsymbol{\beta})$. From (3.125), we write the one-step Gauss–Newton estimate as

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \Delta(\boldsymbol{\beta}_{(0)}), \tag{3.126}$$

where $\Delta(\boldsymbol{\beta}_{(0)})$ denotes the right-hand side of (3.125). Gauss–Newton estimation is accomplished by replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ in (3.126). This process is repeated by calculating, at iteration $j = 2, 3, \ldots$,

$$\boldsymbol{\beta}_{(j)} = \boldsymbol{\beta}_{(j-1)} + \Delta(\boldsymbol{\beta}_{(j-1)})$$

until convergence.

Example 3.31 Gauss–Newton for an MA(1)

Consider an invertible MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the truncated errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{3.127}$$

where we condition on $w_0(\theta) = 0$. Taking derivatives,

$$-\frac{\partial w_t(\theta)}{\partial \theta} = w_{t-1}(\theta) + \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \ldots, n, \tag{3.128}$$

where $\partial w_0(\theta)/\partial\theta = 0$. Using the notation of (3.123), we can also write (3.128) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{3.129}$$

where $z_0(\theta) = 0$.

Let $\theta_{(0)}$ be an initial estimate of θ, for example, the estimate given in Example 3.28. Then, the Gauss–Newton procedure for conditional least squares is given by

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^{n} z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^{n} z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \ldots, \tag{3.130}$$

where the values in (3.130) are calculated recursively using (3.127) and (3.129). The calculations are stopped when $|\theta_{(j+1)} - \theta_{(j)}|$, or $|Q(\theta_{(j+1)}) - Q(\theta_{(j)})|$, are smaller than some preset amount.

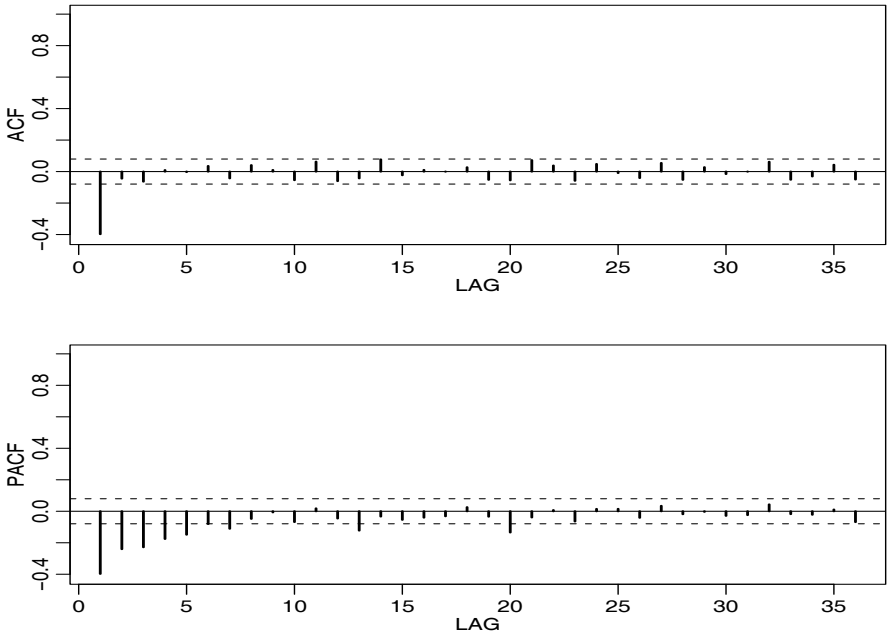

Fig. 3.8. ACF and PACF of transformed glacial varves.

Example 3.32 Fitting the Glacial Varve Series

Consider the series of glacial varve thicknesses from Massachusetts for $n = 634$ years, as analyzed in Example 2.6 and in Problem 2.8, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}) = \log\left(\frac{x_t}{x_{t-1}}\right),$$

which can be interpreted as being approximately the percentage change in the thickness.

The sample ACF and PACF, shown in Figure 3.8, confirm the tendency of $\nabla \log(x_t)$ to behave as a first-order moving average process as the ACF has only a significant peak at lag one and the PACF decreases exponentially. Using Table 3.1, this sample behavior fits that of the MA(1) very well.

The results of eleven iterations of the Gauss–Newton procedure, (3.130), starting with $\theta_{(0)} = -.10$ are given in Table 3.2. The final estimate is $\widehat{\theta} = \theta_{(11)} = -.773$; interim values and the corresponding value of the conditional sum of squares, $S_c(\theta)$ given in (3.121), are also displayed in the table. The final estimate of the error variance is $\widehat{\sigma}_w^2 = 148.98/632 = .236$ with 632 degrees of freedom (one is lost in differencing). The value of the sum of the squared derivatives at convergence is $\sum_{t=1}^{n} z_t^2(\theta_{(11)}) = 369.73$, and con-

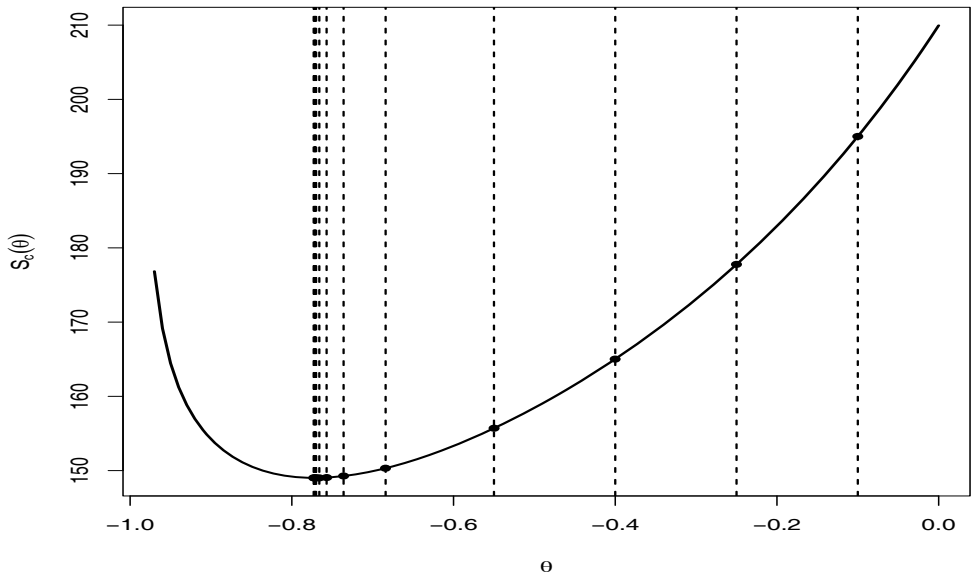

Fig. 3.9. Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 3.32. Vertical lines indicate the values of the parameter obtained via Gauss–Newton; see Table 3.2 for the actual values.

Table 3.2. Gauss–Newton Results for Example 3.32

j	$\theta_{(j)}$	$S_c(\theta_{(j)})$	$\sum_{t=1}^{n} z_t^2(\theta_{(j)})$
0	−0.100	195.0010	183.3464
1	−0.250	177.7614	163.3038
2	−0.400	165.0027	161.6279
3	−0.550	155.6723	182.6432
4	−0.684	150.2896	247.4942
5	−0.736	149.2283	304.3125
6	−0.757	149.0272	337.9200
7	−0.766	148.9885	355.0465
8	−0.770	148.9812	363.2813
9	−0.771	148.9804	365.4045
10	−0.772	148.9799	367.5544
11	−0.773	148.9799	369.7314

sequently, the estimated standard error of $\widehat{\theta}$ is $\sqrt{.236/369.73} = .025$;[7] this leads to a t-value of $-.773/.025 = -30.92$ with 632 degrees of freedom.

Figure 3.9 displays the conditional sum of squares, $S_c(\theta)$ as a function of θ, as well as indicating the values of each step of the Gauss–Newton algorithm. Note that the Gauss–Newton procedure takes large steps toward

[7] To estimate the standard error, we are using the standard regression results from (2.9) as an approximation

the minimum initially, and then takes very small steps as it gets close to the minimizing value. When there is only one parameter, as in this case, it would be easy to evaluate $S_c(\theta)$ on a grid of points, and then choose the appropriate value of θ from the grid search. It would be difficult, however, to perform grid searches when there are many parameters.

In the general case of causal and invertible ARMA(p,q) models, maximum likelihood estimation and conditional and unconditional least squares estimation (and Yule–Walker estimation in the case of AR models) all lead to optimal estimators. The proof of this general result can be found in a number of texts on theoretical time series analysis (for example, Brockwell and Davis, 1991, or Hannan, 1970, to mention a few). We will denote the ARMA coefficient parameters by $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$.

Property 3.10 Large Sample Distribution of the Estimators

Under appropriate conditions, for causal and invertible ARMA processes, the maximum likelihood, the unconditional least squares, and the conditional least squares estimators, each initialized by the method of moments estimator, all provide optimal estimators of σ_w^2 and $\boldsymbol{\beta}$, in the sense that $\widehat{\sigma}_w^2$ is consistent, and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the best asymptotic normal distribution. In particular, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_w^2 \, \boldsymbol{\Gamma}_{p,q}^{-1}\right). \tag{3.131}$$

The asymptotic variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$ is the inverse of the information matrix. In particular, the $(p+q)\times(p+q)$ matrix $\Gamma_{p,q}$, has the form

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}. \tag{3.132}$$

The $p \times p$ matrix $\Gamma_{\phi\phi}$ is given by (3.100), that is, the ij-th element of $\Gamma_{\phi\phi}$, for $i,j = 1,\ldots,p$, is $\gamma_x(i-j)$ from an AR(p) process, $\phi(B)x_t = w_t$. Similarly, $\Gamma_{\theta\theta}$ is a $q \times q$ matrix with the ij-th element, for $i,j = 1,\ldots,q$, equal to $\gamma_y(i-j)$ from an AR(q) process, $\theta(B)y_t = w_t$. The $p \times q$ matrix $\Gamma_{\phi\theta} = \{\gamma_{xy}(i-j)\}$, for $i = 1,\ldots,p$; $j = 1,\ldots,q$; that is, the ij-th element is the cross-covariance between the two AR processes given by $\phi(B)x_t = w_t$ and $\theta(B)y_t = w_t$. Finally, $\Gamma_{\theta\phi} = \Gamma_{\phi\theta}'$ is $q \times p$.

Further discussion of Property 3.10, including a proof for the case of least squares estimators for AR(p) processes, can be found in Appendix B, §B.3.

Example 3.33 Some Specific Asymptotic Distributions

The following are some specific cases of Property 3.10.

AR(1): $\gamma_x(0) = \sigma_w^2/(1-\phi^2)$, so $\sigma_w^2 \Gamma_{1,0}^{-1} = (1-\phi^2)$. Thus,

$$\widehat{\phi} \sim \text{AN}\left[\phi, n^{-1}(1-\phi^2)\right]. \tag{3.133}$$

AR(2): The reader can verify that

$$\gamma_x(0) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_w^2}{(1-\phi_2)^2-\phi_1^2}$$

and $\gamma_x(1) = \phi_1\gamma_x(0) + \phi_2\gamma_x(1)$. From these facts, we can compute $\Gamma_{2,0}^{-1}$. In particular, we have

$$\begin{pmatrix}\widehat{\phi}_1\\ \widehat{\phi}_2\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix},\ n^{-1}\begin{pmatrix}1-\phi_2^2 & -\phi_1(1+\phi_2)\\ \text{sym} & 1-\phi_2^2\end{pmatrix}\right]. \tag{3.134}$$

MA(1): In this case, write $\theta(B)y_t = w_t$, or $y_t+\theta y_{t-1} = w_t$. Then, analogous to the AR(1) case, $\gamma_y(0) = \sigma_w^2/(1-\theta^2)$, so $\sigma_w^2\Gamma_{0,1}^{-1} = (1-\theta^2)$. Thus,

$$\widehat{\theta} \sim \text{AN}\left[\theta, n^{-1}(1-\theta^2)\right]. \tag{3.135}$$

MA(2): Write $y_t+\theta_1 y_{t-1}+\theta_2 y_{t-2} = w_t$, so , analogous to the AR(2) case, we have

$$\begin{pmatrix}\widehat{\theta}_1\\ \widehat{\theta}_2\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix},\ n^{-1}\begin{pmatrix}1-\theta_2^2 & \theta_1(1+\theta_2)\\ \text{sym} & 1-\theta_2^2\end{pmatrix}\right]. \tag{3.136}$$

ARMA(1,1): To calculate $\Gamma_{\phi\theta}$, we must find $\gamma_{xy}(0)$, where $x_t-\phi x_{t-1} = w_t$ and $y_t+\theta y_{t-1} = w_t$. We have

$$\begin{aligned}\gamma_{xy}(0) &= \text{cov}(x_t, y_t) = \text{cov}(\phi x_{t-1}+w_t, -\theta y_{t-1}+w_t)\\ &= -\phi\theta\gamma_{xy}(0) + \sigma_w^2.\end{aligned}$$

Solving, we find, $\gamma_{xy}(0) = \sigma_w^2/(1+\phi\theta)$. Thus,

$$\begin{pmatrix}\widehat{\phi}\\ \widehat{\theta}\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\phi\\ \theta\end{pmatrix},\ n^{-1}\begin{bmatrix}(1-\phi^2)^{-1} & (1+\phi\theta)^{-1}\\ \text{sym} & (1-\theta^2)^{-1}\end{bmatrix}^{-1}\right]. \tag{3.137}$$

Example 3.34 Overfitting Caveat

The asymptotic behavior of the parameter estimators gives us an additional insight into the problem of fitting ARMA models to data. For example, suppose a time series follows an AR(1) process and we decide to fit an AR(2) to the data. Do any problems occur in doing this? More generally, why not simply fit large-order AR models to make sure that we capture the dynamics of the process? After all, if the process is truly an AR(1), the other autoregressive parameters will not be significant. The answer is that if we overfit, we obtain less efficient, or less precise parameter estimates. For example, if we fit an AR(1) to an AR(1) process, for large n, $\text{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large n, $\text{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$

because $\phi_2 = 0$. Thus, the variance of ϕ_1 has been inflated, making the estimator less precise.

We do want to mention, however, that overfitting can be used as a diagnostic tool. For example, if we fit an AR(2) model to the data and are satisfied with that model, then adding one more parameter and fitting an AR(3) should lead to approximately the same model as in the AR(2) fit. We will discuss model diagnostics in more detail in §3.8.

The reader might wonder, for example, why the asymptotic distributions of $\widehat{\phi}$ from an AR(1) and $\widehat{\theta}$ from an MA(1) are of the same form; compare (3.133) to (3.135). It is possible to explain this unexpected result heuristically using the intuition of linear regression. That is, for the normal regression model presented in §2.2 with no intercept term, $x_t = \beta z_t + w_t$, we know $\widehat{\beta}$ is normally distributed with mean β, and from (2.9),

$$\mathrm{var}\left\{\sqrt{n}\left(\widehat{\beta}-\beta\right)\right\} = n\sigma_w^2\left(\sum_{t=1}^{n} z_t^2\right)^{-1} = \sigma_w^2\left(n^{-1}\sum_{t=1}^{n} z_t^2\right)^{-1}.$$

For the causal AR(1) model given by $x_t = \phi x_{t-1} + w_t$, the intuition of regression tells us to expect that, for n large,

$$\sqrt{n}\left(\widehat{\phi}-\phi\right)$$

is approximately normal with mean zero and with variance given by

$$\sigma_w^2\left(n^{-1}\sum_{t=2}^{n} x_{t-1}^2\right)^{-1}.$$

Now, $n^{-1}\sum_{t=2}^{n} x_{t-1}^2$ is the sample variance (recall that the mean of x_t is zero) of the x_t, so as n becomes large we would expect it to approach $\mathrm{var}(x_t) = \gamma(0) = \sigma_w^2/(1-\phi^2)$. Thus, the large sample variance of $\sqrt{n}\left(\widehat{\phi}-\phi\right)$ is

$$\sigma_w^2\gamma_x(0)^{-1} = \sigma_w^2\left(\frac{\sigma_w^2}{1-\phi^2}\right)^{-1} = (1-\phi^2);$$

that is, (3.133) holds.

In the case of an MA(1), we may use the discussion of Example 3.31 to write an approximate regression model for the MA(1). That is, consider the approximation (3.129) as the regression model

$$z_t(\widehat{\theta}) = -\theta z_{t-1}(\widehat{\theta}) + w_{t-1},$$

where now, $z_{t-1}(\widehat{\theta})$ as defined in Example 3.31, plays the role of the regressor. Continuing with the analogy, we would expect the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}-\theta\right)$ to be normal, with mean zero, and approximate variance

$$\sigma_w^2 \left(n^{-1} \sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta}) \right)^{-1}.$$

As in the AR(1) case, $n^{-1}\sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta})$ is the sample variance of the $z_t(\widehat{\theta})$ so, for large n, this should be $\mathrm{var}\{z_t(\theta)\} = \gamma_z(0)$, say. But note, as seen from (3.129), $z_t(\theta)$ is approximately an AR(1) process with parameter $-\theta$. Thus,

$$\sigma_w^2 \gamma_z(0)^{-1} = \sigma_w^2 \left(\frac{\sigma_w^2}{1-(-\theta)^2} \right)^{-1} = (1-\theta^2),$$

which agrees with (3.135). Finally, the asymptotic distributions of the AR parameter estimates and the MA parameter estimates are of the same form because in the MA case, the "regressors" are the differential processes $z_t(\theta)$ that have AR structure, and it is this structure that determines the asymptotic variance of the estimators. For a rigorous account of this approach for the general case, see Fuller (1996, Theorem 5.5.4).

In Example 3.32, the estimated standard error of $\widehat{\theta}$ was .025. In that example, we used regression results to estimate the standard error as the square root of

$$n^{-1}\widehat{\sigma}_w^2 \left(n^{-1} \sum_{t=1}^{n} z_t^2(\widehat{\theta}) \right)^{-1} = \frac{\widehat{\sigma}_w^2}{\sum_{t=1}^{n} z_t^2(\widehat{\theta})},$$

where $n = 632$, $\widehat{\sigma}_w^2 = .236$, $\sum_{t=1}^{n} z_t^2(\widehat{\theta}) = 369.73$ and $\widehat{\theta} = -.773$. Using (3.135), we could have also calculated this value using the asymptotic approximation, the square root of $(1-(-.773)^2)/632$, which is also .025.

If n is small, or if the parameters are close to the boundaries, the asymptotic approximations can be quite poor. The bootstrap can be helpful in this case; for a broad treatment of the bootstrap, see Efron and Tibshirani (1994). We discuss the case of an AR(1) here and leave the general discussion for Chapter 6. For now, we give a simple example of the bootstrap for an AR(1) process.

Example 3.35 Bootstrapping an AR(1)

We consider an AR(1) model with a regression coefficient near the boundary of causality and an error process that is symmetric but not normal. Specifically, consider the causal model

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t, \tag{3.138}$$

where $\mu = 50$, $\phi = .95$, and w_t are iid double exponential with location zero, and scale parameter $\beta = 2$. The density of w_t is given by

$$f(w) = \frac{1}{2\beta} \exp\{-|w|/\beta\} \quad -\infty < w < \infty.$$

In this example, $E(w_t) = 0$ and $\mathrm{var}(w_t) = 2\beta^2 = 8$. Figure 3.10 shows $n = 100$ simulated observations from this process. This particular realization

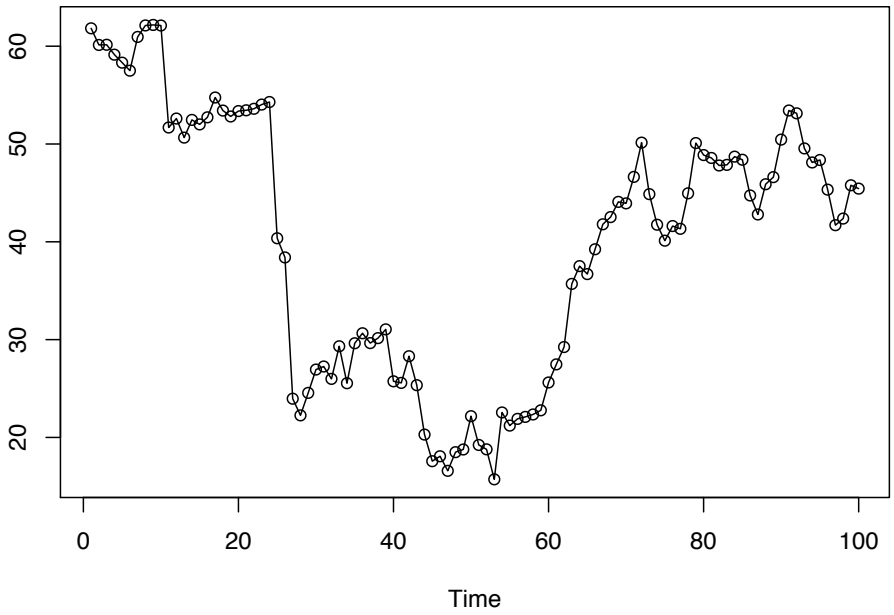

Fig. 3.10. One hundred observations generated from the model in Example 3.35.

is interesting; the data look like they were generated from a nonstationary process with three different mean levels. In fact, the data were generated from a well-behaved, albeit non-normal, stationary and causal model. To show the advantages of the bootstrap, we will act as if we do not know the actual error distribution and we will proceed as if it were normal; of course, this means, for example, that the normal based MLE of ϕ will not be the actual MLE because the data are not normal.

Using the data shown in Figure 3.10, we obtained the Yule–Walker estimates $\widehat{\mu} = 40.05$, $\widehat{\phi} = .96$, and $s_w^2 = 15.30$, where s_w^2 is the estimate of $\text{var}(w_t)$. Based on Property 3.10, we would say that $\widehat{\phi}$ is approximately normal with mean ϕ (which we supposedly do not know) and variance $(1-\phi^2)/100$, which we would approximate by $(1-.96^2)/100 = .03^2$.

To assess the finite sample distribution of $\widehat{\phi}$ when $n = 100$, we simulated 1000 realizations of this AR(1) process and estimated the parameters via Yule–Walker. The finite sampling density of the Yule–Walker estimate of ϕ, based on the 1000 repeated simulations, is shown in Figure 3.11. Clearly the sampling distribution is not close to normality for this sample size. The mean of the distribution shown in Figure 3.11 is .89, and the variance of the distribution is $.05^2$; these values are considerably different than the asymptotic values. Some of the quantiles of the finite sample distribution are .79 (5%), .86 (25%), .90 (50%), .93 (75%), and .95 (95%). The R code to perform the simulation and plot the histogram is as follows:

```
set.seed(111)
phi.yw = rep(NA, 1000)
for (i in 1:1000){
```

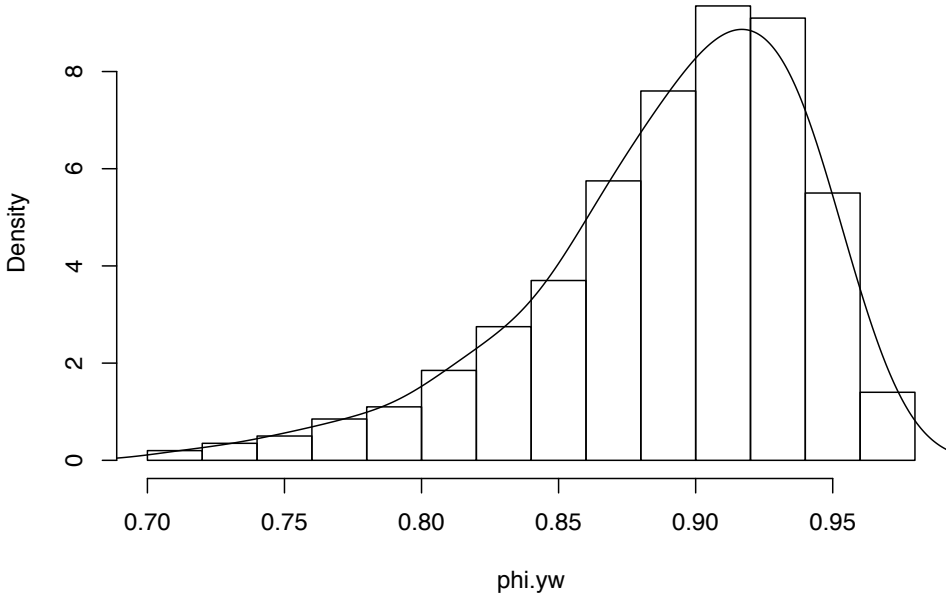

Fig. 3.11. Finite sample density of the Yule–Walker estimate of ϕ in Example 3.35.

```
  e = rexp(150, rate=.5); u = runif(150,-1,1); de = e*sign(u)
  x = 50 + arima.sim(n=100,list(ar=.95), innov=de, n.start=50)
  phi.yw[i] = ar.yw(x, order=1)$ar }
hist(phi.yw, prob=TRUE, main="")
lines(density(phi.yw, bw=.015))
```

Before discussing the bootstrap, we first investigate the sample innovation process, $x_t - x_t^{t-1}$, with corresponding variances P_t^{t-1}. For the AR(1) model in this example,

$$x_t^{t-1} = \mu + \phi(x_{t-1} - \mu), \qquad t = 2, \ldots, 100.$$

From this, it follows that

$$P_t^{t-1} = E(x_t - x_t^{t-1})^2 = \sigma_w^2, \qquad t = 2, \ldots, 100.$$

When $t = 1$, we have

$$x_1^0 = \mu \quad \text{and} \quad P_1^0 = \sigma_w^2/(1-\phi^2).$$

Thus, the innovations have zero mean but different variances; in order that all of the innovations have the same variance, σ_w^2, we will write them as

$$\begin{aligned} \epsilon_1 &= (x_1 - \mu)\sqrt{(1-\phi^2)} \\ \epsilon_t &= (x_t - \mu) - \phi(x_{t-1} - \mu), \quad \text{for} \quad t = 2, \ldots, 100. \end{aligned} \tag{3.139}$$

From these equations, we can write the model in terms of the ϵ_t as

$$x_1 = \mu + \epsilon_1/\sqrt{(1-\phi^2)}$$
$$x_t = \mu + \phi(x_{t-1} - \mu) + \epsilon_t \quad \text{for} \quad t = 2, \ldots, 100. \tag{3.140}$$

Next, replace the parameters with their estimates in (3.139), that is, $\widehat{\mu} = 40.048$ and $\widehat{\phi} = .957$, and denote the resulting sample innovations as $\{\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_{100}\}$. To obtain one bootstrap sample, first randomly sample, with replacement, $n = 100$ values from the set of sample innovations; call the sampled values $\{\epsilon_1^*, \ldots, \epsilon_{100}^*\}$. Now, generate a bootstrapped data set sequentially by setting

$$x_1^* = 40.048 + \epsilon_1^*/\sqrt{(1-.957^2)}$$
$$x_t^* = 40.048 + .957(x_{t-1}^* - 40.048) + \epsilon_t^*, \quad t = 2, \ldots, n. \tag{3.141}$$

Next, estimate the parameters as if the data were x_t^*. Call these estimates $\widehat{\mu}(1)$, $\widehat{\phi}(1)$, and $s_w^2(1)$. Repeat this process a large number, B, of times, generating a collection of bootstrapped parameter estimates, $\{\widehat{\mu}(b), \widehat{\phi}(b), s_w^2(b), b = 1, \ldots, B\}$. We can then approximate the finite sample distribution of an estimator from the bootstrapped parameter values. For example, we can approximate the distribution of $\widehat{\phi} - \phi$ by the empirical distribution of $\widehat{\phi}(b) - \widehat{\phi}$, for $b = 1, \ldots, B$.

Figure 3.12 shows the bootstrap histogram of 200 bootstrapped estimates of ϕ using the data shown in Figure 3.10. In addition, Figure 3.12 shows a density estimate based on the bootstrap histogram, as well as the asymptotic normal density that would have been used based on Property 3.10. Clearly, the bootstrap distribution of $\widehat{\phi}$ is closer to the distribution of $\widehat{\phi}$ shown in Figure 3.11 than to the asymptotic normal approximation. In particular, the mean of the distribution of $\widehat{\phi}(b)$ is .92 with a variance of $.05^2$. Some quantiles of this distribution are .83 (5%), .90 (25%), .93 (50%), .95 (75%), and .98 (95%).

To perform a similar bootstrap exercise in R, use the following commands. We note that the R estimation procedure is conditional on the first observation, so the first residual is not returned. To get around this problem, we simply fix the first observation and bootstrap the remaining data. The simulated data are available in the file `ar1boot`, but you can simulate your own data as was done in the code that produced Figure 3.11.

```
x = ar1boot
m = mean(x)   # estimate of mu
fit = ar.yw(x, order=1)
phi = fit$ar  # estimate of phi
nboot = 200   # number of bootstrap replicates
resids = fit$resid[-1]  # the first resid is NA
x.star = x    # initialize x*
phi.star.yw = rep(NA, nboot)
for (i in 1:nboot) {
  resid.star = sample(resids, replace=TRUE)
```

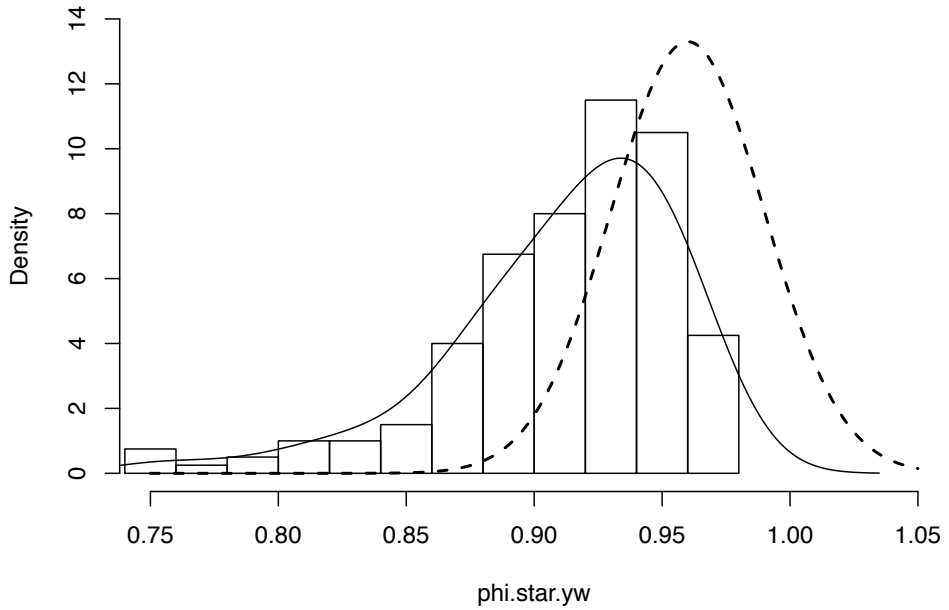

Fig. 3.12. Bootstrap histogram of $\widehat{\phi}$ based on 200 bootstraps; a density estimate based on the histogram (solid line) and the corresponding asymptotic normal density (dashed line).

```
  for (t in 1:99){ x.star[t+1] = m + phi*(x.star[t]-m) +
    resid.star[t] }
  phi.star.yw[i] = ar.yw(x.star, order=1)$ar }
hist(phi.star.yw, 10, main="", prob=TRUE, ylim=c(0,14),
    xlim=c(.75,1.05))
lines(density(phi.star.yw, bw=.02))
u = seq(.75, 1.05, by=.001)
lines(u, dnorm(u, mean=.96, sd=.03), lty="dashed", lwd=2)
```

3.7 Integrated Models for Nonstationary Data

In Chapters 1 and 2, we saw that if x_t is a random walk, $x_t = x_{t-1} + w_t$, then by differencing x_t, we find that $\nabla x_t = w_t$ is stationary. In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. For example, in §2.2 we considered the model

$$x_t = \mu_t + y_t, \tag{3.142}$$

where $\mu_t = \beta_0 + \beta_1 t$ and y_t is stationary. Differencing such a process will lead to a stationary process:

$$\nabla x_t = x_t - x_{t-1} = \beta_1 + y_t - y_{t-1} = \beta_1 + \nabla y_t.$$

Another model that leads to first differencing is the case in which μ_t in (3.142) is stochastic and slowly varying according to a random walk. That is,

$$\mu_t = \mu_{t-1} + v_t$$

where v_t is stationary. In this case,

$$\nabla x_t = v_t + \nabla y_t,$$

is stationary. If μ_t in (3.142) is a k-th order polynomial, $\mu_t = \sum_{j=0}^{k} \beta_j t^j$, then (Problem 3.27) the differenced series $\nabla^k y_t$ is stationary. Stochastic trend models can also lead to higher order differencing. For example, suppose

$$\mu_t = \mu_{t-1} + v_t \quad \text{and} \quad v_t = v_{t-1} + e_t,$$

where e_t is stationary. Then, $\nabla x_t = v_t + \nabla y_t$ is not stationary, but

$$\nabla^2 x_t = e_t + \nabla^2 y_t$$

is stationary.

The integrated ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing.

Definition 3.11 *A process x_t is said to be* **ARIMA(p, d, q)** *if*

$$\nabla^d x_t = (1 - B)^d x_t$$

is ARMA(p, q). In general, we will write the model as

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t. \tag{3.143}$$

If $E(\nabla^d x_t) = \mu$, we write the model as

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B) w_t,$$

where $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$.

Because of the nonstationarity, care must be taken when deriving forecasts. For the sake of completeness, we discuss this issue briefly here, but we stress the fact that both the theoretical and computational aspects of the problem are best handled via state-space models. We discuss the theoretical details in Chapter 6. For information on the state-space based computational aspects in R, see the ARIMA help files (`?arima` and `?predict.Arima`); our scripts `sarima` and `sarima.for` are basically front ends for these R scripts.

It should be clear that, since $y_t = \nabla^d x_t$ is ARMA, we can use §3.5 methods to obtain forecasts of y_t, which in turn lead to forecasts for x_t. For example, if $d = 1$, given forecasts y_{n+m}^n for $m = 1, 2, \ldots$, we have $y_{n+m}^n = x_{n+m}^n - x_{n+m-1}^n$, so that

$$x^n_{n+m} = y^n_{n+m} + x^n_{n+m-1}$$

with initial condition $x^n_{n+1} = y^n_{n+1} + x_n$ (noting $x^n_n = x_n$).

It is a little more difficult to obtain the prediction errors P^n_{n+m}, but for large n, the approximation used in §3.5, equation (3.86), works well. That is, the mean-squared prediction error can be approximated by

$$P^n_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi^{*2}_j, \tag{3.144}$$

where ψ^*_j is the coefficient of z^j in $\psi^*(z) = \theta(z)/\phi(z)(1-z)^d$.

To better understand integrated models, we examine the properties of some simple cases; Problem 3.29 covers the ARIMA$(1,1,0)$ case.

Example 3.36 Random Walk with Drift

To fix ideas, we begin by considering the random walk with drift model first presented in Example 1.11, that is,

$$x_t = \delta + x_{t-1} + w_t,$$

for $t = 1, 2, \ldots$, and $x_0 = 0$. Technically, the model is not ARIMA, but we could include it trivially as an ARIMA$(0,1,0)$ model. Given data $x_1, \ldots, x_n$, the one-step-ahead forecast is given by

$$x^n_{n+1} = E(x_{n+1} \mid x_n, \ldots, x_1) = E(\delta + x_n + w_{n+1} \mid x_n, \ldots, x_1) = \delta + x_n.$$

The two-step-ahead forecast is given by $x^n_{n+2} = \delta + x^n_{n+1} = 2\delta + x_n$, and consequently, the m-step-ahead forecast, for $m = 1, 2, \ldots$, is

$$x^n_{n+m} = m\,\delta + x_n, \tag{3.145}$$

To obtain the forecast errors, it is convenient to recall equation (1.4), i.e., $x_n = n\,\delta + \sum_{j=1}^n w_j$, in which case we may write

$$x_{n+m} = (n+m)\,\delta + \sum_{j=1}^{n+m} w_j = m\,\delta + x_n + \sum_{j=n+1}^{n+m} w_j.$$

From this it follows that the m-step-ahead prediction error is given by

$$P^n_{n+m} = E(x_{n+m} - x^n_{n+m})^2 = E\Big(\sum_{j=n+1}^{n+m} w_j\Big)^2 = m\,\sigma_w^2. \tag{3.146}$$

Hence, unlike the stationary case (see Example 3.22), as the forecast horizon grows, the prediction errors, (3.146), increase without bound and the forecasts follow a straight line with slope δ emanating from x_n. We note that

(3.144) is exact in this case because $\psi^*(z) = 1/(1-z) = \sum_{j=0}^{\infty} z^j$ for $|z| < 1$, so that $\psi_j^* = 1$ for all j.

The w_t are Gaussian, so estimation is straightforward because the differenced data, say $y_t = \nabla x_t$, are independent and identically distributed normal variates with mean δ and variance σ_w^2. Consequently, optimal estimates of δ and σ_w^2 are the sample mean and variance of the y_t, respectively.

Example 3.37 IMA$(1,1)$ and EWMA

The ARIMA(0,1,1), or IMA(1,1) model is of interest because many economic time series can be successfully modeled this way. In addition, the model leads to a frequently used, and abused, forecasting method called exponentially weighted moving averages (EWMA). We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1}, \tag{3.147}$$

with $|\lambda| < 1$, for $t = 1, 2, \ldots$, and $x_0 = 0$, because this model formulation is easier to work with here, and it leads to the standard representation for EWMA. We could have included a drift term in (3.147), as was done in the previous example, but for the sake of simplicity, we leave it out of the discussion. If we write

$$y_t = w_t - \lambda w_{t-1},$$

we may write (3.147) as $x_t = x_{t-1} + y_t$. Because $|\lambda| < 1$, y_t has an invertible representation, $y_t + \sum_{j=1}^{\infty} \lambda^j y_{t-j} = w_t$, and substituting $y_t = x_t - x_{t-1}$, we may write

$$x_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{t-j} + w_t. \tag{3.148}$$

as an approximation for large t (put $x_t = 0$ for $t \le 0$). Verification of (3.148) is left to the reader (Problem 3.28). Using the approximation (3.148), we have that the approximate one-step-ahead predictor, using the notation of §3.5, is

$$\begin{aligned} \tilde{x}_{n+1} &= \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j} \\ &= (1-\lambda)x_n + \lambda \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n-j} \\ &= (1-\lambda)x_n + \lambda \tilde{x}_n. \end{aligned} \tag{3.149}$$

From (3.149), we see that the new forecast is a linear combination of the old forecast and the new observation. Based on (3.149) and the fact that we only observe $x_1, \ldots, x_n$, and consequently $y_1, \ldots, y_n$ (because $y_t = x_t - x_{t-1}$; $x_0 = 0$), the truncated forecasts are

$$\tilde{x}_{n+1}^n = (1-\lambda)x_n + \lambda \tilde{x}_n^{n-1}, \quad n \ge 1, \tag{3.150}$$

with $\tilde{x}_1^0 = x_1$ as an initial value. The mean-square prediction error can be approximated using (3.144) by noting that $\psi^*(z) = (1-\lambda z)/(1-z) = 1 + (1-\lambda)\sum_{j=1}^{\infty} z^j$ for $|z| < 1$; consequently, for large n, (3.144) leads to

$$P_{n+m}^n \approx \sigma_w^2[1 + (m-1)(1-\lambda)^2].$$

In EWMA, the parameter $1-\lambda$ is often called the smoothing parameter and is restricted to be between zero and one. Larger values of λ lead to smoother forecasts. This method of forecasting is popular because it is easy to use; we need only retain the previous forecast value and the current observation to forecast the next time period. Unfortunately, as previously suggested, the method is often abused because some forecasters do not verify that the observations follow an IMA$(1,1)$ process, and often arbitrarily pick values of λ. In the following, we show how to generate 100 observations from an IMA(1,1) model with $\lambda = -\theta = .8$ and then calculate and display the fitted EWMA superimposed on the data. This is accomplished using the Holt-Winters command in R (see the help file `?HoltWinters` for details; no output is shown):

```
set.seed(666)
x = arima.sim(list(order = c(0,1,1), ma = -0.8), n = 100)
(x.ima = HoltWinters(x, beta=FALSE, gamma=FALSE))  # α below is 1 − λ
  Smoothing parameter: alpha: 0.1663072
plot(x.ima)
```

3.8 Building ARIMA Models

There are a few basic steps to fitting ARIMA models to time series data. These steps involve plotting the data, possibly transforming the data, identifying the dependence orders of the model, parameter estimation, diagnostics, and model choice. First, as with any data analysis, we should construct a time plot of the data, and inspect the graph for any anomalies. If, for example, the variability in the data grows with time, it will be necessary to transform the data to stabilize the variance. In such cases, the Box–Cox class of power transformations, equation (2.37), could be employed. Also, the particular application might suggest an appropriate transformation. For example, suppose a process evolves as a fairly small and stable percent-change, such as an investment. For example, we might have

$$x_t = (1 + p_t)x_{t-1},$$

where x_t is the value of the investment at time t and p_t is the percentage-change from period $t-1$ to t, which may be negative. Taking logs we have

$$\log(x_t) = \log(1 + p_t) + \log(x_{t-1}),$$

or

$$\nabla \log(x_t) = \log(1 + p_t).$$

If the percent change p_t stays relatively small in magnitude, then $\log(1+p_t) \approx p_t$[8] and, thus,

$$\nabla \log(x_t) \approx p_t,$$

will be a relatively stable process. Frequently, $\nabla \log(x_t)$ is called the return or growth rate. This general idea was used in Example 3.32, and we will use it again in Example 3.38.

After suitably transforming the data, the next step is to identify preliminary values of the autoregressive order, p, the order of differencing, d, and the moving average order, q. We have already addressed, in part, the problem of selecting d. A time plot of the data will typically suggest whether any differencing is needed. If differencing is called for, then difference the data once, $d = 1$, and inspect the time plot of ∇x_t. If additional differencing is necessary, then try differencing again and inspect a time plot of $\nabla^2 x_t$. Be careful not to overdifference because this may introduce dependence where none exists. For example, $x_t = w_t$ is serially uncorrelated, but $\nabla x_t = w_t - w_{t-1}$ is MA(1). In addition to time plots, the sample ACF can help in indicating whether differencing is needed. Because the polynomial $\phi(z)(1 - z)^d$ has a unit root, the sample ACF, $\widehat{\rho}(h)$, will not decay to zero fast as h increases. Thus, a slow decay in $\widehat{\rho}(h)$ is an indication that differencing may be needed.

When preliminary values of d have been settled, the next step is to look at the sample ACF and PACF of $\nabla^d x_t$ for whatever values of d have been chosen. Using Table 3.1 as a guide, preliminary values of p and q are chosen. Recall that, if $p = 0$ and $q > 0$, the ACF cuts off after lag q, and the PACF tails off. If $q = 0$ and $p > 0$, the PACF cuts off after lag p, and the ACF tails off. If $p > 0$ and $q > 0$, both the ACF and PACF will tail off. Because we are dealing with estimates, it will not always be clear whether the sample ACF or PACF is tailing off or cutting off. Also, two models that are seemingly different can actually be very similar. With this in mind, we should not worry about being so precise at this stage of the model fitting. At this stage, a few preliminary values of p, d, and q should be at hand, and we can start estimating the parameters.

Example 3.38 Analysis of GNP Data

In this example, we consider the analysis of quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations. The data are real U.S. gross national product in billions of chained 1996 dollars and have been seasonally adjusted. The data were obtained from the Federal Reserve Bank of St. Louis (`http://research.stlouisfed.org/`). Figure 3.13 shows a plot of the data, say, y_t. Because strong trend hides any other effect, it is not clear from Figure 3.13 that the variance is increasing with time. For the purpose of demonstration, the sample ACF of the data is displayed in Figure 3.14. Figure 3.15

[8] $\log(1+p) = p - \frac{p^2}{2} + \frac{p^3}{3} - \cdots$ for $-1 < p \leq 1$. If p is a small percent-change, then the higher-order terms in the expansion are negligible.

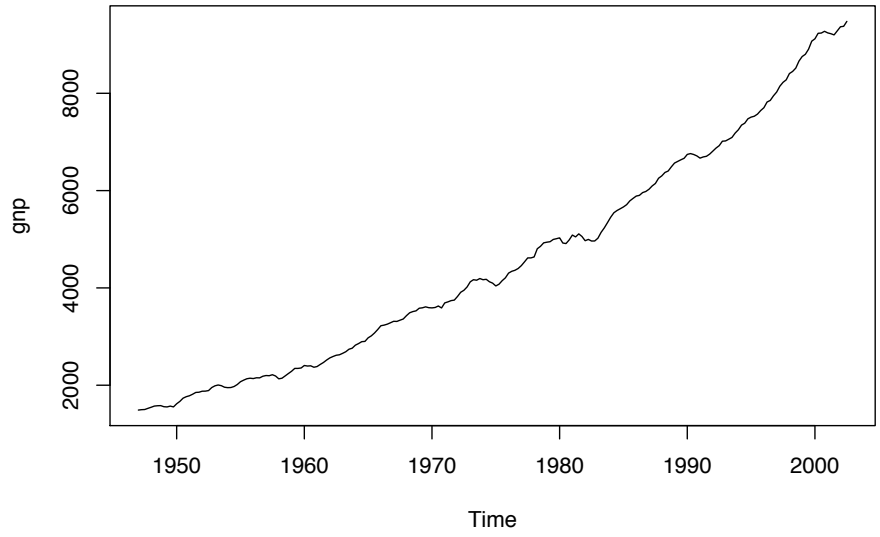

Fig. 3.13. Quarterly U.S. GNP from 1947(1) to 2002(3).

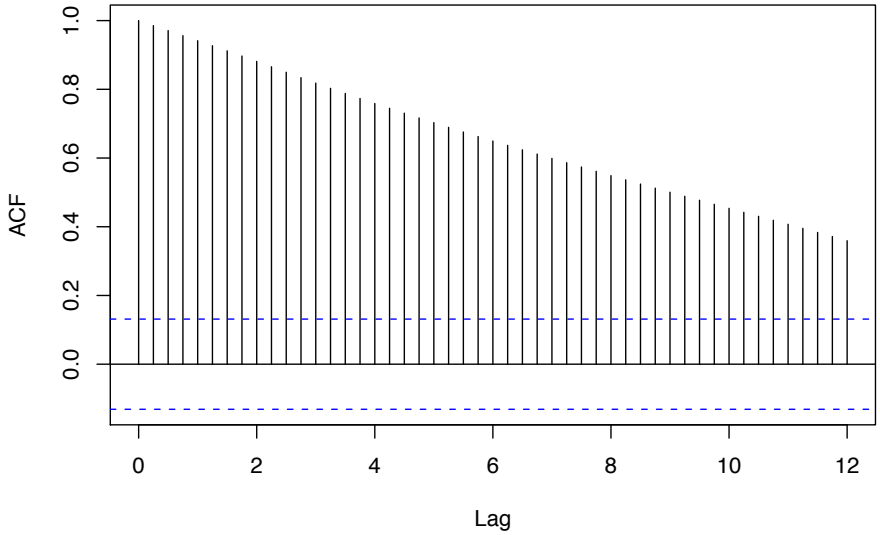

Fig. 3.14. Sample ACF of the GNP data. Lag is in terms of years.

shows the first difference of the data, ∇y_t, and now that the trend has been removed we are able to notice that the variability in the second half of the data is larger than in the first half of the data. Also, it appears as though a trend is still present after differencing. The growth rate, say, $x_t = \nabla \log(y_t)$, is plotted in Figure 3.16, and, appears to be a stable process. Moreover, we may interpret the values of x_t as the percentage quarterly growth of U.S. GNP.

The sample ACF and PACF of the quarterly growth rate are plotted in Figure 3.17. Inspecting the sample ACF and PACF, we might feel that the ACF is cutting off at lag 2 and the PACF is tailing off. This would suggest

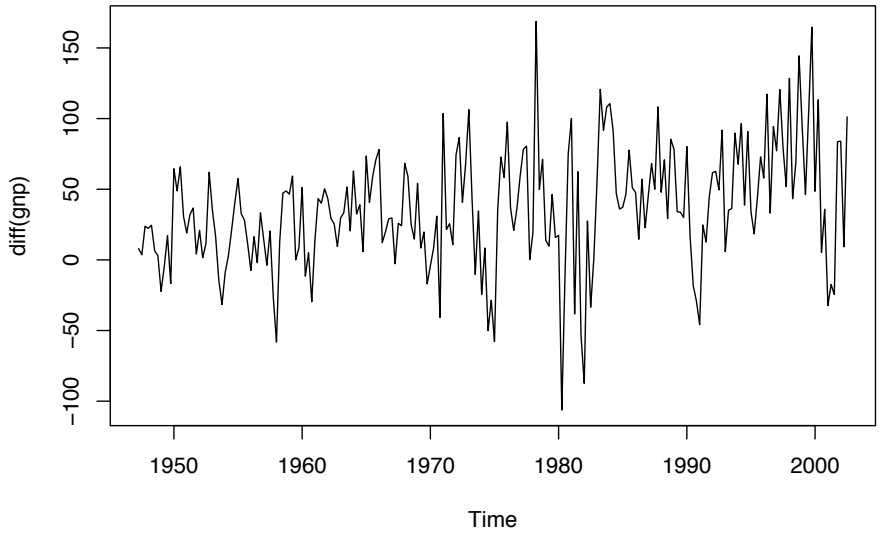

Fig. 3.15. First difference of the U.S. GNP data.

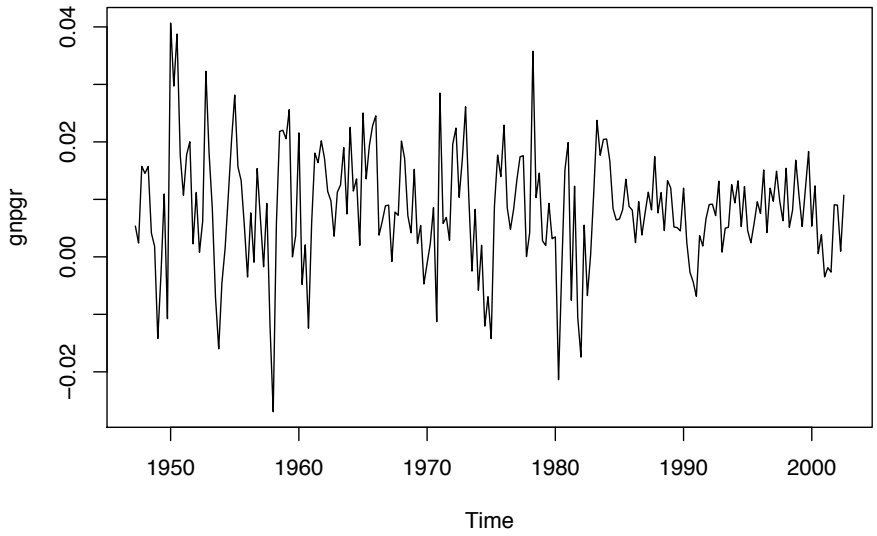

Fig. 3.16. U.S. GNP quarterly growth rate.

the GNP growth rate follows an MA(2) process, or log GNP follows an ARIMA$(0, 1, 2)$ model. Rather than focus on one model, we will also suggest that it appears that the ACF is tailing off and the PACF is cutting off at lag 1. This suggests an AR(1) model for the growth rate, or ARIMA$(1, 1, 0)$ for log GNP. As a preliminary analysis, we will fit both models.

Using MLE to fit the MA(2) model for the growth rate, x_t, the estimated model is

$$x_t = .008_{(.001)} + .303_{(.065)}\widehat{w}_{t-1} + .204_{(.064)}\widehat{w}_{t-2} + \widehat{w}_t, \tag{3.151}$$

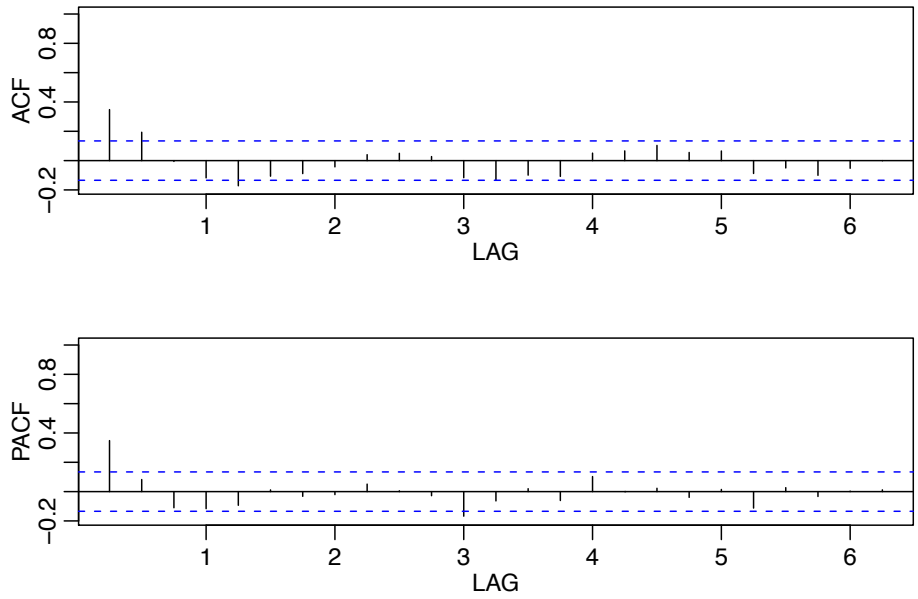

Fig. 3.17. Sample ACF and PACF of the GNP quarterly growth rate. Lag is in terms of years.

where $\widehat{\sigma}_w = .0094$ is based on 219 degrees of freedom. The values in parentheses are the corresponding estimated standard errors. All of the regression coefficients are significant, including the constant. *We make a special note of this because, as a default, some computer packages do not fit a constant in a differenced model.* That is, these packages assume, by default, that there is no drift. In this example, not including a constant leads to the wrong conclusions about the nature of the U.S. economy. Not including a constant assumes the average quarterly growth rate is zero, whereas the U.S. GNP average quarterly growth rate is about 1% (which can be seen easily in Figure 3.16). We leave it to the reader to investigate what happens when the constant is not included.

The estimated AR(1) model is

$$x_t = .008_{(.001)}\,(1 - .347) + .347_{(.063)}\,x_{t-1} + \widehat{w}_t, \tag{3.152}$$

where $\widehat{\sigma}_w = .0095$ on 220 degrees of freedom; note that the constant in (3.152) is $.008\,(1 - .347) = .005$.

We will discuss diagnostics next, but assuming both of these models fit well, how are we to reconcile the apparent differences of the estimated models (3.151) and (3.152)? In fact, the fitted models are nearly the same. To show this, consider an AR(1) model of the form in (3.152) without a constant term; that is,

$$x_t = .35x_{t-1} + w_t,$$

and write it in its causal form, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where we recall $\psi_j = .35^j$. Thus, $\psi_0 = 1, \psi_1 = .350, \psi_2 = .123, \psi_3 = .043, \psi_4 = .015, \psi_5 = .005, \psi_6 =$

$.002, \psi_7 = .001, \psi_8 = 0, \psi_9 = 0, \psi_{10} = 0$, and so forth. Thus,

$$x_t \approx .35w_{t-1} + .12w_{t-2} + w_t,$$

which is similar to the fitted MA(2) model in (3.152).

The analysis can be performed in R as follows.

```
plot(gnp)
acf2(gnp, 50)
gnpgr = diff(log(gnp))  # growth rate
plot(gnpgr)
acf2(gnpgr, 24)
sarima(gnpgr, 1, 0, 0)  # AR(1)
sarima(gnpgr, 0, 0, 2)  # MA(2)
ARMAtoMA(ar=.35, ma=0, 10)  # prints psi-weights
```

The next step in model fitting is diagnostics. This investigation includes the analysis of the residuals as well as model comparisons. Again, the first step involves a time plot of the innovations (or residuals), $x_t - \widehat{x}_t^{t-1}$, or of the standardized innovations

$$e_t = \left(x_t - \widehat{x}_t^{t-1}\right) \,/\, \sqrt{\widehat{P}_t^{t-1}}, \tag{3.153}$$

where $\widehat{x}_t^{t-1}$ is the one-step-ahead prediction of x_t based on the fitted model and $\widehat{P}_t^{t-1}$ is the estimated one-step-ahead error variance. If the model fits well, the standardized residuals should behave as an iid sequence with mean zero and variance one. The time plot should be inspected for any obvious departures from this assumption. Unless the time series is Gaussian, it is not enough that the residuals are uncorrelated. For example, it is possible in the non-Gaussian case to have an uncorrelated process for which values contiguous in time are highly dependent. As an example, we mention the family of GARCH models that are discussed in Chapter 5.

Investigation of marginal normality can be accomplished visually by looking at a histogram of the residuals. In addition to this, a normal probability plot or a Q-Q plot can help in identifying departures from normality. See Johnson and Wichern (1992, Chapter 4) for details of this test as well as additional tests for multivariate normality.

There are several tests of randomness, for example the runs test, that could be applied to the residuals. We could also inspect the sample autocorrelations of the residuals, say, $\widehat{\rho}_e(h)$, for any patterns or large values. Recall that, for a white noise sequence, the sample autocorrelations are approximately independently and normally distributed with zero means and variances $1/n$. Hence, a good check on the correlation structure of the residuals is to plot $\widehat{\rho}_e(h)$ versus h along with the error bounds of $\pm 2/\sqrt{n}$. The residuals from a model fit, however, will not quite have the properties of a white noise sequence and the variance of $\widehat{\rho}_e(h)$ can be much less than $1/n$. Details can be found in Box and Pierce (1970) and McLeod (1978). This part of the diagnostics can be viewed

as a visual inspection of $\widehat{\rho}_e(h)$ with the main concern being the detection of obvious departures from the independence assumption.

In addition to plotting $\widehat{\rho}_e(h)$, we can perform a general test that takes into consideration the magnitudes of $\widehat{\rho}_e(h)$ as a group. For example, it may be the case that, individually, each $\widehat{\rho}_e(h)$ is small in magnitude, say, each one is just slightly less that $2/\sqrt{n}$ in magnitude, but, collectively, the values are large. The Ljung–Box–Pierce Q-statistic given by

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\widehat{\rho}_e^2(h)}{n-h} \tag{3.154}$$

can be used to perform such a test. The value H in (3.154) is chosen somewhat arbitrarily, typically, $H = 20$. Under the null hypothesis of model adequacy, asymptotically ($n \to \infty$), $Q \sim \chi^2_{H-p-q}$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1-\alpha)$-quantile of the χ^2_{H-p-q} distribution. Details can be found in Box and Pierce (1970), Ljung and Box (1978), and Davies et al. (1977). The basic idea is that if w_t is white noise, then by Property 1.1, $n\widehat{\rho}_w^2(h)$, for $h = 1, \ldots, H$, are asymptotically independent χ^2_1 random variables. This means that $n\sum_{h=1}^{H}\widehat{\rho}_w^2(h)$ is approximately a χ^2_H random variable. Because the test involves the ACF of residuals from a model fit, there is a loss of $p+q$ degrees of freedom; the other values in (3.154) are used to adjust the statistic to better match the asymptotic chi-squared distribution.

Example 3.39 Diagnostics for GNP Growth Rate Example

We will focus on the MA(2) fit from Example 3.38; the analysis of the AR(1) residuals is similar. Figure 3.18 displays a plot of the standardized residuals, the ACF of the residuals, a boxplot of the standardized residuals, and the p-values associated with the Q-statistic, (3.154), at lags $H = 3$ through $H = 20$ (with corresponding degrees of freedom $H - 2$).

Inspection of the time plot of the standardized residuals in Figure 3.18 shows no obvious patterns. Notice that there are outliers, however, with a few values exceeding 3 standard deviations in magnitude. The ACF of the standardized residuals shows no apparent departure from the model assumptions, and the Q-statistic is never significant at the lags shown. The normal Q-Q plot of the residuals shows departure from normality at the tails due to the outliers that occurred primarily in the 1950s and the early 1980s.

The model appears to fit well except for the fact that a distribution with heavier tails than the normal distribution should be employed. We discuss some possibilities in Chapters 5 and 6. The diagnostics shown in Figure 3.18 are a by-product of the `sarima` command from the previous example.[9]

[9] The script `tsdiag` is available in R to run diagnostics for an ARIMA object, however, the script has errors and we do not recommend using it.

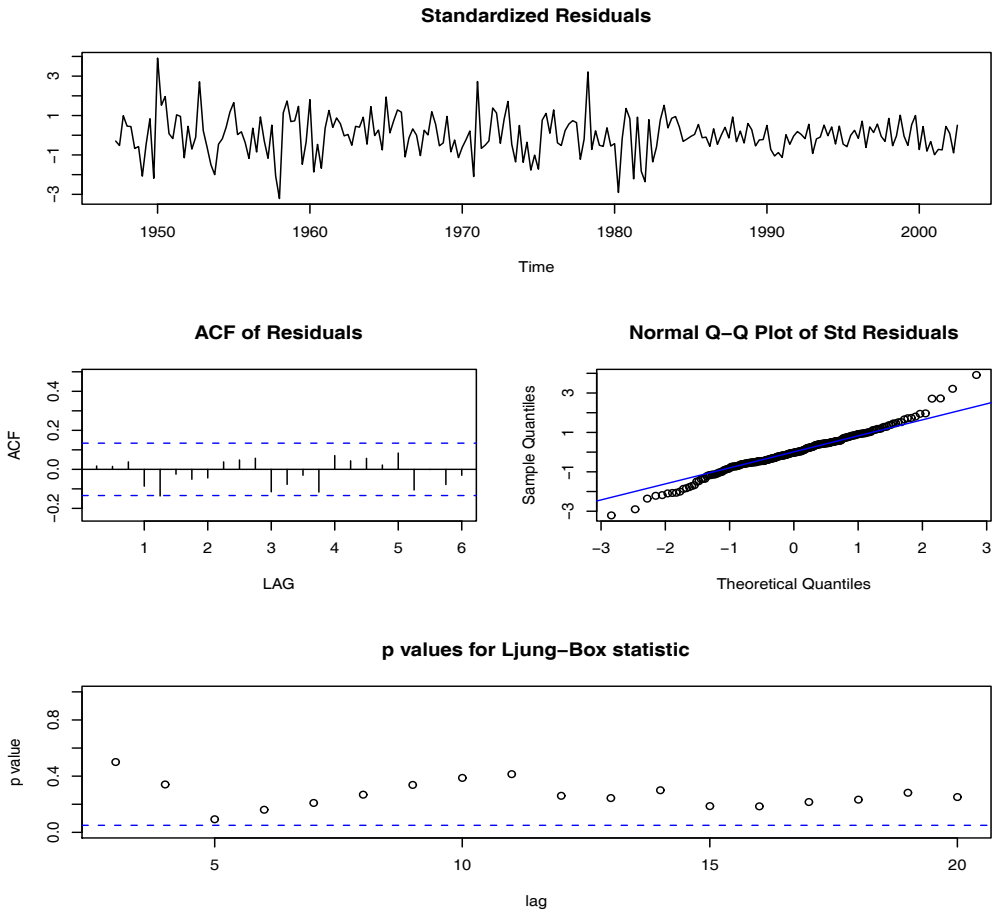

Fig. 3.18. Diagnostics of the residuals from MA(2) fit on GNP growth rate.

Example 3.40 Diagnostics for the Glacial Varve Series

In Example 3.32, we fit an ARIMA$(0,1,1)$ model to the logarithms of the glacial varve data and there appears to be a small amount of autocorrelation left in the residuals and the Q-tests are all significant; see Figure 3.19.

To adjust for this problem, we fit an ARIMA$(1,1,1)$ to the logged varve data and obtained the estimates

$$\widehat{\phi} = .23_{(.05)},\ \widehat{\theta} = -.89_{(.03)},\ \widehat{\sigma}_w^2 = .23.$$

Hence the AR term is significant. The Q-statistic p-values for this model are also displayed in Figure 3.19, and it appears this model fits the data well.

As previously stated, the diagnostics are byproducts of the individual `sarima` runs. We note that we did not fit a constant in either model because there is no apparent drift in the differenced, logged varve series. This fact can be verified by noting the constant is not significant when the command `no.constant=TRUE` is removed in the code:

```
sarima(log(varve), 0, 1, 1, no.constant=TRUE)    # ARIMA(0,1,1)
sarima(log(varve), 1, 1, 1, no.constant=TRUE)    # ARIMA(1,1,1)
```