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.

R.H. Shumway and D.S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-7865-3_3,

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:

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

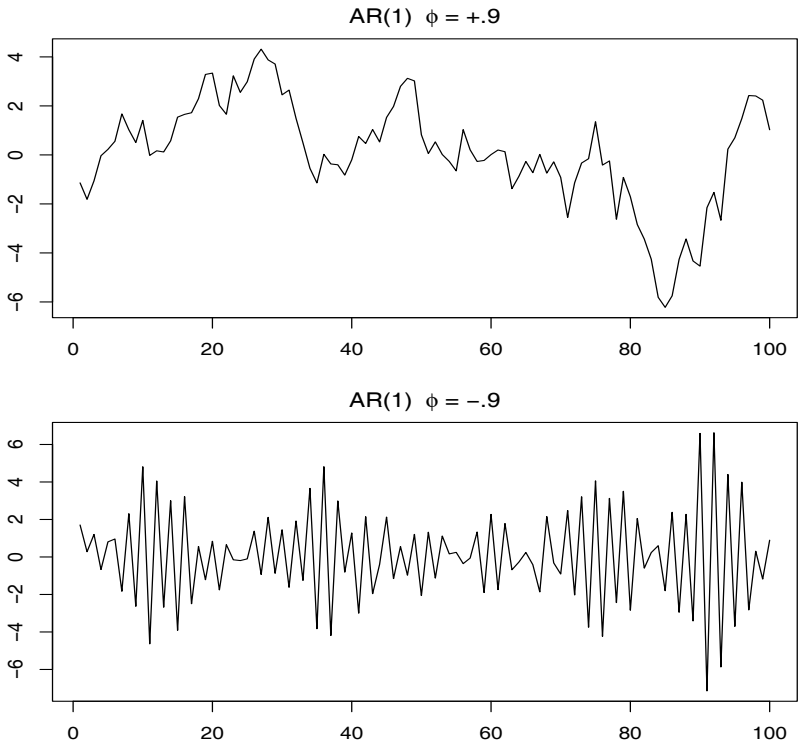

Fig. 3.1. Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).

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

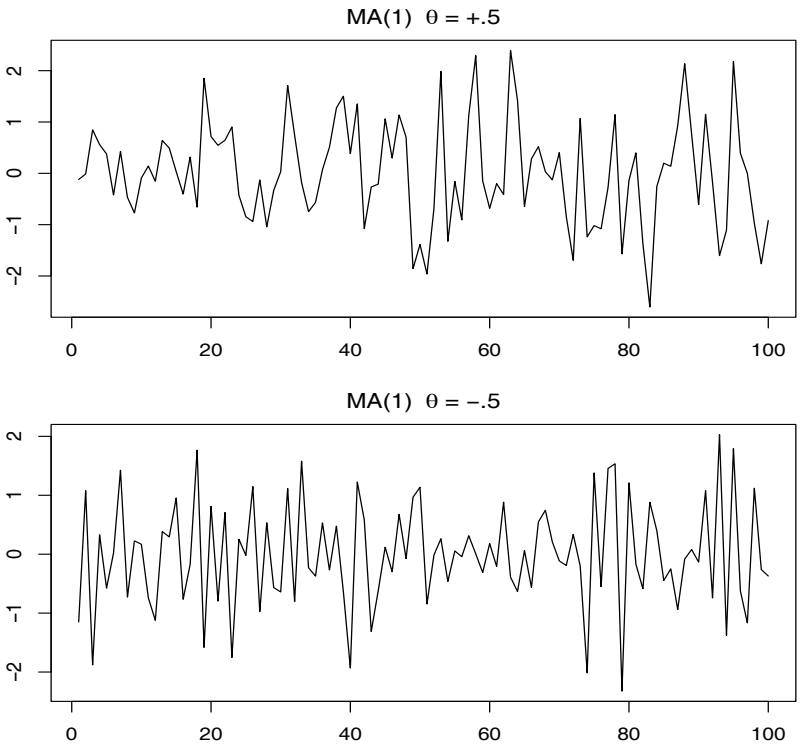

Fig. 3.2. Simulated MA(1) models: $\theta = .5$ (top); $\theta = -.5$ (bottom).

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 Figure 3.2 where $\theta = .5$ is smoother than the series in Figure 3.2, where $\theta = -.5$.

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

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

Examples 3.2, 3.5, and 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. We leave the details of the equivalence to the reader (Problem 3.5).

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 3.3 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 3.4.

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 3.3:

```
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:

```
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

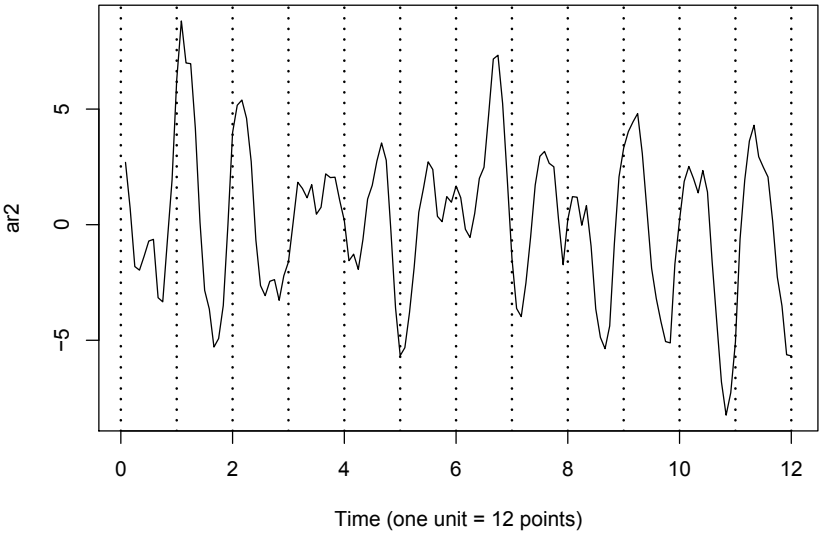

Fig. 3.3. Simulated AR(2) model, $n = 144$ with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

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

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

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

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 $j \ge 1$, 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

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

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

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

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

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

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,

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

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.

[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 $E(x_{t+h} - \sum_{j=1}^{h-1} \alpha_j x_{t+j})^2$.

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

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

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

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

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 $h > p$,

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

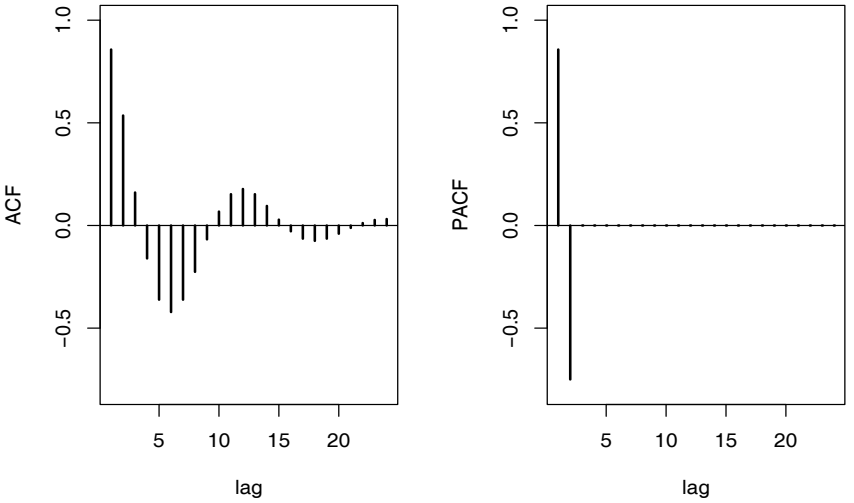

Fig. 3.4. The ACF and PACF of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

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 3.4 shows the ACF and the PACF of the AR(2) model presented in Example 3.10.

To reproduce Figure 3.4 in R, use the following commands:

```
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.

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

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 3.5 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

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

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

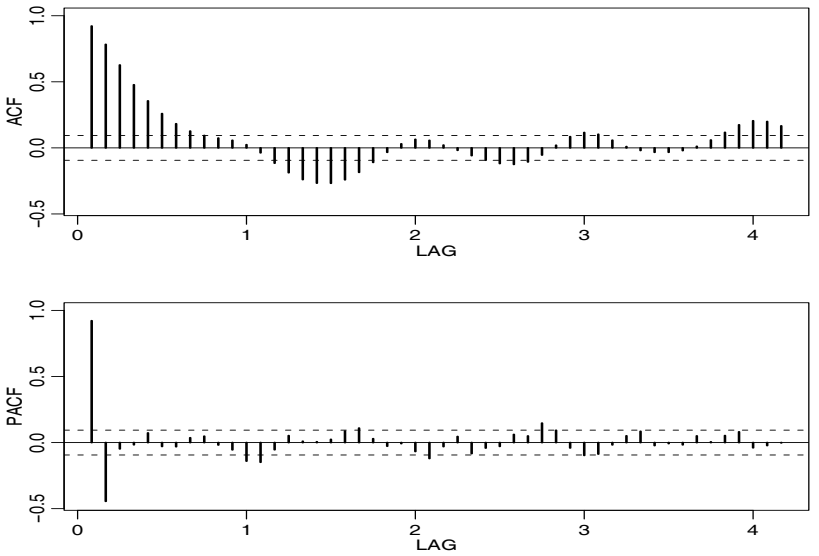

Fig. 3.5. ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).

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

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

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

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,* $x^n_{n+m} = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k$*, of* x_{n+m}*, for* $m \geq 1$*, is found by solving*

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

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

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

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

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

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

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

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_{n+1} 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),

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

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,

$$\begin{aligned} \phi_{11} &= \rho(1) = \frac{\phi_1}{1-\phi_2} \\ \phi_{22} &= \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2 \\ \phi_{21} &= \rho(1)[1 - \phi_2] = \phi_1 \\ \phi_{33} &= \frac{\rho(3) - \phi_1\rho(2) - \phi_2\rho(1)}{1 - \phi_1\rho(1) - \phi_2\rho(2)} = 0. \end{aligned}$$

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

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

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

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

or

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

The prediction equations can again be written in matrix notation as

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

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

The mean square m-step-ahead prediction error is

$$P_{n+m}^{n} = E\left(x_{n+m} - x_{n+m}^{n}\right)^2 = \gamma(0) - \boldsymbol{\gamma}_n^{(m)'} \Gamma_n^{-1} \boldsymbol{\gamma}_n^{(m)}. \tag{3.76}$$

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, $x_t - x_t^{t-1}$, for $t = 1, \ldots, n$, using the fact that the innovations $x_t - x_t^{t-1}$ and $x_s - x_s^{s-1}$ 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

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

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,

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

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

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

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 3.6 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

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

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_{n+m}, of x_{n+m}, and there is no need for approximations. That is, for $n > p$, $\widetilde{x}^n_{n+m} = \widetilde{x}_{n+m} = x^n_{n+m}$. Also, in this case, the one-step-ahead prediction error is $E(x_{n+1} - x^n_{n+1})^2 = \sigma_w^2$. 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

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

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

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

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

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

For $m \geq 2$, we have

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

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 $\widetilde{w}^n_0 = 0$, $x_0 = 0$, and then iterate the errors forward in time

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

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

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

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

$$x^n_{n+m} \pm c_{\frac{\alpha}{2}} \sqrt{P^n_{n+m}}, \tag{3.93}$$

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 3.6 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

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

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 $\widehat{\sigma}^2_w = 89.72$,

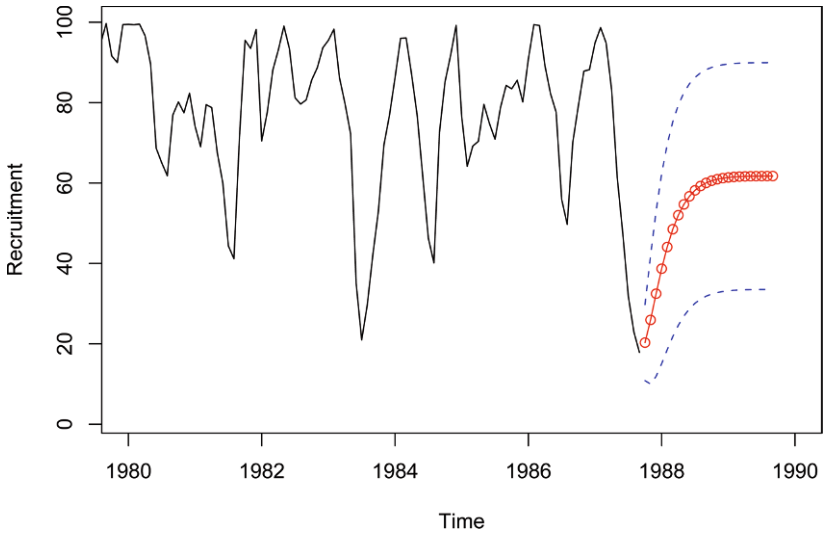

Fig. 3.6. Twenty-four month forecasts for the Recruitment series. The actual data shown are from about January 1980 to September 1987, and then the forecasts plus and minus one standard error are displayed.

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 3.6, use the following commands:

```
1 regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE)
2 fore = predict(regr, n.ahead=24)
3 ts.plot(rec, fore$pred, col=1:2, xlim=c(1980,1990),
    ylab="Recruitment")
4 lines(fore$pred, type="p", col=2)
5 lines(fore$pred+fore$se, lty="dashed", col=4)
6 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

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

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, $\alpha_j \equiv \phi_{nj}^{(m)}$, for $j = 1, \ldots, n$, where the $\phi_{nj}^{(m)}$ are given by (3.75). Finally, the backcasts are given by

$$x_{1-m}^n = \phi_{n1}^{(m)} x_1 + \cdots + \phi_{nn}^{(m)} x_n, \quad m = 1, 2, \ldots. \tag{3.97}$$

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

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

[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.

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 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 3.3 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=FALSE)[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]

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

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 $w_t \sim$ 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

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

[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.

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

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

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

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

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

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

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

where

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

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

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

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

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

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 Examples 3.17 and 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

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

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

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

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

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

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

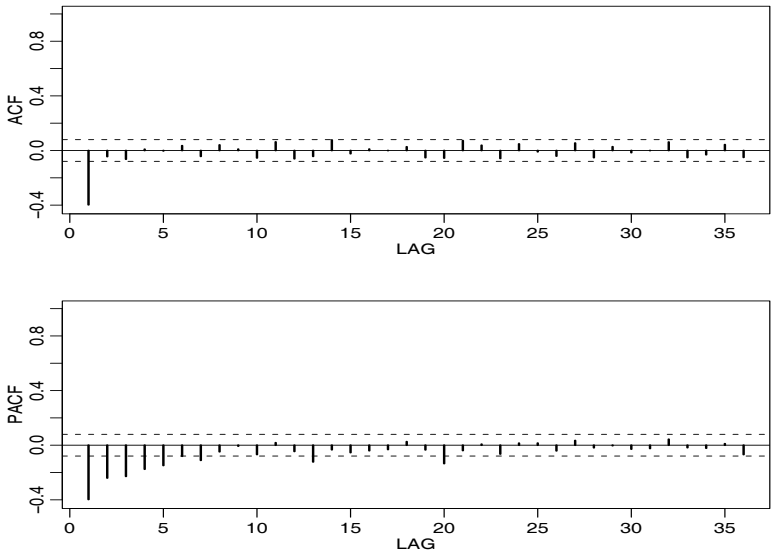

Fig. 3.7. ACF and PACF of transformed glacial varves.

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

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.

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.

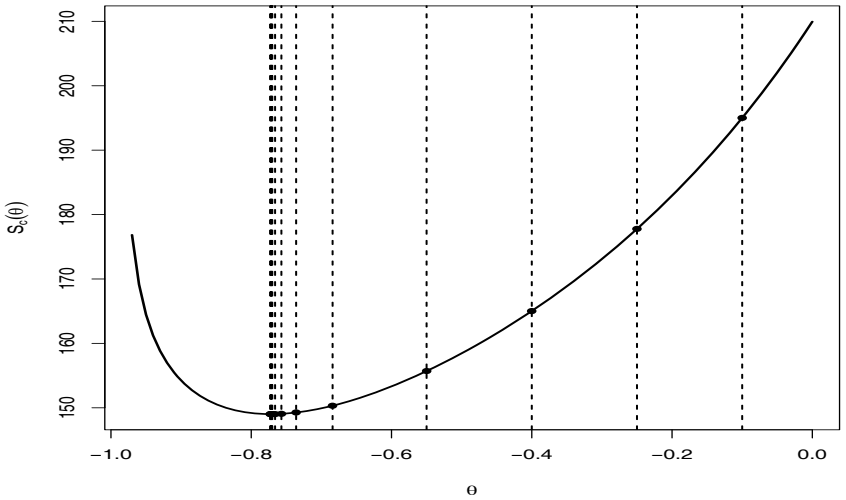

Fig. 3.8. Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 3.32. Vertical lines indicate the values of the parameter obtained via Gauss–Newton; see Table 3.2 for the actual values.

The sample ACF and PACF, shown in Figure 3.7, 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 consequently, 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 3.8 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 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.

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

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

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, $\mathrm{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large n, $\mathrm{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$ 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.

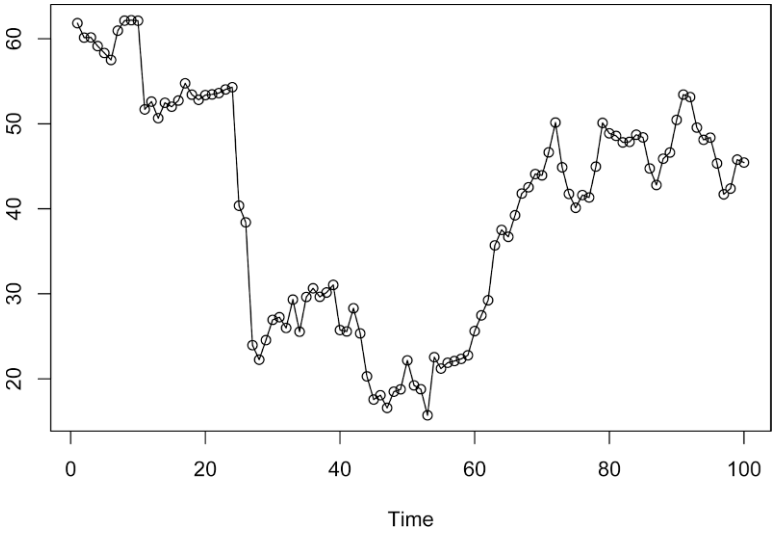

Fig. 3.9. One hundred observations generated from the model in Example 3.35.

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 $\text{var}(w_t) = 2\beta^2 = 8$. Figure 3.9 shows $n = 100$ simulated observations from this process. This particular realization 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 3.9, 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 $(1 - .96^2)/100 = .03^2$.

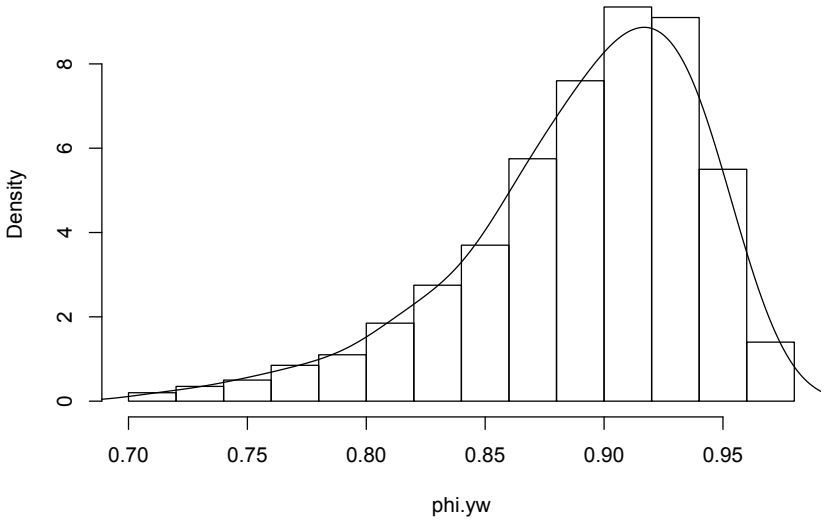

Fig. 3.10. Finite sample density of the Yule–Walker estimate of ϕ in Example 3.35.

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 3.10. Clearly the sampling distribution is not close to normality for this sample size. The mean of the distribution shown in Figure 3.10 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){
  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

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

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

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

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 3.11 shows the bootstrap histogram of 200 bootstrapped estimates of ϕ using the data shown in Figure 3.9. In addition, Figure 3.11 shows a density estimate based on the bootstrap histogram, as well as the asymptotic normal density that would have been used based on Proposition 3.10. Clearly, the bootstrap distribution of $\widehat{\phi}$ is closer to the distribution of $\widehat{\phi}$ shown in Figure 3.10 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,

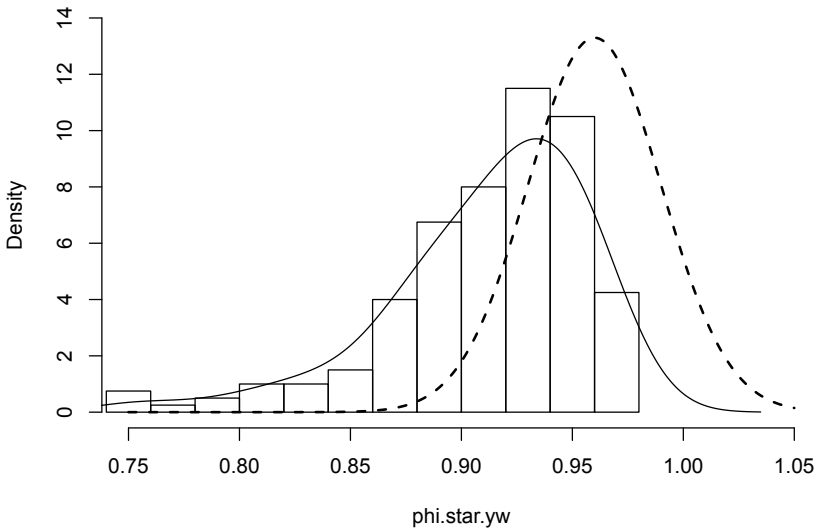

Fig. 3.11. Bootstrap histogram of $\widehat{\phi}$ based on 200 bootstraps; a density estimate based on the histogram (solid line) and the corresponding asymptotic normal density (dashed line).

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 3.10.

```
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)
  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_{n+m} for $m = 1, 2, \ldots$, we have $y^n_{n+m} = x^n_{n+m} - x^n_{n+m-1}$, 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

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

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

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

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

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

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):

```
1 set.seed(666)
2 x = arima.sim(list(order = c(0,1,1), ma = -0.8), n = 100)
3 (x.ima = HoltWinters(x, beta=FALSE, gamma=FALSE))  # α below is 1 − λ
    Smoothing parameter: alpha: 0.1663072
4 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

[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.

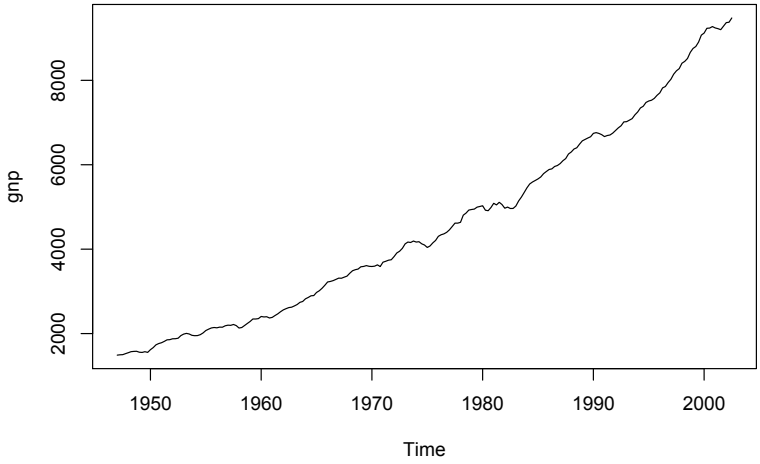

Fig. 3.12. Quarterly U.S. GNP from 1947(1) to 2002(3).

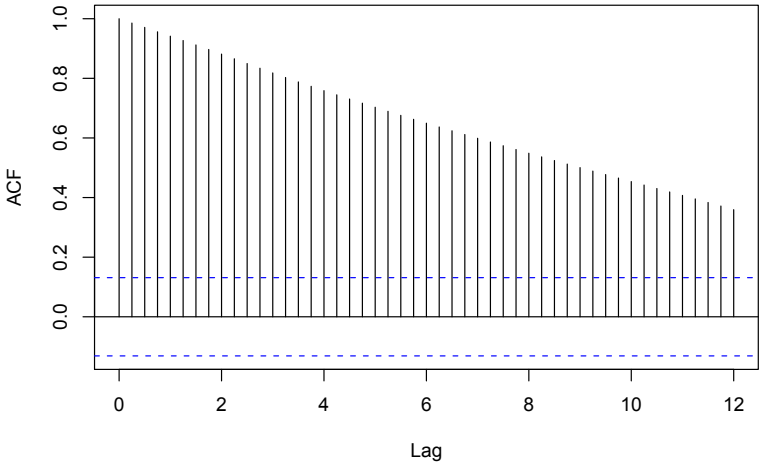

Fig. 3.13. Sample ACF of the GNP data. Lag is in terms of years.

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 3.12 shows a plot of the data, say, y_t. Because strong trend hides any other effect, it is not clear from Figure 3.12 that the variance is increasing with time. For the purpose of demonstration, the sample ACF of the data is displayed in Figure 3.13. Figure 3.14 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

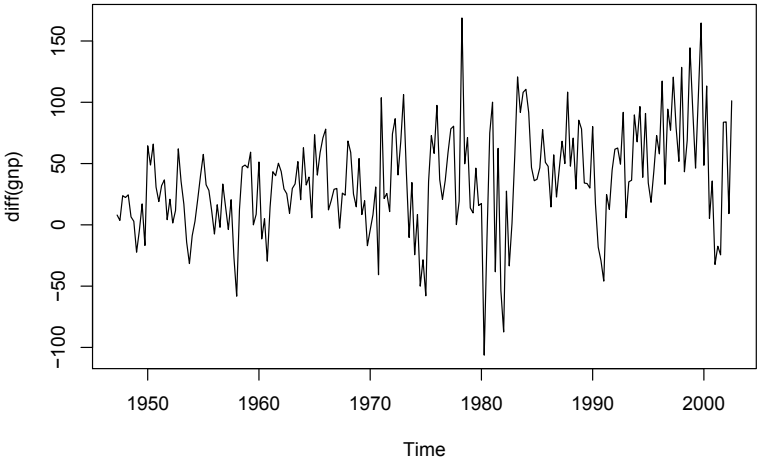

Fig. 3.14. First difference of the U.S. GNP data.

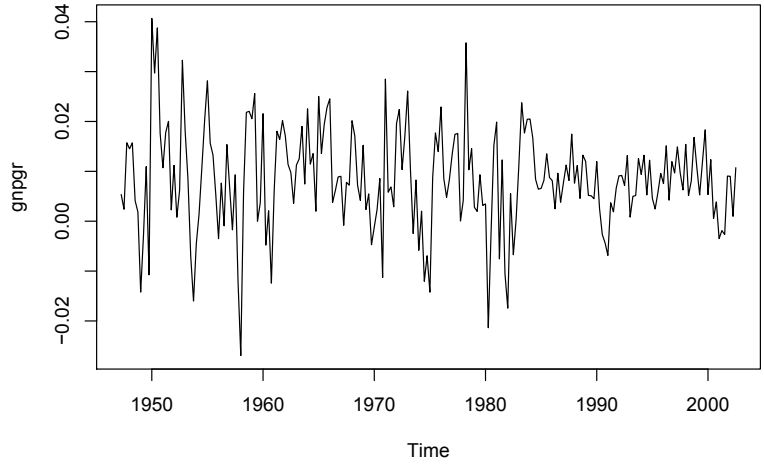

Fig. 3.15. U.S. GNP quarterly growth rate.

rate, say, $x_t = \nabla \log(y_t)$, is plotted in Figure 3.15, 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 3.16. 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 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

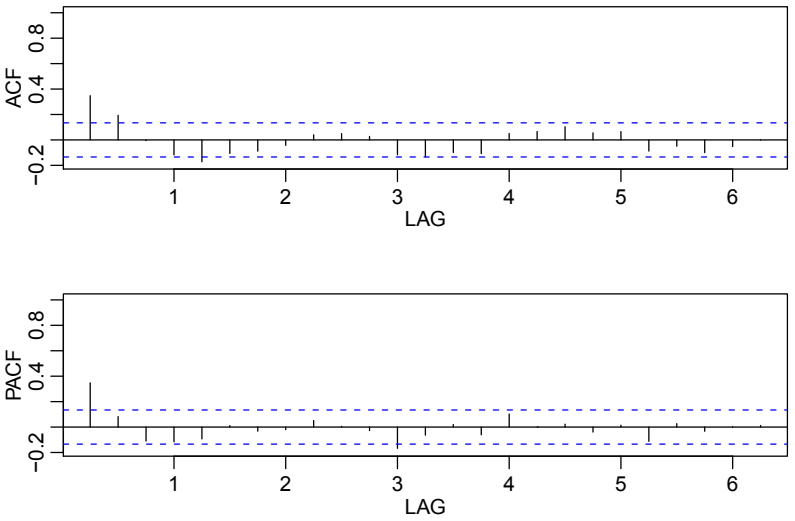

Fig. 3.16. Sample ACF and PACF of the GNP quarterly growth rate. Lag is in terms of years.

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

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

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 3.15). 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 3.17 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 3.17 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

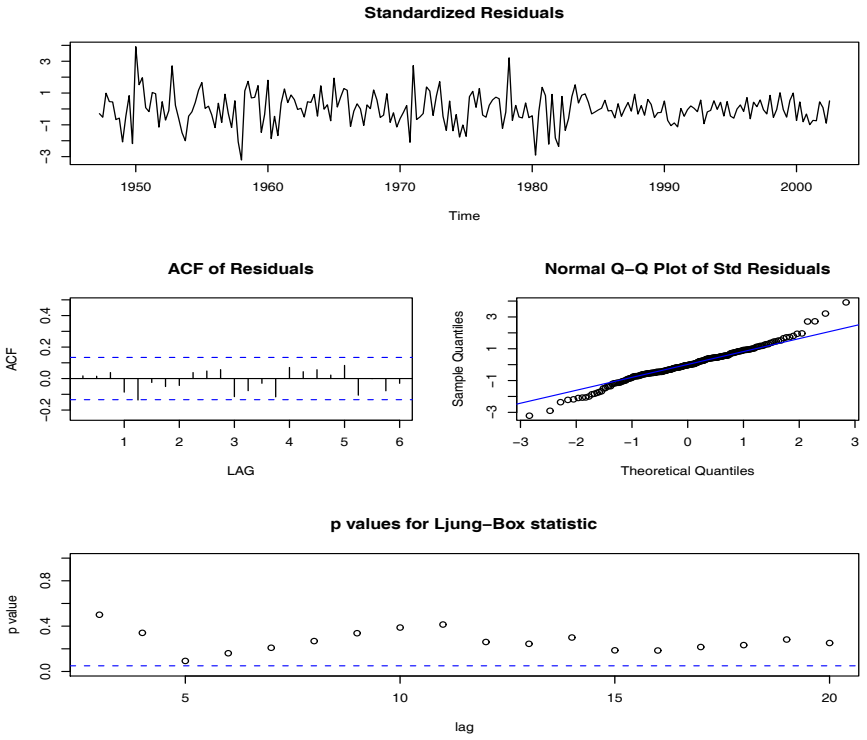

Fig. 3.17. Diagnostics of the residuals from MA(2) fit on GNP growth rate.

some possibilities in Chapters 5 and 6. The diagnostics shown in Figure 3.17 are a by-product of the `sarima` command from the previous example.[9]

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 3.18.

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 3.18, 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

[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.

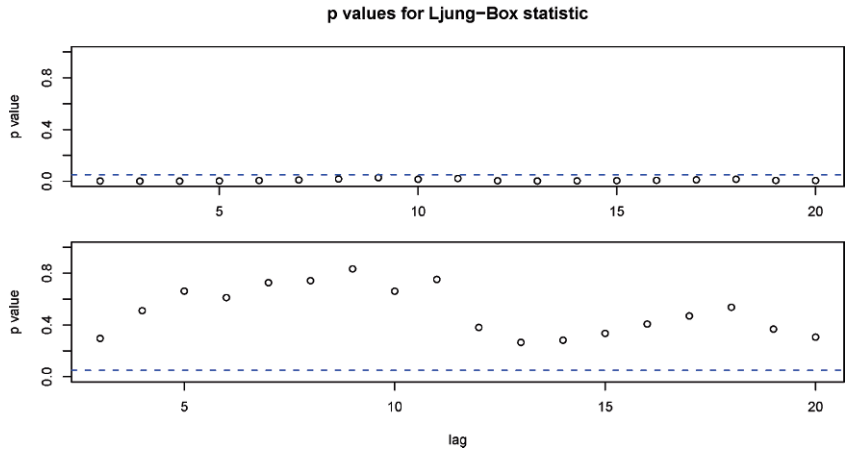

Fig. 3.18. Q-statistic p-values for the ARIMA$(0,1,1)$ fit [top] and the ARIMA$(1,1,1)$ fit [bottom] to the logged varve data.

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:

```
1 sarima(log(varve), 0, 1, 1, no.constant=TRUE)    # ARIMA(0,1,1)
2 sarima(log(varve), 1, 1, 1, no.constant=TRUE)    # ARIMA(1,1,1)
```

In Example 3.38, we have two competing models, an AR(1) and an MA(2) on the GNP growth rate, that each appear to fit the data well. In addition, we might also consider that an AR(2) or an MA(3) might do better for forecasting. Perhaps combining both models, that is, fitting an ARMA$(1,2)$ to the GNP growth rate, would be the best. As previously mentioned, we have to be concerned with overfitting the model; it is not always the case that more is better. Overfitting leads to less-precise estimators, and adding more parameters may fit the data better but may also lead to bad forecasts. This result is illustrated in the following example.

Example 3.41 A Problem with Overfitting

Figure 3.19 shows the U.S. population by official census, every ten years from 1910 to 1990, as points. If we use these nine observations to predict the future population, we can use an eight-degree polynomial so the fit to the nine observations is perfect. The model in this case is

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_8 t^8 + w_t.$$

The fitted line, which is plotted in the figure, passes through the nine observations. The model predicts that the population of the United States will be close to zero in the year 2000, and will cross zero sometime in the year 2002!

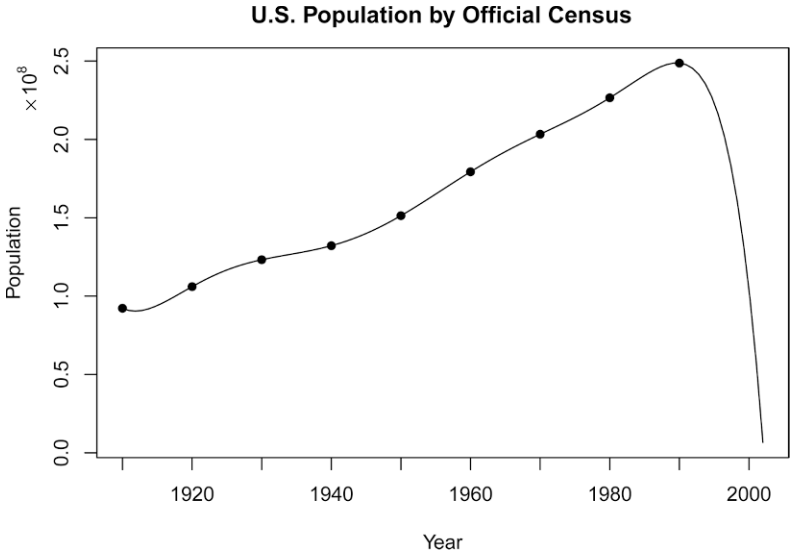

Fig. 3.19. A perfect fit and a terrible forecast.

The final step of model fitting is model choice or model selection. That is, we must decide which model we will retain for forecasting. The most popular techniques, AIC, AICc, and BIC, were described in §2.2 in the context of regression models.

Example 3.42 Model Choice for the U.S. GNP Series

Returning to the analysis of the U.S. GNP data presented in Examples 3.38 and 3.39, recall that two models, an AR(1) and an MA(2), fit the GNP growth rate well. To choose the final model, we compare the AIC, the AICc, and the BIC for both models. These values are a byproduct of the `sarima` runs displayed at the end of Example 3.38, but for convenience, we display them again here (recall the growth rate data are in `gnpgr`):

```
1 sarima(gnpgr, 1, 0, 0)  # AR(1)
      $AIC: -8.294403   $AICc: -8.284898   $BIC: -9.263748
2 sarima(gnpgr, 0, 0, 2)  # MA(2)
      $AIC: -8.297693   $AICc: -8.287854   $BIC: -9.251711
```

The AIC and AICc both prefer the MA(2) fit, whereas the BIC prefers the simpler AR(1) model. It is often the case that the BIC will select a model of smaller order than the AIC or AICc. It would not be unreasonable in this case to retain the AR(1) because pure autoregressive models are easier to work with.

3.9 Multiplicative Seasonal ARIMA Models

In this section, we introduce several modifications made to the ARIMA model to account for seasonal and nonstationary behavior. Often, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag s. For example, with monthly economic data, there is a strong yearly component occurring at lags that are multiples of $s = 12$, because of the strong connections of all activity to the calendar year. Data taken quarterly will exhibit the yearly repetitive period at $s = 4$ quarters. Natural phenomena such as temperature also have strong components corresponding to seasons. Hence, the natural variability of many physical, biological, and economic processes tends to match with seasonal fluctuations. Because of this, it is appropriate to introduce autoregressive and moving average polynomials that identify with the seasonal lags. The resulting pure seasonal autoregressive moving average model, say, ARMA$(P, Q)_s$, then takes the form

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \tag{3.155}$$

with the following definition.

Definition 3.12 *The operators*

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{3.156}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \tag{3.157}$$

are the **seasonal autoregressive operator** *and the* **seasonal moving average operator** *of orders P and Q, respectively, with seasonal period s.*

Analogous to the properties of nonseasonal ARMA models, the pure seasonal ARMA$(P, Q)_s$ is causal only when the roots of $\Phi_P(z^s)$ lie outside the unit circle, and it is invertible only when the roots of $\Theta_Q(z^s)$ lie outside the unit circle.

Example 3.43 A Seasonal ARMA Series

A first-order seasonal autoregressive moving average series that might run over months could be written as

$$(1 - \Phi B^{12})x_t = (1 + \Theta B^{12})w_t$$

or

$$x_t = \Phi x_{t-12} + w_t + \Theta w_{t-12}.$$

This model exhibits the series x_t in terms of past lags at the multiple of the yearly seasonal period $s = 12$ months. It is clear from the above form that estimation and forecasting for such a process involves only straightforward modifications of the unit lag case already treated. In particular, the causal condition requires $|\Phi| < 1$, and the invertible condition requires $|\Theta| < 1$.

Table 3.3. Behavior of the ACF and PACF for Pure SARMA Models

	$\text{AR}(P)_s$	$\text{MA}(Q)_s$	$\text{ARMA}(P,Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots$, are zero.

For the first-order seasonal ($s = 12$) MA model, $x_t = w_t + \Theta w_{t-12}$, it is easy to verify that

$$\begin{aligned} \gamma(0) &= (1 + \Theta^2)\sigma^2 \\ \gamma(\pm 12) &= \Theta\sigma^2 \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

Thus, the only nonzero correlation, aside from lag zero, is

$$\rho(\pm 12) = \Theta/(1 + \Theta^2).$$

For the first-order seasonal ($s = 12$) AR model, using the techniques of the nonseasonal AR(1), we have

$$\begin{aligned} \gamma(0) &= \sigma^2/(1 - \Phi^2) \\ \gamma(\pm 12k) &= \sigma^2\Phi^k/(1 - \Phi^2) \quad k = 1, 2, \ldots \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

In this case, the only non-zero correlations are

$$\rho(\pm 12k) = \Phi^k, \quad k = 0, 1, 2, \ldots \ .$$

These results can be verified using the general result that $\gamma(h) = \Phi\gamma(h - 12)$, for $h \geq 1$. For example, when $h = 1$, $\gamma(1) = \Phi\gamma(11)$, but when $h = 11$, we have $\gamma(11) = \Phi\gamma(1)$, which implies that $\gamma(1) = \gamma(11) = 0$. In addition to these results, the PACF have the analogous extensions from nonseasonal to seasonal models.

As an initial diagnostic criterion, we can use the properties for the pure seasonal autoregressive and moving average series listed in Table 3.3. These properties may be considered as generalizations of the properties for nonseasonal models that were presented in Table 3.1.

In general, we can combine the seasonal and nonseasonal operators into a multiplicative seasonal autoregressive moving average model, denoted by $\text{ARMA}(p, q) \times (P, Q)_s$, and write

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \tag{3.158}$$

as the overall model. Although the diagnostic properties in Table 3.3 are not strictly true for the overall mixed model, the behavior of the ACF and PACF tends to show rough patterns of the indicated form. In fact, for mixed models, we tend to see a mixture of the facts listed in Tables 3.1 and 3.3. In fitting such models, focusing on the seasonal autoregressive and moving average components first generally leads to more satisfactory results.

Example 3.44 A Mixed Seasonal Model

Consider an ARMA$(0,1) \times (1,0)_{12}$ model

$$x_t = \Phi x_{t-12} + w_t + \theta w_{t-1},$$

where $|\Phi| < 1$ and $|\theta| < 1$. Then, because x_{t-12}, w_t, and w_{t-1} are uncorrelated, and x_t is stationary, $\gamma(0) = \Phi^2\gamma(0) + \sigma_w^2 + \theta^2\sigma_w^2$, or

$$\gamma(0) = \frac{1+\theta^2}{1-\Phi^2}\,\sigma_w^2.$$

In addition, multiplying the model by x_{t-h}, $h > 0$, and taking expectations, we have $\gamma(1) = \Phi\gamma(11) + \theta\sigma_w^2$, and $\gamma(h) = \Phi\gamma(h-12)$, for $h \geq 2$. Thus, the ACF for this model is

$$\begin{aligned} \rho(12h) &= \Phi^h \quad h = 1, 2, \ldots \\ \rho(12h-1) &= \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h \quad h = 0, 1, 2, \ldots, \\ \rho(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

The ACF and PACF for this model, with $\Phi = .8$ and $\theta = -.5$, are shown in Figure 3.20. These type of correlation relationships, although idealized here, are typically seen with seasonal data.

To reproduce Figure 3.20 in R, use the following commands:

```
phi = c(rep(0,11),.8)
ACF = ARMAacf(ar=phi, ma=-.5, 50)[-1]      # [-1] removes 0 lag
PACF = ARMAacf(ar=phi, ma=-.5, 50, pacf=TRUE)
par(mfrow=c(1,2))
plot(ACF,  type="h", xlab="lag", ylim=c(-.4,.8));  abline(h=0)
plot(PACF, type="h", xlab="lag", ylim=c(-.4,.8));  abline(h=0)
```

Seasonal nonstationarity can occur, for example, when the process is nearly periodic in the season. For example, with average monthly temperatures over the years, each January would be approximately the same, each February would be approximately the same, and so on. In this case, we might think of average monthly temperature x_t as being modeled as

$$x_t = S_t + w_t,$$

where S_t is a seasonal component that varies slowly from one year to the next, according to a random walk,

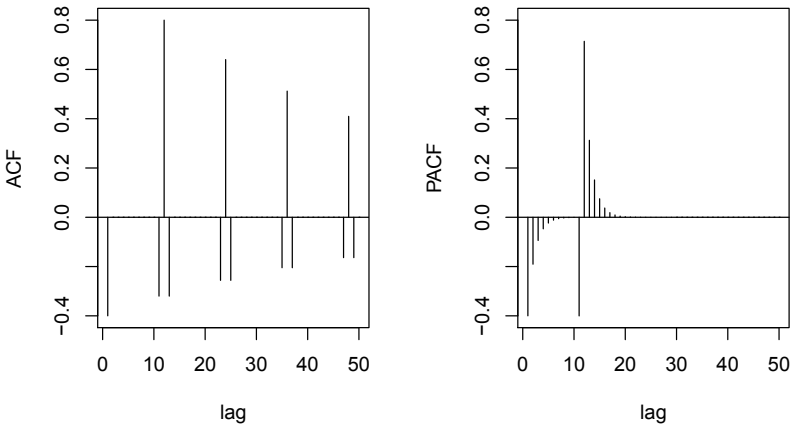

Fig. 3.20. ACF and PACF of the mixed seasonal ARMA model $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

$$S_t = S_{t-12} + v_t.$$

In this model, w_t and v_t are uncorrelated white noise processes. The tendency of data to follow this type of model will be exhibited in a sample ACF that is large and decays very slowly at lags $h = 12k$, for $k = 1, 2, \ldots$. If we subtract the effect of successive years from each other, we find that

$$(1 - B^{12})x_t = x_t - x_{t-12} = v_t + w_t - w_{t-12}.$$

This model is a stationary $\text{MA}(1)_{12}$, and its ACF will have a peak only at lag 12. In general, seasonal differencing can be indicated when the ACF decays slowly at multiples of some season s, but is negligible between the periods. Then, a seasonal difference of order D is defined as

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \tag{3.159}$$

where $D = 1, 2, \ldots$, takes positive integer values. Typically, $D = 1$ is sufficient to obtain seasonal stationarity. Incorporating these ideas into a general model leads to the following definition.

Definition 3.13 *The multiplicative* **seasonal autoregressive integrated moving average** *model, or* **SARIMA** *model is given by*

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t, \tag{3.160}$$

where w_t is the usual Gaussian white noise process. The general model is denoted as $\mathbf{ARIMA}(p, d, q) \times (P, D, Q)_s$. *The ordinary autoregressive and moving average components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively [see (3.5) and (3.18)], and the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ [see (3.156) and (3.157)] of orders P and Q and ordinary and seasonal difference components by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.*

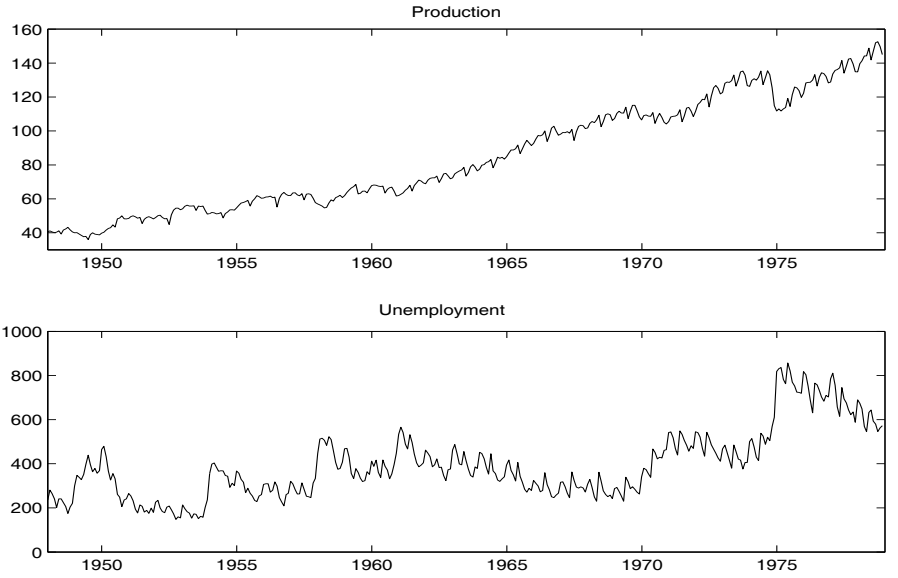

Fig. 3.21. Values of the Monthly Federal Reserve Board Production Index and Unemployment (1948-1978, $n = 372$ months).

Example 3.45 An SARIMA Model

Consider the following model, which often provides a reasonable representation for seasonal, nonstationary, economic time series. We exhibit the equations for the model, denoted by ARIMA$(0, 1, 1) \times (0, 1, 1)_{12}$ in the notation given above, where the seasonal fluctuations occur every 12 months. Then, the model (3.160) becomes

$$(1 - B^{12})(1 - B)x_t = (1 + \Theta B^{12})(1 + \theta B)w_t. \tag{3.161}$$

Expanding both sides of (3.161) leads to the representation

$$(1 - B - B^{12} + B^{13})x_t = (1 + \theta B + \Theta B^{12} + \Theta\theta B^{13})w_t,$$

or in difference equation form

$$x_t = x_{t-1} + x_{t-12} - x_{t-13} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}.$$

Note that the multiplicative nature of the model implies that the coefficient of w_{t-13} is the product of the coefficients of w_{t-1} and w_{t-12} rather than a free parameter. The multiplicative model assumption seems to work well with many seasonal time series data sets while reducing the number of parameters that must be estimated.

Selecting the appropriate model for a given set of data from all of those represented by the general form (3.160) is a daunting task, and we usually

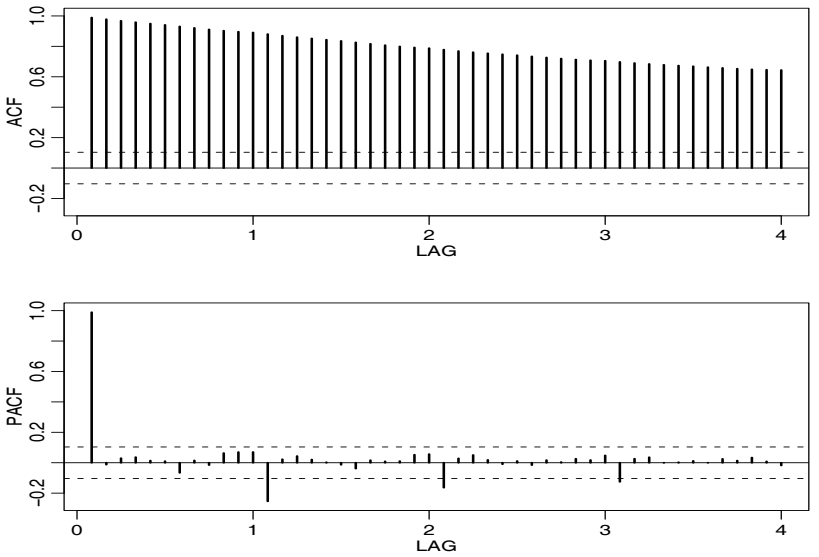

Fig. 3.22. ACF and PACF of the production series.

think first in terms of finding difference operators that produce a roughly stationary series and then in terms of finding a set of simple autoregressive moving average or multiplicative seasonal ARMA to fit the resulting residual series. Differencing operations are applied first, and then the residuals are constructed from a series of reduced length. Next, the ACF and the PACF of these residuals are evaluated. Peaks that appear in these functions can often be eliminated by fitting an autoregressive or moving average component in accordance with the general properties of Tables 3.1 and 3.2. In considering whether the model is satisfactory, the diagnostic techniques discussed in §3.8 still apply.

Example 3.46 The Federal Reserve Board Production Index

A problem of great interest in economics involves first identifying a model within the Box–Jenkins class for a given time series and then producing forecasts based on the model. For example, we might consider applying this methodology to the Federal Reserve Board Production Index shown in Figure 3.21. For demonstration purposes only, the ACF and PACF for this series are shown in Figure 3.22. We note that the trend in the data, the slow decay in the ACF, and the fact that the PACF at the first lag is nearly 1, all indicate nonstationary behavior.

Following the recommended procedure, a first difference was taken, and the ACF and PACF of the first difference

$$\nabla x_t = x_t - x_{t-1}$$

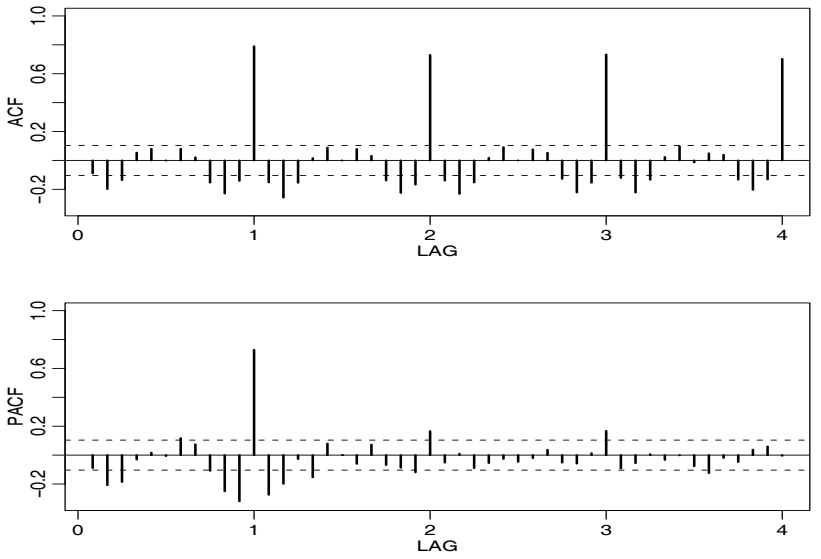

Fig. 3.23. ACF and PACF of differenced production, $(1 - B)x_t$.

are shown in Figure 3.23. Noting the peaks at seasonal lags, $h = 1s, 2s, 3s, 4s$ where $s = 12$ (i.e., $h = 12, 24, 36, 48$) with relatively slow decay suggests a seasonal difference. Figure 3.24 shows the ACF and PACF of the seasonal difference of the differenced production, say,

$$\nabla_{12}\nabla x_t = (1 - B^{12})(1 - B)x_t.$$

First, concentrating on the seasonal ($s = 12$) lags, the characteristics of the ACF and PACF of this series tend to show a strong peak at $h = 1s$ in the autocorrelation function, with smaller peaks appearing at $h = 2s, 3s$, combined with peaks at $h = 1s, 2s, 3s, 4s$ in the partial autocorrelation function. It appears that either

(i) the ACF is cutting off after lag $1s$ and the PACF is tailing off in the seasonal lags,
(ii) the ACF is cutting off after lag $3s$ and the PACF is tailing off in the seasonal lags, or
(iii) the ACF and PACF are both tailing off in the seasonal lags.

Using Table 3.3, this suggests either (i) an SMA of order $Q = 1$, (ii) an SMA of order $Q = 3$, or (iii) an SARMA of orders $P = 2$ (because of the two spikes in the PACF) and $Q = 1$.

Next, inspecting the ACF and the PACF at the within season lags, $h = 1, \ldots, 11$, it appears that either (a) both the ACF and PACF are tailing off, or (b) that the PACF cuts off at lag 2. Based on Table 3.1, this result indicates that we should either consider fitting a model (a) with both $p > 0$ and $q > 0$ for the nonseasonal components, say $p = 1, q = 1$, or (b) $p =$

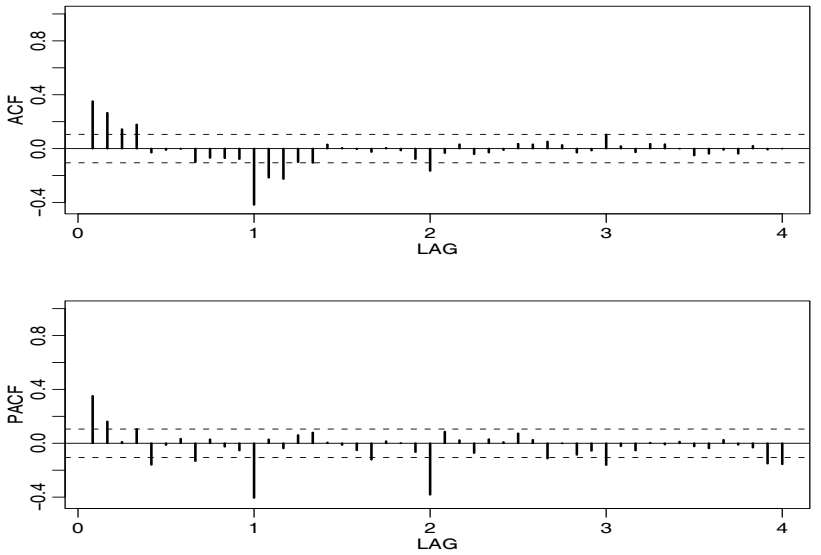

Fig. 3.24. ACF and PACF of first differenced and then seasonally differenced production, $(1-B)(1-B^{12})x_t$.

$2, q = 0$. It turns out that there is little difference in the results for case (a) and (b), but that (b) is slightly better, so we will concentrate on case (b).

Fitting the three models suggested by these observations we obtain:

(i) ARIMA$(2,1,0) \times (0,1,1)_{12}$:
AIC= 1.372, AICc= 1.378, BIC= .404

(ii) ARIMA$(2,1,0) \times (0,1,3)_{12}$:
AIC= 1.299, AICc= 1.305, BIC= .351

(iii) ARIMA$(2,1,0) \times (2,1,1)_{12}$:
AIC= 1.326, AICc= 1.332, BIC= .379

The ARIMA$(2,1,0) \times (0,1,3)_{12}$ is the preferred model, and the fitted model in this case is

$$(1 - .30_{(.05)}B - .11_{(.05)}B^2)\nabla_{12}\nabla\widehat{x}_t = (1 - .74_{(.05)}B^{12} - .14_{(.06)}B^{24} + .28_{(.05)}B^{36})\widehat{w}_t$$

with $\widehat{\sigma}_w^2 = 1.312$.

The diagnostics for the fit are displayed in Figure 3.25. We note the few outliers in the series as exhibited in the plot of the standardized residuals and their normal Q-Q plot, and a small amount of autocorrelation that still remains (although not at the seasonal lags) but otherwise, the model fits well. Finally, forecasts based on the fitted model for the next 12 months are shown in Figure 3.26.

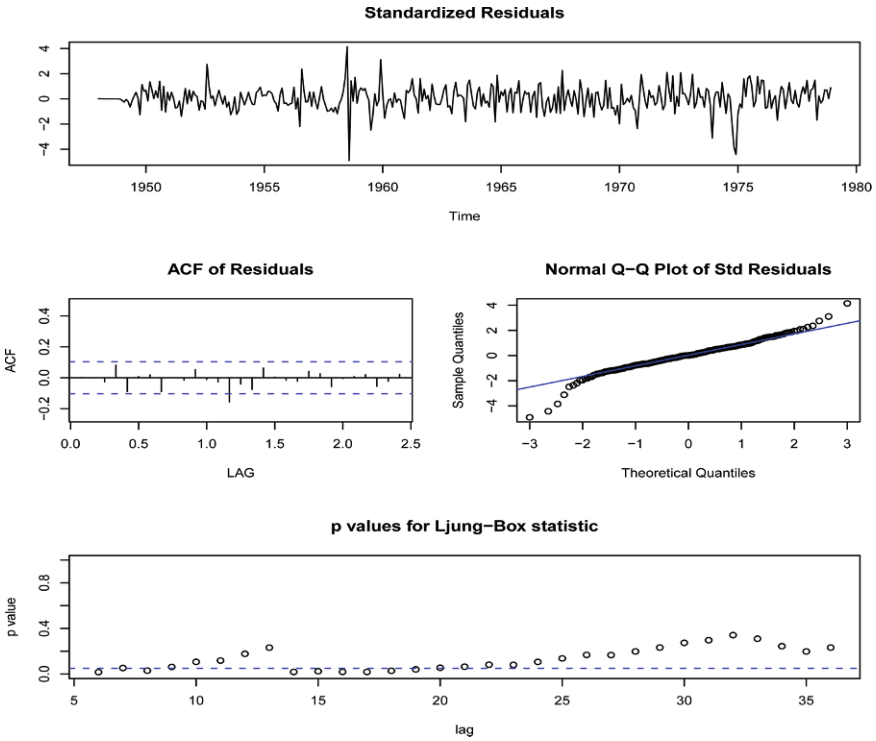

Fig. 3.25. Diagnostics for the ARIMA$(2,1,0) \times (0,1,3)_{12}$ fit on the Production Index.

The following R code can be used to perform the analysis.

```
acf2(prodn, 48)
acf2(diff(prodn), 48)
acf2(diff(diff(prodn), 12), 48)
sarima(prodn, 2, 1, 1, 0, 1, 3, 12)  # fit model (ii)
sarima.for(prodn, 12, 2, 1, 1, 0, 1, 3, 12) # forecast
```

Problems

Section 3.2

3.1 For an MA(1), $x_t = w_t + \theta w_{t-1}$, show that $|\rho_x(1)| \leq 1/2$ for any number θ. For which values of θ does $\rho_x(1)$ attain its maximum and minimum?

3.2 Let w_t be white noise with variance σ_w^2 and let $|\phi| < 1$ be a constant. Consider the process $x_1 = w_1$, and

$$x_t = \phi x_{t-1} + w_t, \quad t = 2, 3, \ldots .$$

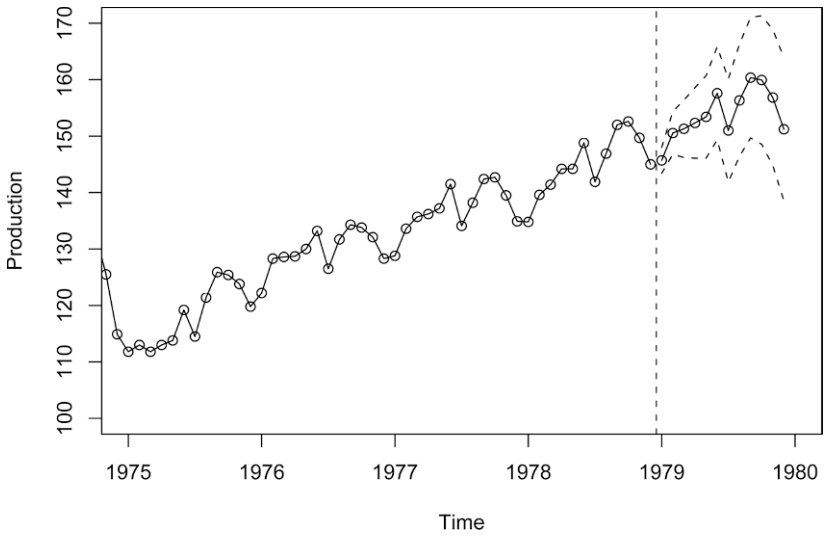

Fig. 3.26. Forecasts and limits for production index. The vertical dotted line separates the data from the predictions.

(a) Find the mean and the variance of $\{x_t,\ t = 1, 2, \ldots\}$. Is x_t stationary?
(b) Show

$$\text{corr}(x_t, x_{t-h}) = \phi^h \left[\frac{\text{var}(x_{t-h})}{\text{var}(x_t)}\right]^{1/2}$$

for $h \geq 0$.
(c) Argue that for large t,

$$\text{var}(x_t) \approx \frac{\sigma_w^2}{1-\phi^2}$$

and

$$\text{corr}(x_t, x_{t-h}) \approx \phi^h, \quad h \geq 0,$$

so in a sense, x_t is "asymptotically stationary."
(d) Comment on how you could use these results to simulate n observations of a stationary Gaussian AR(1) model from simulated iid N(0,1) values.
(e) Now suppose $x_1 = w_1/\sqrt{1-\phi^2}$. Is this process stationary?

3.3 Verify the calculations made in Example 3.3:

(a) Let $x_t = \phi x_{t-1} + w_t$ where $|\phi| > 1$ and $w_t \sim$ iid $\text{N}(0, \sigma_w^2)$. Show $E(x_t) = 0$ and $\gamma_x(h) = \sigma_w^2 \phi^{-2}\, \phi^{-h}/(1-\phi^{-2})$.
(b) Let $y_t = \phi^{-1} y_{t-1} + v_t$ where $v_t \sim$ iid $\text{N}(0, \sigma_w^2 \phi^{-2})$ and ϕ and σ_w are as in part (a). Argue that y_t is causal with the same mean function and autocovariance function as x_t.

3.4 Identify the following models as ARMA(p, q) models (watch out for parameter redundancy), and determine whether they are causal and/or invertible:

(a) $x_t = .80x_{t-1} - .15x_{t-2} + w_t - .30w_{t-1}$.
(b) $x_t = x_{t-1} - .50x_{t-2} + w_t - w_{t-1}$.

3.5 Verify the causal conditions for an AR(2) model given in (3.28). That is, show that an AR(2) is causal if and only if (3.28) holds.

Section 3.3

3.6 For the AR(2) model given by $x_t = -.9x_{t-2} + w_t$, find the roots of the autoregressive polynomial, and then sketch the ACF, $\rho(h)$.

3.7 For the AR(2) series shown below, use the results of Example 3.9 to determine a set of difference equations that can be used to find the ACF $\rho(h)$, $h = 0, 1, \ldots$; solve for the constants in the ACF using the initial conditions. Then plot the ACF values to lag 10 (use `ARMAacf` as a check on your answers).

(a) $x_t + 1.6x_{t-1} + .64x_{t-2} = w_t$.
(b) $x_t - .40x_{t-1} - .45x_{t-2} = w_t$.
(c) $x_t - 1.2x_{t-1} + .85x_{t-2} = w_t$.

Section 3.4

3.8 Verify the calculations for the autocorrelation function of an ARMA$(1, 1)$ process given in Example 3.13. Compare the form with that of the ACF for the ARMA$(1, 0)$ and the ARMA$(0, 1)$ series. Plot (or sketch) the ACFs of the three series on the same graph for $\phi = .6$, $\theta = .9$, and comment on the diagnostic capabilities of the ACF in this case.

3.9 Generate $n = 100$ observations from each of the three models discussed in Problem 3.8. Compute the sample ACF for each model and compare it to the theoretical values. Compute the sample PACF for each of the generated series and compare the sample ACFs and PACFs with the general results given in Table 3.1.

Section 3.5

3.10 Let x_t represent the cardiovascular mortality series (`cmort`) discussed in Chapter 2, Example 2.2.

(a) Fit an AR(2) to x_t using linear regression as in Example 3.17.
(b) Assuming the fitted model in (a) is the true model, find the forecasts over a four-week horizon, x^n_{n+m}, for $m = 1, 2, 3, 4$, and the corresponding 95% prediction intervals.

3.11 Consider the MA(1) series

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

where w_t is white noise with variance σ_w^2.

(a) Derive the minimum mean-square error one-step forecast based on the infinite past, and determine the mean-square error of this forecast.
(b) Let $\widetilde{x}_{n+1}^n$ be the truncated one-step-ahead forecast as given in (3.92). Show that

$$E\left[(x_{n+1} - \widetilde{x}_{n+1}^n)^2\right] = \sigma^2(1 + \theta^{2+2n}).$$

Compare the result with (a), and indicate how well the finite approximation works in this case.

3.12 In the context of equation (3.63), show that, if $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then Γ_n is positive definite.

3.13 Suppose x_t is stationary with zero mean and recall the definition of the PACF given by (3.55) and (3.56). That is, let

$$\epsilon_t = x_t - \sum_{i=1}^{h-1} a_i x_{t-i}$$

and

$$\delta_{t-h} = x_{t-h} - \sum_{j=1}^{h-1} b_j x_{t-j}$$

be the two residuals where $\{a_1, \ldots, a_{h-1}\}$ and $\{b_1, \ldots, b_{h-1}\}$ are chosen so that they minimize the mean-squared errors

$$E[\epsilon_t^2] \quad \text{and} \quad E[\delta_{t-h}^2].$$

The PACF at lag h was defined as the cross-correlation between ϵ_t and δ_{t-h}; that is,

$$\phi_{hh} = \frac{E(\epsilon_t \delta_{t-h})}{\sqrt{E(\epsilon_t^2) E(\delta_{t-h}^2)}}.$$

Let R_h be the $h \times h$ matrix with elements $\rho(i-j), i, j = 1, \ldots, h$, and let $\boldsymbol{\rho}_h = (\rho(1), \rho(2), \ldots, \rho(h))'$ be the vector of lagged autocorrelations, $\rho(h) = \mathrm{corr}(x_{t+h}, x_t)$. Let $\widetilde{\boldsymbol{\rho}}_h = (\rho(h), \rho(h-1), \ldots, \rho(1))'$ be the reversed vector. In addition, let x_t^h denote the BLP of x_t given $\{x_{t-1}, \ldots, x_{t-h}\}$:

$$x_t^h = \alpha_{h1} x_{t-1} + \cdots + \alpha_{hh} x_{t-h},$$

as described in Property 3.3. Prove

$$\phi_{hh} = \frac{\rho(h) - \widetilde{\boldsymbol{\rho}}_{h-1}' R_{h-1}^{-1} \boldsymbol{\rho}_h}{1 - \widetilde{\boldsymbol{\rho}}_{h-1}' R_{h-1}^{-1} \widetilde{\boldsymbol{\rho}}_{h-1}} = \alpha_{hh}.$$

In particular, this result proves Property 3.4.

Hint: Divide the prediction equations [see (3.63)] by $\gamma(0)$ and write the matrix equation in the partitioned form as

$$\begin{pmatrix} R_{h-1} & \widetilde{\boldsymbol{\rho}}_{h-1} \\ \widetilde{\boldsymbol{\rho}}_{h-1}' & \rho(0) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \alpha_{hh} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\rho}_{h-1} \\ \rho(h) \end{pmatrix},$$

where the $h \times 1$ vector of coefficients $\boldsymbol{\alpha} = (\alpha_{h1}, \ldots, \alpha_{hh})'$ is partitioned as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \alpha_{hh})'$.

3.14 Suppose we wish to find a prediction function $g(x)$ that minimizes

$$MSE = E[(y - g(x))^2],$$

where x and y are jointly distributed random variables with density function $f(x, y)$.

(a) Show that MSE is minimized by the choice

$$g(x) = E(y \mid x).$$

Hint:

$$MSE = \int \left[\int (y - g(x))^2 f(y|x) dy \right] f(x) dx.$$

(b) Apply the above result to the model

$$y = x^2 + z,$$

where x and z are independent zero-mean normal variables with variance one. Show that $MSE = 1$.

(c) Suppose we restrict our choices for the function $g(x)$ to linear functions of the form

$$g(x) = a + bx$$

and determine a and b to minimize MSE. Show that $a = 1$ and

$$b = \frac{E(xy)}{E(x^2)} = 0$$

and $MSE = 3$. What do you interpret this to mean?

3.15 For an AR(1) model, determine the general form of the m-step-ahead forecast x_{t+m}^t and show

$$E[(x_{t+m} - x_{t+m}^t)^2] = \sigma_w^2 \frac{1 - \phi^{2m}}{1 - \phi^2}.$$

3.16 Consider the ARMA(1,1) model discussed in Example 3.7, equation (3.27); that is, $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Show that truncated prediction as defined in (3.91) is equivalent to truncated prediction using the recursive formula (3.92).

3.17 Verify statement (3.87), that for a fixed sample size, the ARMA prediction errors are correlated.

Section 3.6

3.18 Fit an AR(2) model to the cardiovascular mortality series (`cmort`) discussed in Chapter 2, Example 2.2. using linear regression and using Yule–Walker.

(a) Compare the parameter estimates obtained by the two methods.
(b) Compare the estimated standard errors of the coefficients obtained by linear regression with their corresponding asymptotic approximations, as given in Property 3.10.

3.19 Suppose $x_1, \ldots, x_n$ are observations from an AR(1) process with $\mu = 0$.

(a) Show the backcasts can be written as $x_t^n = \phi^{1-t}x_1$, for $t \leq 1$.
(b) In turn, show, for $t \leq 1$, the backcasted errors are

$$\widehat{w}_t(\phi) = x_t^n - \phi x_{t-1}^n = \phi^{1-t}(1-\phi^2)x_1.$$

(c) Use the result of (b) to show $\sum_{t=-\infty}^{1} \widehat{w}_t^2(\phi) = (1-\phi^2)x_1^2$.
(d) Use the result of (c) to verify the unconditional sum of squares, $S(\phi)$, can be written as $\sum_{t=-\infty}^{n} \widehat{w}_t^2(\phi)$.
(e) Find x_t^{t-1} and r_t for $1 \leq t \leq n$, and show that

$$S(\phi) = \sum_{t=1}^{n} (x_t - x_t^{t-1})^2 \,/\, r_t.$$

3.20 Repeat the following numerical exercise three times. Generate $n = 500$ observations from the ARMA model given by

$$x_t = .9x_{t-1} + w_t - .9w_{t-1},$$

with $w_t \sim$ iid $\mathrm{N}(0,1)$. Plot the simulated data, compute the sample ACF and PACF of the simulated data, and fit an ARMA$(1,1)$ model to the data. What happened and how do you explain the results?

3.21 Generate 10 realizations of length $n = 200$ each of an ARMA(1,1) process with $\phi = .9, \theta = .5$ and $\sigma^2 = 1$. Find the MLEs of the three parameters in each case and compare the estimators to the true values.

3.22 Generate $n = 50$ observations from a Gaussian AR(1) model with $\phi = .99$ and $\sigma_w = 1$. Using an estimation technique of your choice, compare the approximate asymptotic distribution of your estimate (the one you would use for inference) with the results of a bootstrap experiment (use $B = 200$).

3.23 Using Example 3.31 as your guide, find the Gauss–Newton procedure for estimating the autoregressive parameter, ϕ, from the AR(1) model, $x_t = \phi x_{t-1} + w_t$, given data $x_1, \ldots, x_n$. Does this procedure produce the unconditional or the conditional estimator? *Hint:* Write the model as $w_t(\phi) = x_t - \phi x_{t-1}$; your solution should work out to be a non-recursive procedure.

3.24 Consider the stationary series generated by

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

where $E(x_t) = \mu$, $|\theta| < 1, |\phi| < 1$ and the w_t are iid random variables with zero mean and variance σ_w^2.

(a) Determine the mean as a function of α for the above model. Find the autocovariance and ACF of the process x_t, and show that the process is weakly stationary. Is the process strictly stationary?

(b) Prove the limiting distribution as $n \to \infty$ of the sample mean,

$$\bar{x} = n^{-1} \sum_{t=1}^{n} x_t,$$

is normal, and find its limiting mean and variance in terms of α, ϕ, θ, and σ_w^2. (Note: This part uses results from Appendix A.)

3.25 A problem of interest in the analysis of geophysical time series involves a simple model for observed data containing a signal and a reflected version of the signal with unknown amplification factor a and unknown time delay δ. For example, the depth of an earthquake is proportional to the time delay δ for the P wave and its reflected form pP on a seismic record. Assume the signal, say s_t, is white and Gaussian with variance σ_s^2, and consider the generating model

$$x_t = s_t + a s_{t-\delta}.$$

(a) Prove the process x_t is stationary. If $|a| < 1$, show that

$$s_t = \sum_{j=0}^{\infty} (-a)^j x_{t-\delta j}$$

is a mean square convergent representation for the signal s_t, for $t = 1, \pm 1, \pm 2, \ldots$.

(b) If the time delay δ is assumed to be known, suggest an approximate computational method for estimating the parameters a and σ_s^2 using maximum likelihood and the Gauss–Newton method.

(c) If the time delay δ is an unknown integer, specify how we could estimate the parameters including δ. Generate a $n = 500$ point series with $a = .9$, $\sigma_w^2 = 1$ and $\delta = 5$. Estimate the integer time delay δ by searching over $\delta = 3, 4, \ldots, 7$.

3.26 *Forecasting with estimated parameters:* Let $x_1, x_2, \ldots, x_n$ be a sample of size n from a causal AR(1) process, $x_t = \phi x_{t-1} + w_t$. Let $\widehat{\phi}$ be the Yule–Walker estimator of ϕ.

(a) Show $\widehat{\phi} - \phi = O_p(n^{-1/2})$. See Appendix A for the definition of $O_p(\cdot)$.
(b) Let x_{n+1}^n be the one-step-ahead forecast of x_{n+1} given the data $x_1, \ldots, x_n$, based on the known parameter, ϕ, and let $\widehat{x}_{n+1}^n$ be the one-step-ahead forecast when the parameter is replaced by $\widehat{\phi}$. Show $x_{n+1}^n - \widehat{x}_{n+1}^n = O_p(n^{-1/2})$.

Section 3.7

3.27 Suppose

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_q t^q + x_t, \quad \beta_q \neq 0,$$

where x_t is stationary. First, show that $\nabla^k x_t$ is stationary for any $k = 1, 2, \ldots$, and then show that $\nabla^k y_t$ is not stationary for $k < q$, but is stationary for $k \geq q$.

3.28 Verify that the IMA(1,1) model given in (3.147) can be inverted and written as (3.148).

3.29 For the ARIMA$(1, 1, 0)$ model with drift, $(1 - \phi B)(1 - B)x_t = \delta + w_t$, let $y_t = (1 - B)x_t = \nabla x_t$.

(a) Noting that y_t is AR(1), show that, for $j \geq 1$,

$$y_{n+j}^n = \delta\,[1 + \phi + \cdots + \phi^{j-1}] + \phi^j\, y_n.$$

(b) Use part (a) to show that, for $m = 1, 2, \ldots$,

$$x_{n+m}^n = x_n + \frac{\delta}{1-\phi}\Big[m - \frac{\phi(1-\phi^m)}{(1-\phi)}\Big] + (x_n - x_{n-1})\frac{\phi(1-\phi^m)}{(1-\phi)}.$$

Hint: From (a), $x_{n+j}^n - x_{n+j-1}^n = \delta\frac{1-\phi^j}{1-\phi} + \phi^j(x_n - x_{n-1})$. Now sum both sides over j from 1 to m.
(c) Use (3.144) to find P_{n+m}^n by first showing that $\psi_0^* = 1$, $\psi_1^* = (1 + \phi)$, and $\psi_j^* - (1 + \phi)\psi_{j-1}^* + \phi\psi_{j-2}^* = 0$ for $j \geq 2$, in which case $\psi_j^* = \frac{1-\phi^{j+1}}{1-\phi}$, for $j \geq 1$. Note that, as in Example 3.36, equation (3.144) is exact here.

3.30 For the logarithm of the glacial varve data, say, x_t, presented in Example 3.32, use the first 100 observations and calculate the EWMA, $\widetilde{x}_{t+1}^t$, given in (3.150) for $t = 1, \ldots, 100$, using $\lambda = .25, .50$, and $.75$, and plot the EWMAs and the data superimposed on each other. Comment on the results.

Section 3.8

3.31 In Example 3.39, we presented the diagnostics for the MA(2) fit to the GNP growth rate series. Using that example as a guide, complete the diagnostics for the AR(1) fit.

3.32 Crude oil prices in dollars per barrel are in `oil`; see Appendix R for more details. Fit an ARIMA(p, d, q) model to the growth rate performing all necessary diagnostics. Comment.

3.33 Fit an ARIMA(p, d, q) model to the global temperature data `gtemp` performing all of the necessary diagnostics. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

3.34 One of the series collected along with particulates, temperature, and mortality described in Example 2.2 is the sulfur dioxide series, `so2`. Fit an ARIMA(p, d, q) model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast the data into the future four time periods ahead (about one month) and calculate 95% prediction intervals for each of the four forecasts. Comment.

Section 3.9

3.35 Consider the ARIMA model

$$x_t = w_t + \Theta w_{t-2}.$$

(a) Identify the model using the notation ARIMA$(p, d, q) \times (P, D, Q)_s$.
(b) Show that the series is invertible for $|\Theta| < 1$, and find the coefficients in the representation

$$w_t = \sum_{k=0}^{\infty} \pi_k x_{t-k}.$$

(c) Develop equations for the m-step ahead forecast, $\widetilde{x}_{n+m}$, and its variance based on the infinite past, $x_n, x_{n-1}, \ldots$.

3.36 Plot (or sketch) the ACF of the seasonal ARIMA$(0, 1) \times (1, 0)_{12}$ model with $\Phi = .8$ and $\theta = .5$.

3.37 Fit a seasonal ARIMA model of your choice to the unemployment data (`unemp`) displayed in Figure 3.21. Use the estimated model to forecast the next 12 months.

3.38 Fit a seasonal ARIMA model of your choice to the U.S. Live Birth Series (`birth`). Use the estimated model to forecast the next 12 months.

3.39 Fit an appropriate seasonal ARIMA model to the log-transformed Johnson and Johnson earnings series (`jj`) of Example 1.1. Use the estimated model to forecast the next 4 quarters.

The following problems require supplemental material given in Appendix B.

3.40 Suppose $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$, where $\phi_p \neq 0$ and w_t is white noise such that w_t is uncorrelated with $\{x_k; k < t\}$. Use the Projection Theorem to show that, for $n > p$, the BLP of x_{n+1} on $\overline{\text{sp}}\{x_k, k \leq n\}$ is

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

3.41 Use the Projection Theorem to derive the Innovations Algorithm, Property 3.6, equations (3.77)-(3.79). Then, use Theorem B.2 to derive the m-step-ahead forecast results given in (3.80) and (3.81).

3.42 Consider the series $x_t = w_t - w_{t-1}$, where w_t is a white noise process with mean zero and variance σ_w^2. Suppose we consider the problem of predicting x_{n+1}, based on only $x_1, \ldots, x_n$. Use the Projection Theorem to answer the questions below.

(a) Show the best linear predictor is

$$x_{n+1}^n = -\frac{1}{n+1} \sum_{k=1}^{n} k\, x_k.$$

(b) Prove the mean square error is

$$E(x_{n+1} - x_{n+1}^n)^2 = \frac{n+2}{n+1}\, \sigma_w^2.$$

3.43 Use Theorem B.2 and B.3 to verify (3.116).

3.44 Prove Theorem B.2.

3.45 Prove Property 3.2.

4

Spectral Analysis and Filtering

4.1 Introduction

The notion that a time series exhibits repetitive or regular behavior over time is of fundamental importance because it distinguishes time series analysis from classical statistics, which assumes complete independence over time. We have seen how dependence over time can be introduced through models that describe in detail the way certain empirical data behaves, even to the extent of producing forecasts based on the models. It is natural that models based on predicting the present as a regression on the past, such as are provided by the celebrated ARIMA or state-space forms, will be attractive to statisticians, who are trained to view nature in terms of linear models. In fact, the difference equations used to represent these kinds of models are simply the discrete versions of linear differential equations that may, in some instances, provide the ideal physical model for a certain phenomenon. An alternate version of the way nature behaves exists, however, and is based on a decomposition of an empirical series into its regular components.

In this chapter, we argue, the concept of regularity of a series can best be expressed in terms of periodic variations of the underlying phenomenon that produced the series, expressed as Fourier frequencies being driven by sines and cosines. Such a possibility was discussed in Chapters 1 and 2. From a regression point of view, we may imagine a system responding to various driving frequencies by producing linear combinations of sine and cosine functions. Expressed in these terms, the time domain approach may be thought of as regression of the present on the past, whereas the frequency domain approach may be considered as regression of the present on periodic sines and cosines. The frequency domain approaches are the focus of this chapter and Chapter 7. To illustrate the two methods for generating series with a single primary periodic component, consider Figure 1.9, which was generated from a simple second-order autoregressive model, and the middle and bottom panels of Figure 1.11, which were generated by adding a cosine wave with a period of 50 points to white noise. Both series exhibit strong periodic fluctuations,

R.H. Shumway and D.S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-7865-3_4,

illustrating that both models can generate time series with regular behavior. As discussed in Example 2.8, a fundamental objective of spectral analysis is to identify the dominant frequencies in a series and to find an explanation of the system from which the measurements were derived.

Of course, the primary justification for any alternate model must lie in its potential for explaining the behavior of some empirical phenomenon. In this sense, an explanation involving only a few kinds of primary oscillations becomes simpler and more physically meaningful than a collection of parameters estimated for some selected difference equation. It is the tendency of observed data to show periodic kinds of fluctuations that justifies the use of frequency domain methods. Many of the examples in §1.2 are time series representing real phenomena that are driven by periodic components. The speech recording of the syllable *aa...hh* in Figure 1.3 contains a complicated mixture of frequencies related to the opening and closing of the glottis. Figure 1.5 shows the monthly SOI, which we later explain as a combination of two kinds of periodicities, a seasonal periodic component of 12 months and an El Niño component of about three to five years. Of fundamental interest is the return period of the El Niño phenomenon, which can have profound effects on local climate. Also of interest is whether the different periodic components of the new fish population depend on corresponding seasonal and El Niño-type oscillations. We introduce the coherence as a tool for relating the common periodic behavior of two series. Seasonal periodic components are often pervasive in economic time series; this phenomenon can be seen in the quarterly earnings series shown in Figure 1.1. In Figure 1.6, we see the extent to which various parts of the brain will respond to a periodic stimulus generated by having the subject do alternate left and right finger tapping. Figure 1.7 shows series from an earthquake and a nuclear explosion. The relative amounts of energy at various frequencies for the two phases can produce statistics, useful for discriminating between earthquakes and explosions.

In this chapter, we summarize an approach to handling correlation generated in stationary time series that begins by transforming the series to the frequency domain. This simple linear transformation essentially matches sines and cosines of various frequencies against the underlying data and serves two purposes as discussed in Examples 2.8 and 2.9. The periodogram that was introduced in Example 2.9 has its population counterpart called the power spectrum, and its estimation is a main goal of spectral analysis. Another purpose of exploring this topic is statistical convenience resulting from the periodic components being nearly uncorrelated. This property facilitates writing likelihoods based on classical statistical methods.

An important part of analyzing data in the frequency domain, as well as the time domain, is the investigation and exploitation of the properties of the time-invariant linear filter. This special linear transformation is used similarly to linear regression in conventional statistics, and we use many of the same terms in the time series context. We have previously mentioned the coherence as a measure of the relation between two series at a given frequency, and

we show later that this coherence also measures the performance of the best linear filter relating the two series. Linear filtering can also be an important step in isolating a signal embedded in noise. For example, the lower panels of Figure 1.11 contain a signal contaminated with an additive noise, whereas the upper panel contains the pure signal. It might also be appropriate to ask whether a linear filter transformation exists that could be applied to the lower panel to produce a series closer to the signal in the upper panel. The use of filtering for reducing noise will also be a part of the presentation in this chapter. We emphasize, throughout, the analogy between filtering techniques and conventional linear regression.

Many frequency scales will often coexist, depending on the nature of the problem. For example, in the Johnson & Johnson data set in Figure 1.1, the predominant frequency of oscillation is one cycle per year (4 quarters), or .25 cycles per observation. The predominant frequency in the SOI and fish populations series in Figure 1.5 is also one cycle per year, but this corresponds to 1 cycle every 12 months, or .083 cycles per observation. For simplicity, we measure frequency, ω, at cycles per time point and discuss the implications of certain frequencies in terms of the problem context. Of descriptive interest is the *period* of a time series, defined as the number of points in a cycle, i.e., $1/\omega$. Hence, the predominant period of the Johnson & Johnson series is $1/.25$ or 4 quarters per cycle, whereas the predominant period of the SOI series is 12 months per cycle.

4.2 Cyclical Behavior and Periodicity

As previously mentioned, we have already encountered the notion of periodicity in numerous examples in Chapters 1, 2 and 3. The general notion of periodicity can be made more precise by introducing some terminology. In order to define the rate at which a series oscillates, we first define a cycle as one complete period of a sine or cosine function defined over a unit time interval. As in (1.5), we consider the periodic process

$$x_t = A\cos(2\pi\omega t + \phi) \tag{4.1}$$

for $t = 0, \pm 1, \pm 2, \ldots$, where ω is a frequency index, defined in cycles per unit time with A determining the height or *amplitude* of the function and ϕ, called the *phase*, determining the start point of the cosine function. We can introduce random variation in this time series by allowing the amplitude and phase to vary randomly.

As discussed in Example 2.8, for purposes of data analysis, it is easier to use a trigonometric identity[1] and write (4.1) as

[1] $\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$.

$$x_t = U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t), \tag{4.2}$$

where $U_1 = A\cos\phi$ and $U_2 = -A\sin\phi$ are often taken to be normally distributed random variables. In this case, the amplitude is $A = \sqrt{U_1^2 + U_2^2}$ and the phase is $\phi = \tan^{-1}(-U_2/U_1)$. From these facts we can show that if, and only if, in (4.1), A and ϕ are independent random variables, where A^2 is chi-squared with 2 degrees of freedom, and ϕ is uniformly distributed on $(-\pi, \pi)$, then U_1 and U_2 are independent, standard normal random variables (see Problem 4.2).

The above random process is also a function of its frequency, defined by the parameter ω. The frequency is measured in cycles per unit time, or in cycles per point in the above illustration. For $\omega = 1$, the series makes one cycle per time unit; for $\omega = .50$, the series makes a cycle every two time units; for $\omega = .25$, every four units, and so on. In general, for data that occur at discrete time points will need at least two points to determine a cycle, so the highest frequency of interest is .5 cycles per point. This frequency is called the *folding frequency* and defines the highest frequency that can be seen in discrete sampling. Higher frequencies sampled this way will appear at lower frequencies, called *aliases*; an example is the way a camera samples a rotating wheel on a moving automobile in a movie, in which the wheel appears to be rotating at a different rate. For example, movies are recorded at 24 frames per second. If the camera is filming a wheel that is rotating at the rate of 24 cycles per second (or 24 Hertz), the wheel will appear to stand still (that's about 110 miles per hour in case you were wondering).

Consider a generalization of (4.2) that allows mixtures of periodic series with multiple frequencies and amplitudes,

$$x_t = \sum_{k=1}^{q} \left[U_{k1} \cos(2\pi\omega_k t) + U_{k2} \sin(2\pi\omega_k t) \right], \tag{4.3}$$

where U_{k1}, U_{k2}, for $k = 1, 2, \ldots, q$, are independent zero-mean random variables with variances σ_k^2, and the ω_k are distinct frequencies. Notice that (4.3) exhibits the process as a sum of independent components, with variance σ_k^2 for frequency ω_k. Using the independence of the Us and the trig identity in footnote 1, it is easy to show[2] (Problem 4.3) that the autocovariance function of the process is

$$\gamma(h) = \sum_{k=1}^{q} \sigma_k^2 \cos(2\pi\omega_k h), \tag{4.4}$$

and we note the autocovariance function is the sum of periodic components with weights proportional to the variances σ_k^2. Hence, x_t is a mean-zero stationary processes with variance

[2] For example, for x_t in (4.2) we have $\mathrm{cov}(x_{t+h}, x_t) = \sigma^2\{\cos(2\pi\omega[t+h])\cos(2\pi\omega t) + \sin(2\pi\omega[t+h])\sin(2\pi\omega t)\} = \sigma^2\cos(2\pi\omega h)$, noting that $\mathrm{cov}(U_1, U_2) = 0$.

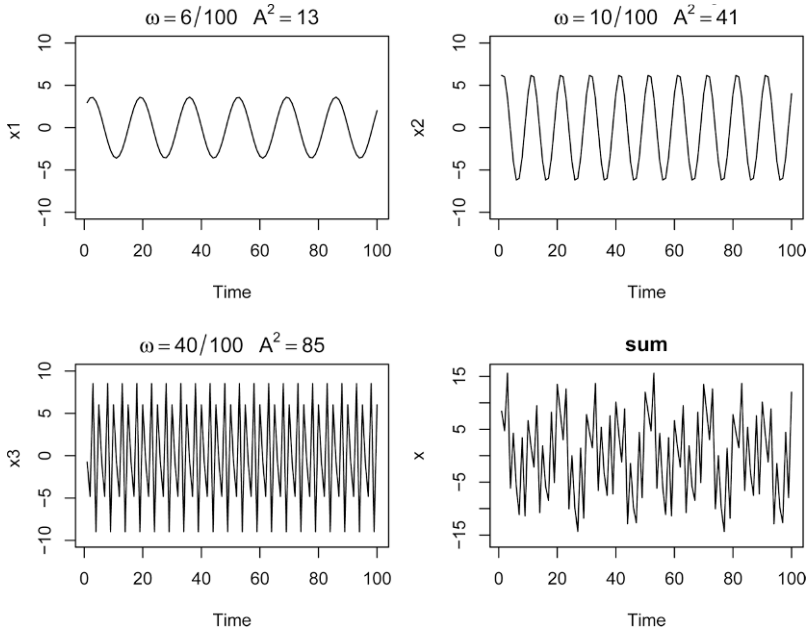

Fig. 4.1. Periodic components and their sum as described in Example 4.1.

$$\gamma(0) = E(x_t^2) = \sum_{k=1}^{q} \sigma_k^2, \tag{4.5}$$

which exhibits the overall variance as a sum of variances of each of the component parts.

Example 4.1 A Periodic Series

Figure 4.1 shows an example of the mixture (4.3) with $q = 3$ constructed in the following way. First, for $t = 1, \ldots, 100$, we generated three series

$$x_{t1} = 2\cos(2\pi t\, 6/100) + 3\sin(2\pi t\, 6/100)$$
$$x_{t2} = 4\cos(2\pi t\, 10/100) + 5\sin(2\pi t\, 10/100)$$
$$x_{t3} = 6\cos(2\pi t\, 40/100) + 7\sin(2\pi t\, 40/100)$$

These three series are displayed in Figure 4.1 along with the corresponding frequencies and squared amplitudes. For example, the squared amplitude of x_{t1} is $A^2 = 2^2 + 3^2 = 13$. Hence, the maximum and minimum values that x_{t1} will attain are $\pm\sqrt{13} = \pm 3.61$.

Finally, we constructed

$$x_t = x_{t1} + x_{t2} + x_{t3}$$

and this series is also displayed in Figure 4.1. We note that x_t appears to behave as some of the periodic series we saw in Chapters 1 and 2. The

systematic sorting out of the essential frequency components in a time series, including their relative contributions, constitutes one of the main objectives of spectral analysis.

The R code to reproduce Figure 4.1 is

```
x1 = 2*cos(2*pi*1:100*6/100) + 3*sin(2*pi*1:100*6/100)
x2 = 4*cos(2*pi*1:100*10/100) + 5*sin(2*pi*1:100*10/100)
x3 = 6*cos(2*pi*1:100*40/100) + 7*sin(2*pi*1:100*40/100)
x = x1 + x2 + x3
par(mfrow=c(2,2))
plot.ts(x1, ylim=c(-10,10), main=expression(omega==6/100~~~A^2==13))
plot.ts(x2, ylim=c(-10,10), main=expression(omega==10/100~~~A^2==41))
plot.ts(x3, ylim=c(-10,10), main=expression(omega==40/100~~~A^2==85))
plot.ts(x,  ylim=c(-16,16), main="sum")
```

Example 4.2 The Scaled Periodogram for Example 4.1

In §2.3, Example 2.9, we introduced the periodogram as a way to discover the periodic components of a time series. Recall that the scaled periodogram is given by

$$P(j/n) = \left(\frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi t j/n)\right)^2 + \left(\frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi t j/n)\right)^2, \qquad (4.6)$$

and it may be regarded as a measure of the squared correlation of the data with sinusoids oscillating at a frequency of $\omega_j = j/n$, or j cycles in n time points. Recall that we are basically computing the regression of the data on the sinusoids varying at the fundamental frequencies, j/n. As discussed in Example 2.9, the periodogram may be computed quickly using the fast Fourier transform (FFT), and there is no need to run repeated regressions.

The scaled periodogram of the data, x_t, simulated in Example 4.1 is shown in Figure 4.2, and it clearly identifies the three components x_{t1}, x_{t2}, and x_{t3} of x_t. Note that

$$P(j/n) = P(1 - j/n), \quad j = 0, 1, \ldots, n-1,$$

so there is a mirroring effect at the folding frequency of $1/2$; consequently, the periodogram is typically not plotted for frequencies higher than the folding frequency. In addition, note that the heights of the scaled periodogram shown in the figure are

$$P(6/100) = 13, \quad P(10/100) = 41, \quad P(40/100) = 85,$$

$P(j/n) = P(1-j/n)$ and $P(j/n) = 0$ otherwise. These are exactly the values of the squared amplitudes of the components generated in Example 4.1. This outcome suggests that the periodogram may provide some insight into the variance components, (4.5), of a real set of data.

Assuming the simulated data, `x`, were retained from the previous example, the R code to reproduce Figure 4.2 is

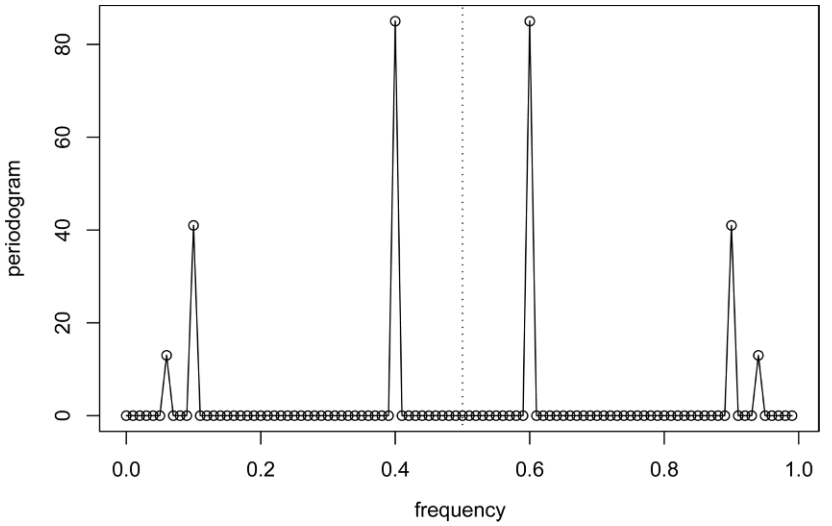

Fig. 4.2. Periodogram of the data generated in Example 4.1.

```
P = abs(2*fft(x)/100)^2;  Fr = 0:99/100
plot(Fr, P, type="o", xlab="frequency", ylab="periodogram")
```

If we consider the data x_t in Example 4.1 as a color (waveform) made up of primary colors x_{t1}, x_{t2}, x_{t3} at various strengths (amplitudes), then we might consider the periodogram as a prism that decomposes the color x_t into its primary colors (spectrum). Hence the term *spectral analysis*.

Another fact that may be of use in understanding the periodogram is that for any time series sample $x_1, \ldots, x_n$, where n is odd, we may write, *exactly*

$$x_t = a_0 + \sum_{j=1}^{(n-1)/2} \left[a_j \cos(2\pi t\, j/n) + b_j \sin(2\pi t\, j/n) \right], \tag{4.7}$$

for $t = 1, \ldots, n$ and suitably chosen coefficients. If n is even, the representation (4.7) can be modified by summing to $(n/2 - 1)$ and adding an additional component given by $a_{n/2} \cos(2\pi t\, 1/2) = a_{n/2}(-1)^t$. The crucial point here is that (4.7) is exact for any sample. Hence (4.3) may be thought of as an approximation to (4.7), the idea being that many of the coefficients in (4.7) may be close to zero. Recall from Example 2.9 that

$$P(j/n) = a_j^2 + b_j^2, \tag{4.8}$$

so the scaled periodogram indicates which components in (4.7) are large in magnitude and which components are small. We also saw (4.8) in Example 4.2.

The periodogram, which was introduced in Schuster (1898) and used in Schuster (1906) for studying the periodicities in the sunspot series (shown in

Figure 4.31 in the Problems section) is a sample based statistic. In Example 4.2, we discussed the fact that the periodogram may be giving us an idea of the variance components associated with each frequency, as presented in (4.5), of a time series. These variance components, however, are population parameters. The concepts of population parameters and sample statistics, as they relate to spectral analysis of time series can be generalized to cover stationary time series and that is the topic of the next section.

4.3 The Spectral Density

The idea that a time series is composed of periodic components, appearing in proportion to their underlying variances, is fundamental in the spectral representation given in Theorem C.2 of Appendix C. The result is quite technical because it involves stochastic integration; that is, integration with respect to a stochastic process. The essence of Theorem C.2 is that (4.3) is approximately true for any stationary time series. In other words, we have the following.

Property 4.1 Spectral Representation of a Stationary Process

In nontechnical terms, Theorem C.2 states that any stationary time series may be thought of, approximately, as the random superposition of sines and cosines oscillating at various frequencies.

Given that (4.3) is approximately true for all stationary time series, the next question is whether a meaningful representation for its autocovariance function, like the one displayed in (4.4), also exists. The answer is yes, and this representation is given in Theorem C.1 of Appendix C. The following example will help explain the result.

Example 4.3 A Periodic Stationary Process

Consider a periodic stationary random process given by (4.2), with a fixed frequency ω_0, say,

$$x_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t),$$

where U_1 and U_2 are independent zero-mean random variables with equal variance σ^2. The number of time periods needed for the above series to complete one cycle is exactly $1/\omega_0$, and the process makes exactly ω_0 cycles per point for $t = 0, \pm 1, \pm 2, \ldots$. It is easily shown that[3]

$$\begin{aligned}\gamma(h) &= \sigma^2 \cos(2\pi\omega_0 h) = \frac{\sigma^2}{2} e^{-2\pi i\omega_0 h} + \frac{\sigma^2}{2} e^{2\pi i\omega_0 h} \\ &= \int_{-1/2}^{1/2} e^{2\pi i\omega h} dF(\omega)\end{aligned}$$

[3] Some identities may be helpful here: $e^{i\alpha} = \cos(\alpha) + i\sin(\alpha)$ and consequently, $\cos(\alpha) = (e^{i\alpha} + e^{-i\alpha})/2$ and $\sin(\alpha) = (e^{i\alpha} - e^{-i\alpha})/2i$.

using a Riemann–Stieltjes integration, where $F(\omega)$ is the function defined by

$$F(\omega) = \begin{cases} 0 & \omega < -\omega_0, \\ \sigma^2/2 & -\omega_0 \le \omega < \omega_0, \\ \sigma^2 & \omega \ge \omega_0. \end{cases}$$

The function $F(\omega)$ behaves like a cumulative distribution function for a discrete random variable, except that $F(\infty) = \sigma^2 = \mathrm{var}(x_t)$ instead of one. In fact, $F(\omega)$ is a cumulative distribution function, not of probabilities, but rather of variances associated with the frequency ω_0 in an analysis of variance, with $F(\infty)$ being the total variance of the process x_t. Hence, we term $F(\omega)$ the *spectral distribution function.*

Theorem C.1 in Appendix C states that a representation such as the one given in Example 4.3 always exists for a stationary process. In particular, if x_t is stationary with autocovariance $\gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)]$, then there exists a unique monotonically increasing function $F(\omega)$, called the spectral distribution function, that is bounded, with $F(-\infty) = F(-1/2) = 0$, and $F(\infty) = F(1/2) = \gamma(0)$ such that

$$\gamma(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h}\, dF(\omega). \tag{4.9}$$

A more important situation we use repeatedly is the one covered by Theorem C.3, where it is shown that, subject to absolute summability of the autocovariance, the spectral distribution function is absolutely continuous with $dF(\omega) = f(\omega)\, d\omega$, and the representation (4.9) becomes the motivation for the property given below.

Property 4.2 The Spectral Density

If the autocovariance function, $\gamma(h)$, of a stationary process satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \tag{4.10}$$

then it has the representation

$$\gamma(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h}\, f(\omega)\, d\omega \quad h = 0, \pm 1, \pm 2, \ldots \tag{4.11}$$

as the inverse transform of the spectral density, which has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)\mathrm{e}^{-2\pi i\omega h} \quad -1/2 \le \omega \le 1/2. \tag{4.12}$$

This spectral density is the analogue of the probability density function; the fact that $\gamma(h)$ is non-negative definite ensures

$$f(\omega) \geq 0$$

for all ω (see Appendix C, Theorem C.3 for details). It follows immediately from (4.12) that

$$f(\omega) = f(-\omega) \quad \text{and} \quad f(\omega) = f(1-\omega),$$

verifying the spectral density is an even function of period one. Because of the evenness, we will typically only plot $f(\omega)$ for $\omega \geq 0$. In addition, putting $h = 0$ in (4.11) yields

$$\gamma(0) = \text{var}(x_t) = \int_{-1/2}^{1/2} f(\omega)\, d\omega,$$

which expresses the total variance as the integrated spectral density over all of the frequencies. We show later on, that a linear filter can isolate the variance in certain frequency intervals or bands.

Analogous to probability theory, $\gamma(h)$ in (4.11) is the characteristic function[4] of the spectral density $f(\omega)$ in (4.12). These facts should make it clear that, when the conditions of Property 4.2 are satisfied, *the autocovariance function, $\gamma(h)$, and the spectral density function, $f(\omega)$, contain the same information*. That information, however, is expressed in different ways. The autocovariance function expresses information in terms of lags, whereas the spectral density expresses the same information in terms of cycles. Some problems are easier to work with when considering lagged information and we would tend to handle those problems in the time domain. Nevertheless, other problems are easier to work with when considering periodic information and we would tend to handle those problems in the spectral domain.

We note that the autocovariance function, $\gamma(h)$, in (4.11) and the spectral density, $f(\omega)$, in (4.12) are Fourier transform pairs. In particular, this means that if $f(\omega)$ and $g(\omega)$ are two spectral densities for which

$$\gamma_f(h) = \int_{-1/2}^{1/2} f(\omega) \mathrm{e}^{2\pi i \omega h}\, d\omega = \int_{-1/2}^{1/2} g(\omega) \mathrm{e}^{2\pi i \omega h}\, d\omega = \gamma_g(h) \tag{4.13}$$

for all $h = 0, \pm 1, \pm 2, \ldots$, then

$$f(\omega) = g(\omega). \tag{4.14}$$

We also mention, at this point, that we have been focusing on the frequency ω, expressed in cycles per point rather than the more common (in statistics)

[4] If $M_X(\lambda) = E(e^{\lambda X})$ for $\lambda \in \mathbb{R}$ is the moment generating function of random variable X, then $\varphi_X(\lambda) = M_X(i\lambda)$ is the characteristic function.

alternative $\lambda = 2\pi\omega$ that would give radians per point. Finally, the absolute summability condition, (4.10), is not satisfied by (4.4), the example that we have used to introduce the idea of a spectral representation. The condition, however, is satisfied for ARMA models.

It is illuminating to examine the spectral density for the series that we have looked at in earlier discussions.

Example 4.4 White Noise Series

As a simple example, consider the theoretical power spectrum of a sequence of uncorrelated random variables, w_t, with variance σ_w^2. A simulated set of data is displayed in the top of Figure 1.8. Because the autocovariance function was computed in Example 1.16 as $\gamma_w(h) = \sigma_w^2$ for $h = 0$, and zero, otherwise, it follows from (4.12), that

$$f_w(\omega) = \sigma_w^2$$

for $-1/2 \leq \omega \leq 1/2$. Hence the process contains equal power at all frequencies. This property is seen in the realization, which seems to contain all different frequencies in a roughly equal mix. In fact, the name white noise comes from the analogy to white light, which contains all frequencies in the color spectrum at the same level of intensity. Figure 4.3 shows a plot of the white noise spectrum for $\sigma_w^2 = 1$.

If x_t is ARMA, its spectral density can be obtained explicitly using the fact that it is a linear process, i.e., $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. In the following property, we exhibit the form of the spectral density of an ARMA model. The proof of the property follows directly from the proof of a more general result, Property 4.7 given on page 222, by using the additional fact that $\psi(z) = \theta(z)/\phi(z)$; recall Property 3.1.

Property 4.3 The Spectral Density of ARMA

If x_t is ARMA(p, q), $\phi(B)x_t = \theta(B)w_t$, its spectral density is given by

$$f_x(\omega) = \sigma_w^2 \frac{|\theta(\mathrm{e}^{-2\pi i\omega})|^2}{|\phi(\mathrm{e}^{-2\pi i\omega})|^2} \tag{4.15}$$

where $\phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k$ and $\theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k$.

Example 4.5 Moving Average

As an example of a series that does not have an equal mix of frequencies, we consider a moving average model. Specifically, consider the MA(1) model given by

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

A sample realization is shown in the top of Figure 3.2 and we note that the series has less of the higher or faster frequencies. The spectral density will verify this observation.

The autocovariance function is displayed in Example 3.4 on page 90, and for this particular example, we have

$$\gamma(0) = (1 + .5^2)\sigma_w^2 = 1.25\sigma_w^2; \quad \gamma(\pm 1) = .5\sigma_w^2; \quad \gamma(\pm h) = 0 \text{ for } h > 1.$$

Substituting this directly into the definition given in (4.12), we have

$$\begin{aligned} f(\omega) &= \sum_{h=-\infty}^{\infty} \gamma(h)\, \mathrm{e}^{-2\pi i\omega h} = \sigma_w^2 \left[1.25 + .5\left(\mathrm{e}^{-2\pi i\omega} + \mathrm{e}^{2\pi\omega}\right)\right] \\ &= \sigma_w^2 \left[1.25 + \cos(2\pi\omega)\right]. \end{aligned} \tag{4.16}$$

We can also compute the spectral density using Property 4.3, which states that for an MA, $f(\omega) = \sigma_w^2|\theta(\mathrm{e}^{-2\pi i\omega})|^2$. Because $\theta(z) = 1 + .5z$, we have

$$\begin{aligned} |\theta(\mathrm{e}^{-2\pi i\omega})|^2 &= |1 + .5\mathrm{e}^{-2\pi i\omega}|^2 = (1 + .5\mathrm{e}^{-2\pi i\omega})(1 + .5\mathrm{e}^{2\pi i\omega}) \\ &= 1.25 + .5\left(\mathrm{e}^{-2\pi i\omega} + \mathrm{e}^{2\pi\omega}\right) \end{aligned}$$

which leads to agreement with (4.16).

Plotting the spectrum for $\sigma_w^2 = 1$, as in the middle of Figure 4.3, shows the lower or slower frequencies have greater power than the higher or faster frequencies.

Example 4.6 A Second-Order Autoregressive Series

We now consider the spectrum of an AR(2) series of the form

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

for the special case $\phi_1 = 1$ and $\phi_2 = -.9$. Figure 1.9 on page 14 shows a sample realization of such a process for $\sigma_w = 1$. We note the data exhibit a strong periodic component that makes a cycle about every six points.

To use Property 4.3, note that $\theta(z) = 1$, $\phi(z) = 1 - z + .9z^2$ and

$$\begin{aligned} |\phi(e^{-2\pi i\omega})|^2 &= (1 - \mathrm{e}^{-2\pi i\omega} + .9\mathrm{e}^{-4\pi i\omega})(1 - \mathrm{e}^{2\pi i\omega} + .9\mathrm{e}^{4\pi i\omega}) \\ &= 2.81 - 1.9(\mathrm{e}^{2\pi i\omega} + \mathrm{e}^{-2\pi i\omega}) + .9(\mathrm{e}^{4\pi i\omega} + \mathrm{e}^{-4\pi i\omega}) \\ &= 2.81 - 3.8\cos(2\pi\omega) + 1.8\cos(4\pi\omega). \end{aligned}$$

Using this result in (4.15), we have that the spectral density of x_t is

$$f_x(\omega) = \frac{\sigma_w^2}{2.81 - 3.8\cos(2\pi\omega) + 1.8\cos(4\pi\omega)}.$$

Setting $\sigma_w = 1$, the bottom of Figure 4.3 displays $f_x(\omega)$ and shows a strong power component at about $\omega = .16$ cycles per point or a period between six and seven cycles per point and very little power at other frequencies. In this case, modifying the white noise series by applying the second-order AR

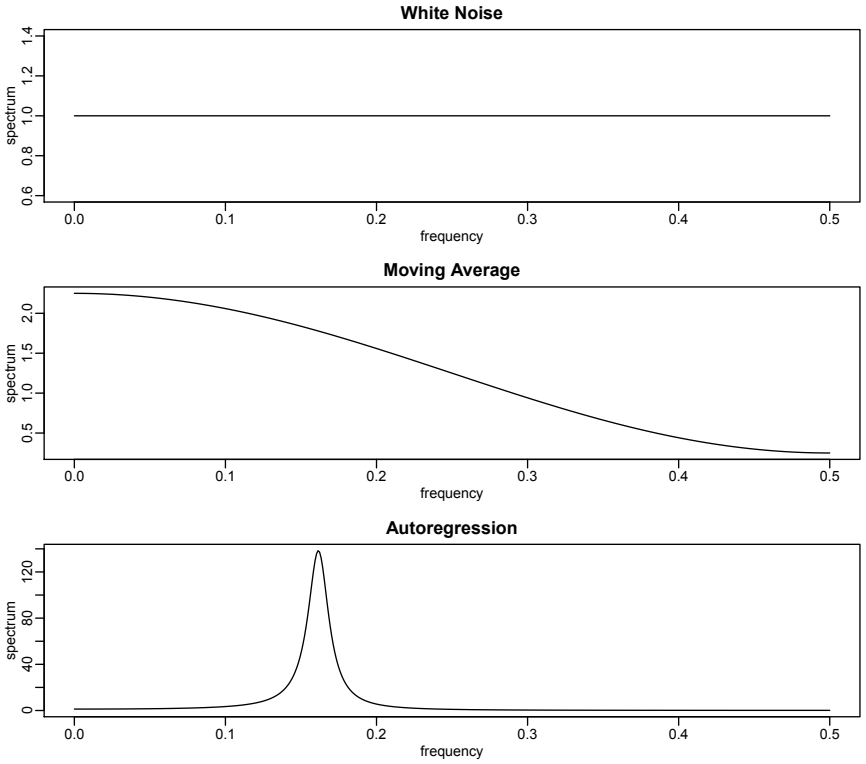

Fig. 4.3. Theoretical spectra of white noise (top), a first-order moving average (middle), and a second-order autoregressive process (bottom).

operator has concentrated the power or variance of the resulting series in a very narrow frequency band.

The spectral density can also be obtained from first principles, without having to use Property 4.3. Because $w_t = x_t - x_{t-1} + .9x_{t-2}$ in this example, we have

$$\begin{aligned}\gamma_w(h) &= \mathrm{cov}(w_{t+h}, w_t)\\ &= \mathrm{cov}(x_{t+h} - x_{t+h-1} + .9x_{t+h-2},\, x_t - x_{t-1} + .9x_{t-2})\\ &= 2.81\gamma_x(h) - 1.9[\gamma_x(h+1) + \gamma_x(h-1)] + .9[\gamma_x(h+2) + \gamma_x(h-2)]\end{aligned}$$

Now, substituting the spectral representation (4.11) for $\gamma_x(h)$ in the above equation yields

$$\begin{aligned}\gamma_w(h) &= \int_{-1/2}^{1/2} \big[2.81 - 1.9(\mathrm{e}^{2\pi i\omega} + \mathrm{e}^{-2\pi i\omega}) + .9(\mathrm{e}^{4\pi i\omega} + \mathrm{e}^{-4\pi i\omega})\big]\mathrm{e}^{2\pi i\omega h} f_x(\omega)d\omega\\ &= \int_{-1/2}^{1/2} \big[2.81 - 3.8\cos(2\pi\omega) + 1.8\cos(4\pi\omega)\big]\mathrm{e}^{2\pi i\omega h} f_x(\omega)d\omega.\end{aligned}$$

If the spectrum of the white noise process, w_t, is $g_w(\omega)$, the uniqueness of the Fourier transform allows us to identify

$$g_w(\omega) = [2.81 - 3.8\cos(2\pi\omega) + 1.8\cos(4\pi\omega)]\, f_x(\omega).$$

But, as we have already seen, $g_w(\omega) = \sigma_w^2$, from which we deduce that

$$f_x(\omega) = \frac{\sigma_w^2}{2.81 - 3.8\cos(2\pi\omega) + 1.8\cos(4\pi\omega)}$$

is the spectrum of the autoregressive series.

To reproduce Figure 4.3, use the `spec.arma` script (see §R.1):

```
par(mfrow=c(3,1))
spec.arma(log="no", main="White Noise")
spec.arma(ma=.5, log="no", main="Moving Average")
spec.arma(ar=c(1,-.9), log="no", main="Autoregression")
```

The above examples motivate the use of the power spectrum for describing the theoretical variance fluctuations of a stationary time series. Indeed, the interpretation of the spectral density function as the variance of the time series over a given frequency band gives us the intuitive explanation for its physical meaning. The plot of the function $f(\omega)$ over the frequency argument ω can even be thought of as an analysis of variance, in which the columns or block effects are the frequencies, indexed by ω.

Example 4.7 Every Explosion has a Cause (cont)

In Example 3.3, we discussed the fact that explosive models have causal counterparts. In that example, we also indicated that it was easier to show this result in general in the spectral domain. In this example, we give the details for an AR(1) model, but the techniques used here will indicate how to generalize the result.

As in Example 3.3, we suppose that $x_t = 2x_{t-1} + w_t$, where $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. Then, the spectral density of x_t is

$$f_x(\omega) = \sigma_w^2\, |1 - 2e^{-2\pi i\omega}|^{-2}. \tag{4.17}$$

But,

$$|1 - 2e^{-2\pi i\omega}| = |1 - 2e^{2\pi i\omega}| = |(2e^{2\pi i\omega})\,(\tfrac{1}{2}e^{-2\pi i\omega} - 1)| = 2\,|1 - \tfrac{1}{2}e^{-2\pi i\omega}|.$$

Thus, (4.17) can be written as

$$f_x(\omega) = \tfrac{1}{4}\sigma_w^2\, |1 - \tfrac{1}{2}e^{-2\pi i\omega}|^{-2},$$

which implies that $x_t = \frac{1}{2}x_{t-1} + v_t$, with $v_t \sim$ iid $\mathrm{N}(0, \frac{1}{4}\sigma_w^2)$ is an equivalent form of the model.

4.4 Periodogram and Discrete Fourier Transform

We are now ready to tie together the periodogram, which is the sample-based concept presented in §4.2, with the spectral density, which is the population-based concept of §4.3.

Definition 4.1 *Given data* $x_1, \ldots, x_n$*, we define the* **discrete Fourier transform (DFT)** *to be*

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} \tag{4.18}$$

for $j = 0, 1, \ldots, n-1$*, where the frequencies* $\omega_j = j/n$ *are called the* **Fourier** *or* **fundamental frequencies**.

If n is a highly composite integer (i.e., it has many factors), the DFT can be computed by the fast Fourier transform (FFT) introduced in Cooley and Tukey (1965). Also, different packages scale the FFT differently, so it is a good idea to consult the documentation. R computes the DFT defined in (4.18) without the factor $n^{-1/2}$, but with an additional factor of $e^{2\pi i \omega_j}$ that can be ignored because we will be interested in the squared modulus of the DFT. Sometimes it is helpful to exploit the inversion result for DFTs which shows the linear transformation is one-to-one. For the inverse DFT we have,

$$x_t = n^{-1/2} \sum_{j=0}^{n-1} d(\omega_j) \mathrm{e}^{2\pi i \omega_j t} \tag{4.19}$$

for $t = 1, \ldots, n$. The following example shows how to calculate the DFT and its inverse in R for the data set $\{1, 2, 3, 4\}$; note that R writes a complex number $z = a + ib$ as `a+bi`.

```
1 (dft = fft(1:4)/sqrt(4))
   [1]  5+0i  -1+1i  -1+0i  -1-1i
2 (idft = fft(dft, inverse=TRUE)/sqrt(4))
   [1] 1+0i  2+0i  3+0i  4+0i
3 (Re(idft))  # keep it real
   [1] 1  2  3  4
```

We now define the periodogram as the squared modulus[5] of the DFT.

Definition 4.2 *Given data* $x_1, \ldots, x_n$*, we define the* **periodogram** *to be*

$$I(\omega_j) = |d(\omega_j)|^2 \tag{4.20}$$

for $j = 0, 1, 2, \ldots, n-1$.

[5] Recall that if $z = a + ib$, then $\bar{z} = a - ib$, and $|z|^2 = z\bar{z} = a^2 + b^2$.

Note that $I(0) = n\bar{x}^2$, where $\bar{x}$ is the sample mean. In addition, because $\sum_{t=1}^{n} \exp(-2\pi i t \frac{j}{n}) = 0$ for $j \neq 0$,[6] we can write the DFT as

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} (x_t - \bar{x}) e^{-2\pi i \omega_j t} \tag{4.21}$$

for $j \neq 0$. Thus, for $j \neq 0$,

$$\begin{aligned} I(\omega_j) = |d(\omega_j)|^2 &= n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} (x_t - \bar{x})(x_s - \bar{x}) e^{-2\pi i \omega_j (t-s)} \\ &= n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}) e^{-2\pi i \omega_j h} \\ &= \sum_{h=-(n-1)}^{n-1} \widehat{\gamma}(h) e^{-2\pi i \omega_j h} \end{aligned} \tag{4.22}$$

where we have put $h = t - s$, with $\widehat{\gamma}(h)$ as given in (1.34).[7]

Recall, $P(\omega_j) = (4/n) I(\omega_j)$ where $P(\omega_j)$ is the scaled periodogram defined in (4.6). Henceforth we will work with $I(\omega_j)$ instead of $P(\omega_j)$. In view of (4.22), the periodogram, $I(\omega_j)$, is the sample version of $f(\omega_j)$ given in (4.12). That is, we may think of the periodogram as the "sample spectral density" of x_t.

It is sometimes useful to work with the real and imaginary parts of the DFT individually. To this end, we define the following transforms.

Definition 4.3 *Given data $x_1, \ldots, x_n$, we define the* **cosine transform**

$$d_c(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t \cos(2\pi \omega_j t) \tag{4.23}$$

and the **sine transform**

$$d_s(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t \sin(2\pi \omega_j t) \tag{4.24}$$

where $\omega_j = j/n$ for $j = 0, 1, \ldots, n-1$.

We note that $d(\omega_j) = d_c(\omega_j) - i\, d_s(\omega_j)$ and hence

$$I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j). \tag{4.25}$$

We have also discussed the fact that spectral analysis can be thought of as an analysis of variance. The next example examines this notion.

[6] $\sum_{t=1}^{n} z^t = z \frac{1-z^n}{1-z}$ for $z \neq 1$.

[7] Note that (4.22) can be used to obtain $\widehat{\gamma}(h)$ by taking the inverse DFT of $I(\omega_j)$. This approach was used in Example 1.27 to obtain a two-dimensional ACF.

Example 4.8 Spectral ANOVA

Let $x_1, \ldots, x_n$ be a sample of size n, where for ease, n is odd. Then, recalling Example 2.9 on page 67 and the discussion around (4.7) and (4.8),

$$x_t = a_0 + \sum_{j=1}^{m} \left[a_j \cos(2\pi\omega_j t) + b_j \sin(2\pi\omega_j t) \right], \tag{4.26}$$

where $m = (n-1)/2$, is exact for $t = 1, \ldots, n$. In particular, using multiple regression formulas, we have $a_0 = \bar{x}$,

$$a_j = \frac{2}{n} \sum_{t=1}^{n} x_t \cos(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_c(\omega_j)$$

$$b_j = \frac{2}{n} \sum_{t=1}^{n} x_t \sin(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_s(\omega_j).$$

Hence, we may write

$$(x_t - \bar{x}) = \frac{2}{\sqrt{n}} \sum_{j=1}^{m} \left[d_c(\omega_j) \cos(2\pi\omega_j t) + d_s(\omega_j) \sin(2\pi\omega_j t) \right]$$

for $t = 1, \ldots, n$. Squaring both sides and summing we obtain

$$\sum_{t=1}^{n} (x_t - \bar{x})^2 = 2 \sum_{j=1}^{m} \left[d_c^2(\omega_j) + d_s^2(\omega_j) \right] = 2 \sum_{j=1}^{m} I(\omega_j)$$

using the results of Problem 2.10(d) on page 81. Thus, we have partitioned the sum of squares into harmonic components represented by frequency ω_j with the periodogram, $I(\omega_j)$, being the mean square regression. This leads to the ANOVA table for n odd:

Source	df	SS	MS
ω_1	2	$2I(\omega_1)$	$I(\omega_1)$
ω_2	2	$2I(\omega_2)$	$I(\omega_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_m	2	$2I(\omega_m)$	$I(\omega_m)$
Total	$n-1$	$\sum_{t=1}^{n}(x_t - \bar{x})^2$	

This decomposition means that if the data contain some strong periodic components, the periodogram values corresponding to those frequencies (or near those frequencies) will be large. On the other hand, the corresponding values of the periodogram will be small for periodic components not present in the data.

The following is an R example to help explain this concept. We consider $n = 5$ observations given by $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 2, x_5 = 1$. Note that

the data complete one cycle, but not in a sinusoidal way. Thus, we should expect the $\omega_1 = 1/5$ component to be relatively large but not exhaustive, and the $\omega_2 = 2/5$ component to be small.

```
x = c(1, 2, 3, 2, 1)
c1 = cos(2*pi*1:5*1/5);  s1 = sin(2*pi*1:5*1/5)
c2 = cos(2*pi*1:5*2/5);  s2 = sin(2*pi*1:5*2/5)
omega1 = cbind(c1, s1);  omega2 = cbind(c2, s2)
anova(lm(x~omega1+omega2))    # ANOVA Table
```

```
              Df   Sum Sq  Mean Sq
omega1         2  2.74164  1.37082
omega2         2   .05836   .02918
Residuals      0   .00000
```

```
abs(fft(x))^2/5       # the periodogram (as a check)
```

```
[1] 16.2  1.37082  .029179  .029179  1.37082
 #  I(0)   I(1/5)   I(2/5)   I(3/5)   I(4/5)
```

Note that $\bar{x} = 1.8$, and $I(0) = 16.2 = 5 \times 1.8^2 (= n\bar{x}^2)$. Also, note that

$$I(1/5) = 1.37082 = \texttt{Mean Sq}(\omega_1) \quad \text{and} \quad I(2/5) = .02918 = \texttt{Mean Sq}(\omega_2)$$

and $I(j/5) = I(1 - j/5)$, for $j = 3, 4$. Finally, we note that the sum of squares associated with the residuals (SSE) is zero, indicating an exact fit.

We are now ready to present some large sample properties of the periodogram. First, let μ be the mean of a stationary process x_t with absolutely summable autocovariance function $\gamma(h)$ and spectral density $f(\omega)$. We can use the same argument as in (4.22), replacing $\bar{x}$ by μ in (4.21), to write

$$I(\omega_j) = n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \mu)(x_t - \mu) e^{-2\pi i \omega_j h} \tag{4.27}$$

where ω_j is a non-zero fundamental frequency. Taking expectation in (4.27) we obtain

$$E\left[I(\omega_j)\right] = \sum_{h=-(n-1)}^{n-1} \left(\frac{n - |h|}{n}\right) \gamma(h) e^{-2\pi i \omega_j h}. \tag{4.28}$$

For any given $\omega \neq 0$, choose a sequence of fundamental frequencies $\omega_{j:n} \to \omega$[8] from which it follows by (4.28) that, as $n \to \infty$[9]

$$E\left[I(\omega_{j:n})\right] \to f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i h \omega}. \tag{4.29}$$

[8] By this we mean $\omega_{j:n} = j_n/n$, where $\{j_n\}$ is a sequence of integers chosen so that j_n/n is the closest Fourier frequency to ω; consequently, $|j_n/n - \omega| \leq \frac{1}{2n}$.

[9] From Definition 4.2 we have $I(0) = n\bar{x}^2$, so the analogous result of (4.29) for the case $\omega = 0$ is $E[I(0)] - n\mu^2 = n \operatorname{var}(\bar{x}) \to f(0)$ as $n \to \infty$.

In other words, under absolute summability of $\gamma(h)$, the spectral density is the long-term average of the periodogram.

To examine the asymptotic distribution of the periodogram, we note that if x_t is a normal time series, the sine and cosine transforms will also be jointly normal, because they are linear combinations of the jointly normal random variables $x_1, x_2, \ldots, x_n$. In that case, the assumption that the covariance function satisfies the condition

$$\theta = \sum_{h=-\infty}^{\infty} |h||\gamma(h)| \ < \infty \tag{4.30}$$

is enough to obtain simple large sample approximations for the variances and covariances. Using the same argument used to develop (4.28) we have

$$\mathrm{cov}[d_c(\omega_j), d_c(\omega_k)] = n^{-1} \sum_{s=1}^{n} \sum_{t=1}^{n} \gamma(s-t) \cos(2\pi\omega_j s) \cos(2\pi\omega_k t), \tag{4.31}$$

$$\mathrm{cov}[d_c(\omega_j), d_s(\omega_k)] = n^{-1} \sum_{s=1}^{n} \sum_{t=1}^{n} \gamma(s-t) \cos(2\pi\omega_j s) \sin(2\pi\omega_k t), \tag{4.32}$$

and

$$\mathrm{cov}[d_s(\omega_j), d_s(\omega_k)] = n^{-1} \sum_{s=1}^{n} \sum_{t=1}^{n} \gamma(s-t) \sin(2\pi\omega_j s) \sin(2\pi\omega_k t), \tag{4.33}$$

where the variance terms are obtained by setting $\omega_j = \omega_k$ in (4.31) and (4.33). In Appendix C, §C.2, we show the terms in (4.31)-(4.33) have interesting properties under assumption (4.30), namely, for $\omega_j, \omega_k \neq 0$ or $1/2$,

$$\mathrm{cov}[d_c(\omega_j), d_c(\omega_k)] = \begin{cases} f(\omega_j)/2 + \varepsilon_n & \omega_j = \omega_k, \\ \varepsilon_n & \omega_j \neq \omega_k, \end{cases} \tag{4.34}$$

$$\mathrm{cov}[d_s(\omega_j), d_s(\omega_k)] = \begin{cases} f(\omega_j)/2 + \varepsilon_n & \omega_j = \omega_k, \\ \varepsilon_n & \omega_j \neq \omega_k, \end{cases} \tag{4.35}$$

and

$$\mathrm{cov}[d_c(\omega_j), d_s(\omega_k)] = \varepsilon_n, \tag{4.36}$$

where the error term ε_n in the approximations can be bounded,

$$|\varepsilon_n| \leq \theta/n, \tag{4.37}$$

and θ is given by (4.30). If $\omega_j = \omega_k = 0$ or $1/2$ in (4.34), the multiplier $1/2$ disappears; note that $d_s(0) = d_s(1/2) = 0$, so (4.35) does not apply.

Example 4.9 Covariance of Sine and Cosine Transforms

For the three-point moving average series of Example 1.9 and $n = 256$ observations, the theoretical covariance matrix of the vector $\boldsymbol{d} = (d_c(\omega_{26}), d_s(\omega_{26}), d_c(\omega_{27}), d_s(\omega_{27}))'$ is

$$\mathrm{cov}(\boldsymbol{d}) = \begin{pmatrix} .3752 & -.0009 & -.0022 & -.0010 \\ -.0009 & .3777 & -.0009 & .0003 \\ -.0022 & -.0009 & .3667 & -.0010 \\ -.0010 & .0003 & -.0010 & .3692 \end{pmatrix}.$$

The diagonal elements can be compared with half the theoretical spectral values of $\frac{1}{2}f(\omega_{26}) = .3774$ for the spectrum at frequency $\omega_{26} = 26/256$, and of $\frac{1}{2}f(\omega_{27}) = .3689$ for the spectrum at $\omega_{27} = 27/256$. Hence, the cosine and sine transforms produce nearly uncorrelated variables with variances approximately equal to one half of the theoretical spectrum. For this particular case, the uniform bound is determined from $\theta = 8/9$, yielding $|\varepsilon_{256}| \leq .0035$ for the bound on the approximation error.

If $x_t \sim \mathrm{iid}(0, \sigma^2)$, then it follows from (4.30)-(4.36), Problem 2.10(d), and a central limit theorem[10] that

$$d_c(\omega_{j:n}) \sim \mathrm{AN}(0, \sigma^2/2) \quad \text{and} \quad d_s(\omega_{j:n}) \sim \mathrm{AN}(0, \sigma^2/2) \tag{4.38}$$

jointly and independently, and independent of $d_c(\omega_{k:n})$ and $d_s(\omega_{k:n})$ provided $\omega_{j:n} \to \omega_1$ and $\omega_{k:n} \to \omega_2$ where $0 < \omega_1 \neq \omega_2 < 1/2$. We note that in this case, $f_x(\omega) = \sigma^2$. In view of (4.38), it follows immediately that as $n \to \infty$,

$$\frac{2I(\omega_{j:n})}{\sigma^2} \xrightarrow{d} \chi^2_2 \quad \text{and} \quad \frac{2I(\omega_{k:n})}{\sigma^2} \xrightarrow{d} \chi^2_2 \tag{4.39}$$

with $I(\omega_{j:n})$ and $I(\omega_{k:n})$ being asymptotically independent, where χ^2_ν denotes a chi-squared random variable with ν degrees of freedom.

Using the central limit theory of §C.2, it is fairly easy to extend the results of the iid case to the case of a linear process.

Property 4.4 Distribution of the Periodogram Ordinates

If

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \tag{4.40}$$

where $w_t \sim iid(0, \sigma_w^2)$, and (4.30) holds, then for any collection of m distinct frequencies $\omega_j \in (0, 1/2)$ with $\omega_{j:n} \to \omega_j$

[10] If $Y_j \sim \mathrm{iid}(0, \sigma^2)$ and $\{a_j\}$ are constants for which $\sum_{j=1}^n a_j^2 / \max_{1 \leq j \leq n} a_j^2 \to \infty$ as $n \to \infty$, then $\sum_{j=1}^n a_j Y_j \sim \mathrm{AN}\Big(0, \sigma^2 \sum_{j=1}^n a_j^2\Big)$. AN is read *asymptotically normal* and is explained in Definition A.5; convergence in distribution ($\xrightarrow{d}$) is explained in Definition A.4.

$$\frac{2I(\omega_{j:n})}{f(\omega_j)} \xrightarrow{d} \text{iid } \chi^2_2 \tag{4.41}$$

provided $f(\omega_j) > 0$, *for* $j = 1, \ldots, m$.

This result is stated more precisely in Theorem C.7 of §C.3. Other approaches to large sample normality of the periodogram ordinates are in terms of cumulants, as in Brillinger (1981), or in terms of mixing conditions, such as in Rosenblatt (1956a). Here, we adopt the approach used by Hannan (1970), Fuller (1996), and Brockwell and Davis (1991).

The distributional result (4.41) can be used to derive an approximate confidence interval for the spectrum in the usual way. Let $\chi^2_\nu(\alpha)$ denote the lower α probability tail for the chi-squared distribution with ν degrees of freedom; that is,

$$\Pr\{\chi^2_\nu \le \chi^2_\nu(\alpha)\} = \alpha. \tag{4.42}$$

Then, an approximate $100(1-\alpha)\%$ confidence interval for the spectral density function would be of the form

$$\frac{2\ I(\omega_{j:n})}{\chi^2_2(1-\alpha/2)} \le f(\omega) \le \frac{2\ I(\omega_{j:n})}{\chi^2_2(\alpha/2)}. \tag{4.43}$$

Often, nonstationary trends are present that should be eliminated before computing the periodogram. Trends introduce extremely low frequency components in the periodogram that tend to obscure the appearance at higher frequencies. For this reason, it is usually conventional to center the data prior to a spectral analysis using either mean-adjusted data of the form $x_t - \bar{x}$ to eliminate the zero or d-c component or to use detrended data of the form $x_t - \widehat{\beta}_1 - \widehat{\beta}_2 t$ to eliminate the term that will be considered a half cycle by the spectral analysis. Note that higher order polynomial regressions in t or nonparametric smoothing (linear filtering) could be used in cases where the trend is nonlinear.

As previously indicated, it is often convenient to calculate the DFTs, and hence the periodogram, using the fast Fourier transform algorithm. The FFT utilizes a number of redundancies in the calculation of the DFT when n is highly composite; that is, an integer with many factors of $2, 3$, or 5, the best case being when $n = 2^p$ is a factor of 2. Details may be found in Cooley and Tukey (1965). To accommodate this property, we can pad the centered (or detrended) data of length n to the next highly composite integer n' by adding zeros, i.e., setting $x^c_{n+1} = x^c_{n+2} = \cdots = x^c_{n'} = 0$, where x^c_t denotes the centered data. This means that the fundamental frequency ordinates will be $\omega_j = j/n'$ instead of j/n. We illustrate by considering the periodogram of the SOI and Recruitment series, as has been given in Figure 1.5 of Chapter 1. Recall that they are monthly series and $n = 453$ months. To find n' in R, use the command `nextn(453)` to see that $n' = 480$ will be used in the spectral analyses by default [use `help(spec.pgram)` to see how to override this default].

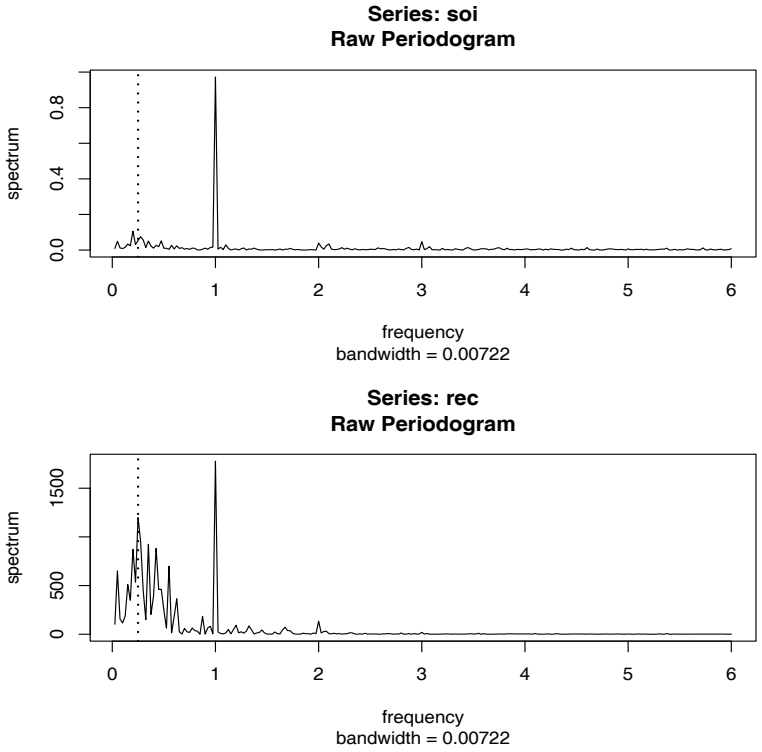

Fig. 4.4. Periodogram of SOI and Recruitment, $n = 453$ ($n' = 480$), where the frequency axis is labeled in multiples of $\Delta = 1/12$. Note the common peaks at $\omega = 1\Delta = 1/12$, or one cycle per year (12 months), and $\omega = \frac{1}{4}\Delta = 1/48$, or one cycle every four years (48 months).

Example 4.10 Periodogram of SOI and Recruitment Series

Figure 4.4 shows the periodograms of each series, where the frequency axis is labeled in multiples of $\Delta = 1/12$. As previously indicated, the centered data have been padded to a series of length 480. We notice a narrow-band peak at the obvious yearly (12 month) cycle, $\omega = 1\Delta = 1/12$. In addition, there is considerable power in a wide band at the lower frequencies that is centered around the four-year (48 month) cycle $\omega = \frac{1}{4}\Delta = 1/48$ representing a possible El Niño effect. This wide band activity suggests that the possible El Niño cycle is irregular, but tends to be around four years on average. We will continue to address this problem as we move to more sophisticated analyses.

Noting $\chi^2_2(.025) = .05$ and $\chi^2_2(.975) = 7.38$, we can obtain approximate 95% confidence intervals for the frequencies of interest. For example, the periodogram of the SOI series is $I_S(1/12) = .97$ at the yearly cycle. An approximate 95% confidence interval for the spectrum $f_S(1/12)$ is then

$$[2(.97)/7.38,\ 2(.97)/.05] = [.26,\ 38.4],$$

which is too wide to be of much use. We do notice, however, that the lower value of .26 is higher than any other periodogram ordinate, so it is safe to say that this value is significant. On the other hand, an approximate 95% confidence interval for the spectrum at the four-year cycle, $f_S(1/48)$, is

$$[2(.05)/7.38,\ 2(.05)/.05] = [.01,\ 2.12],$$

which again is extremely wide, and with which we are unable to establish significance of the peak.

We now give the R commands that can be used to reproduce Figure 4.4. To calculate and graph the periodogram, we used the `spec.pgram` command in R. We note that the value of Δ is the reciprocal of the value of `frequency` used in `ts()` when making the data a time series object. If the data are not time series objects, `frequency` is set to 1. Also, we set `log="no"` because R will plot the periodogram on a $\log_{10}$ scale by default. Figure 4.4 displays a `bandwidth` and by default, R tapers the data (which we override in the commands below). We will discuss bandwidth and tapering in the next section, so ignore these concepts for the time being.

```
par(mfrow=c(2,1))
soi.per = spec.pgram(soi, taper=0, log="no")
abline(v=1/4, lty="dotted")
rec.per = spec.pgram(rec, taper=0, log="no")
abline(v=1/4, lty="dotted")
```

The confidence intervals for the SOI series at the yearly cycle, $\omega = 1/12 = 40/480$, and the possible El Niño cycle of four years $\omega = 1/48 = 10/480$ can be computed in R as follows:

```
soi.per$spec[40]  # 0.97223; soi pgram at freq 1/12 = 40/480
soi.per$spec[10]  # 0.05372; soi pgram at freq 1/48 = 10/480
#  conf intervals  -  returned value:
U = qchisq(.025,2)    #  0.05063
L = qchisq(.975,2)    #  7.37775
2*soi.per$spec[10]/L  #  0.01456
2*soi.per$spec[10]/U  #  2.12220
2*soi.per$spec[40]/L  #  0.26355
2*soi.per$spec[40]/U  # 38.40108
```

The example above makes it clear that the periodogram as an estimator is susceptible to large uncertainties, and we need to find a way to reduce the variance. Not surprisingly, this result follows if we think about the periodogram, $I(\omega_j)$ as an estimator of the spectral density $f(\omega)$ and realize that it is the sum of squares of only two random variables for any sample size. The solution to this dilemma is suggested by the analogy with classical statistics where we look for independent random variables with the same variance and average the squares of these common variance observations. Independence and equality of variance do not hold in the time series case, but the covariance

structure of the two adjacent estimators given in Example 4.9 suggests that for neighboring frequencies, these assumptions are approximately true.

4.5 Nonparametric Spectral Estimation

To continue the discussion that ended the previous section, we introduce a frequency band, $\mathcal{B}$, of $L << n$ contiguous fundamental frequencies, centered around frequency $\omega_j = j/n$, which is chosen close to a frequency of interest, ω. For frequencies of the form $\omega^* = \omega_j + k/n$, let

$$\mathcal{B} = \left\{\omega^*:\ \omega_j - \frac{m}{n} \le \omega^* \le \omega_j + \frac{m}{n}\right\}, \tag{4.44}$$

where

$$L = 2m + 1 \tag{4.45}$$

is an odd number, chosen such that the spectral values in the interval $\mathcal{B}$,

$$f(\omega_j + k/n), \quad k = -m, \ldots, 0, \ldots, m$$

are approximately equal to $f(\omega)$. This structure can be realized for large sample sizes, as shown formally in §C.2.

We now define an averaged (or smoothed) periodogram as the average of the periodogram values, say,

$$\bar{f}(\omega) = \frac{1}{L}\sum_{k=-m}^{m} I(\omega_j + k/n), \tag{4.46}$$

over the band $\mathcal{B}$. Under the assumption that the spectral density is fairly constant in the band $\mathcal{B}$, and in view of (4.41) we can show that under appropriate conditions,[11] for large n, the periodograms in (4.46) are approximately distributed as independent $f(\omega)\chi^2_2/2$ random variables, for $0 < \omega < 1/2$, as long as we keep L fairly small relative to n. This result is discussed formally in §C.2. Thus, under these conditions, $L\bar{f}(\omega)$ is the sum of L approximately independent $f(\omega)\chi^2_2/2$ random variables. It follows that, for large n,

$$\frac{2L\bar{f}(\omega)}{f(\omega)} \dot{\sim} \chi^2_{2L} \tag{4.47}$$

where $\dot{\sim}$ means *is approximately distributed as.*

In this scenario, where we smooth the periodogram by simple averaging, it seems reasonable to call the width of the frequency interval defined by (4.44),

[11] The conditions, which are sufficient, are that x_t is a linear process, as described in Property 4.4, with $\sum_j \sqrt{|j|}\,|\psi_j| < \infty$, and w_t has a finite fourth moment.

$$B_w = \frac{L}{n}, \tag{4.48}$$

the bandwidth.[12] The concept of the bandwidth, however, becomes more complicated with the introduction of spectral estimators that smooth with unequal weights. Note (4.48) implies the degrees of freedom can be expressed as

$$2L = 2B_w n, \tag{4.49}$$

or twice the time-bandwidth product. The result (4.47) can be rearranged to obtain an approximate $100(1-\alpha)\%$ confidence interval of the form

$$\frac{2L\bar{f}(\omega)}{\chi^2_{2L}(1-\alpha/2)} \le f(\omega) \le \frac{2L\bar{f}(\omega)}{\chi^2_{2L}(\alpha/2)} \tag{4.50}$$

for the true spectrum, $f(\omega)$.

Many times, the visual impact of a spectral density plot will be improved by plotting the logarithm of the spectrum instead of the spectrum (the log transformation is the variance stabilizing transformation in this situation). This phenomenon can occur when regions of the spectrum exist with peaks of interest much smaller than some of the main power components. For the log spectrum, we obtain an interval of the form

$$\big[\log \bar{f}(\omega) + \log 2L - \log \chi^2_{2L}(1-\alpha/2), \\ \log \bar{f}(\omega) + \log 2L - \log \chi^2_{2L}(\alpha/2)\big]. \tag{4.51}$$

We can also test hypotheses relating to the equality of spectra using the fact that the distributional result (4.47) implies that the ratio of spectra based on roughly independent samples will have an approximate $F_{2L,2L}$ distribution. The independent estimators can either be from different frequency bands or from different series.

If zeros are appended before computing the spectral estimators, we need to adjust the degrees of freedom and an approximation is to replace $2L$ by $2Ln/n'$. Hence, we define the adjusted degrees of freedom as

$$df = \frac{2Ln}{n'} \tag{4.52}$$

[12] The bandwidth value used in R is based on Grenander (1951). The basic idea is that bandwidth can be related to the standard deviation of the weighting distribution. For the uniform distribution on the frequency range $-m/n$ to m/n, the standard deviation is $L/n\sqrt{12}$ (using a continuity correction). Consequently, in the case of (4.46), R will report a bandwidth of $L/n\sqrt{12}$, which amounts to dividing our definition by $\sqrt{12}$. Note that in the extreme case $L = n$, we would have $B_w = 1$ indicating that everything was used in the estimation; in this case, R would report a bandwidth of $1/\sqrt{12}$. There are many definitions of bandwidth and an excellent discussion may be found in Percival and Walden (1993, §6.7).

and use it instead of $2L$ in the confidence intervals (4.50) and (4.51). For example, (4.50) becomes

$$\frac{df\,\bar{f}(\omega)}{\chi^2_{df}(1-\alpha/2)} \le f(\omega) \le \frac{df\,\bar{f}(\omega)}{\chi^2_{df}(\alpha/2)}. \tag{4.53}$$

A number of assumptions are made in computing the approximate confidence intervals given above, which may not hold in practice. In such cases, it may be reasonable to employ resampling techniques such as one of the parametric bootstraps proposed by Hurvich and Zeger (1987) or a nonparametric local bootstrap proposed by Paparoditis and Politis (1999). To develop the bootstrap distributions, we assume that the contiguous DFTs in a frequency band of the form (4.44) all came from a time series with identical spectrum $f(\omega)$. This, in fact, is exactly the same assumption made in deriving the large-sample theory. We may then simply resample the L DFTs in the band, with replacement, calculating a spectral estimate from each bootstrap sample. The sampling distribution of the bootstrap estimators approximates the distribution of the nonparametric spectral estimator. For further details, including the theoretical properties of such estimators, see Paparoditis and Politis (1999).

Before proceeding further, we pause to consider computing the average periodograms for the SOI and Recruitment series, as shown in Figure 4.5.

Example 4.11 Averaged Periodogram for SOI and Recruitment

Generally, it is a good idea to try several bandwidths that seem to be compatible with the general overall shape of the spectrum, as suggested by the periodogram. The SOI and Recruitment series periodograms, previously computed in Figure 4.4, suggest the power in the lower El Niño frequency needs smoothing to identify the predominant overall period. Trying values of L leads to the choice $L = 9$ as a reasonable value, and the result is displayed in Figure 4.5. In our notation, the bandwidth in this case is $B_w = 9/480 = .01875$ cycles per month for the spectral estimator. This bandwidth means we are assuming a relatively constant spectrum over about $.01875/.5 = 3.75\%$ of the entire frequency interval $(0, 1/2)$. To obtain the bandwidth, $B_w = .01875$, from the one reported by R in Figure 4.5, we can multiply $.065\Delta$ (the frequency scale is in increments of Δ) by $\sqrt{12}$ as discussed in footnote 12 on page 197.

The smoothed spectra shown in Figure 4.5 provide a sensible compromise between the noisy version, shown in Figure 4.4, and a more heavily smoothed spectrum, which might lose some of the peaks. An undesirable effect of averaging can be noticed at the yearly cycle, $\omega = 1\Delta$, where the narrow band peaks that appeared in the periodograms in Figure 4.4 have been flattened and spread out to nearby frequencies. We also notice, and have marked, the appearance of harmonics of the yearly cycle, that is, frequencies of the form $\omega = k\Delta$ for $k = 1, 2, \ldots$. Harmonics typically occur when a periodic component is present, but not in a sinusoidal fashion; see Example 4.12.

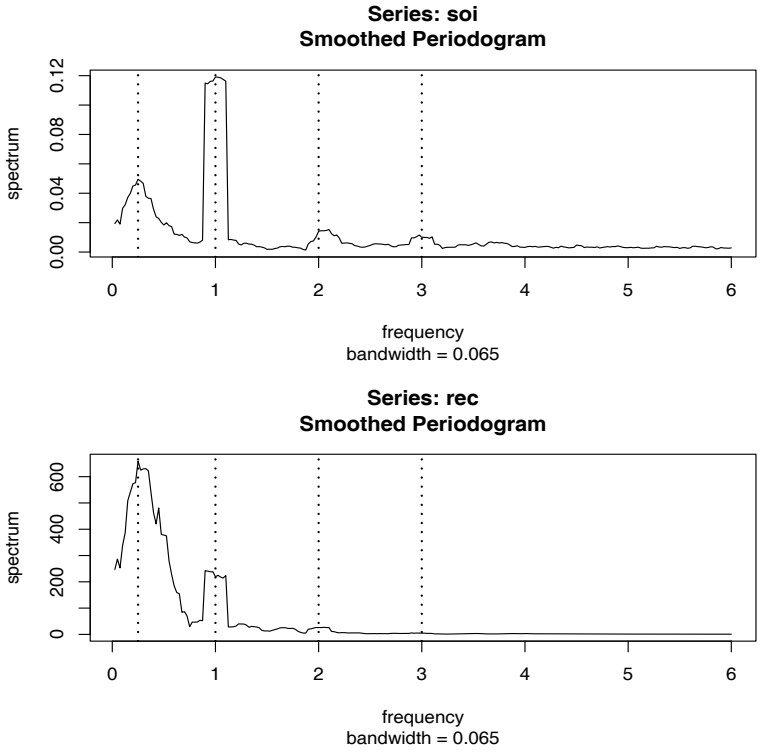

Fig. 4.5. The averaged periodogram of the SOI and Recruitment series $n = 453$, $n' = 480$, $L = 9$, $df = 17$, showing common peaks at the four year period, $\omega = \frac{1}{4}\Delta = 1/48$ cycles/month, the yearly period, $\omega = 1\Delta = 1/12$ cycles/month and some of its harmonics $\omega = k\Delta$ for $k = 2, 3$.

Figure 4.5 can be reproduced in R using the following commands. The basic call is to the function `spec.pgram`. To compute averaged periodograms, use the Daniell kernel, and specify m, where $L = 2m + 1$ ($L = 9$ and $m = 4$ in this example). We will explain the kernel concept later in this section, specifically just prior to Example 4.13.

```
par(mfrow=c(2,1))
k = kernel("daniell", 4)
soi.ave = spec.pgram(soi, k, taper=0, log="no")
abline(v=c(.25,1,2,3), lty=2)
# Repeat above lines using rec in place of soi on line 3
soi.ave$bandwidth       # 0.0649519 = reported bandwidth
soi.ave$bandwidth*(1/12)*sqrt(12)       # 0.01875 = Bw
```

The adjusted degrees of freedom are $df = 2(9)(453)/480 \approx 17$. We can use this value for the 95% confidence intervals, with $\chi^2_{df}(.025) = 7.56$ and $\chi^2_{df}(.975) = 30.17$. Substituting into (4.53) gives the intervals in Table 4.1 for the two frequency bands identified as having the maximum power. To

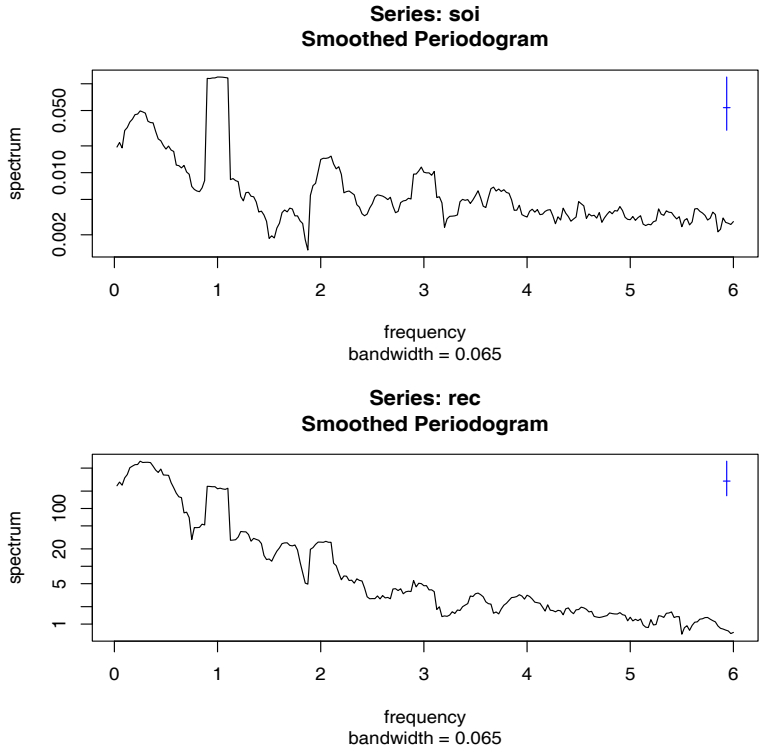

Fig. 4.6. Figure 4.5 with the average periodogram ordinates plotted on a $\log_{10}$ scale. The display in the upper right-hand corner represents a generic 95% confidence interval.

examine the two peak power possibilities, we may look at the 95% confidence intervals and see whether the lower limits are substantially larger than adjacent baseline spectral levels. For example, the El Niño frequency of 48 months has lower limits that exceed the values the spectrum would have if there were simply a smooth underlying spectral function without the peaks. The relative distribution of power over frequencies is different, with the SOI having less power at the lower frequency, relative to the seasonal periods, and the recruit series having relatively more power at the lower or El Niño frequency.

The entries in Table 4.1 for SOI can be obtained in R as follows:

```
df = soi.ave$df          #  df = 16.9875  (returned values)
U = qchisq(.025, df)     #  U = 7.555916
L = qchisq(.975, df)     #  L = 30.17425
soi.ave$spec[10]         #  0.0495202
soi.ave$spec[40]         #  0.1190800
# intervals
df*soi.ave$spec[10]/L    # 0.0278789
df*soi.ave$spec[10]/U    # 0.1113333
```

Table 4.1. Confidence Intervals for the Spectra of the SOI and Recruitment Series

Series	ω	Period	Power	Lower	Upper
SOI	1/48	4 years	.05	.03	.11
	1/12	1 year	.12	.07	.27
Recruits	1/48	4 years	6.59	3.71	14.82
$\times 10^2$	1/12	1 year	2.19	1.24	4.93

```
df*soi.ave$spec[40]/L    # 0.0670396
df*soi.ave$spec[40]/U    # 0.2677201
# repeat above commands with soi replaced by rec
```

Finally, Figure 4.6 shows the averaged periodograms in Figure 4.5 plotted on a $\log_{10}$ scale. This is the default plot in R, and these graphs can be obtained by removing the statement `log="no"` in the `spec.pgram` call. Notice that the default plot also shows a generic confidence interval of the form (4.51) (with log replaced by $\log_{10}$) in the upper right-hand corner. To use it, imagine placing the tick mark on the averaged periodogram ordinate of interest; the resulting bar then constitutes an approximate 95% confidence interval for the spectrum at that frequency. We note that displaying the estimates on a log scale tends to emphasize the harmonic components.

Example 4.12 Harmonics

In the previous example, we saw that the spectra of the annual signals displayed minor peaks at the harmonics; that is, the signal spectra had a large peak at $\omega = 1\Delta = 1/12$ cycles/month (the one-year cycle) and minor peaks at its harmonics $\omega = k\Delta$ for $k = 2, 3, \ldots$ (two-, three-, and so on, cycles per year). This will often be the case because most signals are not perfect sinusoids (or perfectly cyclic). In this case, the harmonics are needed to capture the non-sinusoidal behavior of the signal. As an example, consider the signal formed in Figure 4.7 from a (fundamental) sinusoid oscillating at two cycles per unit time along with the second through sixth harmonics at decreasing amplitudes. In particular, the signal was formed as

$$\begin{aligned} x_t = &\sin(2\pi 2t) + .5\sin(2\pi 4t) + .4\sin(2\pi 6t) \\ &+ .3\sin(2\pi 8t) + .2\sin(2\pi 10t) + .1\sin(2\pi 12t) \end{aligned} \tag{4.54}$$

for $0 \le t \le 1$. Notice that the signal is non-sinusoidal in appearance and rises quickly then falls slowly.

A figure similar to Figure 4.7 can be generated in R as follows.

```
t = seq(0, 1, by=1/200)
amps = c(1, .5, .4, .3, .2, .1)
x = matrix(0, 201, 6)
for (j in 1:6) x[,j] = amps[j]*sin(2*pi*t*2*j)
x = ts(cbind(x, rowSums(x)), start=0, deltat=1/200)
```

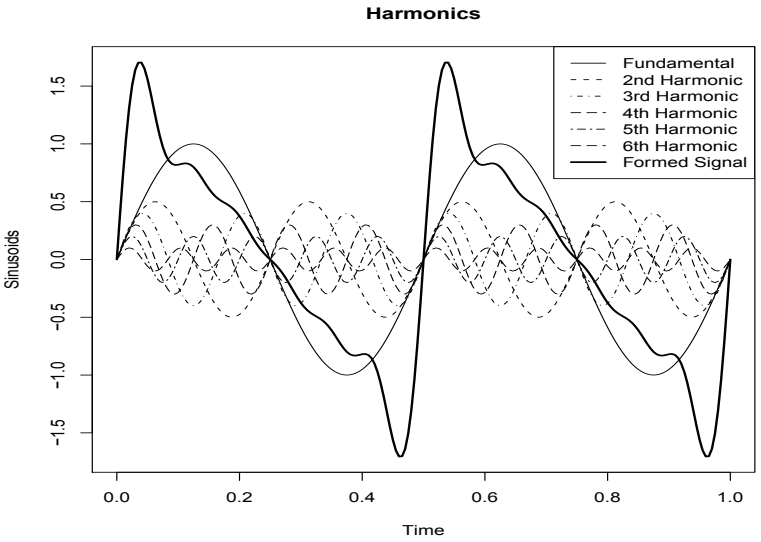

Fig. 4.7. A signal (thick solid line) formed by a fundamental sinusoid (thin solid line) oscillating at two cycles per unit time and its harmonics as specified in (4.54).

```
ts.plot(x, lty=c(1:6, 1), lwd=c(rep(1,6), 2), ylab="Sinusoids")
names = c("Fundamental","2nd Harmonic","3rd Harmonic","4th
    Harmonic", "5th Harmonic", "6th Harmonic", "Formed Signal")
legend("topright", names, lty=c(1:6, 1), lwd=c(rep(1,6), 2))
```

Example 4.11 points out the necessity for having some relatively systematic procedure for deciding whether peaks are significant. The question of deciding whether a single peak is significant usually rests on establishing what we might think of as a baseline level for the spectrum, defined rather loosely as the shape that one would expect to see if no spectral peaks were present. This profile can usually be guessed by looking at the overall shape of the spectrum that includes the peaks; usually, a kind of baseline level will be apparent, with the peaks seeming to emerge from this baseline level. If the lower confidence limit for the spectral value is still greater than the baseline level at some predetermined level of significance, we may claim that frequency value as a statistically significant peak. To be consistent with our stated indifference to the upper limits, we might use a one-sided confidence interval.

An important aspect of interpreting the significance of confidence intervals and tests involving spectra is that typically, more than one frequency will be of interest, so that we will potentially be interested in simultaneous statements about a whole collection of frequencies. For example, it would be unfair to claim in Table 4.1 the two frequencies of interest as being statistically significant and all other potential candidates as nonsignificant at the overall level of $\alpha = .05$. In this case, we follow the usual statistical approach, noting that if K statements $S_1, S_2, \ldots, S_k$ are made at significance level α, i.e.,

$P\{S_k\} = 1 - \alpha$, then the overall probability all statements are true satisfies the Bonferroni inequality

$$P\{\text{all } S_k \text{ true}\} \geq 1 - K\alpha. \tag{4.55}$$

For this reason, it is desirable to set the significance level for testing each frequency at α/K if there are K potential frequencies of interest. If, a priori, potentially $K = 10$ frequencies are of interest, setting $\alpha = .01$ would give an overall significance level of bound of .10.

The use of the confidence intervals and the necessity for smoothing requires that we make a decision about the bandwidth B_w over which the spectrum will be essentially constant. Taking too broad a band will tend to smooth out valid peaks in the data when the constant variance assumption is not met over the band. Taking too narrow a band will lead to confidence intervals so wide that peaks are no longer statistically significant. Thus, we note that there is a conflict here between variance properties or bandwidth stability, which can be improved by increasing B_w and resolution, which can be improved by decreasing B_w. A common approach is to try a number of different bandwidths and to look qualitatively at the spectral estimators for each case.

To address the problem of resolution, it should be evident that the flattening of the peaks in Figures 4.5 and 4.6 was due to the fact that simple averaging was used in computing $\bar{f}(\omega)$ defined in (4.46). There is no particular reason to use simple averaging, and we might improve the estimator by employing a weighted average, say

$$\widehat{f}(\omega) = \sum_{k=-m}^{m} h_k \, I(\omega_j + k/n), \tag{4.56}$$

using the same definitions as in (4.46) but where the weights $h_k > 0$ satisfy

$$\sum_{k=-m}^{m} h_k = 1.$$

In particular, it seems reasonable that the resolution of the estimator will improve if we use weights that decrease as distance from the center weight h_0 increases; we will return to this idea shortly. To obtain the averaged periodogram, $\bar{f}(\omega)$, in (4.56), set $h_k = L^{-1}$, for all k, where $L = 2m + 1$. The asymptotic theory established for $\bar{f}(\omega)$ still holds for $\widehat{f}(\omega)$ provided that the weights satisfy the additional condition that if $m \to \infty$ as $n \to \infty$ but $m/n \to 0$, then

$$\sum_{k=-m}^{m} h_k^2 \to 0.$$

Under these conditions, as $n \to \infty$,

(i) $E\left(\widehat{f}(\omega)\right) \to f(\omega)$

(ii) $\left(\sum_{k=-m}^{m} h_k^2\right)^{-1} \mathrm{cov}\left(\widehat{f}(\omega), \widehat{f}(\lambda)\right) \to f^2(\omega) \quad \text{for } \omega = \lambda \neq 0, 1/2.$

In (ii), replace $f^2(\omega)$ by 0 if $\omega \neq \lambda$ and by $2f^2(\omega)$ if $\omega = \lambda = 0$ or $1/2$.

We have already seen these results in the case of $\bar{f}(\omega)$, where the weights are constant, $h_k = L^{-1}$, in which case $\sum_{k=-m}^{m} h_k^2 = L^{-1}$. The distributional properties of (4.56) are more difficult now because $\widehat{f}(\omega)$ is a weighted linear combination of asymptotically independent χ^2 random variables. An approximation that seems to work well is to replace L by $\left(\sum_{k=-m}^{m} h_k^2\right)^{-1}$. That is, define

$$L_h = \left(\sum_{k=-m}^{m} h_k^2\right)^{-1} \tag{4.57}$$

and use the approximation[13]

$$\frac{2L_h \widehat{f}(\omega)}{f(\omega)} \mathrel{\dot\sim} \chi^2_{2L_h}. \tag{4.58}$$

In analogy to (4.48), we will define the bandwidth in this case to be

$$B_w = \frac{L_h}{n}. \tag{4.59}$$

Using the approximation (4.58) we obtain an approximate $100(1-\alpha)\%$ confidence interval of the form

$$\frac{2L_h \widehat{f}(\omega)}{\chi^2_{2L_h}(1-\alpha/2)} \leq f(\omega) \leq \frac{2L_h \widehat{f}(\omega)}{\chi^2_{2L_h}(\alpha/2)} \tag{4.60}$$

for the true spectrum, $f(\omega)$. If the data are padded to n', then replace $2L_h$ in (4.60) with $df = 2L_h n/n'$ as in (4.52).

An easy way to generate the weights in R is by repeated use of the Daniell kernel. For example, with $m = 1$ and $L = 2m + 1 = 3$, the Daniell kernel has weights $\{h_k\} = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$; applying this kernel to a sequence of numbers, $\{u_t\}$, produces

$$\widehat{u}_t = \tfrac{1}{3}u_{t-1} + \tfrac{1}{3}u_t + \tfrac{1}{3}u_{t+1}.$$

We can apply the same kernel again to the $\widehat{u}_t$,

$$\widehat{\widehat{u}}_t = \tfrac{1}{3}\widehat{u}_{t-1} + \tfrac{1}{3}\widehat{u}_t + \tfrac{1}{3}\widehat{u}_{t+1},$$

which simplifies to

$$\widehat{\widehat{u}}_t = \tfrac{1}{9}u_{t-2} + \tfrac{2}{9}u_{t-1} + \tfrac{3}{9}u_t + \tfrac{2}{9}u_{t+1} + \tfrac{1}{9}u_{t+2}.$$

[13] The approximation proceeds as follows: If $\widehat{f} \mathrel{\dot\sim} c\chi^2_\nu$, where c is a constant, then $E\widehat{f} \approx c\nu$ and $\mathrm{var}\widehat{f} \approx f^2 \sum_k h_k^2 \approx c^2 2\nu$. Solving, $c \approx f \sum_k h_k^2/2 = f/2L_h$ and $\nu \approx 2\left(\sum_k h_k^2\right)^{-1} = 2L_h$.

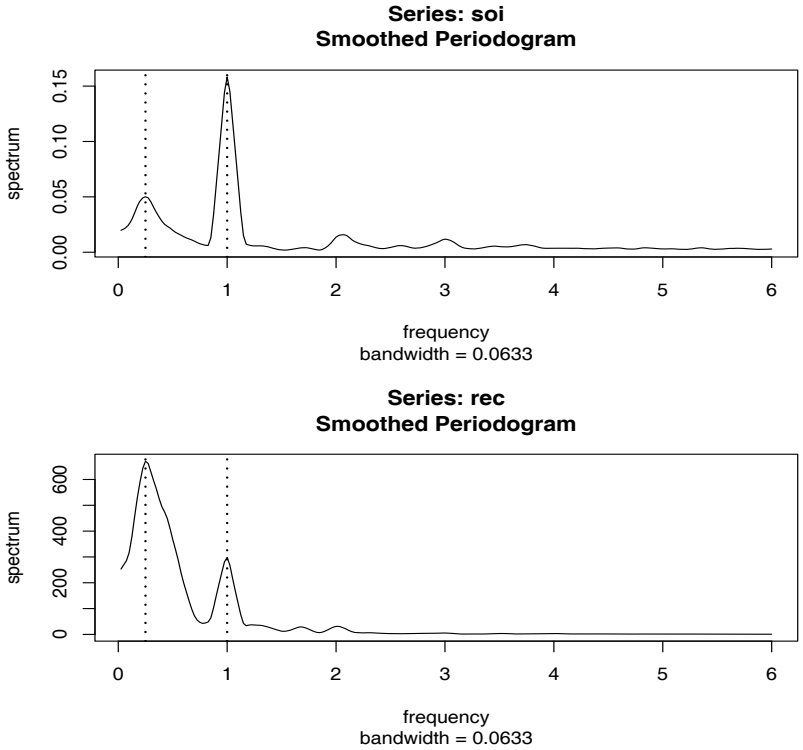

Fig. 4.8. Smoothed spectral estimates of the SOI and Recruitment series; see Example 4.13 for details.

The modified Daniell kernel puts half weights at the end points, so with $m = 1$ the weights are $\{h_k\} = \{\frac{1}{4}, \frac{2}{4}, \frac{1}{4}\}$ and

$$\widehat{u}_t = \tfrac{1}{4}u_{t-1} + \tfrac{1}{2}u_t + \tfrac{1}{4}u_{t+1}.$$

Applying the same kernel again to $\widehat{u}_t$ yields

$$\widehat{\widehat{u}}_t = \tfrac{1}{16}u_{t-2} + \tfrac{4}{16}u_{t-1} + \tfrac{6}{16}u_t + \tfrac{4}{16}u_{t+1} + \tfrac{1}{16}u_{t+2}.$$

These coefficients can be obtained in R by issuing the `kernel` command. For example, `kernel("modified.daniell", c(1,1))` would produce the coefficients of the last example. It is also possible to use different values of m, e.g., try `kernel("modified.daniell", c(1,2))` or `kernel("daniell", c(5,3))`. The other kernels that are currently available in R are the Dirichlet kernel and the Fejér kernel, which we will discuss shortly.

Example 4.13 Smoothed Periodogram for SOI and Recruitment

In this example, we estimate the spectra of the SOI and Recruitment series using the smoothed periodogram estimate in (4.56). We used a modified Daniell kernel twice, with $m = 3$ both times. This yields $L_h =$

$1/\sum_{k=-m}^{m} h_k^2 = 9.232$, which is close to the value of $L = 9$ used in Example 4.11. In this case, the bandwidth is $B_w = 9.232/480 = .019$ and the modified degrees of freedom is $df = 2L_h 453/480 = 17.43$. The weights, h_k, can be obtained and graphed in R as follows:

```
kernel("modified.daniell", c(3,3))
```

```
    coef[-6] = 0.006944 = coef[ 6]
    coef[-5] = 0.027778 = coef[ 5]
    coef[-4] = 0.055556 = coef[ 4]
    coef[-3] = 0.083333 = coef[ 3]
    coef[-2] = 0.111111 = coef[ 2]
    coef[-1] = 0.138889 = coef[ 1]
    coef[ 0] = 0.152778
```

```
plot(kernel("modified.daniell", c(3,3)))  # not shown
```

The resulting spectral estimates can be viewed in Figure 4.8 and we notice that the estimates more appealing than those in Figure 4.5. Figure 4.8 was generated in R as follows; we also show how to obtain df and B_w.

```
par(mfrow=c(2,1))
k = kernel("modified.daniell", c(3,3))
soi.smo = spec.pgram(soi, k, taper=0, log="no")
abline(v=1, lty="dotted");  abline(v=1/4, lty="dotted")
# Repeat above lines with rec replacing soi in line 3
df = soi.smo2$df            # df = 17.42618
Lh = 1/sum(k[-k$m:k$m]^2)   # Lh = 9.232413
Bw = Lh/480                 # Bw = 0.01923419
```

The bandwidth reported by R is .063, which is approximately $B_w/\sqrt{12}\Delta$, where $\Delta = 1/12$ in this example. Reissuing the `spec.pgram` commands with `log="no"` removed will result in a figure similar to Figure 4.6. Finally, we mention that R uses the modified Daniell kernel by default. For example, an easier way to obtain `soi.smo` is to issue the command:

```
soi.smo = spectrum(soi, spans=c(7,7), taper=0)
```

Notice that `spans` is a vector of odd integers, given in terms of $L = 2m + 1$ instead of m. These values give the widths of the modified Daniell smoother to be used to smooth the periodogram.

We are now ready to briefly introduce the concept of *tapering*; a more detailed discussion may be found in Bloomfield (2000, §9.5). Suppose x_t is a mean-zero, stationary process with spectral density $f_x(\omega)$. If we replace the original series by the tapered series

$$y_t = h_t x_t, \tag{4.61}$$

for $t = 1, 2, \ldots, n$, use the modified DFT

$$d_y(\omega_j) = n^{-1/2} \sum_{t=1}^{n} h_t x_t e^{-2\pi i \omega_j t}, \tag{4.62}$$

and let $I_y(\omega_j) = |d_y(\omega_j)|^2$, we obtain (see Problem 4.15)

$$E[I_y(\omega_j)] = \int_{-1/2}^{1/2} W_n(\omega_j - \omega)\ f_x(\omega)\ d\omega \tag{4.63}$$

where

$$W_n(\omega) = |H_n(\omega)|^2 \tag{4.64}$$

and

$$H_n(\omega) = n^{-1/2} \sum_{t=1}^{n} h_t e^{-2\pi i \omega t}. \tag{4.65}$$

The value $W_n(\omega)$ is called a spectral window because, in view of (4.63), it is determining which part of the spectral density $f_x(\omega)$ is being "seen" by the estimator $I_y(\omega_j)$ on average. In the case that $h_t = 1$ for all t, $I_y(\omega_j) = I_x(\omega_j)$ is simply the periodogram of the data and the window is

$$W_n(\omega) = \frac{\sin^2(n\pi\omega)}{n \sin^2(\pi\omega)} \tag{4.66}$$

with $W_n(0) = n$, which is known as the Fejér or modified Bartlett kernel. If we consider the averaged periodogram in (4.46), namely

$$\bar{f}_x(\omega) = \frac{1}{L} \sum_{k=-m}^{m} I_x(\omega_j + k/n),$$

the window, $W_n(\omega)$, in (4.63) will take the form

$$W_n(\omega) = \frac{1}{nL} \sum_{k=-m}^{m} \frac{\sin^2[n\pi(\omega + k/n)]}{\sin^2[\pi(\omega + k/n)]}\,. \tag{4.67}$$

Tapers generally have a shape that enhances the center of the data relative to the extremities, such as a cosine bell of the form

$$h_t = .5\left[1 + \cos\left(\frac{2\pi(t - \bar{t})}{n}\right)\right], \tag{4.68}$$

where $\bar{t} = (n+1)/2$, favored by Blackman and Tukey (1959). In Figure 4.9, we have plotted the shapes of two windows, $W_n(\omega)$, for $n = 480$ and $L = 9$, when (i) $h_t \equiv 1$, in which case, (4.67) applies, and (ii) h_t is the cosine taper in (4.68). In both cases the predicted bandwidth should be $B_w = 9/480 = .01875$ cycles per point, which corresponds to the "width" of the windows shown in Figure 4.9. Both windows produce an integrated average spectrum over this band but the untapered window in the top panels shows considerable ripples over the band and outside the band. The ripples outside the band are called sidelobes and tend to introduce frequencies from outside the interval that may contaminate the desired spectral estimate within the band. For example, a large dynamic range for the values in the spectrum introduces spectra in contiguous frequency intervals several orders of magnitude greater than the value in the interval of interest. This effect is sometimes called leakage. Figure 4.9 emphasizes the suppression of the sidelobes in the Fejér kernel when a cosine taper is used.

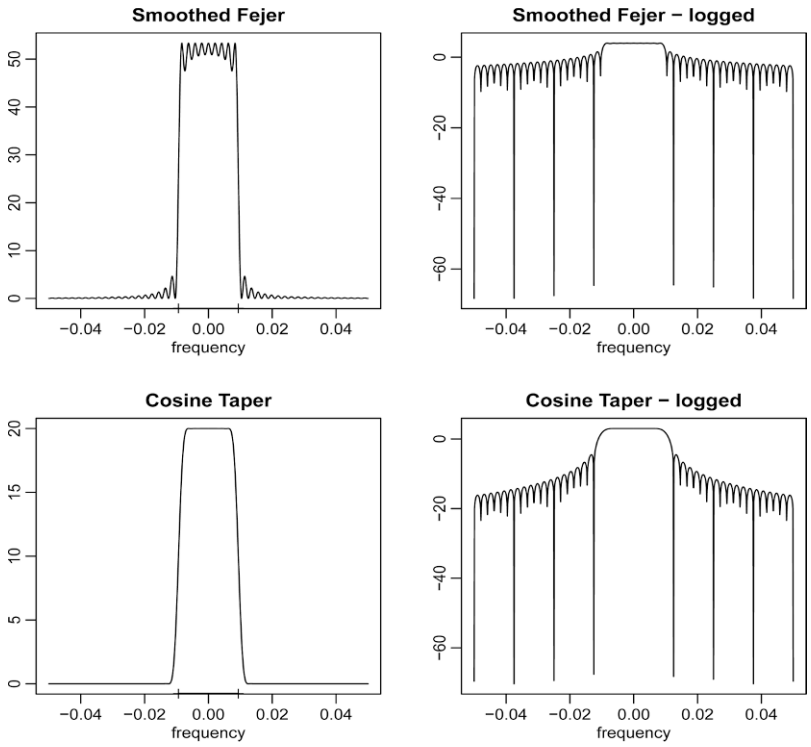

Fig. 4.9. Averaged Fejér window (top row) and the corresponding cosine taper window (bottom row) for $L = 9$, $n = 480$. The extra tic marks on the horizontal axis of the left-hand plots exhibit the predicted bandwidth, $B_w = 9/480 = .01875$.

Example 4.14 The Effect of Tapering the SOI Series

In this example, we examine the effect of tapering on the estimate of the spectrum of the SOI series. The results for the Recruitment series are similar. Figure 4.10 shows two spectral estimates plotted on a log scale. The degree of smoothing here is the same as in Example 4.13. The dashed line in Figure 4.10 shows the estimate without any tapering and hence it is the same as the estimated spectrum displayed in the top of Figure 4.8. The solid line shows the result with full tapering. Notice that the tapered spectrum does a better job in separating the yearly cycle ($\omega = 1$) and the El Niño cycle ($\omega = 1/4$).

The following R session was used to generate Figure 4.10. We note that, by default, R tapers 10% of each end of the data and leaves the middle 80% of the data alone. To instruct R not to taper, we must specify `taper=0`. For full tapering, we use the argument `taper=.5` to instruct R to taper 50% of each end of the data.

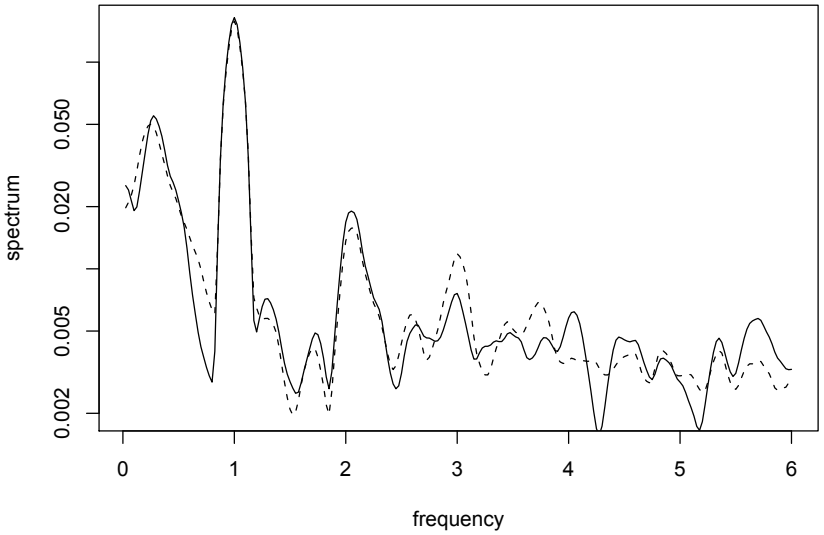

Fig. 4.10. Smoothed spectral estimates of the SOI without tapering (dashed line) and with full tapering (solid line); see Example 4.14 for details.

```
s0 = spectrum(soi, spans=c(7,7), taper=0, plot=FALSE)
s50 = spectrum(soi, spans=c(7,7), taper=.5, plot=FALSE)
plot(s0$freq, s0$spec, log="y", type="l", lty=2, ylab="spectrum",
     xlab="frequency")     # dashed line
lines(s50$freq, s50$spec)  # solid line
```

We close this section with a brief discussion of lag window estimators. First, consider the periodogram, $I(\omega_j)$, which was shown in (4.22) to be

$$I(\omega_j) = \sum_{|h|<n} \widehat{\gamma}(h) e^{-2\pi i \omega_j h}.$$

Thus, (4.56) can be written as

$$\begin{aligned} \widehat{f}(\omega) &= \sum_{|k|\le m} h_k \, I(\omega_j + k/n) \\ &= \sum_{|k|\le m} h_k \sum_{|h|<n} \widehat{\gamma}(h) e^{-2\pi i(\omega_j + k/n)h} \\ &= \sum_{|h|<n} g(h/n)\, \widehat{\gamma}(h) e^{-2\pi i \omega_j h}. \end{aligned} \tag{4.69}$$

where $g(h/n) = \sum_{|k|\le m} h_k \exp(-2\pi i k h/n)$. Equation (4.69) suggests estimators of the form

$$\widetilde{f}(\omega) = \sum_{|h|\le r} w(h/r)\, \widehat{\gamma}(h) e^{-2\pi i \omega h} \tag{4.70}$$

where $w(\cdot)$ is a weight function, called the lag window, that satisfies

(i) $w(0) = 1$
(ii) $|w(x)| \le 1$ and $w(x) = 0$ for $|x| > 1$,
(iii) $w(x) = w(-x)$.

Note that if $w(x) = 1$ for $|x| < 1$ and $r = n$, then $\widetilde{f}(\omega_j) = I(\omega_j)$, the periodogram. This result indicates the problem with the periodogram as an estimator of the spectral density is that it gives too much weight to the values of $\widehat{\gamma}(h)$ when h is large, and hence is unreliable [e.g, there is only one pair of observations used in the estimate $\widehat{\gamma}(n-1)$, and so on]. The smoothing window is defined to be

$$W(\omega) = \sum_{h=-r}^{r} w(h/r)e^{-2\pi i\omega h}, \tag{4.71}$$

and it determines which part of the periodogram will be used to form the estimate of $f(\omega)$. The asymptotic theory for $\widehat{f}(\omega)$ holds for $\widetilde{f}(\omega)$ under the same conditions and provided $r \to \infty$ as $n \to \infty$ but with $r/n \to 0$. We have

$$E\{\widetilde{f}(\omega)\} \to f(\omega), \tag{4.72}$$

$$\frac{n}{r}\mathrm{cov}\left(\widetilde{f}(\omega), \widetilde{f}(\lambda)\right) \to f^2(\omega)\int_{-1}^{1} w^2(x)dx \qquad \omega = \lambda \ne 0, 1/2. \tag{4.73}$$

In (4.73), replace $f^2(\omega)$ by 0 if $\omega \ne \lambda$ and by $2f^2(\omega)$ if $\omega = \lambda = 0$ or $1/2$.

Many authors have developed various windows and Brillinger (2001, Ch 3) and Brockwell and Davis (1991, Ch 10) are good sources of detailed information on this topic. We mention a few.

The rectangular lag window, which gives uniform weight in (4.70),

$$w(x) = 1, \quad |x| \le 1,$$

corresponds to the Dirichlet smoothing window given by

$$W(\omega) = \frac{\sin(2\pi r + \pi)\omega}{\sin(\pi\omega)}. \tag{4.74}$$

This smoothing window takes on negative values, which may lead to estimates of the spectral density that are negative a various frequencies. Using (4.73) in this case, for large n we have

$$\mathrm{var}\{\widetilde{f}(\omega)\} \approx \frac{2r}{n}f^2(\omega).$$

The Parzen lag window is defined to be

$$w(x) = \begin{cases} 1 - 6x + 6|x|^3 & |x| < 1/2, \\ 2(1-|x|)^3 & 1/2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

This leads to an approximate smoothing window of

$$W(\omega) = \frac{6}{\pi r^3} \frac{\sin^4(r\omega/4)}{\sin^4(\omega/2)}.$$

For large n, the variance of the estimator is approximately

$$\text{var}\{\widetilde{f}(\omega)\} \approx .539 f^2(\omega)/n.$$

The Tukey-Hanning lag window has the form

$$w(x) = \frac{1}{2}(1 + \cos(x)), \quad |x| \leq 1$$

which leads to the smoothing window

$$W(\omega) = \frac{1}{4} D_r(2\pi\omega - \pi/r) + \frac{1}{2} D_r(2\pi\omega) + \frac{1}{4} D_r(2\pi\omega + \pi/r)$$

where $D_r(\omega)$ is the Dirichlet kernel in (4.74). The approximate large sample variance of the estimator is

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{3r}{4n} f^2(\omega).$$

The triangular lag window, also known as the Bartlett or Fejér window, given by

$$w(x) = 1 - |x|, \quad |x| \leq 1$$

leads to the Fejér smoothing window:

$$W(\omega) = \frac{\sin^2(\pi r\omega)}{r \sin^2(\pi\omega)}.$$

In this case, (4.73) yields

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{2r}{3n} f^2(\omega).$$

The idealized rectangular smoothing window, also called the Daniell window, is given by

$$W(\omega) = \begin{cases} r & |\omega| \leq 1/2r, \\ 0 & \text{otherwise}, \end{cases}$$

and leads to the sinc lag window, namely

$$w(x) = \frac{\sin(\pi x)}{\pi x}, \quad |x| \leq 1.$$

From (4.73) we have

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{r}{n} f^2(\omega).$$

For lag window estimators, the width of the idealized rectangular window that leads to the same asymptotic variance as a given lag window estimator is sometimes called the equivalent bandwidth. For example, the bandwidth of the idealized rectangular window is $b_r = 1/r$ and the asymptotic variance is $\frac{1}{nb_r} f^2$. The asymptotic variance of the triangular window is $\frac{2r}{3n} f^2$, so setting $\frac{1}{nb_r} f^2 = \frac{2r}{3n} f^2$ and solving we get $b_r = 3/2r$ as the equivalent bandwidth.

4.6 Parametric Spectral Estimation

The methods of §4.5 lead to estimators generally referred to as nonparametric spectra because no assumption is made about the parametric form of the spectral density. In Property 4.3, we exhibited the spectrum of an ARMA process and we might consider basing a spectral estimator on this function, substituting the parameter estimates from an ARMA(p, q) fit on the data into the formula for the spectral density $f_x(\omega)$ given in (4.15). Such an estimator is called a parametric spectral estimator. For convenience, a parametric spectral estimator is obtained by fitting an AR(p) to the data, where the order p is determined by one of the model selection criteria, such as AIC, AICc, and BIC, defined in (2.19)-(2.21). Parametric autoregressive spectral estimators will often have superior resolution in problems when several closely spaced narrow spectral peaks are present and are preferred by engineers for a broad variety of problems (see Kay, 1988). The development of autoregressive spectral estimators has been summarized by Parzen (1983).

If $\widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p$ and $\widehat{\sigma}_w^2$ are the estimates from an AR(p) fit to x_t, then based on Property 4.3, a parametric spectral estimate of $f_x(\omega)$ is attained by substituting these estimates into (4.15), that is,

$$\widehat{f}_x(\omega) = \frac{\widehat{\sigma}_w^2}{|\widehat{\phi}(\mathrm{e}^{-2\pi i\omega})|^2}, \tag{4.75}$$

where

$$\widehat{\phi}(z) = 1 - \widehat{\phi}_1 z - \widehat{\phi}_2 z^2 - \cdots - \widehat{\phi}_p z^p. \tag{4.76}$$

The asymptotic distribution of the autoregressive spectral estimator has been obtained by Berk (1974) under the conditions $p \to \infty$, $p^3/n \to 0$ as $p,\ n \to \infty$, which may be too severe for most applications. The limiting results imply a confidence interval of the form

$$\frac{\widehat{f}_x(\omega)}{(1 + Cz_{\alpha/2})} \le f_x(\omega) \le \frac{\widehat{f}_x(\omega)}{(1 - Cz_{\alpha/2})}, \tag{4.77}$$

where $C = \sqrt{2p/n}$ and $z_{\alpha/2}$ is the ordinate corresponding to the upper $\alpha/2$ probability of the standard normal distribution. If the sampling distribution is to be checked, we suggest applying the bootstrap estimator to get the sampling distribution of $\widehat{f}_x(\omega)$ using a procedure similar to the one used for $p = 1$ in

Example 3.35. An alternative for higher order autoregressive series is to put the AR(p) in state-space form and use the bootstrap procedure discussed in §6.7.

An interesting fact about rational spectra of the form (4.15) is that any spectral density can be approximated, arbitrarily close, by the spectrum of an AR process.

Property 4.5 AR Spectral Approximation

Let $g(\omega)$ be the spectral density of a stationary process. Then, given $\epsilon > 0$, there is a time series with the representation

$$x_t = \sum_{k=1}^{p} \phi_k x_{t-k} + w_t$$

where w_t is white noise with variance σ_w^2, such that

$$|f_x(\omega) - g(\omega)| < \epsilon \quad \text{forall } \omega \in [-1/2, 1/2].$$

Moreover, p is finite and the roots of $\phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k$ are outside the unit circle.

One drawback of the property is that it does not tell us how large p must be before the approximation is reasonable; in some situations p may be extremely large. Property 4.5 also holds for MA and for ARMA processes in general, and a proof of the result may be found in Fuller (1996, Ch 4). We demonstrate the technique in the following example.

Example 4.15 Autoregressive Spectral Estimator for SOI

Consider obtaining results comparable to the nonparametric estimators shown in Figure 4.5 for the SOI series. Fitting successively higher order AR(p) models for $p = 1, 2, \ldots, 30$ yields a minimum BIC at $p = 15$ and a minimum AIC at $p = 16$, as shown in Figure 4.11. We can see from Figure 4.11 that BIC is very definite about which model it chooses; that is, the minimum BIC is very distinct. On the other hand, it is not clear what is going to happen with AIC; that is, the minimum is not so clear, and there is some concern that AIC will start decreasing after $p = 30$. Minimum AICc selects the $p = 15$ model, but suffers from the same uncertainty as AIC. The spectra of the two cases are almost identical, as shown in Figure 4.12, and we note the strong peaks at 52 months and 12 months corresponding to the nonparametric estimators obtained in §4.5. In addition, the harmonics of the yearly period are evident in the estimated spectrum.

To perform a similar analysis in R, the command `spec.ar` can be used to fit the best model via AIC and plot the resulting spectrum. A quick way to obtain the AIC values is to run the `ar` command as follows.

```
1 spaic = spec.ar(soi, log="no", ylim=c(0,.3))  # min AIC spec
2 text(frequency(soi)*1/52,.07,substitute(omega==1/52)) # El Nino
      Cycle
```

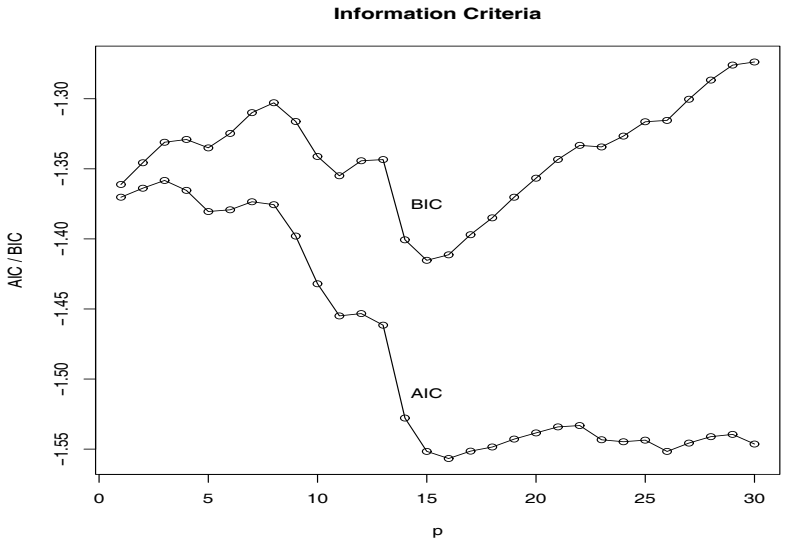

Fig. 4.11. Model selection criteria AIC and BIC as a function of order p for autoregressive models fitted to the SOI series.

```
text(frequency(soi)*1/12,.29,substitute(omega==1/12)) # Yearly Cycle
sp16 = spec.ar(soi,order=16, log="no", plot=F)
lines(sp16$freq, sp16$spec, lty="dashed")    # ar16 spec
(soi.ar = ar(soi, order.max=30))    # estimates and AICs
dev.new()
plot(1:30, soi.ar$aic[-1], type="o")   # plot AICs
```

R works only with the AIC in this case. To generate Figure 4.11 we used the following code to obtain AIC, AICc, and BIC. Because AIC and AICc are nearly identical in this example, we only graphed AIC and BIC+1; we added 1 to the BIC to reduce white space in the graphic.

```
n = length(soi)
AIC = rep(0, 30) -> AICc -> BIC
for (k in 1:30){
  fit = ar(soi, order=k, aic=FALSE)
  sigma2 = var(fit$resid, na.rm=TRUE)
  BIC[k] = log(sigma2) + (k*log(n)/n)
  AICc[k] = log(sigma2) + ((n+k)/(n-k-2))
  AIC[k] = log(sigma2) + ((n+2*k)/n)    }
IC = cbind(AIC, BIC+1)
ts.plot(IC, type="o", xlab="p", ylab="AIC / BIC")
text(15, -1.5, "AIC");  text(15, -1.38, "BIC")
```

Finally, it should be mentioned that any parametric spectrum, say $f(\omega;\boldsymbol{\theta})$, depending on the vector parameter $\boldsymbol{\theta}$ can be estimated via the Whittle likelihood (Whittle, 1961), using the approximate properties of the discrete Fourier

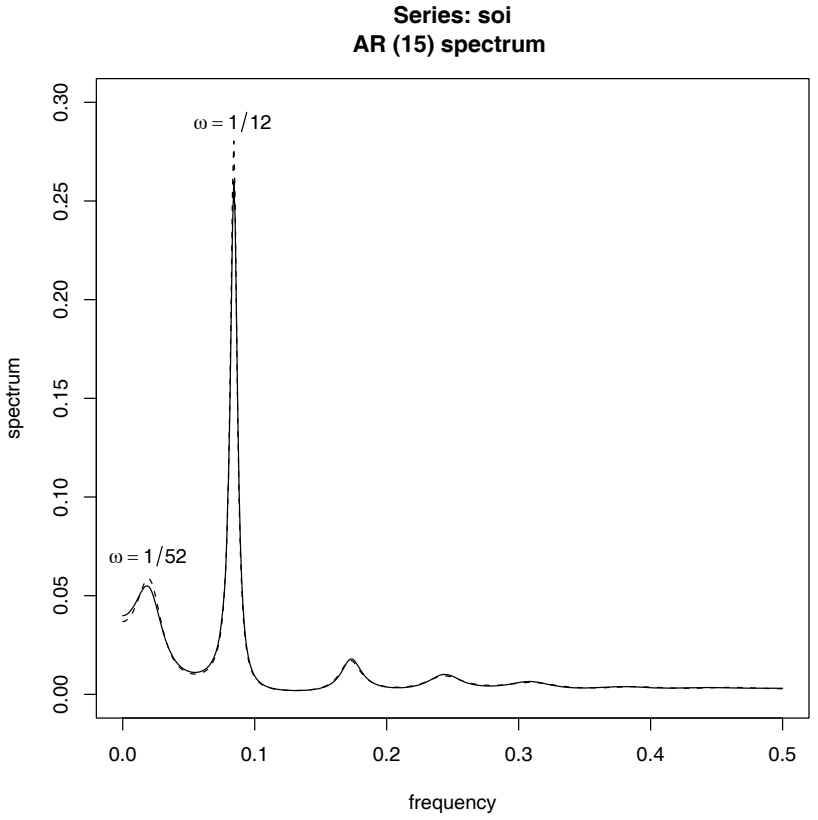

Fig. 4.12. Autoregressive spectral estimators for the SOI series using models selected by AIC ($p = 16$, solid line) and by BIC and AICc ($p = 15$, dashed line). The first peak corresponds to the El Niño period of 52 months.

transform derived in Appendix C. We have that the DFTs, $d(\omega_j)$, are approximately complex normally distributed with mean zero and variance $f(\omega_j;\boldsymbol{\theta})$ and are approximately independent for $\omega_j \neq \omega_k$. This implies that an approximate log likelihood can be written in the form

$$\ln L(\boldsymbol{x};\boldsymbol{\theta}) \approx -\sum_{0<\omega_j<1/2}\left(\ln f_x(\omega_j;\boldsymbol{\theta}) + \frac{|d(\omega_j)|^2}{f_x(\omega_j;\boldsymbol{\theta})}\right), \tag{4.78}$$

where the sum is sometimes expanded to include the frequencies $\omega_j = 0, 1/2$. If the form with the two additional frequencies is used, the multiplier of the sum will be unity, except for the purely real points at $\omega_j = 0, 1/2$ for which the multiplier is $1/2$. For a discussion of applying the Whittle approximation to the problem of estimating parameters in an ARMA spectrum, see Anderson (1978). The Whittle likelihood is especially useful for fitting long memory models that will be discussed in Chapter 5.

4.7 Multiple Series and Cross-Spectra

The notion of analyzing frequency fluctuations using classical statistical ideas extends to the case in which there are several jointly stationary series, for example, x_t and y_t. In this case, we can introduce the idea of a correlation indexed by frequency, called the coherence. The results in Appendix C, §C.2, imply the covariance function

$$\gamma_{xy}(h) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]$$

has the representation

$$\gamma_{xy}(h) = \int_{-1/2}^{1/2} f_{xy}(\omega) \mathrm{e}^{2\pi i \omega h} \, d\omega \quad h = 0, \pm 1, \pm 2, ..., \tag{4.79}$$

where the cross-spectrum is defined as the Fourier transform

$$f_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \mathrm{e}^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2, \tag{4.80}$$

assuming that the cross-covariance function is absolutely summable, as was the case for the autocovariance. The cross-spectrum is generally a complex-valued function, and it is often written as[14]

$$f_{xy}(\omega) = c_{xy}(\omega) - i q_{xy}(\omega), \tag{4.81}$$

where

$$c_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \cos(2\pi\omega h) \tag{4.82}$$

and

$$q_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \sin(2\pi\omega h) \tag{4.83}$$

are defined as the cospectrum and quadspectrum, respectively. Because of the relationship $\gamma_{yx}(h) = \gamma_{xy}(-h)$, it follows, by substituting into (4.80) and rearranging, that

$$f_{yx}(\omega) = \overline{f_{xy}(\omega)}. \tag{4.84}$$

This result, in turn, implies that the cospectrum and quadspectrum satisfy

$$c_{yx}(\omega) = c_{xy}(\omega) \tag{4.85}$$

and

$$q_{yx}(\omega) = -q_{xy}(\omega). \tag{4.86}$$

[14] For this section, it will be useful to recall the facts $e^{-i\alpha} = \cos(\alpha) - i\sin(\alpha)$ and if $z = a + ib$, then $\overline{z} = a - ib$.

An important example of the application of the cross-spectrum is to the problem of predicting an output series y_t from some input series x_t through a linear filter relation such as the three-point moving average considered below. A measure of the strength of such a relation is the squared coherence function, defined as

$$\rho^2_{y\cdot x}(\omega) = \frac{|f_{yx}(\omega)|^2}{f_{xx}(\omega) f_{yy}(\omega)}, \tag{4.87}$$

where $f_{xx}(\omega)$ and $f_{yy}(\omega)$ are the individual spectra of the x_t and y_t series, respectively. Although we consider a more general form of this that applies to multiple inputs later, it is instructive to display the single input case as (4.87) to emphasize the analogy with conventional squared correlation, which takes the form

$$\rho^2_{yx} = \frac{\sigma^2_{yx}}{\sigma^2_x \sigma^2_y},$$

for random variables with variances σ^2_x and σ^2_y and covariance $\sigma_{yx} = \sigma_{xy}$. This motivates the interpretation of squared coherence and the squared correlation between two time series at frequency ω.

Example 4.16 Three-Point Moving Average

As a simple example, we compute the cross-spectrum between x_t and the three-point moving average $y_t = (x_{t-1}+x_t+x_{t+1})/3$, where x_t is a stationary input process with spectral density $f_{xx}(\omega)$. First,

$$\begin{aligned}
\gamma_{xy}(h) &= \operatorname{cov}(x_{t+h}, y_t) = \tfrac{1}{3}\operatorname{cov}(x_{t+h}, x_{t-1} + x_t + x_{t+1}) \\
&= \frac{1}{3}\big(\gamma_{xx}(h+1) + \gamma_{xx}(h) + \gamma_{xx}(h-1)\big) \\
&= \frac{1}{3}\int_{-1/2}^{1/2} \big(\mathrm{e}^{2\pi i\omega} + 1 + \mathrm{e}^{-2\pi i\omega}\big)\, \mathrm{e}^{2\pi i\omega h} f_{xx}(\omega)\, d\omega \\
&= \frac{1}{3}\int_{-1/2}^{1/2} [1 + 2\cos(2\pi\omega)] f_{xx}(\omega) \mathrm{e}^{2\pi i\omega h}\, d\omega,
\end{aligned}$$

where we have use (4.11). Using the uniqueness of the Fourier transform, we argue from the spectral representation (4.79) that

$$f_{xy}(\omega) = \tfrac{1}{3}\left[1 + 2\cos(2\pi\omega)\right]\, f_{xx}(\omega)$$

so that the cross-spectrum is real in this case. From Example 4.5, the spectral density of y_t is

$$\begin{aligned}
f_{yy}(\omega) &= \tfrac{1}{9}[3 + 4\cos(2\pi\omega) + 2\cos(4\pi\omega)] f_{xx}(\omega) \\
&= \tfrac{1}{9}\,\left[1 + 2\cos(2\pi\omega)\right]^2\, f_{xx}(\omega),
\end{aligned}$$

using the identity $\cos(2\alpha) = 2\cos^2(\alpha) - 1$ in the last step. Substituting into (4.87) yields the squared coherence between x_t and y_t as unity over all

frequencies. This is a characteristic inherited by more general linear filters, as will be shown in Problem 4.23. However, if some noise is added to the three-point moving average, the coherence is not unity; these kinds of models will be considered in detail later.

Property 4.6 Spectral Representation of a Vector Stationary Process

If the elements of the $p \times p$ autocovariance function matrix

$$\Gamma(h) = E[(\boldsymbol{x}_{t+h} - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})']$$

of a p-dimensional stationary time series, $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, has elements satisfying

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty \tag{4.88}$$

for all $j, k = 1, \ldots, p$, then $\Gamma(h)$ has the representation

$$\Gamma(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} \; f(\omega) \; d\omega \quad h = 0, \pm 1, \pm 2, ..., \tag{4.89}$$

as the inverse transform of the spectral density matrix, $f(\omega) = \{f_{jk}(\omega)\}$, for $j, k = 1, \ldots, p$, with elements equal to the cross-spectral components. The matrix $f(\omega)$ has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h)\mathrm{e}^{-2\pi i\omega h} \quad -1/2 \leq \omega \leq 1/2. \tag{4.90}$$

Example 4.17 Spectral Matrix of a Bivariate Process

Consider a jointly stationary bivariate process (x_t, y_t). We arrange the autocovariances in the matrix

$$\Gamma(h) = \begin{pmatrix} \gamma_{xx}(h) & \gamma_{xy}(h) \\ \gamma_{yx}(h) & \gamma_{yy}(h) \end{pmatrix}.$$

The spectral matrix would be given by

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix},$$

where the Fourier transform (4.89) and (4.90) relate the autocovariance and spectral matrices.

The extension of spectral estimation to vector series is fairly obvious. For the vector series $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, we may use the vector of DFTs, say $\boldsymbol{d}(\omega_j) = (d_1(\omega_j), d_2(\omega_j), \ldots, d_p(\omega_j))'$, and estimate the spectral matrix by

$$\bar{f}(\omega) = L^{-1} \sum_{k=-m}^{m} I(\omega_j + k/n) \tag{4.91}$$

where now

$$I(\omega_j) = \boldsymbol{d}(\omega_j)\, \boldsymbol{d}^*(\omega_j) \tag{4.92}$$

is a $p \times p$ complex matrix.[15]

Again, the series may be tapered before the DFT is taken in (4.91) and we can use weighted estimation,

$$\widehat{f}(\omega) = \sum_{k=-m}^{m} h_k\, I(\omega_j + k/n) \tag{4.93}$$

where $\{h_k\}$ are weights as defined in (4.56). The estimate of squared coherence between two series, y_t and x_t is

$$\widehat{\rho}^2_{y\cdot x}(\omega) = \frac{|\widehat{f}_{yx}(\omega)|^2}{\widehat{f}_{xx}(\omega)\widehat{f}_{yy}(\omega)}. \tag{4.94}$$

If the spectral estimates in (4.94) are obtained using equal weights, we will write $\bar{\rho}^2_{y\cdot x}(\omega)$ for the estimate.

Under general conditions, if $\rho^2_{y\cdot x}(\omega) > 0$ then

$$|\widehat{\rho}_{y\cdot x}(\omega)| \sim AN\left(|\rho_{y\cdot x}(\omega)|, \left(1 - \rho^2_{y\cdot x}(\omega)\right)^2 / 2L_h\right) \tag{4.95}$$

where L_h is defined in (4.57); the details of this result may be found in Brockwell and Davis (1991, Ch 11). We may use (4.95) to obtain approximate confidence intervals for the squared coherency $\rho^2_{y\cdot x}(\omega)$.

We can test the hypothesis that $\rho^2_{y\cdot x}(\omega) = 0$ if we use $\bar{\rho}^2_{y\cdot x}(\omega)$ for the estimate with $L > 1$,[16] that is,

$$\bar{\rho}^2_{y\cdot x}(\omega) = \frac{|\bar{f}_{yx}(\omega)|^2}{\bar{f}_{xx}(\omega)\bar{f}_{yy}(\omega)}. \tag{4.96}$$

In this case, under the null hypothesis, the statistic

$$F = \frac{\bar{\rho}^2_{y\cdot x}(\omega)}{(1 - \bar{\rho}^2_{y\cdot x}(\omega))}(L-1) \tag{4.97}$$

has an approximate F-distribution with 2 and $2L-2$ degrees of freedom. When the series have been extended to length n', we replace $2L-2$ by $df-2$,

[15] If Z is a complex matrix, then $Z^* = \overline{Z}'$ denotes the conjugate transpose operation. That is, Z^* is the result of replacing each element of Z by its complex conjugate and transposing the resulting matrix.

[16] If $L = 1$ then $\bar{\rho}^2_{y\cdot x}(\omega) \equiv 1$.

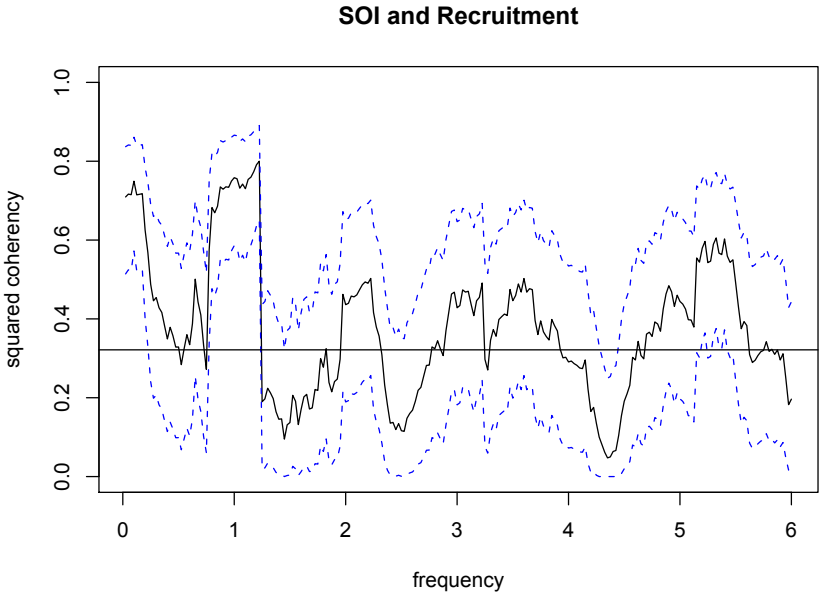

Fig. 4.13. Squared coherency between the SOI and Recruitment series; $L = 19$, $n = 453$, $n' = 480$, and $\alpha = .001$. The horizontal line is $C_{.001}$.

where df is defined in (4.52). Solving (4.97) for a particular significance level α leads to

$$C_\alpha = \frac{F_{2,2L-2}(\alpha)}{L - 1 + F_{2,2L-2}(\alpha)} \tag{4.98}$$

as the approximate value that must be exceeded for the original squared coherence to be able to reject $\rho^2_{y\cdot x}(\omega) = 0$ at an a priori specified frequency.

Example 4.18 Coherence Between SOI and Recruitment

Figure 4.13 shows the squared coherence between the SOI and Recruitment series over a wider band than was used for the spectrum. In this case, we used $L = 19$, $df = 2(19)(453/480) \approx 36$ and $F_{2,df-2}(.001) \approx 8.53$ at the significance level $\alpha = .001$. Hence, we may reject the hypothesis of no coherence for values of $\bar{\rho}^2_{y\cdot x}(\omega)$ that exceed $C_{.001} = .32$. We emphasize that this method is crude because, in addition to the fact that the F-statistic is approximate, we are examining the squared coherence across all frequencies with the Bonferroni inequality, (4.55), in mind. Figure 4.13 also exhibits confidence bands as part of the R plotting routine. We emphasize that these bands are only valid for ω where $\rho^2_{y\cdot x}(\omega) > 0$.

In this case, the seasonal frequency and the El Niño frequencies ranging between about 3 and 7 year periods are strongly coherent. Other frequencies are also strongly coherent, although the strong coherence is less impressive because the underlying power spectrum at these higher frequencies is fairly

small. Finally, we note that the coherence is persistent at the seasonal harmonic frequencies.

This example may be reproduced using the following R commands.

```
sr=spec.pgram(cbind(soi,rec),kernel("daniell",9),taper=0,plot=FALSE)
sr$df                    # df = 35.8625
f = qf(.999, 2, sr$df-2)  # = 8.529792
C = f/(18+f)              # = 0.318878
plot(sr, plot.type = "coh", ci.lty = 2)
abline(h = C)
```

4.8 Linear Filters

Some of the examples of the previous sections have hinted at the possibility the distribution of power or variance in a time series can be modified by making a linear transformation. In this section, we explore that notion further by defining a linear filter and showing how it can be used to extract signals from a time series. The linear filter modifies the spectral characteristics of a time series in a predictable way, and the systematic development of methods for taking advantage of the special properties of linear filters is an important topic in time series analysis.

A linear filter uses a set of specified coefficients a_j, for $j = 0, \pm 1, \pm 2, \ldots$, to transform an input series, x_t, producing an output series, y_t, of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |a_j| < \infty. \tag{4.99}$$

The form (4.99) is also called a convolution in some statistical contexts. The coefficients, collectively called the *impulse response function*, are required to satisfy absolute summability so y_t in (4.99) exists as a limit in mean square and the infinite Fourier transform

$$A_{yx}(\omega) = \sum_{j=-\infty}^{\infty} a_j \, \mathrm{e}^{-2\pi i \omega j}, \tag{4.100}$$

called the *frequency response function*, is well defined. We have already encountered several linear filters, for example, the simple three-point moving average in Example 4.16, which can be put into the form of (4.99) by letting $a_{-1} = a_0 = a_1 = 1/3$ and taking $a_t = 0$ for $|j| \geq 2$.

The importance of the linear filter stems from its ability to enhance certain parts of the spectrum of the input series. To see this, assuming that x_t is stationary with spectral density $f_{xx}(\omega)$, the autocovariance function of the filtered output y_t in (4.99) can be derived as

$$\begin{aligned}
\gamma_{yy}(h) &= \mathrm{cov}(y_{t+h}, y_t) \\
&= \mathrm{cov}\left(\sum_r a_r x_{t+h-r}, \sum_s a_s x_{t-s}\right) \\
&= \sum_r \sum_s a_r \gamma_{xx}(h-r+s) a_s \\
&= \sum_r \sum_s a_r \left[\int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega(h-r+s)} f_{xx}(\omega) d\omega\right] a_s \\
&= \int_{-1/2}^{1/2} \left(\sum_r a_r \mathrm{e}^{-2\pi i\omega r}\right)\left(\sum_s a_s \mathrm{e}^{2\pi i\omega s}\right) \mathrm{e}^{2\pi i\omega h} f_{xx}(\omega)\ d\omega \\
&= \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} |A_{yx}(\omega)|^2 f_{xx}(\omega)\ d\omega,
\end{aligned}$$

where we have first replaced $\gamma_{xx}(\cdot)$ by its representation (4.11) and then substituted $A_{yx}(\omega)$ from (4.100). The computation is one we do repeatedly, exploiting the uniqueness of the Fourier transform. Now, because the left-hand side is the Fourier transform of the spectral density of the output, say, $f_{yy}(\omega)$, we get the important filtering property as follows.

Property 4.7 Output Spectrum of a Filtered Stationary Series

The spectrum of the filtered output y_t in (4.99) is related to the spectrum of the input x_t by

$$f_{yy}(\omega) = |A_{yx}(\omega)|^2\ f_{xx}(\omega), \tag{4.101}$$

where the frequency response function $A_{yx}(\omega)$ is defined in (4.100).

The result (4.101) enables us to calculate the exact effect on the spectrum of any given filtering operation. This important property shows the spectrum of the input series is changed by filtering and the effect of the change can be characterized as a frequency-by-frequency multiplication by the squared magnitude of the frequency response function. Again, an obvious analogy to a property of the variance in classical statistics holds, namely, if x is a random variable with variance σ_x^2, then $y = ax$ will have variance $\sigma_y^2 = a^2\sigma_x^2$, so the variance of the linearly transformed random variable is changed by multiplication by a^2 in much the same way as the linearly filtered spectrum is changed in (4.101).

Finally, we mention that Property 4.3, which was used to get the spectrum of an ARMA process, is just a special case of Property 4.7 where in (4.99), $x_t = w_t$ is white noise, in which case $f_{xx}(\omega) = \sigma_w^2$, and $a_j = \psi_j$, in which case

$$A_{yx}(\omega) = \psi(\mathrm{e}^{-2\pi i\omega}) = \theta(\mathrm{e}^{-2\pi i\omega})\ /\ \phi(\mathrm{e}^{-2\pi i\omega}).$$

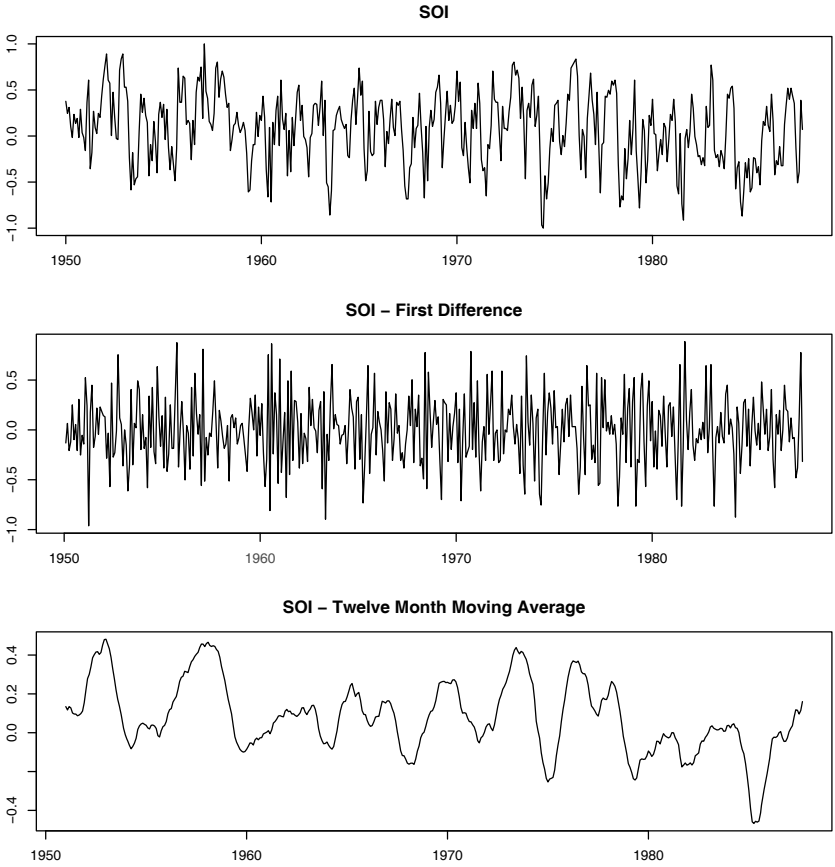

Fig. 4.14. SOI series (top) compared with the differenced SOI (middle) and a centered 12-month moving average (bottom).

Example 4.19 First Difference and Moving Average Filters

We illustrate the effect of filtering with two common examples, the first difference filter

$$y_t = \nabla x_t = x_t - x_{t-1}$$

and the symmetric moving average filter

$$y_t = \tfrac{1}{24}\big(x_{t-6} + x_{t+6}\big) + \tfrac{1}{12}\sum_{r=-5}^{5} x_{t-r},$$

which is a modified Daniell kernel with $m = 6$. The results of filtering the SOI series using the two filters are shown in the middle and bottom panels of Figure 4.14. Notice that the effect of differencing is to roughen the series because it tends to retain the higher or faster frequencies. The centered

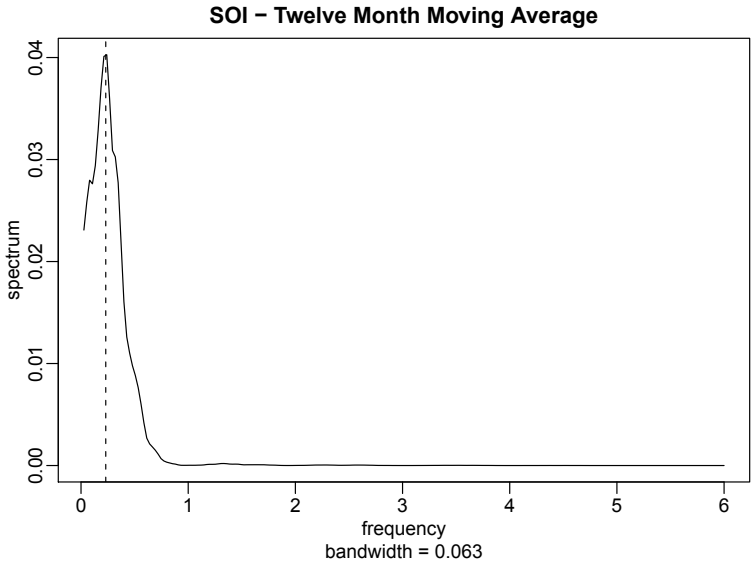

Fig. 4.15. Spectral analysis of SOI after applying a 12-month moving average filter. The vertical line corresponds to the 52-month cycle.

moving average smoothes the series because it retains the lower frequencies and tends to attenuate the higher frequencies. In general, differencing is an example of a *high-pass filter* because it retains or passes the higher frequencies, whereas the moving average is a *low-pass filter* because it passes the lower or slower frequencies.

Notice that the slower periods are enhanced in the symmetric moving average and the seasonal or yearly frequencies are attenuated. The filtered series makes about 9 cycles in the length of the data (about one cycle every 52 months) and the moving average filter tends to enhance or extract the signal that is associated with El Niño. Moreover, by the low-pass filtering of the data, we get a better sense of the El Niño effect and its irregularity. Figure 4.15 shows the results of a spectral analysis on the low-pass filtered SOI series. It is clear that all high frequency behavior has been removed and the El Niño cycle is accentuated; the dotted vertical line in the figure corresponds to the 52 months cycle.

Now, having done the filtering, it is essential to determine the exact way in which the filters change the input spectrum. We shall use (4.100) and (4.101) for this purpose. The first difference filter can be written in the form (4.99) by letting $a_0 = 1, a_1 = -1$, and $a_r = 0$ otherwise. This implies that

$$A_{yx}(\omega) = 1 - \mathrm{e}^{-2\pi i\omega},$$

and the squared frequency response becomes

$$|A_{yx}(\omega)|^2 = (1 - \mathrm{e}^{-2\pi i\omega})(1 - \mathrm{e}^{2\pi i\omega}) = 2[1 - \cos(2\pi\omega)]. \tag{4.102}$$

The top panel of Figure 4.16 shows that the first difference filter will attenuate the lower frequencies and enhance the higher frequencies because the multiplier of the spectrum, $|A_{yx}(\omega)|^2$, is large for the higher frequencies and small for the lower frequencies. Generally, the slow rise of this kind of filter does not particularly recommend it as a procedure for retaining only the high frequencies.

For the centered 12-month moving average, we can take $a_{-6} = a_6 = 1/24$, $a_k = 1/12$ for $-5 \leq k \leq 5$ and $a_k = 0$ elsewhere. Substituting and recognizing the cosine terms gives

$$A_{yx}(\omega) = \tfrac{1}{12}\Big[1 + \cos(12\pi\omega) + 2\sum_{k=1}^{5}\cos(2\pi\omega k)\Big]. \tag{4.103}$$

Plotting the squared frequency response of this function as in Figure 4.16 shows that we can expect this filter to cut most of the frequency content above .05 cycles per point. This corresponds to eliminating periods shorter than $T = 1/.05 = 20$ points. In particular, this drives down the yearly components with periods of $T = 12$ months and enhances the El Niño frequency, which is somewhat lower. The filter is not completely efficient at attenuating high frequencies; some power contributions are left at higher frequencies, as shown in the function $|A_{yx}(\omega)|^2$ and in the spectrum of the moving average shown in Figure 4.3.

The following R session shows how to filter the data, perform the spectral analysis of this example, and plot the squared frequency response curve of the difference filter.

```
par(mfrow=c(3,1))
plot(soi)  # plot data
plot(diff(soi))  # plot first difference
k = kernel("modified.daniell", 6)  # filter weights
plot(soif <- kernapply(soi, k))    # plot 12 month filter
dev.new()
spectrum(soif, spans=9, log="no")  # spectral analysis
abline(v=12/52, lty="dashed")
dev.new()
w = seq(0, .5, length=500)  # frequency response
FR = abs(1-exp(2i*pi*w))^2
plot(w, FR, type="l")
```

The two filters discussed in the previous example were different in that the frequency response function of the first difference was complex-valued, whereas the frequency response of the moving average was purely real. A short derivation similar to that used to verify (4.101) shows, when x_t and y_t are related by the linear filter relation (4.99), the cross-spectrum satisfies

$$f_{yx}(\omega) = A_{yx}(\omega) f_{xx}(\omega),$$

so the frequency response is of the form

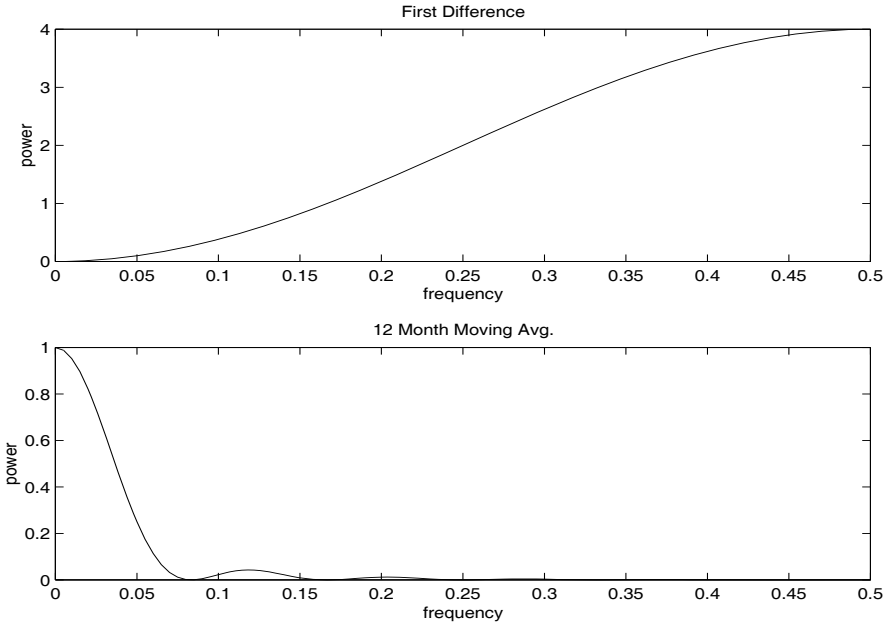

Fig. 4.16. Squared frequency response functions of the first difference and 12-month moving average filters.

$$A_{yx}(\omega) = \frac{f_{yx}(\omega)}{f_{xx}(\omega)} \tag{4.104}$$

$$= \frac{c_{yx}(\omega)}{f_{xx}(\omega)} - i\frac{q_{yx}(\omega)}{f_{xx}(\omega)}, \tag{4.105}$$

where we have used (4.81) to get the last form. Then, we may write (4.105) in polar coordinates as

$$A_{yx}(\omega) = |A_{yx}(\omega)|\ \exp\{-i\ \phi_{yx}(\omega)\}, \tag{4.106}$$

where the amplitude and phase of the filter are defined by

$$|A_{yx}(\omega)| = \frac{\sqrt{c_{yx}^2(\omega) + q_{yx}^2(\omega)}}{f_{xx}(\omega)} \tag{4.107}$$

and

$$\phi_{yx}(\omega) = \tan^{-1}\left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)}\right). \tag{4.108}$$

A simple interpretation of the phase of a linear filter is that it exhibits time delays as a function of frequency in the same way as the spectrum represents the variance as a function of frequency. Additional insight can be gained by considering the simple delaying filter

$$y_t = Ax_{t-D},$$

where the series gets replaced by a version, amplified by multiplying by A and delayed by D points. For this case,

$$f_{yx}(\omega) = A\mathrm{e}^{-2\pi i\omega D} f_{xx}(\omega),$$

and the amplitude is $|A|$, and the phase is

$$\phi_{yx}(\omega) = -2\pi\omega D,$$

or just a linear function of frequency ω. For this case, applying a simple time delay causes phase delays that depend on the frequency of the periodic component being delayed. Interpretation is further enhanced by setting

$$x_t = \cos(2\pi\omega t),$$

in which case

$$y_t = A\cos(2\pi\omega t - 2\pi\omega D).$$

Thus, the output series, y_t, has the same period as the input series, x_t, but the amplitude of the output has increased by a factor of $|A|$ and the phase has been changed by a factor of $-2\pi\omega D$.

Example 4.20 Difference and Moving Average Filters

We consider calculating the amplitude and phase of the two filters discussed in Example 4.19. The case for the moving average is easy because $A_{yx}(\omega)$ given in (4.103) is purely real. So, the amplitude is just $|A_{yx}(\omega)|$ and the phase is $\phi_{yx}(\omega) = 0$. In general, symmetric ($a_j = a_{-j}$) filters have zero phase. The first difference, however, changes this, as we might expect from the example above involving the time delay filter. In this case, the squared amplitude is given in (4.102). To compute the phase, we write

$$\begin{aligned} A_{yx}(\omega) &= 1 - e^{-2\pi i\omega} = e^{-i\pi\omega}(e^{i\pi\omega} - e^{-i\pi\omega}) \\ &= 2ie^{-i\pi\omega}\sin(\pi\omega) = 2\sin^2(\pi\omega) + 2i\cos(\pi\omega)\sin(\pi\omega) \\ &= \frac{c_{yx}(\omega)}{f_{xx}(\omega)} - i\frac{q_{yx}(\omega)}{f_{xx}(\omega)}, \end{aligned}$$

so

$$\phi_{yx}(\omega) = \tan^{-1}\left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)}\right) = \tan^{-1}\left(\frac{\cos(\pi\omega)}{\sin(\pi\omega)}\right).$$

Noting that

$$\cos(\pi\omega) = \sin(-\pi\omega + \pi/2)$$

and that

$$\sin(\pi\omega) = \cos(-\pi\omega + \pi/2),$$

we get

$$\phi_{yx}(\omega) = -\pi\omega + \pi/2,$$

and the phase is again a linear function of frequency.

The above tendency of the frequencies to arrive at different times in the filtered version of the series remains as one of two annoying features of the difference type filters. The other weakness is the gentle increase in the frequency response function. If low frequencies are really unimportant and high frequencies are to be preserved, we would like to have a somewhat sharper response than is obvious in Figure 4.16. Similarly, if low frequencies are important and high frequencies are not, the moving average filters are also not very efficient at passing the low frequencies and attenuating the high frequencies. Improvement is possible by using longer filters, obtained by approximations to the infinite inverse Fourier transform. The design of filters will be discussed in §4.10 and §4.11.

We will occasionally use results for multivariate series $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ that are comparable to the simple property shown in (4.101). Consider the matrix filter

$$\boldsymbol{y}_t = \sum_{j=-\infty}^{\infty} A_j \boldsymbol{x}_{t-j}, \tag{4.109}$$

where $\{A_j\}$ denotes a sequence of $q \times p$ matrices such that $\sum_{j=-\infty}^{\infty} \|A_j\| < \infty$ and $\|\cdot\|$ denotes any matrix norm, $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ is a $p \times 1$ stationary vector process with mean vector $\boldsymbol{\mu}_x$ and $p \times p$, matrix covariance function $\Gamma_{xx}(h)$ and spectral matrix $f_{xx}(\omega)$, and $\boldsymbol{y}_t$ is the $q \times 1$ vector output process. Then, we can obtain the following property.

Property 4.8 Output Spectral Matrix of a Linearly Filtered Stationary Vector Series

The spectral matrix of the filtered output $\boldsymbol{y}_t$ in (4.109) is related to the spectrum of the input $\boldsymbol{x}_t$ by

$$f_{yy}(\omega) = \mathcal{A}(\omega) f_{xx}(\omega) \mathcal{A}^*(\omega), \tag{4.110}$$

where the matrix frequency response function $\mathcal{A}(\omega)$ is defined by

$$\mathcal{A}(\omega) = \sum_{j=-\infty}^{\infty} A_j \exp(-2\pi i \omega j). \tag{4.111}$$

4.9 Dynamic Fourier Analysis and Wavelets

If a time series, x_t, is stationary, its second-order behavior remains the same, regardless of the time t. It makes sense to match a stationary time series with sines and cosines because they, too, behave the same forever. Indeed, based on the Spectral Representation Theorem (Appendix C, §C.1), we may regard a stationary series as the superposition of sines and cosines that oscillate at various frequencies. As seen in this text, however, many time series are not stationary. Typically, the data are coerced into stationarity via transformations, or we restrict attention to parts of the data where stationarity appears

to adhere. In some cases, the nonstationarity of a time series is of interest. That is to say, it is the local behavior of the process, and not the global behavior of the process, that is of concern to the investigator. As a case in point, we mention the explosion and earthquake series first presented in Example 1.7 (see Figure 1.7). The following example emphasizes the importance of dynamic (or time-frequency) Fourier analysis.

Example 4.21 Dynamic Spectral Analysis of Seismic Traces

Consider the earthquake and explosion series displayed in Figure 1.7; it should be apparent that the dynamics of the series are changing with time. The goal of this analysis is to summarize the spectral behavior of the signal as it evolves over time.

First, a spectral analysis is performed on a short section of the data. Then, the section is shifted, and a spectral analysis is performed on the new section. This process is repeated until the end of the data, and the results are shown an image in Figures 4.17 and 4.18; in the images, darker areas correspond to higher power. Specifically, in this example, let x_t, for $t = 1, \ldots, 2048$, represent the series of interest. Then, the sections of the data that were analyzed were $\{x_{t_k+1}, \ldots, x_{t_k+256}\}$, for $t_k = 128k$, and $k = 0, 1, \ldots, 14$; e.g., the first section analyzed is $\{x_1, \ldots, x_{256}\}$, the second section analyzed is $\{x_{129}, \ldots, x_{384}\}$, and so on. Each section of 256 observations was tapered using a cosine bell, and spectral estimation was performed using a repeated Daniell kernel with weights $\frac{1}{9}\{1, 2, 3, 2, 1\}$; see page 204. The sections overlap each other, however, this practice is not necessary and sometimes not desirable.[17]

The results of the dynamic analysis are shown as the estimated spectra for frequencies up to 10 Hz (the folding frequency is 20 Hz) for each starting location (time), $t_k = 128k$, with $k = 0, 1, \ldots, 14$. The S component for the earthquake shows power at the low frequencies only, and the power remains strong for a long time. In contrast, the explosion shows power at higher frequencies than the earthquake, and the power of the signals (P and S waves) does not last as long as in the case of the earthquake.

The following is an R session that corresponds to the analysis of the explosion series. The images are generated using `filled.contour()` on the log of the power; this, as well as using a gray scale and limiting the number of levels was done to produce a decent black-and-white graphic. The images look better in color, so we advise removing the `nlevels=...` and the `col=gray(...)` parts of the code. We also include the code for obtaining a

[17] A number of technical problems exist in this setting because the process of interest is nonstationary and we have not specified the nature of the nonstationarity. In addition, overlapping intervals complicate matters by introducing another layer of dependencies among the spectra. Consequently, the spectral estimates of contiguous sections are dependent in a non-trivial way that we have not specified. Nevertheless, as seen from this example, dynamic spectral analysis can be a helpful tool in summarizing the local behavior of a time series.

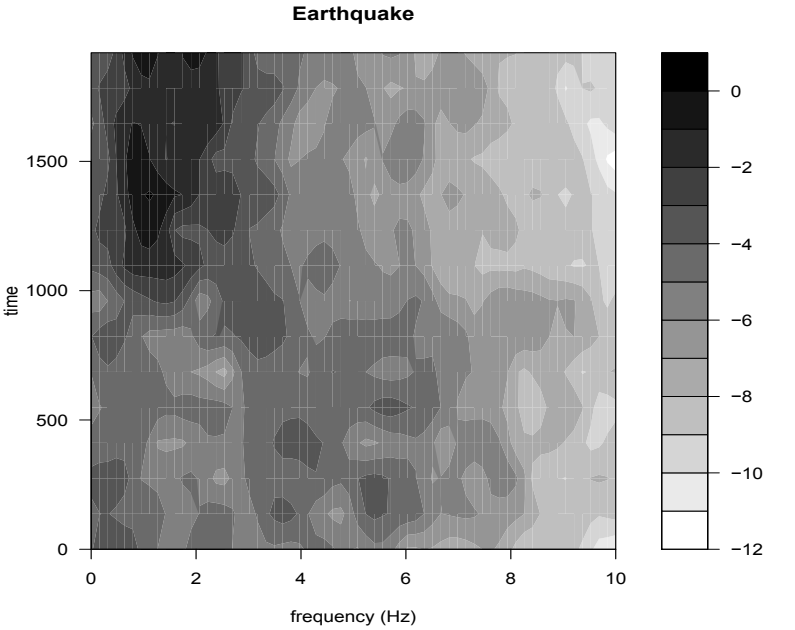

Fig. 4.17. Time-frequency image for the dynamic Fourier analysis of the earthquake series shown in Figure 1.7.

three-dimensional graphic to display the information, however, the graphic is not exhibited in the text.

```
nobs = length(EXP6)  # number of observations
wsize = 256   # window size
overlap = 128  # overlap
ovr = wsize-overlap
nseg = floor(nobs/ovr)-1;  # number of segments
krnl = kernel("daniell", c(1,1))  # kernel
ex.spec = matrix(0, wsize/2, nseg)
 for (k in 1:nseg) {
  a = ovr*(k-1)+1
  b = wsize+ovr*(k-1)
  ex.spec[,k] = spectrum(EXP6[a:b], krnl, taper=.5, plot=F)$spec }
x = seq(0, 10,  len = nrow(ex.spec)/2)
y = seq(0, ovr*nseg, len = ncol(ex.spec))
z = ex.spec[1:(nrow(ex.spec)/2),]
filled.contour(x , y, log(z), ylab="time",xlab="frequency (Hz)",
     nlevels=12, col=gray(11:0/11), main="Explosion")
persp(x, y, z, zlab="Power",xlab="frequency (Hz)",ylab="time",
     ticktype="detailed", theta=25,d=2, main="Explosion")  # not shown
```

One way to view the time-frequency analysis of Example 4.21 is to consider it as being based on local transforms of the data x_t of the form

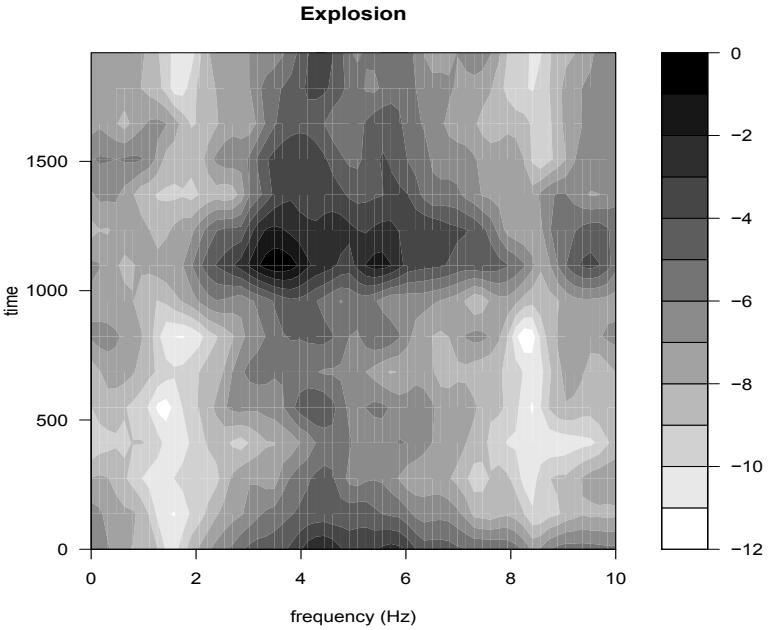

Fig. 4.18. Time-frequency image for the dynamic Fourier analysis of the explosion series shown in Figure 1.7.

$$d_{j,k} = n^{-1/2} \sum_{t=1}^{n} x_t \psi_{j,k}(t), \tag{4.112}$$

where

$$\psi_{j,k}(t) = \begin{cases} (n/m)^{1/2} h_t \, e^{-2\pi i t j/m} & t \in [t_k + 1, t_k + m], \\ 0 & \text{otherwise}, \end{cases} \tag{4.113}$$

where h_t is a taper and m is some fraction of n. In Example 4.21, $n = 2048$, $m = 256$, $t_k = 128k$, for $k = 0, 1, \ldots, 14$, and h_t was a cosine bell taper over 256 points. In (4.112) and (4.113), j indexes frequency, $\omega_j = j/m$, for $j = 1, 2, \ldots, [m/2]$, and k indexes the location, or time shift, of the transform. In this case, the transforms are based on tapered cosines and sines that have been zeroed out over various regions in time. The key point here is that the transforms are based on *local* sinusoids. Figure 4.19 shows an example of four local, tapered cosine functions at various frequencies. In that figure, the length of the data is considered to be one, and the cosines are localized to a fourth of the data length.

In addition to dynamic Fourier analysis as a method to overcome the restriction of stationarity, researchers have sought various alternative methods. A recent, and successful, alternative is wavelet analysis. The website

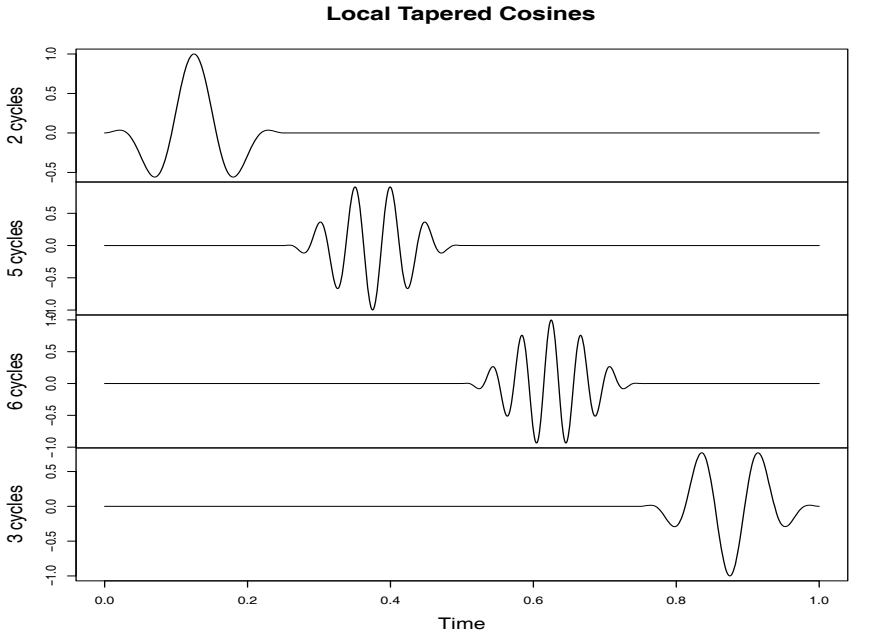

Fig. 4.19. Local, tapered cosines at various frequencies.

`http://www.wavelet.org` is devoted to wavelets, which includes information about books, technical papers, software, and links to other sites. In addition, we mention the monograph on wavelets by Daubechies (1992), the text by Percival and Walden (2000), and we note that many statistical software manufacturers have wavelet modules that sit on top of their base package. In this section, we rely primarily on the S-PLUS wavelets module (with a manual written by Bruce and Gao, 1996), however, we will present some R code where possible. The basic idea of wavelet analysis is to imitate dynamic Fourier analysis, but with functions (wavelets) that may be better suited to capture the local behavior of nonstationary time series.

Wavelets come in families generated by a father wavelet, ϕ, and a mother wavelet, ψ. The father wavelets are used to capture the smooth, low-frequency nature of the data, whereas the mother wavelets are used to capture the detailed, and high-frequency nature of the data. The father wavelet integrates to one, and the mother wavelet integrates to zero

$$\int \phi(t)dt = 1 \quad \text{and} \quad \int \psi(t)dt = 0. \tag{4.114}$$

For a simple example, consider the Haar function,

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2, \\ -1, & 1/2 \le t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.115}$$

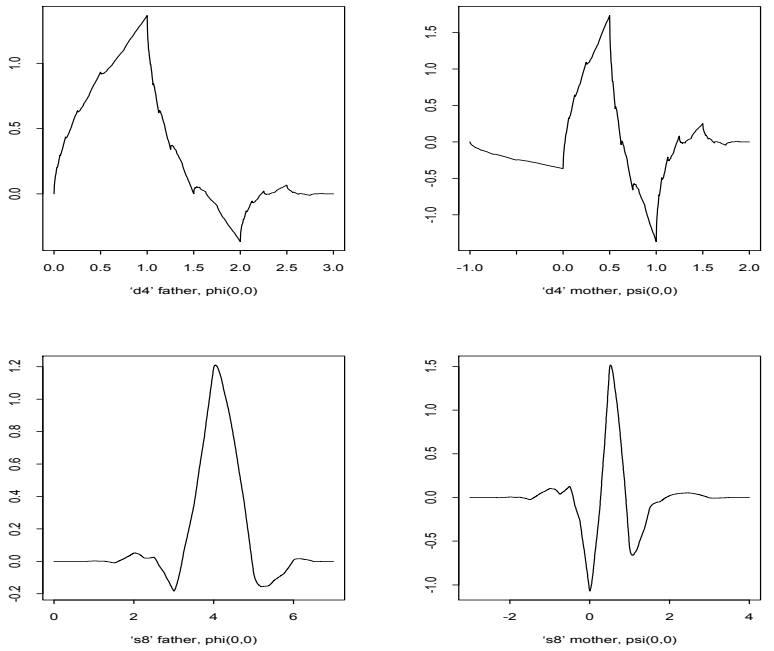

Fig. 4.20. Father and mother daublet4 wavelets (top row); father and mother symmlet8 wavelets (bottom row).

The father in this case is $\phi(t) = 1$ for $t \in [0, 1)$ and zero otherwise. The Haar functions are useful for demonstrating properties of wavelets, but they do not have good time-frequency localization properties. Figure 4.20 displays two of the more commonly used wavelets that are available with the S-PLUS wavelets module, the *daublet4* and *symmlet8* wavelets, which are described in detail in Daubechies (1992). The number after the name refers to the width and smoothness of the wavelet; for example, the symmlet10 wavelet is wider and smoother than the symmlet8 wavelet. Daublets are one of the first type of continuous orthogonal wavelets with compact support, and symmlets were constructed to be closer to symmetry than daublets. In general, wavelets do not have an analytical form, but instead they are generated using numerical methods.

Figure 4.20 was generated in S-PLUS using the wavelet module as follows:[18]

```
1 d4f <- wavelet("d4", mother=F)
```

[18] At this time, the R packages available for wavelet analysis are not extensive enough for our purposes, hence we will rely on S-PLUS for some of the demonstrations. We will provide R code when possible, and that will be based on the `wavethresh` package (version 4.2-1) that accompanies Nason (2008).

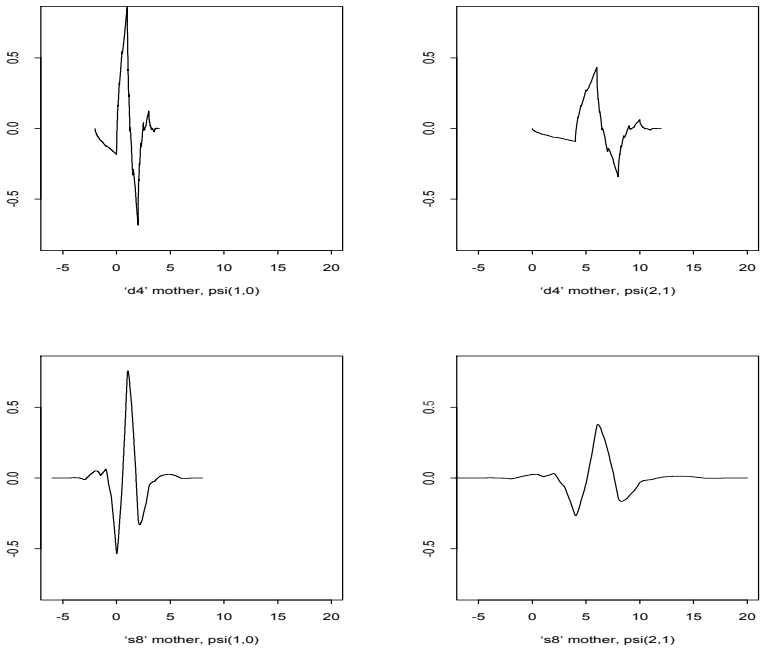

Fig. 4.21. Scaled and translated daublet4 wavelets, $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ (top row); scaled and translated symmlet8 wavelets, $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ (bottom row).

```
d4m <- wavelet("d4")
s8f <- wavelet("s8", mother=F)
s8m <- wavelet("s8")
par(mfrow=c(2,2))
plot(d4f);  plot(d4m)
plot(s8f);  plot(s8m)
```

It is possible to draw some wavelets in R using the `wavethresh` package. In that package, daublets are called DaubExPhase and symmlets are called DaubLeAsymm. The following R session displays some of the available wavelets (this will produce a figure similar to Figure 4.20) and it assumes the `wavethresh` package has been downloaded and installed (see Appendix R, §R.2, for details on installing packages). The `filter.number` determines the width and smoothness of the wavelet.

```
library(wavethresh)
par(mfrow=c(2,2))
draw(filter.number=4, family="DaubExPhase", enhance=FALSE, main="")
draw(filter.number=8, family="DaubExPhase", enhance=FALSE, main="")
draw(filter.number=4, family="DaubLeAsymm", enhance=FALSE, main="")
draw(filter.number=8, family="DaubLeAsymm", enhance=FALSE, main="")
```

When we depart from periodic functions, such as sines and cosines, the precise meaning of frequency, or cycles per unit time, is lost. When using wavelets, we typically refer to scale rather than frequency. The orthogonal wavelet decomposition of a time series, x_t, for $t = 1, \ldots, n$ is

$$\begin{aligned} x_t = &\sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) \\ &+ \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \cdots + \sum_k d_{1,k}\psi_{1,k}(t), \end{aligned} \tag{4.116}$$

where J is the number of scales, and k ranges from one to the number of coefficients associated with the specified component (see Example 4.22). In (4.116), the wavelet functions $\phi_{J,k}(t), \psi_{J,k}(t), \psi_{J-1,k}(t), \ldots, \psi_{1,k}(t)$ are generated from the father wavelet, $\phi(t)$, and the mother wavelet, $\psi(t)$, by translation (shift) and scaling:

$$\phi_{J,k}(t) = 2^{-J/2}\phi\left(\frac{t - 2^J k}{2^J}\right), \tag{4.117}$$

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(\frac{t - 2^j k}{2^j}\right), \quad j = 1, \ldots, J. \tag{4.118}$$

The choice of dyadic shifts and scales is arbitrary but convenient. The shift or translation parameter is $2^j k$, and scale parameter is 2^j. The wavelet functions are spread out and shorter for larger values of j (or scale parameter 2^j) and tall and narrow for small values of the scale. Figure 4.21 shows $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ generated from the daublet4 (top row), and the symmlet8 (bottom row) mother wavelets. We may think of $1/2^j$ (or 1/scale) in wavelet analysis as being the analogue of frequency ($\omega_j = j/n$) in Fourier analysis. For example, when $j = 1$, the scale parameter of 2 is akin to the Nyquist frequency of 1/2, and when $j = 6$, the scale parameter of 2^6 is akin to a low frequency ($1/2^6 \approx 0.016$). In other words, larger values of the scale refer to slower, smoother (or coarser) movements of the signal, and smaller values of the scale refer to faster, choppier (or finer) movements of the signal. Figure 4.21 was generated in S-PLUS using the wavelet module as follows:

```
d4.1 <- wavelet("d4", level=1, shift=0)
d4.2 <- wavelet("d4", level=2, shift=1)
s8.1 <- wavelet("s8", level=1, shift=0)
s8.2 <- wavelet("s8", level=2, shift=1)
par(mfrow=c(2,2))
plot(d4.1, ylim=c(-.8,.8), xlim=c(-6,20))
plot(d4.2, ylim=c(-.8,.8), xlim=c(-6,20))
plot(s8.1, ylim=c(-.8,.8), xlim=c(-6,20))
plot(s8.2, ylim=c(-.8,.8), xlim=c(-6,20))
```

The discrete wavelet transform (DWT) of the data x_t are the coefficients $s_{J,k}$ and $d_{j,k}$ for $j = J, J-1, \ldots, 1$, in (4.116). To some degree of approxima-

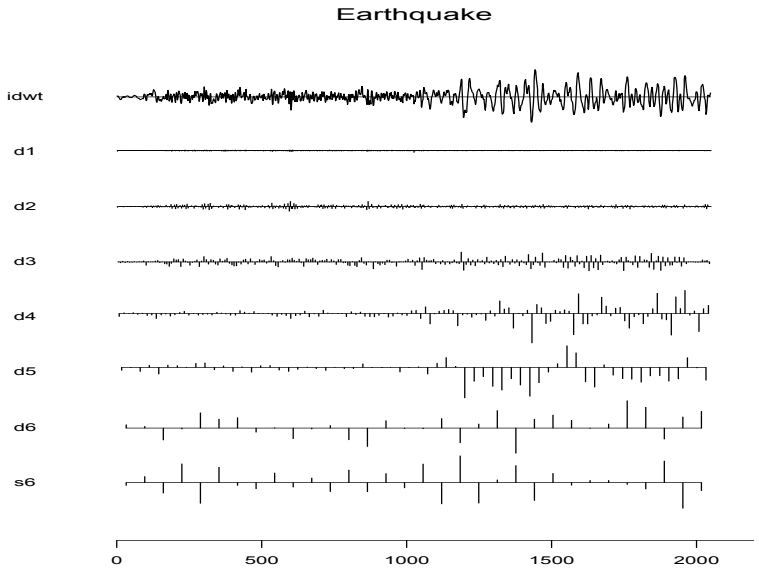

Fig. 4.22. Discrete wavelet transform of the earthquake series using the symmlet8 wavelets, and $J = 6$ levels of scale.

tion, they are given by[19]

$$s_{J,k} = n^{-1/2} \sum_{t=1}^{n} x_t \phi_{J,k}(t), \tag{4.119}$$

$$d_{j,k} = n^{-1/2} \sum_{t=1}^{n} x_t \psi_{j,k}(t) \quad j = J, J-1, \ldots, 1. \tag{4.120}$$

It is the magnitudes of the coefficients that measure the importance of the corresponding wavelet term in describing the behavior of x_t. As in Fourier analysis, the DWT is not computed as shown but is calculated using a fast algorithm. The $s_{J,k}$ are called the smooth coefficients because they represent the smooth behavior of the data. The $d_{j,k}$ are called the detail coefficients because they tend to represent the finer, more high-frequency nature, of the data.

Example 4.22 Wavelet Analysis of Earthquake and Explosion

Figures 4.22 and 4.23 show the DWTs, based on the symmlet8 wavelet basis, for the earthquake and explosion series, respectively. Each series is of

[19] The actual DWT coefficients are defined via a set of filters whose coefficients are close to what you would get by sampling the father and mother wavelets, but not exactly so; see the discussion surrounding Figures 471 and 478 in Percival and Walden (2000).

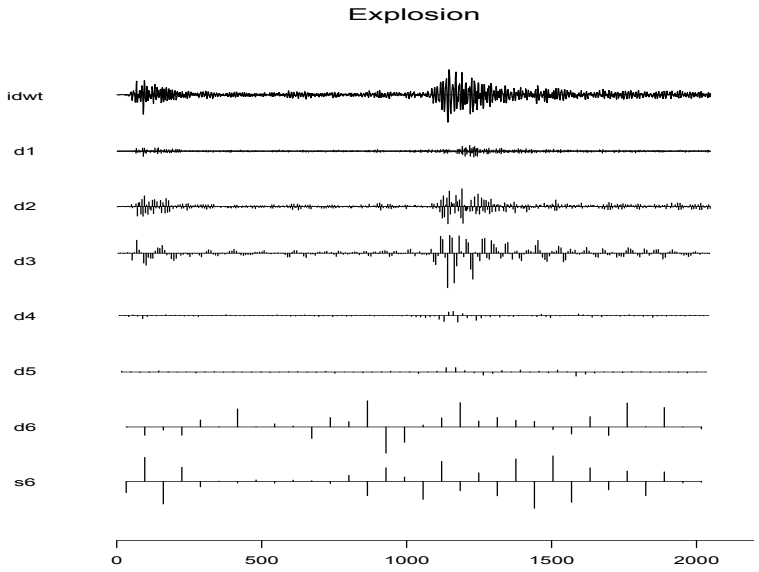

Fig. 4.23. Discrete wavelet transform of the explosion series using the symmlet8 wavelets and $J = 6$ levels of scale.

length $n = 2^{11} = 2048$, and in this example, the DWTs are calculated using $J = 6$ levels. In this case, $n/2 = 2^{10} = 1024$ values are in $d1 = \{d_{1,k};\ k = 1, \ldots, 2^{10}\}$, $n/2^2 = 2^9 = 512$ values are in $d2 = \{d_{2,k};\ k = 1, \ldots, 2^9\}$, and so on, until finally, $n/2^6 = 2^5 = 32$ values are in $d6$ and in $s6$. The detail values $d_{1,k}, \ldots, d_{6,k}$ are plotted at the same scale, and hence, the relative importance of each value can be seen from the graph. The smooth values $s_{6,k}$ are typically larger than the detail values and plotted on a different scale. The top of Figures 4.22 and 4.23 show the inverse DWT (IDWT) computed from all of the coefficients. The displayed IDWT is a reconstruction of the data, and it reproduces the data except for round-off error.

Comparing the DWTs, the earthquake is best represented by wavelets with larger scale than the explosion. One way to measure the importance of each level, $d1, d2, \ldots, d6, s6$, is to evaluate the proportion of the total power (or energy) explained by each. The total power of a time series x_t, for $t = 1, \ldots, n$, is $TP = \sum_{t=1}^{n} x_t^2$. The total power associated with each level of scale is (recall $n = 2^{11}$),

$$TP_6^s = \sum_{k=1}^{n/2^6} s_{6,k}^2 \quad \text{and} \quad TP_j^d = \sum_{k=1}^{n/2^j} d_{j,k}^2, \quad j = 1, \ldots, 6.$$

Because we are working with an orthogonal basis, we have

Table 4.2. Fraction of Total Power

Component	Earthquake	Explosion
s6	0.009	0.002
d6	0.043	0.002
d5	0.377	0.007
d4	0.367	0.015
d3	0.160	0.559
d2	0.040	0.349
d1	0.003	0.066

$$TP = TP_6^s + \sum_{j=1}^{6} TP_j^d,$$

and the proportion of the total power explained by each level of detail would be the ratios TP_j^d/TP for $j = 1, \ldots, 6$, and for the smooth level, it would be TP_6^s/TP. These values are listed in Table 4.2. From that table nearly 80% of the total power of the earthquake series is explained by the higher scale details $d4$ and $d5$, whereas 90% of the total power is explained by the smaller scale details $d2$ and $d3$ for the explosion.

Figures 4.24 and 4.25 show the time-scale plots (or scalograms) based on the DWT of the earthquake series and the explosion series, respectively. These figures are the wavelet analog of the time-frequency plots shown in Figures 4.17 and 4.18. The power axis represents the magnitude of each value d_{jk} or $s_{6,k}$. The time axis matches the time axis in the DWTs shown in Figures 4.22 and 4.23, and the scale axis is plotted as 1/scale, listed from the coarsest scale to the finest scale. On the 1/scale axis, the coarsest scale values, represented by the smooth coefficients $s6$, are plotted over the range $[0, 2^{-6})$, the coarsest detail values, $d6$, are plotted over $[2^{-6}, 2^{-5})$, and so on. In these figures, we did not plot the finest scale values, $d1$, so the finest scale values exhibited in Figures 4.24 and 4.25 are in $d2$, which are plotted over the range $[2^{-2}, 2^{-1})$.

The conclusions drawn from these plots are the same as those drawn from Figures 4.17 and 4.18. That is, the S wave for the earthquake shows power at the high scales (or low 1/scale) only, and the power remains strong for a long time. In contrast, the explosion shows power at smaller scales (or higher 1/scale) than the earthquake, and the power of the signals (P and S waves) do not last as long as in the case of the earthquake.

Assuming the data files EQ5 and EXP6 have been read into S-PLUS, the analyses of this example can performed using the S-PLUS wavelets module (which must be loaded prior to the analyses) as follows:

```
1 eq <- scale(EQ5)
2 ex <- scale(EXP6)
3 eq.dwt <- dwt(eq)
```

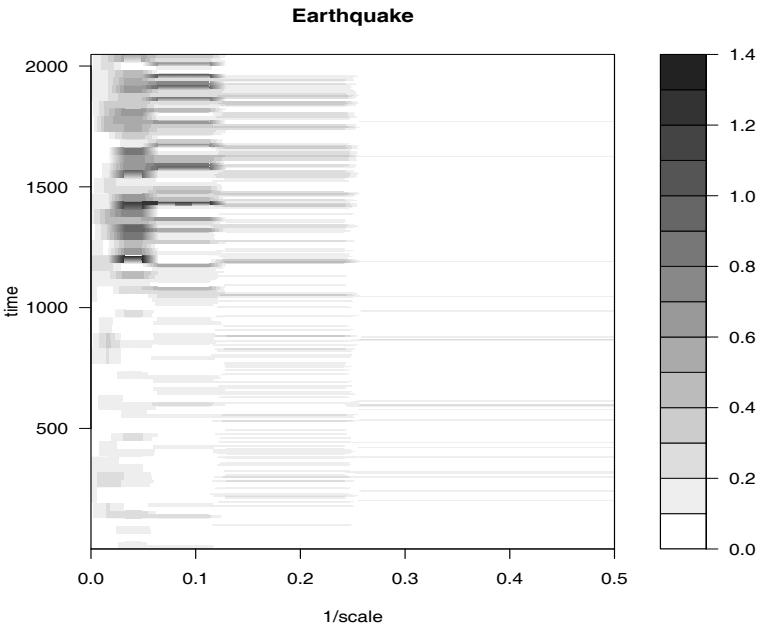

Fig. 4.24. Time-scale image (scalogram) of the earthquake series.

```
ex.dwt <- dwt(ex)
plot(eq.dwt)
plot(ex.dwt)
# energy distributions (Table 4.2)
dotchart(eq.dwt)  # a graphic
summary(eq.dwt)  # numerical details
dotchart(ex.dwt)
summary(ex.dwt)
# time scale plots
time.scale.plot(eq.dwt)
time.scale.plot(ex.dwt)
```

Similar analyses may be performed in R using the `wavelets`, `wavethresh`, or `waveslim` packages. We exhibit the analysis for the earthquake series using `wavesthresh`, assuming it has been downloaded and installed.[20]

```
library(wavethresh)
eq = scale(EQ5)  # standardize the series
ex = scale(EXP6)
eq.dwt = wd(eq, filter.number=8)
ex.dwt = wd(ex, filter.number=8)
```

[20] In `wavethresh`, the transforms are denoted by the resolution rather than the scale. If the series is of length $n = 2^p$, then resolution $p - i$ corresponds to level i for $i = 1, \ldots, p$.

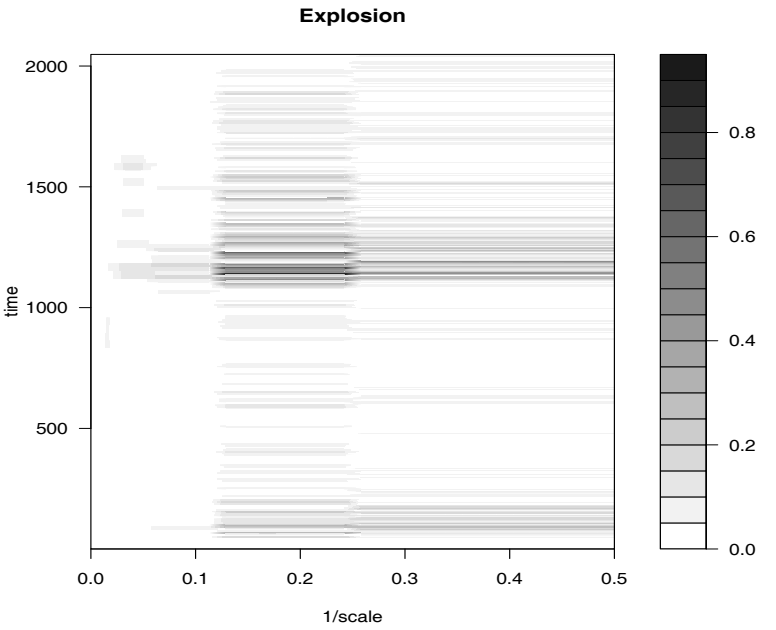

Fig. 4.25. Time-scale image (scalogram) of the explosion series.

```
# plot the wavelet transforms
par(mfrow = c(1,2))
plot(eq.dwt, main="Earthquake")
plot(ex.dwt, main="Explosion")
# total power
TPe = rep(NA,11)  # for the earthquake series
for (i in 0:10){TPe[i+1] = sum(accessD(eq.dwt, level=i)^2)}
TotEq = sum(TPe)  # check with sum(eq^2)
TPx = rep(NA,11)  # for the explosion series
for (i in 0:10){TPx[i+1] = sum(accessD(ex.dwt, level=i)^2)}
TotEx = sum(TPx)  # check with sum(ex^2)
# make a nice table
Power = round(cbind(11:1, 100*TPe/TotEq, 100*TPx/TotEx), digits=3)
colnames(Power) = c("Level", "EQ(%)", "EXP(%)")
Power
```

Wavelets can be used to perform nonparametric smoothing along the lines first discussed in §2.4, but with an emphasis on localized behavior. Although a considerable amount of literature exists on this topic, we will present the basic ideas. For further information, we refer the reader to Donoho and Johnstone (1994, 1995). As in §2.4, we suppose the data x_t can be written in terms of a signal plus noise model as

$$x_t = s_t + \epsilon_t. \tag{4.121}$$

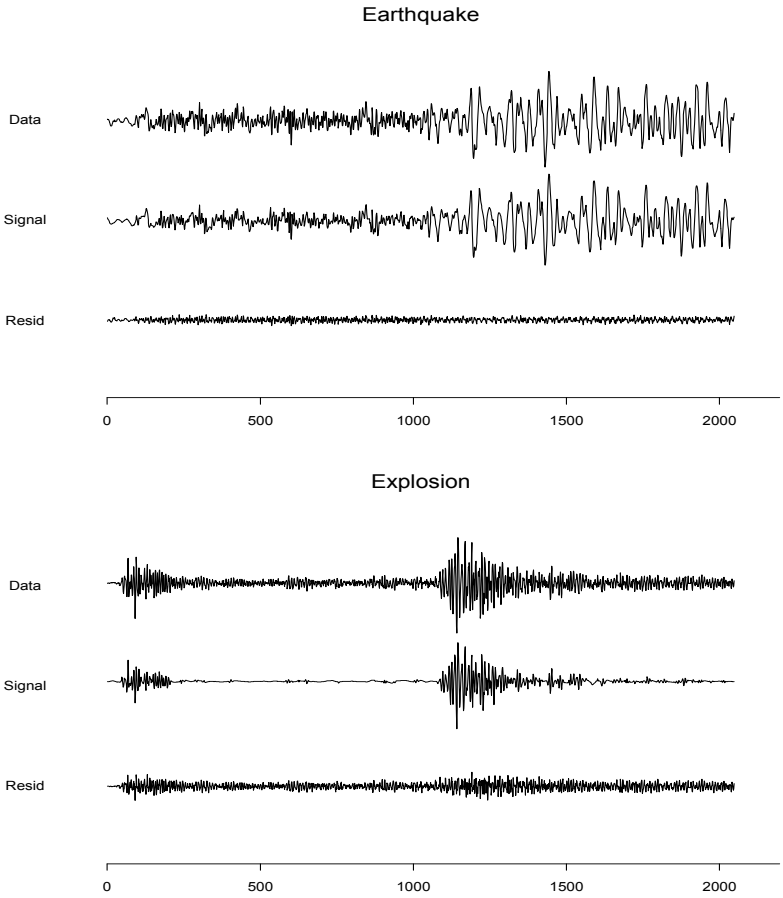

Fig. 4.26. Waveshrink estimates of the earthquake and explosion signals.

The goal here is to remove the noise from the data, and obtain an estimate of the signal, s_t, without having to specify a parametric form of the signal. The technique based on wavelets is referred to as waveshrink.

The basic idea behind waveshrink is to shrink the wavelet coefficients in the DWT of x_t toward zero in an attempt to denoise the data and then to estimate the signal via (4.116) with the new coefficients. One obvious way to shrink the coefficients toward zero is to simply zero out any coefficient smaller in magnitude than some predetermined value, λ. Such a shrinkage rule is discontinuous and sometimes it is preferable to use a continuous shrinkage function. One such method, termed soft shrinkage, proceeds as follows. If the value of a coefficient is a, we set that coefficient to zero if $|a| \leq \lambda$, and to $\text{sign}(a)(|a| - \lambda)$ if $|a| > \lambda$. The choice of a shrinkage method is based on

the goal of the signal extraction. This process entails choosing a value for the shrinkage threshold, λ, and we may wish to use a different threshold value, say, λ_j, for each level of scale $j = 1, \ldots, J$. One particular method that works well if we are interested in a relatively high degree of smoothness in the estimate is to choose $\lambda = \widehat{\sigma}_\epsilon \sqrt{2 \log n}$ for all scale levels, where $\widehat{\sigma}_\epsilon$ is an estimate of the scale of the noise, σ_ϵ. Typically a robust estimate of σ_ϵ is used, e.g., the median of the absolute deviations of the data from the median (MAD). For other thresholding techniques or for a better understanding of waveshrink, see Donoho and Johnstone (1994, 1995), or the S-PLUS wavelets module manual (Bruce and Gao, 1996, Ch 6).

Example 4.23 Waveshrink Analysis of Earthquake and Explosion

Figure 4.26 shows the results of a waveshrink analysis on the earthquake and explosion series. In this example, soft shrinkage was used with a universal threshold of $\lambda = \widehat{\sigma}_\epsilon \sqrt{2 \log n}$ where $\widehat{\sigma}_\epsilon$ is the MAD. Figure 4.26 displays the data x_t, the estimated signal $\widehat{s}_t$, as well as the residuals $x_t - \widehat{s}_t$. According to this analysis, the earthquake is mostly signal and characterized by prolonged energy, whereas the explosion is comprised of short bursts of energy.

Figure 4.26 was generated in S-PLUS using the wavelets module. For example, the analysis of the earthquake series was performed as follows.

```
eq.dwt <- dwt(eq)
eq.shrink <- waveshrink(eq.dwt, shrink.rule="universal",
    shrink.fun="soft")
```

In R, using the `wavethresh` package, use the following commands for the earthquake series.

```
library(wavethresh)
eq = scale(EQ5)
par(mfrow=c(3,1))
eq.dwt = wd(eq, filter.number=8)
eq.smo = wr(threshold(eq.dwt, levels=5:10))
ts.plot(eq, main="Earthquake", ylab="Data")
ts.plot(eq.smo, ylab="Signal")
ts.plot(eq-eq.smo, ylab="Resid")
```

4.10 Lagged Regression Models

One of the intriguing possibilities offered by the coherence analysis of the relation between the SOI and Recruitment series discussed in Example 4.18 would be extending classical regression to the analysis of lagged regression models of the form

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \tag{4.122}$$

where v_t is a stationary noise process, x_t is the observed input series, and y_t is the observed output series. We are interested in estimating the filter

coefficients β_r relating the adjacent lagged values of x_t to the output series y_t.

In the case of SOI and Recruitment series, we might identify the El Niño driving series, SOI, as the input, x_t, and y_t, the Recruitment series, as the output. In general, there will be more than a single possible input series and we may envision a $q \times 1$ vector of driving series. This multivariate input situation is covered in Chapter 7. The model given by (4.122) is useful under several different scenarios, corresponding to different assumptions that can be made about the components.

We assume that the inputs and outputs have zero means and are jointly stationary with the 2×1 vector process $(x_t, y_t)'$ having a spectral matrix of the form

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}. \tag{4.123}$$

Here, $f_{xy}(\omega)$ is the cross-spectrum relating the input x_t to the output y_t, and $f_{xx}(\omega)$ and $f_{yy}(\omega)$ are the spectra of the input and output series, respectively. Generally, we observe two series, regarded as input and output and search for regression functions $\{\beta_t\}$ relating the inputs to the outputs. We assume all autocovariance functions satisfy the absolute summability conditions of the form (4.30).

Then, minimizing the mean squared error

$$MSE = E\left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\right)^2 \tag{4.124}$$

leads to the usual orthogonality conditions

$$E\left[\left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\right) x_{t-s}\right] = 0 \tag{4.125}$$

for all $s = 0, \pm 1, \pm 2, \ldots$. Taking the expectations inside leads to the normal equations

$$\sum_{r=-\infty}^{\infty} \beta_r \, \gamma_{xx}(s-r) = \gamma_{yx}(s) \tag{4.126}$$

for $s = 0, \pm 1, \pm 2, \ldots$. These equations might be solved, with some effort, if the covariance functions were known exactly. If data (x_t, y_t) for $t = 1, ..., n$ are available, we might use a finite approximation to the above equations with $\widehat{\gamma}_{xx}(h)$ and $\widehat{\gamma}_{yx}(h)$ substituted into (4.126). If the regression vectors are essentially zero for $|s| \geq M/2$, and $M < n$, the system (4.126) would be of full rank and the solution would involve inverting an $(M-1) \times (M-1)$ matrix.

A frequency domain approximate solution is easier in this case for two reasons. First, the computations depend on spectra and cross-spectra that can be estimated from sample data using the techniques of §4.6. In addition, no matrices will have to be inverted, although the frequency domain ratio will

have to be computed for each frequency. In order to develop the frequency domain solution, substitute the representation (4.89) into the normal equations, using the convention defined in (4.123). The left side of (4.126) can then be written in the form

$$\int_{-1/2}^{1/2} \sum_{r=-\infty}^{\infty} \beta_r \, \mathrm{e}^{2\pi i\omega(s-r)} \, f_{xx}(\omega) \, d\omega = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega s} B(\omega) f_{xx}(\omega) \, d\omega,$$

where

$$B(\omega) = \sum_{r=-\infty}^{\infty} \beta_r \, \mathrm{e}^{-2\pi i\omega r} \tag{4.127}$$

is the Fourier transform of the regression coefficients β_t. Now, because $\gamma_{yx}(s)$ is the inverse transform of the cross-spectrum $f_{yx}(\omega)$, we might write the system of equations in the frequency domain, using the uniqueness of the Fourier transform, as

$$B(\omega) f_{xx}(\omega) = f_{yx}(\omega), \tag{4.128}$$

which then become the analogs of the usual normal equations. Then, we may take

$$\widehat{B}(\omega_k) = \frac{\widehat{f}_{yx}(\omega_k)}{\widehat{f}_{xx}(\omega_k)} \tag{4.129}$$

as the estimator for the Fourier transform of the regression coefficients, evaluated at some subset of fundamental frequencies $\omega_k = k/M$ with $M << n$. Generally, we assume smoothness of $B(\cdot)$ over intervals of the form $\{\omega_k + \ell/n;\ \ell = -(L-1)/2, \ldots, (L-1)/2\}$. The inverse transform of the function $\widehat{B}(\omega)$ would give $\widehat{\beta}_t$, and we note that the discrete time approximation can be taken as

$$\widehat{\beta}_t = M^{-1} \sum_{k=0}^{M-1} \widehat{B}(\omega_k) \mathrm{e}^{2\pi i\omega_k t} \tag{4.130}$$

for $t = 0, \pm 1, \pm 2, \ldots, \pm(M/2 - 1)$. If we were to use (4.130) to define $\widehat{\beta}_t$ for $|t| \geq M/2$, we would end up with a sequence of coefficients that is periodic with a period of M. In practice we define $\widehat{\beta}_t = 0$ for $|t| \geq M/2$ instead. Problem 4.32 explores the error resulting from this approximation.

Example 4.24 Lagged Regression for SOI and Recruitment

The high coherence between the SOI and Recruitment series noted in Example 4.18 suggests a lagged regression relation between the two series. A natural direction for the implication in this situation is implied because we feel that the sea surface temperature or SOI should be the input and the Recruitment series should be the output. With this in mind, let x_t be the SOI series and y_t the Recruitment series.

Although we think naturally of the SOI as the input and the Recruitment as the output, two input-output configurations are of interest. With SOI as the input, the model is

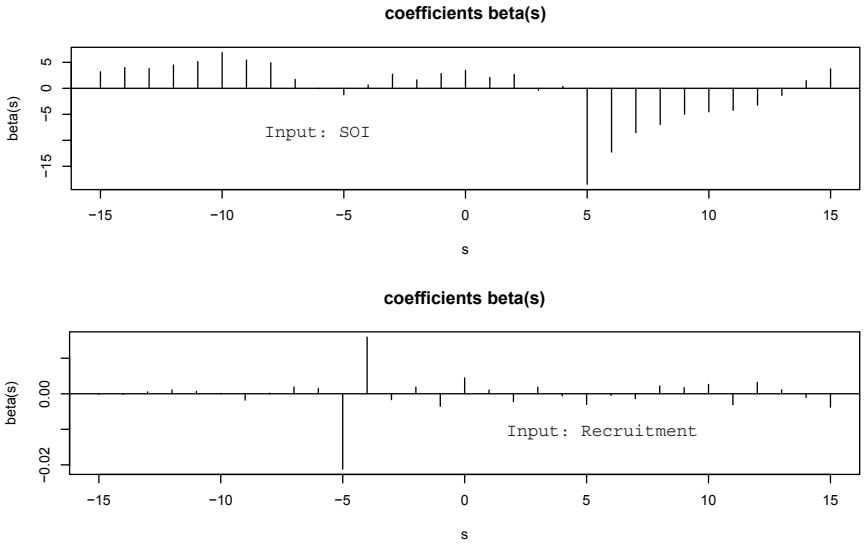

Fig. 4.27. Estimated impulse response functions relating SOI to Recruitment (top) and Recruitment to SOI (bottom) $L = 15, M = 32$.

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r} + w_t$$

whereas a model that reverses the two roles would be

$$x_t = \sum_{r=-\infty}^{\infty} b_r y_{t-r} + v_t,$$

where w_t and v_t are white noise processes. Even though there is no plausible environmental explanation for the second of these two models, displaying both possibilities helps to settle on a parsimonious transfer function model.

Based on the script `LagReg` (see Appendix R, §R.1), the estimated regression or impulse response function for SOI, with $M = 32$ and $L = 15$ is

```
1 LagReg(soi, rec, L=15, M=32, threshold=6)

           lag s    beta(s)
     [1,]      5 -18.479306
     [2,]      6 -12.263296
     [3,]      7  -8.539368
     [4,]      8  -6.984553
The prediction equation is
rec(t) = alpha + sum_s[ beta(s)*soi(t-s) ], where alpha = 65.97
MSE = 414.08
```

Note the negative peak at a lag of five points in the top of Figure 4.27; in this case, SOI is the input series. The fall-off after lag five seems to be

approximately exponential and a possible model is

$$y_t = 66 - 18.5x_{t-5} - 12.3x_{t-6} - 8.5x_{t-7} - 7x_{t-8} + w_t.$$

If we examine the inverse relation, namely, a regression model with the Recruitment series y_t as the input, the bottom of Figure 4.27 implies a much simpler model,

```
2 LagReg(rec, soi, L=15, M=32, inverse=TRUE, threshold=.01)

           lag s     beta(s)
     [1,]     4  0.01593167
     [2,]     5 -0.02120013
The prediction equation is
soi(t) = alpha + sum_s[ beta(s)*rec(t+s) ], where alpha = 0.41
MSE = 0.07
```

depending on only two coefficients, namely,

$$x_t = .41 + .016y_{t+4} - .02y_{t+5} + v_t.$$

Multiplying both sides by $50B^5$ and rearranging, we have

$$(1 - .8B)y_t = 20.5 - 50B^5x_t + \epsilon_t,$$

where ϵ_t is white noise, as our final, parsimonious model.

The example shows we can get a clean estimator for the transfer functions relating the two series if the coherence $\widehat{\rho}^2_{xy}(\omega)$ is large. The reason is that we can write the minimized mean squared error (4.124) as

$$MSE = E\Big[\big(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\big)y_t\Big] = \gamma_{yy}(0) - \sum_{r=-\infty}^{\infty} \beta_r \gamma_{xy}(-r),$$

using the result about the orthogonality of the data and error term in the Projection theorem. Then, substituting the spectral representations of the autocovariance and cross-covariance functions and identifying the Fourier transform (4.127) in the result leads to

$$\begin{aligned} MSE &= \int_{-1/2}^{1/2} [f_{yy}(\omega) - B(\omega)f_{xy}(\omega)]\, d\omega \\ &= \int_{-1/2}^{1/2} f_{yy}(\omega)[1 - \rho^2_{yx}(\omega)]d\omega, \end{aligned} \tag{4.131}$$

where $\rho^2_{yx}(\omega)$ is just the squared coherence given by (4.87). The similarity of (4.131) to the usual mean square error that results from predicting y from x is obvious. In that case, we would have

$$E(y - \beta x)^2 = \sigma^2_y(1 - \rho^2_{xy})$$

for jointly distributed random variables x and y with zero means, variances σ_x^2 and σ_y^2, and covariance $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$. Because the mean squared error in (4.131) satisfies $MSE \geq 0$ with $f_{yy}(\omega)$ a non-negative function, it follows that the coherence satisfies

$$0 \leq \rho_{xy}^2(\omega) \leq 1$$

for all ω. Furthermore, Problem 4.33 shows the squared coherence is one when the output are linearly related by the filter relation (4.122), and there is no noise, i.e., $v_t = 0$. Hence, the multiple coherence gives a measure of the association or correlation between the input and output series as a function of frequency.

The matter of verifying that the F-distribution claimed for (4.97) will hold when the sample coherence values are substituted for theoretical values still remains. Again, the form of the F-statistic is exactly analogous to the usual t-test for no correlation in a regression context. We give an argument leading to this conclusion later using the results in Appendix C, §C.3. Another question that has not been resolved in this section is the extension to the case of multiple inputs $x_{t1}, x_{t2}, \ldots, x_{tq}$. Often, more than just a single input series is present that can possibly form a lagged predictor of the output series y_t. An example is the cardiovascular mortality series that depended on possibly a number of pollution series and temperature. We discuss this particular extension as a part of the multivariate time series techniques considered in Chapter 7.

4.11 Signal Extraction and Optimum Filtering

A model closely related to regression can be developed by assuming again that

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \tag{4.132}$$

but where the βs are known and x_t is some unknown random signal that is uncorrelated with the noise process v_t. In this case, we observe only y_t and are interested in an estimator for the signal x_t of the form

$$\widehat{x}_t = \sum_{r=-\infty}^{\infty} a_r y_{t-r}. \tag{4.133}$$

In the frequency domain, it is convenient to make the additional assumptions that the series x_t and v_t are both mean-zero stationary series with spectra $f_{xx}(\omega)$ and $f_{vv}(\omega)$, often referred to as the signal spectrum and noise spectrum, respectively. Often, the special case $\beta_t = \delta_t$, in which δ_t is the Kronecker delta, is of interest because (4.132) reduces to the simple signal plus noise model

$$y_t = x_t + v_t \tag{4.134}$$

in that case. In general, we seek the set of filter coefficients a_t that minimize the mean squared error of estimation, say,

$$MSE = E\left[\left(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r}\right)^2\right]. \tag{4.135}$$

This problem was originally solved by Kolmogorov (1941) and by Wiener (1949), who derived the result in 1941 and published it in classified reports during World War II.

We can apply the orthogonality principle to write

$$E\left[\left(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r}\right) y_{t-s}\right] = 0$$

for $s = 0, \pm 1, \pm 2, \ldots$, which leads to

$$\sum_{r=-\infty}^{\infty} a_r \gamma_{yy}(s-r) = \gamma_{xy}(s),$$

to be solved for the filter coefficients. Substituting the spectral representations for the autocovariance functions into the above and identifying the spectral densities through the uniqueness of the Fourier transform produces

$$A(\omega) f_{yy}(\omega) = f_{xy}(\omega), \tag{4.136}$$

where $A(\omega)$ and the optimal filter a_t are Fourier transform pairs for $B(\omega)$ and β_t. Now, a special consequence of the model is that (see Problem 4.23)

$$f_{xy}(\omega) = \overline{B(\omega)} f_{xx}(\omega) \tag{4.137}$$

and

$$f_{yy}(\omega) = |B(\omega)|^2 f_{xx}(\omega) + f_{vv}(\omega), \tag{4.138}$$

implying the optimal filter would be Fourier transform of

$$A(\omega) = \frac{\overline{B(\omega)}}{\left(|B(\omega)|^2 + \frac{f_{vv}(\omega)}{f_{xx}(\omega)}\right)}, \tag{4.139}$$

where the second term in the denominator is just the inverse of the signal to noise ratio, say,

$$\mathrm{SNR}(\omega) = \frac{f_{xx}(\omega)}{f_{vv}(\omega)}. \tag{4.140}$$

The result shows the optimum filters can be computed for this model if the signal and noise spectra are both known or if we can assume knowledge

of the signal-to-noise ratio SNR(ω) as function of frequency. In Chapter 7, we show some methods for estimating these two parameters in conjunction with random effects analysis of variance models, but we assume here that it is possible to specify the signal-to-noise ratio a priori. If the signal-to-noise ratio is known, the optimal filter can be computed by the inverse transform of the function $A(\omega)$. It is more likely that the inverse transform will be intractable and a finite filter approximation like that used in the previous section can be applied to the data. In this case, we will have

$$a_t^M = M^{-1} \sum_{k=0}^{M-1} A(\omega_k) \mathrm{e}^{2\pi i \omega_k t} \tag{4.141}$$

as the estimated filter function. It will often be the case that the form of the specified frequency response will have some rather sharp transitions between regions where the signal-to-noise ratio is high and regions where there is little signal. In these cases, the shape of the frequency response function will have ripples that can introduce frequencies at different amplitudes. An aesthetic solution to this problem is to introduce tapering as was done with spectral estimation in (4.61)-(4.68). We use below the tapered filter $\tilde{a}_t = h_t a_t$ where h_t is the cosine taper given in (4.68). The squared frequency response of the resulting filter will be $|\tilde{A}(\omega)|^2$, where

$$\tilde{A}(\omega) = \sum_{t=-\infty}^{\infty} a_t h_t \mathrm{e}^{-2\pi i \omega t}. \tag{4.142}$$

The results are illustrated in the following example that extracts the El Niño component of the sea surface temperature series.

Example 4.25 Estimating the El Niño Signal via Optimal Filters

Figure 4.5 shows the spectrum of the SOI series, and we note that essentially two components have power, the El Niño frequency of about .02 cycles per month (the four-year cycle) and a yearly frequency of about .08 cycles per month (the annual cycle). We assume, for this example, that we wish to preserve the lower frequency as signal and to eliminate the higher order frequencies, and in particular, the annual cycle. In this case, we assume the simple signal plus noise model

$$y_t = x_t + v_t,$$

so that there is no convolving function β_t. Furthermore, the signal-to-noise ratio is assumed to be high to about .06 cycles per month and zero thereafter. The optimal frequency response was assumed to be unity to .05 cycles per point and then to decay linearly to zero in several steps. Figure 4.28 shows the coefficients as specified by (4.141) with $M = 64$, as well as the frequency response function given by (4.142), of the cosine tapered coefficients; recall Figure 4.9, where we demonstrated the need for tapering to avoid severe

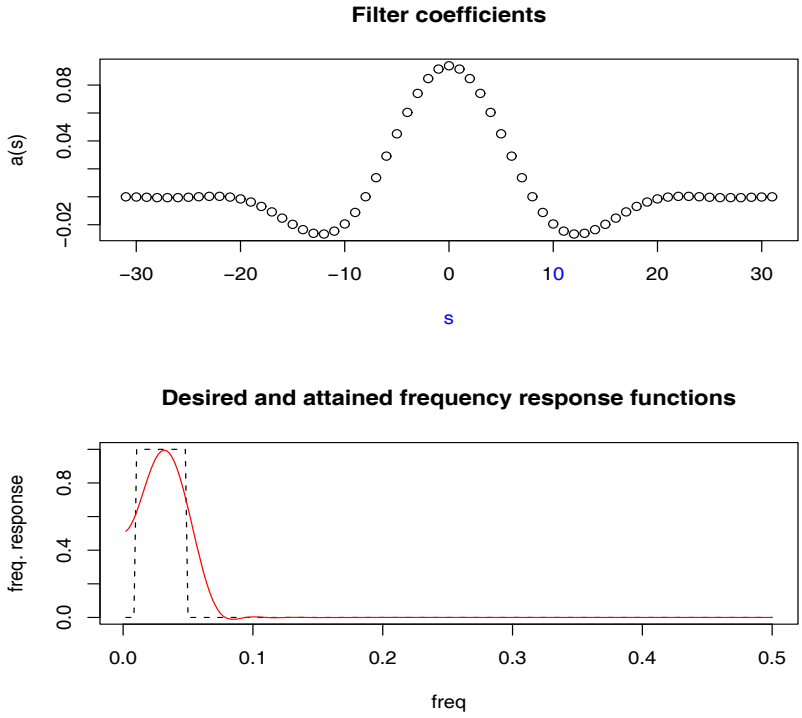

Fig. 4.28. Filter coefficients (top) and frequency response functions (bottom) for designed SOI filters.

ripples in the window. The constructed response function is compared to the ideal window in Figure 4.28.

Figure 4.29 shows the original and filtered SOI index, and we see a smooth extracted signal that conveys the essence of the underlying El Niño signal. The frequency response of the designed filter can be compared with that of the symmetric 12-month moving average applied to the same series in Example 4.19. The filtered series, shown in Figure 4.14, shows a good deal of higher frequency chatter riding on the smoothed version, which has been introduced by the higher frequencies that leak through in the squared frequency response, as in Figure 4.16.

The analysis can be replicated using the script `SigExtract`; see Appendix R, §R.1, for details.

```
1 SigExtract(soi, L=9, M=64, max.freq=.05)
```

The design of finite filters with a specified frequency response requires some experimentation with various target frequency response functions and we have only touched on the methodology here. The filter designed here, sometimes called a low-pass filter reduces the high frequencies and keeps or passes the low frequencies. Alternately, we could design a high-pass filter to keep high

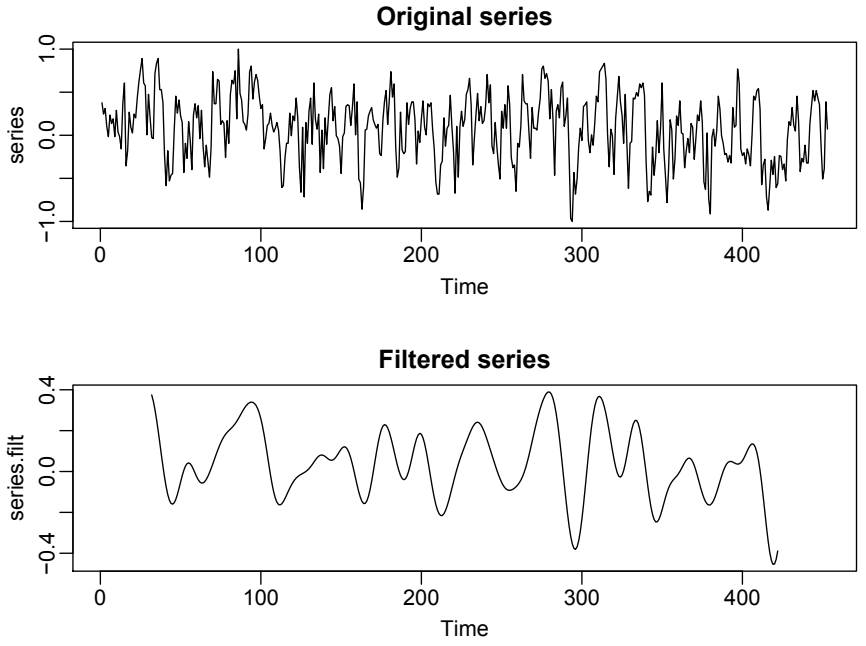

Fig. 4.29. Original SOI series (top) compared to filtered version showing the estimated El Niño temperature signal (bottom).

frequencies if that is where the signal is located. An example of a simple high-pass filter is the first difference with a frequency response that is shown in Figure 4.16. We can also design band-pass filters that keep frequencies in specified bands. For example, seasonal adjustment filters are often used in economics to reject seasonal frequencies while keeping both high frequencies, lower frequencies, and trend (see, for example, Grether and Nerlove, 1970).

The filters we have discussed here are all symmetric two-sided filters, because the designed frequency response functions were purely real. Alternatively, we may design recursive filters to produce a desired response. An example of a recursive filter is one that replaces the input x_t by the filtered output

$$y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + x_t - \sum_{k=1}^{q} \theta_k x_{t-k}. \tag{4.143}$$

Note the similarity between (4.143) and the ARIMA$(p, 1, q)$ model, in which the white noise component is replaced by the input. Transposing the terms involving y_t and using the basic linear filter result in Property 4.7 leads to

$$f_y(\omega) = \frac{|\theta(\mathrm{e}^{-2\pi i\omega})|^2}{|\phi(\mathrm{e}^{-2\pi i\omega})|^2} f_x(\omega), \tag{4.144}$$

where

$$\phi(\mathrm{e}^{-2\pi i\omega}) = 1 - \sum_{k=1}^{p} \phi_k \mathrm{e}^{-2\pi ik\omega}$$

and

$$\theta(\mathrm{e}^{-2\pi i\omega}) = 1 - \sum_{k=1}^{q} \theta_k \mathrm{e}^{-2\pi ik\omega}.$$

Recursive filters such as those given by (4.144) distort the phases of arriving frequencies, and we do not consider the problem of designing such filters in any detail.

4.12 Spectral Analysis of Multidimensional Series

Multidimensional series of the form $x_{\boldsymbol{s}}$, where $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)'$ is an r-dimensional vector of spatial coordinates or a combination of space and time coordinates, were introduced in §1.7. The example given there, shown in Figure 1.15, was a collection of temperature measurements taking on a rectangular field. These data would form a two-dimensional process, indexed by row and column in space. In that section, the multidimensional autocovariance function of an r-dimensional stationary series was given as $\gamma_x(\boldsymbol{h}) = E[x_{\boldsymbol{s}+\boldsymbol{h}} x_{\boldsymbol{s}}]$, where the multidimensional lag vector is $\boldsymbol{h} = (h_1, h_2, \ldots, h_r)'$.

The multidimensional wavenumber spectrum is given as the Fourier transform of the autocovariance, namely,

$$f_x(\boldsymbol{\omega}) = \sum_{\boldsymbol{h}} \gamma_x(\boldsymbol{h}) \mathrm{e}^{-2\pi i \boldsymbol{\omega}'\boldsymbol{h}}. \tag{4.145}$$

Again, the inverse result

$$\gamma_x(\boldsymbol{h}) = \int_{-1/2}^{1/2} f_x(\boldsymbol{\omega}) e^{2\pi i \boldsymbol{\omega}'\boldsymbol{h}} d\boldsymbol{\omega} \tag{4.146}$$

holds, where the integral is over the multidimensional range of the vector $\boldsymbol{\omega}$. The wavenumber argument is exactly analogous to the frequency argument, and we have the corresponding intuitive interpretation as the cycling rate ω_i per distance traveled s_i in the i-th direction.

Two-dimensional processes occur often in practical applications, and the representations above reduce to

$$f_x(\omega_1, \omega_2) = \sum_{h_1=-\infty}^{\infty} \sum_{h_2=-\infty}^{\infty} \gamma_x(h_1, h_2) \mathrm{e}^{-2\pi i(\omega_1 h_1 + \omega_2 h_2)} \tag{4.147}$$

and

$$\gamma_x(h_1, h_2) = \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} f_x(\omega_1, \omega_2) \mathrm{e}^{2\pi i(\omega_1 h_1 + \omega_2 h_2)} d\omega_1 \, d\omega_2 \tag{4.148}$$

in the case $r = 2$. The notion of linear filtering generalizes easily to the two-dimensional case by defining the impulse response function a_{s_1,s_2} and the spatial filter output as

$$y_{s_1,s_2} = \sum_{u_1} \sum_{u_2} a_{u_1,u_2} x_{s_1-u_1,s_2-u_2}. \tag{4.149}$$

The spectrum of the output of this filter can be derived as

$$f_y(\omega_1, \omega_2) = |A(\omega_1, \omega_2)|^2 f_x(\omega_1, \omega_2), \tag{4.150}$$

where

$$A(\omega_1, \omega_2) = \sum_{u_1} \sum_{u_2} a_{u_1,u_2} \mathrm{e}^{-2\pi i(\omega_1 u_1 + \omega_2 u_2)}. \tag{4.151}$$

These results are analogous to those in the one-dimensional case, described by Property 4.7.

The multidimensional DFT is also a straightforward generalization of the univariate expression. In the two-dimensional case with data on a rectangular grid, $\{x_{s_1,s_2};\ s_1 = 1, ..., n_1,\ s_2 = 1, ..., n_2\}$, we will write, for $-1/2 \le \omega_1, \omega_2 \le 1/2$,

$$d(\omega_1, \omega_2) = (n_1 n_2)^{-1/2} \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} x_{s_1,s_2} \mathrm{e}^{-2\pi i(\omega_1 s_1 + \omega_2 s_2)} \tag{4.152}$$

as the two-dimensional DFT, where the frequencies ω_1, ω_2 are evaluated at multiples of $(1/n_1, 1/n_2)$ on the spatial frequency scale. The two-dimensional wavenumber spectrum can be estimated by the smoothed sample wavenumber spectrum

$$\bar{f}_x(\omega_1, \omega_2) = (L_1 L_2)^{-1} \sum_{\ell_1,\ell_2} \left| d(\omega_1 + \ell_1/n_1, \omega_2 + \ell_2/n_2) \right|^2, \tag{4.153}$$

where the sum is taken over the grid $\{-m_j \le \ell_j \le m_j;\ j = 1, 2\}$, where $L_1 = 2m_1 + 1$ and $L_2 = 2m_2 + 1$. The statistic

$$\frac{2L_1 L_2 \bar{f}_x(\omega_1, \omega_2)}{f_x(\omega_1, \omega_2)} \dot{\sim} \chi^2_{2L_1 L_2} \tag{4.154}$$

can be used to set confidence intervals or make approximate tests against a fixed assumed spectrum $f_0(\omega_1, \omega_2)$. We may also extend this analysis to weighted estimation and window estimation as discussed in §4.5.

Example 4.26 Soil Surface Temperatures

As an example, consider the periodogram of the two-dimensional temperature series shown in Figure 1.15 and analyzed by Bazza et al. (1988). We recall the spatial coordinates in this case will be (s_1, s_2), which define the spatial coordinates rows and columns so that the frequencies in the two

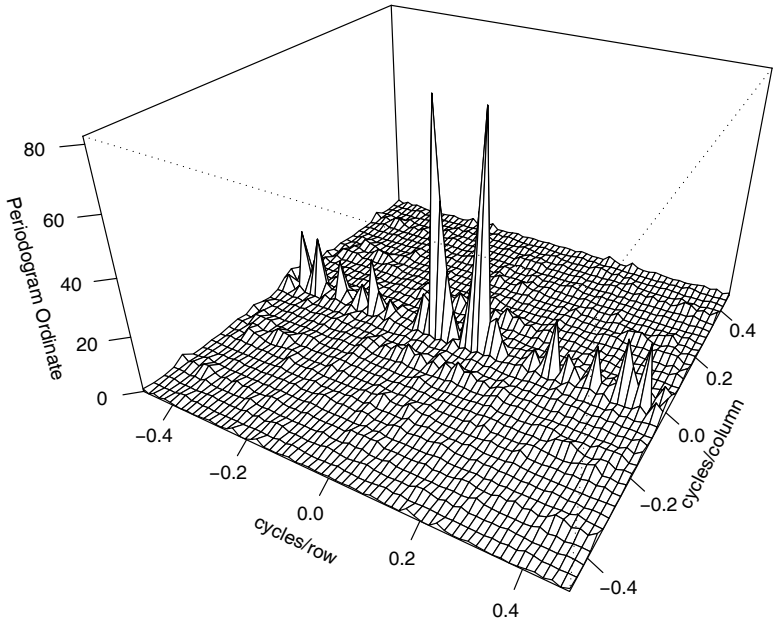

Fig. 4.30. Two-dimensional periodogram of soil temperature profile showing peak at .0625 cycles/row. The period is 16 rows, and this corresponds to 16×17 ft $= 272$ ft.

directions will be expressed as cycles per row and cycles per column. Figure 4.30 shows the periodogram of the two-dimensional temperature series, and we note the ridge of strong spectral peaks running over rows at a column frequency of zero. An obvious periodic component appears at frequencies of .0625 and $-.0625$ cycles per row, which corresponds to 16 rows or about 272 ft. On further investigation of previous irrigation patterns over this field, treatment levels of salt varied periodically over columns. This analysis is extended in Problem 4.17, where we recover the salt treatment profile over rows and compare it to a signal, computed by averaging over columns.

Figure 4.30 may be reproduced in R as follows. In the code for this example, the periodogram is computed in one step as `per`; the rest of the code is simply manipulation to obtain a nice graphic.

```
per = abs(fft(soiltemp-mean(soiltemp))/sqrt(64*36))^2
per2 = cbind(per[1:32,18:2], per[1:32,1:18])
per3 = rbind(per2[32:2,],per2)
par(mar=c(1,2.5,0,0)+.1)
persp(-31:31/64, -17:17/36, per3, phi=30, theta=30, expand=.6,
    ticktype="detailed", xlab="cycles/row", ylab="cycles/column",
    zlab="Periodogram Ordinate")
```

Another application of two-dimensional spectral analysis of agricultural field trials is given in McBratney and Webster (1981), who used it to detect ridge and furrow patterns in yields. The requirement for regular, equally spaced samples on fairly large grids has tended to limit enthusiasm for strict two-dimensional spectral analysis. An exception is when a propagating signal from a given velocity and azimuth is present so predicting the wavenumber spectrum as a function of velocity and azimuth becomes feasible (see Shumway et al., 1999).

Problems

Section 4.2

4.1 Repeat the simulations and analyses in Examples 4.1 and 4.2 with the following changes:

(a) Change the sample size to $n = 128$ and generate and plot the same series as in Example 4.1:

$$\begin{aligned} x_{t1} &= 2\cos(2\pi\,.06\,t) + 3\sin(2\pi\,.06\,t), \\ x_{t2} &= 4\cos(2\pi\,.10\,t) + 5\sin(2\pi\,.10\,t), \\ x_{t3} &= 6\cos(2\pi\,.40\,t) + 7\sin(2\pi\,.40\,t), \\ x_t &= x_{t1} + x_{t2} + x_{t3}. \end{aligned}$$

What is the major difference between these series and the series generated in Example 4.1? (Hint: The answer is *fundamental*. But if your answer is the series are longer, you may be punished severely.)

(b) As in Example 4.2, compute and plot the periodogram of the series, x_t, generated in (a) and comment.

(c) Repeat the analyses of (a) and (b) but with $n = 100$ (as in Example 4.1), and adding noise to x_t; that is

$$x_t = x_{t1} + x_{t2} + x_{t3} + w_t$$

where $w_t \sim$ iid $\mathrm{N}(0, 25)$. That is, you should simulate and plot the data, and then plot the periodogram of x_t and comment.

4.2 With reference to equations (4.1) and (4.2), let $Z_1 = U_1$ and $Z_2 = -U_2$ be independent, standard normal variables. Consider the polar coordinates of the point (Z_1, Z_2), that is,

$$A^2 = Z_1^2 + Z_2^2 \quad \text{and} \quad \phi = \tan^{-1}(Z_2/Z_1).$$

(a) Find the joint density of A^2 and ϕ, and from the result, conclude that A^2 and ϕ are independent random variables, where A^2 is a chi-squared random variable with 2 df, and ϕ is uniformly distributed on $(-\pi, \pi)$.

(b) Going in reverse from polar coordinates to rectangular coordinates, suppose we assume that A^2 and ϕ are independent random variables, where A^2 is chi-squared with 2 df, and ϕ is uniformly distributed on $(-\pi, \pi)$. With $Z_1 = A\cos(\phi)$ and $Z_2 = A\sin(\phi)$, where A is the positive square root of A^2, show that Z_1 and Z_2 are independent, standard normal random variables.

4.3 Verify (4.4).

Section 4.3

4.4 A time series was generated by first drawing the white noise series w_t from a normal distribution with mean zero and variance one. The observed series x_t was generated from

$$x_t = w_t - \theta w_{t-1}, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where θ is a parameter.

(a) Derive the theoretical mean value and autocovariance functions for the series x_t and w_t. Are the series x_t and w_t stationary? Give your reasons.
(b) Give a formula for the power spectrum of x_t, expressed in terms of θ and ω.

4.5 A first-order autoregressive model is generated from the white noise series w_t using the generating equations

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

where ϕ, for $|\phi| < 1$, is a parameter and the w_t are independent random variables with mean zero and variance σ_w^2.

(a) Show that the power spectrum of x_t is given by

$$f_x(\omega) = \frac{\sigma_w^2}{1 + \phi^2 - 2\phi\cos(2\pi\omega)}.$$

(b) Verify the autocovariance function of this process is

$$\gamma_x(h) = \frac{\sigma_w^2\ \phi^{|h|}}{1 - \phi^2},$$

$h = 0, \pm 1, \pm 2, \ldots$, by showing that the inverse transform of $\gamma_x(h)$ is the spectrum derived in part (a).

4.6 In applications, we will often observe series containing a signal that has been delayed by some unknown time D, i.e.,

$$x_t = s_t + As_{t-D} + n_t,$$

where s_t and n_t are stationary and independent with zero means and spectral densities $f_s(\omega)$ and $f_n(\omega)$, respectively. The delayed signal is multiplied by some unknown constant A.

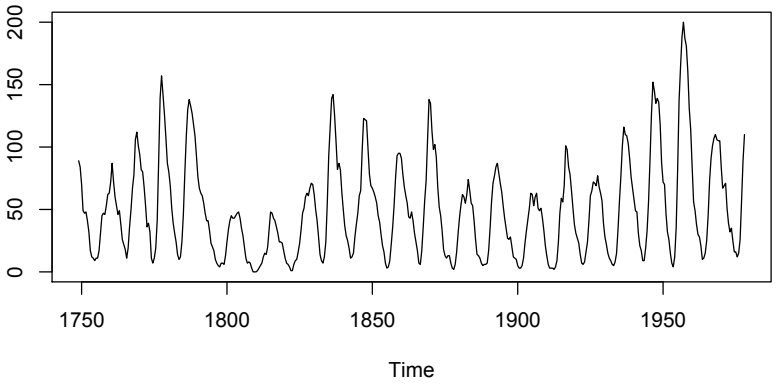

Fig. 4.31. Smoothed 12-month sunspot numbers (`sunspotz`) sampled twice per year.

(a) Prove
$$f_x(\omega) = [1 + A^2 + 2A\cos(2\pi\omega D)]f_s(\omega) + f_n(\omega).$$
(b) How could the periodicity expected in the spectrum derived in (a) be used to estimate the delay D? (Hint: Consider the case where $f_n(\omega) = 0$; i.e., there is no noise.)

4.7 Suppose x_t and y_t are stationary zero-mean time series with x_t independent of y_s for all s and t. Consider the product series
$$z_t = x_t y_t.$$
Prove the spectral density for z_t can be written as
$$f_z(\omega) = \int_{-1/2}^{1/2} f_x(\omega - \nu) f_y(\nu)\, d\nu.$$

Section 4.4

4.8 Figure 4.31 shows the biyearly smoothed (12-month moving average) number of sunspots from June 1749 to December 1978 with $n = 459$ points that were taken twice per year; the data are contained in `sunspotz`. With Example 4.10 as a guide, perform a periodogram analysis identifying the predominant periods and obtaining confidence intervals for the identified periods. Interpret your findings.

4.9 The levels of salt concentration known to have occurred over rows, corresponding to the average temperature levels for the soil science data considered in Figures 1.15 and 1.16, are in `salt` and `saltemp`. Plot the series and then identify the dominant frequencies by performing separate spectral analyses on the two series. Include confidence intervals for the dominant frequencies and interpret your findings.

4.10 Let the observed series x_t be composed of a periodic signal and noise so it can be written as

$$x_t = \beta_1 \cos(2\pi\omega_k t) + \beta_2 \sin(2\pi\omega_k t) + w_t,$$

where w_t is a white noise process with variance σ_w^2. The frequency ω_k is assumed to be known and of the form k/n in this problem. Suppose we consider estimating β_1, β_2 and σ_w^2 by least squares, or equivalently, by maximum likelihood if the w_t are assumed to be Gaussian.

(a) Prove, for a fixed ω_k, the minimum squared error is attained by

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = 2n^{-1/2} \begin{pmatrix} d_c(\omega_k) \\ d_s(\omega_k) \end{pmatrix},$$

where the cosine and sine transforms (4.23) and (4.24) appear on the right-hand side.

(b) Prove that the error sum of squares can be written as

$$SSE = \sum_{t=1}^{n} x_t^2 - 2I_x(\omega_k)$$

so that the value of ω_k that minimizes squared error is the same as the value that maximizes the periodogram $I_x(\omega_k)$ estimator (4.20).

(c) Under the Gaussian assumption and fixed ω_k, show that the F-test of no regression leads to an F-statistic that is a monotone function of $I_x(\omega_k)$.

4.11 Prove the convolution property of the DFT, namely,

$$\sum_{s=1}^{n} a_s x_{t-s} = \sum_{k=0}^{n-1} d_A(\omega_k) d_x(\omega_k) \exp\{2\pi\omega_k t\},$$

for $t = 1, 2, \ldots, n$, where $d_A(\omega_k)$ and $d_x(\omega_k)$ are the discrete Fourier transforms of a_t and x_t, respectively, and we assume that $x_t = x_{t+n}$ is periodic.

Section 4.5

4.12 Repeat Problem 4.8 using a nonparametric spectral estimation procedure. In addition to discussing your findings in detail, comment on your choice of a spectral estimate with regard to smoothing and tapering.

4.13 Repeat Problem 4.9 using a nonparametric spectral estimation procedure. In addition to discussing your findings in detail, comment on your choice of a spectral estimate with regard to smoothing and tapering.

4.14 The periodic behavior of a time series induced by echoes can also be observed in the spectrum of the series; this fact can be seen from the results stated in Problem 4.6(a). Using the notation of that problem, suppose we observe $x_t = s_t + As_{t-D} + n_t$, which implies the spectra satisfy $f_x(\omega) = [1 + A^2 + 2A\cos(2\pi\omega D)]f_s(\omega) + f_n(\omega)$. If the noise is negligible ($f_n(\omega) \approx 0$) then $\log f_x(\omega)$ is approximately the sum of a periodic component, $\log[1 + A^2 + 2A\cos(2\pi\omega D)]$, and $\log f_s(\omega)$. Bogart et al. (1962) proposed treating the detrended log spectrum as a pseudo time series and calculating its spectrum, or *cepstrum*, which should show a peak at a *quefrency* corresponding to $1/D$. The cepstrum can be plotted as a function of quefrency, from which the delaty D can be estimated.

For the speech series presented in Example 1.3, estimate the pitch period using cepstral analysis as follows. The data are in `speech`.

(a) Calculate and display the log-periodogram of the data. Is the periodogram periodic, as predicted?
(b) Perform a cepstral (spectral) analysis on the detrended logged periodogram, and use the results to estimate the delay D. How does your answer compare with the analysis of Example 1.24, which was based on the ACF?

4.15 Use Property 4.2 to verify (4.63). Then verify (4.66) and (4.67).

4.16 Consider two time series

$$x_t = w_t - w_{t-1},$$

$$y_t = \tfrac{1}{2}(w_t + w_{t-1}),$$

formed from the white noise series w_t with variance $\sigma_w^2 = 1$.

(a) Are x_t and y_t jointly stationary? Recall the cross-covariance function must also be a function only of the lag h and cannot depend on time.
(b) Compute the spectra $f_y(\omega)$ and $f_x(\omega)$, and comment on the difference between the two results.
(c) Suppose sample spectral estimators $\bar{f}_y(.10)$ are computed for the series using $L = 3$. Find a and b such that

$$P\left\{a \le \bar{f}_y(.10) \le b\right\} = .90.$$

This expression gives two points that will contain 90% of the sample spectral values. Put 5% of the area in each tail.

Section 4.6

4.17 Analyze the coherency between the temperature and salt data discussed in Problem 4.9. Discuss your findings.

4.18 Consider two processes

$$x_t = w_t \quad \text{and} \quad y_t = \phi x_{t-D} + v_t$$

where w_t and v_t are independent white noise processes with common variance σ^2, ϕ is a constant, and D is a fixed integer delay.

(a) Compute the coherency between x_t and y_t.
(b) Simulate $n = 1024$ normal observations from x_t and y_t for $\phi = .9$, $\sigma^2 = 1$, and $D = 0$. Then estimate and plot the coherency between the simulated series for the following values of L and comment:
(i) $L = 1$, (ii) $L = 3$, (iii) $L = 41$, and (iv) $L = 101$.

Section 4.7

4.19 For the processes in Problem 4.18:

(a) Compute the phase between x_t and y_t.
(b) Simulate $n = 1024$ observations from x_t and y_t for $\phi = .9$, $\sigma^2 = 1$, and $D = 1$. Then estimate and plot the phase between the simulated series for the following values of L and comment:
(i) $L = 1$, (ii) $L = 3$, (iii) $L = 41$, and (iv) $L = 101$.

4.20 Consider the bivariate time series records containing monthly U.S. production as measured by the Federal Reserve Board Production Index and monthly unemployment as given in Figure 3.21.

(a) Compute the spectrum and the log spectrum for each series, and identify statistically significant peaks. Explain what might be generating the peaks. Compute the coherence, and explain what is meant when a high coherence is observed at a particular frequency.
(b) What would be the effect of applying the filter

$$u_t = x_t - x_{t-1} \quad \text{followed by} \quad v_t = u_t - u_{t-12}$$

to the series given above? Plot the predicted frequency responses of the simple difference filter and of the seasonal difference of the first difference.
(c) Apply the filters successively to one of the two series and plot the output. Examine the output after taking a first difference and comment on whether stationarity is a reasonable assumption. Why or why not? Plot after taking the seasonal difference of the first difference. What can be noticed about the output that is consistent with what you have predicted from the frequency response? Verify by computing the spectrum of the output after filtering.

4.21 Determine the theoretical power spectrum of the series formed by combining the white noise series w_t to form

$$y_t = w_{t-2} + 4w_{t-1} + 6w_t + 4w_{t+1} + w_{t+2}.$$

Determine which frequencies are present by plotting the power spectrum.

4.22 Let $x_t = \cos(2\pi\omega t)$, and consider the output

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k},$$

where $\sum_k |a_k| < \infty$. Show

$$y_t = |A(\omega)| \cos(2\pi\omega t + \phi(\omega)),$$

where $|A(\omega)|$ and $\phi(\omega)$ are the amplitude and phase of the filter, respectively. Interpret the result in terms of the relationship between the input series, x_t, and the output series, y_t.

4.23 Suppose x_t is a stationary series, and we apply two filtering operations in succession, say,

$$y_t = \sum_r a_r x_{t-r} \quad \text{then} \quad z_t = \sum_s b_s y_{t-s}.$$

(a) Show the spectrum of the output is

$$f_z(\omega) = |A(\omega)|^2 |B(\omega)|^2 f_x(\omega),$$

where $A(\omega)$ and $B(\omega)$ are the Fourier transforms of the filter sequences a_t and b_t, respectively.

(b) What would be the effect of applying the filter

$$u_t = x_t - x_{t-1} \quad \text{followed by} \quad v_t = u_t - u_{t-12}$$

to a time series?

(c) Plot the predicted frequency responses of the simple difference filter and of the seasonal difference of the first difference. Filters like these are called seasonal adjustment filters in economics because they tend to attenuate frequencies at multiples of the monthly periods. The difference filter tends to attenuate low-frequency trends.

4.24 Suppose we are given a stationary zero-mean series x_t with spectrum $f_x(\omega)$ and then construct the derived series

$$y_t = a y_{t-1} + x_t, \quad t = \pm 1, \pm 2, \ldots.$$

(a) Show how the theoretical $f_y(\omega)$ is related to $f_x(\omega)$.

(b) Plot the function that multiplies $f_x(\omega)$ in part (a) for $a = .1$ and for $a = .8$. This filter is called a recursive filter.

Section 4.8

4.25 Often, the periodicities in the sunspot series are investigated by fitting an autoregressive spectrum of sufficiently high order. The main periodicity is often stated to be in the neighborhood of 11 years. Fit an autoregressive spectral estimator to the sunspot data using a model selection method of your choice. Compare the result with a conventional nonparametric spectral estimator found in Problem 4.8.

4.26 Fit an autoregressive spectral estimator to the Recruitment series and compare it to the results of Example 4.13.

4.27 Suppose a sample time series with $n = 256$ points is available from the first-order autoregressive model. Furthermore, suppose a sample spectrum computed with $L = 3$ yields the estimated value $\bar{f}_x(1/8) = 2.25$. Is this sample value consistent with $\sigma_w^2 = 1, \phi = .5$? Repeat using $L = 11$ if we just happen to obtain the same sample value.

4.28 Suppose we wish to test the noise alone hypothesis $H_0 : x_t = n_t$ against the signal-plus-noise hypothesis $H_1 : x_t = s_t + n_t$, where s_t and n_t are uncorrelated zero-mean stationary processes with spectra $f_s(\omega)$ and $f_n(\omega)$. Suppose that we want the test over a band of $L = 2m + 1$ frequencies of the form $\omega_{j:n} + k/n$, for $k = 0, \pm 1, \pm 2, \ldots, \pm m$ near some fixed frequency ω. Assume that both the signal and noise spectra are approximately constant over the interval.

(a) Prove the approximate likelihood-based test statistic for testing H_0 against H_1 is proportional to

$$T = \sum_k |d_x(\omega_{j:n} + k/n)|^2 \left(\frac{1}{f_n(\omega)} - \frac{1}{f_s(\omega) + f_n(\omega)} \right).$$

(b) Find the approximate distributions of T under H_0 and H_1.
(c) Define the false alarm and signal detection probabilities as $P_F = P\{T > K|H_0\}$ and $P_d = P\{T > k|H_1\}$, respectively. Express these probabilities in terms of the signal-to-noise ratio $f_s(\omega)/f_n(\omega)$ and appropriate chi-squared integrals.

Section 4.9

4.29 Repeat the dynamic Fourier analysis of Example 4.21 on the remaining seven earthquakes and seven explosions in the data file `eqexp`. Do the conclusions about the difference between earthquakes and explosions stated in the example still seem valid?

4.30 Repeat the wavelet analyses of Examples 4.22 and 4.23 on all earthquake and explosion series in the data file `eqexp`. Do the conclusions about the difference between earthquakes and explosions stated in Examples 4.22 and 4.23 still seem valid?

4.31 Using Examples 4.21-4.23 as a guide, perform a dynamic Fourier analysis and wavelet analyses (dwt and waveshrink analysis) on the event of unknown origin that took place near the Russian nuclear test facility in Novaya Zemlya. State your conclusion about the nature of the event at Novaya Zemlya.

Section 4.10

4.32 Consider the problem of approximating the filter output

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}, \quad \sum_{-\infty}^{\infty} |a_k| < \infty,$$

by

$$y_t^M = \sum_{|k|<M/2} a_k^M x_{t-k}$$

for $t = M/2 - 1, M/2, \ldots, n - M/2$, where x_t is available for $t = 1, \ldots, n$ and

$$a_t^M = M^{-1} \sum_{k=0}^{M-1} A(\omega_k) \exp\{2\pi i \omega_k t\}$$

with $\omega_k = k/M$. Prove

$$E\{(y_t - y_t^M)^2\} \le 4\gamma_x(0) \Bigg(\sum_{|k|\ge M/2} |a_k| \Bigg)^2.$$

4.33 Prove the squared coherence $\rho^2_{y\cdot x}(\omega) = 1$ for all ω when

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r},$$

that is, when x_t and y_t can be related exactly by a linear filter.

4.34 The data set `climhyd`, contains 454 months of measured values for six climatic variables: (i) air temperature [`Temp`], (ii) dew point [`DewPt`], (iii) cloud cover [`CldCvr`], (iv) wind speed [`WndSpd`], (v) precipitation [`Precip`], and (vi) inflow [`Inflow`], at Lake Shasta in California; the data are displayed in Figure 7.3. We would like to look at possible relations among the weather factors and between the weather factors and the inflow to Lake Shasta.

(a) First transform the inflow and precipitation series as follows: $I_t = \log i_t$, where i_t is inflow, and $P_t = \sqrt{p}_t$, where p_t is precipitation. Then, compute the square coherencies between all the weather variables and transformed inflow and argue that the strongest determinant of the inflow series is (transformed) precipitation. [*Tip:* If `x` contains multiple time series, then the easiest way to display all the squared coherencies is to first make an object of class *spec*; e.g., `u = spectrum(x, span=c(7,7), plot=FALSE)` and then plot the coherencies suppressing the confidence intervals, `plot(u, ci=-1, plot.type="coh")`.]

(b) Fit a lagged regression model of the form

$$I_t = \beta_0 + \sum_{j=0}^{\infty} \beta_j P_{t-j} + w_t,$$

using thresholding, and then comment of the predictive ability of precipitation for inflow.

Section 4.11

4.35 Consider the *signal plus noise* model

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t,$$

where the signal and noise series, x_t and v_t are both stationary with spectra $f_x(\omega)$ and $f_v(\omega)$, respectively. Assuming that x_t and v_t are independent of each other for all t, verify (4.137) and (4.138).

4.36 Consider the model

$$y_t = x_t + v_t,$$

where

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

such that v_t is Gaussian white noise and independent of x_t with $\mathrm{var}(v_t) = \sigma_v^2$, and w_t is Gaussian white noise and independent of v_t, with $\mathrm{var}(w_t) = \sigma_w^2$, and $|\phi| < 1$ and $Ex_0 = 0$. Prove that the spectrum of the observed series y_t is

$$f_y(\omega) = \frac{\sigma^2 |1 - \theta e^{-2\pi i\omega}|^2}{|1 - \phi e^{-2\pi i\omega}|^2},$$

where

$$\theta = \frac{c \pm \sqrt{c^2 - 4}}{2}, \quad \sigma^2 = \frac{\sigma_v^2 \phi}{\theta},$$

and

$$c = \frac{\sigma_w^2 + \sigma_v^2(1 + \phi^2)}{\sigma_v^2 \phi}.$$

4.37 Consider the same model as in the preceding problem.

(a) Prove the optimal smoothed estimator of the form

$$\widehat{x}_t = \sum_{s=-\infty}^{\infty} a_s y_{t-s}$$

has

$$a_s = \frac{\sigma_w^2}{\sigma^2} \frac{\theta^{|s|}}{1-\theta^2}.$$

(b) Show the mean square error is given by

$$E\{(x_t - \widehat{x}_t)^2\} = \frac{\sigma_v^2 \sigma_w^2}{\sigma^2(1-\theta^2)}.$$

(c) Compare mean square error of the estimator in part (b) with that of the optimal finite estimator of the form

$$\widehat{x}_t = a_1 y_{t-1} + a_2 y_{t-2}$$

when $\sigma_v^2 = .053, \sigma_w^2 = .172$, and $\phi_1 = .9$.

Section 4.12

4.38 Consider the two-dimensional linear filter given as the output (4.149).

(a) Express the two-dimensional autocovariance function of the output, say, $\gamma_y(h_1, h_2)$, in terms of an infinite sum involving the autocovariance function of $x_{\boldsymbol{s}}$ and the filter coefficients a_{s_1,s_2}.

(b) Use the expression derived in (a), combined with (4.148) and (4.151) to derive the spectrum of the filtered output (4.150).

The following problems require supplemental material from Appendix C

4.39 Let w_t be a Gaussian white noise series with variance σ_w^2. Prove that the results of Theorem C.4 hold without error for the DFT of w_t.

4.40 Show that condition (4.40) implies (C.19) by showing

$$n^{-1/2} \sum_{h \geq 0} h\,|\gamma(h)| \leq \sigma_w^2 \sum_{k \geq 0} |\psi_k| \sum_{j \geq 0} \sqrt{j}\,|\psi_j|.$$

4.41 Prove Lemma C.4.

4.42 Finish the proof of Theorem C.5.

4.43 For the zero-mean complex random vector $\boldsymbol{z} = \boldsymbol{x}_c - i\boldsymbol{x}_s$, with $\mathrm{cov}(\boldsymbol{z}) = \Sigma = C - iQ$, with $\Sigma = \Sigma^*$, define

$$w = 2\mathrm{Re}(\boldsymbol{a}^* \boldsymbol{z}),$$

where $\boldsymbol{a} = \boldsymbol{a}_c - i\boldsymbol{a}_s$ is an arbitrary non-zero complex vector. Prove

$$\mathrm{cov}(w) = 2\boldsymbol{a}^* \Sigma \boldsymbol{a}.$$

Recall * denotes the complex conjugate transpose.