

Free Texts in Statistics

Robert H. Shumway
David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

EZ – Third Edition

Robert H. Shumway
David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

EZ - Third Edition

Version: 2015.04.25

Contents

Preface to the EZ Third Edition	vii
1 Characteristics of Time Series	1
1.1 Introduction	1
1.2 The Nature of Time Series Data	1
1.3 Time Series Statistical Models	8
1.4 Measures of Dependence: Autocorrelation and Cross-Correlation	13
1.5 Stationary Time Series	17
1.6 Estimation of Correlation	23
Problems	28
2 Time Series Regression and Exploratory Data Analysis	33
2.1 Classical Regression in the Time Series Context	33
2.2 Exploratory Data Analysis	42
2.3 Smoothing in the Time Series Context	51
Problems	54
3 ARIMA Models	57
3.1 Introduction	57
3.2 Autoregressive Moving Average Models	57
3.3 Autocorrelation and Partial Autocorrelation	67
3.4 Estimation	74
3.5 Forecasting	81
3.6 Maximum Likelihood Estimation **	87
3.7 Integrated Models for Nonstationary Data	91
3.8 Building ARIMA Models	95
3.9 Regression with Autocorrelated Errors	103
3.10 Multiplicative Seasonal ARIMA Models	106
Problems	115

4	Spectral Analysis and Filtering	119
4.1	Introduction	119
4.2	Cyclical Behavior and Periodicity	119
4.3	The Spectral Density	126
4.4	Periodogram and Discrete Fourier Transform	129
4.5	Nonparametric Spectral Estimation	135
4.6	Parametric Spectral Estimation	148
4.7	Linear Filters	150
4.8	Multiple Series and Cross-Spectra	155
	Problems	160
Appendix R:	R Supplement	167
R.1	First Things First	167
R.1.1	The Package for the Text: <code>astsa</code>	167
R.1.2	Other Packages	168
R.2	Getting Started	168
R.2.1	Basic Statistics	171
R.2.2	Word	173
R.3	Time Series Primer	174
R.3.1	Included Data Sets	180
R.3.2	Included Scripts	182
References		185
Index		199

Preface to the EZ Third Edition

The EZ version can be used for an introductory time series course where the prerequisite is an understanding of linear regression analysis, although it assumes basic high school math skills (trigonometry, complex numbers, polynomials, calculus, and so on). For readers who are a little rusty on the basic skills, the WikiBook on K-12 mathematics http://en.wikibooks.org/wiki/Subject:K-12_mathematics may be a useful resource; in particular, we mention the book covering **calculus**. We occasionally use matrix notation. For readers lacking this skill, see the high school page on **matrices**.

A calculus based introduction to probability course is an essential co-requisite for this course. Readers should be familiar with most of the content of **basic probability facts** (a pdf file), which is a short introduction to the necessary material.

The EZ version follows the first four chapters of regular version of the text with some sections removed or rearranged, and with the addition of a few examples to take the place of removed sections. Appendix R has been updated for the use of the R package for the text **astsa**; all other appendices have been removed. *Because of the rearrangement of some sections, the code listing on the website for the text will not match the EZ version.* Also, the problem sets in this version differ from the regular version.

Although chapters 5 – 7 were removed, the section from Chapter 5 on regression with autocorrelated errors was moved to Chapter 3 (along with the associated exercises). In Chapter 3, there are two estimation sections, one that uses only least squares and method of moments, and a double-starred section on MLE that comes after the section on forecasting. The MLE section requires skills obtained in a course on basic mathematical statistics and may be skipped.

Internal links are **dark red**, external links are **magenta**, R code is in **blue typewriter font**, output is **purple** and comments are **# green**.

Characteristics of Time Series

1.1 Introduction

The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that these adjacent observations are independent and identically distributed. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis.

Historically, time series methods were applied to problems in the physical and environmental sciences. This fact accounts for the basic engineering flavor permeating the language of time series analysis. In our view, the first step in any time series investigation always involves careful scrutiny of the recorded data plotted over time. Before looking more closely at the particular statistical methods, it is appropriate to mention that two separate, but not necessarily mutually exclusive, approaches to time series analysis exist, commonly identified as the *time domain approach* ([Chapter 3](#)) and the *frequency domain approach* ([Chapter 4](#)).

1.2 The Nature of Time Series Data

The following examples illustrate some of the common kinds of time series data as well as some of the statistical questions that might be asked about such data.

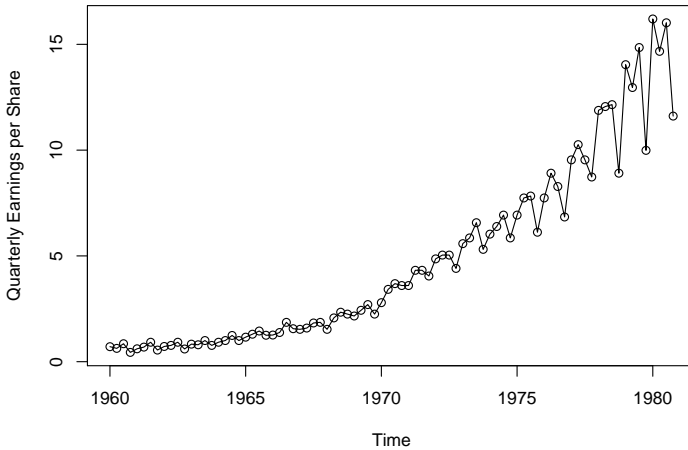


Fig. 1.1. Johnson & Johnson quarterly earnings per share, 1960-I to 1980-IV.

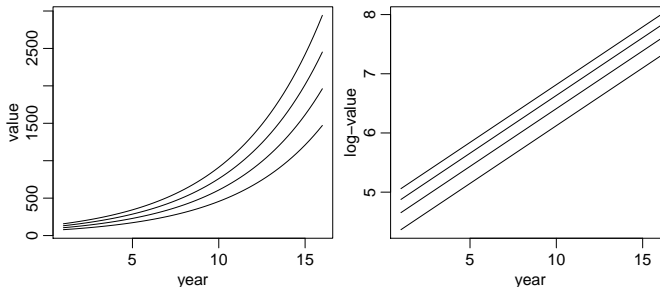


Fig. 1.2. *Left:* Quarterly value of initial deposits of \$75, \$100, \$125, and \$150 over 15 years, with a quarterly growth rate of 5%; $x_t = (1 + .05)x_{t-1}$. *Right:* Logs of the quarterly values; $\log(x_t) = \log(1 + .05) + \log(x_{t-1})$.

Example 1.1 Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the increasing underlying trend and variability, and a somewhat regular oscillation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 2 (see Problem 2.1) using regression techniques. Also, compare Figure 1.1 with Figure 1.2.

To use package `astsa`, and then plot the data for this example using R, type the following.

```
require(astsa)      # ** SEE FOOTNOTE
plot(jj, type="o", ylab="Quarterly Earnings per Share")
```

** Throughout the text, we assume that the R package for the book, `astsa`, has been downloaded and installed. See Appendix R (Section R.1.1) for further details. All the data files used in the (original) text are loaded when `astsa` is loaded.

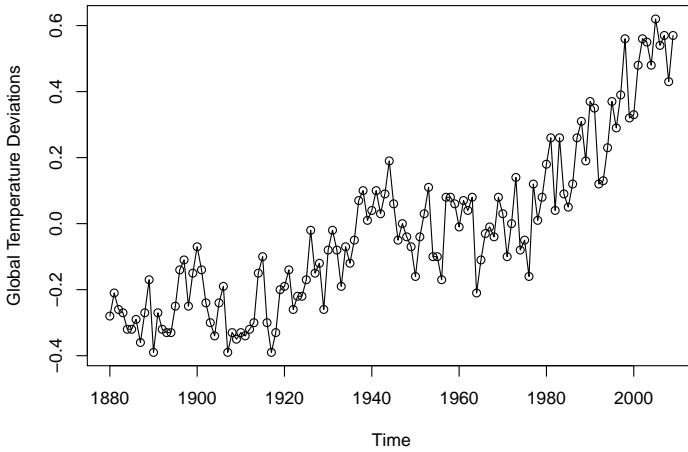


Fig. 1.3. Yearly average global temperature deviations (1880–2009) in $^{\circ}\text{C}$.

Example 1.2 Global Warming

Consider the global temperature series record shown in [Figure 1.3](#). The data are the global mean land–ocean temperature index from 1880 to 2009, with the base period 1951–1980. In particular, the data are deviations, measured in degrees centigrade, from the 1951–1980 average, and are an update of Hansen et al. (2006). We note an apparent upward trend in the series during the latter part of the twentieth century that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970. The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface.

The R code for this example is similar to the code in [Example 1.1](#):

```
plot(gtemp, type="o", ylab="Global Temperature Deviations")
```

Example 1.3 Speech Data

More involved questions develop in applications to the physical sciences. [Figure 1.4](#) shows a small .1 second (1000 point) sample of recorded speech for the phrase *aaa...hhh*, and we note the repetitive nature of the signal and the rather regular periodicities. One current problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase *aaa...hhh*. Spectral analysis can be used in this context to produce a signature of this phrase that can be compared with signatures of various library syllables to look for a match. One can immediately notice the rather regular repetition of small wavelets. The separation between the packets is known as the pitch period and represents the response of the vocal tract filter to a periodic sequence of pulses stimulated by the opening and closing of the glottis.

In R, you can reproduce [Figure 1.4](#) as follows:

```
plot(speech)
```

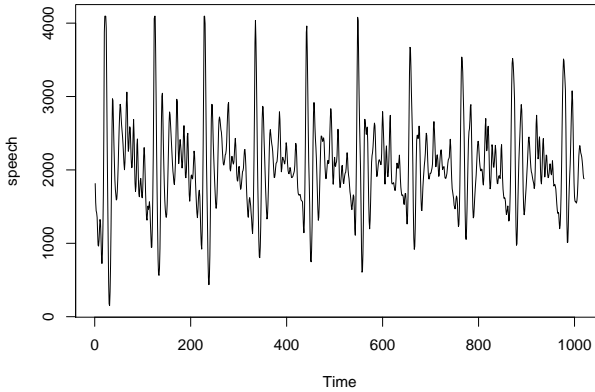


Fig. 1.4. Speech recording of the syllable *aaa...hhh* sampled at 10,000 points per second with $n = 1020$ points.

Example 1.4 New York Stock Exchange

As an example of financial time series data, [Figure 1.5](#) shows the daily returns (or percent change) of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. It is easy to spot the crash of October 19, 1987 in the figure. The data shown in [Figure 1.5](#) are typical of return data. The mean of the series appears to be stable with an average return of approximately zero, however, the volatility (or variability) of data changes over time. In fact, the data show volatility clustering; that is, highly volatile periods tend to be clustered together. A problem in the analysis of these type of financial data is to forecast the volatility of future returns. For example, GARCH models have been developed to handle these problems. The R code for this example is similar to the previous examples:

```
plot(nyse, ylab="NYSE Returns")
```

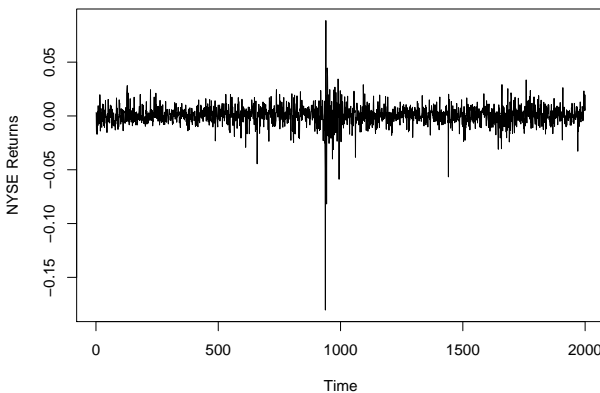


Fig. 1.5. Returns of the NYSE. The data are daily value weighted market returns from February 2, 1984 to December 31, 1991 (2000 trading days). The crash of October 19, 1987 occurs at $t = 938$.

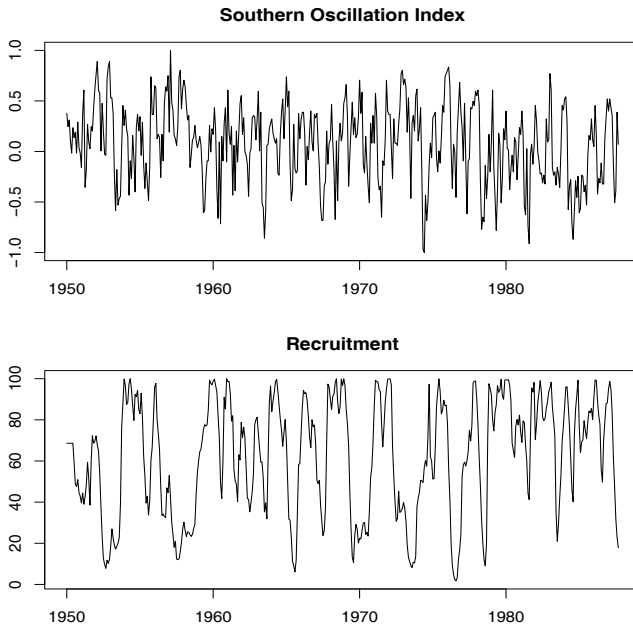


Fig. 1.6. Monthly SOI and Recruitment (estimated new fish), 1950-1987.

Example 1.5 El Niño and Fish Population

We may also be interested in analyzing several time series at once. **Figure 1.6** shows monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated Recruitment (an index of the number of new fish). Both series are for a period of 453 months ranging over the years 1950–1987. SOI measures changes in air pressure related to sea surface temperatures in the central Pacific Ocean. The central Pacific warms every three to seven years due to the El Niño effect, which has been blamed, in particular, for the 1997 floods in the midwestern portions of the United States. Both series in **Figure 1.6** tend to exhibit repetitive behavior, with regularly repeating cycles that are easily visible. This periodic behavior is of interest because underlying processes of interest may be regular and the rate or frequency of oscillation characterizing the behavior of the underlying series would help to identify them. The Recruitment series also shows several kinds of oscillations. The study of the kinds of cycles and their strengths is the subject of Chapter 4. The two series also tend to be somewhat related; it is easy to imagine that somehow the fish population is dependent on the SOI.

The following R code will reproduce **Figure 1.6**:

```
par(mfrow = c(2,1)) # set up the graphics
plot(soi, ylab="", xlab="", main="Southern Oscillation Index")
plot(rec, ylab="", xlab="", main="Recruitment")
```

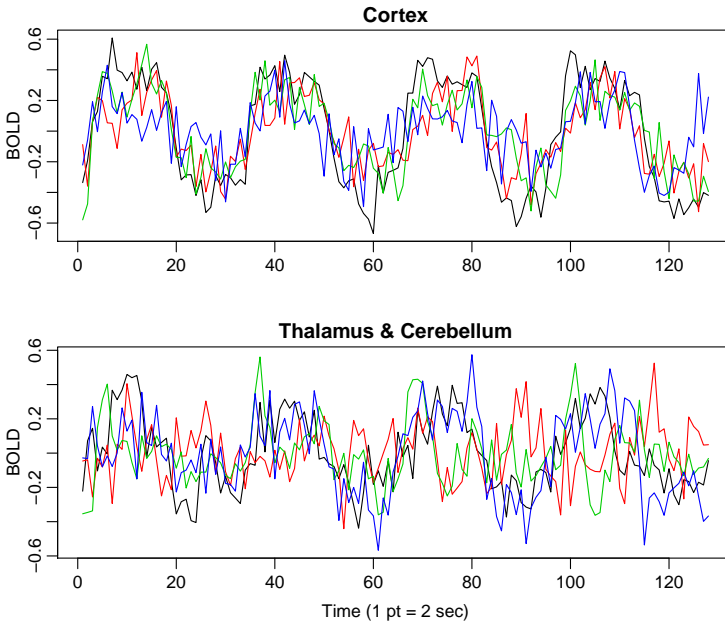


Fig. 1.7. fMRI data from various locations in the cortex, thalamus, and cerebellum; $n = 128$ points, one observation taken every 2 seconds.

Example 1.6 fMRI Imaging

A fundamental problem in classical statistics occurs when we are given a collection of independent series or vectors of series, generated under varying experimental conditions or treatment configurations. Such a set of series is shown in [Figure 1.7](#), where we observe data collected from various locations in the brain via functional magnetic resonance imaging (fMRI). In this example, five subjects were given periodic brushing on the hand. The stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$). For this example, we averaged the results over subjects (these were evoked responses, and all subjects were in phase). The series shown in [Figure 1.7](#) are consecutive measures of blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. Notice that the periodicities appear strongly in the motor cortex series and less strongly in the thalamus and cerebellum. The fact that one has series from different areas of the brain suggests testing whether the areas are responding differently to the brush stimulus.

Use the following R commands to plot the data:

```
par(mfrow=c(2,1), mar=c(3,2,1,0)+.5, mgp=c(1.6,.6,0))
ts.plot(fmri1[,2:5], col=1:4, ylab="BOLD", xlab="", main="Cortex")
ts.plot(fmri1[,6:9], col=1:4, ylab="BOLD", xlab="", main="Thalam & Cereb")
mtext("Time (1 pt = 2 sec)", side=1, line=2)
```

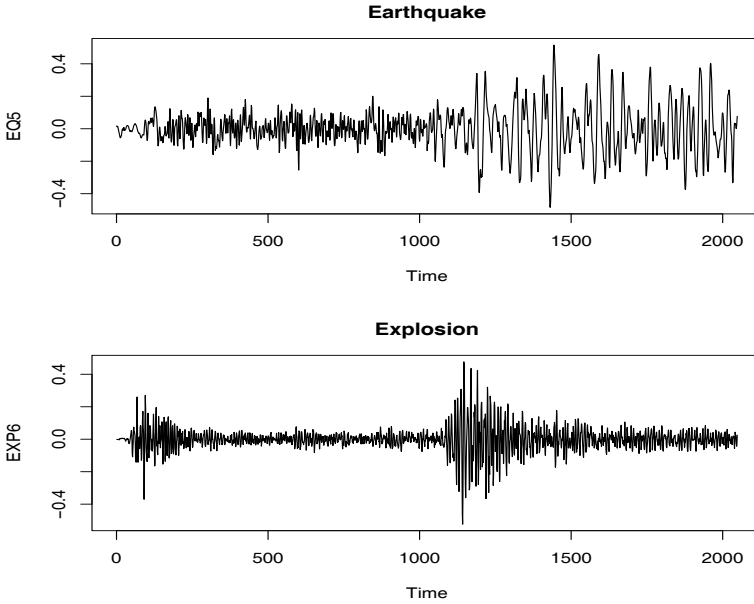


Fig. 1.8. Arrival phases from an earthquake (top) and explosion (bottom) at 40 points per second.

Example 1.7 Earthquakes and Explosions

As a final example, the series in [Figure 1.8](#) represent two phases or arrivals along the surface, denoted by P ($t = 1, \dots, 1024$) and S ($t = 1025, \dots, 2048$), at a seismic recording station. The recording instruments in Scandinavia are observing earthquakes and mining explosions with one of each shown in [Figure 1.8](#). The general problem of interest is in distinguishing or discriminating between waveforms generated by earthquakes and those generated by explosions. Features that may be important are the rough amplitude ratios of the first phase P to the second phase S, which tend to be smaller for earthquakes than for explosions. In the case of the two events in [Figure 1.8](#), the ratio of maximum amplitudes appears to be somewhat less than .5 for the earthquake and about 1 for the explosion. Otherwise, note a subtle difference exists in the periodic nature of the S phase for the earthquake. We can again think about spectral analysis of variance for testing the equality of the periodic components of earthquakes and explosions. We would also like to be able to classify future P and S components from events of unknown origin.

To plot the data as in this example, use the following commands in R:

```
par(mfrow=c(2,1))
plot(EQ5, main="Earthquake")
plot(EXP6, main="Explosion")
```

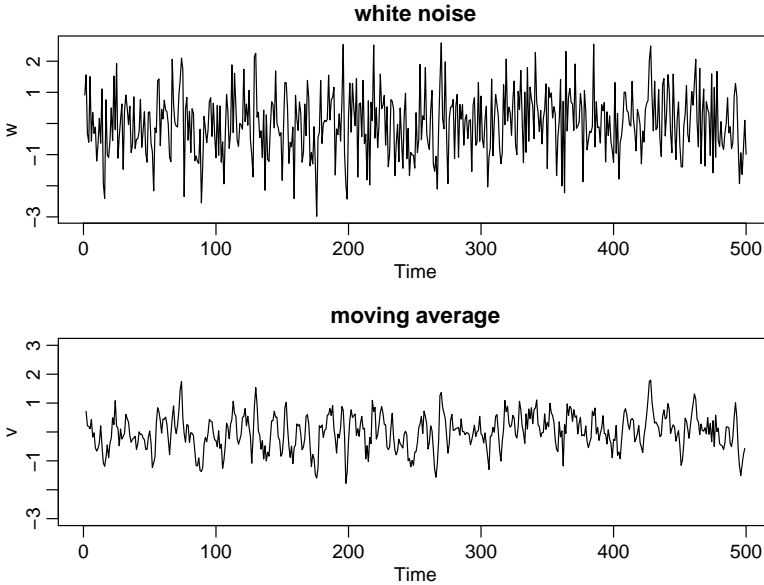


Fig. 1.9. Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom).

1.3 Time Series Statistical Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section.

The fundamental visual characteristic distinguishing the different series shown in [Example 1.1](#) – [Example 1.7](#) is their differing degrees of smoothness. One possible explanation for this smoothness is that adjacent points in time are correlated, so the value of the series at time t , say, x_t , depends in some way on the past values x_{t-1}, x_{t-2}, \dots . This model expresses a fundamental way in which we might think about generating realistic-looking time series.

Example 1.8 White Noise

A simple kind of generated series might be a collection of uncorrelated random variables, w_t , with mean 0 and finite variance σ_w^2 . The time series generated from uncorrelated variables is used as a model for noise in engineering applications where it is called *white noise*; we shall sometimes denote this process as $w_t \sim wn(0, \sigma_w^2)$. The designation white originates from the analogy with white light and indicates that all possible periodic oscillations are present with equal strength.

We will, at times, also require the noise to be independent and identically distributed (iid) random variables with mean 0 and variance σ_w^2 . We shall distinguish this case by saying white independent noise, or by writing $w_t \sim$

$\text{iid}(0, \sigma_w^2)$. A particularly useful white noise series is Gaussian white noise, wherein the w_t are independent normal random variables, with mean 0 and variance σ_w^2 ; or more succinctly, $w_t \sim \text{iid } N(0, \sigma_w^2)$. Figure 1.9 shows in the upper panel a collection of 500 such random variables, with $\sigma_w^2 = 1$, plotted in the order in which they were drawn. The resulting series bears a slight resemblance to the explosion in Figure 1.8 but is not smooth enough to serve as a plausible model for any of the other experimental series. The plot tends to show visually a mixture of many different kinds of oscillations in the white noise series.

If the stochastic behavior of all time series could be explained in terms of the white noise model, classical statistical methods would suffice. Two ways of introducing serial correlation and more smoothness into time series models are given in Example 1.9 and Example 1.10.

Example 1.9 Moving Averages and Filtering

We might replace the white noise series w_t by a moving average that smooths the series. For example, consider replacing w_t in Example 1.8 by an average of its current value and its immediate neighbors in the past and future. That is, let

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1}), \quad (1.1)$$

which leads to the series shown in the lower panel of Figure 1.9. Inspecting the series shows a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster oscillations are taken out. We begin to notice a similarity to the SOI in Figure 1.6, or perhaps, to some of the fMRI series in Figure 1.7.

To reproduce Figure 1.9 in R use the following commands. A linear combination of values in a time series such as in (1.1) is referred to, generically, as a filtered series; hence the command `filter`.

```
w = rnorm(500,0,1) # 500 N(0,1) variates
v = filter(w, sides=2, rep(1/3,3)) # moving average
par(mfrow=c(2,1))
plot.ts(w, main="white noise")
plot.ts(v, ylim=c(-3,3), main="moving average")
```

The speech series in Figure 1.4 and the Recruitment series in Figure 1.6, as well as some of the MRI series in Figure 1.7, differ from the moving average series because one particular kind of oscillatory behavior seems to predominate, producing a sinusoidal type of behavior. A number of methods exist for generating series with this quasi-periodic behavior; we illustrate a popular one based on the autoregressive model considered in Chapter 3.

Example 1.10 Autoregressions

Suppose we consider the white noise series w_t of Example 1.8 as input and calculate the output using the second-order equation

$$x_t = x_{t-1} - .9x_{t-2} + w_t \quad (1.2)$$

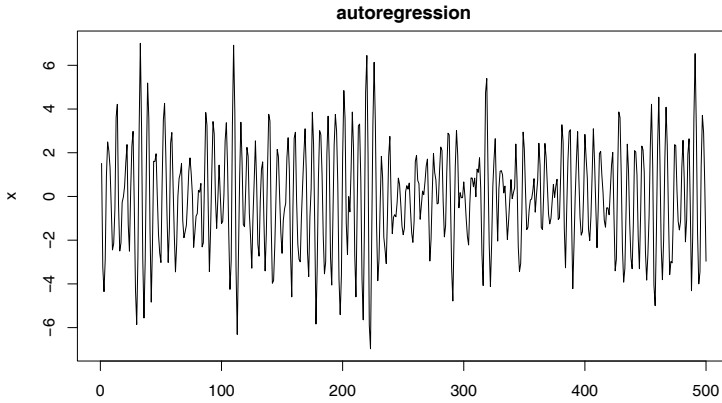


Fig. 1.10. Autoregressive series generated from model (1.2).

successively for $t = 1, 2, \dots, 500$. Equation (1.2) represents a regression or prediction of the current value x_t of a time series as a function of the past two values of the series, and, hence, the term autoregression is suggested for this model. A problem with startup values exists here because (1.2) also depends on the initial conditions x_0 and x_{-1} , but, for now, we assume that we are given these values and generate the succeeding values by substituting into (1.2). That is, given w_1, w_2, \dots, w_{500} , and x_0, x_{-1} , we start with $x_1 = x_0 - .9x_{-1} + w_1$, then recursively compute $x_2 = x_1 - .9x_0 + w_2$, then $x_3 = x_2 - .9x_1 + w_3$, and so on. The resulting output series is shown in Figure 1.10, and we note the periodic behavior of the series, which is similar to that displayed by the speech series in Figure 1.4. The autoregressive model above and its generalizations can be used as an underlying model for many observed series and will be studied in detail in Chapter 3.

One way to simulate and plot data from the model (1.2) in R is to use the following commands (another way is to use `arma.sim`). The initial conditions are set equal to zero, so we let the filter run an extra 50 values to avoid startup problems.

```
w = rnorm(550,0,1) # 50 extra to avoid startup problems
x = filter(w, filter=c(1,-.9), method="recursive")[-(1:50)]
plot.ts(x, main="autoregression")
```

Example 1.11 Random Walk with Drift

A model for analyzing trend such as seen in the global temperature data in Figure 1.3, is the random walk with drift model given by

$$x_t = \delta + x_{t-1} + w_t \quad (1.3)$$

for $t = 1, 2, \dots$, with initial condition $x_0 = 0$, and where w_t is white noise. The constant δ is called the drift, and when $\delta = 0$, the model is called simply a random walk because the value of the time series at time t is the value of the series at time $t - 1$ plus a completely random movement determined

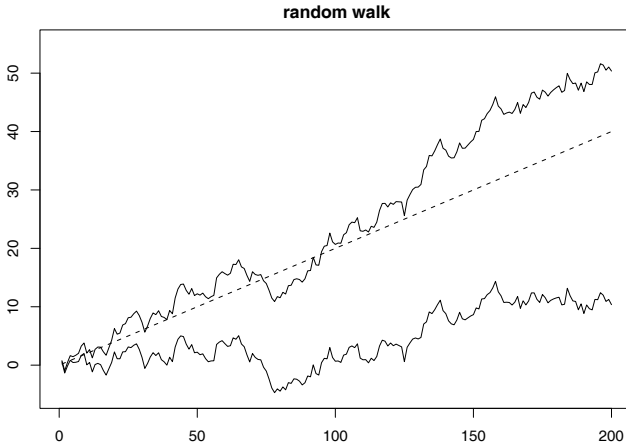


Fig. 1.11. Random walk, $\sigma_w = 1$, with drift $\delta = .2$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and a straight line with slope $.2$ (dashed line).

by w_t . Note that we may rewrite (1.3) as a cumulative sum of white noise variates. That is,

$$x_t = \delta t + \sum_{j=1}^t w_j \quad (1.4)$$

for $t = 1, 2, \dots$; either use induction, or plug (1.4) into (1.3) to verify this statement. Figure 1.11 shows 200 observations generated from the model with $\delta = 0$ and $.2$, and with $\sigma_w = 1$. For comparison, we also superimposed the straight line $.2t$ on the graph.

To reproduce Figure 1.11 in R use the following code (notice the use of multiple commands per line using a semicolon).

```
set.seed(154) # so you can reproduce the results
w = rnorm(200,0,1); x = cumsum(w) # two commands in one line
wd = w +.2; xd = cumsum(wd)
plot.ts(xd, ylim=c(-5,55), main="random walk")
lines(x); lines(.2*(1:200), lty="dashed")
```

Example 1.12 Signal in Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise. For example, it is easy to detect the regular cycle fMRI series displayed on the top of Figure 1.7. Consider the model

$$x_t = 2 \cos(2\pi t/50 + .6\pi) + w_t \quad (1.5)$$

for $t = 1, 2, \dots, 500$, where the first term is regarded as the signal, shown in the upper panel of Figure 1.12. We note that a sinusoidal waveform can be written as

$$A \cos(2\pi\omega t + \phi), \quad (1.6)$$

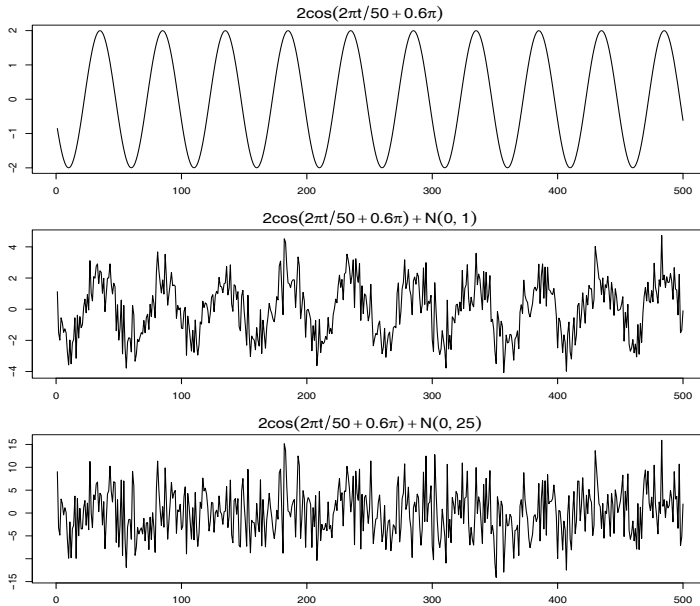


Fig. 1.12. Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel); see (1.5).

where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift. In (1.5), $A = 2$, $\omega = 1/50$ (one cycle every 50 time points), and $\phi = .6\pi$.

An additive noise term was taken to be white noise with $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel), drawn from a normal distribution. Adding the two together obscures the signal, as shown in the lower panels of Figure 1.12. Of course, the degree to which the signal is obscured depends on the amplitude of the signal and the size of σ_w . The ratio of the amplitude of the signal to σ_w (or some function of the ratio) is sometimes called the signal-to-noise ratio (SNR); the larger the SNR, the easier it is to detect the signal. Note that the signal is easily discernible in the middle panel of Figure 1.12, whereas the signal is obscured in the bottom panel. Typically, we will not observe the signal but the signal obscured by noise.

To reproduce Figure 1.12 in R, use the following commands:

```
cs = 2*cos(2*pi*1:500/50 + .6*pi)
w = rnorm(500,0,1)
par(mfrow=c(3,1), mar=c(3,2,2,1), cex.main=1.5)
plot.ts(cs, main=expression(2*cos(2*pi*t/50+.6*pi)))
plot.ts(cs+w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,1)))
plot.ts(cs+5*w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,25)))
```

1.4 Measures of Dependence: Autocorrelation and Cross-Correlation

A complete description of a time series, observed as a collection of n random variables at arbitrary integer time points t_1, t_2, \dots, t_n , for any positive integer n , is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less than the n constants, c_1, c_2, \dots, c_n ; i.e.,

$$F(c_1, c_2, \dots, c_n) = P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_n} \leq c_n). \quad (1.7)$$

Unfortunately, the multidimensional distribution function cannot usually be written easily unless the random variables are jointly normal, in which case the joint density has the well-known form displayed in (1.27).

Although the joint distribution function describes the data completely, it is an unwieldy tool for displaying and analyzing time series data. The distribution function (1.7) must be evaluated as a function of n arguments, so any plotting of the corresponding multivariate density functions is virtually impossible. The marginal distribution functions $F_t(x) = P\{x_t \leq x\}$ or the corresponding marginal density functions $f_t(x) = \frac{\partial F_t(x)}{\partial x}$, when they exist, are often informative for examining the marginal behavior of a series. Another informative marginal descriptive measure is the mean function.

Definition 1.1 *The mean function is defined as*

$$\mu_{xt} = E(x_t) \quad (1.8)$$

provided it exists, where E denotes the usual expected value operator.¹ When no confusion exists about which time series we are referring to, we will drop a subscript and write μ_{xt} as μ_t .

Example 1.13 Mean Function of a Moving Average Series

If w_t denotes a white noise series, then $\mu_{wt} = E(w_t) = 0$ for all t . The top series in Figure 1.9 reflects this, as the series clearly fluctuates around a mean value of zero. Smoothing the series as in Example 1.9 does not change the mean because we can write

$$\mu_{vt} = E(v_t) = \frac{1}{3}[E(w_{t-1}) + E(w_t) + E(w_{t+1})] = 0.$$

Example 1.14 Mean Function of a Random Walk with Drift

Consider the random walk with drift model given in (1.4),

$$x_t = \delta t + \sum_{j=1}^t w_j, \quad t = 1, 2, \dots$$

¹ If x_t is Gaussian with mean μ_t and variance σ_t^2 , abbreviated as $x_t \sim N(\mu_t, \sigma_t^2)$, the marginal density is given by $f_t(x) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp\{-\frac{1}{2\sigma_t^2}(x - \mu_t)^2\}$ for $x \in \mathbb{R}$. In this case, $\mu_t = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx$ and $\sigma_t^2 = E(x_t - \mu_t)^2 = \int_{-\infty}^{\infty} (x - \mu_t)^2 f_t(x) dx$.

Because $E(w_t) = 0$ for all t , and δ is a constant, we have

$$\mu_{xt} = E(x_t) = \delta t + \sum_{j=1}^t E(w_j) = \delta t$$

which is a straight line with slope δ . A realization of a random walk with drift can be compared to its mean function in [Figure 1.11](#).

Example 1.15 Mean Function of Signal Plus Noise

A great many practical applications depend on assuming the observed data have been generated by a fixed signal waveform superimposed on a zero-mean noise process, leading to an additive signal model of the form (1.5). It is clear, because the signal in (1.5) is a fixed function of time, we will have

$$\begin{aligned}\mu_{xt} &= E(x_t) = E[2 \cos(2\pi t/50 + .6\pi) + w_t] \\ &= 2 \cos(2\pi t/50 + .6\pi) + E(w_t) \\ &= 2 \cos(2\pi t/50 + .6\pi),\end{aligned}$$

and the mean function is just the cosine wave.

The lack of independence between two adjacent values x_s and x_t can be assessed numerically, as in classical statistics, using the notions of covariance and correlation. Assuming the variance of x_t is finite, we have the following definition.

Definition 1.2 *The autocovariance function is defined as the second moment product*

$$\gamma_x(s, t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \quad (1.9)$$

for all s and t . When no possible confusion exists about which time series we are referring to, we will drop the subscript and write $\gamma_x(s, t)$ as $\gamma(s, t)$.

Note that $\gamma_x(s, t) = \gamma_x(t, s)$ for all time points s and t . The autocovariance measures the *linear* dependence between two points on the same series observed at different times. Recall from classical statistics that if $\gamma_x(s, t) = 0$, x_s and x_t are not linearly related, but there still may be some dependence structure between them. If, however, x_s and x_t are bivariate normal, $\gamma_x(s, t) = 0$ ensures their independence. It is clear that, for $s = t$, the autocovariance reduces to the (assumed finite) variance, because

$$\gamma_x(t, t) = E[(x_t - \mu_t)^2] = \text{var}(x_t). \quad (1.10)$$

Example 1.16 Autocovariance of White Noise

The white noise series w_t has $E(w_t) = 0$ and

$$\gamma_w(s, t) = \text{cov}(w_s, w_t) = \begin{cases} \sigma_w^2 & s = t, \\ 0 & s \neq t. \end{cases} \quad (1.11)$$

A realization of white noise with $\sigma_w^2 = 1$ is shown in the top panel of [Figure 1.9](#).

Example 1.17 Autocovariance of a Moving Average

Consider applying a three-point moving average to the white noise series w_t of the previous example as in [Example 1.9](#). In this case,

$$\gamma_v(s, t) = \text{cov}(v_s, v_t) = \text{cov} \left\{ \frac{1}{3} (w_{s-1} + w_s + w_{s+1}), \frac{1}{3} (w_{t-1} + w_t + w_{t+1}) \right\}.$$

When $s = t$ we have²

$$\begin{aligned} \gamma_v(t, t) &= \frac{1}{9} \text{cov}\{(w_{t-1} + w_t + w_{t+1}), (w_{t-1} + w_t + w_{t+1})\} \\ &= \frac{1}{9} [\text{cov}(w_{t-1}, w_{t-1}) + \text{cov}(w_t, w_t) + \text{cov}(w_{t+1}, w_{t+1})] \\ &= \frac{3}{9} \sigma_w^2. \end{aligned}$$

When $s = t + 1$,

$$\begin{aligned} \gamma_v(t+1, t) &= \frac{1}{9} \text{cov}\{(w_t + w_{t+1} + w_{t+2}), (w_{t-1} + w_t + w_{t+1})\} \\ &= \frac{1}{9} [\text{cov}(w_t, w_t) + \text{cov}(w_{t+1}, w_{t+1})] \\ &= \frac{2}{9} \sigma_w^2, \end{aligned}$$

using (1.11). Similar computations give $\gamma_v(t-1, t) = 2\sigma_w^2/9$, $\gamma_v(t+2, t) = \gamma_v(t-2, t) = \sigma_w^2/9$, and 0 when $|t-s| > 2$. We summarize the values for all s and t as

$$\gamma_v(s, t) = \begin{cases} \frac{3}{9} \sigma_w^2 & s = t, \\ \frac{2}{9} \sigma_w^2 & |s - t| = 1, \\ \frac{1}{9} \sigma_w^2 & |s - t| = 2, \\ 0 & |s - t| > 2. \end{cases} \quad (1.12)$$

[Example 1.17](#) shows clearly that the smoothing operation introduces a covariance function that decreases as the separation between the two time points increases and disappears completely when the time points are separated by three or more time points. This particular autocovariance is interesting because it only depends on the time separation or lag and not on the absolute location of the points along the series. We shall see later that this dependence suggests a mathematical model for the concept of weak stationarity.

² If the random variables $U = \sum_{j=1}^m a_j X_j$ and $V = \sum_{k=1}^r b_k Y_k$ are linear combinations of random variables $\{X_j\}$ and $\{Y_k\}$, respectively, then $\text{cov}(U, V) = \sum_{j=1}^m \sum_{k=1}^r a_j b_k \text{cov}(X_j, Y_k)$. An easy way to remember this is to treat it like multiplication: $(a_1 x_1 + a_2 x_2)(b_1 y_1 + b_2 y_2) = \dots$. Furthermore, $\text{var}(U) = \text{cov}(U, U)$.

Example 1.18 Autocovariance of a Random Walk

For the random walk model, $x_t = \sum_{j=1}^t w_j$, we have

$$\gamma_x(s, t) = \text{cov}(x_s, x_t) = \text{cov}\left(\sum_{j=1}^s w_j, \sum_{k=1}^t w_k\right) = \min\{s, t\} \sigma_w^2,$$

because the w_t are uncorrelated random variables. Note that, as opposed to the previous examples, the autocovariance function of a random walk depends on the particular time values s and t , and not on the time separation or lag. Also, notice that the variance of the random walk, $\text{var}(x_t) = \gamma_x(t, t) = t \sigma_w^2$, increases without bound as time t increases. The effect of this variance increase can be seen in [Figure 1.11](#) where the processes start to move away from their mean functions δt (note that $\delta = 0$ and $.2$ in that example).

As in classical statistics, it is more convenient to deal with a measure of association between -1 and 1 , and this leads to the following definition.

Definition 1.3 *The autocorrelation function (ACF) is defined as*

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}. \quad (1.13)$$

The ACF measures the linear predictability of the series at time t , say x_t , using only the value x_s . We can show easily that $-1 \leq \rho(s, t) \leq 1$ using the Cauchy–Schwarz inequality.³ If we can predict x_t *perfectly* from x_s through a linear relationship, $x_t = \beta_0 + \beta_1 x_s$, then the correlation will be $+1$ when $\beta_1 > 0$, and -1 when $\beta_1 < 0$. Hence, we have a rough measure of the ability to forecast the series at time t from the value at time s .

Often, we would like to measure the predictability of another series y_t from the series x_s . Assuming both series have finite variances, we have the following definition.

Definition 1.4 *The cross-covariance function between two series, x_t and y_t , is*

$$\gamma_{xy}(s, t) = \text{cov}(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]. \quad (1.14)$$

The cross-covariance function can be scaled to live in $[-1, 1]$:

Definition 1.5 *The cross-correlation function (CCF) is given by*

$$\rho_{xy}(s, t) = \frac{\gamma_{xy}(s, t)}{\sqrt{\gamma_x(s, s)\gamma_y(t, t)}}. \quad (1.15)$$

³ The Cauchy–Schwarz inequality implies $|\gamma(s, t)|^2 \leq \gamma(s, s)\gamma(t, t)$.

We may easily extend the above ideas to the case of more than two series, say, $x_{t1}, x_{t2}, \dots, x_{tr}$; that is, multivariate time series with r components. For example, the extension of (1.14) in this case is

$$\gamma_{jk}(s, t) = E[(x_{sj} - \mu_{sj})(x_{tk} - \mu_{tk})] \quad j, k = 1, 2, \dots, r. \quad (1.16)$$

1.5 Stationary Time Series

The preceding definitions of the mean and autocovariance functions are completely general. Although we have not made any special assumptions about the behavior of the time series, many of the preceding examples have hinted that a sort of regularity may exist over time in the behavior of a time series.

Definition 1.6 *A strictly stationary time series is one for which the probabilistic behavior of every collection of values*

$$\{x_{t_1}, x_{t_2}, \dots, x_{t_k}\}$$

is identical to that of the time shifted set

$$\{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h}\},$$

for all $k = 1, 2, \dots$, all time points t_1, t_2, \dots, t_k , and all time shifts $h = 0, \pm 1, \pm 2, \dots$.

The version of stationarity in Definition 1.6 is too strong for most applications. For example, if we observe the number of hits to a website during a unit of time, then the process would be considered strictly stationary if the probabilistic behavior in the morning is identical to the evening, and the behavior on Sunday is the same as on Monday, and so on. In general, it is difficult to assess strict stationarity from a single data set. Rather than imposing conditions on all possible distributions of a time series, we will use a milder version that imposes conditions only on the first two moments of the series. We now have the following definition.

Definition 1.7 *A weakly stationary time series is a finite variance process where*

- (i) *the mean value function, μ_t , defined in (1.8) is constant and does not depend on time t , and*
- (ii) *the autocovariance function, $\gamma(s, t)$, defined in (1.9) depends on s and t only through their difference $|s - t|$.*

*Henceforth, we will use the term **stationary** to mean weakly stationary; if a process is stationary in the strict sense, we will use the term strictly stationary.*

It should be clear from the discussion of strict stationarity following [Definition 1.6](#) that a strictly stationary, finite variance, time series is also stationary. The converse is not true unless there are further conditions. One important case where stationarity implies strict stationarity is if the time series is Gaussian. We will make this concept more precise at the end of this section.

Note, for example, that a random walk is not stationary because its autocovariance function depends on time; see [Example 1.18](#) and [Problem 1.7](#). Also, the random walk with drift violates both conditions of [Definition 1.7](#) because, in addition, the mean function $\mu_{xt} = \delta t$ is a function of time t .

Because the mean function, $E(x_t) = \mu_t$, of a stationary time series is independent of time t , we will write

$$\mu_t = \mu. \quad (1.17)$$

Also, because the autocovariance function, $\gamma(s, t)$, of a stationary time series, x_t , depends on s and t only through their difference $|s - t|$, we may simplify the notation. Let $s = t + h$, where h represents the time shift or lag. Then

$$\gamma(t + h, t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h, 0)$$

because the time difference between times $t + h$ and t is the same as the time difference between times h and 0 . Thus, the autocovariance function of a stationary time series does not depend on the time argument t . Henceforth, for convenience, we will drop the second argument of $\gamma(h, 0)$.

Definition 1.8 *The autocovariance function of a stationary time series will be written as*

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]. \quad (1.18)$$

Definition 1.9 *The autocorrelation function (ACF) of a stationary time series will be written using (1.13) as*

$$\rho(h) = \frac{\gamma(t + h, t)}{\sqrt{\gamma(t + h, t + h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}. \quad (1.19)$$

The Cauchy–Schwarz inequality shows again that $-1 \leq \rho(h) \leq 1$ for all h , enabling one to assess the relative importance of a given autocorrelation value by comparing with the extreme values -1 and 1 .

Example 1.19 Stationarity of White Noise

The mean and autocovariance functions of the white noise series discussed in [Example 1.8](#) and [Example 1.16](#) are easily evaluated as $\mu_{wt} = 0$ and

$$\gamma_w(h) = \text{cov}(w_{t+h}, w_t) = \begin{cases} \sigma_w^2 & h = 0, \\ 0 & h \neq 0. \end{cases}$$

Thus, white noise satisfies the conditions of [Definition 1.7](#) and is weakly stationary or stationary.

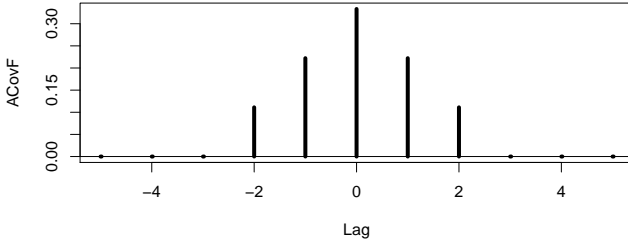


Fig. 1.13. Autocovariance function of a three-point moving average.

Example 1.20 Trend Stationarity

If $x_t = \alpha + \beta t + w_t$, then the mean function is $\mu_{x,t} = E(x_t) = \alpha + \beta t$, which is not independent of time. Therefore, the process is not stationary. The autocovariance function, however, is independent of time, because $\gamma_x(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu_{x,t+h})(x_t - \mu_{x,t})] = E(w_{t+h}w_t) = \gamma_w(h)$, which is given in [Example 1.19](#). Thus, the model may be considered as having stationary behavior around a linear trend; this behavior is sometimes called *trend stationarity*.

Example 1.21 Stationarity of a Moving Average

The three-point moving average process of [Example 1.9](#) is stationary because, from [Example 1.13](#) and [Example 1.17](#), the mean and autocovariance functions $\mu_{vt} = 0$, and

$$\gamma_v(h) = \begin{cases} \frac{3}{9}\sigma_w^2 & h = 0, \\ \frac{2}{9}\sigma_w^2 & h = \pm 1, \\ \frac{1}{9}\sigma_w^2 & h = \pm 2, \\ 0 & |h| > 2 \end{cases}$$

are independent of time t , satisfying the conditions of [Definition 1.7](#). [Figure 1.13](#) shows a plot of the autocovariance as a function of lag h with $\sigma_w^2 = 1$. Interestingly, the autocovariance function is symmetric about lag zero and decays as a function of lag.

The autocovariance function of a stationary process has several useful properties. First, the value at $h = 0$, namely

$$\gamma(0) = E[(x_t - \mu)^2] = \text{var}(x_t). \quad (1.20)$$

Also, the Cauchy–Schwarz inequality implies $|\gamma(h)| \leq \gamma(0)$. Another useful property is that the autocovariance function of a stationary series is symmetric around the origin,

$$\gamma(h) = \gamma(-h) \quad (1.21)$$

for all h . This property follows because shifting the series by h means that

$$\gamma(h) = \gamma(t + h - t)$$

$$\begin{aligned}
&= E[(x_{t+h} - \mu)(x_t - \mu)] \\
&= E[(x_t - \mu)(x_{t+h} - \mu)] \\
&= \gamma(t - (t + h)) \\
&= \gamma(-h),
\end{aligned}$$

which shows how to use the notation as well as proving the result.

When several series are available, a notion of stationarity still applies with additional conditions.

Definition 1.10 *Two time series, say, x_t and y_t , are said to be **jointly stationary** if they are each stationary, and the cross-covariance function*

$$\gamma_{xy}(h) = \text{cov}(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)] \quad (1.22)$$

is a function only of lag h .

Definition 1.11 *The **cross-correlation function (CCF)** of jointly stationary time series x_t and y_t is defined as*

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \quad (1.23)$$

Again, we have the result $-1 \leq \rho_{xy}(h) \leq 1$ which enables comparison with the extreme values -1 and 1 when looking at the relation between x_{t+h} and y_t . The cross-correlation function is not generally symmetric about zero [i.e., typically $\rho_{xy}(h) \neq \rho_{xy}(-h)$]; however, it is the case that

$$\rho_{xy}(h) = \rho_{yx}(-h), \quad (1.24)$$

which can be shown by manipulations similar to those used to show (1.21).

Example 1.22 Joint Stationarity

Consider the two series, x_t and y_t , formed from the sum and difference of two successive values of a white noise process, say,

$$x_t = w_t + w_{t-1} \quad \text{and} \quad y_t = w_t - w_{t-1},$$

where w_t are independent random variables with zero means and variance σ_w^2 . It is easy to show that $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$ and $\gamma_x(1) = \gamma_x(-1) = \sigma_w^2$, $\gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$. Also,

$$\gamma_{xy}(1) = \text{cov}(x_{t+1}, y_t) = \text{cov}(w_{t+1} + w_t, w_t - w_{t-1}) = \sigma_w^2$$

because only one term is nonzero (recall footnote 2 on page 15). Similarly, $\gamma_{xy}(0) = 0$, $\gamma_{xy}(-1) = -\sigma_w^2$. We obtain, using (1.23),

$$\rho_{xy}(h) = \begin{cases} 0 & h = 0, \\ 1/2 & h = 1, \\ -1/2 & h = -1, \\ 0 & |h| \geq 2. \end{cases}$$

Clearly, the autocovariance and cross-covariance functions depend only on the lag separation, h , so the series are jointly stationary.

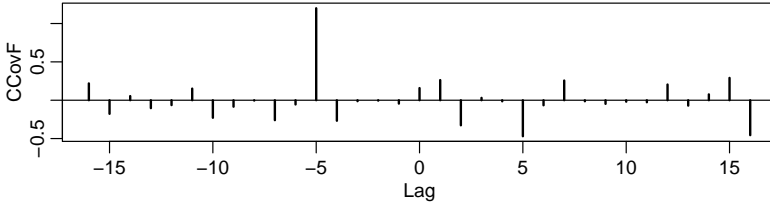


Fig. 1.14. Demonstration of the results of [Example 1.23](#) when $\ell = 5$.

Example 1.23 Prediction Using Cross-Correlation

As a simple example of cross-correlation, consider the problem of determining possible leading or lagging relations between two series x_t and y_t . If the model

$$y_t = Ax_{t-\ell} + w_t$$

holds, the series x_t is said to lead y_t for $\ell > 0$ and is said to lag y_t for $\ell < 0$. Hence, the analysis of leading and lagging relations might be important in predicting the value of y_t from x_t . Assuming, for convenience, that x_t and y_t have zero means, and the noise w_t is uncorrelated with the x_t series, the cross-covariance function can be computed as

$$\begin{aligned}\gamma_{xy}(h) &= \text{cov}(x_{t+h}, y_t) = \text{cov}(x_{t+h}, Ax_{t-\ell} + w_t) \\ &= \text{cov}(x_{t+h}, Ax_{t-\ell}) = A\gamma_x(h + \ell).\end{aligned}$$

Since the largest value of $\gamma_x(h + \ell)$ is $\gamma_x(0)$, i.e., when $h = -\ell$, the cross-covariance function will look like the autocovariance of the input series x_t , and it will have a peak on the negative side if x_t leads y_t and a peak on the positive side if x_t lags y_t . Below is the R code of an example with $\ell = 5$ and $\hat{\gamma}_{xy}(h)$ is shown in [Figure 1.14](#).

```
set.seed(90210)
x = rnorm(100)
y = lag(x,-5) + rnorm(100)
ccf(x,y, ylab='CCovF', type='covariance')
```

The concept of weak stationarity forms the basis for much of the analysis performed with time series. The fundamental properties of the mean and autocovariance functions (1.17) and (1.18) are satisfied by many theoretical models that appear to generate plausible sample realizations. In [Example 1.9](#) and [Example 1.10](#), two series were generated that produced stationary looking realizations, and in [Example 1.21](#), we showed that the series in [Example 1.9](#) was, in fact, weakly stationary. Both examples are special cases of the so-called linear process.

Definition 1.12 A linear process, x_t , is defined to be a linear combination of white noise variates w_t , and is given by⁴

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \quad (1.25)$$

The models we will encounter in Chapter 3 are linear processes. For the linear process (see **Problem 1.11**), we may show that the mean function is $E(x_t) = \mu$, and the autocovariance function is given by

$$\gamma(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j \quad (1.26)$$

for $h \geq 0$; recall that $\gamma(-h) = \gamma(h)$. This method exhibits the autocovariance function of the process in terms of the lagged products of the coefficients. Note that, for **Example 1.9**, we have $\psi_0 = \psi_{-1} = \psi_1 = 1/3$ and the result in **Example 1.21** comes out immediately. The autoregressive series in **Example 1.10** can also be put in this form as we will see in Chapter 3.

Notice that the linear process (1.25) is dependent on the future ($j < 0$), the present ($j = 0$), and the past ($j > 0$). For the purpose of forecasting, a future dependent model will be useless. Consequently, we will focus on processes that do not depend on the future. Such models are called *causal*, and a causal linear process has $\psi_j = 0$ for $j < 0$ in (1.25).

The multivariate normal distribution plays a key role in the analysis of linear time series. Most properties, including strategies for estimation and for forecasting rely on this distribution. Moreover, as previously mentioned, an important case in which a weakly stationary series is also strictly stationary is when the series is a normal or Gaussian series, which we now define.

Definition 1.13 A process, $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$, is said to be a **Gaussian process** if the n -dimensional vectors $\mathbf{x} = (x_{t_1}, x_{t_2}, \dots, x_{t_n})'$, for every collection of distinct time points t_1, t_2, \dots, t_n , and every positive integer n , have a (non-singular⁵) multivariate normal distribution.

Defining the $n \times 1$ mean vector $E(\mathbf{x}) \equiv \boldsymbol{\mu} = (\mu_{t_1}, \mu_{t_2}, \dots, \mu_{t_n})'$ and the $n \times n$ covariance matrix as $\text{var}(\mathbf{x}) \equiv \Gamma = \{\gamma(t_i, t_j); i, j = 1, \dots, n\}$, which is positive definite, the multivariate normal density function can be written as

$$f(\mathbf{x}) = (2\pi)^{-n/2} |\Gamma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \Gamma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (1.27)$$

where $|\cdot|$ denotes the determinant. We list some important items regarding a Gaussian process.

⁴ Let $x_t^n = \mu + \sum_{j=-n}^n \psi_j w_{t-j}$, then $x_t^n \rightarrow x_t$ as $n \rightarrow \infty$ in the sense that the mean squared error goes to zero; i.e., $E(x_t - x_t^n)^2 = \sigma_w^2 \sum_{|j|>n} \psi_j^2 \rightarrow 0$ as $n \rightarrow \infty$.

⁵ No value of the series is a perfect linear combination of others.

- If a Gaussian time series is weakly stationary, then $\mu_t = \mu$ and $\gamma(t_i, t_j) = \gamma(|t_i - t_j|)$, so that the vector $\boldsymbol{\mu}$ and the matrix Γ are independent of time. These facts imply that all the finite distributions, (1.27), of the series $\{x_t\}$ depend only on time lag and not on the actual times, and hence the series must be strictly stationary.
- A result called the *Wold Decomposition* states that a stationary Gaussian time series is a linear process as given in Definition 1.12. Hence, linear processes form the basis of inference and prediction for many time series.
- It is not enough for the marginal distributions to be Gaussian for the process to be Gaussian. It is easy to construct a situation where X and Y are normal, but (X, Y) is not bivariate normal (let X and Z be independent normals and let $Y = Z$ if $XZ > 0$ and $Y = -Z$ if $XZ \leq 0$).

1.6 Estimation of Correlation

Although the theoretical autocorrelation and cross-correlation functions are useful for describing the properties of certain hypothesized models, most of the analyses must be performed using sampled data. This limitation means the sampled points x_1, x_2, \dots, x_n only are available for estimating the mean, autocovariance, and autocorrelation functions. From the point of view of classical statistics, this poses a problem because we will typically not have iid copies of x_t that are available for estimating the covariance and correlation functions. In the usual situation with only one realization, however, the assumption of stationarity becomes critical. Somehow, we must use averages over this single realization to estimate the population means and covariance functions.

Accordingly, if a time series is stationary, the mean function (1.17) $\mu_t = \mu$ is constant so that we can estimate it by the sample mean,

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t. \quad (1.28)$$

The standard error of the estimate is the square root of $\text{var}(\bar{x})$, which can be computed using first principles (recall footnote 2 on page 15), and is given by

$$\begin{aligned} \text{var}(\bar{x}) &= \text{var} \left(\frac{1}{n} \sum_{t=1}^n x_t \right) = \frac{1}{n^2} \text{cov} \left(\sum_{t=1}^n x_t, \sum_{s=1}^n x_s \right) \\ &= \frac{1}{n^2} \left(n\gamma_x(0) + (n-1)\gamma_x(1) + (n-2)\gamma_x(2) + \cdots + \gamma_x(n-1) \right. \\ &\quad \left. + (n-1)\gamma_x(-1) + (n-2)\gamma_x(-2) + \cdots + \gamma_x(1-n) \right) \\ &= \frac{1}{n} \sum_{h=-n}^n \left(1 - \frac{|h|}{n} \right) \gamma_x(h). \end{aligned} \quad (1.29)$$

If the process is white noise, (1.29) reduces to the familiar σ_x^2/n recalling that $\gamma_x(0) = \sigma_x^2$. Note that, in the case of dependence, the standard error of \bar{x} may be smaller or larger than the white noise case depending on the nature of the correlation structure (see [Problem 1.15](#)).

The theoretical autocovariance function, (1.18), is estimated by the sample autocovariance function defined as follows.

Definition 1.14 *The sample autocovariance function is defined as*

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}), \quad (1.30)$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \dots, n-1$.

The sum in (1.30) runs over a restricted range because x_{t+h} is not available for $t+h > n$. The estimator in (1.30) is preferred to the one that would be obtained by dividing by $n-h$ because (1.30) is a non-negative definite function. The autocovariance function, $\gamma(h)$, of a stationary process is non-negative definite ensuring that variances of linear combinations of the variates x_t will never be negative. And, because $\text{var}(a_1 x_{t_1} + \dots + a_n x_{t_n})$ is never negative, the estimate of that variance should also be non-negative. The estimator in (1.30) guarantees this result, but no such guarantee exists if we divide by $n-h$. Note that neither dividing by n nor $n-h$ in (1.30) yields an unbiased estimator of $\gamma(h)$.

Definition 1.15 *The sample autocorrelation function is defined, analogously to (1.19), as*

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \quad (1.31)$$

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data comes from a completely random or white series or whether correlations are statistically significant at some lags.

Property 1.1 Large-Sample Distribution of the ACF

If x_t is white noise, then for n large and under mild conditions, the sample ACF, $\hat{\rho}_x(h)$, for $h = 1, 2, \dots, H$, where H is fixed but arbitrary, is approximately normal with zero mean and standard deviation given by of $\frac{1}{\sqrt{n}}$.

Based on [Property 1.1](#), we obtain a rough method for assessing whether a series is white noise by determining how many values of $\hat{\rho}(h)$ are outside the interval $\pm 2/\sqrt{n}$ (two standard errors); for white noise, approximately 95% of the sample ACFs should be within these limits.⁶ Most statistical packages indicate these bounds by default when displaying the sample ACF; the bounds

⁶ In this text, $z_{.025} = 1.95996398454005423552 \dots$ of normal fame, which is often rounded to 1.96, is rounded to 2.

do not hold for general processes and can be ignored if the interest is other than assessing whiteness. The applications of this property develop because many statistical modeling procedures depend on reducing a time series to a white noise series using various kinds of transformations. After such a procedure is applied, the plotted ACFs of the residuals should then lie roughly within the limits given above.

Example 1.24 A Simulated Time Series

To compare the sample ACF for various sample sizes to the theoretical ACF, consider a contrived set of data generated by tossing a fair coin, letting $x_t = 1$ when a head is obtained and $x_t = -1$ when a tail is obtained. Then, construct y_t as

$$y_t = 5 + x_t - .7x_{t-1}. \quad (1.32)$$

To simulate data, we consider two cases, one with a small sample size ($n = 10$) and another with a moderate sample size ($n = 100$).

```
set.seed(101010)
x1 = 2*rbinom(11, 1, .5) - 1      # simulated sequence of coin tosses
x2 = 2*rbinom(101, 1, .5) - 1
y1 = 5 + filter(x, sides=1, filter=c(1,-.7))[-1]
y2 = 5 + filter(y, sides=1, filter=c(1,-.7))[-1]
plot.ts(y1, type='s'); plot.ts(y2, type='s') # not shown
acf(y1, lag.max=4, plot=FALSE) # se = 1/sqrt(10) = .32
  Autocorrelations of series 'y1', by lag
    0      1      2      3      4
1.000 -0.688  0.425 -0.306 -0.007
acf(y2, lag.max=4, plot=FALSE) # se = 1/sqrt(100) = .1
  Autocorrelations of series 'y2', by lag
    0      1      2      3      4
1.000 -0.480 -0.002 -0.004  0.000
# Note that the sample ACF at lag zero is always 1 (Why?).
```

The theoretical ACF can be obtained from the model (1.32) using the fact that the mean of x_t is zero and the variance of x_t is one. It can be shown that

$$\rho_y(1) = \frac{-.7}{1 + .7^2} = -.47$$

and $\rho_y(h) = 0$ for $|h| > 1$ (Problem 1.20). It is interesting to compare the theoretical ACF with sample ACFs for the realization where $n = 10$ and the other realization where $n = 100$; note the increased variability in the smaller size sample.

Example 1.25 ACF of a Speech Signal

Computing the sample ACF as in the previous example can be thought of as matching the time series h units in the future, say, x_{t+h} against itself, x_t . Figure 1.15 shows the ACF of the speech series of Figure 1.4. The original series appears to contain a sequence of repeating short signals. The ACF confirms this behavior, showing repeating peaks spaced at about 106-109 points. The distance between the repeating signals is known as the *pitch*

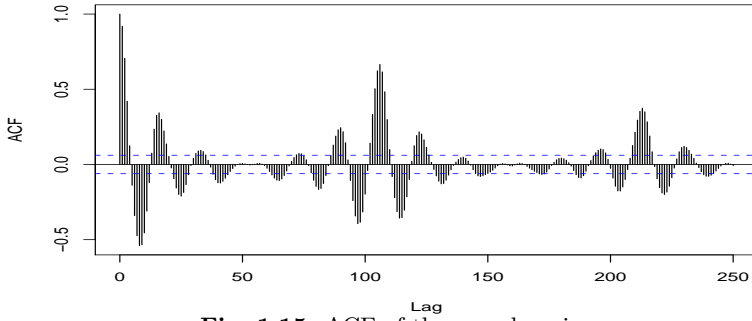


Fig. 1.15. ACF of the speech series.

period and is a fundamental parameter of interest in systems that encode and decipher speech. Because the series is sampled at 10,000 points per second, the pitch period appears to be between .0106 and .0109 seconds. To reproduce Figure 1.15, use `acf(speech, 250)`

Definition 1.16 The estimators for the cross-covariance function, $\gamma_{xy}(h)$, as given in (1.22) and the cross-correlation, $\rho_{xy}(h)$, in (1.23) are given, respectively, by the **sample cross-covariance function**

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \quad (1.33)$$

where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ determines the function for negative lags, and the **sample cross-correlation function**

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}. \quad (1.34)$$

The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data using the property mentioned in Example 1.23 for the theoretical cross-covariance function. Because $-1 \leq \hat{\rho}_{xy}(h) \leq 1$, the practical importance of peaks can be assessed by comparing their magnitudes with their theoretical maximum values. Furthermore, for x_t and y_t independent linear processes of the form (1.25), then under mild conditions, we have the following property.

Property 1.2 Large-Sample Distribution of Cross-Correlation

The large sample distribution of $\hat{\rho}_{xy}(h)$ is normal with mean zero and standard deviation $\frac{1}{\sqrt{n}}$ if at least one of the processes is independent white noise.

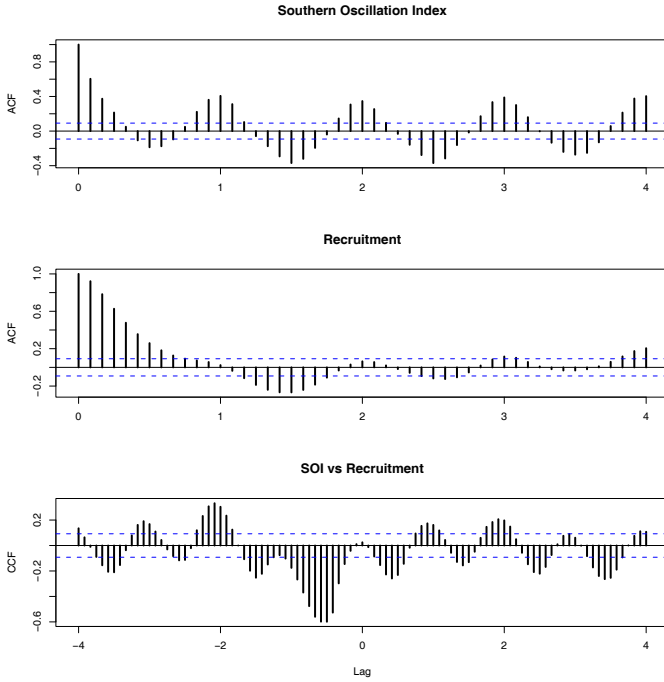


Fig. 1.16. Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment. The lag axes are in terms of seasons (12 months).

Example 1.26 SOI and Recruitment Correlation Analysis

The autocorrelation and cross-correlation functions are also useful for analyzing the joint behavior of two stationary series whose behavior may be related in some unspecified way. In [Example 1.5](#) (see [Figure 1.6](#)), we have considered simultaneous monthly readings of the SOI and the number of new fish (Recruitment) computed from a model. [Figure 1.16](#) shows the autocorrelation and cross-correlation functions (ACFs and CCF) for these two series.

Both of the ACFs exhibit periodicities corresponding to the correlation between values separated by 12 units. Observations 12 months or one year apart are strongly positively correlated, as are observations at multiples such as 24, 36, 48, \dots . Observations separated by six months are negatively correlated, showing that positive excursions tend to be associated with negative excursions six months removed. This appearance is rather characteristic of the pattern that would be produced by a sinusoidal component with a period of 12 months. The cross-correlation function peaks at $h = -6$, showing that the SOI measured at time $t - 6$ months is associated with the Recruitment series at time t . We could say the SOI leads the Recruitment series by six months. The sign of the CCF at $h = -6$ is negative, leading to the conclu-

sion that the two series move in different directions; that is, increases in SOI lead to decreases in Recruitment and vice versa. Again, note the periodicity of 12 months in the CCF.

The flat lines shown on the plots indicate $\pm 2/\sqrt{453}$, so that upper values would be exceeded about 2.5% of the time if the noise were white as specified in **Property 1.1** and **Property 1.2**. Of course, neither series is noise, so we can ignore these lines.

To reproduce **Figure 1.16** in R, use the following commands:

```
par(mfrow=c(3,1))
acf(soi, 48, main="Southern Oscillation Index")
acf(rec, 48, main="Recruitment")
ccf(soi, rec, 48, main="SOI vs Recruitment", ylab="CCF")
```

Problems

1.1 In 25 words or less, and without using symbols, why is stationarity important?

1.2 To compare the earthquake and explosion signals, plot the data displayed in **Figure 1.8** on the same graph using different colors or different line types and comment on the results. (The R code in **Example 1.11** may be of help on how to add lines to existing plots.)

1.3 (a) Generate $n = 100$ observations from the autoregression

$$x_t = -.9x_{t-2} + w_t$$

with $\sigma_w = 1$, using the method described in **Example 1.10**. Next, apply the moving average filter

$$v_t = (x_t + x_{t-1} + x_{t-2} + x_{t-3})/4$$

to x_t , the data you generated. Now plot x_t as a line and superimpose v_t as a dashed line. Note: `v = filter(x, rep(1/4, 4), sides = 1)`

(b) Repeat (a) but with

$$x_t = 2 \cos(2\pi t/4) + w_t,$$

where $w_t \sim \text{iid } N(0, 1)$.

(c) Repeat (a) but where x_t is the log of the Johnson & Johnson data discussed in **Example 1.1**.

(d) What is seasonal adjustment (you can do an internet search)?

(e) State your conclusions (in other words, what did you learn from this exercise).

1.4 Show that the autocovariance function can be written as

$$\gamma(s, t) = E[(x_s - \mu_s)(x_t - \mu_t)] = E(x_s x_t) - \mu_s \mu_t,$$

where $E[x_t] = \mu_t$.

1.5 Consider the time series

$$x_t = \beta_1 + \beta_2 t + w_t,$$

where β_1 and β_2 are regression coefficients, and w_t is a white noise process with variance σ_w^2 .

- (a) Determine whether x_t is stationary.
- (b) Show that the process $y_t = x_t - x_{t-1}$ is stationary.
- (c) Show that the mean of the moving average

$$v_t = \frac{1}{3}(x_{t-1} + x_t + x_{t+1})$$

is $\beta_1 + \beta_2 t$.

1.6 For a moving average process of the form

$$x_t = w_{t-1} + 2w_t + w_{t+1},$$

where w_t are independent with zero means and variance σ_w^2 , determine the autocovariance and autocorrelation functions as a function of lag h and sketch the ACF as a function of h .

1.7 Consider the random walk with drift model

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

for $t = 1, 2, \dots$, with $x_0 = 0$, where w_t is white noise with variance σ_w^2 .

- (a) Show that the model can be written as $x_t = \delta t + \sum_{k=1}^t w_k$.
- (b) Find the mean function and the autocovariance function of x_t .
- (c) Argue that x_t is not stationary.
- (d) Show $\rho_x(t-1, t) = \sqrt{\frac{t-1}{t}} \rightarrow 1$ as $t \rightarrow \infty$. What is the implication of this result?
- (e) Suggest a transformation to make the series stationary, and prove that the transformed series is stationary. (Hint: See [Problem 1.5b](#).)

1.8 Would you treat the global temperature data discussed in [Example 1.2](#) and shown in [Figure 1.3](#) as stationary or non-stationary? Support your answer.

1.9 A time series with a periodic component can be constructed from

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

where U_1 and U_2 are independent random variables with zero means and $E(U_1^2) = E(U_2^2) = \sigma^2$. The constant ω_0 determines the period or time it takes the process to make one complete cycle. Show that this series is weakly stationary with autocovariance function

$$\gamma(h) = \sigma^2 \cos(2\pi\omega_0 h).$$

1.10 Suppose we would like to predict a single stationary series x_t with zero mean and autocorrelation function $\gamma(h)$ at some time in the future, say, $t + \ell$, for $\ell > 0$.

- (a) If we predict using only x_t and some scale multiplier A , show that the mean-square prediction error

$$MSE(A) = E[(x_{t+\ell} - Ax_t)^2]$$

is minimized by the value

$$A = \rho(\ell).$$

- (b) Show that the minimum mean-square prediction error is

$$MSE(A) = \gamma(0)[1 - \rho^2(\ell)].$$

- (c) Show that if $x_{t+\ell} = Ax_t$, then $\rho(\ell) = 1$ if $A > 0$, and $\rho(\ell) = -1$ if $A < 0$.

1.11 Consider the linear process defined in (1.25).

- (a) Verify that the autocovariance function of the process is given by (1.26).
 (b) Use the result to find the autocovariance function of the process in **Problem 1.6**.

1.12 For two jointly stationary series x_t and y_t , verify (1.24).

1.13 Consider the two series

$$x_t = w_t$$

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

where w_t and u_t are independent white noise series with variances σ_w^2 and σ_u^2 , respectively, and θ is an unspecified constant.

- (a) Express the ACF, $\rho_y(h)$, for $h = 0, \pm 1, \pm 2, \dots$ of the series y_t as a function of σ_w^2 , σ_u^2 , and θ .
 (b) Determine the CCF, $\rho_{xy}(h)$ relating x_t and y_t .
 (c) Show that x_t and y_t are jointly stationary.

1.14 Let w_t , for $t = 0, \pm 1, \pm 2, \dots$ be a normal white noise process, and consider the series

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

Determine the mean and autocovariance function of x_t , and state whether it is stationary.

1.15 Suppose $x_t = \mu + w_t + \theta w_{t-1}$, where $w_t \sim wn(0, \sigma_w^2)$.

- (a) Show that mean function is $E(x_t) = \mu$.
 (b) Show that the autocovariance function of x_t is given by $\gamma_x(0) = \sigma_w^2(1 + \theta^2)$, $\gamma_x(\pm 1) = \sigma_w^2\theta$, and $\gamma_x(h) = 0$ otherwise.

- (c) Show that x_t is stationary for all values of $\theta \in \mathbb{R}$.
- (d) Use (1.29) to calculate $\text{var}(\bar{x})$ for estimating μ when (i) $\theta = 1$, (ii) $\theta = 0$, and (iii) $\theta = -1$
- (e) In time series, the sample size n is typically large, so that $\frac{(n-1)}{n} \approx 1$. With this as a consideration, comment on the results of part (d); in particular, how does the accuracy of the estimate of the mean μ change for the three different cases?

1.16 (a) Simulate a series of $n = 500$ Gaussian white noise observations as in [Example 1.8](#) and compute the sample ACF, $\hat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall [Example 1.19](#).]
 (b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

1.17 (a) Simulate a series of $n = 500$ moving average observations as in [Example 1.9](#) and compute the sample ACF, $\hat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall [Example 1.21](#).]
 (b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

1.18 Simulate 500 observations from the AR model specified in [Example 1.10](#) and then plot the sample ACF to lag 50. What does the sample ACF tell you about the approximate cyclic behavior of the data? Hint: Recall [Example 1.25](#) and [Example 1.26](#).

1.19 Simulate a series of $n = 500$ observations from the signal-plus-noise model presented in [Example 1.12](#) with (a) $\sigma_w = 0$, (b) $\sigma_w = 1$ and (c) $\sigma_w = 5$. Compute the sample ACF to lag 100 of the three series you generated and comment.

1.20 For the time series y_t described in [Example 1.24](#), verify the stated result that $\rho_y(1) = -.47$ and $\rho_y(h) = 0$ for $h > 1$.

1.21 In discussing the multivariate normal distribution, [Definition 1.13](#), it is mentioned that it is not enough for the marginal distributions to be Gaussian for the process to be Gaussian. An example was given where X and Y are normal, but (X, Y) is not bivariate normal: Let X and Z be independent normals and let $Y = Z$ if $XZ > 0$ and $Y = -Z$ if $XZ \leq 0$. Demonstrate this example via simulation.

Time Series Regression and Exploratory Data Analysis

2.1 Classical Regression in the Time Series Context

We begin our discussion of linear regression in the time series context by assuming some output or dependent time series, say, x_t , for $t = 1, \dots, n$, is being influenced by a collection of possible inputs or independent series, say, $z_{t1}, z_{t2}, \dots, z_{tq}$, where we first regard the inputs as fixed and known. This assumption, necessary for applying conventional linear regression, will be relaxed later on. We express this relation through the linear regression model

$$x_t = \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_q z_{tq} + w_t, \quad (2.1)$$

where $\beta_1, \beta_2, \dots, \beta_q$ are unknown fixed regression coefficients, and $\{w_t\}$ is a random error or noise process consisting of independent and identically distributed (iid) normal variables with mean zero and variance σ_w^2 ; we will relax the iid assumption later.

Example 2.1 Estimating a Linear Trend

Consider the global temperature data, say x_t , shown in [Figure 1.3](#) and [Figure 2.1](#). As discussed in [Example 1.2](#), there is an apparent upward trend in the series that has been used to argue the global warming hypothesis. We might use simple linear regression to estimate that trend by fitting the model

$$x_t = \beta_1 + \beta_2 t + w_t, \quad t = 1880, 1857, \dots, 2009.$$

This is in the form of the regression model (2.1) when we make the identification $q = 2$, $z_{t1} = 1$ and $z_{t2} = t$. Note that we are making the assumption that the errors, w_t , are an iid normal sequence, which may not be true; the problem of autocorrelated errors is discussed in detail in Chapter 3. Also note that we could have used, for example, $t = 1, \dots, 130$, without affecting the interpretation of the slope coefficient, β_2 ; only the intercept, β_1 , would be affected.

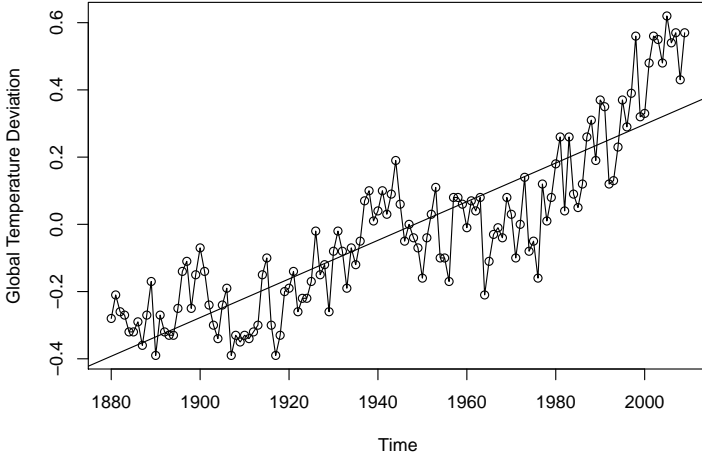


Fig. 2.1. Global temperature deviations shown in Figure 1.3 with fitted linear trend line, which SUCKS as an estimate of trend.

Using simple linear regression, we obtained the estimated coefficients $\hat{\beta}_1 = -11.2$, and $\hat{\beta}_2 = .006$ (with a standard error of .0003) yielding a highly significant estimated increase of .6 degrees centigrade per 100 years. We discuss the precise way in which the solution was accomplished after the example. Finally, Figure 2.1 shows the global temperature data, say x_t , with the estimated trend, say $\hat{x}_t = -11.2 + .006t$, superimposed. *It is apparent that the estimated trend line obtained via simple linear regression does not quite capture the trend of the data and better models will be needed.*

To perform this analysis in R, use the following commands:

```
summary(fit <- lm(gtemp~time(gtemp))) # regress gtemp on time
plot(gtemp, type="o", ylab="Global Temperature Deviation")
abline(fit) # add regression line to the plot
```

The linear model described by (2.1) above can be conveniently written in a more general notation by defining the column vectors $\mathbf{z}_t = (z_{t1}, z_{t2}, \dots, z_{tq})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)'$, where $'$ denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \beta_1 z_{t1} + \dots + \beta_q z_{tq} + w_t = \boldsymbol{\beta}' \mathbf{z}_t + w_t. \quad (2.2)$$

where $w_t \sim \text{iid } N(0, \sigma_w^2)$. In ordinary least squares (OLS), we estimate the unknown coefficient vector $\boldsymbol{\beta}$ by minimizing the error sum of squares

$$Q = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n (x_t - \boldsymbol{\beta}' \mathbf{z}_t)^2, \quad (2.3)$$

with respect to $\beta_1, \beta_2, \dots, \beta_q$. This minimization can be accomplished by differentiating (2.3) with respect to the vector $\boldsymbol{\beta}$ or by using the properties of projections. In the notation above, this procedure gives the *normal equations*

$$\left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}_t' \right) \hat{\boldsymbol{\beta}} = \sum_{t=1}^n \mathbf{z}_t x_t. \quad (2.4)$$

If $\sum_{t=1}^n \mathbf{z}_t \mathbf{z}_t'$ is non-singular, the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \sum_{t=1}^n \mathbf{z}_t x_t.$$

The minimized error sum of squares (2.3), denoted SSE , can be written as

$$SSE = \sum_{t=1}^n (x_t - \hat{\boldsymbol{\beta}}' \mathbf{z}_t)^2. \quad (2.5)$$

The ordinary least squares estimators are unbiased, i.e., $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and have the smallest variance within the class of linear unbiased estimators.

If the errors w_t are normally distributed, $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator for $\boldsymbol{\beta}$ and is normally distributed with

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma_w^2 C, \quad (2.6)$$

where

$$C = \left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \quad (2.7)$$

is a convenient notation for later equations. An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = MSE = \frac{SSE}{n - q}, \quad (2.8)$$

where MSE denotes the *mean squared error*, which is contrasted with the maximum likelihood estimator $\hat{\sigma}_w^2 = SSE/n$. Under the normal assumption, s_w^2 is distributed proportionally to a chi-squared random variable with $n - q$ degrees of freedom, denoted by χ_{n-q}^2 , and independently of $\hat{\boldsymbol{\beta}}$. It follows that

$$t_{n-q} = \frac{(\hat{\beta}_i - \beta_i)}{s_w \sqrt{c_{ii}}} \quad (2.9)$$

has the t-distribution with $n - q$ degrees of freedom; c_{ii} denotes the i -th diagonal element of C , as defined in (2.7).

Various competing models are often of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $r < q$ independent variables, say, $\{z_{t1}, z_{t2}, \dots, z_{tr}\}$ is influencing the dependent variable x_t . The reduced model is

$$x_t = \beta_1 z_{t1} + \dots + \beta_r z_{tr} + w_t \quad (2.10)$$

Table 2.1. Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square
$z_{t,r+1}, \dots, z_{t,q}$	$q - r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q - r)$
Error	$n - q$	SSE	$MSE = SSE/(n - q)$
Total	$n - r$	SSE_r	

where $\beta_1, \beta_2, \dots, \beta_r$ are a subset of coefficients of the original q variables. The null hypothesis in this case is $H_0: \beta_{r+1} = \dots = \beta_q = 0$. We can test the reduced model (2.10) against the full model (2.2) by comparing the error sums of squares under the two models using the F -statistic

$$F_{q-r, n-q} = \frac{(SSE_r - SSE)/(q - r)}{SSE/(n - q)}, \quad (2.11)$$

which has the central F -distribution with $q - r$ and $n - q$ degrees of freedom when (2.10) is the correct model. Note that SSE_r is the error sum of squares under the reduced model (2.10). The statistic, which follows from applying the likelihood ratio criterion, has the improvement per number of parameters added in the numerator compared with the error sum of squares under the full model in the denominator. The information involved in the test procedure is often summarized in an Analysis of Variance (ANOVA) table as given in Table 2.1 for this particular case. The difference in the numerator is often called the regression sum of squares (SSR).

In terms of Table 2.1, it is conventional to write the F -statistic (2.11) as the ratio of the two mean squares, obtaining

$$F_{q-r, n-q} = \frac{MSR}{MSE}, \quad (2.12)$$

where MSR , the *mean squared regression*, is the numerator of (2.11). A special case of interest is $r = 1$ and $z_{t1} \equiv 1$, when the model in (2.10) becomes

$$x_t = \beta_1 + w_t,$$

and we may measure the proportion of variation accounted for by the other variables using

$$R^2 = \frac{SSE_1 - SSE}{SSE_1}, \quad (2.13)$$

where the residual sum of squares under the reduced model

$$SSE_1 = \sum_{t=1}^n (x_t - \bar{x})^2, \quad (2.14)$$

in this case is just the sum of squared deviations from the mean \bar{x} . The measure R^2 is also the *squared multiple correlation* between x_t and the variables $z_{t2}, z_{t3}, \dots, z_{tq}$.

The techniques discussed in the previous paragraph can be used to test various models against one another using the F test given in (2.11), (2.12), and the ANOVA table. These tests have been used in the past in a stepwise manner, where variables are added or deleted when the values from the F -test either exceed or fail to exceed some predetermined levels. The procedure, called stepwise multiple regression, is useful in arriving at a set of useful variables. An alternative is to focus on a procedure for model selection that does not proceed sequentially, but simply evaluates each model on its own merits. Suppose we consider a normal regression model with k coefficients and denote the maximum likelihood estimator for the variance as

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n}, \quad (2.15)$$

where SSE_k denotes the residual sum of squares under the model with k regression coefficients. Then, Akaike (1969, 1973, 1974) suggested measuring the goodness of fit for this particular model by balancing the error of the fit against the number of parameters in the model; we define the following.

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \quad (2.16)$$

where $\hat{\sigma}_k^2$ is given by (2.15) and k is the number of parameters in the model.

The value of k yielding the minimum AIC specifies the best model.¹ The idea is roughly that minimizing $\hat{\sigma}_k^2$ would be a reasonable objective, except that it decreases monotonically as k increases. Therefore, we ought to penalize the error variance by a term proportional to the number of parameters. The choice for the penalty term given by (2.16) is not the only one, and a considerable literature is available advocating different penalty terms. A corrected form, suggested by Sugiura (1978), and expanded by Hurvich and Tsai (1989), can be based on small-sample distributional results for the linear regression model. The corrected form is defined as follows.

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \log \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}, \quad (2.17)$$

where $\hat{\sigma}_k^2$ is given by (2.15), k is the number of parameters in the model, and n is the sample size.

We may also derive a correction term based on Bayesian arguments, as in Schwarz (1978), which leads to the following.

¹ Formally, AIC is defined as $-2 \log L_k + 2k$ where L_k is the maximum value of the likelihood and k is the number of parameters in the model. For the normal regression problem, AIC can be reduced to the form given by (2.16).

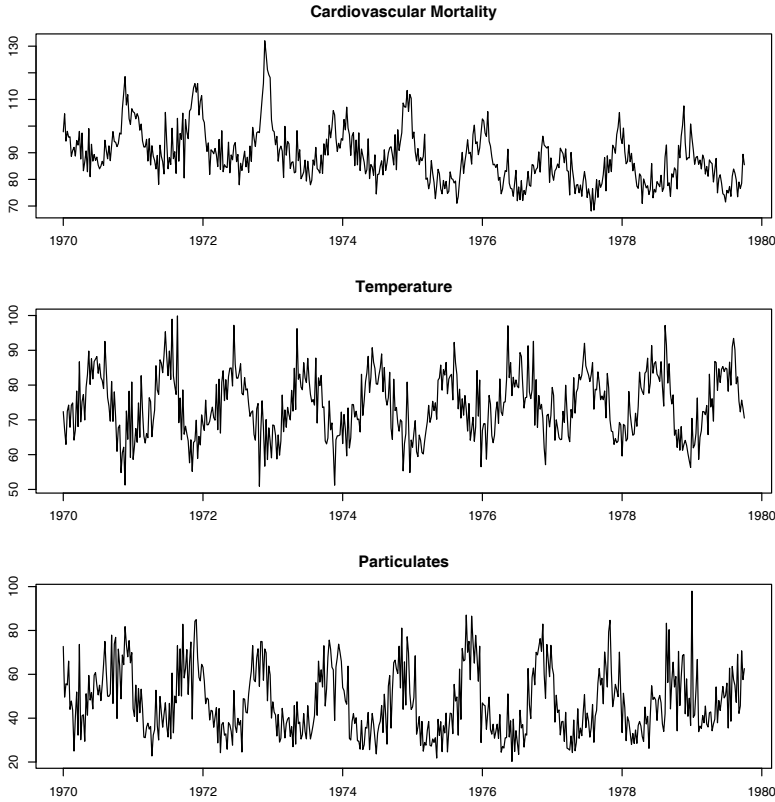


Fig. 2.2. Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970–1979.

Definition 2.3 Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}, \quad (2.18)$$

using the same notation as in *Definition 2.2*.

BIC is also called the Schwarz Information Criterion (SIC); see also Rissanen (1978) for an approach yielding the same statistic based on a minimum description length argument. Various simulation studies have tended to verify that BIC does well at getting the correct order in large samples, whereas AICc tends to be superior in smaller samples where the relative number of parameters is large; see McQuarrie and Tsai (1998) for detailed comparisons. In fitting regression models, two measures that have been used in the past are adjusted R-squared, which is essentially s_w^2 , and Mallows C_p , Mallows (1973), which we do not consider in this context.

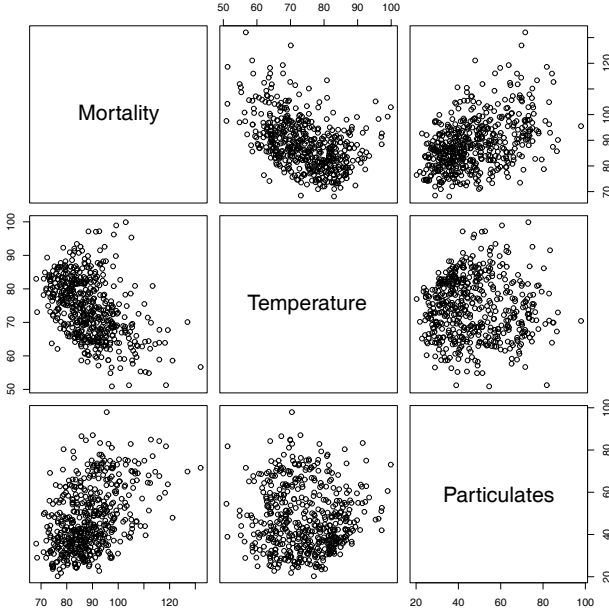


Fig. 2.3. Scatterplot matrix showing plausible relations between mortality, temperature, and pollution.

Example 2.2 Pollution, Temperature and Mortality

The data shown in [Figure 2.2](#) are extracted series from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on weekly mortality in Los Angeles County. Note the strong seasonal components in all of the series, corresponding to winter-summer variations and the downward trend in the cardiovascular mortality over the 10-year period.

A scatterplot matrix, shown in [Figure 2.3](#), indicates a possible linear relation between mortality and the pollutant particulates and a possible relation to temperature. Note the curvilinear shape of the temperature mortality curve, indicating that higher temperatures as well as lower temperatures are associated with increases in cardiovascular mortality.

Based on the scatterplot matrix, we entertain, tentatively, four models where M_t denotes cardiovascular mortality, T_t denotes temperature and P_t denotes the particulate levels. They are

$$M_t = \beta_1 + \beta_2 t + w_t \quad (2.19)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + w_t \quad (2.20)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + \beta_4(T_t - T.)^2 + w_t \quad (2.21)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + \beta_4(T_t - T.)^2 + \beta_5 P_t + w_t \quad (2.22)$$

where we adjust temperature for its mean, $T. = 74.6$, to avoid scaling problems. It is clear that (2.19) is a trend only model, (2.20) is linear temperature,

Table 2.2. Summary Statistics for Mortality Models

Model	k	SSE	df	MSE	R^2	AIC	BIC
(2.19)	2	40,020	506	79.0	.21	5.38	5.40
(2.20)	3	31,413	505	62.2	.38	5.14	5.17
(2.21)	4	27,985	504	55.5	.45	5.03	5.07
(2.22)	5	20,508	503	40.8	.60	4.72	4.77

(2.21) is curvilinear temperature and (2.22) is curvilinear temperature and pollution. We summarize some of the statistics given for this particular case in Table 2.2. The values of R^2 were computed by noting that $SSE_1 = 50,687$ using (2.14).

We note that each model does substantially better than the one before it and that the model including temperature, temperature squared, and particulates does the best, accounting for some 60% of the variability and with the best value for AIC and BIC (because of the large sample size, AIC and AICc are nearly the same). Note that one can compare any two models using the residual sums of squares and (2.11). Hence, a model with only trend could be compared to the full model using $q = 5, r = 2, n = 508$, so

$$F_{3,503} = \frac{(40,020 - 20,508)/3}{20,508/503} = 160,$$

which exceeds $F_{3,503}(.001) = 5.51$. We obtain the best prediction model,

$$\begin{aligned}\widehat{M}_t = & 81.59 - .027_{(.002)}t - .473_{(.032)}(T_t - 74.6) \\ & + .023_{(.003)}(T_t - 74.6)^2 + .255_{(.019)}P_t,\end{aligned}$$

for mortality, where the standard errors, computed from (2.6)-(2.8), are given in parentheses. As expected, a negative trend is present in time as well as a negative coefficient for adjusted temperature. The quadratic effect of temperature can clearly be seen in the scatterplots of Figure 2.3. Pollution weights positively and can be interpreted as the incremental contribution to daily deaths per unit of particulate pollution. It would still be essential to check the residuals $\widehat{w}_t = M_t - \widehat{M}_t$ for autocorrelation (of which there is a substantial amount), but we defer this question to §3.9 when we discuss regression with correlated errors.

Below is the R code to plot the series, display the scatterplot matrix, fit the final regression model (2.22), and compute the corresponding values of AIC and BIC. Finally, the use of `na.action` in `lm()` is to retain the time series attributes for the residuals and fitted values.

```
par(mfrow=c(3,1))
plot(cmort, main="Cardiovascular Mortality", xlab="", ylab="")
plot(tempr, main="Temperature", xlab="", ylab="")
plot(part, main="Particulates", xlab="", ylab="")
dev.new() # open a new graphic device for the scatterplot matrix
```

```

pairs(cbind(Mortality=cmort, Temperature=tempr, Particulates=part))
temp = tempr-mean(tempr) # center temperature
temp2 = temp^2
trend = time(cmort) # time
fit = lm(cmort~ trend + temp + temp2 + part, na.action=NULL)
summary(fit) # regression results
summary(aov(fit)) # ANOVA table (compare to next line)
summary(aov(lm(cmort~cbind(trend, temp, temp2, part)))) # Table 2.1
num = length(cmort) # sample size
AIC(fit)/num - log(2*pi) # AIC as in Def 2.1
BIC(fit)/num - log(2*pi) # BIC as in Def 2.3

```

As previously mentioned, it is possible to include lagged variables in time series regression models and we will continue to discuss this type of problem throughout the text. This concept is explored further in [Problem 2.2](#). The following is a simple example of lagged regression.

Example 2.3 Regression With Lagged Variables

In [Example 1.26](#), we discovered that the Southern Oscillation Index (SOI) measured at time $t - 6$ months is associated with the Recruitment series at time t , indicating that the SOI leads the Recruitment series by six months. Although there is evidence that the relationship is not linear (this is discussed further in [Example 2.7](#)), we may consider the following regression,

$$R_t = \beta_1 + \beta_2 S_{t-6} + w_t, \quad (2.23)$$

where R_t denotes Recruitment for month t and S_{t-6} denotes SOI six months prior. Assuming the w_t sequence is white, the fitted model is

$$\hat{R}_t = 65.79 - 44.28_{(2.78)} S_{t-6} \quad (2.24)$$

with $\hat{\sigma}_w = 22.5$ on 445 degrees of freedom. This result indicates the strong predictive ability of SOI for Recruitment six months in advance. Of course, it is still essential to check the the model assumptions, but again we defer this until later.

Performing lagged regression in R is a little difficult because the series must be aligned prior to running the regression. The easiest way to do this is to create an object that we call `fish` using `ts.intersect`, which aligns the lagged series.

```

fish = ts.intersect( rec, soiL6=lag(soi,-6) )
summary(fit1 <- lm(rec~ soiL6, data=fish, na.action=NULL))

```

The headache of aligning the lagged series can be avoided by using the R package `dynlm`, which must be downloaded and installed.

```

require(dynlm)
summary(fit2 <- dynlm(rec~ L(soi,6)))

```

In the `dynlm` example, `fit2` is similar to a `lm` object, but the time series attributes are retained without any additional commands.

2.2 Exploratory Data Analysis

In general, it is necessary for time series data to be stationary so averaging lagged products over time, as in the previous section, will be a sensible thing to do. With time series data, it is the dependence between the values of the series that is important to measure; we must, at least, be able to estimate autocorrelations with precision. It would be difficult to measure that dependence if the dependence structure is not regular or is changing at every time point. Hence, to achieve any meaningful statistical analysis of time series data, it will be crucial that, if nothing else, the mean and the autocovariance functions satisfy the conditions of stationarity (for at least some reasonable stretch of time) stated in [Definition 1.7](#). Often, this is not the case, and we will mention some methods in this section for playing down the effects of nonstationarity so the stationary properties of the series may be studied.

A number of our examples came from clearly nonstationary series. The Johnson & Johnson series in [Figure 1.1](#) has a mean that increases exponentially over time, and the increase in the magnitude of the fluctuations around this trend causes changes in the covariance function; the variance of the process, for example, clearly increases as one progresses over the length of the series. Also, the global temperature series shown in [Figure 1.3](#) contains some evidence of a trend over time; human-induced global warming advocates seize on this as empirical evidence to advance their hypothesis that temperatures are increasing.

Perhaps the easiest form of nonstationarity to work with is the trend stationary model wherein the process has stationary behavior around a trend. We may write this type of model as

$$x_t = \mu_t + y_t \quad (2.25)$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process. Quite often, strong trend, μ_t , will obscure the behavior of the stationary process, y_t , as we shall see in numerous examples. Hence, there is some advantage to removing the trend as a first step in an exploratory analysis of such time series. The steps involved are to obtain a reasonable estimate of the trend component, say $\hat{\mu}_t$, and then work with the residuals

$$\hat{y}_t = x_t - \hat{\mu}_t. \quad (2.26)$$

Consider the following example.

Example 2.4 Detrending Global Temperature

Here we suppose the model is of the form of [\(2.25\)](#),

$$x_t = \mu_t + y_t,$$

where, as we suggested in the analysis of the global temperature data presented in [Example 2.1](#), a straight line might be a reasonable model for the trend, i.e.,

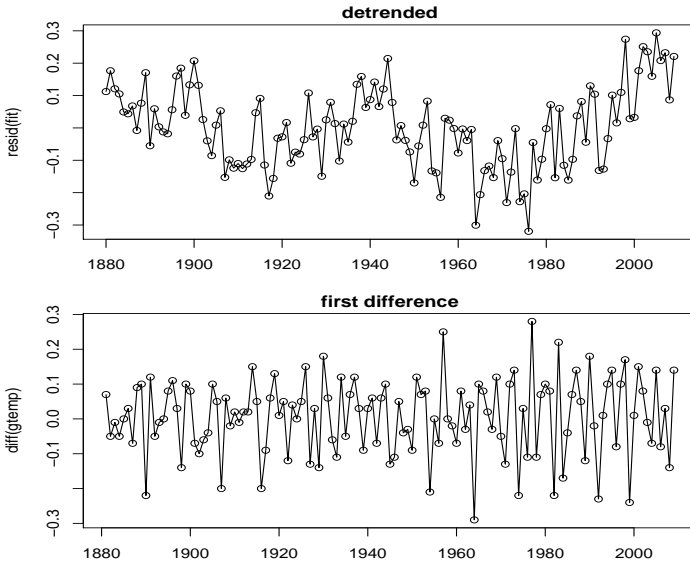


Fig. 2.4. Detrended (top) and differenced (bottom) global temperature series. The original data are shown in [Figure 1.3](#) and [Figure 2.1](#).

$$\mu_t = \beta_1 + \beta_2 t.$$

In that example, we estimated the trend using ordinary least squares² and found

$$\hat{\mu}_t = -11.2 + .006 t.$$

[Figure 2.1](#) shows the data with the estimated trend line superimposed. To obtain the detrended series we simply subtract $\hat{\mu}_t$ from the observations, x_t , to obtain the detrended series

$$\hat{y}_t = x_t + 11.2 - .006 t.$$

The top graph of [Figure 2.4](#) shows the detrended series. [Figure 2.5](#) shows the ACF of the original data (top panel) as well as the ACF of the detrended data (middle panel).

To detrend in the series in R, use the following commands. We also show how to difference and plot the differenced data; we discuss differencing after this example. In addition, we show how to generate the sample ACFs displayed in [Figure 2.5](#).

² Because the error term, y_t , is not assumed to be iid, the reader may feel that weighted least squares is called for in this case. The problem is, we do not know the behavior of y_t and that is precisely what we are trying to assess at this stage. An important result by Grenander and Rosenblatt (1957, Ch 7), however, is that under mild conditions, for polynomial or periodic regression, when the sample size is large, ordinary least squares is equivalent to weighted least squares in that the standard errors of the estimates are the approximately same.

```

fit = lm(gtemp~time(gtemp), na.action=NULL) # regress gtemp on time
par(mfrow=c(2,1))
plot(resid(fit), type="o", main="detrended")
plot(diff(gtemp), type="o", main="first difference")
par(mfrow=c(3,1)) # plot ACFs
acf(gtemp, 48, main="gtemp")
acf(resid(fit), 48, main="detrended")
acf(diff(gtemp), 48, main="first difference")

```

In [Example 1.11](#) and the corresponding [Figure 1.11](#) we saw that a random walk might also be a good model for trend. That is, rather than modeling trend as fixed (as in [Example 2.4](#)), we might model trend as a stochastic component using the random walk with drift model,

$$\mu_t = \delta + \mu_{t-1} + w_t, \quad (2.27)$$

where w_t is white noise and is independent of y_t . If the appropriate model is [\(2.25\)](#), then differencing the data, x_t , yields a stationary process; that is,

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1}. \end{aligned} \quad (2.28)$$

It is easy to show $z_t = y_t - y_{t-1}$ is stationary using [footnote 2](#) of Chapter 1 on page [15](#). That is, because y_t is stationary,

$$\begin{aligned} \gamma_z(h) &= \text{cov}(z_{t+h}, z_t) = \text{cov}(y_{t+h} - y_{t+h-1}, y_t - y_{t-1}) \\ &= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1) \end{aligned}$$

is independent of time; we leave it as an exercise ([Problem 2.5](#)) to show that $x_t - x_{t-1}$ in [\(2.28\)](#) is stationary.

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. One disadvantage, however, is that differencing does not yield an estimate of the stationary process y_t as can be seen in [\(2.28\)](#). If an estimate of y_t is essential, then detrending may be more appropriate. If the goal is to coerce the data to stationarity, then differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, as in [Example 2.4](#). That is, e.g., if $\mu_t = \beta_1 + \beta_2 t$ in the model [\(2.25\)](#), differencing the data produces stationarity (see [Problem 2.4](#)):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_2 + y_t - y_{t-1}.$$

Because differencing plays a central role in time series analysis, it receives its own notation. The first difference is denoted as

$$\nabla x_t = x_t - x_{t-1}. \quad (2.29)$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of [\(2.29\)](#), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we will use often in our discussion of ARIMA models in Chapter 3.

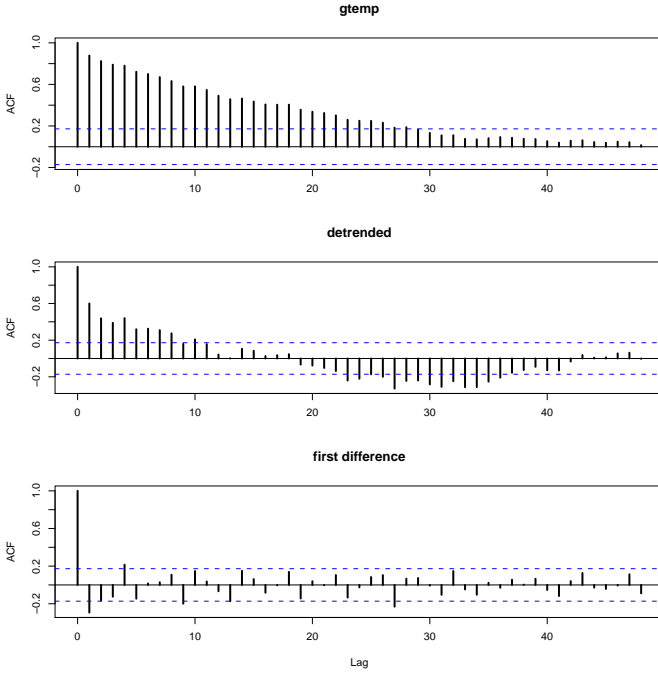


Fig. 2.5. Sample ACFs of the global temperature (top), and of the detrended (middle) and the differenced (bottom) series.

Definition 2.4 We define the **backshift operator** by

$$Bx_t = x_{t-1}$$

and extend it to powers $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$, and so on. Thus,

$$B^k x_t = x_{t-k}. \quad (2.30)$$

The idea of an inverse operator can also be given if we require $B^{-1}B = 1$, so that

$$x_t = B^{-1}Bx_t = B^{-1}x_{t-1}.$$

That is, B^{-1} is the *forward-shift operator*. In addition, it is clear that we may rewrite (2.29) as

$$\nabla x_t = (1 - B)x_t, \quad (2.31)$$

and we may extend the notion further. For example, the second difference becomes

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2} \quad (2.32)$$

by the linearity of the operator. To check, just take the difference of the first difference $\nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$.

Definition 2.5 Differences of order d are defined as

$$\nabla^d = (1 - B)^d, \quad (2.33)$$

where we may expand the operator $(1 - B)^d$ algebraically to evaluate for higher integer values of d . When $d = 1$, we drop it from the notation.

The first difference (2.29) is an example of a linear filter applied to eliminate a trend. Other filters, formed by averaging values near x_t , can produce adjusted series that eliminate other kinds of unwanted fluctuations, as in Chapter 3. The differencing technique is an important component of the ARIMA model discussed in Chapter 3.

Example 2.5 Differencing Global Temperature

The first difference of the global temperature series, also shown in Figure 2.4, produces different results than removing trend by detrending via regression. For example, the differenced series does not contain the long middle cycle we observe in the detrended series. The ACF of this series is also shown in Figure 2.5. In this case it appears that the differenced process shows minimal autocorrelation, which may imply the global temperature series is nearly a random walk with drift. It is interesting to note that if the series is a random walk with drift, the mean of the differenced series, which is an estimate of the drift, is about .0066 (but with a large standard error):

```
mean(diff(gtemp)) # = 0.00659 (drift)
sd(diff(gtemp))/sqrt(length(diff(gtemp))) # = 0.00966 (SE)
```

Often, obvious aberrations are present that can contribute nonstationary as well as nonlinear behavior in observed time series. In such cases, transformations may be useful to equalize the variability over the length of a single series. A particularly useful transformation is

$$y_t = \log x_t, \quad (2.34)$$

which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger. Other possibilities are power transformations in the Box–Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log x_t & \lambda = 0. \end{cases} \quad (2.35)$$

Methods for choosing the power λ are available (see Johnson and Wichern, 1992, §4.7) but we do not pursue them here. Often, transformations are also used to improve the approximation to normality or to improve linearity in predicting the value of one series from another.

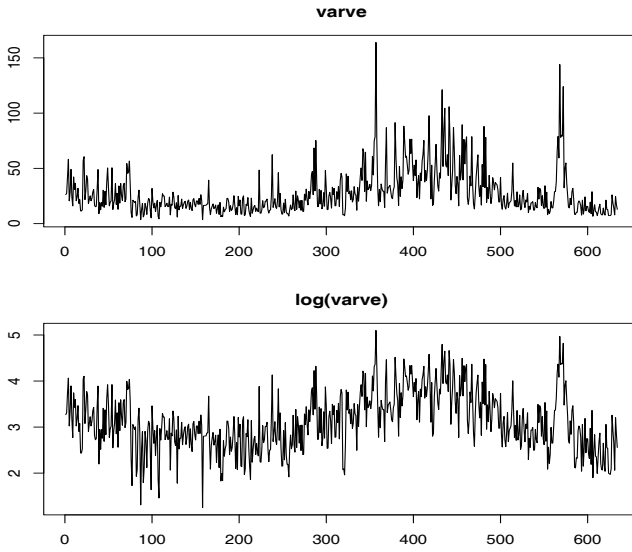


Fig. 2.6. Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with log transformed thicknesses (bottom).

Example 2.6 Paleoclimatic Glacial Varves

Melting glaciers deposit yearly layers of sand and silt during the spring melting seasons, which can be reconstructed yearly over a period ranging from the time deglaciation began in New England (about 12,600 years ago) to the time it ended (about 6,000 years ago). Such sedimentary deposits, called varves, can be used as proxies for paleoclimatic parameters, such as temperature, because, in a warm year, more sand and silt are deposited from the receding glacier. [Figure 2.6](#) shows the thicknesses of the yearly varves collected from one location in Massachusetts for 634 years, beginning 11,834 years ago. For further information, see Shumway and Verosub (1992). Because the variation in thicknesses increases in proportion to the amount deposited, a logarithmic transformation could remove the nonstationarity observable in the variance as a function of time. [Figure 2.6](#) shows the original and transformed varves, and it is clear that this improvement has occurred. We may also plot the histogram of the original and transformed data, as in [Problem 2.6](#), to argue that the approximation to normality is improved. The ordinary first differences (2.31) are also computed in [Problem 2.6](#), and we note that the first differences have a significant negative correlation at lag $h = 1$. Later, in Chapter 5, we will show that perhaps the varve series has long memory and will propose using fractional differencing.

[Figure 2.6](#) was generated in R as follows:

```
par(mfrow=c(2,1))
plot(varve, main="varve", ylab="")
plot(log(varve), main="log(varve)", ylab="")
```

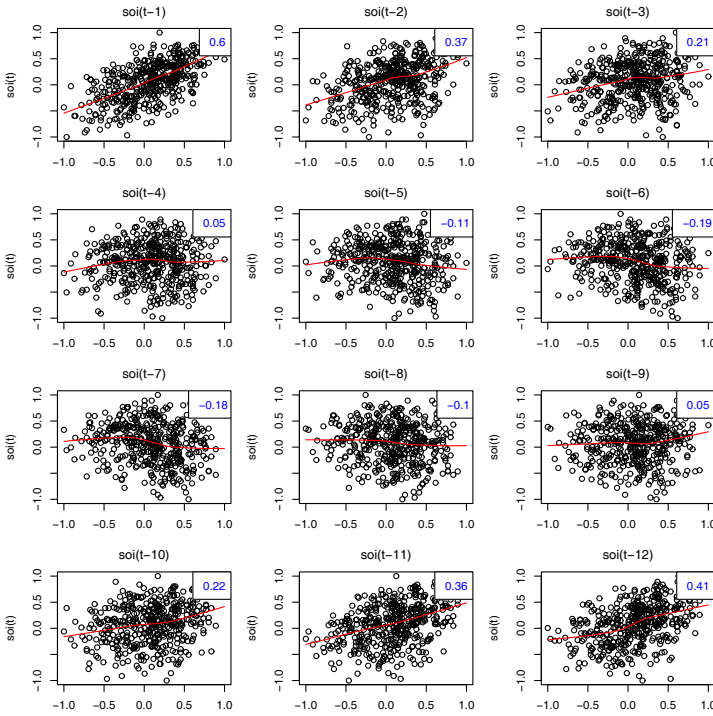


Fig. 2.7. Scatterplot matrix relating current SOI values, S_t , to past SOI values, S_{t-h} , at lags $h = 1, 2, \dots, 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.

Next, we consider another preliminary data processing technique that is used for the purpose of visualizing the relations between series at different lags, namely, scatterplot matrices. In the definition of the ACF, we are essentially interested in relations between x_t and x_{t-h} ; the autocorrelation function tells us whether a substantial linear relation exists between the series and its own lagged values. The ACF gives a profile of the linear correlation at all possible lags and shows which values of h lead to the best predictability. The restriction of this idea to linear predictability, however, may mask a possible nonlinear relation between current values, x_t , and past values, x_{t-h} . This idea extends to two series where one may be interested in examining scatterplots of y_t versus x_{t-h} .

Example 2.7 Scatterplot Matrices, SOI and Recruitment

To check for nonlinear relations of this form, it is convenient to display a lagged scatterplot matrix, as in Figure 2.7, that displays values of the SOI, S_t , on the vertical axis plotted against S_{t-h} on the horizontal axis. The sample autocorrelations are displayed in the upper right-hand corner and superimposed on the scatterplots are locally weighted scatterplot smoothing (lowess) lines that can be used to help discover any nonlinearities. We discuss

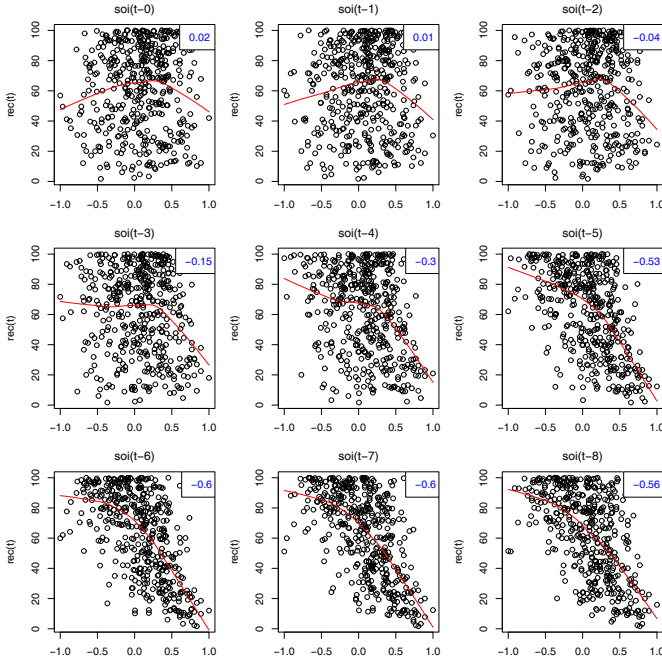


Fig. 2.8. Scatterplot matrix of the Recruitment series, R_t , on the vertical axis plotted against the SOI series, S_{t-h} , on the horizontal axis at lags $h = 0, 1, \dots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a lowess fit.

smoothing in the next section, but for now, think of lowess as a robust method for fitting nonlinear regression.

In [Figure 2.7](#), we notice that the lowess fits are approximately linear, so that the sample autocorrelations are meaningful. Also, we see strong positive linear relations at lags $h = 1, 2, 11, 12$, that is, between S_t and $S_{t-1}, S_{t-2}, S_{t-11}, S_{t-12}$, and a negative linear relation at lags $h = 6, 7$. These results match up well with peaks noticed in the ACF in [Figure 1.16](#).

Similarly, we might want to look at values of one series, say Recruitment, denoted R_t plotted against another series at various lags, say the SOI, S_{t-h} , to look for possible nonlinear relations between the two series. Because, for example, we might wish to predict the Recruitment series, R_t , from current or past values of the SOI series, S_{t-h} , for $h = 0, 1, 2, \dots$ it would be worthwhile to examine the scatterplot matrix. [Figure 2.8](#) shows the lagged scatterplot of the Recruitment series R_t on the vertical axis plotted against the SOI index S_{t-h} on the horizontal axis. In addition, the figure exhibits the sample cross-correlations as well as lowess fits.

[Figure 2.8](#) shows a fairly strong nonlinear relationship between Recruitment, R_t , and the SOI series at $S_{t-5}, S_{t-6}, S_{t-7}, S_{t-8}$, indicating the SOI series tends to lead the Recruitment series and the coefficients are negative,

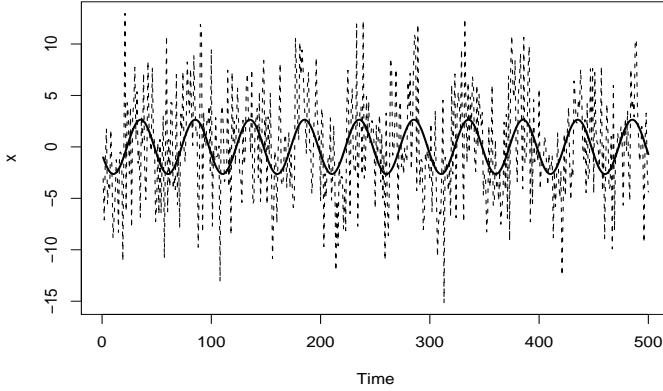


Fig. 2.9. Data generated by (2.36) [dashed line] with the fitted [solid] line, (2.38), superimposed.

implying that increases in the SOI lead to decreases in the Recruitment. The nonlinearity observed in the scatterplots (with the help of the superimposed lowess fits) indicate that the behavior between Recruitment and the SOI is different for positive values of SOI than for negative values of SOI.

Simple scatterplot matrices for one series can be obtained in R using the `lag.plot` command. Figure 2.7 and Figure 2.8 may be reproduced using the following scripts provided with the text (see Appendix R for details):

```
lag1.plot(soi, 12)      # Figure 2.7
lag2.plot(soi, rec, 8)  # Figure 2.8
```

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis and the periodogram; this material may be thought of as an introduction to spectral analysis, which we discuss in detail in Chapter 4. In Example 1.12, we briefly discussed the problem of identifying cyclic or periodic signals in time series. A number of the time series we have seen so far exhibit periodic behavior. For example, the data from the pollution study example shown in Figure 2.2 exhibit strong yearly cycles. Also, the Johnson & Johnson data shown in Figure 1.1 make one cycle every year (four quarters) on top of an increasing trend and the speech data in Figure 1.3 is highly repetitive. The monthly SOI and Recruitment series in Figure 1.7 show strong yearly cycles, but hidden in the series are clues to the El Niño cycle.

Example 2.8 Using Regression to Discover a Signal in Noise

In Example 1.12, we generated $n = 500$ observations from the model

$$x_t = A \cos(2\pi\omega t + \phi) + w_t, \quad (2.36)$$

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.12 on page 12. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown

parameters. In this case the parameters appear in (2.36) in a nonlinear way, so we use a trigonometric identity³ and write

$$A \cos(2\pi\omega t + \phi) = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t),$$

where $\beta_1 = A \cos(\phi)$ and $\beta_2 = -A \sin(\phi)$. Now the model (2.36) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1 \cos(2\pi t/50) + \beta_2 \sin(2\pi t/50) + w_t. \quad (2.37)$$

Using linear regression on the generated data, the fitted model is

$$\hat{x}_t = -.71_{(.30)} \cos(2\pi t/50) - 2.55_{(.30)} \sin(2\pi t/50) \quad (2.38)$$

with $\hat{\sigma}_w = 4.68$, where the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are $\beta_1 = 2 \cos(.6\pi) = -.62$ and $\beta_2 = -2 \sin(.6\pi) = -1.90$. Because the parameter estimates are significant and close to the actual values, it is clear that we are able to detect the signal in the noise using regression, even though the signal appears to be obscured by the noise in the bottom panel of Figure 1.12. Figure 2.9 shows data generated by (2.36) with the fitted line, (2.38), superimposed.

To reproduce the analysis and Figure 2.9 in R, use the following commands:

```
set.seed(1000) # so you can reproduce these results
x = 2*cos(2*pi*1:500/50 + .6*pi) + rnorm(500,0,5)
z1 = cos(2*pi*1:500/50); z2 = sin(2*pi*1:500/50)
summary(fit <- lm(x~0+z1+z2)) # zero to exclude the intercept
plot.ts(x, lty="dashed"); lines(fitted(fit), lwd=2)
```

We will discuss this and related approaches in more detail in Chapter 4.

2.3 Smoothing in the Time Series Context

In §1.4, we introduced the concept of smoothing a time series, and in Example 1.9, we discussed using a moving average to smooth white noise. This method is useful in discovering certain traits in a time series, such as long-term trend and seasonal components. In particular, if x_t represents the observations, then

$$m_t = \sum_{j=-k}^k a_j x_{t-j}, \quad (2.39)$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^k a_j = 1$ is a symmetric moving average of the data.

³ $\cos(\alpha \pm \beta) = \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta)$.

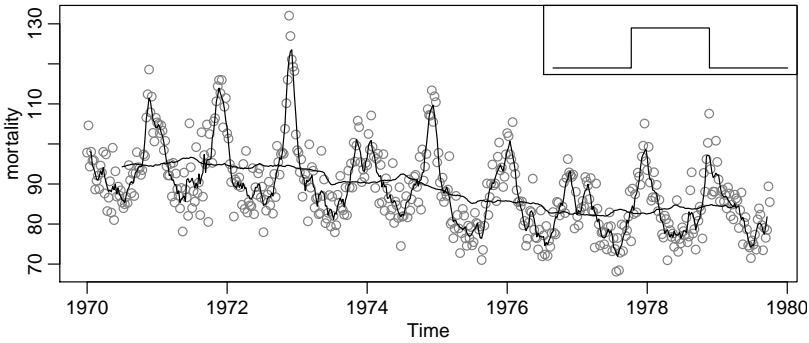


Fig. 2.10. The weekly cardiovascular mortality series discussed in [Example 2.2](#) smoothed using a five-week moving average and a 53-week moving average. The insert shows the shape of the moving average (“boxcar”) kernel described in (2.41).

Example 2.9 Moving Average Smoother

For example, [Figure 2.10](#) shows the weekly mortality series discussed in [Example 2.2](#), a five-point moving average (which is essentially a monthly average with $k = 2$) that helps bring out the seasonal component and a 53-point moving average (which is essentially a yearly average with $k = 26$) that helps bring out the (negative) trend in cardiovascular mortality. Often, the series with the seasonal removed is called the *seasonally adjusted* data. In this example, the weights, $a_{-k}, \dots, a_0, \dots, a_k$, we used were all the same, and equal to $1/(2k + 1)$.⁴

To reproduce [Figure 2.10](#) in R:

```
ma5 = filter(cmort, sides=2, rep(1,5)/5)
ma53 = filter(cmort, sides=2, rep(1,53)/53)
plot(cmort, type="p", ylab="mortality")
lines(ma5); lines(ma53)
```

Although the moving average smoother does a good job in highlighting the trend and seasonal components of the data, it might be considered too choppy. We can obtain a smoother fit using the normal distribution for the weights, instead of boxcar-type weights of (2.39).

Example 2.10 Kernel Smoothing

Kernel smoothing is a moving average smoother that uses a weight function, or kernel, to average the observations. [Figure 2.11](#) shows kernel smoothing of the mortality series, where m_t is now

$$m_t = \sum_{i=1}^n w_i(t)x_i, \quad (2.40)$$

⁴ Sometimes, the end weights, a_{-k} and a_k are set equal to half the value of the other weights.

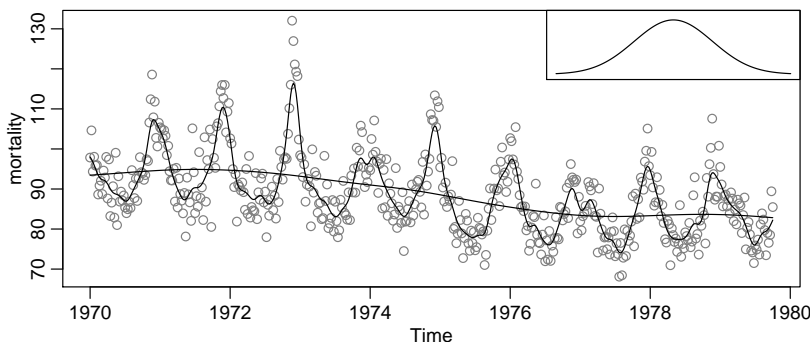


Fig. 2.11. Kernel smoothers of the mortality data. The insert shows the shape of the normal kernel.

where

$$w_i(t) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^n K\left(\frac{t-j}{b}\right) \quad (2.41)$$

are the weights and $K(\cdot)$ is a kernel function. This estimator, which was originally explored by Parzen (1962) and Rosenblatt (1956b), is often called the Nadaraya–Watson estimator (Watson, 1966); typically, the normal kernel, $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, is used.

To implement this in R, use the `ksmooth` function; a rule of thumb for choosing the bandwidth is to divide it by 2. The wider the bandwidth, b , the smoother the result. In Figure 2.11, for the seasonal component the value of b is $b = 8/52$ (roughly $b/2 = 4$ -week averages), and for the trend component, $b = 104/52 = 2$ (roughly yearly averages).

Figure 2.11 can be reproduced in R (or S-PLUS) as follows.

```
plot(cmort, type="p", ylab="mortality")
lines(ksmooth(time(cmort), cmort, "normal", bandwidth=8/52))
lines(ksmooth(time(cmort), cmort, "normal", bandwidth=2))
```

Example 2.11 Lowess

Another approach to smoothing a time plot is nearest neighbor regression. The technique is based on k -nearest neighbors regression, wherein one uses only the data $\{x_{t-k/2}, \dots, x_t, \dots, x_{t+k/2}\}$ to predict x_t via regression, and then sets $m_t = \hat{x}_t$.

Lowess is a method of smoothing that is rather complex, but the basic idea is close to nearest neighbor regression. Figure 2.12 shows smoothing of mortality using the R or S-PLUS function `lowess` (see Cleveland, 1979). First, a certain proportion of nearest neighbors to x_t are included in a weighting scheme; values closer to x_t in time get more weight. Then, a robust weighted regression is used to predict x_t and obtain the smoothed values m_t . The larger the fraction of nearest neighbors included, the smoother the fit will be. In Figure 2.12, the smoother uses about two-thirds of the data (the default) to obtain an estimate of the trend component, and the seasonal component uses 2% of the data.

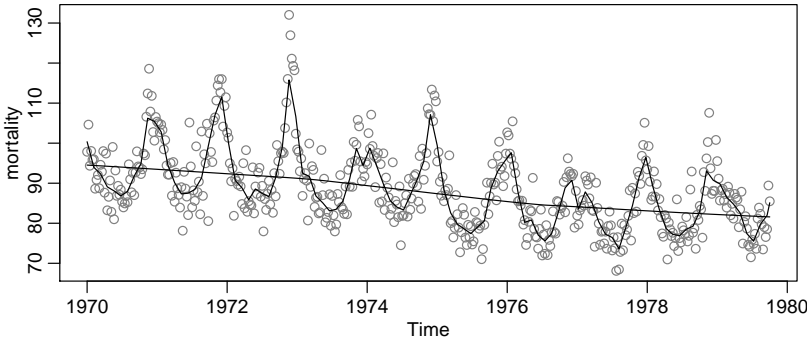


Fig. 2.12. Locally weighted scatterplot smoothers (*lowess*) of the mortality data.

Figure 2.12 can be reproduced in R or S-PLUS as follows.

```
plot(cmort, type="p", ylab="mortality")
lines(lowess(cmort))
lines(lowess(cmort, f=.02))
```

Example 2.12 Smoothing One Series as a Function of Another

In addition to smoothing time plots, smoothing techniques can be applied to smoothing a time series as a function of another time series. In this example, we smooth the scatterplot of two contemporaneously measured time series, mortality as a function of temperature. In [Example 2.2](#), we discovered a nonlinear relationship between mortality and temperature. Continuing along these lines, [Figure 2.13](#) shows scatterplots of mortality, M_t , and temperature, T_t , along with M_t smoothed as a function of T_t using lowess. Note that mortality increases at extreme temperatures, but in an asymmetric way; mortality is higher at colder temperatures than at hotter temperatures. The minimum mortality rate seems to occur at approximately 80° F.

Figure 2.13 can be reproduced in R as follows.

```
plot(tempr, cmort, xlab="Temperature", ylab="Mortality")
lines(lowess(tempr, cmort))
```

Problems

2.1 For the Johnson & Johnson data, say y_t , shown in [Figure 1.1](#), let $x_t = \log(y_t)$.

(a) Fit the regression model

$$x_t = \beta t + \alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t) + w_t$$

where $Q_i(t) = 1$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and zero otherwise. The $Q_i(t)$'s are called indicator variables. We will assume for

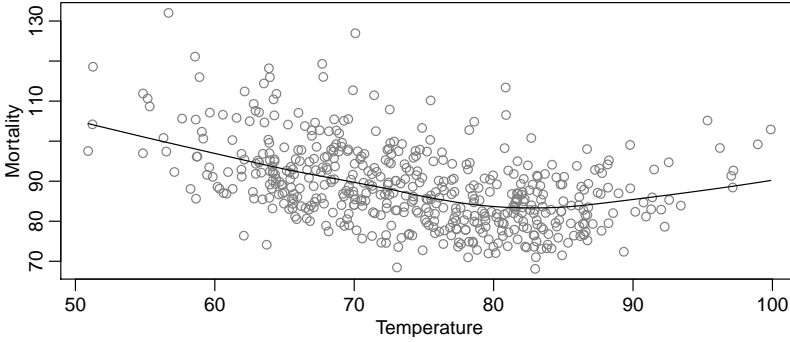


Fig. 2.13. Smooth of mortality as a function of temperature using lowess.

now that w_t is a Gaussian white noise sequence. What is the interpretation of the parameters β , α_1 , α_2 , α_3 , and α_4 ? (*Detailed code is given in Appendix R on page 177.*)

- (b) What happens if you include an intercept term in the model in (a)?
- (c) Graph the data, x_t , and superimpose the fitted values, say \hat{x}_t , on the graph. Examine the residuals, $x_t - \hat{x}_t$, and state your conclusions. Does it appear that the model fits the data well (do the residuals look white)?

2.2 For the mortality data examined in [Example 2.2](#):

- (a) Add another component to the regression in (2.22) that accounts for the particulate count four weeks prior; that is, add P_{t-4} to the regression in (2.22). State your conclusion.
- (b) Using AIC and BIC, is the model in (a) an improvement over the final model in [Example 2.2](#)?

2.3 (a) Generate *four* series that are random walk with drift, (1.4), of length $n = 100$ with $\delta = .01$ and $\sigma_w = 1$. Call the data x_t for $t = 1, \dots, 100$. Fit the regression $x_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01 t$) and the fitted line, $\hat{x}_t = \hat{\beta} t$, on the same graph.

- (b) Generate *four* series of length $n = 100$ that are linear trend plus noise, say $y_t = .01 t + w_t$, where t and w_t are as in part (a). Fit the regression $y_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01 t$) and the fitted line, $\hat{y}_t = \hat{\beta} t$, on the same graph.
- (c) Comment (what did you learn from this assignment).

The following R code may be useful for doing (a) and (b) at the same time.

```
par(mfrow=c(4,2), mar=c(2.5,2.5,0,0)+.5, mgp=c(1.6,.6,0)) # set up
for (i in 1:4){
  x = ts(cumsum(rnorm(100,.01,1))) # data
  y = ts(.01*1:100 + rnorm(100))
  regx = lm(x~0+time(x), na.action=NULL) # regressions
  regy = lm(y~0+time(y), na.action=NULL)
```

```

plot(x)                                     # plots
  lines(.01*time(x), col="red", lty="dashed") # true mean
  abline(regx, col="blue")                  # fitted line
plot(y)
  lines(.01*time(y), col="red", lty="dashed")
  abline(regy, col="blue")
}

```

2.4 Consider a process consisting of a quadratic trend with an additive noise term consisting of independent random variables w_t with zero means and variances σ_w^2 , that is,

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + w_t,$$

where $\beta_0, \beta_1, \beta_2$ are fixed constants.

- (a) Prove x_t is nonstationary.
- (b) If $\beta_2 = 0$ (that is, $x_t = \beta_0 + \beta_1 t + w_t$), let $y_t = \nabla x_t = x_t - x_{t-1}$ and show that y_t is stationary by finding its mean function and autocovariance function.
- (c) If $\beta_2 \neq 0$, show that the first difference series ∇x_t is not stationary, but $\nabla^2 x_t$ is stationary.

2.5 Show (2.28) is stationary.

2.6 The glacial varve record plotted in Figure 2.6 exhibits some nonstationarity that can be improved by transforming to logarithms and some additional nonstationarity that can be corrected by differencing the logarithms.

- (a) Argue that the glacial varves series, say x_t , exhibits heteroscedasticity by computing the sample variance over the first half and the second half of the data. Argue that the transformation $y_t = \log x_t$ stabilizes the variance over the series. Plot the histograms of x_t and y_t to see whether the approximation to normality is improved by transforming the data.
- (b) Plot the series y_t . Do any time intervals, of the order 100 years, exist where one can observe behavior comparable to that observed in the global temperature records in Figure 1.3?
- (c) Examine the sample ACF of y_t and comment.
- (d) Compute the difference $u_t = y_t - y_{t-1}$, examine its time plot and sample ACF, and argue that differencing the logged varve data produces a reasonably stationary series. Can you think of a practical interpretation for u_t ? Hint: For $|p|$ close to zero, $\log(1+p) \approx p$; let $p = (y_t - y_{t-1})/y_{t-1}$.

2.7 Use the three different smoothing techniques described in Example 2.9, Example 2.10, and Example 2.11, to estimate the trend in the global temperature series displayed in Figure 1.3. Comment.

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.

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 moving average (MA) models. Often, these models are combined to form the autoregressive moving average (ARMA) model. 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). Seasonal data, such as the data discussed in [Example 1.1](#) and [Example 1.5](#) lead to seasonal autoregressive integrated moving average (SARIMA) models. The Box–Jenkins method for identifying a plausible models is given in this chapter along with techniques for parameter estimation and forecasting for these models.

3.2 Autoregressive Moving Average 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, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, 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 1.10](#) 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, \dots, x_n . For example, the lagged scatterplot matrix for the Southern Oscillation Index (SOI), shown in [Figure 2.7](#), gives a distinct indication that the values at lags 1, 2, and 12, are linearly associated with the current value. We will make this notion more precise in our discussion of forecasting.

Definition 3.1 *An autoregressive model of order p , abbreviated $\mathbf{AR}(p)$, is of the form*

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t, \quad (3.1)$$

where x_t is stationary, and $\phi_1, \phi_2, \dots, \phi_p$ 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),

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

or write

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t, \quad (3.2)$$

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

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}, \dots, x_{t-p} , are random components, whereas \mathbf{z}_t was assumed to be fixed. A useful form follows by using the backshift operator (2.30) to write the $\mathbf{AR}(p)$ model, (3.1), as

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)x_t = w_t, \quad (3.3)$$

or even more concisely as

$$\phi(B)x_t = w_t. \quad (3.4)$$

Example 3.1 The AR(1) Model

Consider the first-order model, AR(1), given by $x_t = \phi x_{t-1} + w_t$. Provided that $|\phi| < 1$ we can represent an AR(1) model as a linear process given by¹

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \quad (3.5)$$

Representation (3.5) is called the stationary solution of the model. In fact, by simple substitution,

$$\underbrace{\sum_{j=0}^{\infty} \phi^j w_{t-j}}_{x_t} = \phi \underbrace{\left(\sum_{k=0}^{\infty} \phi^k w_{t-1-k} \right)}_{x_{t-1}} + w_t.$$

Using (3.5), it is easy to see that the AR(1) process is stationary with mean

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

and autocovariance function,

$$\begin{aligned} \gamma(h) &= \text{cov}(x_{t+h}, x_t) = E \left[\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j} \right) \left(\sum_{k=0}^{\infty} \phi^k w_{t-k} \right) \right] \\ &= E \left[(w_{t+h} + \cdots + \phi^h w_t + \phi^{h+1} w_{t-1} + \cdots) (w_t + \phi w_{t-1} + \cdots) \right] \quad (3.6) \\ &= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j} \phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1 - \phi^2}, \quad h \geq 0. \end{aligned}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only exhibit the autocovariance function for $h \geq 0$. From (3.6), the ACF of an AR(1) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \quad h \geq 0. \quad (3.7)$$

Example 3.2 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

¹ Iterate backward, $x_t = \phi x_{t-1} + w_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t = \phi^2 x_{t-2} + \phi w_{t-1} + w_t = \cdots = \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j}$. If $|\phi| < 1$ and $\sup_t E(x_t^2) < \infty$, then $\lim_{k \rightarrow \infty} E \left(x_t - \sum_{j=0}^{k-1} \phi^j w_{t-j} \right)^2 = \lim_{k \rightarrow \infty} \phi^{2k} E(x_{t-k}^2) = 0$.

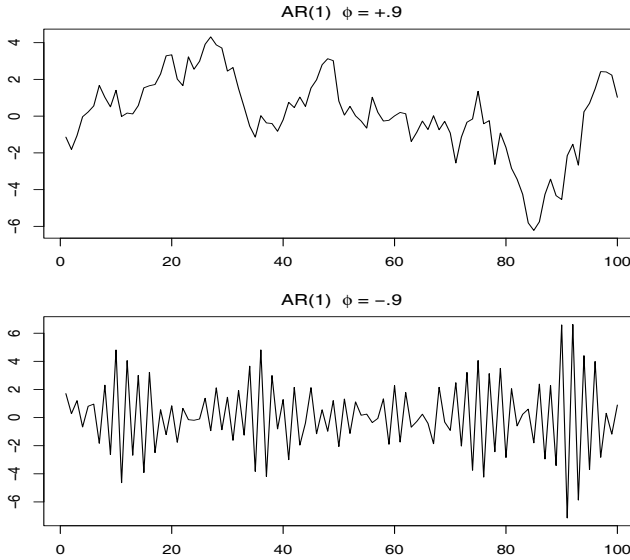


Fig. 3.1. Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).

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

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

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

Example 3.3 Explosive AR Models and Causality

For an AR(1), we have not addressed what happens if $|\phi| > 1$. Because $|\phi|^j$ increases without bound as $j \rightarrow \infty$, it is clear that (3.5) will not be valid in this case. We can, however, modify the approach to obtain a stationary model as follows. Write $x_{t+1} = \phi x_t + w_{t+1}$ and then rearrange it so that $x_t = \phi^{-1}x_{t+1} - \phi^{-1}w_{t+1}$. Now iterate *forward* k steps ahead in time

$$x_t = \phi^{-k}x_{t+k} - \sum_{j=1}^{k-1} \phi^{-j}w_{t+j}.$$

Because $|\phi|^{-1} < 1$, this result suggests the stationary future dependent AR(1) model

$$x_t = - \sum_{j=1}^{\infty} \phi^{-j} w_{t+j}. \quad (3.8)$$

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 this example, the process is stationary, but it is also future dependent, and not causal.

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

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \quad (3.9)$$

where there are q lags in the moving average and $\theta_1, \theta_2, \dots, \theta_q$ ($\theta_q \neq 0$) 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.

As in the AR(p) case, the MA(q) model may be written as

$$x_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t, \quad (3.10)$$

or more concisely as

$$x_t = \theta(B) w_t, \quad (3.11)$$

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \dots, \theta_q$.

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

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

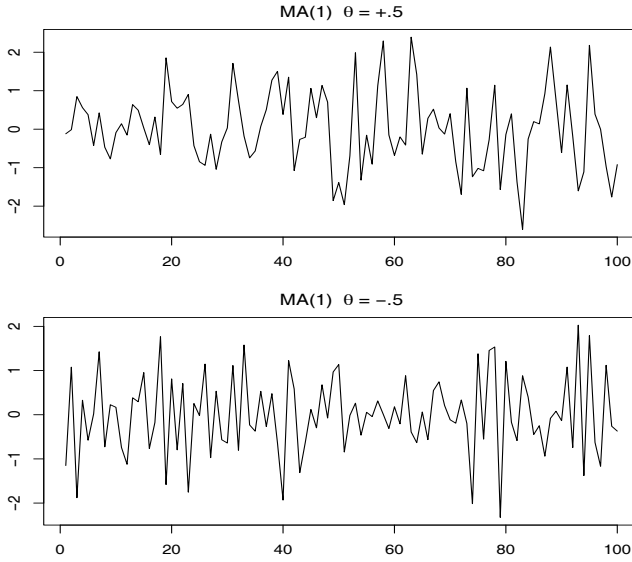


Fig. 3.2. Simulated MA(1) models: $\theta = .5$ (top); $\theta = -.5$ (bottom).

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

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}, \dots . Contrast this with the case of the AR(1) 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 where $\theta = .5$ is smoother than the series where $\theta = -.5$.

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

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

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

Definition 3.3 A time series $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ is **ARMA**(p, q) if it is stationary and

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, \quad (3.12)$$

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 $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$ and write the model as

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}. \quad (3.13)$$

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 ARMA model may be seen as a regression of the present outcome (x_t) on the past outcomes (x_{t-1}, \dots, x_{t-p}), with correlated errors. That is,

$$x_t = \beta_0 + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \epsilon_t,$$

where $\epsilon_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$, although we call the regression parameters ϕ instead of β .

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). Using (3.3) and (3.10), the ARMA(p, q) model in (3.12) may be written in concise form as

$$\phi(B)x_t = \theta(B)w_t. \quad (3.14)$$

Example 3.5 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

$$x_t = .5x_{t-1} - .5w_{t-1} + w_t, \quad (3.15)$$

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 (3.15)], 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.

Example 3.5 points out the need to be careful when fitting ARMA models to data. For example, if a process is truly white noise, it is possible to fit a significant ARMA(k, k) model to the data. That is, it is possible to obtain a seemingly complicated dynamic description of simple white noise. Consider the following example.

Example 3.6 Parameter Redundancy (cont)

Although we have not yet discussed estimation, we present the following example demonstrating the problem of over parameterization. We generated 150 iid standard normals and then fit an ARMA(1, 1) to the data. Note that $\hat{\phi} = -.96$ and $\hat{\theta} = .95$, and both are significant. Below is the R code (note that the estimate called ‘intercept’ is really the estimate of the mean).

```
set.seed(8675309)      # Jenny, I got your number
x = rnorm(150)          # generate iid N(0,1)s
arima(x, order=c(1,0,1)) # estimation
Coefficients:
      ar1      ma1 intercept <- misnomer
-0.9595  0.9527    0.0462
s.e.    0.1688  0.1750    0.0727
```

Thus, forgetting the (non-significant) mean estimate, the fitted model looks like

$$(1 + .96B)x_t = (1 + .95B)w_t$$

Definition 3.4 Causality and Invertibility

It is sometimes advantageous to write the model in its causal form:

$$x_t = \phi(B)^{-1}\theta(B)w_t = \psi(B)w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}, \quad (3.16)$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ ($\psi_0 = 1$), or in its **invertible form**:

$$w_t = \theta(B)^{-1}\phi(B)x_t = \pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}. \quad (3.17)$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ ($\pi_0 = 1$), assuming the representations are well-defined.

Note that the parameters ψ_j may be obtained by matching coefficients of B in $\phi(B)\psi(B) = \theta(B)$. Likewise, the parameters π_j may be obtained by matching coefficients of B in $\phi(B) = \pi(B)\theta(B)$. We note that it is not always possible to solve these relationships and some restrictions apply, as follows.

Property 3.1 Causality and Invertibility (existence)

Let

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad \text{and} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

be the AR and MA polynomials obtained by replacing the backshift operator B in (3.3) and (3.10) by a complex number z . 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 (3.16) can be determined by solving ($\psi_0 = 1$)

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

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 (3.17) can be determined by solving ($\pi_0 = 1$)

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

We demonstrate the property in the following example.

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 notice that

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

and

$$\theta(B) = (1 + B + .25B^2) = (1 + .5B)^2$$

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

$$x_t = .9x_{t-1} + .5w_{t-1} + w_t. \quad (3.18)$$

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

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

Rearranging, we get

$$1 + (\psi_1 - .9)z + (\psi_2 - .9\psi_1)z^2 + \cdots + (\psi_j - .9\psi_{j-1})z^j + \cdots = 1 + .5z.$$

* $\phi(z)$ can't be zero in here... you wouldn't want to divide by zero, would you?

† $\theta(z)$ can't be zero in here.

The coefficients of z on the left and right sides must be the same, so we get $\psi_1 - .9 = .5$ or $\psi_1 = 1.4$, and $\psi_j - .9\psi_{j-1} = 0$ for $j > 1$. Thus, $\psi_j = 1.4(.9)^{j-1}$ for $j \geq 1$ and (3.18) can be written as

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

The values of ψ_j may be calculated in R as follows:

```
ARMAtoMA(ar = .9, ma = .5, 10) # first 10 psi-weights
[1] 1.40 1.26 1.13 1.02 0.92 0.83 0.74 0.67 0.60 0.54
```

The invertible representation using [Property 3.1](#) is obtained by matching coefficients in $\theta(z)\pi(z) = \phi(z)$,

$$(1 + .5z)(1 + \pi_1 z + \pi_2 z^2 + \pi_3 z^3 + \cdots) = 1 - .9z.$$

In this case, the π -weights are given by $\pi_j = (-1)^j 1.4(.5)^{j-1}$, for $j \geq 1$, and hence, we can also write (3.18) as

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

The values of π_j may be calculated in R as follows by reversing the roles of w_t and x_t ; i.e., write the model as $w_t = -.5w_{t-1} + x_t - .9x_{t-1}$:

```
ARMAtoMA(ar = -.5, ma = -.9, 10) # first 10 pi-weights
[1] -1.400 .700 -.350 .175 -.087 .044 -.022 .011 -.006 .003
```

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, we must have $\phi(z) \neq 0$ for $|z| \leq 1$. If we solve $\phi(z) = 1 - \phi z = 0$, we find that the root (or zero) occurs at $z_0 = 1/\phi$, so that $|z_0| > 1$ only if $|\phi| < 1$. For higher order models, the relationship between roots and parameters is not so simple.

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. That is, if z_1 and z_2 are the roots, then $|z_1| > 1$ and $|z_2| > 1$. 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. In terms of the coefficients, the equivalent condition is

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \text{and} \quad |\phi_2| < 1. \quad (3.19)$$

This causality condition specifies a triangular region in the parameter space; see [Figure 3.3](#).

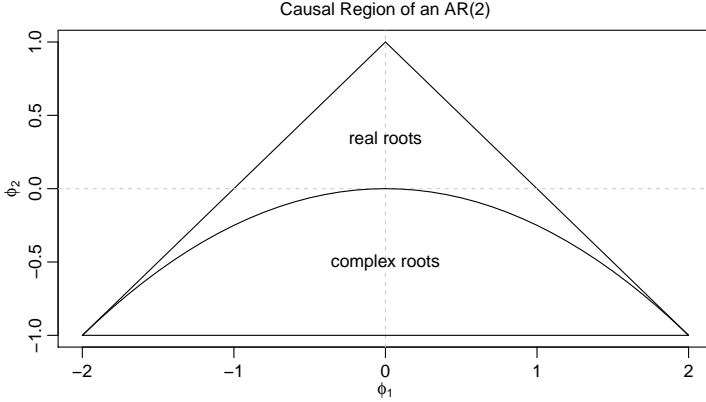


Fig. 3.3. Causal region for an AR(2) in terms of the parameters.

3.3 Autocorrelation and Partial Autocorrelation

We begin by exhibiting the ACF of an MA(q) process.

Example 3.9 ACF of an MA(q)

The model is $x_t = \theta(B)w_t$, 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

$$\begin{aligned} \gamma(h) &= \text{cov}(x_{t+h}, x_t) = \text{cov}\left(\sum_{j=0}^q \theta_j w_{t+h-j}, \sum_{k=0}^q \theta_k w_{t-k}\right) \\ &= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \leq h \leq q \\ 0 & h > q. \end{cases} \end{aligned} \quad (3.20)$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only display the values for $h \geq 0$. The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (3.20) by $\gamma(0)$ yields the ACF of an MA(q):

$$\rho(h) = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \theta_1^2 + \cdots + \theta_q^2} & 1 \leq h \leq q \\ 0 & h > q. \end{cases} \quad (3.21)$$

Example 3.10 ACF of an AR(p) and ARMA(p, q)

For an AR(p) or ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, write it as

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

or

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}. \quad (3.22)$$

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

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}, \quad h \geq 0, \quad (3.23)$$

so that the ACF is given by

$$\rho(h) = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+h}}{\sum_{j=0}^{\infty} \psi_j^2}, \quad h \geq 0. \quad (3.24)$$

Unlike the MA(q), the ACF of an AR(p) or an ARMA(p, q) does not cut off at any lag, so using the ACF to help identify the order of an AR or ARMA is difficult. Also, (3.24) is not appealing in that it provides little information about the appearance of the ACF of various models.

Example 3.11 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

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2), \quad h = 1, 2, \dots \quad (3.25)$$

In (3.25), we used the fact that $E(x_t) = 0$ and for $h > 0$,

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

Divide (3.25) through by $\gamma(0)$ to obtain a recursion for the ACF:

$$\rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \quad h = 1, 2, \dots \quad (3.26)$$

The initial conditions are $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1 - \phi_2)$, which is obtained by evaluating (3.26) for $h = 1$ and noting that $\rho(1) = \rho(-1)$.

Equations such as (3.26) are called difference equations, and the solutions are fairly simple expressions. First, the polynomial associated with (3.26) is

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

where the power of z is the power of the backshift, B ; i.e., (3.26) is $(1 - \phi_1 B - \phi_2 B^2)\rho(h) = 0$. In general, z is a complex number. Let z_1 and z_2 be the roots (or zeros) of the associated polynomial, i.e., $\phi(z_1) = \phi(z_2) = 0$. For a causal model, the roots are outside the unit circle: $|z_1| > 1$ and $|z_2| > 1$. Now, consider the solutions:

(i) When z_1 and z_2 are distinct, then

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

so $\rho(h) \rightarrow 0$ exponentially fast as $h \rightarrow \infty$. The constants c_1 and c_2 are obtained by solving for them using the initial conditions given above.

For example, when $h = 0$, we have $1 = c_1 + c_2$, and so on.

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

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

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

In case (i) with complex roots, $z_2 = \bar{z}_1$ are a complex conjugate pair, and $c_2 = \bar{c}_1$ [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 $h \rightarrow \infty$, but it does so in a sinusoidal fashion. The implication of this result is shown in [Example 3.12](#).

Example 3.12 An AR(2) with Complex Roots

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

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

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

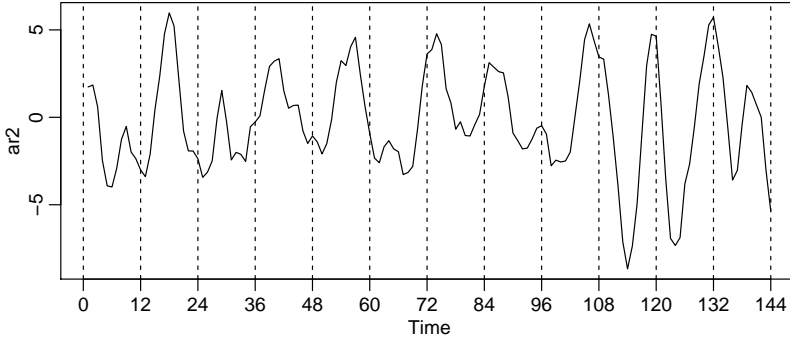


Fig. 3.4. Simulated AR(2) model, $n = 144$ with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

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

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

To reproduce **Figure 3.4**:

```
set.seed(8675309)
ar2 = arima.sim(list(order=c(2,0,0), ar=c(1.5,-.75)), n = 144)
plot(ar2, axes=FALSE, xlab="Time")
axis(2); axis(1, at=seq(0,144,by=12)); box()
abline(v=seq(0,144,by=12), lty=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)
```

In general, the behavior of the ACF of an $\text{AR}(p)$ or an $\text{ARMA}(p, q)$ when $p \geq 2$ will be similar to the $\text{AR}(2)$ case. When $p = 1$, the behavior is like the $\text{AR}(1)$ case.

Example 3.13 The ACF of an $\text{ARMA}(1, 1)$

Consider the $\text{ARMA}(1, 1)$ process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$, where $|\phi| < 1$. Using the theory of difference equations, we can show that the ACF is given by

$$\rho(h) = \frac{(1 + \theta\phi)(\phi + \theta)}{1 + 2\theta\phi + \theta^2} \phi^{h-1}, \quad h \geq 1. \quad (3.27)$$

Notice that the general pattern of $\rho(h)$ in (3.27) is not different from that of an $\text{AR}(1)$ given in (3.7). Hence, it is unlikely that we will be able to tell the difference between an $\text{ARMA}(1, 1)$ and an $\text{AR}(1)$ based solely on an ACF estimated from a sample. This consideration will lead us to the partial autocorrelation function.

THE PARTIAL AUTOCORRELATION FUNCTION (PACF)

In (3.21), we saw that 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).

Recall that if X , Y , and Z are random variables, then the partial correlation between X and Y given Z is obtained by regressing X on Z to obtain \hat{X} , regressing Y on Z to obtain \hat{Y} , and then calculating

$$\rho_{XY|Z} = \text{corr}\{X - \hat{X}, Y - \hat{Y}\}.$$

The idea is that $\rho_{XY|Z}$ measures the correlation between X and Y with the linear effect of Z removed (or partialled out). If the variables are multivariate normal, then this definition coincides with $\rho_{XY|Z} = \text{corr}(X, Y | Z)$.

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

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

This result follows from causality because x_{t-2} involves $\{w_{t-2}, w_{t-3}, \dots\}$, 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 partialling out) the effect of x_{t-1} . That is, we consider the correlation between $x_t - \phi x_{t-1}$ and $x_{t-2} - \phi x_{t-1}$, because it is the correlation between x_t and x_{t-2} with the linear dependence of each on x_{t-1} removed. In this way, we have broken the dependence chain between x_t and x_{t-2} . In fact,

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

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

Definition 3.5 *The partial autocorrelation function (PACF) of a stationary process, x_t , denoted ϕ_{hh} , for $h = 1, 2, \dots$, is*

$$\phi_{11} = \text{corr}(x_1, x_0) = \rho(1) \quad (3.28)$$

and

$$\phi_{hh} = \text{corr}(x_h - \hat{x}_h, x_0 - \hat{x}_0), \quad h \geq 2, \quad (3.29)$$

where \hat{x}_h is the regression of x_h on $\{x_1, x_2, \dots, x_{h-1}\}$ and \hat{x}_0 is the regression of x_0 on $\{x_1, x_2, \dots, x_{h-1}\}$.

Thus, due to the stationarity, the PACF, ϕ_{hh} , is the correlation between x_{t+h} and x_t with the linear dependence of everything between them, namely $\{x_{t+1}, \dots, x_{t+h-1}\}$, on each, removed.

Example 3.14 The PACF of an AR(p)

The model can be written as

$$x_{t+h} = \sum_{j=1}^p \phi_j x_{t+h-j} + w_{t+h},$$

where the roots of $\phi(z)$ are outside the unit circle. When $h > p$, the regression of x_{t+h} on $\{x_{t+1}, \dots, x_{t+h-1}\}$, is

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

Although we have not proved this result, it should be obvious that it is so. Thus, when $h > p$,

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

because, by causality, $x_t - \hat{x}_t$ depends only on $\{w_{t+h-1}, w_{t+h-2}, \dots\}$. When $h \leq p$, ϕ_{pp} is not zero, and $\phi_{11}, \dots, \phi_{p-1,p-1}$ are not necessarily zero. We will see later that, in fact, $\phi_{pp} = \phi_p$. Figure 3.5 shows the ACF and the PACF of the AR(2) model presented in Example 3.12.

To reproduce Figure 3.5 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)
```

We also have the following large sample result for the PACF, which may be compared to the similar result for the ACF given in Property 1.1.

Property 3.2 Large Sample Distribution of the PACF

If the time series is a causal AR(p) process and the sample size n is large, then $\sqrt{n} \hat{\phi}_{hh}$ is approximately $N(0, 1)$, for $h > p$. This result also holds for $p = 0$, wherein the process is white noise.

Example 3.15 The PACF of an MA(q)

For an 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 $|\theta| < 1$, it can be shown that

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

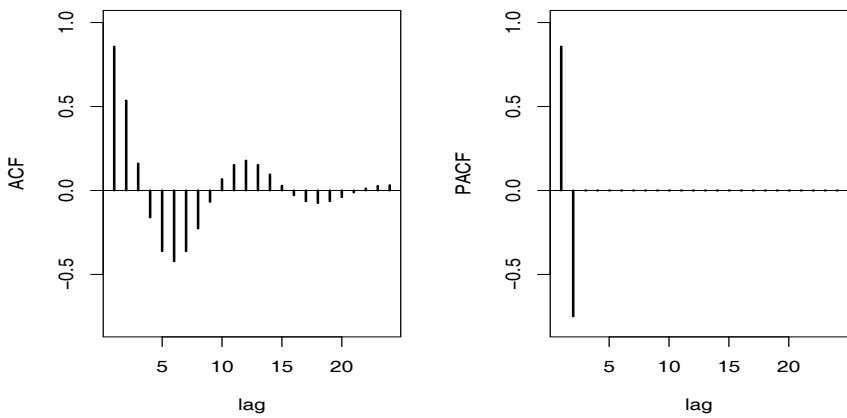


Fig. 3.5. The ACF and PACF of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

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

We do not have to compute the PACF by performing numerous regressions first. The computations are done via a recursive formula that we will discuss later in [Property 3.3](#).

The PACF for MA models behaves much like the ACF for AR models. Also, the PACF for AR models behaves much like the ACF for MA models. Because an invertible ARMA model has an infinite AR representation, the PACF will not cut off. We may summarize these results in [Table 3.1](#).

Example 3.16 Preliminary Analysis of the Recruitment Series

We consider the problem of modeling the Recruitment series shown in [Figure 1.6](#). There are 453 months of observed recruitment ranging over the years 1950-1987. The ACF and the PACF given in [Figure 3.6](#) are consistent with the behavior of an AR(2). The ACF has cycles corresponding roughly to a 12-month period, and the PACF has large values for $h = 1, 2$ and then is essentially zero for higher order lags. Based on [Table 3.1](#), these results suggest that a second-order ($p = 2$) autoregressive model might provide a good fit. Although we will discuss estimation in detail in §3.6, we ran a regression (see §2.2) using the data triplets $\{(x; z_1, z_2) : (x_3; x_2, x_1), (x_4; x_3, x_2), \dots, (x_{453}; x_{452}, x_{451})\}$ to fit the model

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

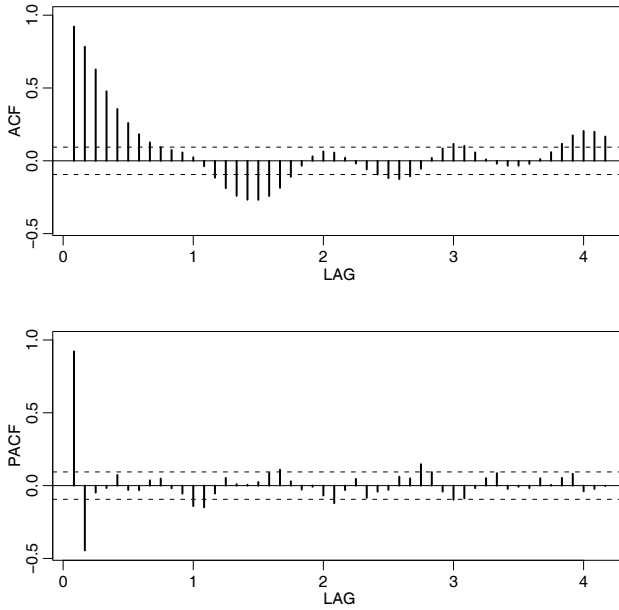


Fig. 3.6. ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).

for $t = 3, 4, \dots, 453$. The values of the estimates were $\hat{\phi}_0 = 6.74_{(1.11)}$, $\hat{\phi}_1 = 1.35_{(.04)}$, $\hat{\phi}_2 = -.46_{(.04)}$, and $\hat{\sigma}_w^2 = 89.72$, 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
```

3.4 Estimation

Throughout this section, we assume we have n observations, x_1, \dots, 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, ϕ_1, \dots, ϕ_p , $\theta_1, \dots, \theta_q$, 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),

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

similar to [Example 3.11](#), we have the following result:

Definition 3.6 *The Yule–Walker equations are given by*

$$\rho(h) = \phi_1 \rho(h-1) + \cdots + \phi_p \rho(h-p), \quad h = 1, 2, \dots, p, \quad (3.30)$$

$$\sigma_w^2 = \gamma(0) [1 - \phi_1 \rho(1) - \cdots - \phi_p \rho(p)]. \quad (3.31)$$

The estimators obtained by replacing $\gamma(0)$ with its estimate, $\hat{\gamma}(0)$ and $\rho(h)$ with its estimate, $\hat{\rho}(h)$, are called the *Yule–Walker estimators*. For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\hat{\sigma}_w^2$ is close to the true value of σ_w^2 .

Example 3.17 Yule–Walker Estimation for an AR(2) Process

The data shown in [Figure 3.4](#) were $n = 144$ simulated observations from the AR(2) model $x_t = 1.5x_{t-1} - .75x_{t-2} + w_t$, where $w_t \sim \text{iid } N(0, 1)$. Using the same simulated data, we have

```
ar.yw(ar2, order=2)
Coefficients:
      1      2
1.4471 -0.7160
sigma^2 estimated as 1.561
```

Example 3.18 Yule–Walker Estimation of the Recruitment Series

In [Example 3.16](#) 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 3.16](#).

```
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)
```

In the case of AR(p) models, the Yule–Walker estimators are optimal estimators, but this is not true for MA(q) or ARMA(p, q) models. 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 3.19 Method of Moments Estimation for an MA(1)

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

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

Two solutions exist, so we would pick the invertible one. If $|\hat{\rho}(1)| \leq \frac{1}{2}$, 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 $|\hat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\hat{\rho}(1) = .507$ 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)
```

The preferred method of estimation is maximum likelihood estimation (MLE), which determines the values of the parameters that are most *likely* to have produced the observations. MLE is discussed in [Section 3.6](#). For ARMA models, this is closely related to least squares.

LEAST SQUARES ESTIMATION

We now discuss least squares for ARMA(p, q) models via Gauss–Newton. Write the model parameters as $\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, and for the ease of discussion, we will put $\mu = 0$. Now, write the model in terms of the errors

$$w_t(\beta) = x_t - \sum_{j=1}^p \phi_j x_{t-j} - \sum_{k=1}^q \theta_k w_{t-k}(\beta), \quad (3.32)$$

emphasizing the dependence of the errors on the parameters (recall that $w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$ by invertibility, and the π_j are complicated functions of β).

For *conditional least squares*, we approximate the residual sum of squares by conditioning on x_1, \dots, x_p (if $p > 0$) and $w_p = w_{p-1} = w_{p-2} = \dots = w_{1-q} = 0$ (if $q > 0$), in which case, given β , we may evaluate (3.32) for $t = p+1, p+2, \dots, n$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\beta) = \sum_{t=p+1}^n w_t^2(\beta). \quad (3.33)$$

Minimizing $S_c(\beta)$ with respect to β yields the conditional least squares estimates.

If $q = 0$, the problem is linear regression and no iterative technique is needed to minimize $S_c(\phi_1, \dots, \phi_p)$. For example, for an AR(1), $x_t = \phi x_{t-1} + w_t$, the conditional sum of squares is

$$S_c(\phi) = \sum_{t=2}^n w_t^2(\phi) = \sum_{t=2}^n (x_t - \phi x_{t-1})^2.$$

Note that we have to start at $t = 2$ because x_0 is not observed. The conditional least squares estimate of ϕ follows from simple linear regression wherein,

$$\hat{\phi} = \frac{\sum_{t=2}^n x_t x_{t-1}}{\sum_{t=2}^n x_{t-1}^2},$$

which is nearly $\hat{\rho}(1)$.

If $q > 0$, the problem becomes nonlinear regression and we will have to rely on numerical optimization. Gauss–Newton uses an iterative method for solving the problem of minimizing (3.33). We demonstrate the method for an MA(1).

Example 3.20 Gauss–Newton for an MA(1)

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

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \dots, n, \quad (3.34)$$

where we condition on $w_0(\theta) = 0$. Our goal is to find the value of θ that minimizes $S_c(\theta) = \sum_{t=1}^n w_t^2(\theta)$, which is a nonlinear function of θ .

Let $\theta_{(0)}$ be an initial estimate of θ . For example, we could use method of moments. The first-order Taylor expansion³ of $w_t(\theta)$ at $\theta_{(0)}$ is

$$w_t(\theta) \approx w_t(\theta_{(0)}) - (\theta - \theta_{(0)}) z_t(\theta_{(0)}), \quad (3.35)$$

where

$$z_t(\theta_{(0)}) = - \left. \frac{\partial w_t(\theta)}{\partial \theta} \right|_{\theta=\theta_{(0)}}.$$

Taking derivatives,

$$- \frac{\partial w_t(\theta)}{\partial \theta} = w_{t-1}(\theta) + \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \dots, n, \quad (3.36)$$

where $\partial w_0(\theta)/\partial \theta = 0$. Using the notation of (3.35), we can also write (3.36) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \dots, n, \quad (3.37)$$

where $z_0(\theta) = 0$. This implies that the derivative sequence is an AR process, which we may easily compute given a value of θ .

³ [Newton's method](#) and [Taylor expansion](#) (links to WikiBooks K-12 calculus book).

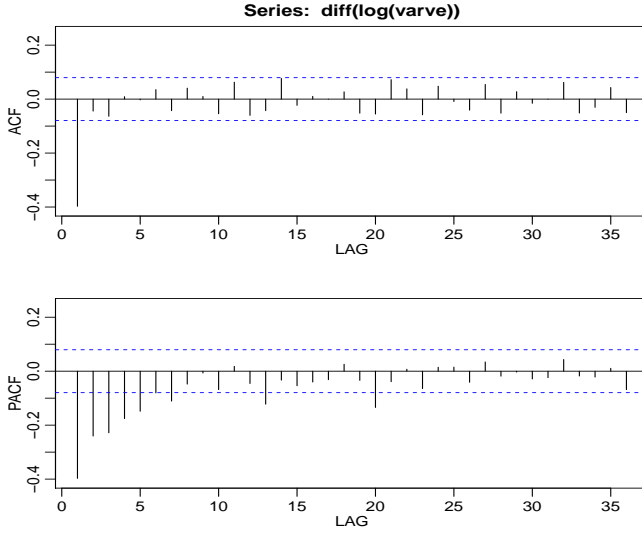


Fig. 3.7. ACF and PACF of transformed glacial varves.

The linear approximation of $S_c(\theta)$ is found by replacing $w_t(\theta)$ by its linear approximation in (3.35),

$$Q(\theta) = \sum_{t=1}^n \left[\underbrace{w_t(\theta_{(0)})}_{y_t} - \underbrace{(\theta - \theta_{(0)})}_{\beta} \underbrace{z_t(\theta_{(0)})}_{z_t} \right]^2 \quad (3.38)$$

and this is the quantity that we will minimize. The problem is now simple linear regression (“ $y_t = \beta z_t + \epsilon_t$ ”), so that

$$(\theta - \theta_{(0)}) = \sum_{t=1}^n z_t(\theta_{(0)}) w_t(\theta_{(0)}) / \sum_{t=1}^n z_t^2(\theta_{(0)}),$$

or

$$\hat{\theta} = \theta_{(0)} + \sum_{t=1}^n z_t(\theta_{(0)}) w_t(\theta_{(0)}) / \sum_{t=1}^n z_t^2(\theta_{(0)}).$$

Consequently, the Gauss–Newton procedure in this case is, on iteration $j+1$, set

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^n z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^n z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \dots, \quad (3.39)$$

where the values in (3.39) are calculated recursively using (3.34) and (3.37). 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 3.21 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 2.6, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

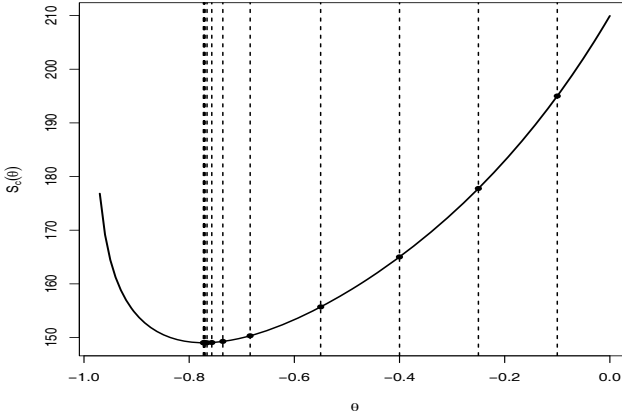


Fig. 3.8. Conditional sum of squares versus values of the moving average parameter for the glacial varve example, [Example 3.21](#). Vertical lines indicate the values of the parameter obtained via Gauss–Newton; see [Table 3.2](#) for the actual values.

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

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

The sample ACF and PACF, shown in [Figure 3.7](#), 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.

Because $\hat{\rho}(1) = -.4$, it should be clear that $\hat{\theta}$ is negative. Recalling [Example 3.4](#), the method of moments estimate is found by setting $-.4 = \hat{\theta}/(1+\hat{\theta}^2)$, which yields two solutions $\hat{\theta} = -.5$ and -2 . Since $-.5$ is the invertible solution, we could initialize the Gauss–Newton procedure at this value. However, for most ARMA models, it is not necessary to be so precise as long as the procedure is started in a causal and invertible region. For demonstration purposes, we will start the procedure at $\theta_{(0)} = -.1$.

The results of eleven iterations of the Gauss–Newton procedure, [\(3.39\)](#), are given in [Table 3.2](#). The final estimate is $\hat{\theta} = \theta_{(11)} = -.773$; interim values and the corresponding value of the conditional sum of squares, $S_c(\theta)$ given in [\(3.33\)](#), are also displayed in the table. The final estimate of the error variance is $\hat{\sigma}_w^2 = 148.98/632 = .236$ 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 $\hat{\theta}$ is $\sqrt{.236/369.73} = .025$;⁴ this leads to a t -value of $-.773/.025 = -30.92$ with 632 degrees of freedom.

⁴ To estimate the standard error, we are using the standard regression results from [\(2.6\)](#) as an approximation

Table 3.2. Gauss–Newton Results for **Example 3.21**

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

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.

In the general case of causal and invertible ARMA(p, q) models, maximum likelihood estimation and nonlinear least squares estimation (and Yule–Walker estimation in the case of AR models) all lead to optimal estimators.

Example 3.22 Some Specific Asymptotic Distributions ⁵

AR(1):

$$\hat{\phi} \sim \text{AN} [\phi, n^{-1}(1 - \phi^2)] . \quad (3.40)$$

AR(2):

$$\begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, n^{-1} \begin{pmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ \text{sym} & 1 - \phi_2^2 \end{pmatrix} \right] . \quad (3.41)$$

MA(1):

$$\hat{\theta} \sim \text{AN} [\theta, n^{-1}(1 - \theta^2)] . \quad (3.42)$$

MA(2):

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, n^{-1} \begin{pmatrix} 1 - \theta_2^2 & \theta_1(1 + \theta_2) \\ \text{sym} & 1 - \theta_2^2 \end{pmatrix} \right] . \quad (3.43)$$

⁵ $X_n \sim \text{AN}(\mu_n, \sigma_n^2)$ if $\Pr\{(X_n - \mu_n)/\sigma_n \leq z\} \rightarrow \Pr\{Z \leq z\}$ as $n \rightarrow \infty$, where $Z \sim N(0, 1)$.

ARMA(1,1):

$$\begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \phi \\ \theta \end{pmatrix}, n^{-1} \begin{bmatrix} (1 - \phi^2)^{-1} & (1 + \phi\theta)^{-1} \\ \text{sym} & (1 - \theta^2)^{-1} \end{bmatrix}^{-1} \right]. \quad (3.44)$$

Example 3.23 Overfitting Caveat

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

3.5 Forecasting

In forecasting, the goal is to predict future values of a time series, x_{n+m} , $m = 1, 2, \dots$, based on the data collected to the present, $\mathbf{x} = \{x_n, x_{n-1}, \dots, x_1\}$. Throughout this section, we will assume x_t is stationary and the model parameters are known. When the parameters are unknown, we replace them with their estimates.

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

$$x_{n+1}^n = \phi_{n1}x_n + \phi_{n2}x_{n-1} + \dots + \phi_{nn}x_1, \quad (3.45)$$

and we must solve for the coefficients to minimize the *mean squared prediction error* given by

$$P_{n+1}^n = E [x_{n+1} - x_{n+1}^n]^2 = E [x_{n+1} - (\phi_{n1}x_n + \dots + \phi_{nn}x_1)]^2.$$

Generally, the coefficients depend on the sample size n , and this is made explicit. The second subscript indicates the distance between the future and

the data values. For example, ϕ_{nj} is the coefficient of x_{n+1-j} because it is j time periods away from x_{n+1} .

The coefficients $\{\phi_{n1}, \phi_{n2}, \dots, \phi_{nn}\}$ satisfy

$$\sum_{j=1}^n \phi_{nj} \gamma(k-j) = \gamma(k), \quad k = 1, \dots, n, \quad (3.46)$$

which is similar to the normal equations obtained in least squares regression. It should be clear that given data, we could substitute $\gamma(k)$ with its estimate $\hat{\gamma}(k)$ into (3.46) and then solve for the ϕ s.

The prediction equations (3.46) can be written in matrix notation as

$$\Gamma_n \phi_n = \gamma_n, \quad (3.47)$$

where $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^n$ is an $n \times n$ matrix, $\phi_n = (\phi_{n1}, \dots, \phi_{nn})'$ is an $n \times 1$ vector, and $\gamma_n = (\gamma(1), \dots, \gamma(n))'$ is an $n \times 1$ vector. For ARMA models, Γ_n is nonsingular, and the unique elements of ϕ_n are given by

$$\phi_n = \Gamma_n^{-1} \gamma_n. \quad (3.48)$$

The mean square one-step-ahead prediction error is

$$P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n. \quad (3.49)$$

Example 3.24 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 (3.46) or (3.48), 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., $x_3^2 = \phi_{21} x_2 + \phi_{22} x_1$. It should be apparent from the model that $x_3^2 = \phi_1 x_2 + \phi_2 x_1$. In general, 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 $j = 3, 4, \dots, n$.

As in Example 3.24, it should be clear that, if the time series is an AR(p) process, then, for $n \geq p$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \dots + \phi_p x_{n-p+1}. \quad (3.50)$$

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 (3.48) 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 3.3 The Durbin–Levinson Algorithm

Equations (3.48) and (3.49) can be solved iteratively as follows:

$$\phi_{00} = 0, \quad P_1^0 = \gamma(0). \quad (3.51)$$

For $n \geq 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi_{n-1,k} \rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k} \rho(k)}, \quad P_{n+1}^n = P_n^{n-1} (1 - \phi_{nn}^2), \quad (3.52)$$

where, for $n \geq 2$, $\phi_{nk} = \phi_{n-1,k} - \phi_{nn} \phi_{n-1,n-k}$, $k = 1, 2, \dots, n-1$.

An important consequence of the Durbin–Levinson algorithm is as follows.

Property 3.4 Iterative Solution for the PACF

The PACF of a stationary process x_t , can be obtained iteratively via (3.52) as ϕ_{nn} , for $n = 1, 2, \dots$.

Using Property 3.4 and putting $n = p$ in (3.45) and (3.50), it follows that for an AR(p) model,

$$\begin{aligned} x_{p+1}^p &= \phi_{p1} x_p + \phi_{p2} x_{p-1} + \dots + \phi_{pp} x_1 \\ &= \phi_1 x_p + \phi_2 x_{p-1} + \dots + \phi_p x_1. \end{aligned} \quad (3.53)$$

Result (3.53) 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 3.14.

So far, we have concentrated on one-step-ahead prediction, but m -step-ahead prediction is similar. The BLP of x_{n+m} for any $m \geq 1$ given data $\{x_1, \dots, x_n\}$ yields the m -step-ahead predictor

$$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \dots + \phi_{nn}^{(m)} x_1, \quad (3.54)$$

where $\{\phi_{n1}^{(m)}, \phi_{n2}^{(m)}, \dots, \phi_{nn}^{(m)}\}$ satisfy

$$\sum_{j=1}^n \phi_{nj}^{(m)} \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \dots, n. \quad (3.55)$$

The prediction equations (3.55) can again be written in matrix notation as

$$\Gamma_n \phi_n^{(m)} = \gamma_n^{(m)}, \quad (3.56)$$

where $\gamma_n^{(m)} = (\gamma(m), \dots, \gamma(m+n-1))'$, and $\phi_n^{(m)} = (\phi_{n1}^{(m)}, \dots, \phi_{nn}^{(m)})'$ are $n \times 1$ vectors.

The mean square m -step-ahead prediction error is

$$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = \gamma(0) - \gamma_n^{(m)'} \Gamma_n^{-1} \gamma_n^{(m)}. \quad (3.57)$$

FORECASTING ARMA PROCESSES

Because ARMA models are invertible; i.e., $w_t = x_t + \sum_{j=1}^{\infty} \pi_j x_{t-j}$, we may write

$$x_{n+m} = - \sum_{j=1}^{\infty} \pi_j x_{n+m-j} + w_{n+m}.$$

If we had the infinite history $\{x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots\}$, of the data available, we would predict x_{n+m} by

$$x_{n+m}^n = - \sum_{j=1}^{\infty} \pi_j x_{n+m-j}^n,$$

successively for $m = 1, 2, \dots$. In this case, $x_t^n = x_t$ for $t = n, n-1, \dots$. We only have the actual data $\{x_n, x_{n-1}, \dots, x_1\}$ available, but a practical solution is to truncate the forecasts as

$$x_{n+m}^n = - \sum_{j=1}^{n+m-1} \pi_j x_{n+m-j}^n,$$

with $x_t^n = x_t$ for $1 \leq t \leq n$. For an AR(p) model, truncated forecasts are exact when $n > p$. For ARMA models in general, as long as n is large, the approximation works well because the π -weights are going to zero exponentially fast. The truncated forecasts have a simple representation, as follows.

Property 3.5 Truncated Prediction for ARMA *For ARMA(p, q) models, the truncated predictors are calculated recursively for $m = 1, 2, \dots$, as*

$$x_{n+m}^n = \phi_1 x_{n+m-1}^n + \dots + \phi_p x_{n+m-p}^n + \theta_1 w_{n+m-1}^n + \dots + \theta_q w_{n+m-q}^n, \quad (3.58)$$

where $x_t^n = x_t$ for $1 \leq t \leq n$ and $x_t^n = 0$ for $t \leq 0$. The truncated prediction errors are given by: $w_t^n = 0$ for $t \leq 0$ or $t > n$, and

$$w_t^n = \phi(B)x_t^n - \theta_1 w_{t-1}^n - \dots - \theta_q w_{t-q}^n$$

for $1 \leq t \leq n$.

Example 3.25 Forecasting an ARMA(1, 1) Series

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

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

Then, based on (3.58), the one-step-ahead truncated forecast is

$$x_{n+1}^n = \phi x_n + 0 + \theta w_n^n.$$

For $m \geq 2$, we have

$$x_{n+m}^n = \phi x_{n+m-1}^n,$$

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

To calculate 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 $t = 1, \dots, n$. For truncated forecasting using (3.58), put $w_0^n = 0$, $x_0 = 0$, and then iterate the errors forward in time

$$w_t^n = x_t - \phi x_{t-1} - \theta w_{t-1}^n, \quad t = 1, \dots, n.$$

For large n , it can be shown that the mean squared prediction error for ARMA(p, q) models is approximately

$$P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \quad (3.59)$$

To establish (3.59), by invertibility, the w_n, w_{n-1}, \dots are functions of the observations x_n, x_{n-1}, \dots . Letting $w_{n+k}^n = E(w_{n+k} \mid x_n, x_{n-1}, \dots)$, we have $w_{n+k}^n = w_{n+k}$ for $k \leq 0$, but for future errors, $w_{n+k}^n = E w_{n+k} = 0$ for $k \geq 1$. Thus

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{m+n-j} \quad \text{and} \quad x_{n+m}^n = \sum_{j=0}^{\infty} \psi_j w_{m+n-j}^n = \sum_{j=m}^{\infty} \psi_j w_{m+n-j},$$

so that $E[x_{n+m} - x_{n+m}^n]^2 = E\left[\sum_{j=0}^{m-1} \psi_j w_{m+n-j}\right]^2$, which is (3.59).

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

$$x_{n+m}^n \pm c_{\frac{\alpha}{2}} \sqrt{P_{n+m}^n}, \quad (3.60)$$

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 Johnson and Wichern, 1992, Chapter 5].

Example 3.26 Forecasting the Recruitment Series

In Example 3.16 we fit an AR(2) model to the Recruitment series using OLS. Here, we use MLE:

```
sarima(rec,2,0,0)           # fit model
Coefficients:
      ar1      ar2      xmean
    1.3512  -0.4612  61.8585
s.e.    0.0416    0.0417    4.0039
sigma^2 estimated as 89.33
61.8585*(1-1.3512+.4612)    # get constant
[1] 6.804435
```

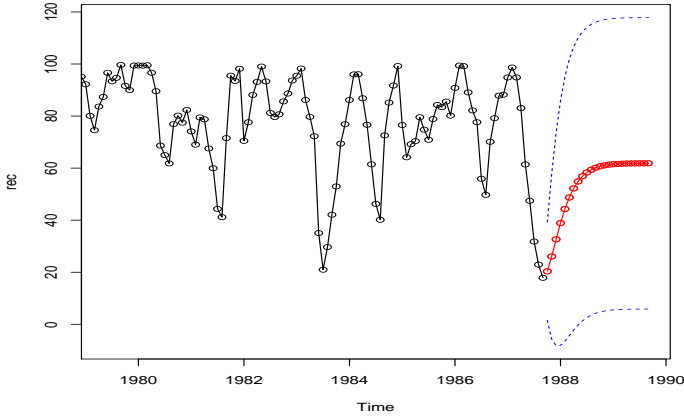


Fig. 3.9. Twenty-four month forecasts for the Recruitment series. The actual data shown are from about January 1980 to September 1987, and then the forecasts plus and minus one standard error are displayed.

The results are nearly the same as using OLS. Using the parameter estimates as the actual parameter values, the forecasts are calculated as

$$x_{n+m}^n = 6.80 + 1.35x_{n+m-1}^n - .46x_{n+m-2}^n$$

for $n = 453$ and $m = 1, 2, \dots, 12$. Recall that $x_t^s = x_t$ when $t \leq s$. The forecasts errors P_{n+m}^n are calculated using (3.59). Here, $\hat{\sigma}_w^2 = 89.33$, and the ψ -weights satisfy $\psi_j = 1.35\psi_{j-1} - .46\psi_{j-2}$ for $j \geq 2$, where $\psi_0 = 1$ and $\psi_1 = 1.35$. Thus, for $n = 453$,

$$P_{n+1}^n = 89.33, \quad P_{n+2}^n = 89.33(1 + 1.35^2), \quad P_{n+3}^n = 89.33(1 + 1.35^2 + [1.35^2 - .46]^2),$$

and so on.

Figure 3.9 shows the result of forecasting the Recruitment series over a 24-month horizon, $m = 1, 2, \dots, 24$, obtained in R as

`sarima.for(rec, 24, 2, 0, 0)`

Note how the forecast levels off to the mean quickly and the prediction intervals are wide and become constant.

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

$$x_{1-m}^n = \sum_{j=1}^n \alpha_{nj}^{(m)} x_j. \quad (3.61)$$

Analogous to (3.55), the prediction equations (assuming $\mu_x = 0$) are

$$\sum_{j=1}^n \alpha_{nj}^{(m)} \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \dots, n. \quad (3.62)$$

These equations are precisely the prediction equations for forward prediction. That is, $\alpha_{nj}^{(m)} \equiv \phi_{nj}^{(m)}$, for $j = 1, \dots, n$, where the $\phi_{nj}^{(m)}$ are given by (3.56). Finally, the backcasts are given by

$$x_{1-m}^n = \phi_{n1}^{(m)} x_1 + \dots + \phi_{nn}^{(m)} x_n, \quad m = 1, 2, \dots \quad (3.63)$$

The implication of this result is that to backcast, you simply reverse the order of the data and forecast the reversed data set.

3.6 Maximum Likelihood Estimation **

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

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

The conditional distribution of x_t given x_{t-1}, \dots, x_1 is Gaussian with mean x_t^{t-1} and variance P_t^{t-1} . Recall from [Property 3.3](#) that $P_1^0 = \gamma_x(0)$, and

$$\begin{aligned} P_t^{t-1} &= P_{t-1}^{t-2}(1 - \phi_{t-1,t-1}^2) = P_{t-2}^{t-3}(1 - \phi_{t-2,t-2}^2)(1 - \phi_{t-1,t-1}^2) \\ &= \dots = \gamma_x(0) \prod_{j=1}^{t-1} (1 - \phi_{jj}^2), \end{aligned}$$

for $t = 2, \dots, n$. For ARMA models, $\gamma_x(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$, in which case we may write

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

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

$$L(\boldsymbol{\beta}, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} [r_1(\boldsymbol{\beta})r_2(\boldsymbol{\beta}) \dots r_n(\boldsymbol{\beta})]^{-1/2} \exp \left[-\frac{S(\boldsymbol{\beta})}{2\sigma_w^2} \right], \quad (3.64)$$

where

$$S(\boldsymbol{\beta}) = \sum_{t=1}^n \left[\frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t(\boldsymbol{\beta})} \right]. \quad (3.65)$$

Both x_t^{t-1} and r_t are functions of $\boldsymbol{\beta}$ alone, and we make that fact explicit in (3.64)-(3.65). Given values for $\boldsymbol{\beta}$ and σ_w^2 , the likelihood may be evaluated using

** This section may be skipped without harming any living things.

the techniques of §3.5. Maximum likelihood estimation would now proceed by maximizing (3.64) with respect to β and σ_w^2 . As in the usual normal case, the estimate of the variance is

$$\hat{\sigma}_w^2 = n^{-1}S(\hat{\beta}), \quad (3.66)$$

where $\hat{\beta}$ is the value of β that minimizes the *profile* or *concentrated* likelihood

$$l(\beta) = \log [n^{-1}S(\beta)] + n^{-1} \sum_{t=1}^n \log r_t(\beta). \quad (3.67)$$

The minimization of (3.67) involves the use of a numerical optimization routine such as the methods described in the following example.

Example 3.27 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 $l(\beta)$ be a criterion function of k parameters $\beta = (\beta_1, \dots, \beta_k)$ that we wish to minimize with respect to β . For example, consider the likelihood function given by (3.67). Suppose $l(\hat{\beta})$ is the extremum that we are interested in finding, and $\hat{\beta}$ is found by solving $\partial l(\beta)/\partial \beta_j = 0$, for $j = 1, \dots, k$. Let $l^{(1)}(\beta)$ denote the $k \times 1$ vector of partials

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

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

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

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

$$\mathbf{0} = l^{(1)}(\hat{\beta}) \approx l^{(1)}(\beta_{(0)}) - l^{(2)}(\beta_{(0)}) [\hat{\beta} - \beta_{(0)}].$$

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

$$\beta_{(1)} = \beta_{(0)} + [l^{(2)}(\beta_{(0)})]^{-1} l^{(1)}(\beta_{(0)}).$$

The Newton–Raphson algorithm proceeds by iterating this result, replacing $\beta_{(0)}$ by $\beta_{(1)}$ to get $\beta_{(2)}$, and so on, until convergence. Thus, at iteration $j = 1, 2, \dots$, we have

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

Under appropriate conditions, the sequence of estimators, $\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \dots$, will converge to $\hat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$.

For MLE, the criterion function used is $l(\boldsymbol{\beta})$ given by (3.67); $l^{(1)}(\boldsymbol{\beta})$ 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 $I(\boldsymbol{\beta}) \equiv E[l^{(2)}(\boldsymbol{\beta})]$, which is called the *information matrix*. Under appropriate conditions, the inverse of the information matrix is the asymptotic variance-covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$. This is sometimes approximated by the inverse of the Hessian at $\hat{\boldsymbol{\beta}}$. That is,

$$\hat{\boldsymbol{\beta}} \sim \text{AN}(\boldsymbol{\beta}, I(\hat{\boldsymbol{\beta}})^{-1}) \quad \text{and} \quad \hat{\boldsymbol{\beta}} \sim \text{AN}(\boldsymbol{\beta}, l^{(2)}(\hat{\boldsymbol{\beta}})^{-1}).$$

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 3.28 Estimation for an AR(1)

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

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{3.68}$$

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

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

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 $x_t \mid x_{t-1} \sim N(\mu + \phi(x_{t-1} - \mu), \sigma_w^2)$, 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

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

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

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}(1-\phi^2)^{1/2} \exp \left[-\frac{S(\mu, \phi)}{2\sigma_w^2} \right], \quad (3.69)$$

where

$$S(\mu, \phi) = (1-\phi^2)(x_1 - \mu)^2 + \sum_{t=2}^n [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2. \quad (3.70)$$

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 (3.69) 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

$$\hat{\sigma}_w^2 = n^{-1}S(\hat{\mu}, \hat{\phi}), \quad (3.71)$$

where $\hat{\mu}$ and $\hat{\phi}$ are the MLEs of μ and ϕ , respectively. If we replace n in (3.71) by $n-2$, we would obtain the unconditional least squares estimate of σ_w^2 .

If, in (3.69), we take logs, replace σ_w^2 by $\hat{\sigma}_w^2$, and ignore constants, $\hat{\mu}$ and $\hat{\phi}$ are the values that minimize the criterion function

$$l(\mu, \phi) = \log [n^{-1}S(\mu, \phi)] - n^{-1} \log(1-\phi^2); \quad (3.72)$$

that is, $l(\mu, \phi) \propto -2 \log L(\mu, \phi, \hat{\sigma}_w^2)$.⁶ Because (3.70) and (3.72) 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

$$\begin{aligned} L(\mu, \phi, \sigma_w^2 \mid x_1) &= \prod_{t=2}^n f_w [(x_t - \mu) - \phi(x_{t-1} - \mu)] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2} \exp \left[-\frac{S_c(\mu, \phi)}{2\sigma_w^2} \right], \end{aligned} \quad (3.73)$$

where the conditional sum of squares is

$$S_c(\mu, \phi) = \sum_{t=2}^n [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2. \quad (3.74)$$

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

The conditional MLE of σ_w^2 is

$$\hat{\sigma}_w^2 = S_c(\hat{\mu}, \hat{\phi}) / (n - 1), \quad (3.75)$$

and $\hat{\mu}$ and $\hat{\phi}$ 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

$$S_c(\mu, \phi) = \sum_{t=2}^n [x_t - (\alpha + \phi x_{t-1})]^2. \quad (3.76)$$

The problem is now the linear regression problem stated in §2.2. Following the results from least squares estimation, we have $\hat{\alpha} = \bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}$, 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

$$\hat{\mu} = \frac{\bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}}{1 - \hat{\phi}} \quad (3.77)$$

$$\hat{\phi} = \frac{\sum_{t=2}^n (x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^n (x_{t-1} - \bar{x}_{(1)})^2}. \quad (3.78)$$

From (3.77) and (3.78), we see that $\hat{\mu} \approx \bar{x}$ and $\hat{\phi} \approx \hat{\rho}(1)$. 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 (3.75) to be equivalent to the least squares estimator, that is, divide $S_c(\hat{\mu}, \hat{\phi})$ by $(n - 3)$ instead of $(n - 1)$ in (3.75).

For general AR(p) models, maximum likelihood estimation, unconditional least squares, and conditional least squares follow analogously to the AR(1) example. For the AR(1) model, recall that $x_1^0 = \mu$ and $x_t^{t-1} = \mu + \phi(x_{t-1} - \mu)$, for $t = 2, \dots, n$. 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 $t = 2, \dots, n$. Hence, the likelihood presented in (3.69) is identical to the innovations form of the likelihood given by (3.64). Moreover, the generic $S(\boldsymbol{\beta})$ in (3.65) is $S(\mu, \phi)$ given in (3.70) and the generic $l(\boldsymbol{\beta})$ in (3.67) is $l(\mu, \phi)$ in (3.72).

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

$$x_t = \mu_t + y_t, \quad (3.79)$$

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 (3.79) 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 (3.79) is quadratic, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$, then the differenced series $\nabla^2 y_t$ 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. The basic idea is that if differencing the data at some order d produces an ARMA process, then the original process is said to be ARIMA.

Definition 3.7 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

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t. \quad (3.80)$$

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

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

$$x_{n+m}^n = y_{n+m}^n + x_{n+m-1}^n$$

with initial condition $x_{n+1}^n = y_{n+1}^n + x_n$ (noting $x_n^n = x_n$).

It is a little more difficult to obtain the prediction errors P_{n+m}^n , but for large n , the approximation used in §3.5, equation (3.59), works well. That is, the mean-squared prediction error can be approximated by

$$P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^{*2}, \quad (3.81)$$

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

To better understand forecasting integrated models, we examine the properties of some simple cases.

Example 3.29 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 $t = 1, 2, \dots$, 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, \dots, x_n , the one-step-ahead forecast is given by

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

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

$$x_{n+m}^n = m\delta + x_n, \quad (3.82)$$

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

$$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = E\left(\sum_{j=n+1}^{n+m} w_j\right)^2 = m\sigma_w^2. \quad (3.83)$$

Unlike the stationary case, as the forecast horizon grows, the prediction errors, (3.83), increase without bound and the forecasts follow a straight line with slope δ emanating from x_n .

Example 3.30 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. The model leads to a frequently used forecasting method called exponentially weighted moving averages (EWMA). We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1}, \quad (3.84)$$

with $|\lambda| < 1$, for $t = 1, 2, \dots$, 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 (3.84), 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 (3.84) 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

$$x_t = \sum_{j=1}^{\infty} (1 - \lambda) \lambda^{j-1} x_{t-j} + w_t. \quad (3.85)$$

as an approximation for large t (put $x_t = 0$ for $t \leq 0$). Verification of (3.85) is left to the reader (Problem 3.11). Using the approximation (3.85), we have that the approximate one-step-ahead predictor is

$$x_{n+1}^n = (1 - \lambda)x_n + \lambda x_n^{n-1}, \quad (3.86)$$

because $x_n^{n-1} = \sum_{j=1}^{\infty} (1 - \lambda) \lambda^{j-1} x_{n-j}$ and $w_{n+1}^n = 0$. From (3.86), we see that the new forecast is a linear combination of the old forecast and the new observation. The mean-square prediction error can be approximated using (3.81) 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 , (3.81) 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. 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). The results are displayed in Figure 3.10

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

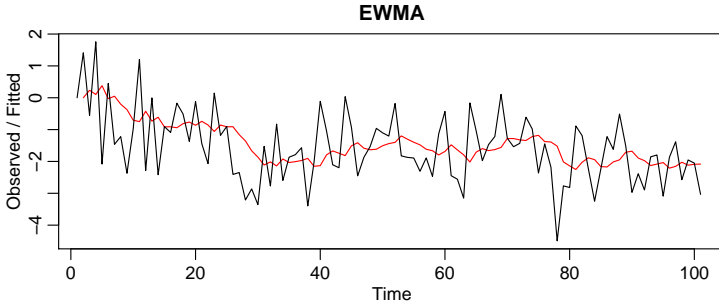


Fig. 3.10. Output for Example 3.30: Simulated data with an EWMA superimposed.

3.8 Building ARIMA Models

There are a few basic steps to fitting ARIMA models to time series data. *These steps involve (i) plotting the data, (ii) possibly transforming the data, (iii) identifying the dependence orders of the model, (iv) parameter estimation, (v) diagnostics, (vi) 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 (2.35), 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. Then a reasonable model is

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

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. Recall Figure 1.2 where $p_t \equiv .05$ and was compared to Figure 1.1. 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 $\log(1 + p_t) \approx p_t$ ⁷ 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 3.21, and we will use it again in Example 3.31.

⁷ $\log(1 + p) = p - \frac{p^2}{2} + \frac{p^3}{3} - \dots$ for $-1 < p \leq 1$. If p is a small percent-change, then the higher-order terms in the expansion are negligible.

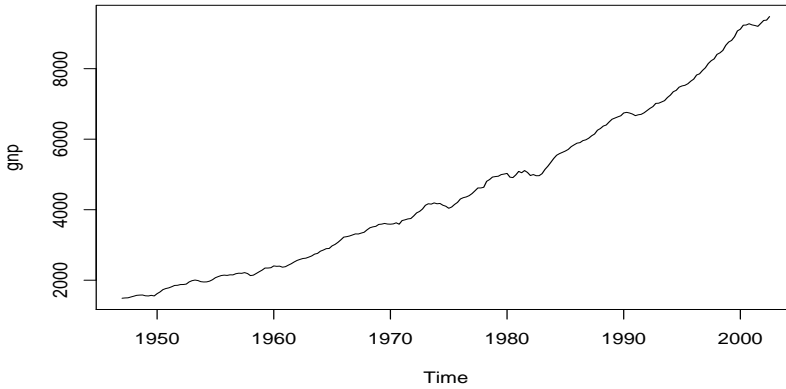


Fig. 3.11. Quarterly U.S. GNP from 1947(1) to 2002(3).

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. In addition to time plots, the sample ACF can help in indicating whether differencing is needed. Recall that the ACF of an ARMA model should decay exponentially fast to zero. Slow decay in the sample ACF may be an indication that differencing is needed.

The next step is to look at the sample ACF and PACF of (possibly differenced) data. Using Table 3.1 as a guide, preliminary values of p and q are chosen. Recall that 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, q) should be at hand, and we can start fitting models.

Example 3.31 Analysis of GNP Data

In this example, we consider the analysis of quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations. The data are real U.S. gross national product in billions of chained 1996 dollars and have been seasonally adjusted. The data were obtained from the Federal Reserve Bank of St. Louis (<http://research.stlouisfed.org/>). Figure 3.11 shows a plot of the data, say, y_t . Because strong trend hides any other effect, it is not clear from Figure 3.11 that the variance is increasing with time. For the purpose of demonstration, the sample ACF of the data is displayed in Figure 3.12. Figure 3.13 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

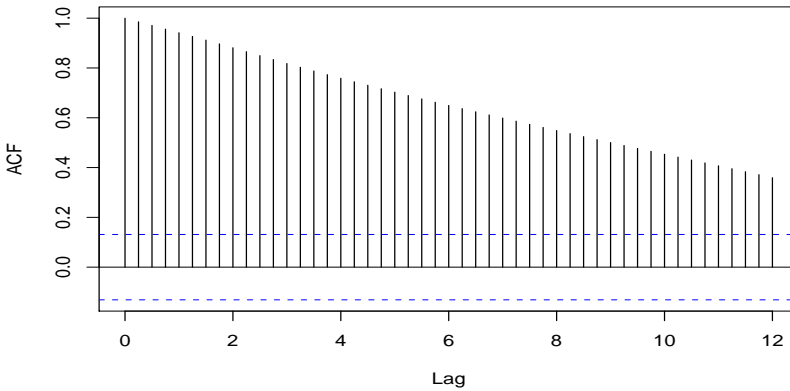


Fig. 3.12. Sample ACF of the GNP data. Lag is in terms of years.

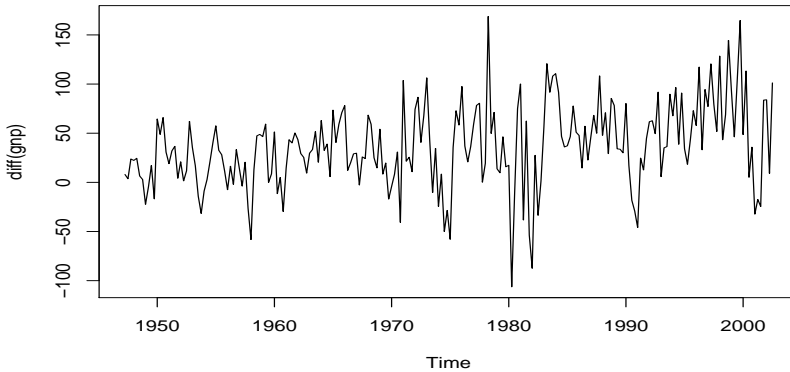


Fig. 3.13. First difference of the U.S. GNP data.

present after differencing. The growth rate, say, $x_t = \nabla \log(y_t)$, is plotted in [Figure 3.14](#), 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 3.15](#). 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 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

$$x_t = .008_{(.001)} + .303_{(.065)}\hat{w}_{t-1} + .204_{(.064)}\hat{w}_{t-2} + \hat{w}_t, \quad (3.87)$$

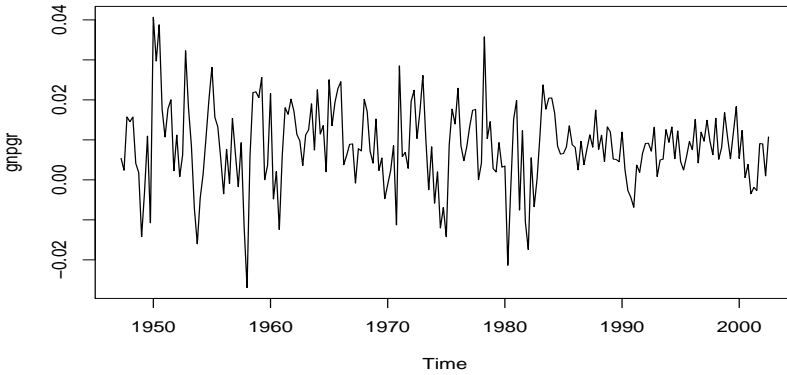


Fig. 3.14. U.S. GNP quarterly growth rate.

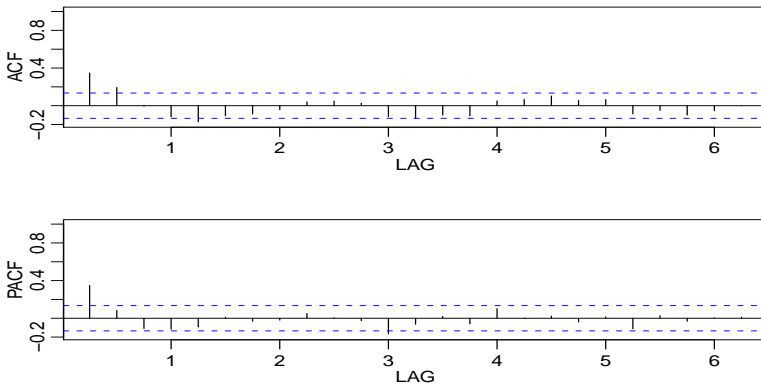


Fig. 3.15. Sample ACF and PACF of the GNP quarterly growth rate. Lag is in terms of years.

where $\hat{\sigma}_w = .0094$ 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 3.14](#)). We leave it to the reader to investigate what happens when the constant is not included.

The estimated AR(1) model is

$$x_t = .008_{(.001)} (1 - .347) + .347_{(.063)} x_{t-1} + \hat{w}_t, \quad (3.88)$$

where $\hat{\sigma}_w = .0095$ on 220 degrees of freedom; note that the constant in (3.88) 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 (3.87) and (3.88)? In fact, the fitted models are nearly the same. To show this, consider an AR(1) model of the form in (3.88) 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 (3.87).

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), $x_t - \hat{x}_t^{t-1}$, or of the standardized innovations

$$e_t = (x_t - \hat{x}_t^{t-1}) / \sqrt{\hat{P}_t^{t-1}}, \quad (3.89)$$

where \hat{x}_t^{t-1} is the one-step-ahead prediction of x_t based on the fitted model and \hat{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. It is possible in the non-Gaussian case to have an uncorrelated process for which contiguous values are highly dependent. As an example, we mention the family of GARCH models.

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, $\hat{\rho}_e(h)$, 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 $\hat{\rho}_e(h)$ 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 $\hat{\rho}_e(h)$ 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 $\hat{\rho}_e(h)$ with the main concern being the detection of obvious departures from the independence assumption.

In addition to plotting $\hat{\rho}_e(h)$, we can perform a general test of whiteness that takes into consideration the magnitudes of $\hat{\rho}_e(h)$ as a group. The Ljung–Box–Pierce Q-statistic given by

$$Q = n(n+2) \sum_{h=1}^H \frac{\hat{\rho}_e^2(h)}{n-h} \quad (3.90)$$

can be used to perform such a test. The value H in (3.90) is chosen somewhat arbitrarily, typically, $H = 20$. Under the null hypothesis of model adequacy, asymptotically ($n \rightarrow \infty$), $Q \sim \chi_{H-p-q}^2$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1-\alpha)$ -quantile of the χ_{H-p-q}^2 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](#), $n\hat{\rho}_w^2(h)$, for $h = 1, \dots, H$, are asymptotically independent χ_1^2 random variables. This means that $n \sum_{h=1}^H \hat{\rho}_w^2(h)$ is approximately a χ_H^2 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 (3.90) are used to adjust the statistic to better match the asymptotic chi-squared distribution.

Example 3.32 Diagnostics for GNP Growth Rate Example

We will focus on the MA(2) fit from [Example 3.31](#); the analysis of the AR(1) residuals is similar. [Figure 3.16](#) 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, (3.90), 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 3.16](#) 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.

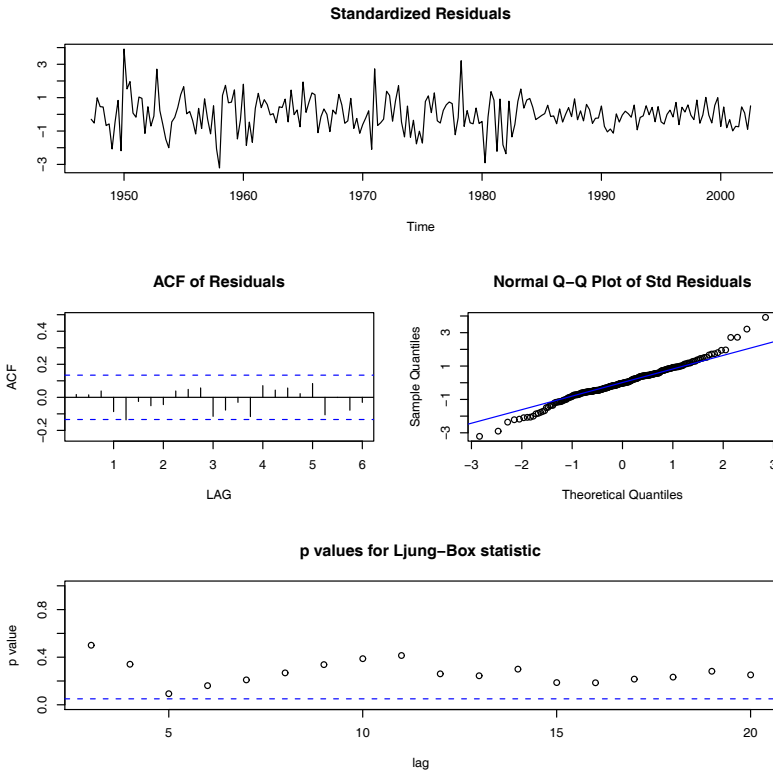


Fig. 3.16. Diagnostics of the residuals from MA(2) fit on GNP growth rate.

The model appears to fit well except for the fact that a distribution with heavier tails than the normal distribution should be employed. The diagnostics shown in [Figure 3.16](#) are a by-product of the `sarima` command from the previous example.

Example 3.33 Diagnostics for the Glacial Varve Series

In [Example 3.21](#), 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 3.17](#).

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

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

Hence the AR term is significant. The Q-statistic p-values for this model are also displayed in [Figure 3.17](#), 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

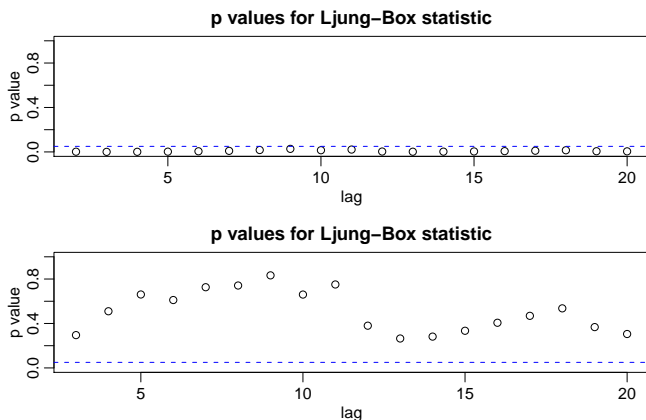


Fig. 3.17. Q-statistic p -values for the ARIMA(0,1,1) fit [top] and the ARIMA(1,1,1) fit [bottom] to the logged varve data.

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:

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

In [Example 3.31](#), 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 3.34 A Problem with Overfitting

[Figure 3.18](#) 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!

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

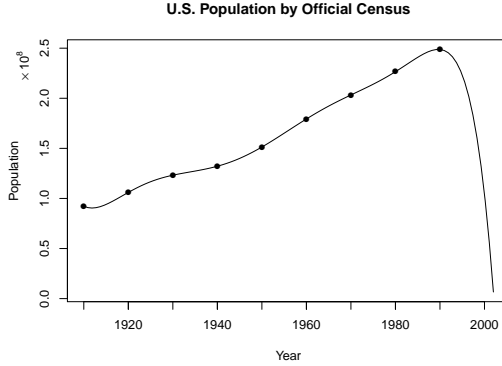


Fig. 3.18. A perfect fit and a terrible forecast.

techniques, AIC, AICc, and BIC, were described in §2.2 in the context of regression models.

Example 3.35 Model Choice for the U.S. GNP Series

Returning to the analysis of the U.S. GNP data presented in [Example 3.31](#) and [Example 3.32](#), 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 3.31](#), but for convenience, we display them again here (recall the growth rate data are in `gnpgr`):

```
sarima(gnpgr, 1, 0, 0) # AR(1)
    $AIC: -8.294403    $AICc: -8.284898    $BIC: -9.263748
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. In this case, it is reasonable to retain the AR(1) because pure autoregressive models are easier to work with.

3.9 Regression with Autocorrelated Errors

In §2.2, we covered the classical regression model with uncorrelated errors w_t . In this section, we discuss the modifications that might be considered when the errors are correlated. That is, consider the regression model

$$y_t = \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + x_t = \sum_{j=1}^r \beta_j z_{tj} + x_t \quad (3.91)$$

where x_t is a process with some covariance function $\gamma_x(s, t)$. In ordinary least squares, the assumption is that x_t is white Gaussian noise, in which case

$\gamma_x(s, t) = 0$ for $s \neq t$ and $\gamma_x(t, t) = \sigma^2$, independent of t . If this is not the case, then weighted least squares should be used.⁸

In the time series case, it is often possible to assume a stationary covariance structure for the error process x_t that corresponds to a linear process and try to find an ARMA representation for x_t . For example, if we have a pure AR(p) error, then

$$\phi(B)x_t = w_t,$$

and $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ is the linear transformation that, when applied to the error process, produces the white noise w_t . Multiplying the regression equation through by the transformation $\phi(B)$ yields,

$$\underbrace{\phi(B)y_t}_{y_t^*} = \sum_{j=1}^r \beta_j \underbrace{\phi(B)z_{tj}}_{z_{tj}^*} + \underbrace{\phi(B)x_t}_{w_t},$$

and we are back to the linear regression model where the observations have been transformed so that $y_t^* = \phi(B)y_t$ is the dependent variable, $z_{tj}^* = \phi(B)z_{tj}$ for $j = 1, \dots, r$, are the independent variables, but the β s are the same as in the original model. For example, if $p = 1$, then $y_t^* = y_t - \phi y_{t-1}$ and $z_{tj}^* = z_{tj} - \phi z_{t-1,j}$.

In the AR case, we may set up the least squares problem as minimizing the error sum of squares

$$S(\phi, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\phi(B)y_t - \sum_{j=1}^r \beta_j \phi(B)z_{tj} \right]^2$$

with respect to all the parameters, $\phi = \{\phi_1, \dots, \phi_p\}$ and $\beta = \{\beta_1, \dots, \beta_r\}$. Of course, this is done using numerical methods.

If the error process is ARMA(p, q), i.e., $\phi(B)x_t = \theta(B)w_t$, then in the above discussion, we transform by $\pi(B)x_t = w_t$, where, recalling (3.17), $\pi(B) = \theta(B)^{-1}\phi(B)$. In this case the error sum of squares also depends on $\theta = \{\theta_1, \dots, \theta_q\}$:

$$S(\phi, \theta, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\pi(B)y_t - \sum_{j=1}^r \beta_j \pi(B)z_{tj} \right]^2$$

At this point, the main problem is that we do not typically know the behavior of the noise x_t prior to the analysis. An easy way to tackle this problem was first presented in Cochrane and Orcutt (1949), and with the advent of cheap computing is modernized below:

⁸ Write the model in vector notation as $\mathbf{y} = Z\beta + \mathbf{x}$, and let $\Gamma = \{\gamma_x(s, t)\}$. Then $\Gamma^{-1/2}\mathbf{y} = \Gamma^{-1/2}Z\beta + \Gamma^{-1/2}\mathbf{x}$, so that we can write the model as $\mathbf{y}^* = Z^*\beta + \delta$, where the covariance matrix of δ is the identity. It follows that the weighted estimate of β is $\hat{\beta}_w = (Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}\mathbf{y}$, and the variance-covariance matrix of the estimator is $\text{var}(\hat{\beta}_w) = (Z'\Gamma^{-1}Z)^{-1}$. If x_t is white noise, then $\Gamma = \sigma^2 I$ and these results reduce to the usual least squares results.

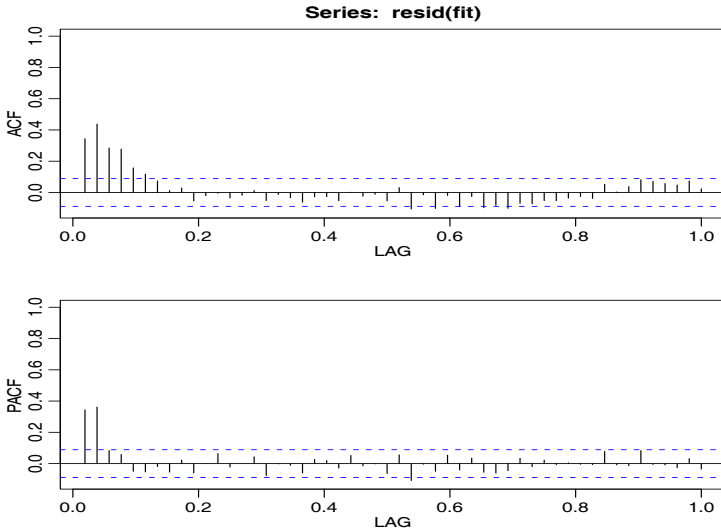


Fig. 3.19. Sample ACF and PACF of the mortality residuals indicating an AR(2) process.

- (i) First, run an ordinary regression of y_t on z_{t1}, \dots, z_{tr} (acting as if the errors are uncorrelated). Retain the residuals, $\hat{x}_t = y_t - \sum_{j=1}^r \hat{\beta}_j z_{tj}$.
- (ii) Identify ARMA model(s) for the residuals \hat{x}_t .
- (iii) Run weighted least squares (or MLE) on the regression model with autocorrelated errors using the model specified in step (ii).
- (iv) Inspect the residuals \hat{w}_t for whiteness, and adjust the model if necessary.

Example 3.36 Mortality, Temperature and Pollution

We consider the analyses presented in [Example 2.2](#), relating mean adjusted temperature T_t , and particulate levels P_t to cardiovascular mortality M_t . We consider the regression model

$$M_t = \beta_1 + \beta_2 t + \beta_3 T_t + \beta_4 T_t^2 + \beta_5 P_t + x_t, \quad (3.92)$$

where, for now, we assume that x_t is white noise. The sample ACF and PACF of the residuals from the ordinary least squares fit of (3.92) are shown in [Figure 3.19](#), and the results suggest an AR(2) model for the residuals.

Our next step is to fit the correlated error model (3.92), but where x_t is AR(2),

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

and w_t is white noise. The model can be fit using the `arma` function as follows (partial output shown).

```
trend = time(cmort); temp = tempr - mean(tempr); temp2 = temp^2
fit = lm(cmort~trend + temp + temp2 + part, na.action=NULL)
acf2(resid(fit), 52) # implies AR2
```

```

sarima(cmort, 2,0,0, xreg=cbind(trend,temp,temp2,part))
Coefficients:
      ar1      ar2  intercept      trend      temp      temp2      part
    0.3848  0.4326   80.2116  -1.5165  -0.0190  0.0154  0.1545
s.e.  0.0436  0.0400    1.8072   0.4226   0.0495  0.0020  0.0272
sigma^2 estimated as 26.01: loglikelihood = -1549.04, aic = 3114.07

```

The residual analysis output from `sarima` (not shown) shows no obvious departure of the residuals from whiteness.

3.10 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, $\text{ARMA}(P, Q)_s$, then takes the form

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \quad (3.93)$$

where the operators

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps} \quad (3.94)$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs} \quad (3.95)$$

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 $\text{ARMA}(P, Q)_s$ 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 3.37 A Seasonal AR Series

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

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

or

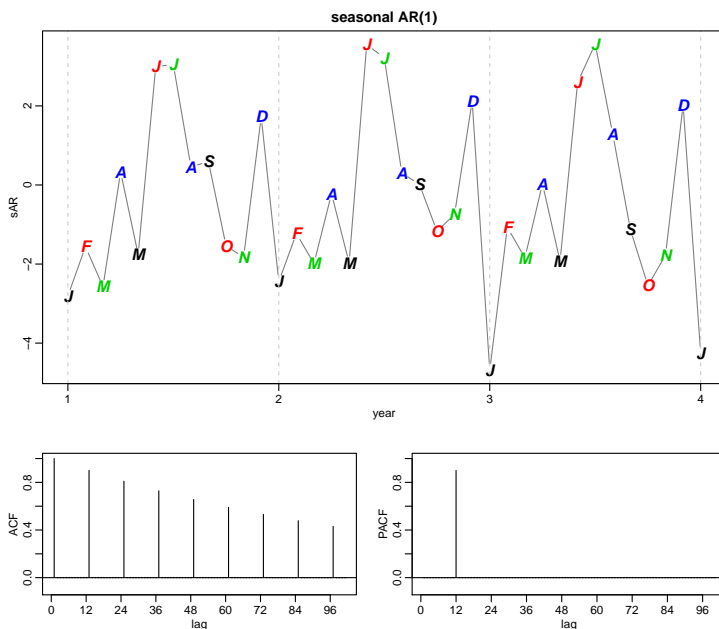


Fig. 3.20. Data generated from a seasonal ($s = 12$) AR(1), and the true ACF and PACF of the model $x_t = .9x_{t-12} + w_t$.

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

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 $|\Phi| < 1$.

We simulated 3 years of data from the model with $\Phi = .9$, and exhibit the *theoretical* ACF and PACF of the model. See [Figure 3.20](#).

```
set.seed(666)
phi = c(rep(0,11),.9)
sAR = arima.sim(list(order=c(12,0,0), ar=phi), n=37)
sAR = ts(sAR, freq=12)
layout(matrix(c(1,2, 1,3), nc=2))
par(mar=c(3,3,2,1), mgp=c(1.6,.6,0))
plot(sAR, axes=FALSE, main='seasonal AR(1)', xlab='year', type='c')
Months = c("J","F","M","A","M","J","J","A","S","O","N","D")
points(sAR, pch=Months, cex=1.25, font=4, col=1:4)
axis(1, 1:4); abline(v=1:4, lty=2, col='#cccccc')
axis(2); box()
ACF = ARMAacf(ar=phi, ma=0, 100)
PACF = ARMAacf(ar=phi, ma=0, 100, pacf=TRUE)
plot(ACF,type="h", xlab="lag", ylim=c(-.1,1)); abline(h=0)
plot(PACF, type="h", xlab="lag", ylim=c(-.1,1)); abline(h=0)
```

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

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

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \dots$, 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

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

In this case, the only non-zero correlations are

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

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. These results are demonstrated in [Figure 3.20](#).

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

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \quad (3.96)$$

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 Table 3.1 and Table 3.3. In fitting such models, focusing on the seasonal autoregressive and moving average components first generally leads to more satisfactory results.

Example 3.38 A Mixed Seasonal Model

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

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

The ACF and PACF for this model, with $\Phi = .8$ and $\theta = -.5$, are shown in Figure 3.21. These type of correlation relationships, although idealized here, are typically seen with seasonal data.

To reproduce Figure 3.21 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 persistence occurs 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 a little from one year to the next, according to a random walk,

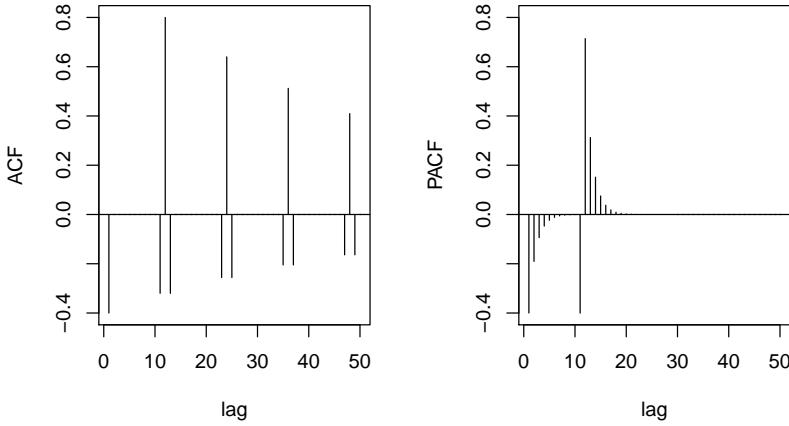


Fig. 3.21. ACF and PACF of the mixed seasonal ARMA model $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

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

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 $k = 1, 2, \dots$. 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 $MA(1)_{12}$, 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

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \quad (3.97)$$

where $D = 1, 2, \dots$, 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 3.8 *The multiplicative seasonal autoregressive integrated moving average model, or SARIMA model is given by*

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t, \quad (3.98)$$

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, and the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ of orders P and Q and ordinary and seasonal difference components by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.

Example 3.39 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 $\text{ARIMA}(0, 1, 1) \times (0, 1, 1)_{12}$ in the notation given above, where the seasonal fluctuations occur every 12 months. Then, with $\delta = 0$, the model (3.98) becomes

$$\nabla_{12}\nabla x_t = \Theta(B^{12})\theta(B)w_t$$

or

$$(1 - B^{12})(1 - B)x_t = (1 + \Theta B^{12})(1 + \theta B)w_t. \quad (3.99)$$

Expanding both sides of (3.99) 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 (3.98) is a daunting task, and we usually 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 Table 3.1 and Table 3.3. In considering whether the model is satisfactory, the diagnostic techniques discussed in §3.8 still apply.

Example 3.40 Air Passengers

We consider the R data set `AirPassengers`, which are the monthly totals of international airline passengers, 1949 to 1960, taken from Box & Jenkins (1970). Various plots of the data and transformed data are shown in Figure 3.22 and were obtained as follows:

```
x = AirPassengers
lx = log(x); dlx = diff(lx); ddx = diff(dlx, 12)
plot.ts(cbind(x, lx, dlx, ddx), main="")
# of interest for showing seasonal RW (not shown here):
dev.new()
par(mfrow=c(2,1), mar=c(3,3,1,1), mgp=c(1.6,.6,0))
monthplot(dlx); monthplot(ddlx)
```

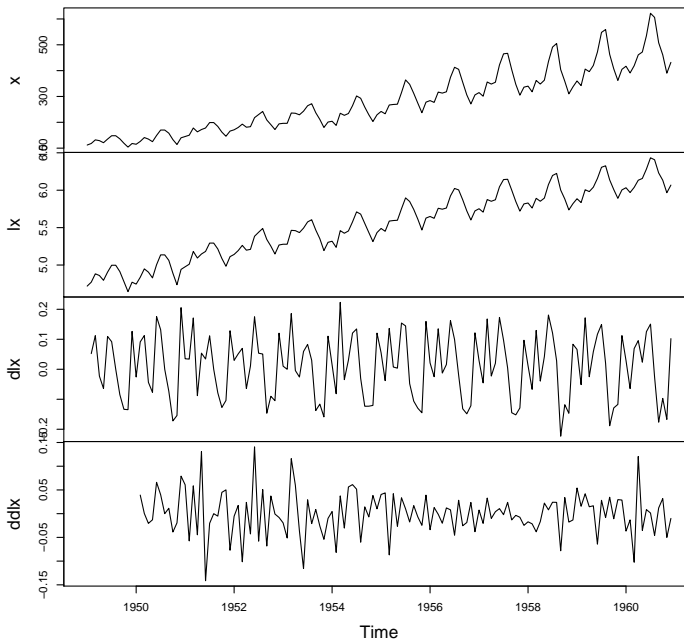


Fig. 3.22. R data set **AirPassengers**, which are the monthly totals of international airline passengers \mathbf{x} , and the transformed data: $\mathbf{lx} = \log x_t$, $\mathbf{dlx} = \nabla \log x_t$, and $\mathbf{ddlx} = \nabla_{12} \nabla \log x_t$.

Note that \mathbf{x} is the original series, which shows trend plus increasing variance. The logged data are in \mathbf{lx} , and the transformation stabilizes the variance. The logged data are then differenced to remove trend, and are stored in \mathbf{dlx} . It is clear there is still persistence in the seasons (i.e., $\mathbf{dlx}_t \approx \mathbf{dlx}_{t-12}$), so that a twelfth-order difference is applied and stored in \mathbf{ddlx} . The transformed data appears to be stationary and we are now ready to fit a model.

The sample ACF and PACF of \mathbf{ddlx} ($\nabla_{12} \nabla \log x_t$) are shown in **Figure 3.23**. The R code is:

```
acf2(ddlx,50)
```

Between Seasons: It appears that at the seasons, the ACF is cutting off a lag $1s$ ($s = 12$), whereas the PACF is tailing off at lags $1s, 2s, 3s, 4s, \dots$. These results implies an $\text{SMA}(1)$, $P = 0$, $Q = 1$, in the season ($s = 12$).

Within Seasons: Inspecting the sample ACF and PACF at the lower lags, it appears as though both are tailing off. This suggests an $\text{ARMA}(1, 1)$ within the seasons, $p = q = 1$.

Thus, we first try an $\text{ARIMA}(1, 1, 1) \times (0, 1, 1)_{12}$ on the logged data:

```
sarima(lx, 1,1,1, 0,1,1,12)
```

Coefficients:

	ar1	ma1	sma1
	0.1960	-0.5784	-0.5643
s.e.	0.2475	0.2132	0.0747
sigma^2 estimated as 0.001341			
\$AIC	-5.5726	\$AICc	-5.556713
		\$BIC	-6.510729

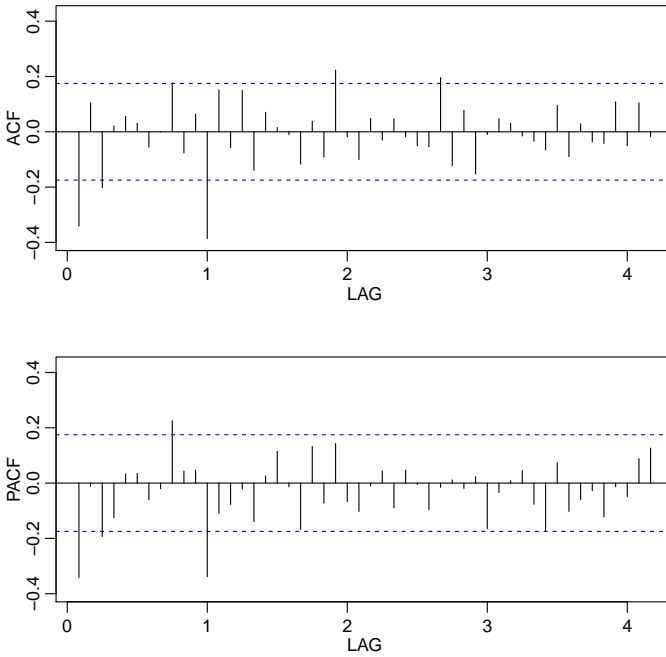


Fig. 3.23. Sample ACF and PACF of $dd1x$ ($\nabla_{12}\nabla \log x_t$).

However, the AR parameter is not significant, so we should try dropping one parameter from the within seasons part. In this case, we try both an $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$ and an $ARIMA(1, 1, 0) \times (0, 1, 1)_{12}$ model:

```
sarima(lx, 0,1,1, 0,1,1,12)
Coefficients:
      ma1      sma1
    -0.4018  -0.5569
s.e.    0.0896   0.0731
sigma^2 estimated as 0.001348
$AIC -5.58133 $AICc -5.56625 $BIC -6.540082
```

```
sarima(lx, 1,1,0, 0,1,1,12)
Coefficients:
      ar1      sma1
    -0.3395  -0.5619
s.e.    0.0822   0.0748
sigma^2 estimated as 0.001367
$AIC -5.567081 $AICc -5.552002 $BIC -6.525834
```

All information criteria prefer the $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$ model, which is the model displayed in (3.99). The residual diagnostics are shown in Figure 3.24, and except for one or two outliers, the model seems to fit well.

Finally, we forecast the logged data out twelve months, and the results are shown in Figure 3.25.

```
sarima.for(lx, 12, 0,1,1, 0,1,1,12)
```

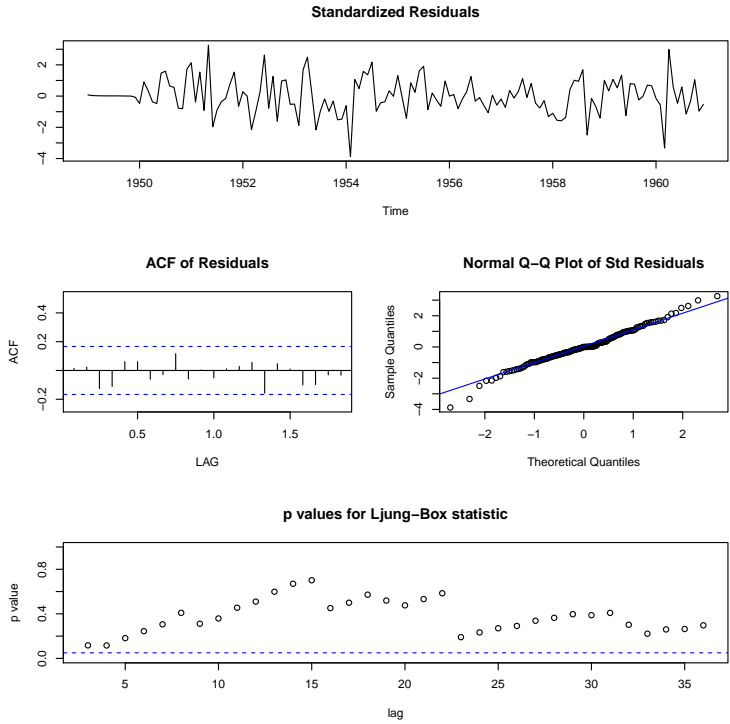


Fig. 3.24. Residual analysis for the $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$ fit to the logged air passengers data set.

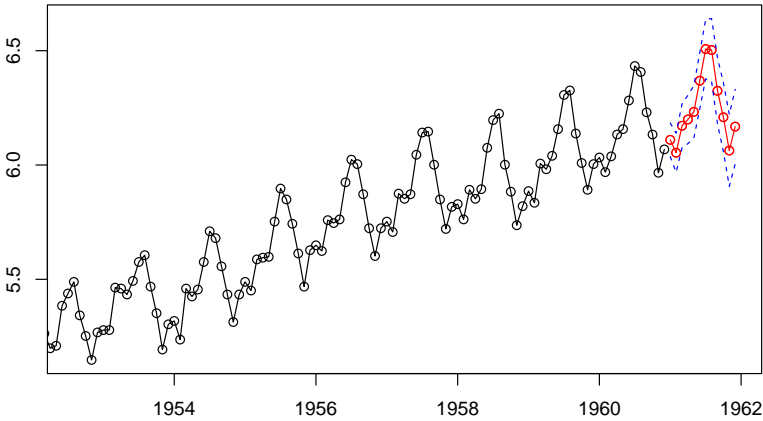


Fig. 3.25. Twelve month forecast using the $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$ model on the logged air passenger data set.

Problems

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

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

- (a) Find the mean and the variance of $\{x_t, t = 1, 2, \dots\}$. 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 $0 \leq h < t$. *Hint:* Write $x_t = \phi^h x_{t-h} + \sum_{j=0}^{h-1} \phi_j w_{t-j}$.

- (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? *Hint:* Show $\text{var}(x_t)$ is constant and use part (b).

3.3 Using [Example 3.7](#) as a guide, identify the following models as ARMA(p, q) models (watch out for parameter redundancy), and determine whether they are causal and/or invertible. If the model is causal, use R to find the first 10 ψ -weights, and if the model is invertible, use R to find the first 10 π -weights.

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

3.4 For the AR(2) model given by $x_t = -.9x_{t-2} + w_t$, follow the R code in [Example 3.12](#) to find the roots of the autoregressive polynomial, find the pseudo period of the process, and then plot the theoretical ACF, $\rho(h)$.

3.5 (a) Compare the *theoretical* ACF and PACF of an ARMA(1,1), an ARMA(1,0), and an ARMA(0,1) series by plotting the ACFs and PACFs of the three series for $\phi = .6$, $\theta = .9$. Comment on the capability of the ACF and PACF to determine the order of the models. *Hint:* See the code for [Example 3.14](#).

- (b) Use `arima.sim` to generate $n = 100$ observations from each of the three models discussed in (a). Compute the sample ACFs and PACFs for each model and compare it to the theoretical values. How do the results compare with the general results given in Table 3.1?
- (c) Repeat (b) but with $n = 500$. Comment.

3.6 Let c_t be the cardiovascular mortality series (`cmort`) discussed in Chapter 2, Example 2.2 and let $x_t = \nabla c_t$ be the differenced data.

- (a) Plot x_t and compare it to the actual data plotted in Figure 2.2. Why does differencing seem reasonable in this case?
- (b) Calculate and plot the sample ACF and PACF of x_t and using Table 3.1, argue that an AR(1) is appropriate for x_t .
- (c) Fit an AR(1) to x_t using maximum likelihood (basically unconditional least squares) as in Section 3.6. The easiest way to do this is to use `sarima` from `astsa`. Comment on the significance of the regression parameter estimates of the model. What is the estimate of the white noise variance?
- (d) Examine the residuals and comment on whether or not you think the residuals are white.
- (e) Assuming the fitted model is the true model, find the forecasts over a four-week horizon, x_{n+m}^n , for $m = 1, 2, 3, 4$, and the corresponding 95% prediction intervals; $n = 508$ here. The easiest way to do this is to use `sarima.for` from `astsa`.
- (f) Show how the values obtained in part (e) were calculated.
- (g) What is the one-step-ahead forecast of the actual value of cardiovascular mortality; i.e., what is c_{n+1}^n ?

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

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

3.8 Repeat the following numerical exercise three times. Generate $n = 500$ iid $N(0, 1)$ observations. Fit an ARMA(1, 1) model to the data. Compare the parameter estimates in each case and explain the results.

3.9 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.

3.10 Using Example 3.20 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, \dots, 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.

3.11 Verify that the IMA(1,1) model given in (3.84) can be inverted and written as (3.85).

3.12 For the logarithm of the glacial varve data, say, x_t , presented in [Example 3.21](#), use the first 100 observations and calculate the EWMA, x_{n+1}^n , discussed in [Example 3.30](#), for $n = 1, \dots, 100$, using $\lambda = .25, .50$, and $.75$, and plot the EWMA's and the data superimposed on each other. Comment on the results.

3.13 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.

3.14 Fit an ARIMA(p, d, q) model to the global temperature data `gtemp2` in `astsa` performing all of the necessary diagnostics. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

3.15 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.

3.16 Let S_t represent the monthly sales data in `sales` ($n = 150$), and let L_t be the leading indicator in `lead`.

- Fit an ARIMA model to S_t , the monthly sales data. Discuss your model fitting in a step-by-step fashion, presenting your (A) initial examination of the data, (B) transformations, if necessary, (C) initial identification of the dependence orders and degree of differencing, (D) parameter estimation, (E) residual diagnostics and model choice.
- Use the CCF and lag plots between ∇S_t and ∇L_t to argue that a regression of ∇S_t on ∇L_{t-3} is reasonable. [Note that in `lag2.plot()`, the first named series is the one that gets lagged.]
- Fit the regression model $\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t$, where x_t is an ARMA process (explain how you decided on your model for x_t). Discuss your results. *R help: If you have to work with various transformations of series in `x` and `y`, first align the data:*

```
dog = ts.intersect( lag(x,-11), diff(y,97) )
xnew = dog[,1] # dog has 2 columns, the first is lag(x,-11) ...
ynew = dog[,2] # ... and the second column is diff(y,97)
plot(dog)      # now you can manipulate xnew and ynew simultaneously
lag2.plot(xnew, ynew, 5)
```

3.17 Redo [Problem 2.2](#) without assuming the error term is white noise.

3.18 Redo [Problem 2.1](#) without assuming the error term is white noise.

3.19 Plot the theoretical ACF of the seasonal ARIMA(0, 1) \times (1, 0)₁₂ model with $\Phi = .8$ and $\theta = .5$ out to lag 50.

3.20 Fit a seasonal ARIMA model of your choice to the unemployment data, `unemp`. Use the estimated model to forecast the next 12 months.

3.21 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.

3.22 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.

Spectral Analysis and Filtering

4.1 Introduction

The cyclic behavior of data is the focus of this chapter. 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 1.6](#) 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

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 $t = 0, \pm 1, \pm 2, \dots$, 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 identity¹ and write (4.1) as

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

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.

If we assume that U_1 and U_2 are iid random variables with mean 0 and variance σ^2 , then x_t in (4.2) is stationary with mean $E(x_t) = 0$ and, writing $c_t = \cos(2\pi\omega t)$ and $s_t = \sin(2\pi\omega t)$, autocovariance function

$$\begin{aligned} \gamma(h) &= \text{cov}(x_{t+h}, x_t) = \text{cov}(U_1 c_{t+h} + U_2 s_{t+h}, U_1 c_t + U_2 s_t) \\ &= \text{cov}(U_1 c_{t+h}, U_1 c_t) + \text{cov}(U_1 c_{t+h}, U_2 s_t) \\ &\quad + \text{cov}(U_2 s_{t+h}, U_1 c_t) + \text{cov}(U_2 s_{t+h}, U_2 s_t) \\ &= \sigma^2 c_{t+h} c_t + 0 + 0 + \sigma^2 s_{t+h} s_t = \sigma^2 \cos(2\pi\omega h), \end{aligned} \quad (4.3)$$

using [footnote 1](#) and noting that $\text{cov}(U_1, U_2) = 0$.

The random process in (4.2) is function of its frequency, ω . 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, we 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.

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

$$x_t = \sum_{k=1}^q [U_{k1} \cos(2\pi\omega_k t) + U_{k2} \sin(2\pi\omega_k t)], \quad (4.4)$$

where U_{k1}, U_{k2} , for $k = 1, 2, \dots, q$, are independent zero-mean random variables with variances σ_k^2 , and the ω_k are distinct frequencies. Notice that (4.4) exhibits the process as a sum of independent components, with variance σ_k^2

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

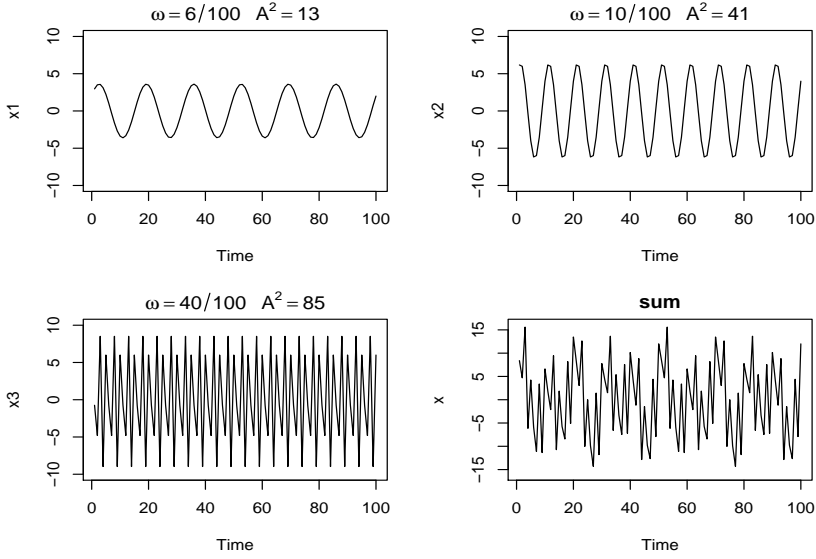


Fig. 4.1. Periodic components and their sum as described in [Example 4.1](#).

for frequency ω_k . As in (4.3), it is easy to show ([Problem 4.2](#)) that the autocovariance function of the process is

$$\gamma(h) = \sum_{k=1}^q \sigma_k^2 \cos(2\pi\omega_k h), \quad (4.5)$$

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

$$\gamma(0) = E(x_t^2) = \sum_{k=1}^q \sigma_k^2, \quad (4.6)$$

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 (4.4) with $q = 3$ constructed in the following way. First, for $t = 1, \dots, 100$, we generated three series

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

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")
```

The model given in (4.4), along with its autocovariance given (4.5), is a population construct. If the model is correct, our next step would be to estimate the variances σ_k^2 and frequencies ω_k that form the model (4.4). In the next example, we consider the problem of estimation of these quantities.

Example 4.2 Estimation and the Periodogram

For any time series sample x_1, \dots, x_n , where n is odd, we may write, *exactly*

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

for $t = 1, \dots, n$ 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 (4.4) 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.

Using the regression results from Chapter 2, the coefficients a_j and b_j are of the form $\sum_{t=1}^n x_t z_{tj} / \sum_{t=1}^n z_{tj}^2$, where z_{tj} is either $\cos(2\pi t j/n)$ or $\sin(2\pi t j/n)$. Using [Problem 4.21](#), $\sum_{t=1}^n z_{tj}^2 = n/2$ when $j/n \neq 0, 1/2$, so the regression coefficients in (4.7) can be written as

$$a_j = \frac{2}{n} \sum_{t=1}^n x_t \cos(2\pi t j/n) \quad \text{and} \quad b_j = \frac{2}{n} \sum_{t=1}^n x_t \sin(2\pi t j/n).$$

We then define the *scaled periodogram* to be

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

because it indicates which frequency components in (4.7) are large in magnitude and which components are small. *The scaled periodogram is the estimate of σ_j^2 corresponding to the sinusoid oscillating at a frequency of $\omega_j = j/n$, or j cycles in n time points.* These particular frequencies are called the *Fourier* or *fundamental frequencies*. Large values of $P(j/n)$ indicate which frequencies $\omega_j = j/n$ are predominant in the series, whereas small values of $P(j/n)$ may be associated with noise.

It is not necessary to run a large regression to obtain the values of a_j and b_j because they can be computed quickly if n is a highly composite integer. Although we will discuss it in more detail in §4.4, the discrete Fourier transform (DFT) is a complex-valued weighted average of the data given by

$$\begin{aligned} d(j/n) &= n^{-1/2} \sum_{t=1}^n x_t \exp(-2\pi i t j/n) \\ &= n^{-1/2} \left(\sum_{t=1}^n x_t \cos(2\pi t j/n) - i \sum_{t=1}^n x_t \sin(2\pi t j/n) \right), \end{aligned} \quad (4.9)$$

for $j = 0, 1, \dots, n-1$, where the frequencies j/n are called the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (4.9) may be computed quickly using the fast Fourier transform (FFT). Note that

$$|d(j/n)|^2 = \frac{1}{n} \left(\sum_{t=1}^n x_t \cos(2\pi t j/n) \right)^2 + \frac{1}{n} \left(\sum_{t=1}^n x_t \sin(2\pi t j/n) \right)^2 \quad (4.10)$$

and it is this quantity that is called the periodogram. We may calculate the scaled periodogram, (4.8), using the periodogram as

$$P(j/n) = \frac{4}{n} |d(j/n)|^2. \quad (4.11)$$

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

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

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

Useful information: Euler's formula: $e^{i\alpha} = \cos(\alpha) + i \sin(\alpha)$. Consequently, $\cos(\alpha) = \frac{e^{i\alpha} + e^{-i\alpha}}{2}$, and $\sin(\alpha) = \frac{e^{i\alpha} - e^{-i\alpha}}{2i}$. Also, $\frac{1}{i} = -i$ because $-i \times i = 1$. If $z = a - ib$ is complex, then $|z|^2 = z\bar{z} = (a - ib)(a + ib) = a^2 + b^2$.

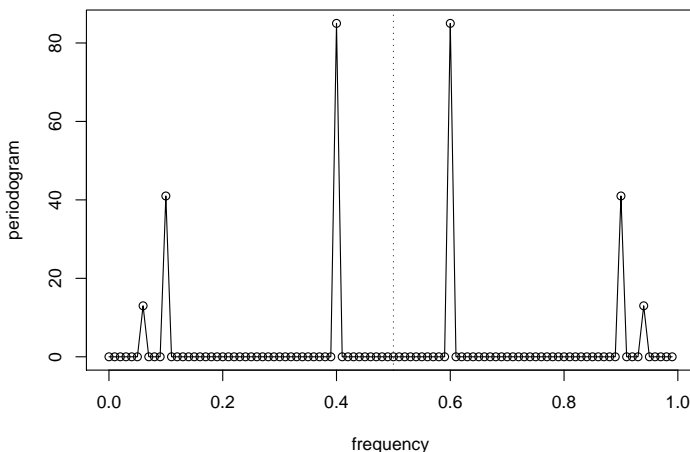


Fig. 4.2. Periodogram of the data generated in [Example 4.1](#).

$$P\left(\frac{6}{100}\right) = P\left(\frac{94}{100}\right) = 13, \quad P\left(\frac{10}{100}\right) = P\left(\frac{90}{100}\right) = 41, \quad P\left(\frac{40}{100}\right) = P\left(\frac{60}{100}\right) = 85,$$

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, (4.6), of a real set of data.

Assuming the simulated data, \mathbf{x} , were retained from the previous example, the R code to reproduce [Figure 4.2](#) is

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

Different packages scale the FFT differently, so it is a good idea to consult the documentation. R computes it without the factor $n^{-1/2}$ and with an additional factor of $e^{2\pi i \omega_j}$ that can be ignored because we will be interested in the squared modulus.

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*.

The following is an example using actual data.

Example 4.3 Star Magnitude

The data in [Figure 4.3](#) are the magnitude of a star taken at midnight for 600 consecutive days. The data are taken from the classic text, *The Calculus of Observations, a Treatise on Numerical Mathematics*, by E.T. Whittaker and G. Robinson, (1923, Blackie & Son, Ltd.).

The periodogram for frequencies less than .08 is also displayed in the figure; the periodogram for frequencies higher than .08 are essentially zero. Note that the 29 day cycle and the 25 day cycle are the most prominent periodic components of the data.

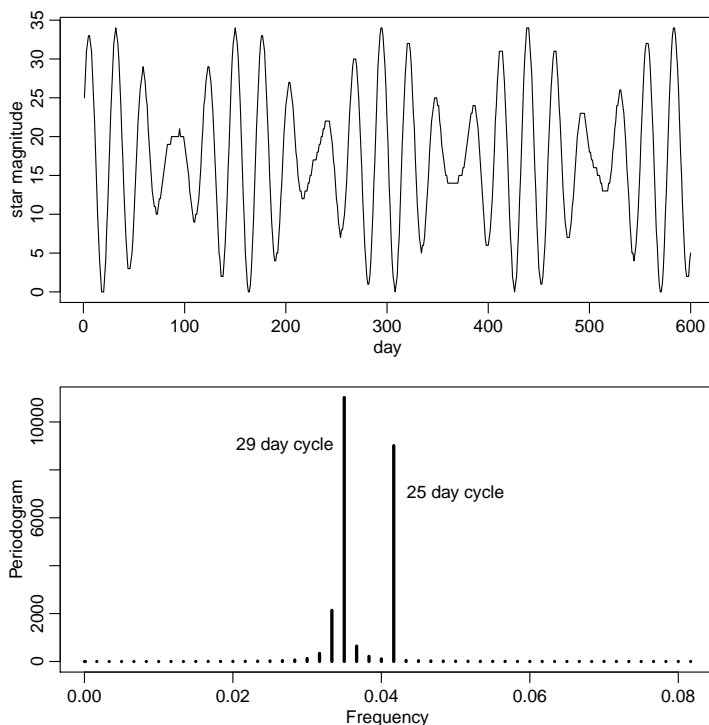


Fig. 4.3. Star magnitudes and part of the corresponding periodogram.

The R code to reproduce **Figure 4.3** is

```
n = length(star)
par(mfrow=c(2,1), mar=c(3,3,1,1), mgp=c(1.6,.6,0))
plot(star, ylab="star magnitude", xlab="day")
Per = Mod(fft(star-mean(star)))^2/n
Freq = (1:n -1)/n
plot(Freq[1:50], Per[1:50], type='h', lwd=3, ylab="Periodogram",
     xlab="Frequency")
u = which.max(Per[1:50])           # 22 freq=21/600=.035 cycles/day
uu = which.max(Per[1:50][-u])     # 25 freq=24/600=.040 cycles/day
1/Freq[22]; 1/Freq[25]           # period = days/cycle
text(.05, 7000, "25 day cycle"); text(.027, 9000, "29 day cycle")
```

The periodogram, which was introduced in Schuster (1898) and used in Schuster (1906) for studying the periodicities in the sunspot series (shown in **Figure 4.22** in the Problems section) is a sample based statistic. In **Example 4.2** and **Example 4.3**, we discussed the fact that the periodogram may be giving us an idea of the variance components associated with each frequency, as presented in (4.6), 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. The result, called the *Spectral Representation Theorem*, is quite technical, but the essence of the theorem is that (4.4) is approximately true for any stationary time series.

The examples in the previous section are not generally realistic because time series are rarely the superposition of a small number of distinct sinusoids. An important situation we use repeatedly is the case when the autocovariance function is absolutely summable, in which case there is a spectral density.

Property 4.1 The Spectral Density

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

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \quad (4.12)$$

then it has the representation

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i \omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \dots, \quad (4.13)$$

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

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2. \quad (4.14)$$

The examples of the previous section were analogues of probability mass functions, or discrete distributions. The pictures of the periodogram in Figure 4.2 and Figure 4.3 are akin to histograms. The spectral density is the analogue of the probability density function, or of continuous distributions.

The fact that $\gamma(h)$ is non-negative definite ensures $f(\omega) \geq 0$ for all ω . It follows immediately from (4.14) that

$$f(\omega) = f(-\omega)$$

verifying the spectral density is an even function. Because of the evenness, we will typically only plot $f(\omega)$ for $\omega \geq 0$. In addition, putting $h = 0$ in (4.13) 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.

We note that the absolute summability condition, (4.12), is not satisfied by (4.5), 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 1.9. 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 (4.14), 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 4.4 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 4.5, by using the additional fact that $\psi(z) = \theta(z)/\phi(z)$. The result is analogous to the fact that if $X = aY$, then $\text{var}(X) = a^2 \text{var}(Y)$.

Property 4.2 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

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

where $\phi(z) = 1 - \sum_{k=1}^p \phi_k z^k$, $\theta(z) = 1 + \sum_{k=1}^q \theta_k z^k$, and $\psi(z) = \sum_{k=0}^{\infty} \psi_k z^k$.

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.

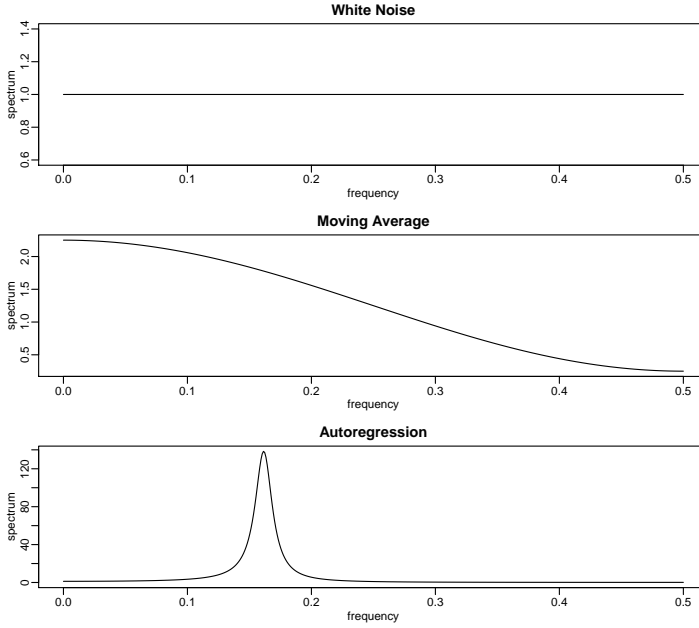


Fig. 4.4. Theoretical spectra of white noise (top), a first-order moving average (middle), and a second-order autoregressive process (bottom).

The autocovariance function is displayed in [Example 3.4](#) on page 61, 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 [\(4.14\)](#), we have

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

We can also compute the spectral density using [Property 4.2](#), which states that for an MA, $f(\omega) = \sigma_w^2 |\theta(e^{-2\pi i \omega})|^2$. Because $\theta(z) = 1 + .5z$, we have

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

which leads to agreement with [\(4.16\)](#).

Plotting the spectrum for $\sigma_w^2 = 1$, as in the middle of [Figure 4.4](#), 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 1.10** 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 4.2**, note that $\theta(z) = 1$, $\phi(z) = 1 - z + .9z^2$ and

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

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

To reproduce **Figure 4.4**, use the `arma.spec` script from `astsa` (see §R.1):

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

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, \dots, x_n , we define the **discrete Fourier transform (DFT)** to be

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega_j t} \quad (4.17)$$

for $j = 0, 1, \dots, n-1$, 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). 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,

$$x_t = n^{-1/2} \sum_{j=0}^{n-1} d(\omega_j) e^{2\pi i \omega_j t} \quad (4.18)$$

for $t = 1, \dots, n$. 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**.

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

We now define the periodogram as the squared modulus² of the DFT.

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

$$I(\omega_j) = |d(\omega_j)|^2 \quad (4.19)$$

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

Note that $I(0) = n\bar{x}^2$, where \bar{x} is the sample mean. Also, for $j \neq 0$,³

$$I(\omega_j) = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) e^{-2\pi i \omega_j h}. \quad (4.20)$$

In view of (4.20), the periodogram, $I(\omega_j)$, is the sample version of $f(\omega_j)$ given in (4.14). That is, we may think of the periodogram as the “sample spectral density” of x_t . Although (4.20) seems like a reasonable estimate of $f(\omega)$, recall from Example 4.2 that $I(\omega_j)$, for any j , is based on only 2 pieces of information (degrees of freedom).

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

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

³ The DFTs of x_t and of $(x_t - \bar{x})$ are the same except at the zero frequency. This follows because $\sum_{t=1}^n \exp(-2\pi i t \frac{j}{n}) = 0$. Consequently, $|d(\omega_j)|^2 = n^{-1} \sum_{t=1}^n \sum_{s=1}^n (x_t - \bar{x})(x_s - \bar{x}) e^{-2\pi i \omega_j (t-s)} = n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}) e^{-2\pi i \omega_j h}$, which is (4.20).

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

$$d_c(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t \cos(2\pi\omega_j t) \quad (4.21)$$

and the **sine transform**

$$d_s(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t \sin(2\pi\omega_j t) \quad (4.22)$$

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

Note that $d_c(\omega_j)$ and $d_s(\omega_j)$ are averages like the sample mean, but with difference weights (the sample mean has weights $\frac{1}{n}$ for each observation). Under appropriate conditions, there is central limit theorem for these quantities. In non-technical terms, the result is similar to the central limit theorem for sample means, that is,

$$d_c(\omega_j) \dot{\sim} N(0, \tfrac{1}{2}f(\omega_j)) \quad \text{and} \quad d_s(\omega_j) \dot{\sim} N(0, \tfrac{1}{2}f(\omega_j)) \quad (4.23)$$

where $\dot{\sim}$ means *approximately distributed as* for n large. Moreover, it can be shown that for large n , $d_c(\omega_j) \perp d_s(\omega_j) \perp d_c(\omega_k) \perp d_s(\omega_k)$, as long as $\omega_j \neq \omega_k$, where \perp is read *is independent of*.

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

$$I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j), \quad (4.24)$$

which for large n is the sum of the squares of two independent normal random variables, which we know has a chi-squared (χ^2) distribution. Thus, for large samples, $I(\omega_j) \dot{\sim} \frac{1}{2}f(\omega_j)\chi_2^2$, or equivalently,

$$\frac{2I(\omega_j)}{f(\omega_j)} \dot{\sim} \chi_2^2, \quad (4.27a)$$

where χ_2^2 is the chi-squared distribution with 2 degrees of freedom. Since the mean and variance of a χ_ν^2 are ν and 2ν , respectively, it follows from (4.27a) that

$$E[I(\omega_j)] \approx f(\omega_j) \quad \text{and} \quad \text{var}[I(\omega_j)] \approx f^2(\omega_j). \quad (4.25)$$

This is bad news because, while the periodogram is approximately unbiased, its variance does not go to zero, and hence it is not consistent. In fact, no matter how large n , the variance of the periodogram does not change. Contrast this with the mean \bar{x} of a random sample of size n for which $E[\bar{x}] = \mu$ and $\text{var}[\bar{x}] = \sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$.

The technical result regarding the large sample distribution of the periodogram under general conditions is given in the following result.

Property 4.3 Distribution of the Periodogram Ordinates*If*

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} \sqrt{|j|} |\psi_j| < \infty \quad (4.26)$$

where $w_t \sim iid(0, \sigma_w^2)$, then for any collection of K distinct frequencies $\omega_k \in (0, 1/2)$ with $\omega_{k:n} \rightarrow \omega_k$ (where $\omega_{k:n}$ is a fundamental frequency) as $n \rightarrow \infty$,

$$\frac{2I(\omega_{k:n})}{f(\omega_k)} \xrightarrow{d} iid \chi_2^2 \quad (4.27)$$

provided $f(\omega_k) > 0$, for $k = 1, \dots, K$.

The distributional result (4.27) can be used to derive an approximate confidence interval for the spectrum in the usual way. Let $\chi_\nu^2(\alpha)$ denote the lower α probability tail for the chi-squared distribution with ν degrees of freedom. Then, an approximate $100(1 - \alpha)\%$ confidence interval for the spectral density function would be of the form

$$\frac{2 I(\omega_{j:n})}{\chi_2^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{2 I(\omega_{j:n})}{\chi_2^2(\alpha/2)}. \quad (4.28)$$

The log transform is the variance stabilizing transformation. In this case, the confidence intervals are of the form

$$[\log I(\omega_{j:n}) + \log 2 - \log \chi_2^2(1 - \alpha/2), \log I(\omega_{j:n}) + \log 2 - \log \chi_2^2(\alpha/2)]. \quad (4.29)$$

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 $x_t - \hat{\beta}_1 - \hat{\beta}_2 t$. 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 $x_{n+1}^c = x_{n+2}^c = \dots = x_{n'}^c = 0$, where x_t^c 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 1.6 of Chapter 1.

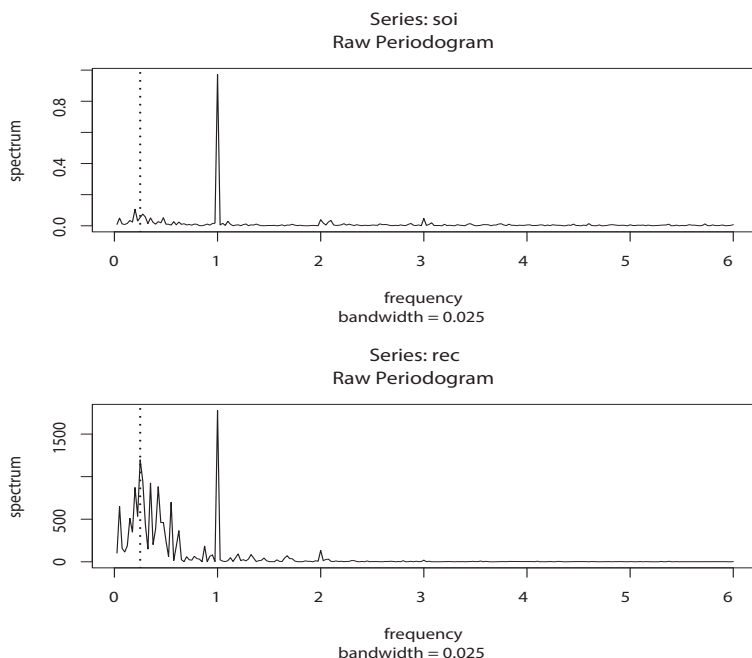


Fig. 4.5. Periodogram of SOI and Recruitment, $n = 453$ ($n' = 480$), where the frequency axis is labeled in multiples of $\Delta = 1/12$. Note the common peaks at $\omega = 1\Delta = 1/12$, or one cycle per year (12 months), and $\omega = \frac{1}{4}\Delta = 1/48$, or one cycle every four years (48 months).

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.

Example 4.7 Periodogram of SOI and Recruitment Series

Figure 4.5 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

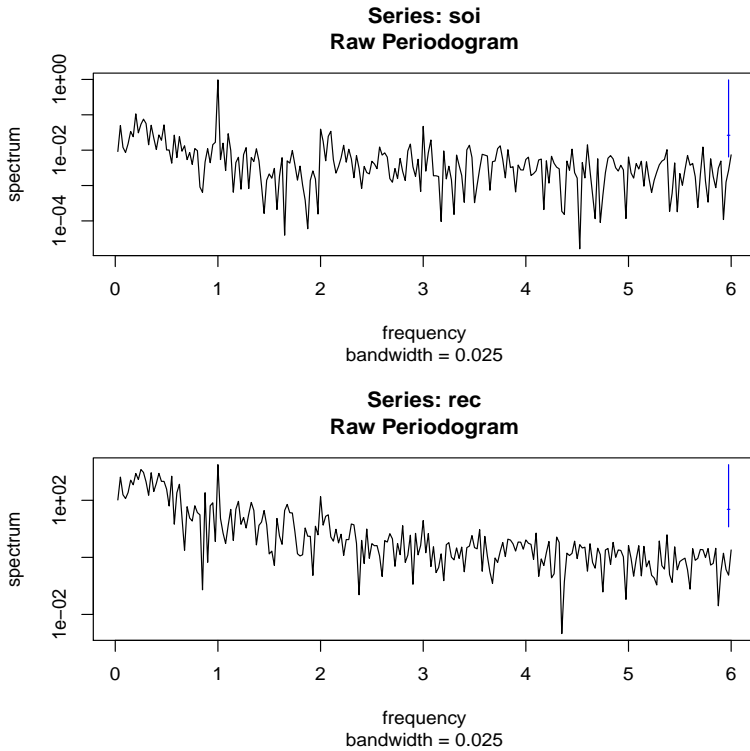


Fig. 4.6. Log-periodogram of SOI and Recruitment. 95% confidence intervals are indicated by the blue line in the upper right corner.

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

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 4.5**. To calculate and graph the periodogram, we used the `mvspec` script from `astsa`, although R's `spec.pgram` can be used. In the code, we set `log="no"` because the script will plot the periodogram on a \log_{10} scale by default. **Figure 4.5** displays a *bandwidth*, which we discuss in the next section.

```
require(astsa) # needed for mvspec() - otherwise use spec.pgram()
par(mfrow=c(2,1))
soi.per = mvspec(soi, log="no"); abline(v=1/4, lty="dotted")
```

```
rec.per = mvspec(rec, 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.

As an analogy to using the periodogram to estimate the spectral density, consider the problem of taking a random sample and then trying to estimate a probability density based on a histogram with many cells. This approach is demonstrated in [Figure 4.7](#).

4.5 Nonparametric Spectral Estimation

To continue the discussion that ended the previous section, we introduce a frequency band, \mathcal{B} , of $L \ll n$ contiguous fundamental frequencies, centered around frequency $\omega_j = j/n$, which is chosen close to a frequency of interest, ω . Let

$$\mathcal{B} = \{\omega_j + k/n : k = 0, \pm 1, \dots, \pm m\}, \quad (4.30)$$

where

$$L = 2m + 1 \quad (4.31)$$

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

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

are approximately equal to $f(\omega)$. For example, to see a small section of the AR(2) spectrum—near the peak—shown in [Figure 4.4](#), use

```
arma.spec(ar=c(1,-.9), xlim=c(.15,.151))
```

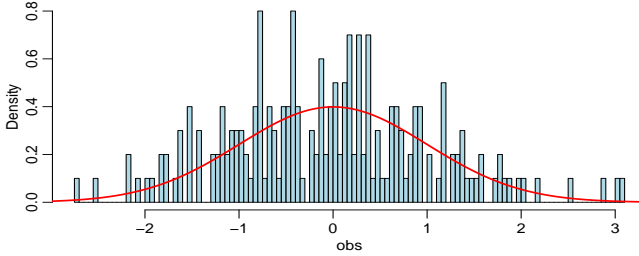


Fig. 4.7. Histogram of a sample of $n = 200$ standard normals with 100 cells and with the standard normal density superimposed. The periodogram is to the spectral density as the histogram is to the normal density.

which is displayed in [Figure 4.8](#).

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

$$\bar{f}(\omega) = \frac{1}{L} \sum_{k=-m}^m I(\omega_j + k/n), \quad (4.32)$$

over the band \mathcal{B} . Under the assumption that the spectral density is fairly constant in the band \mathcal{B} , and in view of (4.27) we can show that under appropriate conditions, for large n , the periodograms in (4.32) 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 . 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 ,

$$\frac{2L\bar{f}(\omega)}{f(\omega)} \sim \chi_{2L}^2. \quad (4.33)$$

Now we have

$$E[\bar{f}(\omega)] \approx f(\omega) \quad \text{and} \quad \text{var}[\bar{f}(\omega)] \approx f^2(\omega)/L, \quad (4.34)$$

which can be compared to (4.25). In this case, we have consistency if we let $L \rightarrow \infty$ as $n \rightarrow \infty$, but L must grow much slower than n , of course (in fact, $L/n \rightarrow 0$ as $n \rightarrow \infty$).

In this scenario, where we smooth the periodogram by simple averaging, the width of the frequency interval defined by (4.30),

$$B_w = \frac{L}{n}, \quad (4.35)$$

is called the *bandwidth*. Note (4.35) implies the degrees of freedom can be expressed as

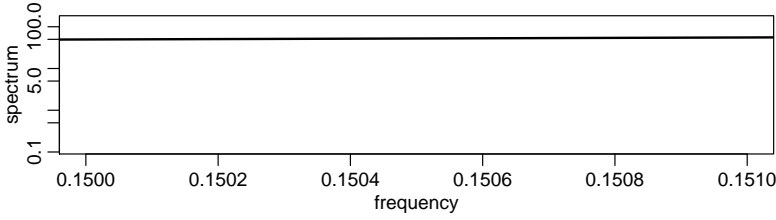


Fig. 4.8. A small section (near the peak) of the AR(2) spectrum shown in [Figure 4.4](#).

$$2L = 2B_w n, \quad (4.36)$$

or twice the time-bandwidth product. The result (4.33) can be rearranged to obtain an approximate $100(1 - \alpha)\%$ confidence interval of the form

$$\frac{2L\bar{f}(\omega)}{\chi_{2L}^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{2L\bar{f}(\omega)}{\chi_{2L}^2(\alpha/2)} \quad (4.37)$$

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

As previously discussed, the visual impact of a spectral density plot will be improved by plotting the logarithm of the spectrum, which 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

$$\begin{aligned} &[\log \bar{f}(\omega) + \log 2L - \log \chi_{2L}^2(1 - \alpha/2), \\ &\quad \log \bar{f}(\omega) + \log 2L - \log \chi_{2L}^2(\alpha/2)]. \end{aligned} \quad (4.38)$$

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

$$df = \frac{2Ln}{n'} \quad (4.39)$$

and use it instead of $2L$ in the confidence intervals (4.37) and (4.38). For example, (4.37) becomes

$$\frac{df\bar{f}(\omega)}{\chi_{df}^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{df\bar{f}(\omega)}{\chi_{df}^2(\alpha/2)}. \quad (4.40)$$

Before proceeding further, we pause to consider computing the average periodograms for the SOI and Recruitment series, as shown in [Figure 4.9](#).

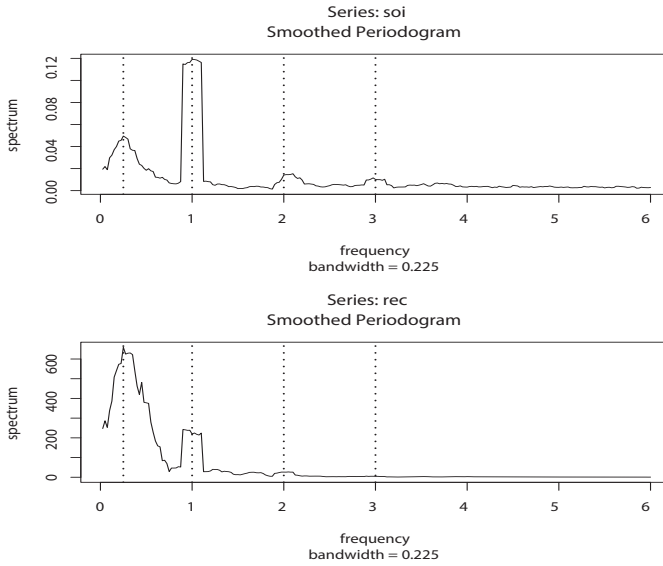


Fig. 4.9. The averaged periodogram of the SOI and Recruitment series $n = 453$, $n' = 480$, $L = 9$, $df = 17$, showing common peaks at the four year period, $\omega = \frac{1}{4}\Delta = 1/48$ (cycle/months), the yearly period, $\omega = 1\Delta = 1/12$ (cycle/months) and some of its harmonics $\omega = k\Delta$ for $k = 2, 3$.

Example 4.8 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 4.5](#), 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 4.9](#).

The smoothed spectra shown in [Figure 4.9](#) provide a sensible compromise between the noisy version, shown in [Figure 4.5](#), 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 4.5](#) 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 $k = 1, 2, \dots$. Harmonics typically occur when a periodic component is present, but not in a sinusoidal fashion; see [Example 4.9](#).

[Figure 4.9](#) can be reproduced in R using the following commands. The basic call is to the function `mvspec`, which is available in `astsa`; alternately, use R's `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).

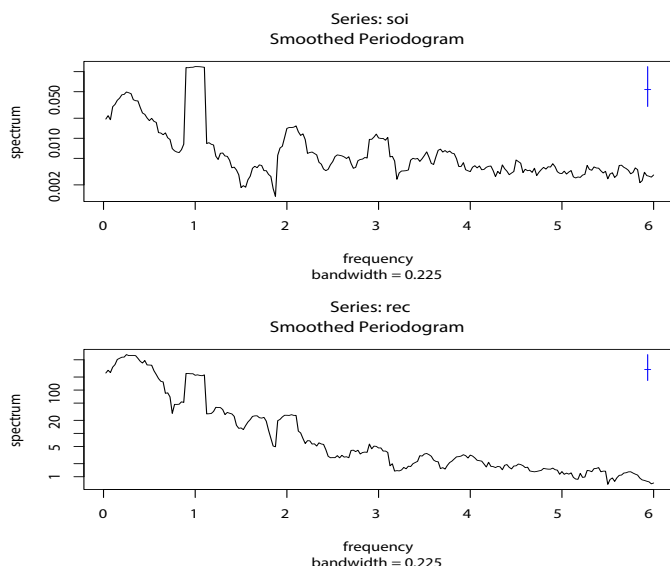


Fig. 4.10. Figure 4.9 with the average periodogram ordinates plotted on a \log_{10} scale. The display in the upper right-hand corner represents a generic 95% confidence interval.

We will explain the kernel concept later in this section, specifically just prior to [Example 4.10](#).

```
par(mfrow=c(2,1))
(k = kernel("daniell", 4))
soi.ave = mvspec(soi, k, 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      # = 0.225
```

The displayed bandwidth (.225) is adjusted for the fact that the frequency scale of the plot is in terms of cycles per year instead of cycles per month (the original unit of the data). Using (4.35), the bandwidth in terms of months is $9/480 = .01875$; the displayed value is simply converted to years, $.01875 \frac{\text{cycles}}{\text{month}} \times 12 \frac{\text{months}}{\text{year}} = .225 \frac{\text{cycles}}{\text{year}}$.

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 (4.40) gives the intervals in [Table 4.1](#) for the two frequency bands identified as having the maximum power. To 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

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 $\times 10^2$	1/48	4 years	6.59	3.71	14.82
	1/12	1 year	2.19	1.24	4.93

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$spec[10]/U     # 0.1113333
df*soi.ave$spec[40]/L     # 0.0670396
df*soi.ave$spec[40]/U     # 0.2677201
# repeat above commands with soi replaced by rec
```

Finally, [Figure 4.10](#) shows the averaged periodograms in [Figure 4.9](#) 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"`. Notice that the default plot also shows a generic confidence interval of the form (4.38) 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 4.9 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 $k = 2, 3, \dots$ (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 4.11](#) 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

$$\begin{aligned}
 x_t = & \sin(2\pi 2t) + .5 \sin(2\pi 4t) + .4 \sin(2\pi 6t) \\
 & + .3 \sin(2\pi 8t) + .2 \sin(2\pi 10t) + .1 \sin(2\pi 12t)
 \end{aligned} \tag{4.41}$$

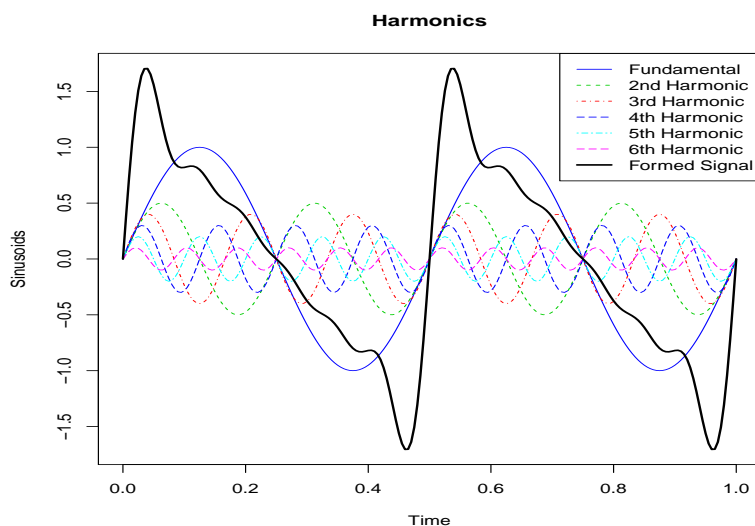


Fig. 4.11. A signal (thick solid line) formed by a fundamental sinusoid (thin solid line) oscillating at two cycles per unit time and its harmonics as specified in (4.41).

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

A figure similar to Figure 4.11 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)
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 4.8 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.

Care must be taken when 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 [Figure 4.9](#) and [Figure 4.10](#) was due to the fact that simple averaging was used in computing $\hat{f}(\omega)$ defined in (4.32). There is no particular reason to use simple averaging, and we might improve the estimator by employing a weighted average, say

$$\hat{f}(\omega) = \sum_{k=-m}^m h_k I(\omega_j + k/n), \quad (4.42)$$

using the same definitions as in (4.32) 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 (4.42), set $h_k = L^{-1}$, for all k , where $L = 2m + 1$. The asymptotic theory established for $\bar{f}(\omega)$ still holds for $\hat{f}(\omega)$ provided that the weights satisfy the additional condition that if $m \rightarrow \infty$ as $n \rightarrow \infty$ but $m/n \rightarrow 0$, then

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

Under these conditions, for n large, we have

$$E[\hat{f}(\omega)] \approx f(\omega) \quad \text{and} \quad \text{var}[\hat{f}(\omega)] \approx f^2(\omega) \sum_{k=-m}^m h_k^2 \quad (4.43)$$

which can be compared to (4.34); as before, we have that $\hat{f}(\omega)$ is consistent. We have already seen this result 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 (4.42) are more difficult now because $\hat{f}(\omega)$ is a weighted linear combination of asymptotically independent χ^2 random variables. An approximation that seems to work well is to replace L by $(\sum_{k=-m}^m h_k^2)^{-1}$. That is, define

$$L_h = \left(\sum_{k=-m}^m h_k^2 \right)^{-1} \quad (4.44)$$

and use the approximation

$$\frac{2L_h \hat{f}(\omega)}{f(\omega)} \sim \chi_{2L_h}^2. \quad (4.45)$$

In analogy to (4.35), we will define the bandwidth in this case to be

$$B_w = \frac{L_h}{n}. \quad (4.46)$$

Using the approximation (4.45) we obtain an approximate $100(1 - \alpha)\%$ confidence interval of the form

$$\frac{2L_h \hat{f}(\omega)}{\chi_{2L_h}^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{2L_h \hat{f}(\omega)}{\chi_{2L_h}^2(\alpha/2)} \quad (4.47)$$

for the true spectrum, $f(\omega)$. If the data are padded to n' , then replace $2L_h$ in (4.47) with $df = 2L_h n/n'$ as in (4.39).

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

$$\hat{u}_t = \frac{1}{3}u_{t-1} + \frac{1}{3}u_t + \frac{1}{3}u_{t+1}.$$

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

$$\hat{\hat{u}}_t = \frac{1}{3}\hat{u}_{t-1} + \frac{1}{3}\hat{u}_t + \frac{1}{3}\hat{u}_{t+1},$$

which simplifies to

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

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

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

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

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

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.

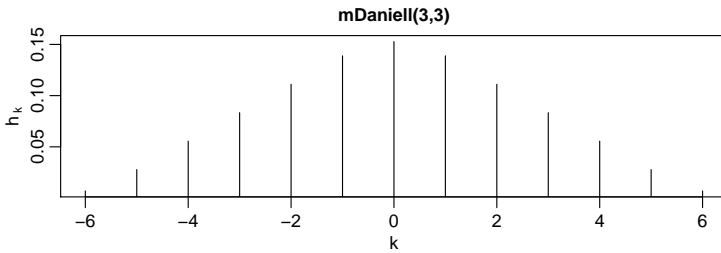


Fig. 4.12. Modified Daniell kernel weights used in [Example 4.10](#)

Example 4.10 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 (4.42). We used a modified Daniell kernel twice, with $m = 3$ both times. This yields $L_h = 1/\sum h_k^2 = 9.232$, which is close to the value of $L = 9$ used in [Example 4.8](#). 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; see [Figure 4.12](#).

```
kernel("modified.daniell", c(3,3))      # for a list
plot(kernel("modified.daniell", c(3,3))) # for a plot
```

The resulting spectral estimates can be viewed in [Figure 4.13](#) and we notice that the estimates more appealing than those in [Figure 4.9](#). [Figure 4.13](#) 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 = mvspec(soi, k, 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.smo$df          # df = 17.42618
soi.smo$bandwidth        # Bw = 0.2308103 = 12*9.232/480
```

Reissuing the `mvspec` commands with `log="no"` removed will result in a figure similar to [Figure 4.10](#); see [Figure 4.13](#). Finally, we mention that the modified Daniell kernel is used by default. For example, an easier way to obtain `soi.smo` is to issue the command:

```
soi.smo = mvspec(soi, spans=c(7,7))
```

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.

TAPERING

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

$$y_t = h_t x_t, \quad (4.48)$$

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

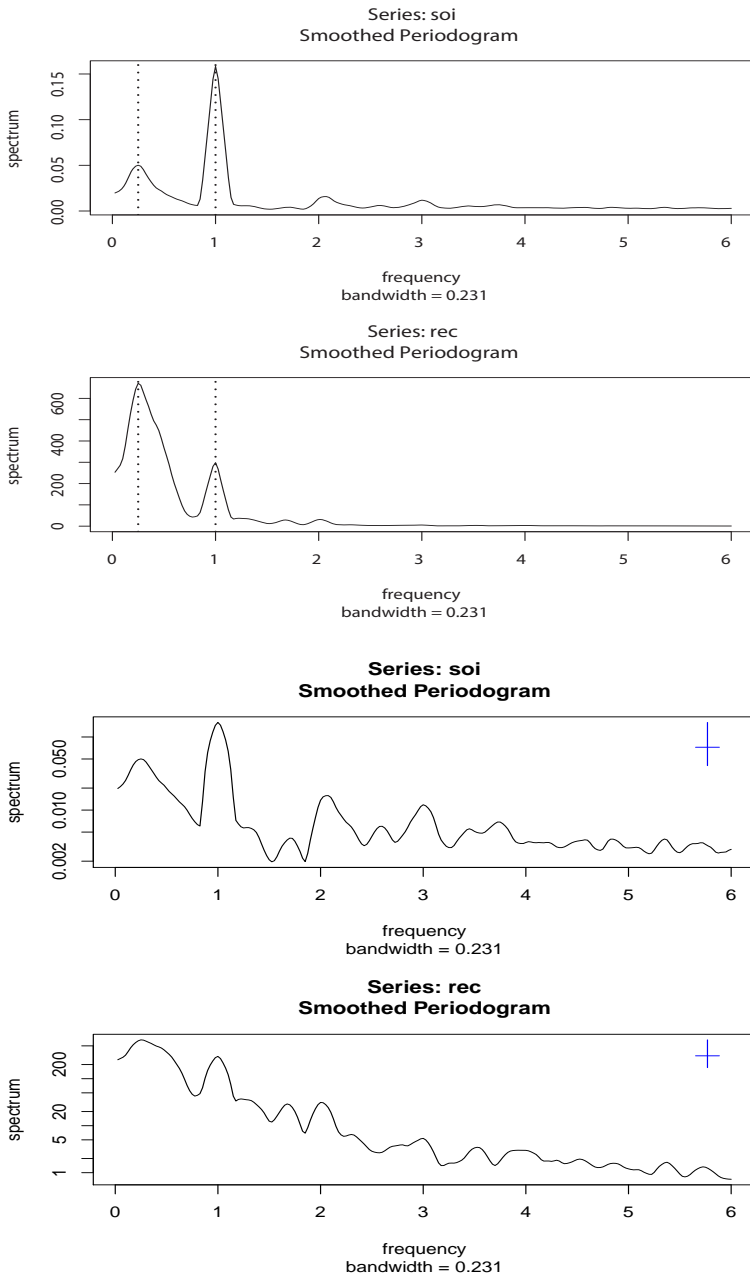


Fig. 4.13. Smoothed spectral estimates of the SOI and Recruitment series; see [Example 4.10](#) for details. The bottom pair are drawn on a log scale. Note that the El Niño cycle is a broad band cycle ranging from about 3 to 7 years. See [Figure 4.18](#) for a graph of the approximate El Niño cycle.

$$d_y(\omega_j) = n^{-1/2} \sum_{t=1}^n h_t x_t e^{-2\pi i \omega_j t}, \quad (4.49)$$

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

$$E[I_y(\omega_j)] = \int_{-1/2}^{1/2} W_n(\omega_j - \omega) f_x(\omega) d\omega. \quad (4.50)$$

The value $W_n(\omega)$ is called a spectral window because, in view of (4.50), 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

$$W_n(\omega) = \frac{\sin^2(n\pi\omega)}{n \sin^2(\pi\omega)} \quad (4.51)$$

with $W_n(0) = n$, which is known as the Fejér or modified Bartlett kernel. If we consider the averaged periodogram in (4.32), namely

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

the window, $W_n(\omega)$, in (4.50) will take the form

$$W_n(\omega) = \frac{1}{nL} \sum_{k=-m}^m \frac{\sin^2[n\pi(\omega + k/n)]}{\sin^2[\pi(\omega + k/n)]}. \quad (4.52)$$

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

$$h_t = .5 \left[1 + \cos \left(\frac{2\pi(t - \bar{t})}{n} \right) \right], \quad (4.53)$$

where $\bar{t} = (n+1)/2$, favored by Blackman and Tukey (1959). In **Figure 4.14**, 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, (4.52) applies, and (ii) h_t is the cosine taper in (4.53). 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 4.14**. 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 4.14** emphasizes the suppression of the sidelobes in the Fejér kernel when a cosine taper is used.

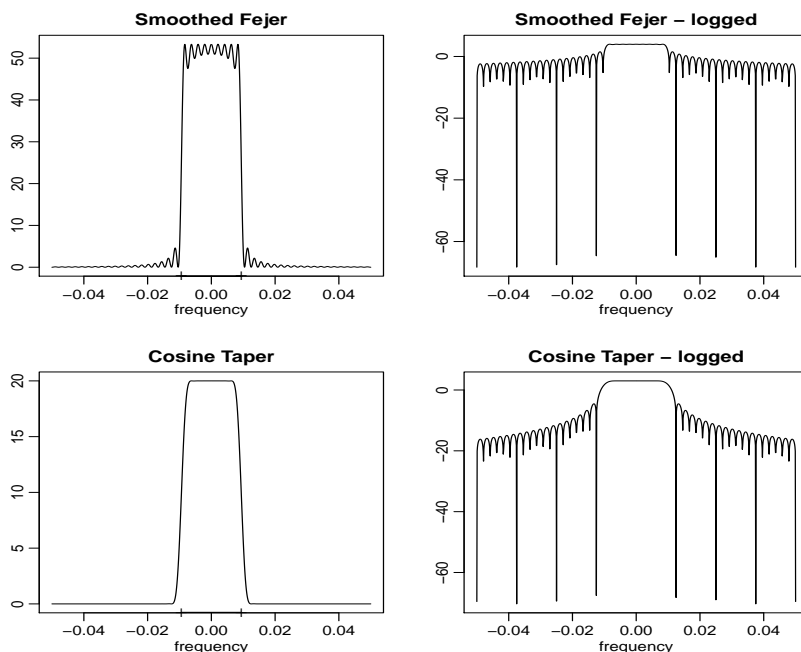


Fig. 4.14. Averaged Fejér window (top row) and the corresponding cosine taper window (bottom row) for $L = 9$, $n = 480$. The extra tic marks on the horizontal axis of the left-hand plots exhibit the predicted bandwidth, $B_w = 9/480 = .01875$.

Example 4.11 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 4.15** shows two spectral estimates plotted on a log scale. The degree of smoothing here is the same as in **Example 4.10**. The dashed line in **Figure 4.15** shows the estimate without any tapering and hence it is the same as the estimated spectrum displayed in the top of **Figure 4.13**. 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 4.15**. We note that, by default, `mvspec` does not taper. For full tapering, we use the argument `taper=.5` to instruct `mvspec` to taper 50% of each end of the data; any value between 0 and .5 is acceptable.

```
s0 = mvspec(soi, spans=c(7,7), plot=FALSE)           # no taper
s50 = mvspec(soi, spans=c(7,7), taper=.5, plot=FALSE) # full taper
plot(s0$freq, s0$spec, log="y", type="l", lty=2, ylab="spectrum",
     xlab="frequency")                                # dashed line
lines(s50$freq, s50$spec)                             # solid line
```

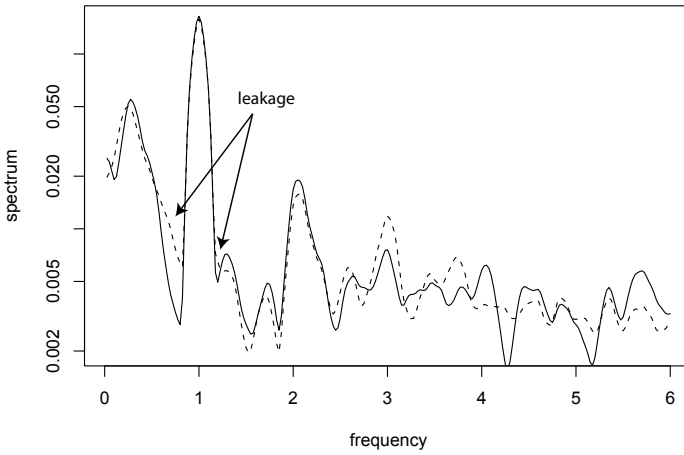


Fig. 4.15. Smoothed spectral estimates of the SOI without tapering (dashed line) and with full tapering (solid line); see [Example 4.11](#) for details.

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 4.2](#), 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 (2.16)–(2.18). 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 $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$ and $\hat{\sigma}_w^2$ are the estimates from an AR(p) fit to x_t , then based on [Property 4.2](#), a parametric spectral estimate of $f_x(\omega)$ is attained by substituting these estimates into (4.15), that is,

$$\hat{f}_x(\omega) = \frac{\hat{\sigma}_w^2}{|\hat{\phi}(e^{-2\pi i\omega})|^2}, \quad (4.54)$$

where

$$\hat{\phi}(z) = 1 - \hat{\phi}_1 z - \hat{\phi}_2 z^2 - \dots - \hat{\phi}_p z^p. \quad (4.55)$$

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 4.4 AR Spectral Approximation

Let $g(\omega)$ be the spectral density of a stationary process, x_t . Then, given $\epsilon > 0$, there is an $AR(p)$ 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

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

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, however, is that the property does not tell us how large p must be before the approximation is reasonable; in some situations p may be extremely large. **Property 4.4** 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 4.12 Autoregressive Spectral Estimator for SOI

Consider obtaining results comparable to the nonparametric estimators shown in **Figure 4.9** for the SOI series. Fitting successively higher order $AR(p)$ models for $p = 1, 2, \dots, 30$ yields a minimum BIC and a minimum AIC at $p = 15$, as shown in **Figure 4.16**. We can see from **Figure 4.16** 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 spectrum is shown in **Figure 4.17**, 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.

```
spaic = spec.ar(soi, log="no")      # min AIC spec
abline(v=frequency(soi)*1/52, lty="dotted") # El Nino Cycle
(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 4.16** 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.

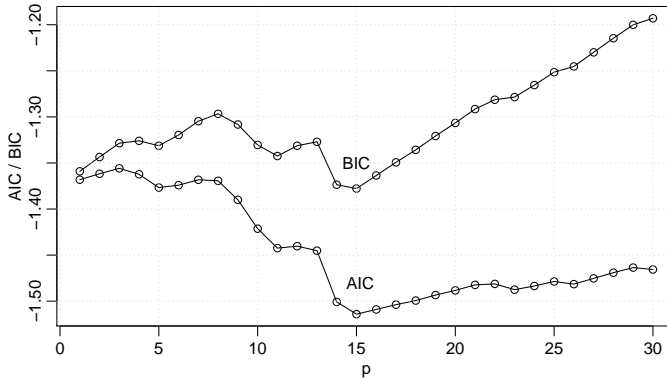


Fig. 4.16. Model selection criteria AIC and BIC as a function of order p for autoregressive models fitted to the SOI series.

```
n = length(soi)
AIC = rep(0, 30) -> AICc -> BIC
for (k in 1:30){
  sigma2 = ar(soi, order=k, aic=FALSE)$var.pred
  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")
```

4.7 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 $j = 0, \pm 1, \pm 2, \dots$, to transform an input series, x_t , producing an output series, y_t , of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}, \quad \sum_{j=-\infty}^{\infty} |a_j| < \infty. \quad (4.56)$$

The form (4.56) 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 (4.56) exists as a limit in mean square and the infinite Fourier transform

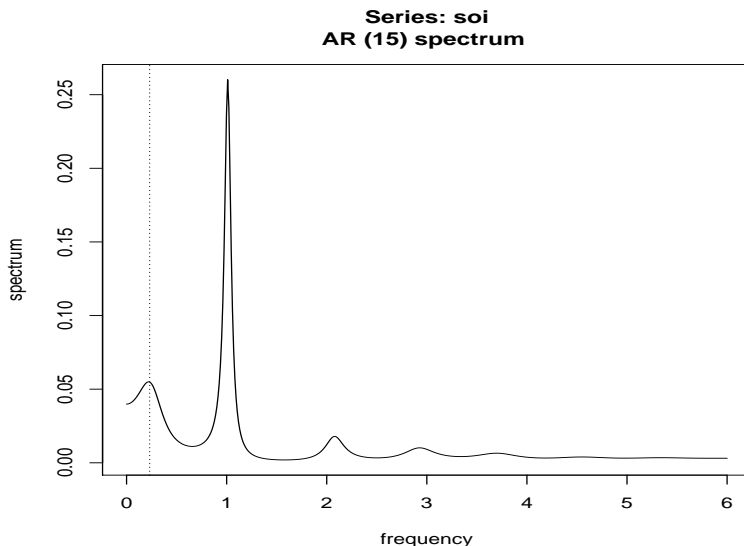


Fig. 4.17. Autoregressive spectral estimator for the SOI series using the AR(15) model selected by AIC, AICc, and BIC. The first peak (marked by a vertical dotted line) corresponds to the El Niño period of 52 months.

$$A_{yx}(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-2\pi i \omega j}, \quad (4.57)$$

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 1.9](#), which can be put into the form of (4.56) 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. We now state the following result.

Property 4.5 Output Spectrum of a Filtered Stationary Series

Assuming existence of spectra, the spectrum of the filtered output y_t in (4.56) is related to the spectrum of the input x_t by

$$f_{yy}(\omega) = |A_{yx}(\omega)|^2 f_{xx}(\omega), \quad (4.58)$$

where the frequency response function $A_{yx}(\omega)$ is defined in (4.57).

The result (4.58) 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

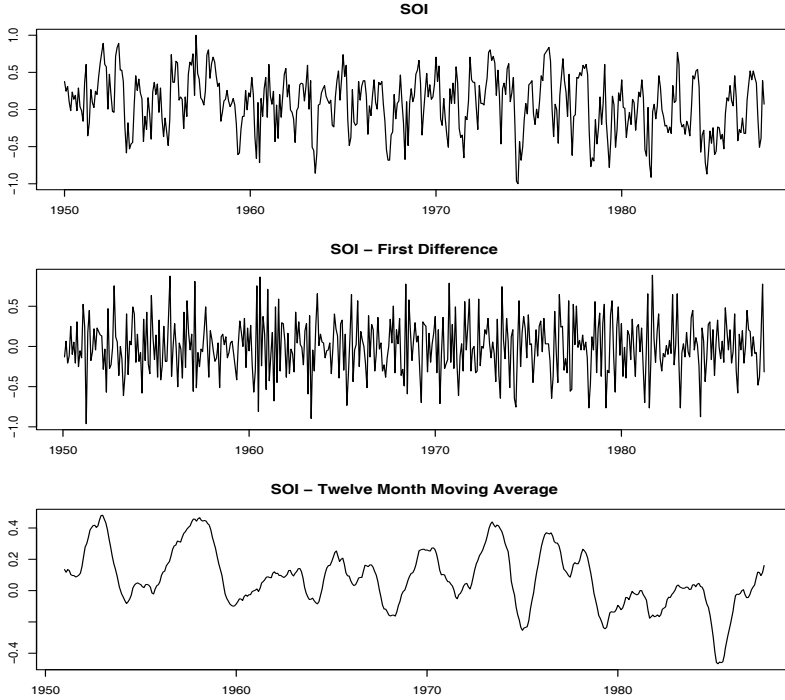


Fig. 4.18. SOI series (top) compared with the differenced SOI (middle) and a centered 12-month moving average (bottom).

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 (4.58).

Finally, we mention that **Property 4.2**, which was used to get the spectrum of an ARMA process, is just a special case of **Property 4.5** where in (4.56), $x_t = w_t$ is white noise, in which case $f_{xx}(\omega) = \sigma_w^2$, and $a_j = \psi_j$, in which case

$$A_{yx}(\omega) = \psi(e^{-2\pi i\omega}) = \theta(e^{-2\pi i\omega}) / \phi(e^{-2\pi i\omega}).$$

Example 4.13 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

$$y_t = \frac{1}{24}(x_{t-6} + x_{t+6}) + \frac{1}{12} \sum_{r=-5}^5 x_{t-r},$$

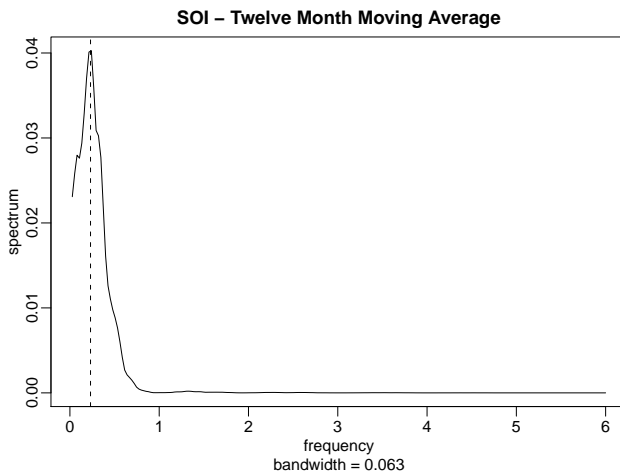


Fig. 4.19. Spectral analysis of SOI after applying a 12-month moving average filter. The vertical line corresponds to the 52-month cycle.

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 4.18](#). Notice that the effect of differencing is to roughen the series because it tends to retain the higher or faster frequencies. The centered moving average smooths 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 4.19](#) 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 [\(4.57\)](#) and [\(4.58\)](#) for this purpose. The first difference filter can be written in the form [\(4.56\)](#) by letting $a_0 = 1$, $a_1 = -1$, and $a_r = 0$ otherwise. This implies that

$$A_{yx}(\omega) = 1 - e^{-2\pi i\omega},$$

and the squared frequency response becomes

$$|A_{yx}(\omega)|^2 = (1 - e^{-2\pi i\omega})(1 - e^{2\pi i\omega}) = 2[1 - \cos(2\pi\omega)]. \quad (4.59)$$

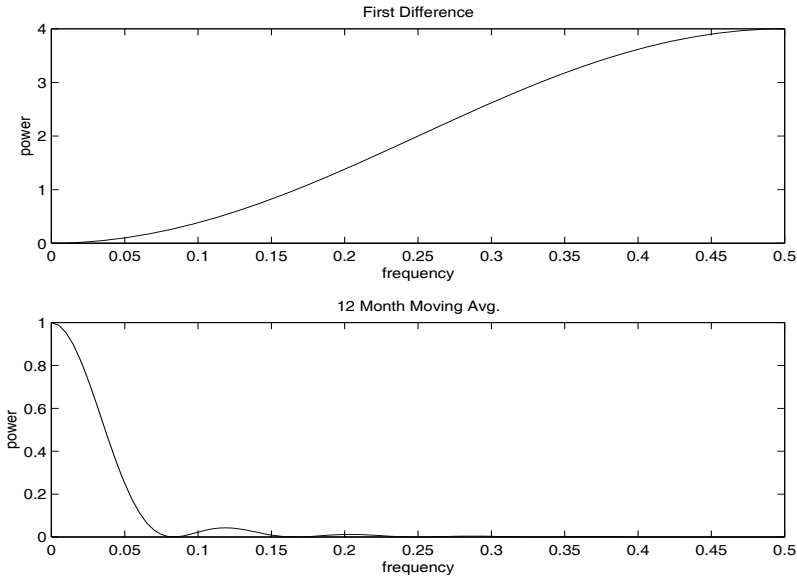


Fig. 4.20. Squared frequency response functions of the first difference and 12-month moving average filters.

The top panel of [Figure 4.20](#) 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

$$A_{yx}(\omega) = \frac{1}{12} \left[1 + \cos(12\pi\omega) + 2 \sum_{k=1}^5 \cos(2\pi\omega k) \right]. \quad (4.60)$$

Plotting the squared frequency response of this function as in [Figure 4.20](#) 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 4.4](#).

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()
mvspec(soif, spans=9, taper=.1, 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")
```

4.8 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 autocovariance function

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

has a spectral representation given by

$$\gamma_{xy}(h) = \int_{-1/2}^{1/2} f_{xy}(\omega) e^{2\pi i \omega h} d\omega \quad h = 0, \pm 1, \pm 2, \dots, \quad (4.61)$$

where the cross-spectrum is defined as the Fourier transform

$$f_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2, \quad (4.62)$$

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 as⁴

$$f_{xy}(\omega) = c_{xy}(\omega) - iq_{xy}(\omega), \quad (4.63)$$

where

$$c_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \cos(2\pi \omega h) \quad (4.64)$$

and

⁴ 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 $\bar{z} = a - ib$.

$$q_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \sin(2\pi\omega h) \quad (4.65)$$

are defined as the cospectrum and quadspectrum, respectively. Because of the relationship $\gamma_{yx}(h) = \gamma_{xy}(-h)$, it follows, by substituting into (4.62) and rearranging, that

$$f_{yx}(\omega) = \overline{f_{xy}(\omega)}. \quad (4.66)$$

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

$$c_{yx}(\omega) = c_{xy}(\omega) \quad (4.67)$$

and

$$q_{yx}(\omega) = -q_{xy}(\omega). \quad (4.68)$$

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

$$\rho_{y \cdot x}^2(\omega) = \frac{|f_{yx}(\omega)|^2}{f_{xx}(\omega)f_{yy}(\omega)}, \quad (4.69)$$

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 (4.69) to emphasize the analogy with conventional squared correlation, which takes the form

$$\rho_{yx}^2 = \frac{\sigma_{yx}^2}{\sigma_x^2 \sigma_y^2},$$

for random variables with variances σ_x^2 and σ_y^2 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 4.14 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,

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

where we have use (4.13). Using the uniqueness of the Fourier transform, we argue from the spectral representation (4.61) that

$$f_{xy}(\omega) = \frac{1}{3} [1 + 2 \cos(2\pi\omega)] f_{xx}(\omega)$$

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

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

using the identity $\cos(2\alpha) = 2 \cos^2(\alpha) - 1$ in the last step. Substituting into (4.69) 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. 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 Process

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

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

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

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty \quad (4.70)$$

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

$$\Gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i \omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \dots, \quad (4.71)$$

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

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2. \quad (4.72)$$

Example 4.15 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 (4.71) and (4.72) relate the autocovariance and spectral matrices.

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

$$\bar{f}(\omega) = L^{-1} \sum_{k=-m}^m I(\omega_j + k/n) \quad (4.73)$$

where now

$$I(\omega_j) = \mathbf{d}(\omega_j) \mathbf{d}^*(\omega_j) \quad (4.74)$$

is a $p \times p$ complex matrix.⁵

Again, the series may be tapered before the DFT is taken in (4.73) and we can use weighted estimation,

$$\hat{f}(\omega) = \sum_{k=-m}^m h_k I(\omega_j + k/n) \quad (4.75)$$

where $\{h_k\}$ are weights as defined in (4.42). The estimate of squared coherence between two series, y_t and x_t is

$$\hat{\rho}_{y \cdot x}^2(\omega) = \frac{|\hat{f}_{yx}(\omega)|^2}{\hat{f}_{xx}(\omega) \hat{f}_{yy}(\omega)}. \quad (4.76)$$

If the spectral estimates in (4.76) are obtained using equal weights, we will write $\bar{\rho}_{y \cdot x}^2(\omega)$ for the estimate.

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

$$|\hat{\rho}_{y \cdot x}(\omega)| \sim AN \left(|\rho_{y \cdot x}(\omega)|, (1 - \rho_{y \cdot x}^2(\omega))^2 / 2L_h \right) \quad (4.77)$$

where L_h is defined in (4.44); the details of this result may be found in Brockwell and Davis (1991, Ch 11). We may use (4.77) to obtain approximate confidence intervals for the squared coherency $\rho_{y \cdot x}^2(\omega)$.

We can test the hypothesis that $\rho_{y \cdot x}^2(\omega) = 0$ if we use $\bar{\rho}_{y \cdot x}^2(\omega)$ for the estimate with $L > 1$,⁶ that is,

⁵ 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.

⁶ If $L = 1$ then $\bar{\rho}_{y \cdot x}^2(\omega) \equiv 1$.

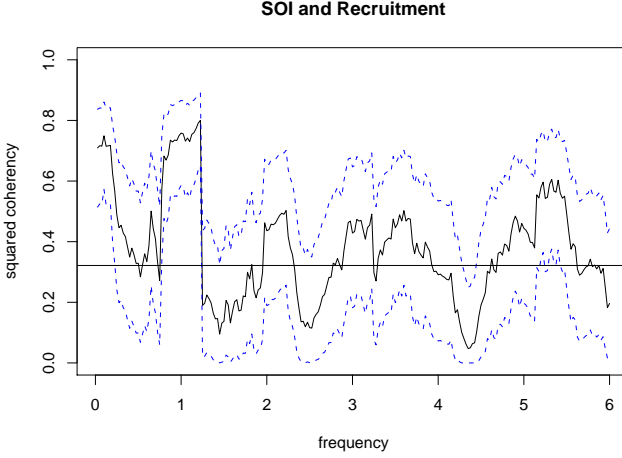


Fig. 4.21. Squared coherency between the SOI and Recruitment series; $L = 19$, $n = 453$, $n' = 480$, and $\alpha = .001$. The horizontal line is $C_{.001}$.

$$\bar{\rho}_{y \cdot x}^2(\omega) = \frac{|\bar{f}_{yx}(\omega)|^2}{\bar{f}_{xx}(\omega)\bar{f}_{yy}(\omega)}. \quad (4.78)$$

In this case, under the null hypothesis, the statistic

$$F = \frac{\bar{\rho}_{y \cdot x}^2(\omega)}{(1 - \bar{\rho}_{y \cdot x}^2(\omega))}(L - 1) \quad (4.79)$$

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 $df - 2$, where df is defined in (4.39). Solving (4.79) for a particular significance level α leads to

$$C_\alpha = \frac{F_{2, 2L-2}(\alpha)}{L - 1 + F_{2, 2L-2}(\alpha)} \quad (4.80)$$

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

Example 4.16 Coherence Between SOI and Recruitment

Figure 4.21 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 $\bar{\rho}_{y \cdot x}^2(\omega)$ 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, (??), in mind. Figure 4.21 also exhibits confidence bands as part of the R plotting routine. We emphasize that these bands are only valid for ω where $\rho_{y \cdot x}^2(\omega) > 0$.

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

Problems

4.1 Repeat the simulations and analyses in [Example 4.1](#) and [Example 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 $w_t \sim \text{iid } N(0, \sigma_w = 5)$. That is, you should simulate and plot the data, and then plot the periodogram of x_t and comment.

4.2 Verify (4.5).

4.3 Consider an MA(1) process

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

where θ is a parameter.

- (a) Derive a formula for the power spectrum of x_t , expressed in terms of θ and ω .
- (b) Sketch the general shape of the spectrum for $\theta > 0$ and for $\theta < 0$.
- (c) How should we interpret the sketches made in part (b)?

4.4 Consider a first-order autoregressive model

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

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

$$\gamma_x(h) = \frac{\sigma_w^2 \phi^{|h|}}{1 - \phi^2},$$

$h = 0, \pm 1, \pm 2, \dots$, by showing that the inverse transform of $\gamma_x(h)$ is the spectrum derived in part (a).

4.5 In applications, we will often observe series containing a signal that has been delayed by some unknown time D , i.e.,

$$x_t = s_t + A s_{t-D} + n_t,$$

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 .

- (a) Find the autocovariance function of x_t and use it to show

$$f_x(\omega) = [1 + A^2 + 2A \cos(2\pi\omega D)]f_s(\omega) + f_n(\omega).$$

- (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, and use the logarithm.)

4.6 Figure 4.22 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 4.7 as a guide, perform a periodogram analysis identifying the predominant periods and obtaining confidence intervals for the identified periods. Interpret your findings.

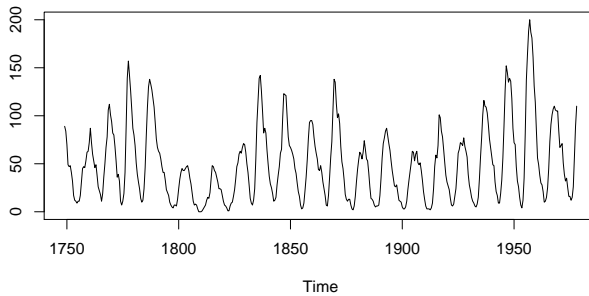


Fig. 4.22. Smoothed 12-month sunspot numbers (`sunspotz`) sampled twice per year.

4.7 The levels of salt concentration known to have occurred over rows, corresponding to the average temperature levels for the soil science 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.

4.8 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 $\omega_k \neq 0, \frac{1}{2}$ is assumed to be known and of the form k/n . Given data x_1, \dots, x_n , suppose we consider estimating β_1, β_2 and σ_w^2 by least squares.

- (a) Use simple regression formulas to show that for a fixed ω_k , the least squares regression coefficients are

$$\hat{\beta}_1 = 2n^{-1/2}d_c(\omega_k) \quad \text{and} \quad \hat{\beta}_2 = 2n^{-1/2}d_s(\omega_k),$$

where the cosine and sine transforms (4.21) and (4.22) appear on the right-hand side. *Hint:* See Problem 4.21.

- (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 (4.19).

- (c) Show that the sum of squares for the regression is given by

$$SSR = 2I_x(\omega_k).$$

- (d) 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 $I_x(\omega_k)$.

4.9 Repeat **Problem 4.6** 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.

4.10 Repeat **Problem 4.7** 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.

4.11 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 4.5(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 *quefreny* corresponding to $1/D$. The cepstrum can be plotted as a function of quefreny, from which the delay 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**.

- Calculate and display the log-periodogram of the data. Is the periodogram periodic, as predicted?
- 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 1.25**, which was based on the ACF?

4.12 Consider two time series

$$x_t = w_t - w_{t-1},$$

$$y_t = \frac{1}{2}(w_t + w_{t-1}),$$

formed from the white noise series w_t with variance $\sigma_w^2 = 1$.

- 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.
- Compute the spectra $f_y(\omega)$ and $f_x(\omega)$, and comment on the difference between the two results.
- Suppose sample spectral estimators $\bar{f}_y(.10)$ are computed for the series using $L = 3$. Find a and b such that

$$P\left\{a \leq \bar{f}_y(.10) \leq b\right\} = .90.$$

This expression gives two points that will contain 90% of the sample spectral values. Put 5% of the area in each tail.

4.13 Analyze the coherency between the temperature and salt data discussed in **Problem 4.7**. Discuss your findings.

4.14 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) $L = 101$.

4.15 For the processes in **Problem 4.14**:

- (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) $L = 101$.

4.16 Consider the bivariate time series records containing monthly U.S. production as measured by the Federal Reserve Board Production Index (`prodn`) and monthly unemployment (`unemp`) that are included with `astsa`.

- (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.

4.17 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 .

4.18 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 4.6**.

4.19 Fit an autoregressive spectral estimator to the Recruitment series and compare it to the results of **Example 4.10**.

4.20 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 $\hat{f}_x(1/8) = 2.25$. 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.

4.21 * This is here for useful information. Verify that for any positive integer n and $j, k = 0, 1, \dots, \lfloor n/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function:

(a) Except for $j = 0$ or $j = n/2$,

$$\sum_{t=1}^n \cos^2(2\pi t j/n) = \sum_{t=1}^n \sin^2(2\pi t j/n) = n/2.$$

(b) When $j = 0$ or $j = n/2$,

$$\sum_{t=1}^n \cos^2(2\pi t j/n) = n \quad \text{but} \quad \sum_{t=1}^n \sin^2(2\pi t j/n) = 0.$$

(c) For $j \neq k$,

$$\sum_{t=1}^n \cos(2\pi t j/n) \cos(2\pi t k/n) = \sum_{t=1}^n \sin(2\pi t j/n) \sin(2\pi t k/n) = 0.$$

(d) Also, for any j and k ,

$$\sum_{t=1}^n \cos(2\pi t j/n) \sin(2\pi t k/n) = 0.$$

* Note, $\sum_{t=1}^n z^t = z \frac{1-z^n}{1-z}$ for $z \neq 1$, and we'll do (a):

$$\begin{aligned} \sum_{t=1}^n \cos^2(2\pi t j/n) &= \frac{1}{4} \sum_{t=1}^n (e^{2\pi i t j/n} + e^{-2\pi i t j/n}) (e^{2\pi i t j/n} + e^{-2\pi i t j/n}) \\ &= \frac{1}{4} \sum_{t=1}^n (e^{4\pi i t j/n} + 1 + 1 + e^{-4\pi i t j/n}) = \frac{n}{2}. \end{aligned}$$

Appendix R

R Supplement

R.1 First Things First

The website for the text is <http://www.stat.pitt.edu/stoffer/tsa3/>. If you do not already have R, point your browser to the Comprehensive R Archive Network (CRAN), <http://cran.r-project.org/> and download and install it. The installation includes help files and some user manuals. You can find helpful tutorials by following CRAN's link to *Contributed Documentation*. Also, beginners will find RStudio (<https://www.rstudio.com/>) helpful for using R.

R.1.1 The Package for the Text: `astsa`

There is an R package for the text called `astsa` (*Applied Statistical Time Series Analysis*), which was the name of the software distributed with the first and second editions of this text, and the original version, Shumway (1988). The package can be obtained from CRAN and its mirrors in the usual way.

To download and install `astsa`, start R and type

```
install.packages("astsa")
```

You will be asked to choose the closest CRAN mirror to you. As with all packages, you have to load `astsa` before you use it by issuing the command

```
require(astsa)
```

This can be automated by creating a `.First` function. For example, start R, type

```
.First <- function(){require(astsa)}
```

and save the workspace when you quit. Now `astsa` will be loaded at every start until you change `.First`. All the data are loaded when the package is loaded. To see the list of data files available for use, type

```
data()
```

R is not consistent with help files across different operating systems. The best help system is the html help, which can be started issuing the command

```
help.start()
```

and then following the *Packages* link to `astsa`. In Windows, you can type

```
help(package=astsa)
```

to get to html help quickly. Another option is to type

```
?astsa
```

and then use the link to the package *Index* at the bottom of the page. A pdf version of the **astsa** manual may be found at CRAN: <http://cran.r-project.org/web/packages/astsa/astsa.pdf>.

R.1.2 Other Packages

We have already discussed how to obtain and load the R package for this text, **astsa**. You may want to use other R packages. As with **astsa**, you have to first download the package and then install it. For example,

```
install.packages("dynlm")
```

will download and install the package **dynamic linear regression** that can be used in Chapter 2; again, you will be asked to choose the closest mirror to you. To use a package, you have to load it at each start up of R, for example:

```
require(dynlm) # load the package
```

To make sure you have the current version of your R packages, periodically update them by issuing the command

```
update.packages(ask=FALSE)
```

You will be asked to choose the nearest repository. Using "**ask=FALSE**" means you won't be asked about updating each package. If you want a list of your packages, run

```
installed.packages()[,1:3]
```

giving the first 3 columns of info: the [1] package name, [2] library path, and [3] version number.

R.2 Getting Started

The best way to use the rest of this appendix is to start up R and enter the example code as it is presented. Also, you can use the results and help files to get a better understanding of how R works (or doesn't work).

The convention throughout the text is that displayed R code is in **blue typewriter font**, output is **purple** and comments are **# green**. Get comfortable, then start her up and try some simple tasks.

```
2+2      # addition
[1] 5
5*5 + 2  # multiplication and addition
[1] 27
5/5 - 3  # division and subtraction
[1] -2
log(exp(pi)) # log, exponential, pi
[1] 3.141593
sin(pi/2) # sinusoids
[1] 1
exp(1)^(-2) # power
[1] 0.1353353
```

```

sqrt(8)    # square root
[1] 2.828427
1:5        # sequences
[1] 1 2 3 4 5
seq(1, 10, by=2) # sequences
[1] 1 3 5 7 9
rep(2,3)   # repeat 2 three times
[1] 2 2 2

```

Next, we'll use *assignment* to make some *objects*:

```

x <- 1 + 2 # put 1 + 2 in object x
x = 1 + 2  # same as above with fewer keystrokes
1 + 2 -> x # same
x          # view object x
[1] 3
(y = 9*3)  # put 9 times 3 in y and view the result
[1] 27
(z = rnorm(5,0,1)) # put 5 standard normals into z and print z
[1] 0.96607946 1.98135811 -0.06064527 0.31028473 0.02046853

```

It is worth pointing out R's *recycling rule* for doing arithmetic. In the code below, `c()` [concatenation] is used to create a vector. Note the use of the semicolon for multiple commands on one line.

```

x = c(1, 2, 3, 4); y = c(2, 4, 6, 8); z = c(10, 20); w = c(8, 3, 2)
x*y # it's 1*2, 2*4, 3*6, 4*8
[1] 2 8 18 32
x/z # it's 1/10, 2/20, 3/10, 4/20
[1] 0.1 0.1 0.3 0.2
x+z # guess
[1] 11 22 13 24
y+w # what happened here?
[1] 10 7 8 16
Warning message:
In y + w : longer object length is not a multiple of
shorter object length

```

To list your objects, remove objects, get help, find out which directory is current (or to change it) or to quit, use the following commands:

```

ls()      # list all objects
[1] "dummy" "mydata" "x" "y" "z"
ls(pattern = "my") # list every object that contains "my"
[1] "dummy" "mydata"
rm(dummy) # remove object "dummy"
rm(list=ls()) # remove almost everything (use with caution)
help.start() # html help and documentation (use it)
data()      # list of available data sets
help(exp)   # specific help (?exp is the same)
getwd()     # get working directory
setwd()     # change working directory
q()         # end the session (keep reading)

```

When you quit, R will prompt you to save an image of your current workspace. Answering “yes” will save all the work you have done so far, and load it up when you next start R. Our suggestion is to answer “yes” most of the time, even though you will also be loading irrelevant past analyses every time you start R. Keep in mind that you can remove items via `rm()`.

To create your own data set, you can make a data vector as follows:

```
mydata = c(1,2,3,2,1)
```

Now you have an object called `mydata` that contains five elements. R calls these objects *vectors* even though they have no dimensions (no rows, no columns); they do have order and length:

```
mydata      # display the data
[1] 1 2 3 2 1
mydata[3]   # the third element
[1] 3
mydata[3:5] # elements three through five
[1] 3 2 1
mydata[-(1:2)] # everything except the first two elements
[1] 3 2 1
length(mydata) # number of elements
[1] 5
dim(mydata)   # no dimensions
NULL
mydata = as.matrix(mydata) # make it a matrix
dim(mydata)   # now it has dimensions
[1] 5 1
```

If you have an external data set, you can use `scan` or `read.table` (or some variant) to input the data. For example, suppose you have an ASCII (text) data file called `dummy.txt` in your working directory, and the file looks like this:

1	2	3	2	1
9	0	2	1	0

```
(dummy = scan("dummy.txt"))      # scan and view it
Read 10 items
[1] 1 2 3 2 1 9 0 2 1 0
(dummy = read.table("dummy.txt")) # read and view it
V1 V2 V3 V4 V5
1  2  3  2  1
9  0  2  1  0
```

There is a difference between `scan` and `read.table`. The former produced a data vector of 10 items while the latter produced a *data frame* with names `V1` to `V5` and two observations per variate. In this case, if you want to list (or use) the second variate, `V2`, you would use

```
dummy$V2
[1] 2 0
```

and so on. You might want to look at the help files `?scan` and `?read.table` now. Data frames (`?data.frame`) are “used as the fundamental data structure by most of R’s modeling software.” Notice that R gave the columns of `dummy` generic names, `V1`, ..., `V5`. You can provide your own names and then use the names to access the data without the use of `$` as in line 4 above.

```
colnames(dummy) = c("Dog", "Cat", "Rat", "Pig", "Man")
```



```
attach(dummy)
Cat
[1] 2 0
Rat*(Pig - Man) # animal arithmetic
[1] 3 2
detach(dummy) # clean up (if desired)
```

R is case sensitive, thus `cat` and `Cat` are different. Also, `cat` is a reserved name (`?cat`) in R, so using `"cat"` instead of `"Cat"` may cause problems later. You may also include a *header* in the data file to avoid using line 5 above. For example, if the file `dummy.txt` looks like this,

Dog	Cat	Rat	Pig	Man
1	2	3	2	1
9	0	2	1	0

then use the following command to read the data.

```
(dummy = read.table("dummy.txt", header=TRUE))
      Dog Cat Rat Pig Man
1      1  2  3  2  1
2      9  0  2  1  0
```

Another popular format for data files is `.csv` (comma separated values). In this case, use `read.csv` instead of `read.table`. The default for `.csv` files is `header=TRUE`; type `?read.table` for further information.

Two commands that are used frequently to manipulate data are `cbind` for *column binding* and `rbind` for *row binding*. The following is an example.

```
x = runif(4) # generate 4 values from a uniform(0,1) into object x
y = runif(4) # generate 4 more and put them into object y
(u = cbind(x,y)) # column bind the two vectors (4 by 2 matrix)
      x      y
[1,] 0.6547304 0.7503984
[2,] 0.8222048 0.1335557
[3,] 0.4555755 0.2151735
[4,] 0.9843289 0.8483795
(u = rbind(x,y)) # row bind the two vectors (2 by 4 matrix)
      [,1]      [,2]      [,3]      [,4]
x 0.6547304 0.8222048 0.4555755 0.9843289
y 0.7503984 0.1335557 0.2151735 0.8483795
```

R.2.1 Basic Statistics

Summary statistics are fairly easy to obtain. We will simulate 25 normals with $\mu = 10$ and $\sigma = 4$ and then perform some basic analyses.

```
set.seed(90210) # so you can reproduce these results
x = rnorm(25, 10, 4) # generate the data
mean(x) # compute the mean
[1] 9.473883
median(x) # compute the median
[1] 9.448511
var(x) # compute the variance
```

```

[1] 13.9267
sd(x) # compute the standard deviation
[1] 3.73185
max(x) # find the largest value
[1] 17.32609
which.max(x) # index of the max (x[25] in this case)
[1] 25
min(x) # find the smallest value
[1] 2.678173
summary(x) # a five number summary (with the mean added)
      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 2.678   7.824   9.449   9.474  11.180  17.330
stem(x) # stem and leaf diagram
The decimal point is 1 digit(s) to the right of the |
 0 | 334
 0 | 5778888899
 1 | 000111234
 1 | 567
par(mfrow=c(1,2)) # multifigure setup (?par for info)
hist(x, col="lightblue", prob=TRUE) # histogram; see Figure R.1
lines(density(x, bw=2)) # fit a density over it
boxplot(x, main="Boxplot of x", col="lightblue") # see Figure R.1

```

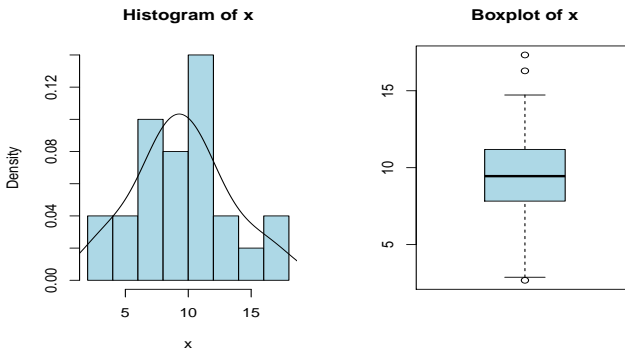


Fig. R.1. Histogram and boxplot of simulated normal data.

We introduced some R graphics without saying much about it. Aside from the manual, there are many books and useful websites that expand on this topic.

Now let's try a simple t-test. Suppose you do not know the command for the t-test. You could guess it.

```
?ttest
```

```

No documentation for 'ttest' in specified packages
and libraries: you could try '??ttest'

```

If you try `??ttest` you will get more information than you will know what to do with. The best thing is to do an internet search on *R ttest*, and you will find what you are looking for right away.

In the following, we will perform a two sample t-test where the data are obtained by simulating from standard normals with $n_1 = 24$ and $n_2 = 37$.

```
t.test(x = rnorm(24), y = rnorm(37), var.equal=TRUE)

Two Sample t-test
data:  rnorm(24) and rnorm(37)
t = -0.8119, df = 59, p-value = 0.4201
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6887137  0.2911483
sample estimates:
 mean of x      mean of y 
-0.196679216  0.002103468

?t.test      # for more options
```

It can't hurt to learn a little about programming in R because you will see some of it along the way. Consider a simple program that we will call **crazy** to produce a graph of a sequence of sample means of increasing sample sizes from a Cauchy distribution with location parameter zero. The code is:

```
crazy <- function(num) {
  x <- rep(NA, num)
  for (n in 1:num) x[n] <- mean(rcauchy(n))
  plot(x, type="l", xlab="sample size", ylab="sample mean")
}
```

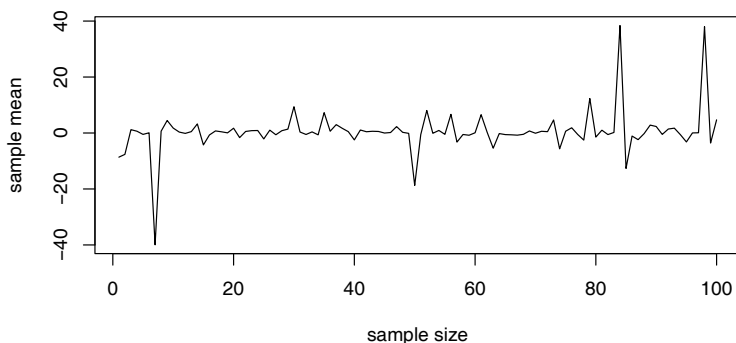
The first line creates the function **crazy** and gives it one argument, **num**, that is the sample size that will end the sequence. Line 2 makes a vector, **x**, of **num** missing values **NA**, that will be used to store the sample means. Line 3 generates **n** random Cauchy variates [**rcauchy(n)**], finds the mean of those values, and puts the result into **x[n]**, the *n*-th value of **x**. The process is repeated in a “do loop” **num** times so that **x[1]** is the sample mean from a sample of size one, **x[2]** is the sample mean from a sample of size two, and so on, until finally, **x[num]** is the sample mean from a sample of size **num**. After the do loop is complete, the fourth line generates a graphic (see [Figure R.2](#)). The fifth line closes the function. To use **crazy** with a limit sample size of 100, for example, type

```
crazy(100)
```

and you will get a graphic that looks like [Figure R.2](#)

R.2.2 Word

Finally, a word of caution: **TRUE** and **FALSE** are reserved words, whereas **T** and **F** are initially set to these. Get in the habit of using the words rather than the letters **T** or **F** because you may get into trouble if you do something like **F = qf(p=.01, df1=3, df2=9)**, so that **F** is no longer **FALSE**, but a quantile of the specified *F*-distribution.

**Fig. R.2.** Crazy example.

R.3 Time Series Primer

In this section, we give a brief introduction on using R for time series. We assume that `astsa` has been loaded.

To create a time series object, use the command `ts`. Related commands are `as.ts` to coerce an object to a time series and `is.ts` to test whether an object is a time series.

First, make a small data set:

```
(mydata = c(1,2,3,2,1)) # make it and view it
[1] 1 2 3 2 1
```

Now make it a time series:

```
(mydata = as.ts(mydata))
Time Series:
Start = 1
End = 5
Frequency = 1
[1] 1 2 3 2 1
```

Make it an annual time series that starts in 1950:

```
(mydata = ts(mydata, start=1950))
Time Series:
Start = 1950
End = 1954
Frequency = 1
[1] 1 2 3 2 1
```

Now make it a quarterly time series that starts in 1950-III:

```
(mydata = ts(mydata, start=c(1950,3), frequency=4))
      Qtr1 Qtr2 Qtr3 Qtr4
1950             1     2
1951     3     2     1
time(mydata) # view the sampled times
      Qtr1      Qtr2      Qtr3      Qtr4
1950             1950.50 1950.75
1951 1951.00 1951.25 1951.50
```

To use part of a time series object, use `window()`:

```
(x = window(mydata, start=c(1951,1), end=c(1951,3)))
```

```
      Qtr1 Qtr2 Qtr3
1951     3     2     1
```

Next, we'll look at lagging and differencing. First make a simple series, x_t :

```
x = ts(1:5)
```

Now, column bind (`cbind`) lagged values of x_t and you will notice that `lag(x)` is *forward* lag, whereas `lag(x, -1)` is *backward* lag (we display the time series attributes in a single row of the output to save space).

```
cbind(x, lag(x), lag(x,-1))
```

	x	lag(x)	lag(x, -1)	
0	NA	1	NA	
1	1	2	NA	
2	2	3	1	
3	3	4	2	2 <- in this row, for example, x is 3,
4	4	5	3	lag(x) is ahead at 4, and
5	5	NA	4	lag(x,-1) is behind at 2
6	NA	NA	5	

Compare `cbind` and `ts.intersect`:

```
ts.intersect(x, lag(x,1), lag(x,-1))
```

	x	lag(x, 1)	lag(x, -1)
2	2	3	1
3	3	4	2
4	4	5	3

To difference a series, $\nabla x_t = x_t - x_{t-1}$, use

```
diff(x)
```

but note that

```
diff(x, 2)
```

is *not* second order differencing, it is $x_t - x_{t-2}$. For second order differencing, that is, $\nabla^2 x_t$, do this:

```
diff(diff(x))
```

and so on for higher order differencing.

For graphing time series, there are a few standard plotting mechanisms that we use repeatedly. If `x` is a time series, then `plot(x)` will produce a time plot. If `x` is not a time series object, then `plot.ts(x)` will coerce it into a time plot as will `ts.plot(x)`. There are differences, which we explore in the following. It would be a good idea to skim the graphical parameters help file (`?par`) while you are here. See [Figure R.3](#) for the resulting graphic.

```
x = -5:5      # x is NOT a time series object
y = 5*cos(x)  # neither is y
par(mfrow=c(3,2)) # multifigure setup: 3 rows, 2 cols
plot(x, main="plot(x)")
plot(x, y, main="plot(x,y)")
plot.ts(x, main="plot.ts(x)")
plot.ts(x, y, main="plot.ts(x,y)")
ts.plot(x, main="ts.plot(x)")
ts.plot(ts(x), ts(y), col=1:2, main="ts.plot(x,y)")
```

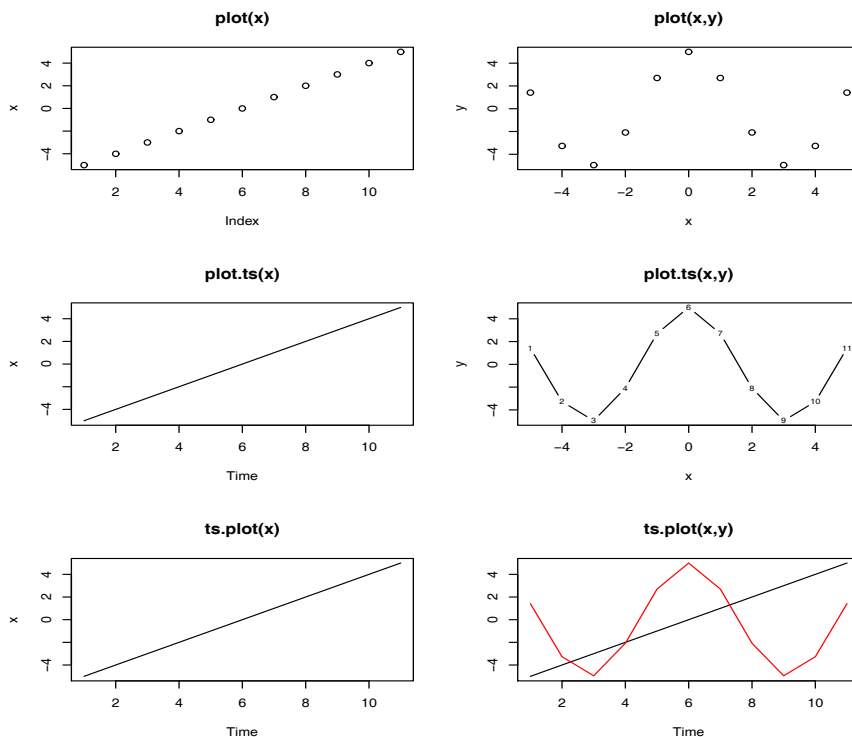


Fig. R.3. Demonstration of different R graphic tools for plotting time series.

We will also make use of regression via `lm()`. First, suppose we want to fit a simple linear regression, $y = \alpha + \beta x + \epsilon$. In R, the formula is written as `y~x`:

```
set.seed(1999)      # so you can reproduce the result
x = rnorm(10,0,1)
y = x + rnorm(10,0,1)
summary(fit <- lm(y~x))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.8851	-0.3867	0.1325	0.3896	0.6561

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.2576	0.1892	1.362	0.2104
x	0.4577	0.2016	2.270	0.0529

Residual standard error: 0.58 on 8 degrees of freedom

Multiple R-squared: 0.3918, Adjusted R-squared: 0.3157

F-statistic: 5.153 on 1 and 8 DF, p-value: 0.05289

```
plot(x, y) # draw a scatterplot of the data (not shown)
abline(fit) # add the fitted line to the plot (not shown)
```

All sorts of information can be extracted from the `lm` object, which we called `fit`. For example,

```
resid(fit)      # will display the residuals (not shown)
fitted(fit)     # will display the fitted values (not shown)
lm(y ~ 0 + x)  # will exclude the intercept (not shown)
```

You have to be careful if you use `lm()` for lagged values of a time series. If you use `lm()`, then what you have to do is “tie” the series together using `ts.intersect`. If you do not tie the series together, they will not be aligned properly. Please read the warning *Using time series* in the `lm()` help file [`help(lm)`]. Here is an example regressing Chapter 2 data, weekly cardiovascular mortality (`cmort`) on particulate pollution (`part`) at the present value and lagged four weeks (`part4`). First, we create a data frame called `ded` that consists of the three series:

```
data(cmort, part) # if these are not already in the workspace
ded = ts.intersect(cmort, part, part4=lag(part,-4), dframe=TRUE)
```

Now the series are all aligned and the regression will work.

```
fit = lm(cmort~part+part4, data=ded, na.action=NULL)
summary(fit)

Call: lm(formula=cmort~part+part4,data=ded,na.action=NULL)
Residuals:
    Min       1Q   Median       3Q      Max
-22.7429  -5.3677  -0.4136   5.2694  37.8539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.01020    1.37498   50.190 < 2e-16
part         0.15140    0.02898    5.225 2.56e-07
part4        0.26297    0.02899    9.071 < 2e-16

Residual standard error: 8.323 on 501 degrees of freedom
Multiple R-Squared:  0.3091,    Adjusted R-squared:  0.3063
F-statistic: 112.1 on 2 and 501 DF,  p-value: < 2.2e-16
```

There was no need to rename `lag(part,-4)` to `part4`, it’s just an example of what you can do. There is a package called `dynlm` that makes it easy to fit lagged regressions. The basic advantage of `dynlm` is that it avoids having to make a data frame; that is, line 2 would be avoided.

In **Problem 2.1**, you are asked to fit a regression model

$$x_t = \beta t + \alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t) + w_t$$

where x_t is logged Johnson & Johnson quarterly earnings ($n = 84$), and $Q_i(t)$ is the indicator of quarter $i = 1, 2, 3, 4$. The indicators can be made using `factor`.

```
trend = time(jj) - 1970 # helps to `center' time
Q = factor(cycle(jj))   # make (Q)uarter factors
reg = lm(log(jj)~0 + trend + Q, na.action=NULL) # no intercept
model.matrix(reg)       # view the model matrix

  trend Q1 Q2 Q3 Q4
1 -10.00 1 0 0 0
2  -9.75 0 1 0 0
3  -9.50 0 0 1 0
4  -9.25 0 0 0 1
```

```

      .      .      .      .      .
      .      .      .      .      .
83    10.50  0  0  1  0
84    10.75  0  0  0  1
summary(reg) # view the results (not shown)

```

The workhorse for ARIMA simulations is `arima.sim`. Here are some examples; no output is shown here so you're on your own.

```

x = arima.sim(list(order=c(1,0,0),ar=.9),n=100)+50 # AR(1) w/mean 50
x = arima.sim(list(order=c(2,0,0),ar=c(1,-.9)),n=100) # AR(2)
x = arima.sim(list(order=c(1,1,1),ar=.9,ma=-.5),n=200) # ARIMA(1,1,1)

```

Next, we'll discuss ARIMA estimation. This gets a bit tricky because R is not user friendly when it comes to fitting ARIMA models. Much of the story is spelled out in the "R Issues" page of the website for the text. In Chapter 3, we use the scripts `acf2`, `sarima`, and `sarima.for` that are included with `astsa` and `tsa3.Rda`. But we will also show you how to use the scripts included with R.

First, we'll fit an ARMA(1,1) model to some simulated data (with diagnostics and forecasting):

```

set.seed(666)
x = 50 + arima.sim(list(order=c(1,0,1), ar=.9, ma=-.5), n=200)
acf(x); pacf(x) # display sample ACF and PACF ... or ...
acf2(x)         # use our script (no output shown)
(x.fit = arima(x, order = c(1, 0, 1))) # fit the model

Call:  arima(x = x, order = c(1, 0, 1))
Coefficients:
      ar1      ma1  intercept
      0.8340  -0.432   49.8960
s.e.   0.0645   0.111   0.2452
sigma^2 estimated as 1.070: log likelihood = -290.79, aic = 589.58

```

Note that the reported `intercept` estimate is an estimate of the mean and *not* the constant. That is, the fitted model is

$$\hat{x}_t - 49.896 = .834(x_{t-1} - 49.896) + \hat{w}_t$$

where $\hat{\sigma}_w^2 = 1.070$. **Incorrect** diagnostics can be accomplished as follows:

```
tsdiag(x.fit, gof.lag=20) # !!!! don't use this !!!!
```

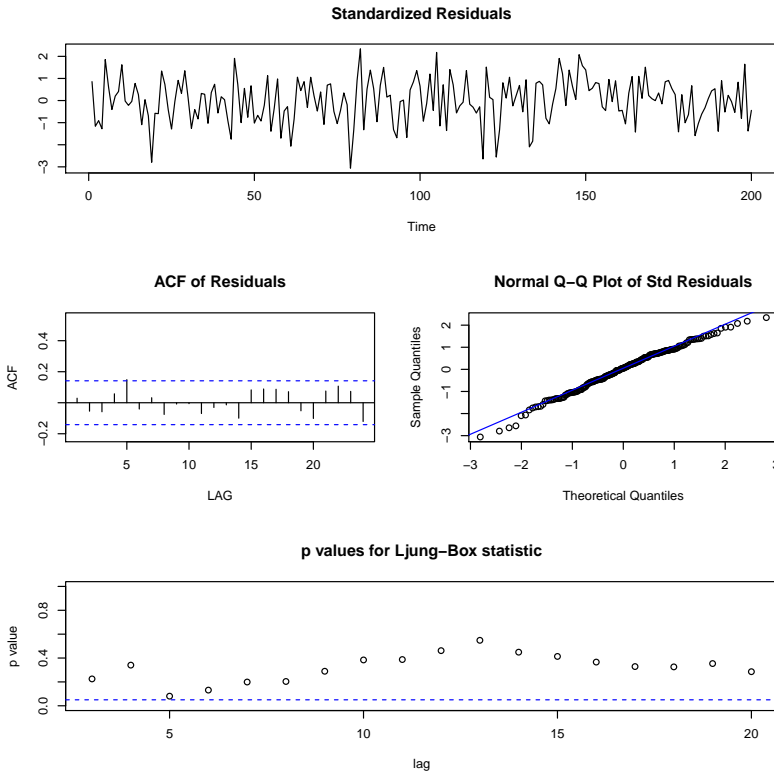
That's right, the Ljung-Box-Pierce test is NOT correct because it does not take into account the fact that the residuals are from a fitted model. If the analysis is repeated using the `sarima` script, a partial output would look like the following (`sarima` will also display the correct diagnostics as a graphic; e.g., see [Figure 3.16](#) on page 101 or [Figure R.4](#)):

```

sarima(x, 1, 0, 1)

Coefficients:
      ar1      ma1    xmean
      0.8340  -0.432  49.8960
s.e.   0.0645   0.111   0.2452
sigma^2 estimated as 1.070: log likelihood = -290.79, aic = 589.58
$AIC [1] 1.097494  $AICc [1] 1.108519  $BIC [1] 0.1469684

```


Fig. R.4. Diagnostics from `sarima`

Forecasting using the script `sarima.for` can be accomplished in one line:

```
sarima.for(x, 10, 1, 0, 1)
```

If you don't use `astsa`, you would obtain and plot the forecasts using the following:

```
x.fore = predict(x.fit, n.ahead=10)
U = x.fore$pred + 2*x.fore$se # x.fore$pred holds predicted values
L = x.fore$pred - 2*x.fore$se # x.fore$se holds stdn errors
miny = min(x,L); maxy = max(x,U)
ts.plot(x, x.fore$pred, col=1:2, ylim=c(miny, maxy))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
```

We close this appendix with a quick spectral analysis. This material is covered in detail in Chapter 4, so we will not discuss this example in much detail here. We will simulate an AR(2) and then estimate the spectrum via nonparametric and parametric methods. No graphics are shown, but we have confidence that you are proficient enough in R to display them yourself.

```
x = arima.sim(list(order=c(2,0,0), ar=c(1,-.9)), n=2^8) # some data
(u = polyroot(c(1,-1,.9))) # x is AR(2) w/complex roots
[1] 0.5555556+0.8958064i 0.5555556-0.8958064i
Arg(u[1])/(2*pi) # dominant frequency around .16
```

```
[1] 0.1616497
par(mfcol=c(2,2))
plot.ts(x)
mvspec(x, spans=c(5,5), taper=.1, log="no") # nonparametric estimate
spec.ar(x, log="no") # parametric spectral estimate
arma.spec(ar=c(1,-.9), log="no") # true spectral density
```

The script `arma.spec` is included in `astsa`. See `spectrum` and `spec.pgram` as alternatives to `mvspec`, which is part of `astsa`. Finally, note the easiest way to get a raw periodogram is:

```
per = abs(fft(x))^2/length(x) # abs() and Mod() are same here
```

R.3.1 Included Data Sets

The data sets included in `astsa`, listed by the chapter in which they are first presented, are as follows.

CHAPTER 1

- jj - Johnson & Johnson quarterly earnings per share, 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980.
- EQ5 - Seismic trace of an earthquake [two phases or arrivals along the surface, the primary wave ($t = 1, \dots, 1024$) and the shear wave ($t = 1025, \dots, 2048$)] recorded at a seismic station.
- EXP6 - Seismic trace of an explosion (similar details as EQ5).
- gtemp - Global mean land-ocean temperature deviations (from 1951-1980 average), measured in degrees centigrade, for the years 1880-2009; data taken from <http://data.giss.nasa.gov/gistemp/graphs/>
- fmri1 - A data frame that consists of fMRI BOLD signals at eight locations (in columns 2-9, column 1 is time period), when a stimulus was applied for 32 seconds and then stopped for 32 seconds. The signal period is 64 seconds and the sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$).
- soi - Southern Oscillation Index (SOI) for a period of 453 months ranging over the years 1950-1987.
- rec - Recruitment (number of new fish) associated with SOI.
- speech - A small .1 second (1000 points) sample of recorded speech for the phrase *aaa...hhh*.
- nyse - Returns of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991.
- soiltemp - A 64×36 matrix of surface soil temperatures.

CHAPTER 2

- oil - Crude oil, WTI spot price FOB (in dollars per barrel), weekly data from 2000 to mid-2010. For definitions and more details, go to http://tonto.eia.doe.gov/dnav/pet/pet_pri_spt_s1_w.htm.
- gas - New York Harbor conventional regular gasoline weekly spot price FOB (in cents per gallon) over the same time period as `oil`.
- varve - Sedimentary deposits from one location in Massachusetts for 634 years, beginning nearly 12,000 years ago.

- cmort** - Average weekly cardiovascular mortality in Los Angeles County; 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.
- tempr** - Temperature series corresponding to **cmort**.
- part** - Particulate series corresponding to **cmort**.
- so2** - Sulfur dioxide series corresponding to **cmort**.
- lap** - LA Pollution-Mortality Study data (1970-1979, weekly data). This is all the data from the study as a multiple time series object. The 11 columns are time series with **names**: (1) Total Mortality - **tmort** (2) Respiratory Mortality - **rmort** (3) Cardiovascular Mortality - **cmort** (4) Temperature - **tempr** (5) Relative Humidity - **rh** (6) Carbon Monoxide - **co** (7) Sulfur Dioxide - **so2** (8) Nitrogen Dioxide - **no2** (9) Hydrocarbons - **hycarb** (10) Ozone - **o3** (11) Particulates - **part**

CHAPTER 3

- prodn** - Monthly Federal Reserve Board Production Index (1948-1978, $n = 372$ months).
- unemp** - Unemployment series corresponding to **prodn**.
- arlboot** - Data not used in this version.
- gnp** - Quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations.
- birth** - Monthly live births (adjusted) in thousands for the United States, 1948-1979.

CHAPTER 4

- sunspotz** - Biannual smoothed (12-month moving average) number of sunspots from June 1749 to December 1978; $n = 459$. The “z” on the end is to distinguish this series from the one included with R (called **sunspots**).
- salt** - Salt profiles taken over a spatial grid set out on an agricultural field, 64 rows at 17-ft spacing.
- saltemp** - Temperature profiles corresponding to **salt**.

CHAPTER 5

- arf** - 1000 simulated observations from an ARFIMA(1, 1, 0) model with $\phi = .75$ and $d = .4$.
- flu** - Monthly pneumonia and influenza deaths per 10,000 people in the United States for 11 years, 1968 to 1978.
- sales** - Sales (with **lead**, a leading indicator), 150 months; taken from Box & Jenkins (1970).
- lead** - See **sales**.
- econ5** - Quarterly U.S. unemployment, GNP, consumption, and government and private investment, from 1948-III to 1988-II.

CHAPTER 6

- ar1miss** - Data from an AR(1) with some values missing at random.
- gtemp2** - Similar to **gtemp** but the data are based only on surface air temperature data obtained from meteorological stations.
- qinfl** - Quarterly inflation rate in the Consumer Price Index from 1953-I to 1980-II, $n = 110$ observations; from Newbold and Bos (1985).
- qintr** - Quarterly interest rate recorded for Treasury bills over the same period as **qinfl**.

WBC - Measurements made for 91 days on the three variables, log(white blood count) [**WBC**], log(platelet) [**PLT**] and hematocrit [**HCT**]; taken from Jones (1984). Missing data code is 0.

PLT - See **WBC**.

HCT - See **WBC**.

blood - **WBC**, **PLT**, **HCT** as a multiple time series with missing data code **NA**.

CHAPTER 7

beamd - Infrasonic signals from a nuclear explosion. This is a data frame consisting of three columns (which are not time series objects) that are data from different channels. The series names are **sensor1**, **sensor2**, **sensor3**.

bnrf1ebv - Nucleotide sequence of the BNRF1 gene of the Epstein-Barr virus (EBV): 1=A, 2=C, 3=G, 4=T.

bnrf1hvs - Nucleotide sequence of the BNRF1 gene of the herpes virus saimiri (HVS); same codes as for EBV.

fmri - Data (as a vector list) from an fMRI experiment in pain, listed by location and stimulus. The specific locations of the brain where the signal was measured were [1] Cortex 1: Primary Somatosensory, Contralateral, [2] Cortex 2: Primary Somatosensory, Ipsilateral, [3] Cortex 3: Secondary Somatosensory, Contralateral, [4] Cortex 4: Secondary Somatosensory, Ipsilateral, [5] Caudate, [6] Thalamus 1: Contralateral, [7] Thalamus 2: Ipsilateral, [8] Cerebellum 1: Contralateral and [9] Cerebellum 2: Ipsilateral. The stimuli (and number of subjects in each condition) are [1] Awake-Brush (5 subjects), [2] Awake-Heat (4 subjects), [3] Awake-Shock (5 subjects), [4] Low-Brush (3 subjects), [5] Low-Heat (5 subjects), and [6] Low-Shock (4 subjects). Issue the command **summary(fmri)** for further details. As an example, **fmri\$L1T6** (location 1, treatment 6) will show the data for the four subjects receiving the Low-Shock treatment at the Cortex 1 location; note that **fmri[[6]]** will display the same data.

climhyd - Lake Shasta inflow data. This is a data frame with column names: **Temp**, **DewPt**, **CldCvr**, **WndSpd**, **Precip**, **Inflow**.

eqexp - This is a data frame of the earthquake and explosion seismic series used throughout the text. The matrix has 17 columns, the first eight are earthquakes, the second eight are explosions, and the last column is the Novaya Zemlya series. The column names are: **EQ1**, **EQ2**, ..., **EQ8**; **EX1**, **EX2**, ..., **EX8**; **NZ**.

R.3.2 Included Scripts

The following scripts are included in **astsa**.

```
lag2.plot(series1, series2, max.lag=0, corr=TRUE, smooth=TRUE)
```

Produces a grid of scatterplots of one series versus another. If (x_t, y_t) is a vector time series, then **lag.plot2(x,y,m)** will generate a grid of scatterplots of x_{t-h} versus y_t for $h = 0, 1, \dots, m$, along with the cross-correlation values (**corr=TRUE**) and a lowess fit (**smooth=TRUE**) assuming x_t is in **x** and y_t is in **y**. Note that the first series, x_t , is the one that gets lagged. If you just want the scatterplots and nothing

else, then use `lag2.plot(x,y,m,corr=FALSE,smooth=FALSE)`. See [Example 2.7](#) for a demonstration.

```
lag1.plot(series, max.lag=1, corr=TRUE, smooth=TRUE)
```

Produces a grid of scatterplots of a series versus lagged values of the series. Similar to `lag2.plot`, the call `lag1.plot(x,m)` will generate a grid of scatterplots of x_{t-h} versus x_t for $h = 1, \dots, m$, along with the autocorrelation values (`corr=TRUE`) and a lowess fit (`smooth=TRUE`). The defaults are the same as `lag2.plot`. See [Example 2.7](#) for a demonstration.

```
acf2(series, max.lag=NULL)
```

Produces a simultaneous plot (and a printout) of the sample ACF and PACF on the same scale. If `x` contains n observations, `acf2(x)` will print and plot the ACF and PACF of `x` to the default lag of $\sqrt{n} + 10$ (unless n is smaller than 50). The number of lags may be specified, e.g., `acf2(x, 33)`. See [Example 3.16](#) on page 73.

```
sarima(series, p, d, q, P=0, D=0, Q=0, S=-1, details=TRUE,
       tol=sqrt(.Machine$double.eps), no.constant=FALSE)
```

Fits ARIMA models including diagnostics in a short command. If your time series is in `x` and you want to fit an $\text{ARIMA}(p, d, q)$ model to the data, the basic call is `sarima(x,p,d,q)`. The results are the parameter estimates, standard errors, AIC, AICc, BIC (as defined in Chapter 2) and diagnostics. To fit a seasonal ARIMA model, the basic call is `sarima(x,p,d,q,P,D,Q,S)`. So, for example, `sarima(x,2,1,0)` will fit an $\text{ARIMA}(2,1,0)$ model to the series in `x`, and `sarima(x,2,1,0,0,1,1,12)` will fit a seasonal $\text{ARIMA}(2,1,0) \times (0,1,1)_{12}$ model to the series in `x`. The `arma` object is stored in `fit: summary(sarima(x,2,1,0)$fit)`; e.g., you may want to inspect the innovations (i.e., the residuals) further.

There are three additional options that can be included in the call.

- `details` turns on/off the output from the nonlinear optimization routine, which is `optim`. The default is `TRUE`, use `details=FALSE` to turn off the output.
- `tol` controls the relative tolerance (`reltol`) used to assess convergence in `sarima` and `sarima.for`. The default is `tol=sqrt(.Machine$double.eps)`, the R default. For details, see the help file for `optim` under the `control` arguments.
- `no.constant` controls whether or not `sarima` includes a constant in the model. In particular, with `sarima`, if there is no differencing ($d = 0$ and $D = 0$) you get the mean estimate. If there's differencing of order one (either $d = 1$ or $D = 1$, but not both), a constant term is included in the model. These may be overridden by setting this to `TRUE`.

```
sarima.for(series, n.ahead, p, d, q, P=0, D=0, Q=0, S=-1,
           tol=sqrt(.Machine$double.eps), no.constant=FALSE)
```

Gives ARIMA forecasts. Similar to `sarima`, to forecast `n.ahead` time points from an ARIMA fit to the data in `x`, the form is `sarima.for(x, n.ahead, p, d, q)` or `sarima.for(x, n.ahead, p, d, q, P, D, Q, S)` for a seasonal model. For example, `sarima.for(x, 5, 1, 0, 1)` will forecast five time points ahead for an ARMA(1,1) fit to `x`. The output prints the forecasts and the standard errors of the forecasts, and supplies a graphic of the forecast with ± 2 prediction error bounds. The options `tol` and `no.constant` are also available.

```
arma.spec(ar=0, ma=0, var.noise=1, n.freq=500, ...)
```

Gives the ARMA spectrum (on a log scale), tests for causality, invertibility, and common zeros. The basic call is `arma.spec(ar, ma)` where `ar` and `ma` are vectors containing the model parameters. Use `log="no"` if you do not want the plot on a log scale. If the model is not causal or invertible an error message is given. If there are common zeros, a spectrum will be displayed and a warning will be given; e.g., `arma.spec(ar=.9, ma=-.9)` will yield a warning and the plot will be the spectrum of white noise. The variance of the noise can be changed with `var.noise`. Finally, the frequencies and the spectral density ordinates are returned invisibly, e.g., `arma.spec(ar=.9)$freq` and `arma.spec(ar=.9)$spec`, if you're interested in the actual values. See [Example 4.6](#) on page 129. With `tsa3.rda`, use `spec.arma(ar=0, ma=0, var.noise=1, n.freq=500, ...)`

```
mvspec(x, spans = NULL, kernel = NULL, taper = 0, pad = 0, fast = TRUE,
       demean = FALSE, detrend = TRUE, plot = TRUE,
       na.action = na.fail, ...)
```

Performs univariate and multivariate spectral estimation. This is `spec.pgram` with a few changes in the defaults and written so you can easily extract the estimate of a multivariate spectral matrix as `fxx`. The bandwidth calculation has been changed to the more practical definition given in the text, $(L_h/n) \times \text{frequency}(x)$. The script can be used in Chapter 4 for the spectral analysis of a univariate time series. Note that the script does not taper by default (`taper=0`); this forces the user to do conscious tapering.

In the multivariate case, if `x` contains a p -variate time series (i.e., the p columns of `x` are time series), and you issue the command `spec = mvspec(x, spans=3)` say, then `spec$fxx` is an array with dimensions `dim=c(p,p,nfreq)`, where `nfreq` is the number of frequencies used. If you print `spec$fxx`, you will see `nfreq` $p \times p$ spectral matrix estimates.

References

- Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, 21, 243-247.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principal. In *2nd Int. Symp. Inform. Theory*, 267-281. B.N. Petrov and F. Csake, eds. Budapest: Akademia Kiado.
- Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.*, AC-19, 716-723.
- Alagón, J. (1989). Spectral discrimination for two groups of time series. *J. Time Series Anal.*, 10, 203-214.
- Alspach, D.L. and H.W. Sorensen (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Trans. Automat. Contr.*, AC-17, 439-447.
- Anderson, B.D.O. and J.B. Moore (1979). *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall.
- Anderson, T.W. (1978). Estimation for autoregressive moving average models in the time and frequency domain. *Ann. Stat.*, 5, 842-865.
- Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley.
- Ansley, C.F. and P. Newbold (1980). Finite sample properties of estimators for autoregressive moving average processes. *J. Econ.*, 13, 159-183.
- Ansley, C.F. and R. Kohn (1982). A geometrical derivation of the fixed interval smoothing algorithm. *Biometrika*, 69, 486-487.
- Antognini, J.F., M.H. Buonocore, E.A. Disbrow, and E. Carstens (1997). Isoflurane anesthesia blunts cerebral responses to noxious and innocuous stimuli: a fMRI study. *Life Sci.*, 61, PL349-PL354.
- Bandettini, A., A. Jesmanowicz, E.C. Wong, and J.S. Hyde (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Res. Med.*, 30, 161-173.
- Bar-Shalom, Y. (1978). Tracking methods in a multi-target environment. *IEEE Trans. Automat. Contr.*, AC-23, 618-626.

- Bar-Shalom, Y. and E. Tse (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11, 4451-4460.
- Bazza, M., R.H. Shumway, and D.R. Nielsen (1988). Two-dimensional spectral analysis of soil surface temperatures. *Hilgardia*, 56, 1-28.
- Bedrick, E.J. and C.-L. Tsai (1994). Model selection for multivariate regression in small samples. *Biometrics*, 50, 226-231.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 289-300.
- Beran, J. (1994). *Statistics for Long Memory Processes*. New York: Chapman and Hall.
- Berk, K.N. (1974). Consistent autoregressive spectral estimates. *Ann. Stat.*, 2, 489-502.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. B*, 36, 192-236.
- Bhat, R.R. (1985). *Modern Probability Theory, 2nd ed.* New York: Wiley.
- Bhattacharya, A. (1943). On a measure of divergence between two statistical populations. *Bull. Calcutta Math. Soc.*, 35, 99-109.
- Billingsley, P. (1999). *Convergence of Probability Measures, (2nd edition)*. New York: Wiley.
- Blackman, R.B. and J.W. Tukey (1959). *The Measurement of Power Spectra from the Point of View of Communications Engineering*. New York: Dover.
- Blight, B.J.N. (1974). Recursive solutions for the estimation of a stochastic parameter. *J. Am. Stat. Assoc.*, 69, 477-481.
- Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.
- Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction, 2nd ed.* New York: Wiley.
- Bloomfield, P. and J.M. Davis (1994). Orthogonal rotation of complex principal components. *Int. J. Climatol.*, 14, 759-775.
- Bogart, B. P., M. J. R. Healy, and J.W. Tukey (1962). The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proc. of the Symposium on Time Series Analysis*, pp. 209-243, Brown University, Providence, USA.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 31, 307- 327.
- Box, G.E.P. and D.A. Pierce (1970). Distributions of residual autocorrelations in autoregressive integrated moving average models. *J. Am. Stat. Assoc.*, 72, 397-402.
- Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis, Forecasting, and Control*. Oakland, CA: Holden-Day.
- Box, G.E.P. and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

- Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994). *Time Series Analysis, Forecasting, and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.
- Breiman, L. and J. Friedman (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Stat. Assoc.*, 80, 580-619.
- Brillinger, D.R. (1973). The analysis of time series collected in an experimental design. In *Multivariate Analysis-III.*, pp. 241-256. P.R. Krishnaiah ed. New York: Academic Press.
- Brillinger, D.R. (1975). *Time Series: Data Analysis and Theory*. New York: Holt, Rinehart & Winston Inc.
- Brillinger, D.R. (1980). Analysis of variance and problems under time series models. In *Handbook of Statistics*, Vol I, pp. 237-278. P.R. Krishnaiah and D.R. Brillinger, eds. Amsterdam: North Holland.
- Brillinger, D.R. (1981, 2001). *Time Series: Data Analysis and Theory*, 2nd ed. San Francisco: Holden-Day. Republished in 2001 by the Society for Industrial and Applied Mathematics, Philadelphia.
- Brockwell, P.J. and R.A. Davis (1991). *Time Series: Theory and Methods*, 2nd ed. New York: Springer-Verlag.
- Bruce, A. and H-Y. Gao (1996). *Applied Wavelet Analysis with S-PLUS*. New York: Springer-Verlag.
- Caines, P.E. (1988). *Linear Stochastic Systems*. New York: Wiley.
- Carlin, B.P., N.G. Polson, and D.S. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Stat. Assoc.*, 87, 493-500.
- Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika*, 81, 541-553.
- Chan, N.H. (2002). *Time Series: Applications to Finance*. New York: Wiley.
- Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of the observations. *Ann. Math. Stat.*, 25, 573-578.
- Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, 74, 829-836.
- Cochrane, D. and G.H. Orcutt (1949). Applications of least squares regression to relationships containing autocorrelated errors. *J. Am. Stat. Assoc.*, 44, 32-61.
- Cooley, J.W. and J.W. Tukey (1965). An algorithm for the machine computation of complex Fourier series. *Math. Comput.*, 19, 297-301.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*. New York: Wiley.
- Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Stat.*, 17, 1749-1766.
- Dargahi-Noubary, G.R. and P.J. Laycock (1981). Spectral ratio discriminants and information theory. *J. Time Series Anal.*, 16, 201-219.
- Danielson, J. (1994). Stochastic volatility in asset prices: Estimation with simulated maximum likelihood. *J. Econometrics*, 61, 375-400.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics.

- Davies, N., C.M. Triggs, and P. Newbold (1977). Significance levels of the Box-Pierce portmanteau statistic in finite samples. *Biometrika*, 64, 517-522.
- Dent, W. and A.-S. Min. (1978). A Monte Carlo study of autoregressive-integrated-moving average processes. *J. Econ.*, 7, 23-55.
- Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39, 1-38.
- Ding, Z., C.W.J. Granger, and R.F. Engle (1993). A long memory property of stock market returns and a new model. *J. Empirical Finance*, 1, 83-106.
- Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-455.
- Donoho, D.L. and I.M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. of Am. Stat. Assoc.*, 90, 1200-1224.
- Durbin, J. (1960). Estimation of parameters in time series regression models. *J. R. Stat. Soc. B*, 22, 139-153.
- Durbin, J. and S.J. Koopman (2001). *Time Series Analysis by State Space Methods* Oxford: Oxford University Press.
- Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987-1007.
- Engle, R.F., D. Nelson, and T. Bollerslev (1994). ARCH Models. In *Handbook of Econometrics*, Vol IV, pp. 2959-3038. R. Engle and D. McFadden, eds. Amsterdam: North Holland.
- Fox, R. and M.S. Taquq (1986). Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Stat.*, 14, 517-532.
- Friedman, J.H. (1984). A Variable Span Smoother. Tech. Rep. No. 5, Lab. for Computational Statistics, Dept. Statistics, Stanford Univ., California.
- Friedman, J.H. and W. Stuetzle. (1981). Projection pursuit regression. *J. Am. Stat. Assoc.*, 76, 817-823.
- Frühwirth-Schnatter, S. (1994). Data Augmentation and Dynamic Linear Models. *J. Time Series Anal.*, 15, 183-202.
- Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series*, 2nd ed. New York: Wiley.
- Gelfand, A.E. and A.F.M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85, 398-409.
- Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-741.
- Geweke, J.F. (1977). The dynamic factor analysis of economic time series models. In *Latent Variables in Socio-Economic Models*, pp 365-383. D. Aigner and A. Goldberger, eds. Amsterdam: North Holland.

- Geweke, J.F. and K.J. Singleton (1981). Latent variable models for time series: A frequency domain approach with an application to the Permanent Income Hypothesis. *J. Econ.*, 17, 287-304.
- Geweke, J.F. and S. Porter-Hudak (1983). The estimation and application of long-memory time series models. *J. Time Series Anal.*, 4, 221-238.
- Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
- Giri, N. (1965). On complex analogues of T^2 and R^2 tests. *Ann. Math. Stat.*, 36, 664-670.
- Goldfeld, S.M. and R.E. Quandt (1973). A Markov model for switching regressions. *J. Econ.*, 1, 3-16.
- Goodman, N.R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Stat.*, 34, 152-177.
- Gordon, K. and A.F.M. Smith (1988). Modeling and monitoring discontinuous changes in time series. In *Bayesian Analysis of Time Series and Dynamic Models*, 359-392.
- Gordon, K. and A.F.M. Smith (1990). Modeling and monitoring biomedical time series. *J. Am. Stat. Assoc.*, 85, 328-337.
- Gouriéroux, C. (1997). *ARCH Models and Financial Applications*. New York: Springer-Verlag.
- Granger, C.W. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Series Anal.*, 1, 15-29.
- Grenander, U. (1951). On empirical spectral analysis of stochastic processes. *Arkiv for Matematik*, 1, 503-531.
- Grenander, U. and M. Rosenblatt (1957). *Statistical Analysis of Stationary Time Series*. New York: Wiley.
- Grether, D.M. and M. Nerlove (1970). Some properties of optimal seasonal adjustment. *Econometrica*, 38, 682-703.
- Gupta, N.K. and R.K. Mehra (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Trans. Automat. Contr.*, AC-19, 774-783.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357-384.
- Hannan, E.J. (1970). *Multiple Time Series*. New York: Wiley.
- Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *J. Royal Statistical Society, B*, 41, 190-195.
- Hannan, E.J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: Wiley.
- Hansen, J., M. Sato, R. Ruedy, K. Lo, D.W. Lea, and M. Medina-Elizade (2006). Global temperature change. *Proc. Natl. Acad. Sci.*, 103, 14288-14293.
- Harrison, P.J. and C.F. Stevens (1976). Bayesian forecasting (with discussion). *J. R. Stat. Soc. B*, 38, 205-247.
- Harvey, A.C. and P.H.J. Todd (1983). Forecasting economic time series with structural and Box-Jenkins models: A case study. *J. Bus. Econ. Stat.*, 1, 299-307.

- Harvey, A.C. and R.G. Pierse (1984). Estimating missing observations in economic time series. *J. Am. Stat. Assoc.*, 79, 125-131.
- Harvey, A.C. (1991). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
- Harvey, A.C. (1993). *Time Series Models*. Cambridge, MA: MIT Press.
- Harvey A.C., E. Ruiz and N. Shephard (1994). Multivariate stochastic volatility models. *Rev. Economic Studies*, 61, 247-264.
- Haslett, J. and A.E. Raftery (1989) Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource (C/R: 89V38 p21-50) *Applied Statistics*, 38, 1-21
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.
- Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, 68, 165-176.
- Hurst, H. (1951). Long term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116, 778-808.
- Hurvich, C.M. and S. Zeger (1987). Frequency domain bootstrap methods for time series. *Tech. Report 87-115*, Department of Statistics and Operations Research, Stern School of Business, New York University.
- Hurvich, C.M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- Hurvich, C.M. and K.I. Beltrao (1993). Asymptotics for the low-frequency ordinates of the periodogram for a long-memory time series. *J. Time Series Anal.*, 14, 455-472.
- Hurvich, C.M., R.S. Deo and J. Brodsky (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *J. Time Series Anal.*, 19, 19-46.
- Hurvich, C.M. and R.S. Deo (1999). Plug-in selection of the number of frequencies in regression estimates of the memory parameter of a long-memory time series. *J. Time Series Anal.*, 20, 331-341.
- Jacquier, E., N.G. Polson, and P.E. Rossi (1994). Bayesian analysis of stochastic volatility models. *J. Bus. Econ. Stat.*, 12, 371-417.
- Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.
- Jenkins, G.M. and D.G. Watts. (1968). *Spectral Analysis and Its Applications*. San Francisco: Holden-Day.
- Johnson, R.A. and D.W. Wichern (1992). *Applied Multivariate Statistical Analysis*, 3rd ed.. Englewood Cliffs, NJ: Prentice-Hall.
- Jones, P.D. (1994). Hemispheric surface air temperature variations: A reanalysis and an update to 1993. *J. Clim.*, 7, 1794-1802.
- Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389-395.
- Jones, R.H. (1984). Fitting multivariate models to unequally spaced data. In *Time Series Analysis of Irregularly Observed Data*, pp. 158-188. E. Parzen, ed. Lecture Notes in Statistics, 25, New York: Springer-Verlag.

- Jones, R.H. (1993). *Longitudinal Data With Serial Correlation : A State-Space Approach*. London: Chapman and Hall.
- Journel, A.G. and C.H. Huijbregts (1978). *Mining Geostatistics*. New York: Academic Press.
- Juang, B.H. and L.R. Rabiner (1985). Mixture autoregressive hidden Markov models for speech signals, *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-33, 1404-1413.
- Kakizawa, Y., R. H. Shumway, and M. Taniguchi (1998). Discrimination and clustering for multivariate time series. *J. Am. Stat. Assoc.*, 93, 328-340.
- Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Trans ASME J. Basic Eng.*, 82, 35-45.
- Kalman, R.E. and R.S. Bucy (1961). New results in filtering and prediction theory. *Trans. ASME J. Basic Eng.*, 83, 95-108.
- Kaufman, L. and P.J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.
- Kay, S.M. (1988). *Modern Spectral Analysis: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall.
- Kazakos, D. and P. Papantoni-Kazakos (1980). Spectral distance measuring between Gaussian processes. *IEEE Trans. Automat. Contr.*, AC-25, 950-959.
- Khatri, C.G. (1965). Classical statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Stat.*, 36, 115-119.
- Kim S., N. Shephard and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Economic Studies*, 65, p.361-393.
- Kitagawa, G. and W. Gersch (1984). A smoothness priors modeling of time series with trend and seasonality. *J. Am. Stat. Assoc.*, 79, 378-389.
- Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion). *J. Am. Stat. Assoc.*, 82, 1032-1041, (C/R: p1041-1063; C/R: V83 p1231).
- Kitagawa, G. and W. Gersch (1996). *Smoothness Priors Analysis of Time Series*. New York: Springer-Verlag.
- Kolmogorov, A.N. (1941). Interpolation und extrapolation von stationären zufälligen Folgen. *Bull. Acad. Sci. U.R.S.S.*, 5, 3-14.
- Krishnaiah, P.R., J.C. Lee, and T.C. Chang (1976). The distribution of likelihood ratio statistics for tests of certain covariance structures of complex multivariate normal populations. *Biometrika*, 63, 543-549.
- Kullback, S. and R.A. Leibler (1951). On information and sufficiency. *Ann. Math. Stat.*, 22, 79-86.
- Kullback, S. (1958). *Information Theory and Statistics*. Gloucester, MA: Peter Smith.
- Lachenbruch, P.A. and M.R. Mickey (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11.
- Lam, P.S. (1990). The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. *J. Monetary Econ.*, 26, 409-432.

- Lay, T. (1997). Research required to support comprehensive nuclear test ban treaty monitoring. *National Research Council Report, National Academy Press*, 2101 Constitution Ave., Washington, DC 20055.
- Levinson, N. (1947). The Wiener (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25, 262-278.
- Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Stat.*, 5, 81-91.
- Ljung, G.M. and G.E.P. Box (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303.
- Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *J. Time Series Anal.*, 6, 35-52.
- Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis, 2nd ed.* Berlin: Springer-Verlag.
- MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1:281-297
- Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, 15, 661-675.
- McBratney, A.B. and R. Webster (1981). Detection of ridge and furrow pattern by spectral analysis of crop yield. *Int. Stat. Rev.*, 49, 45-52.
- McCulloch, R.E. and R.S. Tsay (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J. Am. Stat. Assoc.*, 88, 968-978.
- McDougall, A. J., D.S. Stoffer and D.E. Tyler (1997). Optimal transformations and the spectral envelope for real-valued time series. *J. Stat. Plan. Infer.*, **57**, 195-214.
- McLeod A.I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *J. R. Stat. Soc. B*, 40, 296-302.
- McLeod, A.I. and K.W. Hipel (1978). Preservation of the rescaled adusted range, I. A reassessment of the Hurst phenomenon. *Water Resour. Res.*, 14, 491-508.
- McQuarrie, A.D.R. and C-L. Tsai (1998). *Regression and Time Series Model Selection*, Singapore: World Scientific.
- Meinhold, R.J. and N.D. Singpurwalla (1983). Understanding the Kalman filter. *Am. Stat.*, 37, 123-127.
- Meinhold, R.J. and N.D. Singpurwalla (1989). Robustification of Kalman filter models. *J. Am. Stat. Assoc.*, 84, 479-486.
- Meng X.L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Am. Stat. Assoc.*, 86, 899-909.
- Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21, 1087-1091.
- Mickens, R.E. (1990). *Difference Equations: Theory and Applicatons (2nd ed)*. New York: Springer.
- Nason, G.P. (2008). *Wavelet Methods in Statistics with R*. New York: Springer.
- Newbold, P. and T. Bos (1985). *Stochastic Parameter Regression Models*. Beverly Hills: Sage.

- Ogawa, S., T.M. Lee, A. Nayak and P. Glynn (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.*, 14, 68-78.
- Palma, W. (2007). *Long-Memory Time Series: Theory and Methods*. New York: Wiley.
- Palma, W. and N.H. Chan (1997). Estimation and forecasting of long-memory time series with missing values. *J. Forecast.*, 16, 395-410.
- Paparoditis, E. and Politis, D.N. (1999). The local bootstrap for periodogram statistics. *J. Time Series Anal.*, 20, 193-222.
- Parzen, E. (1962). On estimation of a probability density and mode. *Ann. Math. Stat.*, 35, 1065-1076.
- Parzen, E. (1983). Autoregressive spectral estimation. In *Time Series in the Frequency Domain, Handbook of Statistics*, Vol. 3, pp. 211-243. D.R. Brillinger and P.R. Krishnaiah eds. Amsterdam: North Holland.
- Pawitan, Y. and R.H. Shumway (1989). Spectral estimation and deconvolution for a linear time series model. *J. Time Series Anal.*, 10, 115-129.
- Peña, D. and I. Guttman (1988). A Bayesian approach to robustifying the Kalman filter. In *Bayesian Analysis of Time Series and Dynamic Linear Models*, pp. 227-254. J.C. Spall, ed. New York: Marcel Dekker.
- Percival, D.B. and A.T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques* Cambridge: Cambridge University Press.
- Percival, D.B. and A.T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.
- Phillips, P.C.B. (1987). Time series regression with a unit root. *Econometrica*, 55, 227-301.
- Phillips, P.C.B. and P. Perron (1988). Testing for unit roots in time series regression. *Biometrika*, 75, 335-346.
- Pinsker, M.S. (1964). *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden Day.
- Pole, P.J. and M. West (1988). Nonnormal and nonlinear dynamic Bayesian modeling. In *Bayesian Analysis of Time Series and Dynamic Linear Models*, pp. 167-198. J.C. Spall, ed. New York: Marcel Dekker.
- Press, W.H., S.A. Teukolsky, W. T. Vetterling, and B.P. Flannery (1993). *Numerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge: Cambridge University Press.
- Priestley, M.B., T. Subba-Rao and H. Tong (1974). Applications of principal components analysis and factor analysis in the identification of multi-variable systems. *IEEE Trans. Automat. Contr.*, AC-19, 730-734.
- Priestley, M.B. and T. Subba-Rao (1975). The estimation of factor scores and Kalman filtering for discrete parameter stationary processes. *Int. J. Contr.*, 21, 971-975.
- Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Vol. 1: Univariate Series; Vol 2: Multivariate Series, Prediction and Control. New York: Academic Press.

- Priestley, M.B. (1988). *Nonlinear and Nonstationary Time Series Analysis*. London: Academic Press.
- Quandt, R.E. (1972). A new approach to estimating switching regressions. *J. Am. Stat. Assoc.*, 67, 306-310.
- Rabiner, L.R. and B.H. Juang (1986). An introduction to hidden Markov models, *IEEE Acoust., Speech, Signal Process.*, ASSP-34, 4-16.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.
- Rauch, H.E., F. Tung, and C.T. Striebel (1965). Maximum likelihood estimation of linear dynamic systems. *J. AIAA*, 3, 1445-1450.
- Reinsel, G.C. (1997). *Elements of Multivariate Time Series Analysis, 2nd ed.* New York: Springer-Verlag.
- Renyi, A. (1961). On measures of entropy and information. In *Proceedings of 4th Berkeley Symp. Math. Stat. and Probability*, pp. 547-561, Berkeley: Univ. of California Press.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-471.
- Robinson, P.M. (1995). Gaussian semiparametric estimation of long range dependence. *Ann. Stat.*, 23, 1630-1661.
- Robinson, P.M. (2003). *Time Series With Long Memory*. Oxford: Oxford University Press.
- Rosenblatt, M. (1956a). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.*, 42, 43-47.
- Rosenblatt, M. (1956b). Remarks on some nonparametric estimates of a density functions. *Ann. Math. Stat.*, 27 642-669.
- Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115-124.
- Said E. and D.A. Dickey (1984). Testing for unit roots in autoregressive moving average models of unknown order. *Biometrika*, 71, 599-607.
- Sandmann, G. and S.J. Koopman (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics*, 87, 271-301.
- Sargan, J.D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. In *Econometric Analysis for National Economic Planning*, eds. P. E. Hart, G. Mills and J. K. Whitaker. London: Butterworths. reprinted in *Quantitative Economics and Econometric Analysis*, pp. 275-314, eds. K. F. Wallis and D. F. Hendry (1984). Oxford: Basil Blackwell.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, III, 11-41.
- Schuster, A. (1906). On the periodicities of sunspots. *Phil. Trans. R. Soc., Ser. A*, 206, 69-100.
- Schwarz, F. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6, 461-464.
- Schweppe, F.C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory*, IT-4, 294-305.

- Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, pp 1-100. D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielsen eds. London: Chapman and Hall.
- Shumway, R.H. and W.C. Dean (1968). Best linear unbiased estimation for multivariate stationary processes. *Technometrics*, 10, 523-534.
- Shumway, R.H. (1970). Applied regression and analysis of variance for stationary time series. *J. Am. Stat. Assoc.*, 65, 1527-1546.
- Shumway, R.H. (1971). On detecting a signal in N stationarily correlated noise series. *Technometrics*, 10, 523-534.
- Shumway, R.H. and A.N. Unger (1974). Linear discriminant functions for stationary time series. *J. Am. Stat. Assoc.*, 69, 948-956.
- Shumway, R.H. (1982). Discriminant analysis for time series. In *Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics Vol. 2*, pp. 1-46. P.R. Krishnaiah and L.N. Kanal, eds. Amsterdam: North Holland.
- Shumway, R.H. and D.S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Anal.*, 3, 253-264.
- Shumway, R.H. (1983). Replicated time series regression: An approach to signal estimation and detection. In *Time Series in the Frequency Domain, Handbook of Statistics Vol. 3*, pp. 383-408. D.R. Brillinger and P.R. Krishnaiah, eds. Amsterdam: North Holland.
- Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- Shumway, R.H., R.S. Azari, and Y. Pawitan (1988). Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environ. Res.*, 45, 224-241.
- Shumway, R.H. and D.S. Stoffer (1991). Dynamic linear models with switching. *J. Am. Stat. Assoc.*, 86, 763-769, (Correction: V87 p. 913).
- Shumway, R.H. and K.L. Verosub (1992). State space modeling of paleoclimatic time series. In *Pro. 5th Int. Meeting Stat. Climatol.*. Toronto, pp. 22-26, June, 1992.
- Shumway, R.H., S.E. Kim and R.R. Blandford (1999). Nonlinear estimation for time series observed on arrays. Chapter 7, S. Ghosh, ed. *Asymptotics, Nonparametrics and Time Series*, pp. 227-258. New York: Marcel Dekker.
- Small, C.G. and D.L. McLeish (1994). *Hilbert Space Methods in Probability and Statistical Inference*. New York: Wiley.
- Smith, A.F.M. and M. West (1983). Monitoring renal transplants: An application of the multiprocess Kalman filter. *Biometrics*, 39, 867-878.
- Spliid, H. (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variables. *J. Am. Stat. Assoc.*, 78, 843-849.
- Stoffer, D.S. (1982). Estimation of Parameters in a Linear Dynamic System with Missing Observations. Ph.D. Dissertation. Univ. California, Davis.
- Stoffer, D.S., M. Scher, G. Richardson, N. Day, and P. Coble (1988). A Walsh-Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *J. Am. Stat. Assoc.*, 83, 954-963.

- Stoffer, D.S. and K.D. Wall (1991). Bootstrapping state space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. Am. Stat. Assoc.*, 86, 1024-1033.
- Stoffer, D.S., D.E. Tyler, and A.J. McDougall (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80, 611-622.
- Stoffer, D.S. (1999). Detecting common signals in multiple time series using the spectral envelope. *J. Am. Stat. Assoc.*, 94, 1341-1356.
- Stoffer, D.S. and K.D. Wall (2004). Resampling in State Space Models. In *State Space and Unobserved Component Models Theory and Applications*, Chapter 9, pp. 227-258. Andrew Harvey, Siem Jan Koopman, and Neil Shephard, eds. Cambridge: Cambridge University Press.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Statist, A, Theory Methods*, 7, 13-26.
- Taniguchi, M., M.L. Puri, and M. Kondo (1994). Nonparametric approach for non-Gaussian vector stationary processes. *J. Mult. Anal.*, 56, 259-283.
- Tanner, M. and W.H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.*, 82, 528-554.
- Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes – A study of daily sugar prices, 1961-79. In Anderson, O. D., editor, *Time Series Analysis: Theory and Practice*, Volume 1, pages 203–226. New York: Elsevier/North-Holland.
- Tiao, G.C. and R.S. Tsay (1989). Model specification in multivariate time series (with discussion). *J. Roy. Statist. Soc. B*, 51, 157-213.
- Tiao, G. C. and R.S. Tsay (1994). Some advances in nonlinear and adaptive modeling in time series analysis. *J. Forecast.*, 13, 109-131.
- Tiao, G.C., R.S. Tsay and T. Wang (1993). Usefulness of linear transformations in multivariate time series analysis. *Empir. Econ.*, 18, 567-593.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, 22, 1701-1728.
- Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*. Springer Lecture Notes in Statistics, 21. New York: Springer-Verlag.
- Tong, H. (1990). *Nonlinear Time Series: A Dynamical System Approach*. Oxford: Oxford Univ. Press.
- Tsay, Ruey S. (2002). *Analysis of Financial Time Series*. New York: Wiley.
- Venables, W.N. and B.D. Ripley (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.
- Watson, G.S. (1966). Smooth regression analysis. *Sankhya*, 26, 359-378.
- Weiss, A.A. (1984). ARMA models with ARCH errors. *J. Time Series Anal.*, 5, 129-143.
- West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models 2nd ed.* New York: Springer-Verlag.
- Whittle, P. (1961). Gaussian estimation in stationary time series. *Bull. Int. Stat. Inst.*, 33, 1-26.

- Wiener, N. (1949). *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley.
- Wu, C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.*, 11, 95-103.
- Young, P.C. and D.J. Pedregal (1998). Macro-economic relativity: Government spending, private investment and unemployment in the USA. Centre for Research on Environmental Systems and Statistics, Lancaster University, U.K.
- Yule, G.U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's Sunspot Numbers. *Phil. Trans. R. Soc. Lond.*, A226, 267-298.

Index

- ACF, 17, 19
 - large sample distribution, 26
 - of an AR(1), 65
 - of an AR(2), 74
 - of an ARMA(1,1), 76
 - of an MA(q), 73
 - sample, 25
- AIC, 37, 109, 155
- AICc, 38, 109
- Aliasing, 126
- Amplitude, 125
- AR model, 11, 64
 - conditional sum of squares, 96
 - conditional likelihood, 96
 - likelihood, 95
 - maximum likelihood estimation, 95
 - spectral density, 134
 - unconditional sum of squares, 95
- ARIMA model, 98
 - multiplicative seasonal models, 117
- ARMA model, 68
 - pure seasonal models
 - behavior of ACF and PACF, 115
 - behavior of ACF and PACF, 78
 - causality, 70
 - conditional least squares, 82
 - forecasts
 - prediction intervals, 91
 - Gauss–Newton, 83
 - invertibility, 70
 - likelihood, 92
 - MLE, 93
 - multiplicative seasonal model, 115
 - pure seasonal model, 112
- Autocorrelation function, *see* ACF
- Autocovariance function, 15, 19, 65
 - random sum of sines and cosines, 127
 - sample, 25
- Autoregressive Integrated Moving Average Model, *see* ARIMA model
- Autoregressive models, *see* AR model
- Backcasting, 92
- Backshift operator, 46
- Bandwidth, 143
- Bartlett kernel, 152
- BIC, 38, 109, 155
- BLP
 - m -step-ahead prediction, 89
 - mean square prediction error, 89
 - one-step-ahead prediction, 87
 - one-step-ahead prediction
 - mean square prediction error, 88
- Causal, 70
 - conditions for an AR(2), 72
- CCF, 18, 21
 - large sample distribution, 28
 - sample, 27
- Cepstral analysis, 170
- Coherence, 163
 - estimation, 165
 - hypothesis test, 165
- Complex roots, 75
- Convolution, 157
- Cospectrum, 162

- Cross-correlation function, *see* CCF
- Cross-covariance function, 18
 - sample, 27
- Cross-spectrum, 162
- Cycle, 125

- Daniell kernel, 149, 150
 - modified, 150
- Density function, 14
- Detrending, 33
- DFT, 129
 - inverse, 136
- Differencing, 44–46
- Durbin–Levinson algorithm, 89

- Earthquake series, 7
- Explosion series, 7
- Exponentially Weighted Moving Averages, 99

- Fejér kernel, 152
- FFT, 129
- Filter, 46
 - high-pass, 159
 - linear, 157
 - low-pass, 159
- Folding frequency, 126, 130
- Fourier frequency, 129, 136
- Frequency bands, 133, 142
- Frequency response function, 158
 - of a first difference filter, 158
 - of a moving average filter, 158
- Functional magnetic resonance imaging series, 6
- Fundamental frequency, 129, 136

- Glacial varve series, 47, 84, 107
- Global temperature series, 3, 43, 46
- Growth rate, 102

- Harmonics, 147

- Impulse response function, 157
- Innovations, 105
 - standardized, 105
- Integrated models, 97, 99, 117
 - forecasting, 99
- Invertible, 70

- Johnson & Johnson quarterly earnings series, 2
- Joint distribution function, 14

- LA Pollution – Mortality Study, 38, 56, 59, 111
- Lag, 17, 22
- Lead, 22
- Leakage, 153
 - sidelobe, 153
- Likelihood
 - AR(1) model, 95
 - conditional, 96
- Linear filter, *see* Filter
- Linear process, 23
- Ljung–Box–Pierce statistic, 106
- LSE, 35
 - conditional sum of squares, 96
 - Gauss–Newton, 82
 - unconditional, 95

- MA model, 9, 66
 - autocovariance function, 16, 73
 - Gauss–Newton, 83
 - mean function, 15
 - spectral density, 134
- Mean function, 14
- Method of moments estimators, *see* Yule–Walker
- MLE
 - ARMA model, 93
 - conditional likelihood, 96
 - via Newton–Raphson, 93
 - via scoring, 93

- New York Stock Exchange, 4
- Newton–Raphson, 93
- Normal distribution
 - marginal density, 14
 - multivariate, 24
- NYSE, *see* New York Stock Exchange

- PACF, 77
 - of an MA(1), 79
 - iterative solution, 89
 - large sample results, 79
 - of an AR(p), 77
 - of an MA(q), 79
- Parameter redundancy, 69

- Partial autocorrelation function, *see*
PACF
- Period, 125
- Periodogram, 129, 136
 - distribution, 138
- Phase, 125
- Pitch period, 4
- Quadspectrum, 162
- Random sum of sines and cosines, 126
- Random walk, 12, 15, 99
 - autocovariance function, 17
- Recruitment series, 5, 28, 48, 79, 91, 139, 144, 150, 166
- Regression
 - ANOVA table, 36
 - autocorrelated errors, 109
 - Cochrane-Orcutt procedure, 111
 - model, 33
 - multiple correlation, 37
 - normal equations, 35
 - periodic, 53
 - polynomial, 53
- Return, 4, 102
- Scatterplot matrix, 40, 48
- Scatterplot smoothers, 53
 - kernel, 55
 - lowess, 56, 59
 - nearest neighbors, 56
 - splines, 57, 59
- SIC, 38
- Signal plus noise, 13
 - mean function, 15
- Signal-to-noise ratio, 14
- Smoothing splines, 57
- Southern Oscillation Index, 5, 28, 48, 139, 144, 150, 153, 155, 158, 166
- Spectral density, 132
 - autoregression, 155
 - estimation, 142
 - adjusted degrees of freedom, 144
 - bandwidth stability, 148
 - confidence interval, 143
 - degrees of freedom, 143
 - large sample distribution, 142
 - nonparametric, 154
 - parametric, 154
 - resolution, 148
 - matrix, 164
 - of a filtered series, 158
 - of a moving average, 134
 - of an AR(2), 134
 - of white noise, 133
- Spectral Representation Theorem, 132
 - vector process, 164
- Speech series, 4, 27
- Stationary
 - Gaussian series, 24
 - jointly, 21
 - strictly, 18
 - weakly, 19
- Stochastic process, 9
 - realization, 9
- Stochastic trend, 97
- Taper, 152, 153
 - cosine bell, 152
- Time series, 9
 - multivariate, 18
- Transformation
 - Box-Cox, 47
- Trend stationarity, 20
- U.S. GNP series, 102, 106, 109
- U.S. population series, 108
- Volatility, 4
- White noise, 9
 - autocovariance function, 16
 - Gaussian, 9
- Yule-Walker
 - equations, 80
 - estimators, 81
 - AR(2), 81
 - MA(1), 81