TIME SERIES

© Richard L. Smith

Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Email address: rls@email.unc.edu

VERSION 1.0

11 MAY 1999

PREFACE

These notes have been prepared in conjunction with the course Statistics 133, which I taught in Spring 1999. They cover some of the standard theory and applications of time series analysis, beginning with basic theory about stationary processes, through the theory and applications of ARIMA models and spectral analysis, and leading up to an overview of some of the modern applications of state space modeling in areas such as financial time series. The notes draw on a number of standard references in the field, in particular Brockwell and Davis (1990), which was used as the official text for the course. I have also tried to emphasize computational applications in SPlus. Earlier versions of the notes have been used in courses at Imperial College, Surrey University and Cambridge University.

Beginning with the 2000-2001 academic year, Statistics 133 is to become Statistics 185 with the new name "Time Series and Multivariate Analysis". A separate set of course notes is available for the "Multivariate Analysis" section of Statistics 133.

Richard Smith
Chapel Hill
May 1999

TABLE OF CONTENTS

1. INTRODUCTION TO TIME SERIES ANALYSIS

1.1 Introduction

Time series analysis refers to the branch of statistics where observations are collected sequentially in time, usually but not necessarily at equally-spaced time points, and the analysis relies at least in part on understanding or exploiting the dependence among the observations. The areas of application these days cover almost any area where statistics is applied, but some of the main ones are

• *Economics:* Economic indicators such as unemployment statistics or the retail prices index, and financial series such as currency exchange rates or stock prices, are all examples of time series.

• *Engineering:* Many areas of engineering and signal processing involve series which are sampled at frequent time points and therefore form a time series.

• *Environmental Statistics:* A growing area of application in view of widespread of interest in such topics as global climate change, which involves interpreting time series of temprature or rainfall, and such areas as pollution in the atmosphere, which involve time series of air quality measurments,

• *Medical Statistics:* This is not generally regarded as one of the major areas of time series application, but nevertheless it is common for some adjustment for serial correlation to be needed in the context of medical statistics. Often this arises in the context of many short time series, e.g. data are independent from one patient to the next, but there is a series of measurements on each patient, and these are correlated in time. Such problems are also referred to as *longitudinal data analysis*. This sort of problem, however, involves somewhat different considerations from the ones that will mostly be addressed in these notes.

When we have a time series, what sorts of questions do we want to answer about it? Some of the main ones are:

• *Analysis:* Find a model to describe the time dependence in the data. This is usually a first step to any of the other parts of the subject, but sometimes it is the only step. For example, we might want to know whether the pattern on unemployment statistics in the U.S.A. is different from that in, say, Japan. One could try to answer that question by fitting a model to each series, and then contrasting the two models.

•*Forecasting:* Given a finite sample from the series, forecast the next value or the next several values. Obviously, many applications are of this form.

•*Control:* Of important in many engineering and industrial applications, after making a forecast of the series we might consider how to adjust various control parameters to make the series fit closer to a target.

•*Adjustment:* This refers to a somewhat different kind of problem, where the time series analysis is not the main purpose of the analysis, but where time series correlations may have a significant influence on the analysis being performed. The simplest case arises with the usual linear model,

$$y_t = x_t^T \beta + \eta_t, \tag{1.1}$$

in which y_t is a scalar observation at time t, x_t is a vector of observed covariates, β a vector of unknown parameters, and η_t a random disturbance or errors. The usual linear model assumes the $\{\eta_t\}$ are uncorrelated with mean 0 and common variance σ^2, in which case the least squares estimator $\hat{\beta}$ has variance $(X^T X)^{-1}\sigma^2$, X being the matrix formed from all the regressors $\{x_t\}$. Suppose, however, the errors $\{\eta_t\}$ form a time series of correlated observations. If we knew the correlations, we might apply generalized least squares (GLS) instead of the usual ordinary least squares (OLS). If we do not know the correlations, however, we may decide to do a time series analysis to find something out about them. Typically, the OLS $\hat{\beta}$ behaves well as a point estimator even when there is substantial correlation in the data, but the variance-covariance matrix is quite different from $(X^T X)^{-1}\sigma^2$. Therefore we may need to adjust the estimated variances to allow for serial correlation.

1.2 Techniques of Time Series Analysis

From now on, we shall assume we are dealing with time series sampled at equally spaced time points, where we may assume that the sampling interval is 1 time unit. An important concept is that of a *stationary process*, which will be formally defined in Chapter 2. Loosely, it refers to a series in which all trends and other non-random effects have been removed. Suppose we have a stationary series $\{X_t,\ t = 0, \pm 1, \pm 2, ...\}$. One of the most common models is

$$X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \epsilon_t, \tag{1.2}$$

in which $\phi_1, ..., \phi_p$ are fixed coefficients and $\{\epsilon_t\}$ are independent (or uncorrelated) random disturbances of mean 0 and constant variance σ^2. This is known as the *autoregressive process of order* p, AR(p) for short.

A second example of a stationary time series model is

$$X_t = \sum_{s=0}^{q} \theta_s \epsilon_{t-s}, \tag{1.3}$$

the *moving average process of order* q, or MA(q). We usually assume $\theta_0 = 1$ for identifiability.

It is also possible to combine the autoregressive and moving average structures with a model of form

$$X_t - \sum_{r=1}^{p} \phi_r X_{t-r} = \sum_{s=0}^{q} \theta_s \epsilon_{t-s}, \tag{1.4}$$

usually denoted ARMA (p, q).

In cases where the series is not initially stationary, one common device is to difference the series, i.e. if the original series is $\{Y_t\}$, form first-order differences

$$X_t = \nabla Y_t = Y_t - Y_{t-1}.$$

If this series appears to be stationary then we look for an AR, MA or ARMA representation for $\{X_t\}$. We can iterate this process: the second-order differences are

$$X_t = \nabla^2 Y_t = \nabla(\nabla Y_t) = Y_t - 2Y_{t-1} + Y_{t-2}$$

and so on: in general we can define a d'th order differencing operator ∇^d for any positive integer d. If

$$X_t = \nabla^d Y_t$$

is a stationary processes fitting the ARMA(p, q) model, then the process for $\{Y_t\}$ is called an integrated autoregressive moving average process, or ARIMA(p, d, q).

This scheme represents one of the main avenues towards time series analysis. However, there are at least two quite different approaches:

Spectral Analysis. This refers to methods based on representing a stationary series as a sum of sinusoidal terms via a discrete Fourier transform. Originally such methods were mainly applied in engineering, where Fourier transform methods of signal processing have long been widely accepted, but they are now much more widely used in standard statistical analyses as well. Part of the reason for this has to do with the sampling properties of spectral estimates, which facilitate general nonparametric methods of reconstructing the correlations of the process.

The State Space Approach. The simplest state space model is

$$\begin{aligned} X_t &= S_t + \zeta_t \\ S_t &= \alpha S_{t-1} + \epsilon_t \end{aligned} \tag{1.5}$$

in which $\{S_t\}$ represents an unobserved sequence of *states* of the system, which here follow an AR(1) process, and $\{X_t\}$ represent the observations which are perturbed by a further sequence of (independent or uncorrelated) random variables $\{\zeta_t\}$. Models of this form also arose originally in engineering applications, for which $\{S_t\}$ usually represents some real physical quantity, but similar models are by now used as abstract models in a variety of statistical applications. In the case of (5), it can be shown fairly easily that this is equivalent to an ARMA(1,1) model for $\{X_t\}$, so in this case nothing appears to be gained by taking a state space approach, but more complicated models based on the state space representation go well beyond what is possible within the ARMA framework. The *Kalman Filter* is a widely used computational tool for analysing such models.

The above methods – ARMA models for stationary processes, spectral analysis, and state space models – represent the three main methods in current practice for the analysis of linear time series models. However, there are many alternative approaches to time series analysis which are gradually gaining acceptance. Some examples are

• *Nonlinear models.* For example, the AR model (1.2) may be generalized to

$$X_t = f(X_{t-r}, X_{t-r+1}, ..., X_{t-1}) + \epsilon_t \tag{1.6}$$

for a general nonlinear function f. The fitting of models of this nature is one of the most complex issues in contemporary research in time series analysis. Amongst the applications are a link with nonlinear dynamics and chaos theory: the fitting of a model of the form of (1.6) may be one of the main steps in diagnosing chaotic behaviour in an observational system.

• *Irregularly spaced data.* This is still a relatively unexplored area of time series analysis, but such problems certainly arise in applications, and methods are being devised to deal with them. Such series may, for example, be analyzed by viewing the observed series as a discrete sample from some continuous-time stochastic process, constructing estimates and predictors as appropriate.

• *Generalized Linear Models with dependent errors.* Although models such as the ARMA models may not necessarily assume the data are normally distributed, it is clear that they would not apply very well to, for example, data from a binomial or Poisson distribution. Therefore, there is interest in developing time series models in this kind of setting. The GLM structure is a natural starting point within which to think about such issues.

• *Bayesian analysis.* Most current time series analysis is still based on the classical (frequentist) approach to statistics. However, Bayesian methods of time series analysis have gained much ground in recent years, see for example the book by West and Harrison (1997). The state space approach to time series analysis is especially suitable for Bayesian development and our own presentation of this approach will draw heavily on Bayesian thinking.

• *Long-range dependence.* Sometimes series are stationary but with very slowly decaying correlations; this happens for instance in certain physical processes, and is widely believed to be true in some econometric time series. There is a connection here with the use of fractals as models for time series. One way of modelling such series is based on ARIMA processes with a fractionally differencing coefficient d; it is possible to define such a thing in a way that makes sense! Other methods include a spectral approach.

2. STATIONARY STOCHASTIC PROCESSES

2.1 Definitions

Assume that we have a process $\{X_t,\ t = 0, \pm 1, \pm 2, ..., \}$. There are two widely used definitions of stationarity.

Strict or *Strong Stationarity* is said to hold if, for any positive integer k, and k time points $t_1, ..., t_k$ and any integer lag h, we have that the vectors

$$(X_{t_1}, X_{t_2}, ..., X_{t_k})$$

and

$$(X_{t_1+h}, X_{t_2+h}, ..., X_{t_k+h})$$

have the same joint distribution.

Weak or *Wide-sense* or *Second-order Stationarity* is said to hold if all the variances of the process are finite and we have

$$\begin{aligned} \mathrm{E}(X_t) &= \mu, \\ \mathrm{Cov}(X_t, X_{t+h}) &= \gamma_h, \end{aligned} \tag{2.1}$$

the quantities μ and γ_h being independent of t.

Provided the variances are finite, it is clear that a strictly stationary process must also be second-order stationary. However, the converse is not necessarily true. It is easy to construct examples of random variables which agree in the first- and second-order moment but not to any higher order.

However, in the case of a *Gaussian* process, the two concepts are equivalent. A process $\{X_t,\ t = 0, \pm 1, \pm 2, ...\}$ is said to be Gaussian if, for any k time points $t_1, ..., t_k$, the joint distribution of $X_{t_1}, ..., X_{t_k}$ has a multivariate normal distribution. This distribution is, of course, completely determined once its mean vector and covariance matrix have been specified. Therefore it follows at once that, if a Gaussian process satisfies the second-order stationarity condition, it must also be strictly stationary.

Another useful definition is that of a *linear process*. A process $\{X_t,\ t = 0, \pm 1, \pm 2, ..., \}$ is said to be linear if it has a representation of the form

$$X_t = \mu + \sum_{r=-\infty}^{\infty} c_r \epsilon_{t-r} \tag{2.2}$$

where μ is a common mean, $\{c_r\}$ is a sequence of fixed constants and $\{\epsilon_t\}$ are independent random variables with mean 0 and common variance. We assume $\sum c_r^2 < \infty$ to ensure that

the variances of individual X_t are finite. Such a process is necessarily strictly stationary; if $c_r = 0$ for all $r < 0$ it is said to be *causal* (i.e. in this case the process at time t does not depend on on future, as yet unobserved, values of ϵ_t). A Gaussian process necessarily has a representation as a linear process with normal $\{\epsilon_t\}$, but we may also want to consider non-Gaussian linear processes. The AR, MA and ARMA classes are all special cases of causal linear processes.

2.2 Autocovariances, autocorrelations and spectral representations

For a weakly stationary process of mean 0, the *autocovariance function* is given by

$$\gamma_k = \mathrm{E}\{X_t X_{t+k}\}.$$

It follows from the definition of weak stationarity that this does not depend on t. Also, note that $\gamma_{-k} = \gamma_k$ for all k.

The *autocorrelation function* is

$$\rho_k = \frac{\gamma_k}{\gamma_0}, \ k = 0, \pm 1, \pm 2, ...$$

For any sequence of autocovariances $\{\gamma_k\}$ generated by a stationary process, there exists a function F such that

$$\gamma_k = \int_{(-\pi,\pi]} e^{ik\lambda} dF(\lambda) \tag{2.3}$$

where F is the unique function on $[-\pi, \pi]$ satisfying

(i) $F(-\pi) = 0$,

(ii) F is non-decreasing and right-continuous,

(iii) F has increments symmetric about 0, meaning that for any $0 \leq a < b \leq \pi$ we have

$$F(b) - F(a) = F(-a) - F(-b).$$

Then F is called the *spectral distribution function*, so called because it has many of the properties of a probability distribution function except for $F(\pi) = 1$. Note that the integral (2.3) is a Stieltjes integral reflecting the fact that F may have discontinuities.

However, if F is everywhere continuous and differentiable, with $f(\lambda) = dF(\lambda)/d\lambda$, then f is called the *spectral density function* and (2.3) may be simplified to

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda. \tag{2.4}$$

If $\sum |\gamma_k| < \infty$, then it can be shown that f always eists and is given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{i\lambda k} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \gamma_k \cos(\lambda k). \tag{2.5}$$

The interpretation of F is that, for any $0 \leq \lambda_1 < \lambda_2 \leq \pi$, $F(\lambda_2) - F(\lambda_1)$ measures the contribution to the total variability of the process within the frequency range $\lambda_1 < \lambda \leq \lambda_2$.

Examples

1. White noise: suppose $\gamma_0 = \sigma^2 > 0$ but $\gamma_k = 0$ for all $k \neq 0$.

In this case it is immediately seen that

$$f(\lambda) = \frac{\sigma^2}{2\pi} \quad \text{for all } \lambda,$$

which is independent of λ. The converse also holds, i.e. a process is white noise if and only if its spectral density is constant.

2. Consider the process

$$X_t = \cos(\omega t + U)$$

where U is a random variable uniformly distributed on $(-\pi, \pi)$ and $0 \leq \omega \leq \pi$. We can easily calculate

$$\begin{aligned}
\mathrm{E}(X_t) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega t + u) du = 0, \\
\mathrm{E}(X_t X_{t+k}) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega t + u) \cos(\omega t + \omega k + u) du \\
&= \frac{1}{4\pi} \int_{-\pi}^{\pi} \{\cos(\omega k) + \cos(2\omega t + \omega k + 2u)\} du \\
&= \frac{\cos(\omega k)}{2}.
\end{aligned}$$

Thus we see that $\{X_t\}$ is a stationary process. To find the spectral representation, we want to represent the autocovariance in the form

$$\gamma_k = \frac{\cos(\omega k)}{2} = \int_{(-\pi, \pi]} \cos(\lambda k) dF(\lambda).$$

A suitable F is one that takes jumps of $\frac{1}{4}$ at $\pm\omega$, i.e.

$$F(\lambda) = \begin{cases} 0 & \text{if } -\pi \leq \lambda < -\omega, \\ 1/4 & \text{if } -\omega \leq \lambda < \omega, \\ 1/2 & \text{if } \omega \leq \lambda < \pi. \end{cases}$$

By the uniqueness of F, this is *the* spectral distribution function in this case. Thus, a spectral distribution function which has discontinuities at $\pm\omega$, and is elsewhere flat, corresponds to a single sinusoid which is perfectly predictable once one observation in the series is known. Note that $\sum |\gamma_k| = \infty$ in this case.

An obvious extension of this is to the case where F is flat except for $2k$ discontinuities at $\pm\omega_1, \pm\omega_2, ..., \pm\omega_k$. This corresponds to a process of the form

$$X_t = \sum_{j=1}^{k} a_j \cos(\omega_j t + U_j) \tag{2.6}$$

for constants $a_1, ..., a_k$, in which $U_1, ..., U_k$ are independent random variables each uniformly distributed on $(-\pi, \pi)$.

The restriction to $0 \leq \omega \leq \pi$ in this example is in fact no restriction at all, for the following reason. Suppose we have a process $X_t = \cos(\Omega t + U)$ for general Ω. Then we may write $\Omega = N\pi + \omega$ for some integer N, $\omega \in [0, \pi)$. If N is even then $\cos(\Omega t + U) = \cos(\omega t + U)$ for any integer t, so the frequencies ω and Ω are indistinguishable, or to use the conventional terminology in this situation, they are *aliases*. If N is odd, then $\cos(\Omega t + U) = \cos(\omega t - \pi t + U) = \cos(\pi t - \omega t - U)$. But U has the same distribution as $-U$ so in this case the frequency Ω is aliased to $\pi - \omega$. Thus any frequency Ω is aliased to some frequency in the interval $[0, \pi]$. The upper bound π corresponds to a cycle of period 2 and is the highest frequency detectable with sampling at integral time periods; this is sometimes called *the Nyquist frequency*. The same comment obviously applies to (2.6) in which any k frequencies $\Omega_1, ..., \Omega_k$ may be aliased to frequencies $\omega_1, ..., \omega_k$ in the interval $[0, \pi]$.

3. The AR(1) process

$$X_t = \phi_1 X_{t-1} + \epsilon_t, \tag{2.7}$$

in which $\{\epsilon_t\}$ is an uncorrelated sequence of random variables with mean 0 and common mean σ_ϵ^2, satisfies the relation

$$\mathrm{Var}\{X_t\} = \phi_1^2 \mathrm{Var}\{X_{t-1}\} + \sigma_\epsilon^2$$

so under stationarity, in which $\mathrm{Var}\{X_t\} = \gamma_0 = \sigma_X^2$ independently of t, we have

$$\sigma_X^2 = \frac{1}{1 - \phi_1^2} \sigma_\epsilon^2. \tag{2.8}$$

Note that for (2.8) to make sense we require $|\phi_1| < 1$. This is the *stationarity condition* for an AR(1) process: without this condition the process tends to grow forever and so does not have a stationary distribution. (If $\phi_1 = \pm 1$ then the process is a random walk, which is a recurrent process but does not have a stationary distribution.) We shall see later that all AR processes require some condition of this nature.

Now for the model (2.7) satisfying the stationarity condition $|\phi_1| < 1$, and for $k > 0$, we have

$$\begin{aligned}\gamma_k &= \mathrm{E}\{X_t X_{t-k}\} \\ &= \phi_1 \mathrm{E}\{X_{t-1} X_{t-k}\} + \mathrm{E}\{\epsilon_t X_{t-k}\} \\ &= \phi_1 \gamma_{k-1}.\end{aligned}$$

Since we also have $\gamma_{-k} = \gamma_k$ we deduce

$$\gamma_k = \phi_1^{|k|} \gamma_0, \quad -\infty < k < \infty.$$

Direct application of (2.5) then leads to

$$f(\lambda) = \frac{\gamma_0(1-\phi_1^2)}{\pi(1 - 2\phi_1 \cos\lambda + \phi_1^2)} = \frac{\sigma_\epsilon^2}{\pi(1 - 2\phi_1 \cos\lambda + \phi_1^2)}. \tag{2.9}$$

Fig 2.1: Plot of AR(1) spectral density

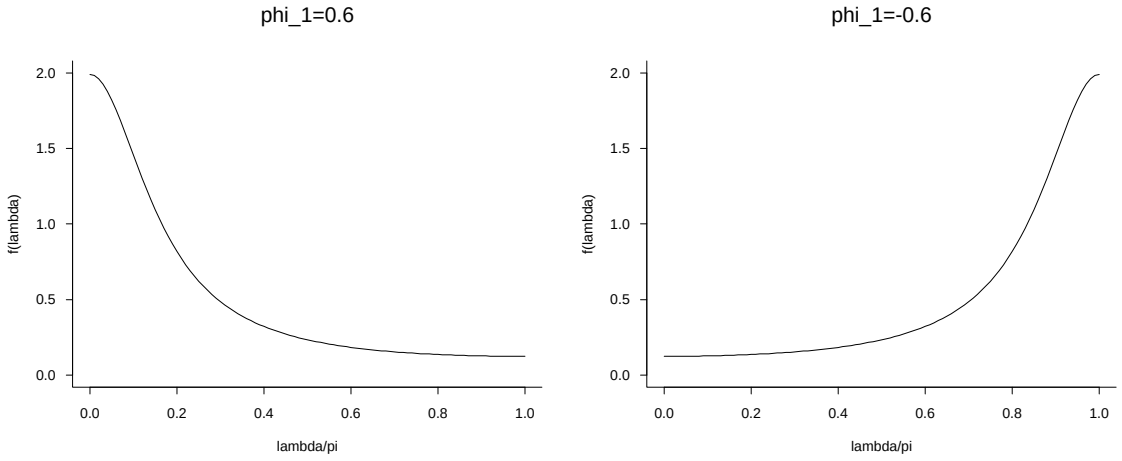


Fig 2.1 shows plots of $f(\lambda)$ for $\phi_1 = \pm 0.6$. In the case $\phi_1 > 0$, the power is concentrated at low frequencies, i.e. corresponding to gradual long-range fluctuations. For $\phi_1 < 0$ the power is concentrated at high frequencies, which reflects the fact that such a process tends to oscillate.

2.3 The Wold decomposition

In general it is possible for write the spectral distribution function F in the form

$$F = F_1 + F_2 \tag{2.10}$$

where F_1 is absolutely continuous and F_2 is a purely discontinuous spectral distribution function. Corresponding to this is a decomposition of the process

$$X_t = U_t + V_t \tag{2.11}$$

into uncorrelated processes U and V in which U has spectral d.f. F_1 and V has spectral d.f. F_2.

As we saw in (2.6), a purely discontinuous spectral distribution function with finitely many jumps corresponds to a mixture of sinusoids, which is a purely deterministic and predictable process. The general result is given by the following:

Theorem. Suppose $F = F_1 + F_2$ as in (2.10) and suppose

$$\int_{-\pi}^{\pi} \log F_1'(\lambda) d\lambda > -\infty. \tag{2.12}$$

Then the decomposition of X into uncorrelated processes U and V as in (2.11) exists, and moreover we have

(i) $U_t = \sum_{r=0}^{\infty} c_r \epsilon_{t-r}$ with $\{\epsilon_r\}$ uncorrelated random variables of mean 0 and common variance; without loss of generality we may take $c_0 = 1$, and we also require $\sum c_r^2 < \infty$,

(ii) V is a deterministic process, i.e. if we know V_s for all $s < t$ then we can predict V_t perfectly.

The sum in (i) is defined in the mean squared sense, i.e.

$$\lim_{R \to \infty} \mathrm{E} \left\{ \left(U_t - \sum_{r=0}^{R} c_r \epsilon_{t-r} \right)^2 \right\} = 0.$$

2.4 Non-Negative Definiteness

One natural question to ask is: given an arbitrary sequence $\{\gamma_k,\ k \geq 0\}$, under what conditions is it the autocovariance function of some stationary process?

It is easily seen that there must be some restrictions, because for any finite sequence of constants $\{c_1, ..., c_T\}$ we have

$$\begin{aligned} 0 &\leq \mathrm{Var} \left(\sum_{t=1}^{T} c_t X_t \right) \\ &= \sum_{s=1}^{T} \sum_{t=1}^{T} c_s c_t \ \mathrm{Cov}(X_s, X_t) \\ &= \sum_{s=1}^{T} \sum_{t=1}^{T} c_s c_t \gamma_{|t-s|}. \end{aligned} \tag{2.13}$$

If (2.13) holds for all sequences $\{c_1, ..., c_T\}$ then we say that $\{\gamma_k\}$ is *non-negative definite*. If the last expression is strictly positive except when $c_1 = ... = c_T = 0$, then it is *positive definite*. It turns out that non-negative definiteness is a necessary and sufficient condition for $\{\gamma_k\}$ to be the autocovariance function of some stationary process. (This result is known as Bochner's Theorem.)

How can we check non-negative definiteness of a given $\{\gamma_k\}$ sequence? A sufficient (and in fact necessary) condition is that the spectral density function defined by (2.5) should be non-negative for all λ. For under this condition we have

$$\begin{aligned}
\sum_s \sum_t c_s c_t \gamma_{|s-t|} &= \sum_s \sum_t c_s c_t \int_{-\pi}^{\pi} e^{i(s-t)\lambda} f(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} \left\{ \sum_s \sum_t c_s c_t e^{i(s-t)\lambda} \right\} f(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} \left| \sum_t c_t e^{it\lambda} \right|^2 f(\lambda) d\lambda \\
&\geq 0.
\end{aligned}$$

Thus, a very general method of constructing autocovariance functions is to take an arbitrary non-negative f and transform it via (2.4).

2.5 Estimating autocovariances and spectral densities

So far, we have discussed autocovariances and spectral densities only as theoretical constructs and have not given any indication about their estimation from data. We do that here, and also discuss a related concept, the *partial autocorrelation function*.

Suppose we have data $\{X_1, ..., X_T\}$. The usual estimate of γ_k for $k > 0$ is given by

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X})(X_{t+k} - \bar{X}) \tag{2.14}$$

where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t.$$

Corresponding to this, we have estimates of the autocorrelations $\rho_k = \gamma_k / \gamma_0$, given by

$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}. \tag{2.15}$$

In (2.14), it might seem more natural to divide by $T-k$ (rather than T) because the sum is taken over $t-k$ terms, but this is not usually done, for two reasons: (a) using the definition (2.14) ensures that the sample autocovariances are non-negative definite, which is evidently a desirable property for them to have, (b) the estimate of r_k in (2.15) often has smaller mean squared error if defined in this way, than it would in the alternative way with the divisor T in (2.14) replaced by $T-k$.

One reason for calculating and plotting the autocovariances or autocorrelations is the following easily verified fact: if $\{X_t\}$ is MA(q), then $\gamma_k = 0$ for $|k| > q$ and so plots of $\{\hat{\gamma}_k\}$ should show a sharp drop to near 0 after the q'th coefficient. This is therefore a diagnostic for an MA(q) process. The corresponding diagnostic for an AP(p) process is based on a different quantity, known as the *partial autocorrelation* function.

The partial autocorrelation of lag k is based on the least-squares regression of X_t on $X_{t-k}, ..., X_{t-1}$. Formally, this is based on postulating the model

$$X_t = \sum_{j=1}^{p} a_{j,k} X_{t-j} + \epsilon_t, \quad t > k,$$

with ϵ_t independent of $X_{t-k}, ..., X_{t-1}$. Least squares estimates of $\{a_{j,k},\ j = 1, ..., k\}$ are obtained by minimization of

$$\sigma_k^2 = \frac{1}{T} \sum_{t=k+1}^{T} \left(X_t - \sum_{j=1}^{p} a_{j,k} X_{t-j} \right)^2 . \tag{2.16}$$

which is (almost) equivalent to solving the equations

$$\hat{\gamma}_\ell = \sum_{j=1}^{k} a_{j,k} \hat{\gamma}_{|j-\ell|}, \quad 1 \le \ell \le k, \tag{2.17}$$

and then calculating the mean sum of squares by substituting in (2.16). In practice, these coefficients may be calculated recursively in k from the *Levinson-Durbin recursion*:

$$\begin{aligned} a_{k,k} &= \frac{\hat{\gamma}_k - \sum_{j=1}^{k-1} a_{j,k-1} \hat{\gamma}_{j-k}}{\sigma_{k-1}^2}, \\ a_{j,k} &= a_{j,k-1} - a_{k,k} a_{k-j,k-1}, \quad 1 \le j \le k-1, \\ \sigma_k^2 &= \sigma_{k-1}^2 (1 - a_{k,k}^2). \end{aligned} \tag{2.18}$$

An obvious measure of how much the k'th order regression improves on that of order $k-1$ is the drop in mean squared residual error, and (2.18) shows that this is determined by the coefficient $a_{k,k}$. We therefore call $a_{k,k}$ the k'th order sample *partial autocorrelation coefficient*. The corresponding population autocorrelation coefficient is, of course, obtained from the same sequence of equations but with γ_k replacing $\hat{\gamma}_k$.

In the case of a Gaussian process, there is an alternative interpretation in that $a_{k,k}$ is the conditional correlation of X_t and X_{t-k} given the intermediate values $X_{t-k+1}, ..., X_{t-1}$. For a non-Gaussian process, this interpretation is no longer valid because conditional expectations are no longer, in general, given by simple linear combinations of the variables being conditioned on.

For both Gaussian and non-Gaussian processes, however, the most important property is the following: if the true process is AR(p), then the population partial autocorrelations of order $k > p$ are all 0, and therefore we would expect the sample partial autocorrelations to drop off sharply after lag p.

For spectral densities, the simplest estimate is given by the *periodogram*

$$I_T(\lambda) = \frac{1}{2\pi T} \left| \sum_{t-1}^{T} X_t e^{i\lambda t} \right|^2 . \tag{2.19}$$

It will be seen in Chapter 5 that the periodogram for fixed λ is an almost unbiased estimator of $f(\lambda)$, provided the underlying process is stationary and its spectral density exists, but that the sample periodogram is too rough to be a good estimator for most practical purposes. Various operations on the periodogram, in particular *smoothing* and *tapering*, will be introduced there, to improve on the raw periodogram as a spectral density estimator.

3. LINEAR FILTERS

3.1 Introduction

Suppose there are two processes X and Y related by

$$Y_t = \sum_{r=-\infty}^{\infty} c_r X_{t-r}, \quad -\infty < t < \infty \quad \text{where} \quad \sum_{r=-\infty}^{\infty} c_r^2 < \infty, \tag{3.1}$$

and suppose their spectral densities are $f_X(\lambda)$ and $f_Y(\lambda)$.

We have

$$\begin{aligned}
\text{Cov}(Y_t, Y_{t+k}) &= \text{Cov}\left(\sum_{r=-\infty}^{\infty} c_r X_{t-r}, \sum_{s=-\infty}^{\infty} c_s X_{t+k-s} \right) \\
&= \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} c_r c_s \gamma_{k+r-s} \\
&= \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} c_r c_s \int_{-\pi}^{\pi} e^{i(k+r-s)\lambda} f_X(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} e^{ik\lambda} \left| \sum c_r e^{ir\lambda} \right|^2 f_X(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi} e^{ik\lambda} f_Y(\lambda) d\lambda.
\end{aligned}$$

Comparison of the last two equations shows that

$$f_Y(\lambda) = \left| C(e^{i\lambda}) \right|^2 f_X(\lambda) \tag{3.2}$$

where $C(z) = \sum c_r z^r$ is the generating function of the filter (we assume this is convergent on $|z| \leq 1$).

Equation (3.2) is the main result of this section, and is an important result in its own right, because it allows us to calculate the effect of applying any linear filter to a given process $\{X_t\}$. For the development which follows, however, we shall mainly be concerned with using this formula to understand better the properties of ARMA processes.

3.2 Application to AR processes

Define a backshift operator B by

$$BX_t = X_{t-1}, \quad B^2 X_t = B(BX_t) = BX_{t-1} = X_{t-2}, \quad \ldots$$

including the identity $IX_T = B^0 X_t = X_t$. Using this notation, we may formally write an AR(p) process as

$$\left(I - \sum_{r=1}^{p} \phi_r B^r \right) X_t = \epsilon_t$$

or in even more compact notation as

$$\phi(B) X = \epsilon$$

where $\phi(z)$ is the generating function $1 - \sum \phi_r z^r$.

Applying (3.2) leads to the formula for the spectral density

$$\left| \phi(e^{i\lambda}) \right|^2 f_X(\lambda) = f_\epsilon(\lambda) = \frac{\sigma_\epsilon^2}{2\pi}$$

and hence

$$f_X(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \cdot \frac{1}{|\phi(e^{i\lambda})|^2}. \tag{3.3}$$

In principle we can now get all the covariances of X by Taylor expanding (3.3) in powers of $e^{i\lambda}$ and using (2.5), but this requires one small assumption: that it is legitimate to Taylor expand the function $1/\phi(e^{i\lambda})$, or in other words that the radius of convergence of $1/\phi(z)$, as a function of the complex variable z, is greater than 1. Since $\phi(z)$ is a polynomial in z, this requires the following statement:

(*) All the zeros of the function $\phi(z)$ lie outside the unit circle in the complex plane.

To see this, note that if the p complex zeros are at $z_1, ..., z_p$, then we can write

$$\phi(z) = \prod_{j=1}^{p} (z - z_j) = \prod_{j=1}^{p} \left\{ -z_j \left(1 - \frac{z}{z_j} \right) \right\}$$

and we can expand

$$\left(1 - \frac{z}{z_j} \right)^{-1} = \sum_{s=0}^{\infty} \left(\frac{z}{z_j} \right)^s$$

if and only if $|z_j| > 1$.

The relation (*) is called the *stationarity condition* for an AR(p) process. It defines exactly what condition is needed on the coefficients $\{\phi_r,\ r = 1, ..., p\}$ to ensure that the process is well-defined and stationary.

For example, for an AR(1) process with $\phi(z) = 1 - \phi_1 z$, we find immediately that $|z_1| = |1/\phi_1| > 1$ is the stationarity condition. Also,

$$\begin{aligned}\left|\phi(e^{i\lambda})\right|^2 &= (1 - \phi_1 e^{i\lambda})(1 - \phi_1 e^{-i\lambda}) \\ &= 1 - \phi_1(e^{i\lambda} + e^{-i\lambda}) + \phi_1^2 \\ &= 1 - 2\phi_1 \cos\lambda + \phi_1^2\end{aligned}$$

which leads directly back to the formula (2.9) which we derived earlier. Note, however, that the present calculation is much more direct than the one that led to (2.9), because in (2.9) we found the spectral density only by first calculating the autocovariance function, whereas here we have gone directly via (3.2). In general it is easier to use (3.2) to calculate the spectral density and then use that to obtain autocovariances, rather than calculate the autocovariances directly.

3.3 The MA process

The MA(q) process

$$X_t = \epsilon_t + \sum_{s=1}^{q} \theta_s \epsilon_{t-s}$$

may similarly be written in operator notation as

$$X_t = \left(1 + \sum_{s=1}^{q} \theta_s B^s\right) \epsilon_t$$

or more compactly

$$X = \theta(B)\epsilon$$

where $\theta(z)$ is the generating function $1 + \sum_1^q b_s z^s$. In this case the spectral density function is

$$f_X(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \cdot |\theta(e^{i\lambda})|^2. \tag{3.4}$$

In this case there is no need for any stationarity condition, since the process is stationary whatever the coefficients $\{b_s\}$, but there is nevertheless a difficulty requiring some restriction on the coefficients. This is most easily seen in the case $q = 1$. In that case we easily calculate the autocovariances to be

$$\gamma_0 = (1 + \theta_1^2)\sigma_\epsilon^2, \quad \gamma_1 = \theta_1 \sigma_\epsilon^2, \quad \gamma_k = 0 \quad \text{for } k > 1,$$

and hence the autocorrelations

$$\rho_0 = 1, \quad \rho_1 = \frac{\theta_1}{1 + \theta_1^2}, \quad \rho_k = 0 \quad \text{for } k > 1. \tag{3.5}$$

Now consider the identical process, but with θ_1 replaced by $1/\theta_1$. It is seen from (3.5) that the autocorrelation function is unchanged by this transformation. In other words, the

two process defined by θ_1 and $1/\theta_1$ are identical for all practical purposes, so that the two processes cannot be distinguished.

As a resolution of this difficulty, it is customary to impose the following *identifiability condition*:

(**) All the zeros of the function $\theta(z)$ lie on or outside the unit circle in the complex plane.

To see why this resolves the problem, suppose we write

$$\theta(z) = \prod_{j=1}^{q} (z - z_j).$$

Then

$$\left|\theta(e^{i\lambda})\right|^2 = \prod_{j=1}^{q} \left\{ \left(e^{i\lambda} - z_j\right) \left(e^{-i\lambda} - z_j\right) \right\}.$$

However, the identity

$$\left(e^{i\lambda} - z_j\right) \left(e^{-i\lambda} - z_j\right) = z_j^2 \left(e^{i\lambda} - \frac{1}{z_j}\right) \left(e^{-i\lambda} - \frac{1}{z_j}\right)$$

shows that there is no change in $|\theta(e^{i\lambda})|$, except for a constant, in replacing any z_j by $1/z_j$. Hence there is no loss of generality in assuming all $|z_j| \geq 1$. If in fact we have $|z_j| > 1$ for all j, then the previous discussion of AR processes shows that we can invert the relation between X_t and ϵ_t to write ϵ_t as an infinite linear combination of the $\{X_s,\ s \leq t\}$. In this slightly strengthened form, the relation (**) is also known as *the invertibility relation* on the coefficients $\{\theta_j,\ 1 \leq j \leq q\}$.

3.4 The ARMA process

The general ARMA(p, q) process may, in similar notation, be written in the form

$$\phi(B)X = \theta(B)\epsilon \tag{3.6}$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are the respective generating functions of the autoregressive and moving average operators. By equating the spectral densities of the two sides of (3.6), we see that the spectral density of X is given by

$$f_X(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \cdot \frac{|\theta(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2}. \tag{3.7}$$

The conditions now required are

(a) the stationarity condition on the coefficients $\{\phi_r,\ 1 \leq r \leq p\}$,

(b) the identifiability condition on the coefficients $\{\theta_s,\ 1 \leq s \leq q\}$,

(c) an additional identifiability condition: *the generating functions $\phi(\cdot)$ and $\theta(\cdot)$ should not have any common zero.*

The reason for condition (c) is that if $\phi(\cdot)$ and $\theta(\cdot)$ have a common zero at z^* say, then it is possible to cancel a common factor $(e^{i\lambda} - z^*)(e^{-i\lambda} - z^*)$ from both the numerator and denominator of (3.7) and so reduce the model to simpler form.

As an example, consider the AR(1) model

$$X_t = 0.8X_{t-1} + \epsilon_t. \tag{3.8}$$

This model also satisfies the equation

$$X_{t-1} = 0.8X_{t-2} + \epsilon_{t-1}, \tag{3.9}$$

so by subtracting, say 0.6 times (3.9) from (3.8), we obtain the model

$$X_t = 1.4X_{t-1} - 0.48X_{t-2} + \epsilon_t - 0.6\epsilon_{t-1}$$

which looks like an ARMA(2,1) model, but of course it is in reality no different from (3.8). In this case, the zeros of $\phi(z) = 1 - 1.4z + 0.48z^2$ are at $1/0.6$ and $1/0.8$, while $\theta(z) = 1 - 0.6z$ is zero at $z = 1/0.6$, i.e. there is a common zero in the two generating functions, and when this is removed the model indeed reduces to (3.8).

3.5 Calculating autocovariances of ARMA models

One application of these formulae is to the calculation of autocovariances of ARMA models. This is useful, e.g. for deciding whether an estimated model provides a good fit to the observed time series, and can also be useful as an initial diagnostic tool.

As an example, consider the ARMA(1,2) process with generating functions

$$\phi(z) = 1 - \phi_1 z, \quad \theta(z) = 1 + \theta_1 z + \theta_2 z^2$$

and $|\phi_1| < 1$. In this case, the generating function of the Wold coefficients $\{c_r,\ r \geq 0\}$ is

$$\begin{aligned} C(z) &= \frac{\theta(z)}{\phi(z)} \\ &= (1 + \theta_1 z + \theta_2 z^2) \sum_{r=0}^{\infty} \phi_1^r z^r \\ &= \sum_{r=0}^{\infty} c_r z^r \end{aligned}$$

so by equating coefficients of z^r, we find

$$c_r = \begin{cases} 1 & \text{if } r = 0, \\ \phi_1 + \theta_1 & \text{if } r = 1, \\ \phi_1^r + \theta_1 \phi_1^{r-1} + \theta_2 \phi_1^{r-2} & \text{if } r \geq 2. \end{cases} \tag{3.10}$$

To compute the covariances, we may use the fact that for $k \geq 0$,

$$\text{Cov}\left(\sum_r c_r \epsilon_{t-r}, \sum_s c_s \epsilon_{t+k-s} \right) = \left(\sum_{r=0}^{\infty} c_r c_{r+k} \right) \sigma_\epsilon^2$$

so that

$$\begin{aligned} \gamma_0 &= \left\{ 1 + (\phi_1 + \theta_1)^2 + \frac{(\phi_1^2 + \theta_1\phi_1 + \theta_2)^2}{1 - \phi_1^2} \right\} \sigma_\epsilon^2, \\ \gamma_1 &= \left\{ \phi_1 + \theta_1 + (\phi_1 + \theta_1)(\phi_1^2 + \theta_1\phi_1 + \theta_2) + \frac{(\phi_1^2 + \theta_1\phi_1 + \theta_2)^2 \phi_1}{1 - \phi_1^2} \right\} \sigma_\epsilon^2, \\ \gamma_k &= \left\{ 1 + (\phi_1 + \theta_1)\phi_1 + \frac{(\phi_1^2 + \theta_1\phi_1 + \theta_2)\phi_1^2}{1 - \phi_1^2} \right\} (\phi_1^2 + \theta_1\phi_1 + \theta_2)\phi_1^{k-2} \sigma_\epsilon^2, \quad k \geq 2. \end{aligned} \tag{3.11}$$

Note that an alternative approach which yields part of the answer is based on the following formula, valid for $k > 2$:

$$\text{Cov}\{X_t - \phi_1 X_{t-1}, X_{t-k}\} = \text{Cov}\{\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}, X_{t-k}\} = 0$$

from which we deduce

$$\gamma_k = \phi_1 \gamma_{k-1}, \quad k > 2. \tag{3.12}$$

Although (3.12) does not yield the full solution (3.11), it may be the most useful part for identification purposes, bacause if it appears from empirical examination of the sample autocorrelations that they are geometrically decaying for $k \geq 2$, that could be taken as providing strong evidence that the process is of ARMA(1,2) form.

4. FITTING ARIMA MODELS

The basic model is ARIMA(p, d, q):

$$\phi(B)\nabla(B)^d X = \theta(B)\epsilon \tag{4.1}$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are autoregressive and moving average operators, of orders p and q respectively, satisfying the stationarity and identifiability conditions of the previous chapter, and $\nabla(B) = I - B$ is the *differencing operator* applied d times, where d is a non-negative integer.

The process of fitting an ARIMA model, as it was made explicit by Box and Jenkins, may be divided into three components,

- Identification
- Estimation
- Verification

which are iterated until a suitable model is identified.

4.1 Identification

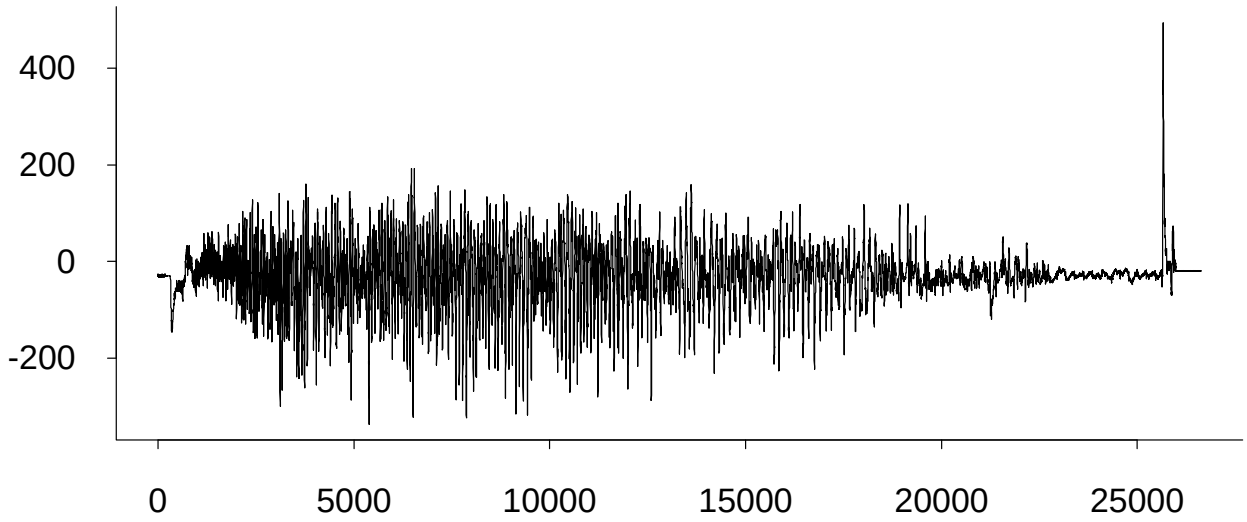
This refers both to the initial preprocessing of the data to make the series stationary, and also to the identification of suitable orders p and q for the ARMA components of the model. The latter identification, however, is always preliminary, since there is plenty of scope to adjust p and q on the basis of the models fitted.

A time series analysis should always begin with a preliminary plot of the data, as an indication of gross features that should guide the analysis. For example, Fig. 4.1 shows a raw time series plot of some electroencephalogram (EEG) data (courtesy of Professor Mike West, Duke University): one would certainly not want to analyse this as a stationary time series, even after differencing, without some initial preprocessing of the data! It might be reasonable to fit ARIMA models to portions of the series.

The main tool for initial preprocessing is differencing, though other methods (e.g. removal of deterministic components by linear regression) are perfectly acceptable, especially when there is some well-defined physical reason for the presence of this component (e.g. annual cycles in meteorological data).

As a guide to the amount of differencing (or other preprocessing) required, the main tool is the autocorrelation function (a.c.f.). With a stationary series, this should decay fairly rapidly to 0. If it fails to do so, then another layer of differencing is usually required. In practice it is rare to go beyond $d = 2$: if the series fails to look stationary after two or at most three applications of differencing, there is probably some more fundamental reason that needs separate investigation.

Fig. 4.1: Plot of EEG data



Once the series is accepted as stationary, the next step is initial identification of p and q. The main tools used for this are the a.c.f. and p.a.c.f. (partial autocorrelation function) plots. In particular,

• An MA(q) series is identified from the property that all values of the a.c.f. after the q'th are negligible,

• An AR(p) series is identified from the property that all values of the p.a.c.f. after the p'th are negligible.

As a guide to what constitutes negligibility, it is worth noting that sample values of the a.c.f. and p.a.c.f. very approximately have standard deviation around $1/\sqrt{T}$ where T is the length of the series. Thus a rule of thumb for treating these values as negligible is based on two standard deviations, or $2/\sqrt{T}$. In S-Plus, lines at $\pm 2/\sqrt{T}$ are shown on the plot as an aid in this process.

4.2 Estimation

(a) AR processes

The standard tool for autoregressive processes is to solve the *Yule-Walker equations*. The are derived from the model relationship

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t.$$

Taking the covariance of both sides with X_{t-k}, we deduce, for $k > 0$,

$$\gamma_k = \sum_{j=1}^{p} \phi_j \gamma_{|j-k|}. \tag{4.2}$$

If we consider equation (4.2) for $1 \leq k \leq p$, we get a system of p equations in p unknowns $\phi_1, ..., \phi_p$, which we can therefore solve in terms of $\gamma_1, ..., \gamma_p$. In practice, of course, we substitute the sample estimates of the autocovariances $\hat{\gamma}_1, ..., \hat{\gamma}_p$ to obtain sample estimates $\hat{\phi}_1, ..., \hat{\phi}_p$. Recall that this is operationally the same as calculating the partial autocorrelations; cf. (2.16)–(2.18). As there, one can use the estimated residual variance $\hat{\sigma}_p^2$ as a guide to the selection of appropriate order p. In particular, one can define an approximate log likelihood

$$-2 \log L = T \log(\hat{\sigma}_p^2)$$

as the basis for a likelihood ratio test statistic. Another widely used measure is the Akaike Information Criterion,

$$AIC = -2 \log L + 2k \tag{4.3}$$

where $k = p$, the number of unknown parameters in the model. The idea is to choose the model (i.e. the value of p) which minimizes AIC. This is a widely used measure in time series analysis, which has the advantage of being very convenient and quick to apply, though as with any automatic procedure, it should not be used in a totally undiscriminating way.

General ARMA processes

Now we turn to the general ARMA process. The idea here is based on numerical maximum likelihood estimation. However, most existing methods do not use exact maximum likelihood but various approximations thereto.

The essential idea behind all maximum likelihood techniques for time series is the *prediction error decomposition*. This is based on the idea that the joint density of any T random variables $X_1, ..., X_T$ may be factored as

$$f(X_1, ..., X_T) = f(X_1) \prod_{t=2}^{T} f(X_t | X_s,\ 1 \leq s \leq t-1). \tag{4.4}$$

Anticipating the notation of Chapter 6, suppose the predictive distribution of X_t given $\{X_s,\ s < t\}$ is normal with mean $\hat{X}_t$ and variance $P_{t-1,1}$. Let us assume this applies to

X_1 as well, i.e. the initial distribution of X_1 is $N(\hat{X}_1, P_{0,1})$. Then (4.4) may be rewritten as a log likelihood in the form

$$\begin{aligned} -2\log L &= -2\log f(X_1, ..., X_T) \\ &= \sum_{t=1}^{T} \left\{ \log 2\pi + \log P_{t-1,1} + \frac{(X_t - \hat{X}_t)^2}{P_{t-1,1}} \right\}. \end{aligned} \tag{4.5}$$

In this, $\hat{X}_t$, $P_{t-1,1}$ and hence L are a function of the unknown parameters $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ of the model, and maximum likelihood estimators may therefore be found by minimizing (4.5) with respect to these parameters. This is carried out by a routine application of numerical nonlinear minimization techniques, subject to the constraints on the parameters due to the stationarity and identifiability conditions. Moreover, the second derivative matrix of $-\log L$, evaluated at the maximum likelihood estimator, is the *observed information matrix*, and its inverse is an approximation to the variance-covariance matrix of the estimators. In particular, we may obtain approximate standard errors for the parameters from this matrix.

In practice the calculation is often simplified if we condition on the first m values of the series, for some small m; in that case, the sum in (4.5) is taken over $m + 1 \leq t \leq T$. For example, in an AR(p) process, for $t > p$ we trivially have

$$\hat{X}_t = \sum_{r=1}^{p} \phi_r X_{t-r}, \quad P_{t-1,1} = \sigma_\epsilon^2,$$

but it is not quite so easy to see how to compute these quantities for $t \leq p$ – this requires some consideration of the stationary distribution of the process and not merely the successive conditional distributions. A similar simplification exists, if not quite so obviously, for the general ARMA process. Therefore, in practice, the exact likelihood function is often replaced by a conditional likelihood function. The main thing to watch here is that, in comparing models with different numbers of parameters, it is important to use the same value of m. S-Plus uses conditional likelihoods and allows the user to fix m.

For comparing different models, we may again define AIC by (4.3), where k is now the total number of unknown parameters in the model, or else compare different models directly by likelihood ratio tests.

For general ARMA models, there are by now a number of approaches to computing the exact likelihood. Harvey and Phillips (1979) showed how to do this by rewriting the model in state space form (Chapter 6 of the present notes), while Ansley (1979) solved the problem directly via a Cholesky decomposition of the covariance matrix. Kohn and Ansley (1985) generalized this to the case of regression with time series errors. The Kohn-Ansley approach appears to be the most efficient when there is no missing data, but the best approach when there is missing data is again based on the Kalman filter (Kohn and Ansley 1986, Harvey 1989).

4.3 Verification

The third step of the Box-Jenkins cycle is to confirm that the model in fact fits the data. There are two basic techniques here:

• Overfitting: i.e. add extra parameters to the model and use likelihood ratio or t tests to check that these are not significant,

• Residual analysis: calculate residuals from the fitted model and plot their residuals and their a.c.f., p.a.c.f., spectral density estimates, etc., to check that they are consistent with white noise.

A useful test statistic for the residuals is the *Box-Pierce test* (also called the *portmanteau test*) which is based on

$$Q = T \sum_{k=1}^{K} r_k^2$$

where K is bigger than $p+q$ but much smaller than T, and r_k is the k'th sample autocorrelation of the residual series. If the fitted model is correct then

$$Q \sim \chi^2_{K-p-q} \quad \text{(approximately)}$$

so we can construct a test based on this.

An alternative method is the *Box-Ljung* procedure which replaces Q by

$$\tilde{Q} = T(T+2) \sum_{k=1}^{K} \frac{r_k^2}{T-k}.$$

This is recommended on the grounds that the distribution of $\tilde{Q}$ is closer to its χ^2_{K-p-q} limit than that of Q.

4.4 Seasonal models

The standard seasonal form of ARIMA model (sometimes called SARIMA) is expressed by the function

$$\phi(B)\Phi(B)\nabla(B)^d \nabla_M(B)^D X = \theta(B)\Theta(B)\epsilon, \tag{4.6}$$

where M is the (assumed known) period of the seasonality, $\nabla_M(B) = I - B^M$, and we

have

$$\begin{aligned}\theta(B) &= 1 + \sum_{j=1}^{q} \theta_j B^j, \\ \Theta(B) &= 1 + \sum_{j=1}^{Q} \Theta_j B^{Mj}, \\ \phi(B) &= 1 - \sum_{j=1}^{p} \phi_j B^j, \\ \Phi(B) &= 1 - \sum_{j=1}^{P} \phi_j B^{Mj}.\end{aligned} \tag{4.7}$$

The idea is to allow a seasonal differencing operator ∇_M, autoregressive operator Φ and moving average operator Θ, of orders D, P and Q respectively, each of which acts at lags which are multiples of M, in addition to the usual operators ∇, ϕ and θ. The whole model is sometimes called SARIMA(p, P, d, D, q, Q). Note that because of the basic algebra of the operators, it does not matter if the order of the various operators in (4.6) is interchanged.

The main steps of identification, estimation and verification in seasonal models are the same as in non-seasonal models. The main difference is that, in examining autocorrelations for both the initial identification and the final verification, particular attention must be paid to the values at or near multiples of the period M. For example, if the estimated autocorrelation $\hat{\gamma}_M$ is large but $\hat{\gamma}_{kM}$ is small for $k > 1$, this might be taken as an indication that $Q = 1$. The sample p.a.c.f. coefficients at multiples of M are used in a similar way for the initial identification of P.

4.5 Periodically correlated processes

SARIMA models are by no means the last word on seasonal data. It should be noted that any seasonal ARMA model is simply a special case of a nonseasonal ARMA model, since both the autoregressive and moving average operators in (4.6) may be expanded out as ordinary (nonseasonal) operators. These models do not allow for seasonal variability in the covariances of the process, but in many practical applications, such variability may be observed from simple plots of the variances and low-order autocorrelations as a function of the time within the cycle (e.g. month of the year in the case of monthly data).

A simple example of a periodically correlated process is the PAR(1) model

$$X_{kM+m} = \phi^{(m)} X_{kM+m-1} + \sigma^{(m)} Z_{kM+m}, \quad 1 \le m \le M,\ k \ge 0, \tag{4.8}$$

in which there are $2M$ parameters $\phi^{(1)}, ..., \phi^{(M)}$, $\sigma^{(1)}, ..., \sigma^{(M)}$ and $\{Z_t\}$ is a white noise process with variance 1. The stationarity condition for this model is

$$\prod_{m=1}^{M} |\phi^{(m)}| < 1. \tag{4.9}$$

A simple extension of this is to allow the $\{Z_t\}$ process in (4.8) to be itself a stationary ARMA process, instead of just white noise. In that case the process is called PARMA.

One can of course think about more general extensions than this, e.g. allowing higher-order PAR terms and also introducing periodic MA terms, but the PARMA model just outlined is already quite complicated and probably good enough for most practical purposes. A paper by Bloomfield, Hurd and Lund (1994) reviewed the theory of periodically correlated processes with reference to a well-known series of stratospheric ozone data (from Arosa, Switzerland), while Lund *et al.* (1995) developed further applications to ozone, temperature and CO_2 series.

We do not have the space here to get into a detailed discussion of periodically correlated models, but mention them as an alternative to the seasonal Box-Jenkins approach. In particular, if just the variances are seasonally dependent (i.e. we write $X_{kM+m} = \sigma_m Z_{kM+m}$ with $\{Z_t\}$ stationary), this may be detected by calculating sample standard deviations for each period m, and if these do appear to be non-constant, dividing through by the estimated σ_m values before fitting a stationary model to $\{Z_t\}$.

4.6 Forecasting in ARMA models

For the discussion of forecasting, it is convenient to strengthen the conditions slightly so as to assume that the moving average operator $\theta(B)$ is invertible, i.e. the zeros of the corresponding polynomial $\theta(z)$ lie outside (not merely "on or outside") the unit circle in the complex plane.

Recall that under the usual stationarity and invertibility conditions, it is possible to expand $C(z) = \theta(z)/\phi(z)$ as a power series $\sum_0^\infty c_r z^r$ to obtain the Wold representation $X_t = \sum_0^\infty c_r \epsilon_{t-r}$. Under invertibility, we can similarly expand

$$D(z) = \frac{\phi(z)}{\theta(z)} = \sum_{r=0}^{\infty} d_r z^r$$

and so rewrite the model as an infinite AR expansion

$$\epsilon_t = \sum_{r=0}^{\infty} d_r X_{t-r}. \tag{4.10}$$

The advantage of (4.10) is that, given the infinite past of the series $\{X_s,\ s < t\}$ we can solve to obtain exactly $\{\epsilon_s,\ s < t\}$.

Now suppose we are interested in forecasting X_{T+k} from observations $\{X_t,\ t \le T\}$. Working with the Wold representation, we may consider forecasts of the form

$$\hat{X}_{T,k} = \sum_{r=0}^{\infty} c_{k,r} \epsilon_{T-r} \tag{4.11}$$

which, in view of the preceding discussion, is something that may be calculated, at least in theory.

Comparing (4.11) with the Wold representation for X_{T+k}, we see that

$$X_{T+k} - \hat{X}_{T,k} = \sum_{r=0}^{k-1} c_r \epsilon_{T+k-r} + \sum_{r=0}^{\infty} (c_{r+k} - c_{k,r}) \epsilon_{T-r}.$$

Hence

$$\mathrm{E}\{(X_{T+k} - \hat{X}_{T,k})^2\} = \left\{ \sum_{r=0}^{k-1} c_r^2 + \sum_{r=0}^{\infty} (c_{r+k} - c_{k,r})^2 \right\} \sigma_\epsilon^2.$$

This expression may be minimized by setting

$$c_{r,k} = c_{r+k} \quad \text{for all } r \geq 0,\ k > 0, \tag{4.12}$$

and then gives rise to the mean squared prediction error

$$\mathrm{E}\{(X_{T+k} - \hat{X}_{T,k})^2\} = \left\{ \sum_{r=0}^{k-1} c_r^2 \right\} \sigma_\epsilon^2. \tag{4.13}$$

This therefore defines our theoretical optimal predictor, and its mean squared error. In practice, one does not usually go through the formality of constructing the Wold and infinite AR representations in this way, Instead, there is an alternative recursive approach, as follows. Define $\hat{X}_{t,k}$ to be the optimal predictor of X_{T+k} given $X_1, ..., X_T$; for $-T+1 \leq k \leq 0$, $\hat{X}_{T,k} = X_{T+k}$. We have the recursive relation

$$\hat{X}_{T,k} = \sum_{r=1}^{p} \phi_r \hat{X}_{T,k-r} + \hat{\epsilon}_{T+k} + \sum_{s=1}^{q} \theta_s \hat{\epsilon}_{T+k-s}. \tag{4.14}$$

For $k \leq 0$, (4.14) allows us to calculate estimates of the one-step prediction errors, $\hat{\epsilon}_t$ for $1 \leq t \leq T$. Then we apply (4.14) with $k > 0$, defining $\hat{\epsilon}_t = 0$ for $t > T$, to calculate the forecasts. Note that the estimates $\{\hat{\epsilon}_t,\ 1 \leq t \leq T\}$ are also the building blocks required to calculate the likelihood function via the prediction error decomposition (4.5). This is because

$$X_{t-1,1} = X_t - \hat{\epsilon}_t.$$

The difficulty that remains is how to start off the recursion (4.14). There are two standard solutions to that:

(a) the *conditional* approach, in which we assume $X_t = \epsilon_t = 0$ for all $t \leq 0$,

(b) the *backcasting* approach, in which we forecast the series in reverse direction to determine estimates of X_0, X_{-1},..., as well as $\epsilon_0 = 0$, $\epsilon_{-1} = 0$, etc.

An alternative superior approach is, however, to recast the model in state space form and apply the Kalman filter (Chapter 6).

Another point to note throughout this discussion is that the "mean squared prediction errors" we have derived are based *solely* on the uncertainties of prediction: they do not make any allowance for errors in model identification. An extended approach which also takes into account the standard errors of the parameter estimates has been given by Ansley and Kohn (1986).

4.7 An example

Our example is the Lake Huron data set from Brockwell and Davis (1991), p. 555. Tthe data are 98 mean levels of Lake Huron in feet (relative to a fixed standard) for the years 1875–1972. The following discussion illustrates the use of various S-Plus functions on this data set.

Let us suppose the data file `huron.dat` is available, containing the observations in sequence. To enter S-Plus, enter the command

```
Splus
```

You should receive a banner message looking something like

```
S-PLUS : Copyright (c) 1988, 1996 MathSoft, Inc.
S : Copyright AT&T.
Version 3.4 Release 1 for Sun SPARC, SunOS 5.3 :  1996
Working data will be in /afs/isis.unc.edu/home/r/l/rls/.Data
>
```

The last symbol (`>`) is the S-Plus prompt. Note that the command to quit S-Plus is `q()`.

A good command to enter next is

```
X11()
```

which opens a graphics window. To read in the data from the file "huron.dat" into S-Plus, where we shall use the data name "hur", type

```
hur<-scan(file='huron.dat')
```

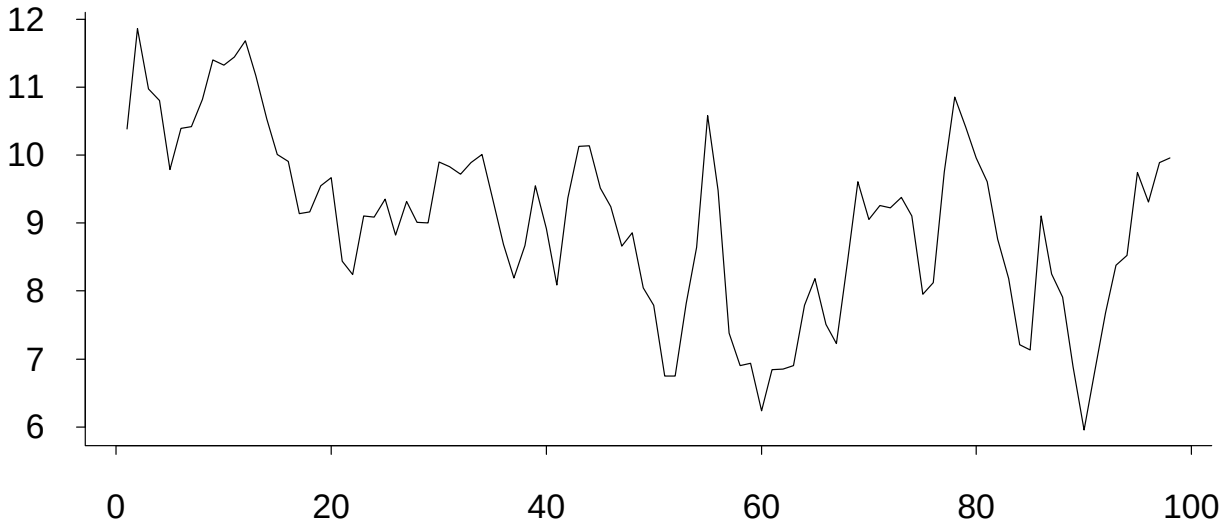
The command

```
tsplot(hur)
```

then produces a so-called time series plot of the data; see Fig. 4.2. The plot shows some evidence of a decreasing trend over the first few years, but there is no visible reason why

this could not be normal fluctuations of a stationary process, so for the time being at least, we do not consider any differencing operation.

Fig. 4.2: Plot of Huron data



A good thing to do next is to plot the a.c.f. and p.a.c.f. functions. The S-Plus commands to do this are

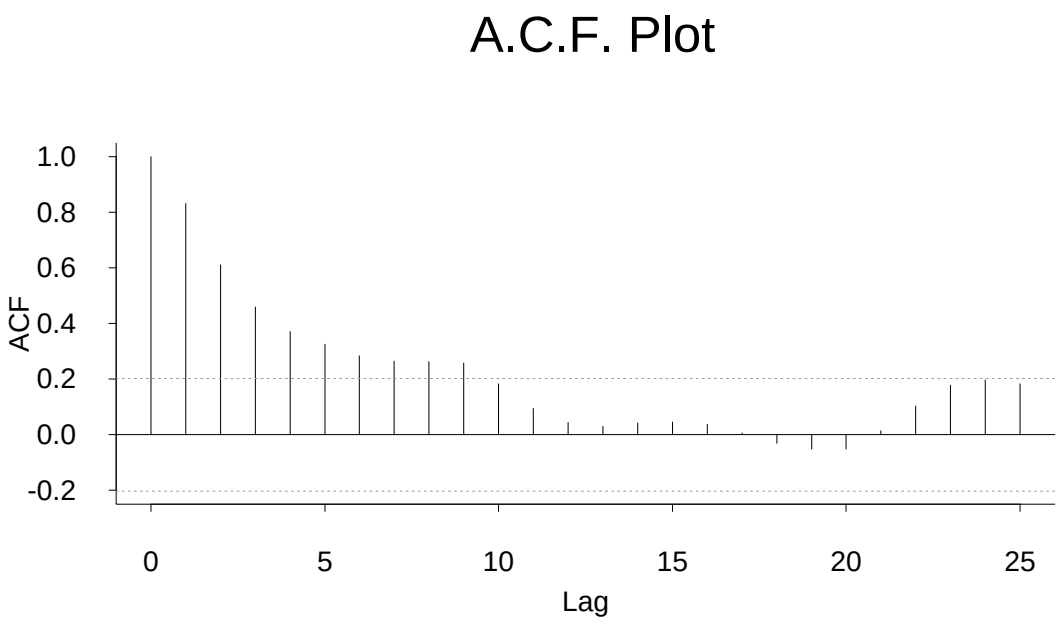
```
acf(hur,25)
```

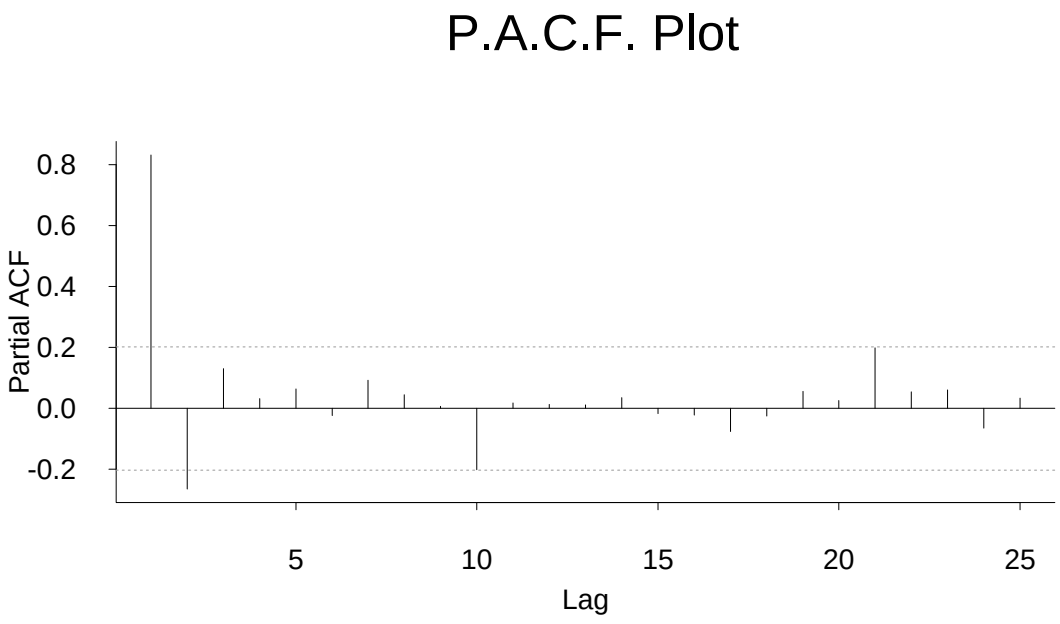
and

```
acf(hur,25,"par")
```

which produces the plots shown in Fig. 4.3 (the "25" here is the number of lags). The a.c.f. plot lies within the approximate 95% confidence bands from lag 10 onwards, while the p.a.c.f. plot shows insignificant values after lag 2. This suggests that the series is stationary with an initial guess of an AR(2) model. In both cases there are some nearly

Fig. 4.3: Autocorrelations of Huron data





significant correlations at higher lags but for the moment we ignore these – it is most likely that they are spurious, but if not, they should still be present after fitting a model to the data.

The next analysis tried used the S-Plus command

```
ar.out<-ar.yw(hur,order.max=10)
```

which creates an S-Plus data object "ar.out" by solving the Yule-Walker equations for every order up to `order.max`. The program then uses AIC to select among these models.

If you type

```
ar.out
```

you will get the full output of the model fit, some of which is shown below:

```
$order:
[1] 2

$ar:

, , 1
[,1]
[1,] 1.0519085
[2,] -0.2649543

$var.pred:
[,1]
[1,] 0.5094762

$aic:
[1] 118.3914795 5.1330872 0.0000000 0.3157349 2.2109680 3.8161316
[7] 5.7627869 6.9114075 8.7189941 10.7152405 8.6430969
```

and so on. You can get individual pieces of this by using the $ attachment to the data file name, e.g. `ar.out$order` for the model order, `ar.out$ar` for the estimated parameters, `ar.out$aic` for the individual AIC values, and so on.

The `ar.yw` procedure fits a series of AR models up to order 10 by the Yule-Walker equations and chooses a "best" model by AIC. In this case the model selected is $p = 2$ and the AIC values, relative to this minimum AIC, are 118.39 (corresponding to $p = 0$), 5.133 ($p = 1$), 0.000 ($p = 2$ – the minimum), 0.3157 ($p = 3$), 2.2110 ($p = 4$), etc. The AIC value for $p = 3$ is small, suggesting that this is a reasonable competitor, but the others are much

larger. The fitted parameters for $p = 2$ are $\phi_1 = 1.0519$, $\phi_2 = -0.2650$; it is apparently not possible to obtain standard errors within this function.

The next step was to try a series of fits using the `arima.mle` function. To use this properly, we first have to center the series about 0 by subtracting the overall sample mean

```
hur<-hur-mean(hur)
```

and then a call of the form

```
hur.mod<-arima.mle(hur,n.cond=5,model=list(order=c(2,0,1)))
```

which requests an ARIMA model fitted to the data set `hur`, where the coefficients following `order=` are the orders p, d and q. Thus in this example $p = 2, d = 0, q = 1$. The option `n.cond=5` fixed the number of initial variables to condition on (m in the discussion following (4.5) above) is set to be 5 – as mentioned earlier, it is important for comparison of different models that this value should be the same for all models. It is also possible to define initial values of the parameters by using the `ar=` and `ma=` options.

The S-Plus analysis gives, amongst other things, the estimated model parameters and their variance-covariance matrix, and the values of `loglik` (=–2 log L) and AIC. There is also an indicator of whether or not the iteration has converged. For the following analysis, the strategy followed has been to start with the AR(1) and MA(1) models and add parameters until it appears from the `loglik` and AIC values that the fit cannot be improved. An important preliminary is to subtract the sample mean from all data points. This analysis was repeated with first differenced data ($d = 1$) to gain some indication of whether differencing would improve the model fit. The results are given in Table 4.1.

It should be noted that some models have an F in the "converged" column, indicating that no convergence has been achieved. In some cases, this is an indication of bad starting values, and the fit should be repeated using the `ar=` and `ma=` options to try different values of the starting parameters. The other reason why the algorithm may fail to converge is because it may hit the boundary contraints imposed by the stationarity and identifiability conditions. That should probably be taken as a warning of a poor choice of model. In the above case, where the algorithm fails to converge, it would appear that this is the reason. For instance, with the ARMA(2,1) model, the final parameters are $\hat{\phi}_1 = 0.0894$, $\hat{\phi}_2 = 0.5762$, $\hat{\theta}_1 = -0.9936$, the last parameter being very close to the boundary at $\theta = 1$.

It appears from this table that the best model is ARMA(1,1) to the undifferenced data. This narrowly inproves on the AR(2) model identified earlier. Moreover, none of the differenced models improves on this, implying that differencing is unnecessary in this example. The ARMA(1,1) parameter fits are $\hat{\phi}_1 = 0.7386$, $\hat{\theta}_1 = -0.3485$ and the variance-covariance matrix is obtained as

$$\begin{pmatrix} 0.006534 & 0.004568 \\ 0.004568 & 0.012639 \end{pmatrix}$$

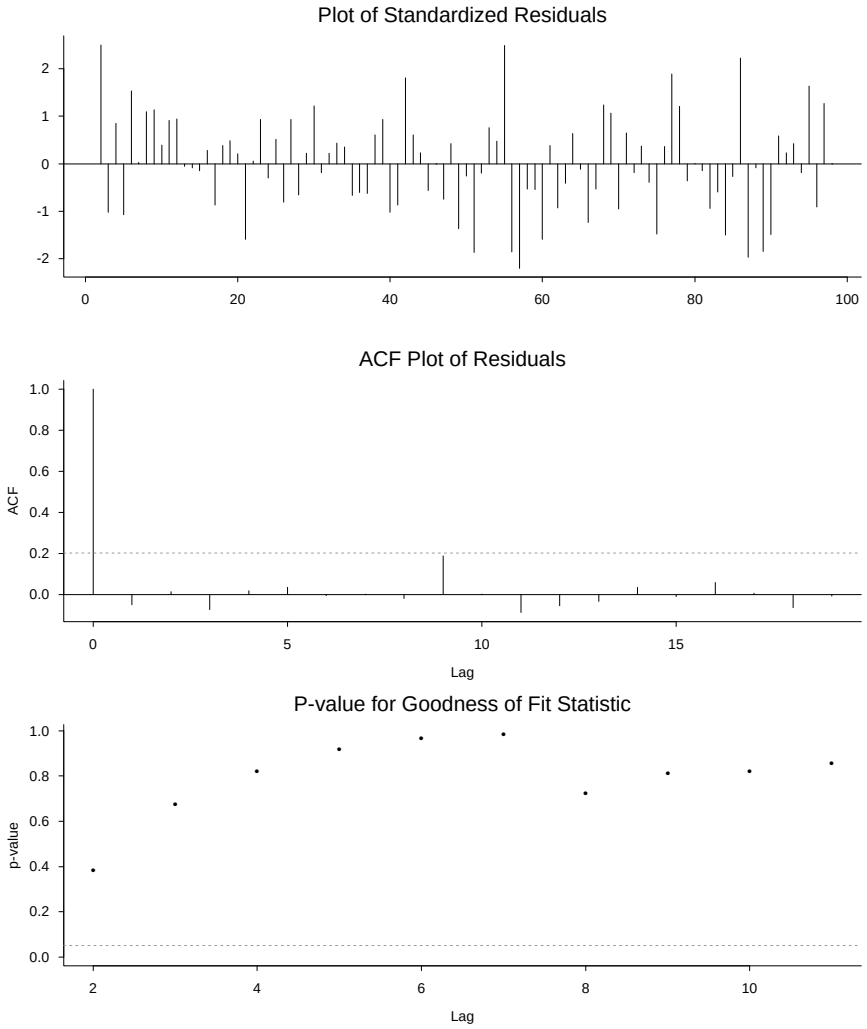
so that the standard errors for $\hat{\phi}_1$ and $\hat{\theta}_1$ are respectively 0.081 and 0.112.

Table 4.1: Model fits to Huron data

Order	loglik	AIC	Converged
(1,0,0)	198.2686	200.2686	T
(0,0,1)	231.1679	233.1679	T
(2,0,0)	191.8183	195.8183	T
(1,0,1)	190.4097	194.4097	T
(2,0,1)	187.5088	193.5088	F
(1,0,2)	190.4012	196.4012	T
(3,0,0)	189.8072	195.8072	T
(3,0,1)	185.7669	193.7669	F
(1,1,0)	201.5242	203.5242	T
(0,1,1)	199.7084	201.7084	T
(1,1,1)	195.5389	199.5389	F
(0,1,2)	196.5308	200.5308	T
(2,1,0)	195.7203	199.7203	T

Finally, the `arima.diag` command was applied to the output from `arima.mle` for the ARMA(1,1) model. The result is Fig. 4.4. This shows, respectively, the standardized residuals from the fitted mode, the a.c.f. of those residuals, and the P-values of the Box-Pierce test for different values of K. The latter are of concern if they go below 0.05 (marked on the plot). It would appear that the model is a good fit.

Fig. 4.4: Huron model diagnostics



5. ESTIMATING SPECTRAL DENSITIES

5.1 Regression on sinusoidal components

The simplest form of spectral analysis consists of regression on a periodic component:

$$Y_t = A\cos\omega t + B\sin\omega t + C + \epsilon_t, \quad 1 \le t \le T,$$

in which $\{\epsilon_t\}$ are uncorrelated with mean 0 and common variance σ^2. We may assume $0 \le \omega \le \pi$ (recall the discussion of aliasing in Chapter 2).

We may rewrite this model in standard vector-matrix notation

$$Y = X\beta + \epsilon$$

in which $Y = (Y_1, ..., Y_T)^T$ is the vector of observations, $\epsilon = (\epsilon_1, ..., \epsilon_T)$ the vector of errors, $\beta = (A\ B\ C)^T$ and X is the $T \times 3$ matrix

$$X = \begin{pmatrix} \cos\omega & \sin\omega & 1 \\ \cos 2\omega & \sin 2\omega & 1 \\ \vdots & \vdots & \vdots \\ \cos T\omega & \sin T\omega & 1 \end{pmatrix}.$$

Then the estimates are given by ordinary least squares as

$$\begin{pmatrix} \hat{A} \\ \hat{B} \\ \hat{C} \end{pmatrix} = (X^TX)^{-1}X^TY = (X^TX)^{-1}\begin{pmatrix} \sum Y_t\cos\omega t \\ \sum Y_t\sin\omega t \\ \sum Y_t \end{pmatrix}$$

and they are unbiased with a covariance matrix is given by the usual formula $(X^TX)^{-1}\sigma^2$. If the $\{\epsilon_t\}$ are independent normal, then the estimates are also independent and normally distributed.

These formulae take a much simpler form if ω is one of the *Fourier frequencies*, defined by $\omega_j = 2\pi j/T$ for some integer j between 0 and $T/2$. We may note the following elementary trigonometric identities:

$$\sum_{t=1}^{T}\cos\omega_j t = \sum_{t=1}^{T}\sin\omega_j t = 0 \quad (j \ne 0),$$

$$\sum_{t=1}^{T}\cos\omega_j t\ \cos\omega_k t = \begin{cases} \frac{T}{2} & \text{when } j = k\ (\ne 0 \text{ or } \frac{T}{2}), \\ T & \text{when } j = k = 0 \text{ or } \frac{T}{2}, \\ 0 & \text{when } j \ne k, \end{cases}$$

$$\sum_{t=1}^{T}\sin\omega_j t\ \sin\omega_k t = \begin{cases} \frac{T}{2} & \text{when } j = k\ (\ne 0 \text{ or } \frac{T}{2}), \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{t=1}^{T}\cos\omega_j t\ \sin\omega_k t = 0 \qquad \text{for all } j, k.$$

For simplicity we do not consider the cases $j = 0$, $j = T/2$, which produce similar but slightly different formulae. For any other Fourier frequency ω_j we have

$$X^T X = \begin{pmatrix} \frac{T}{2} & 0 & 0 \\ 0 & \frac{T}{2} & 0 \\ 0 & 0 & T \end{pmatrix}.$$

We then have

$$\hat{A} = \frac{2}{T} \sum Y_t \cos \omega_j t,$$
$$\hat{B} = \frac{2}{T} \sum Y_t \sin \omega_j t,$$
$$\hat{C} = \bar{Y} = \frac{1}{T} \sum Y_t,$$

and these are uncorrelated estimates with variances $2\sigma^2/T$, $2\sigma^2/T$, σ^2/T, respectively.

A suitable way of testing the significance of the sinusoidal component with frequency ω_j is the reduction in RSS in the above regression,

$$R_T(\omega_j) = \frac{T}{2}(\hat{A}^2 + \hat{B}^2).$$

If the $\{\epsilon_t\}$ are independent normal, then it follows that $\hat{A}$ and $\hat{B}$ are also independent normal each with variance $2\sigma^2/T$, so under the null hypothesis $A = B = 0$ we find that

$$\frac{R_T(\omega_j)}{\sigma^2} \sim \chi^2_2,$$

or equivalently that $R_T(\omega_j)/(2\sigma^2)$ has an exponential distribution with mean 1.

The above theory is easily extended to simultaneous estimation of several periodic components. In particular, if we consider estimation of sinusoidal terms at k Fourier frequencies $\omega_{j_1}, ..., \omega_{j_k}$, the corresponding columns of the X matrix are orthogonal. This means that the point estimates of the coefficients are the same when all k components are estimated simultaneously as when they are estimated one at a time, and also that the parameter estimates, and hence the R_T statistics, are uncorrelated.

Under the null hypothesis that all the A and B coefficients are 0, we have the following result:

(*) *The k test normalized statistics* $R_T(\omega_{j_1})/(2\sigma^2), ..., R_T(\omega_{j_k})/(2\sigma^2)$ *are independent exponentially distributed random variables each with mean 1.*

This is an exact result for all T if the $\{\epsilon_t\}$ are independent normal. It is also valid as an approximation for large T if the $\{\epsilon_t\}$ are independent non-normal. This is because

the Central Limit Theorem guarantees that the parameter estimates are approximately independent normal for large T.

5.2 The periodogram.

The foregoing discussion turns out to be very useful in studying some of the fundamental properties of the periodogram. Recall that the periodogram was defined in (2.19), where it was stated (without any proof) that it is an approximately unbiased estimator of the spectral density f. In this section we shall provide an informal proof of this, and derive some related statistical properties along the way.

In terms of the $R_T(\omega)$ statistics defined above, the periodogram is

$$I_T(\omega) = \frac{R_T(\omega)}{4\pi}.$$

By the property (*), if the process if white noise then the $R_T(\omega_j)$ at Fourier frequencies $\{\omega_j,\ 1 \le j < T/2\}$ are independent exponentially distributed with common mean $\sigma^2/(2\pi) = f(\omega_j)$. This result is exact if the $\{\epsilon_j\}$ are independent normal, and approximate for large T if they are independent non-normal.

The general result is the following:

Theorem. Suppose $Y_t = \sum_{r=0}^{\infty} c_r \epsilon_{t-r}$ is a linear process, with independent $\{\epsilon_t\}$ having mean 0 and common variance σ^2. Suppose this process is stationary with a spectral density

$$f(\omega) = \frac{\sigma^2}{2\pi}|C(e^{i\omega})|^2. \quad (C(z) = \sum_{r=0}^{\infty} c_r z^r.)$$

Then the periodogram ordinates $\{I_T(\omega_j),\ 1 \le j < T/2\}$, are approximately independent and exponentially distributed, with means $\{f(\omega_j),\ 1 \le j < T/2\}$.

Heuristic proof. Let us write

$$\begin{aligned}
\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} Y_t e^{i\omega t} &= \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} \sum_{r=0}^{\infty} c_r \epsilon_{t-r} e^{i\omega t} \\
&= \frac{1}{\sqrt{2\pi T}} \sum_{r=0}^{\infty} c_r e^{i\omega r} \sum_{t=1}^{T} \epsilon_{t-r} e^{i\omega(t-r)} \\
&= \frac{1}{\sqrt{2\pi T}} \sum_{r=0}^{\infty} c_r e^{i\omega r} \sum_{u=1-r}^{T-r} \epsilon_u e^{i\omega u}.
\end{aligned}$$

For large T and fixed r, we may approximate

$$\sum_{u=1-r}^{T-r} \epsilon_u e^{i\omega u} \approx \sum_{u=1}^{T} \epsilon_u e^{i\omega u} \tag{5.1}$$

essentially because we are adding or removing a total of r terms and these are negligible compared with the total length of the sum T.

If we accept the approximation (5.1), we then have

$$\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} Y_t e^{i\omega t} \approx C(e^{i\omega}) \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} \epsilon_t e^{i\omega t}$$

and hence, on taking squared moduli of both sides,

$$I_{T,Y}(\omega) \approx |C(e^{i\omega})|^2 I_{T,\epsilon}(\omega) \tag{5.2}$$

with obvious notation: $I_{T,Y}$ and $I_{T,\epsilon}$ are the periodograms of the Y and ϵ processes respectively.

In combination with the result (*), (5.2) then shows that the periodogram ordinates of Y, evaluated at Fourier frequencies, are approximately independent. Moreover,

$$\mathrm{E}\{I_{T,Y}(\omega_j)\} \approx |C(e^{i\omega_j})|^2 \mathrm{E}\{I_{T,\epsilon}(\omega_j)\} = f(\omega_j).$$

This completes the proof. [A fully rigorous proof is possible under the condition $\sum |c_r| < \infty$, in which case it can be shown that the difference between the two sides in (5.2) converges to 0 in probability, uniformly over all ω, as $T \to \infty$. We shall not attempt to fill in the details of this.]

This theorem is the central result of spectral estimation theory. It is very powerful: for instance, note that we have nowhere attempted to give any remotely similar result for the estimation of sample autocorrelations (and indeed the corresponding results for that problem are much more complicated to state). This is one reason why experienced time series analysts often feel they can gain more information about the process by studying the spectrum than by studying the autocorrelations. However, it also points to some undesirable features of the periodogram: $I_T(\omega)$ for a fixed ω is not a consistent estimate of $f(\omega)$, since it has an approximate exponential distribution with mean $f(\omega)$, and therefore variance $f^2(\omega)$, which does not tend to 0 as $T \to \infty$. Also, the independence of periodogram ordinates at different Fourier frequencies suggests that the sample periodogram, plotted as a function of ω, will be extremely irregular. There are some additional difficulties, which we develop a little later on, with the performance of the sample periodogram in the presence of a sinusoidal variation whose frequency is not one of the Fourier frequencies. These difficulties cause us to introduce two new operations on the periodogram, *smoothing* and *tapering*.

5.3 Smoothing

The idea behind smoothing is to take weighted averages over neighboring frequencies in order to reduce the variability associated with individual periodogram values. However,

such an operation necessarily introduces some bias into the estimation procedure. Theoretical studies focus on the amount of smoothing that is required to obtain an optimum trade-off between bias and variance. In practice, this usually means that there is some subjective judgement to be performed.

The main form of smoothed estimator is given by

$$\hat{f}(\lambda) = \int_{-\pi}^{\pi} \frac{1}{h} K\left(\frac{\lambda - \omega}{h}\right) I_T(\omega) d\omega \tag{5.3}$$

where $I_T(\cdot)$ is the periodogram based on T observations, $K(\cdot)$ is a *kernel function* and h is the *bandwidth*. We usually take $K(\cdot)$ to be a non-negative function, symmetric about 0, and integrating to 1. Thus any symmetric density, such as the normal, will do, though in practice it is more usual to take one of finite range, such as the *Epanechnikov kernel*

$$K(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right), \quad -\sqrt{5} \le t \le \sqrt{5}, \tag{5.4}$$

which is 0 outside $[-\sqrt{5}, \sqrt{5}]$. This choice of kernel function has some optimality properties, though as a practical matter it is generally agreed that these are much less important than the choice of bandwidth h, which effectively controls the range over which the periodogram is smoothed. The following discussion will show how the bandwidth is chosen theoretically.

Consider first the bias of the estimator (5.3). We may write

$$\begin{aligned}
\mathrm{E}\{\hat{f}(\lambda)\} &\approx \int_{-\pi}^{\pi} \frac{1}{h} K\left(\frac{\lambda - \omega}{h}\right) f(\omega) d\omega \\
&\approx \int_{-\infty}^{\infty} K(x) f(\lambda + hx) dx \\
&= \int_{-\infty}^{\infty} K(x) \left\{ f(\lambda) + hx f'(\lambda) + \frac{h^2 \lambda^2}{2} f''(\lambda) + o(h^2) \right\} dx \\
&= f(\lambda) + \frac{h^2 f''}{2} \int x^2 K(x) dx + o(h^2).
\end{aligned}$$

Here, the second line follows from the substitution $\omega = \lambda + hx$ and the change of range from $(-\pi, \pi)$ to $(-\infty, \infty)$ follows from the fact that, with small h, such a transformation does occur asymptotically provided λ is an interior point of the interval $(-\pi, \pi)$. The subsequent expansion obviously assumes that f is at least twice continuously differentiable, so from now on we make that as an explicit assumption. The final result justifies the approximation

$$\text{Bias in } \hat{f}(\lambda) \sim \frac{h^2}{2} f''(\lambda) \int x^2 K(x) dx, \tag{5.5}$$

valid asymptotically as $T \to \infty$ and $h \downarrow 0$.

To examine the variance of (5.3), we should first note that, in practice, (5.3) will be evaluated via a Riemann sum,

$$\hat{f}(\lambda) = \sum_j \int_{\omega_{j-1}}^{\omega_j} \frac{1}{h} K\left(\frac{\lambda-\omega}{h}\right) I_T(\omega) d\omega \approx \frac{2\pi}{T} \sum_j \frac{1}{h} K\left(\frac{\lambda-\omega_j}{h}\right) I_T(\omega_j)$$

where $\omega_j = 2\pi j/T$ is the j'th Fourier frequency. Since the $I_T(\omega_j)$ are asymptotically independent,

$$\begin{aligned} \mathrm{Var}\{\hat{f}(\lambda)\} &\approx \frac{4\pi^2}{T^2} \sum_j \frac{1}{h^2} K^2\left(\frac{\lambda-\omega_j}{h}\right) f^2(\omega_j) \\ &\approx \frac{2\pi}{hT} \int \frac{1}{h} K^2\left(\frac{\lambda-\omega}{h}\right) f^2(\omega) d\omega \\ &= \frac{2\pi}{hT} \int K^2(x) f^2(\lambda + hx) dx \\ &\approx \frac{2\pi}{hT} f^2(\lambda) \int K^2(x) dx. \end{aligned} \tag{5.6}$$

So (5.5) and (5.6) together give the approximate bias and variance of $\hat{f}(\lambda)$. It will be noted that these are of the form

$$\text{Bias} \approx \frac{A}{h^2}, \quad \text{Variance} \approx \frac{B}{hT}$$

where A and B are constants (depending on f and K, but not T or h). Thus the mean squared error (=Bias^2+Variance) is approximately

$$A^2 h^4 + \frac{B}{hT}$$

which is minimized by taking

$$h = \left(\frac{B}{4A^2 T}\right)^{\frac{1}{5}} \tag{5.7}$$

and then leads to the optimal mean squared error

$$\text{MSE} \approx 5 \times 4^{-4/5} A^{2/5} B^{4/5} T^{-4/5}. \tag{5.8}$$

These formulae are rather difficult to apply in practice, because to evaluate A and B requires not only that we know $f(\lambda)$, which we are trying to estimate, but also $f''(\lambda)$, which is even harder. However, there are a number of ways in which these theoretical results are useful in practice, e.g.:

(i) We can try a suite of test cases, such as ARMA models with known spectral densities, to evaluate the optimal bandwidths for them, as a guide to what to do when the true spectral density is unknown.

(ii) The asymptotic results in (5.7) and (5.8) show how the results scale with T, i.e. that the optimal bandwidth scales with $T^{-1/5}$ and the resulting MSE is proportional to $T^{-4/5}$. This gives us some guideline when comparing data sets of different sizes.

(iii) At a more advanced level (well beyond the scope of this course), attempts have been made at "automatic" bandwidth selection based on interactively updating the estimates of the parameters A and B. Such methods can still be somewhat difficult to apply in practice, however.

5.4 Tapering

This is based on a different idea, that of reducing *leakage*.

Consider a series of the form $x_t = \cos(\Omega t + \phi),\ 1 \le t \le T$ where Ω is not necessarily a Fourier frequency. We may write this as the real part of $Ae^{i\Omega t}$, for some complex number A, so for simplicity we consider the complex series $x_t = e^{i\Omega t},\ 1 \le t \le T$. The discrete Fourier transform of this series, evaluated at a frequency ω, is given by

$$\frac{1}{T}\sum x_t e^{-i\omega t} = \frac{1}{T}\sum_{t=1}^{T} e^{i(\Omega-\omega)t}.$$

To evaluate this, we make the following side calculation:

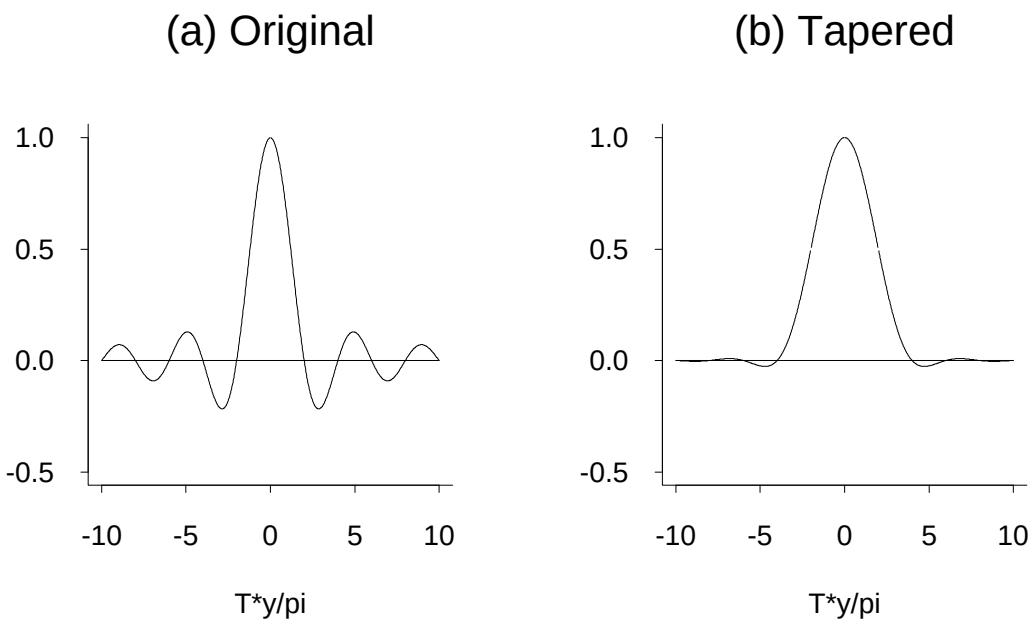
$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} e^{iyt} &= \frac{1}{T}e^{iy}\cdot\frac{e^{iTy}-1}{e^{iy}-1} \\
&= \frac{1}{T}e^{i(T+1)y/2}\cdot\frac{e^{iTy/2}-e^{-iTy/2}}{e^{iy/2}-e^{-iy/2}} \\
&= e^{i(T+1)y/2}\cdot\frac{1}{T}\frac{\sin(Ty/2)}{\sin(y/2)} \\
&= e^{i(T+1)y/2}\cdot D_T(y)
\end{aligned}$$

where $D_T(\cdot)$ is a function known as the *Dirichlet kernel*. We have $D_T(0) = 1$, $D_T(y) = 0$ when $y = 2\pi k/T$ for integer k, but there are significant "side-lobes" to either side of the main peak (Fig. 5.1(a)). These properties enter the periodogram because, in the above case of $x_t = e^{i\Omega t}$, we have

$$I_T(\omega) \propto |D_T(\omega - \Omega)|^2.$$

If Ω is a Fourier frequency, and if I_T is evaluated only at Fourier frequencies $\{\omega_j\}$, then there is no problem, but at non-Fourier frequencies Ω, these side-lobes may represent a significant distortion of the spectrum.

Fig 5.1: Dirichlet kernel, T=200



A solution to the side-lobes problem is *tapering*. Suppose we refine x_t into

$$\begin{aligned}\tilde{x}_t &= \frac{x_t}{2}\left\{1-\cos\left(\frac{2\pi(t-1/2)}{T}\right)\right\} \\ &= \frac{1}{2}e^{i\Omega t} - \frac{1}{4}e^{i\Omega t+2\pi i(t-1/2)/T} - \frac{1}{4}e^{i\Omega t-2\pi i(t-1/2)/T}\end{aligned} \tag{5.9}$$

so that, after a little manipulation, we see that

$$\begin{aligned}&\frac{1}{T}\sum_t \tilde{x}_t e^{-i\omega t} \\ &\quad = e^{i(T+1)(\Omega-\omega)/2}\left\{\frac{1}{2}D_T(\Omega-\omega)+\frac{1}{4}D_T\left(\Omega-\omega+\frac{2\pi}{T}\right)+\frac{1}{4}D_T\left(\Omega-\omega-\frac{2\pi}{T}\right)\right\}.\end{aligned}$$

Thus we see that, modulo a constant 2, the function $D_t(y)$ is replaced by a modified form of Dirichlet kernel

$$\tilde{D}_T(y) = D_T(y)+\frac{1}{2}D_T\left(y+\frac{2\pi}{T}\right)+\frac{1}{2}D_T\left(y-\frac{2\pi}{T}\right)$$

obtained by averaging over neighboring sidelobes. This indeed has a dramatic effect as can be seen from Fig. 5.1(b). There is now a very strong central peak and the neighboring sidelobs, although still present, are so small as to be almost invisible.

The operation that takes x_t into $\tilde{x}_t$ is called *tapering*. In less technical language, the idea behind this is to reduce the end-effects associated with forming a finite Fourier transform. A more general form of taper is to define $\tilde{x}_t = x_t h_t$ where

$$h_t = \begin{cases} \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(t-1/2)}{m}\right)\right\} & 1 \le t \le m, \\ 1 & m+1 \le t \le T-m, \\ \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(T-t+1/2)}{m}\right)\right\} & T-m+1 \le t \le T, \end{cases} \tag{5.10}$$

which is called a $100p\%$ *cosine taper* where usually $p = 2m/T$. (Note, however, that S-Plus departs from this convention by defining $p = m/T$ instead.) Thus the operation (5.9) is a 100% taper; in practice it is usual to apply a 10% or 20% taper in the grounds that this significantly reduces leakage without distorting the original data too much.

Much of the above discussion is taken from Bloomfield (1976), while the recent book by Percival and Walden (1994) goes into the whole issue of tapering in much more detail.

5.5 Examples

The following examples illustrate the S-Plus routine `spec.pgram`.

Fig. 5.2 shows time series plots (`tsplot` command) based on a monthly series of carbon dioxide readings at Mauna Loa, Hawaii, from January 1958 to December 1988 (31 years' complete data). This particular series has become famous in the debate over the greenhouse effect, because it is one of the longest available series of atmospheric CO_2 readings, and it shows a very clear upward trend, as well as strong seasonality.

To study the nature of the trend, Fig. 5.2 also displays residuals from the series after a linear, quadratic or cubic trend have been subtracted (in all three cases, by ordinary least squares without any adjustment for seasonality). The residuals from a linear trend show clear curvature, but there is no evidence of a systematic trend in either of the remaining two plots. Therefore, it appears that among simple polynomial trends, the quadratic trend is the appropriate one to adopt. This is also confirmed by F tests for the significance of the quadratic and cubic components, though of course these should not be taken too seriously in view of the evident seasonality and autocorrelation in the data.

Fig. 5.2: Plots of Mauna Loa data

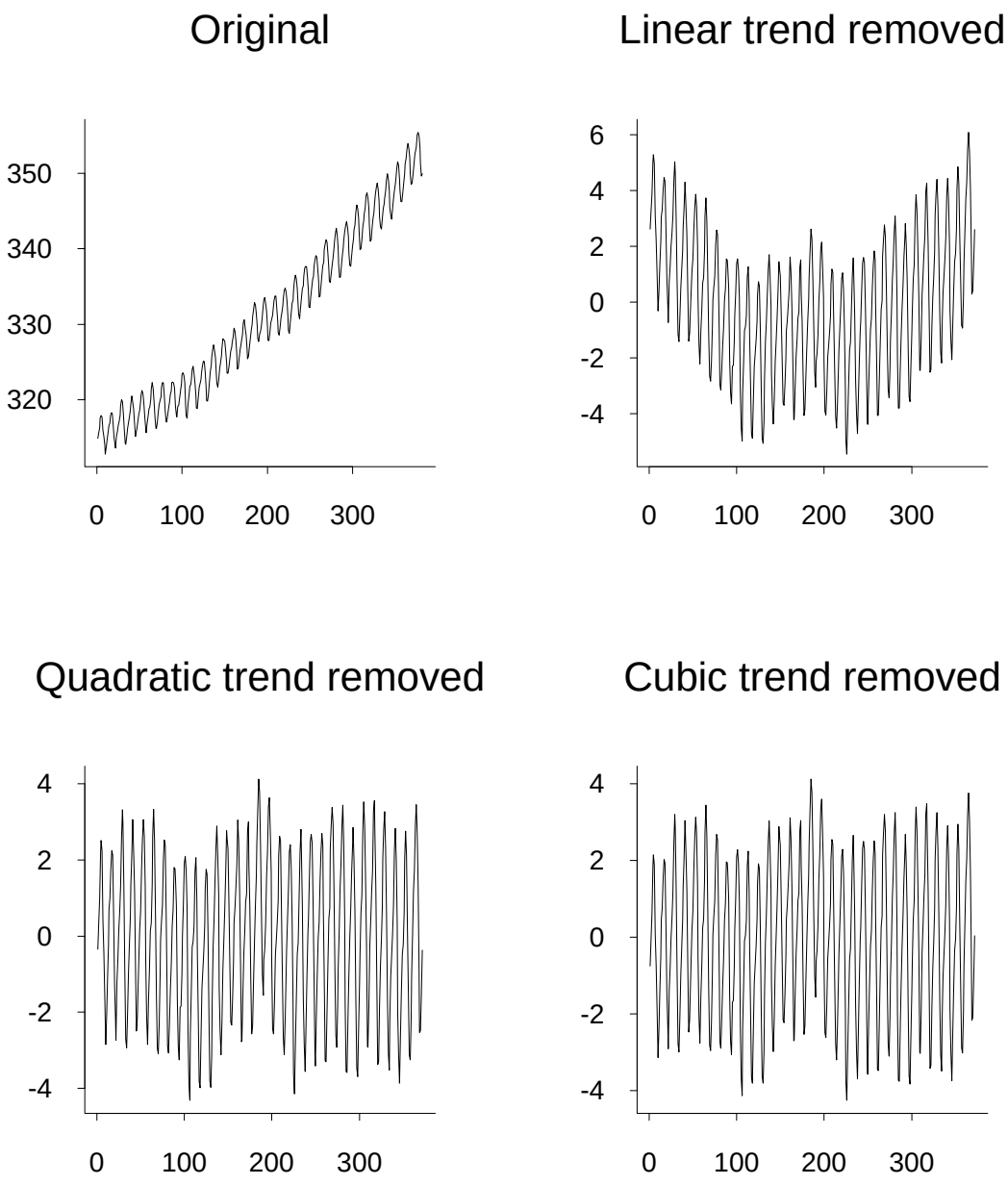


Fig. 5.3 shows spectral plots from the raw data. In accordance with the default option in S-Plus, all of these are based on the data from which a linear trend has been subtracted. Four plots are shown, corresponding to the S-Plus options `spans=1`, which gives the raw periodogram, `spans=c(5,5)`, `spans=c(13,13)` and `spans=c(21,21)`, which correspond to increasing amounts of smoothing. The cross emblem in the upper right corner of the plot represents the bandwidth of the smoother (cross-piece) and the upper and lower bounds of a pointwise 95% confidence interval for the spectral density about the plotted curve (vertical line of the cross). The confidence bands are useful because they give an idea about the significance of the various fluctuations in the plot. For example, in plot (b) it seems clear that the small wiggles at the right hand end of the plot are not significant, which the large peaks at approximately frequencies 1/12, 1/6 etc. do appear to be significant. Plot(b) seems undersmoothed, but by the time we get to plot (d), even the large peaks are merging into one another so this seems clearly overmoothed. The ideal seems somewhere between plots (c) and (d), though this is a subjective judgement.

Fig. 5.4 shows similar plots but based on the series in which a quadratic (rather than linear) trend has been removed from the data. The overall comparison of these plots, particularly with respect to the amount of smoothing required, seems similar to Fig. 5.3, but there is noticably less of a peak near frequency 0. High spectral density at low frequencies is usually taken as an indicator of trends being present in the data, though it can also be an indicator of long-range dependence, a topic which lies beyond the scope of the present notes. In this case, however, it seems clear from Fig. 5.2 that a non-linear trend is present in the data, so this is the most natural explanation.

Apart from its behavior at 0, the most prominent feature of the spectra in both Figs. 5.3 and 5.4 is the seasonal variation, which is apparent from the main peak at frequency 1/12 but also from subsidiary peaks near frequencies 1/6, 1/4 and 1/3. These are the harmonics of the main annual cycle and are to be expected in a series, such as this one, in which there is a strongly defined, but non-sinusoidal, periodic variation.

Fig. 5.5 illustrates the effect of tapering. In this case we have two plots with the same amount of smoothing, but with 0% and 100% ($p = 0.5$ in S-Plus) tapering. It is obvious that the tapered series leads to more sharply defined peaks in the spectral density.

Fig. 5.6 is a time series plots of a series of "global" temperature data produced by the Climatic Research Unit of the University of East Anglia. This is based on monthly values for average land and sea temperature over the northern hemisphere for 1854-1989 (136 years, 1632 data points).

The data are calculated in the form of anomalies, i.e. each month's data average is expressed as a difference from the 1950–1979 mean value for that month. This is a standard device used by climatologists to remove seasonal effects from the data, though as we shall see, it can have a misleading effect.

Fig. 5.3: Spectrum of Mauna Loa data

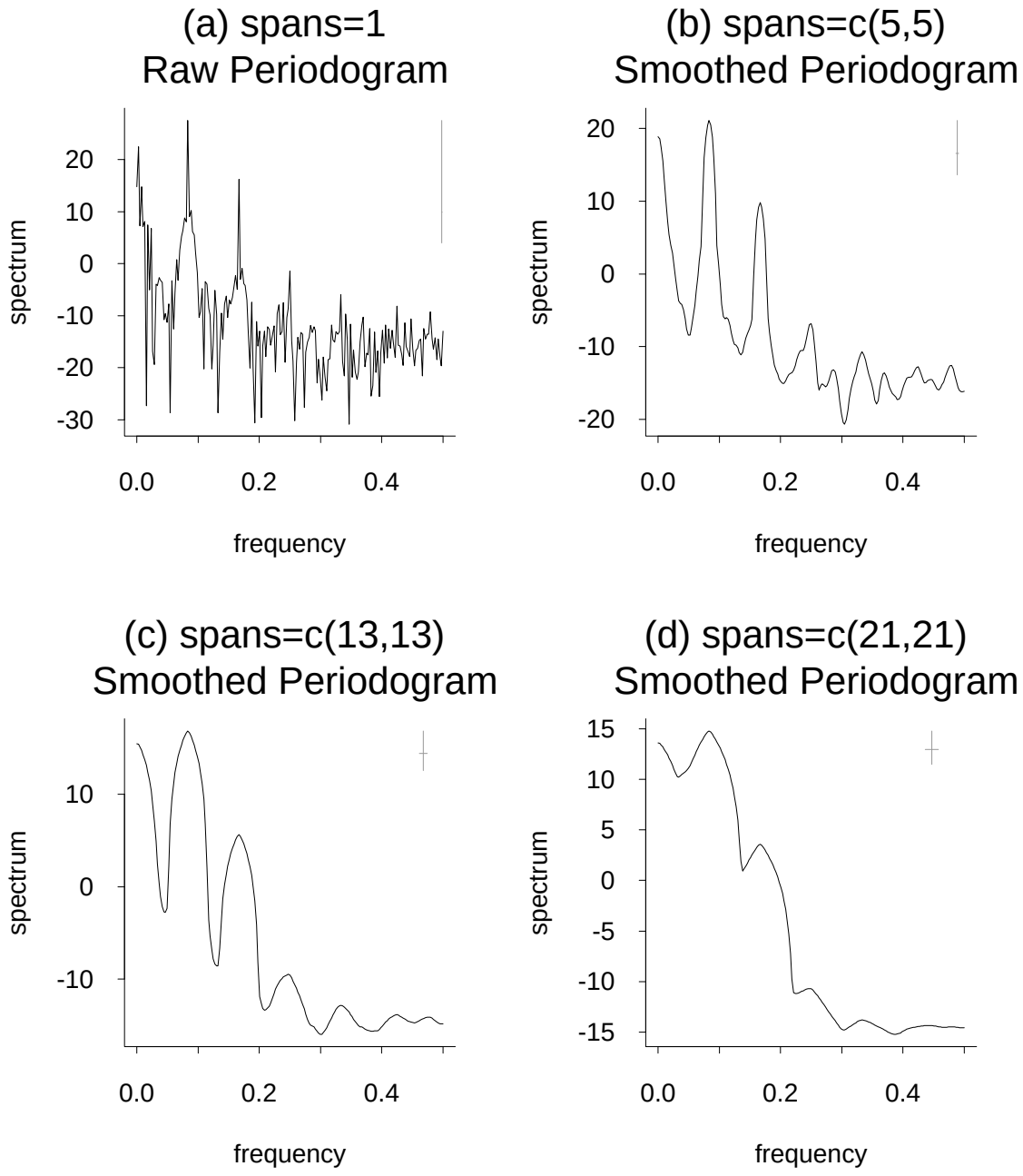


Fig. 5.4: Spectrum of detrended data

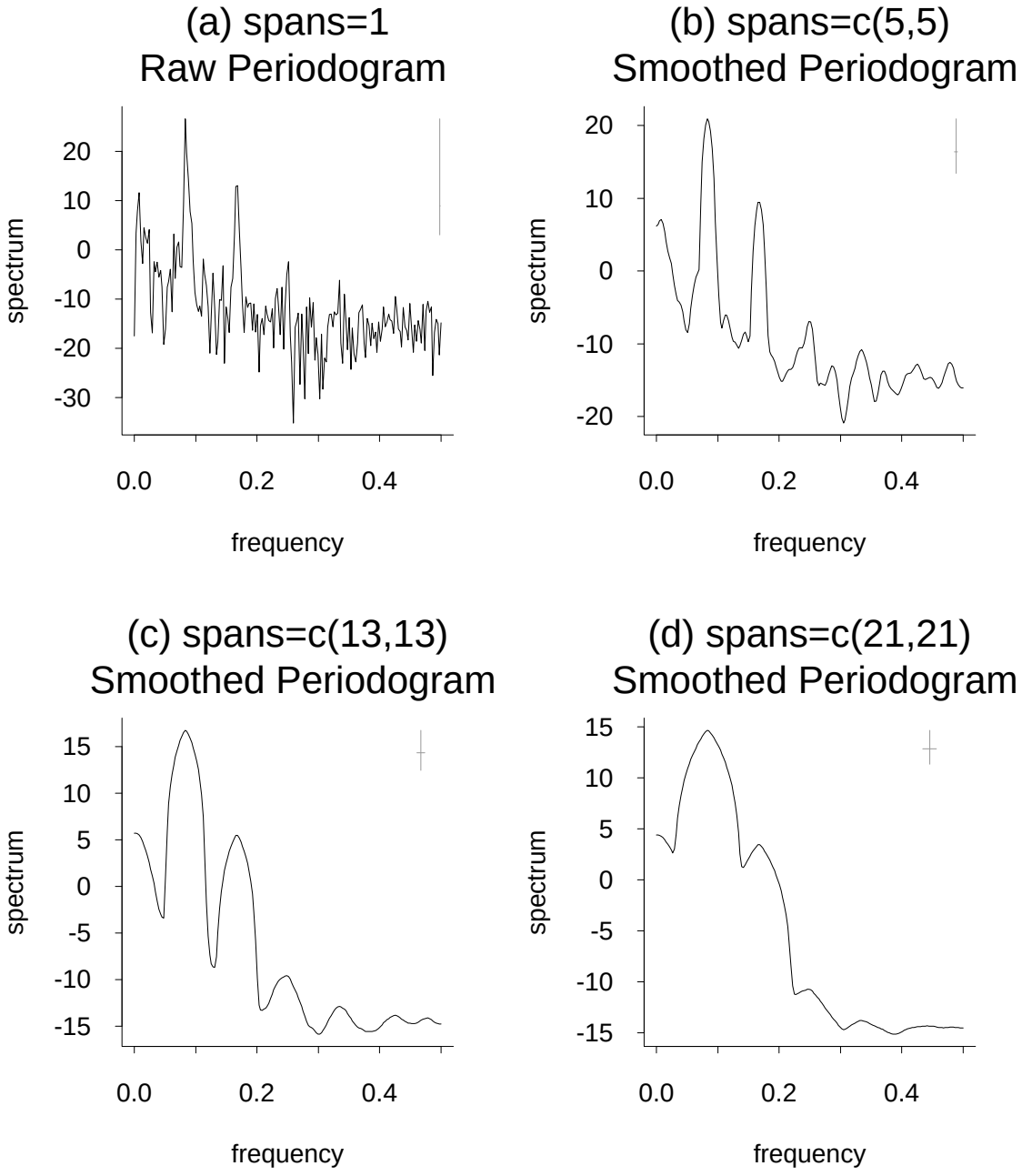
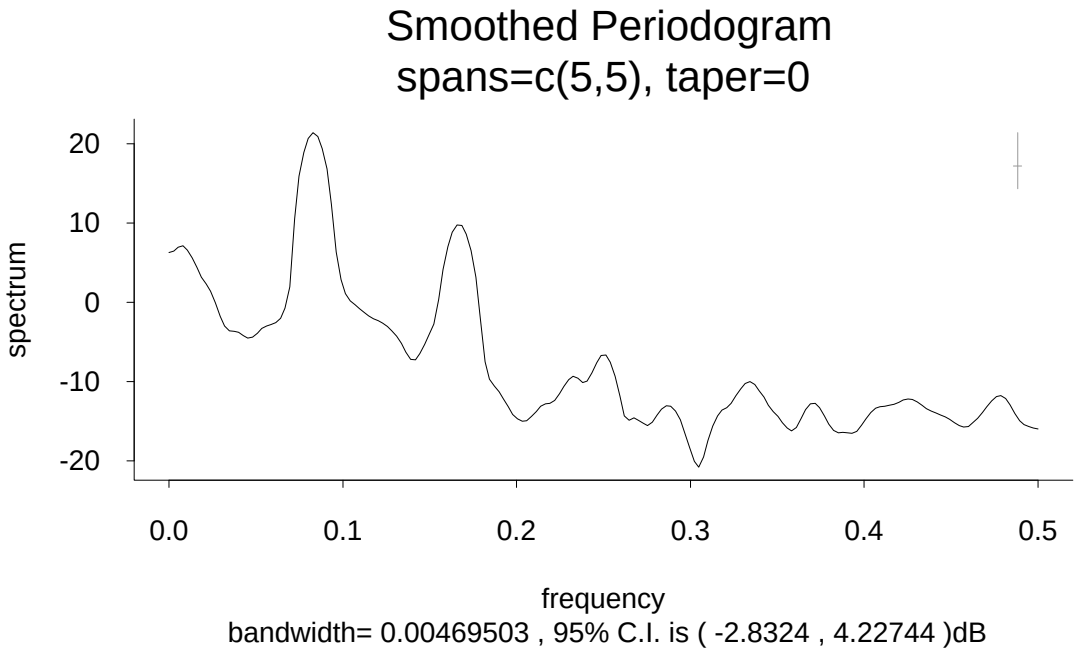


Fig. 5.5: Detrended data: Comparison of tapers



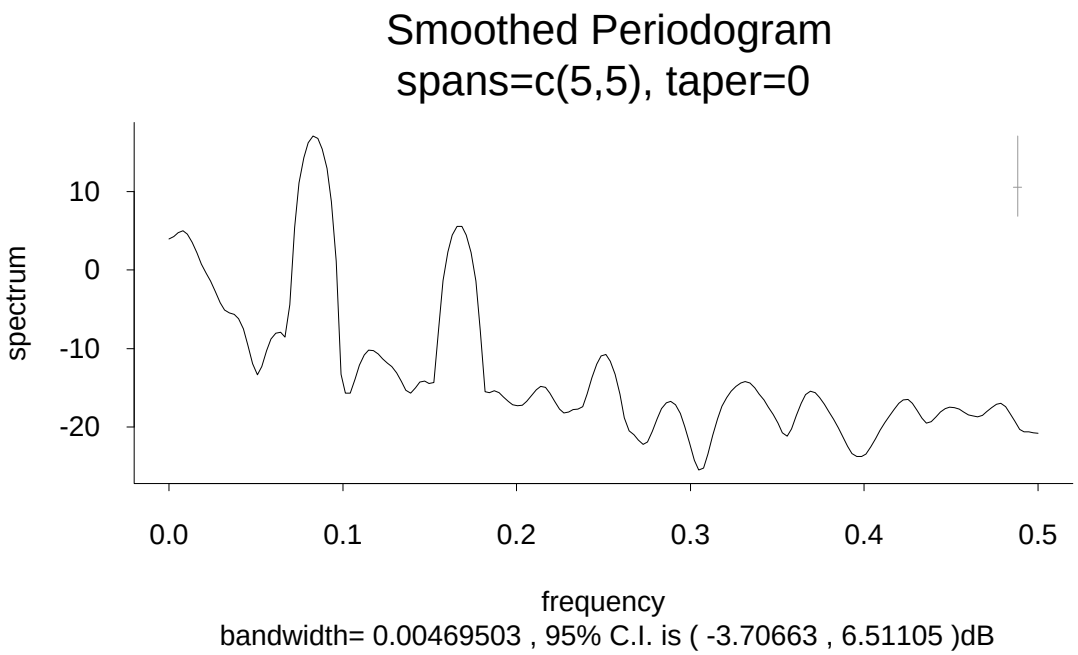
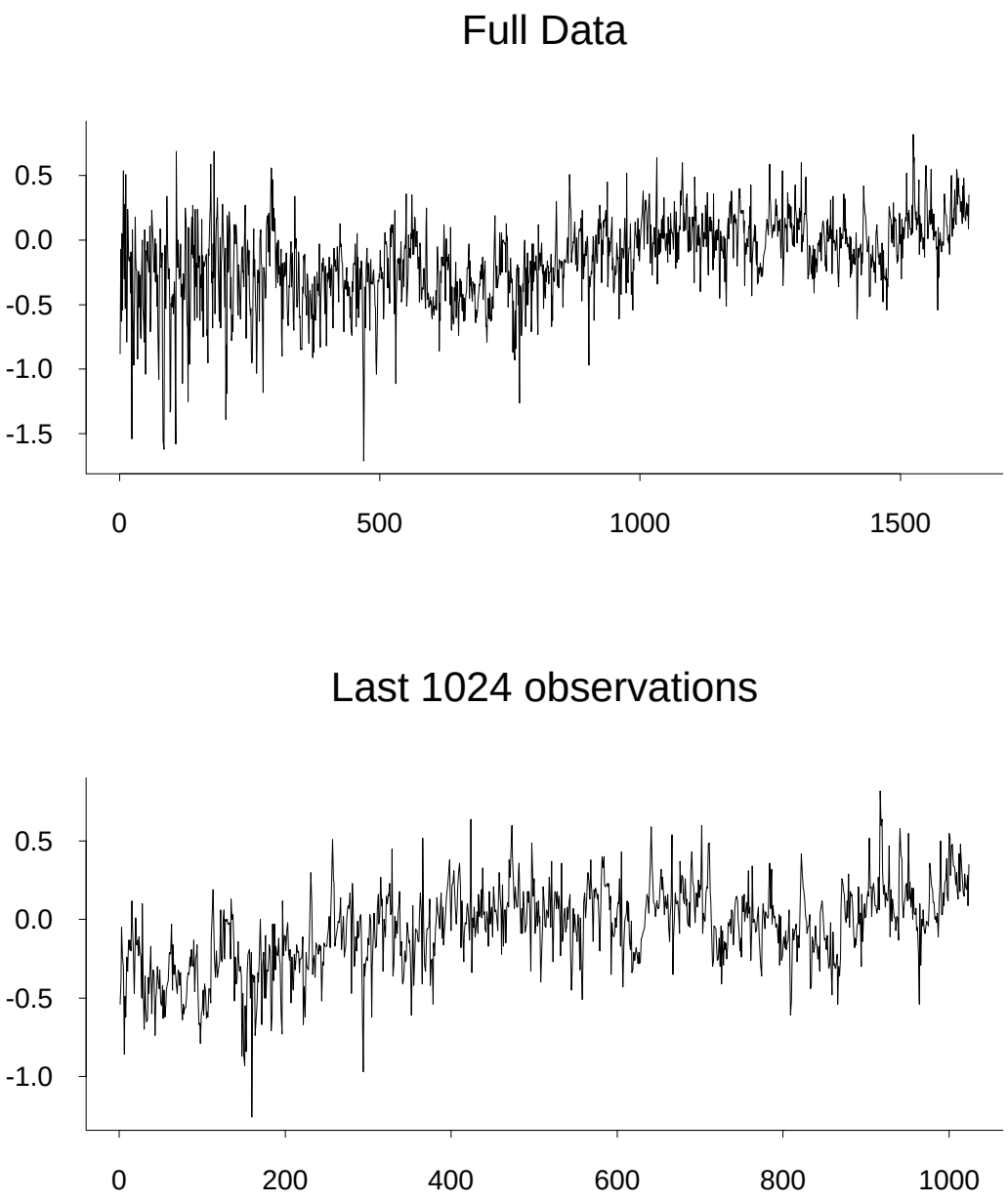


Fig. 5.6: N Hemisphere Temperatures

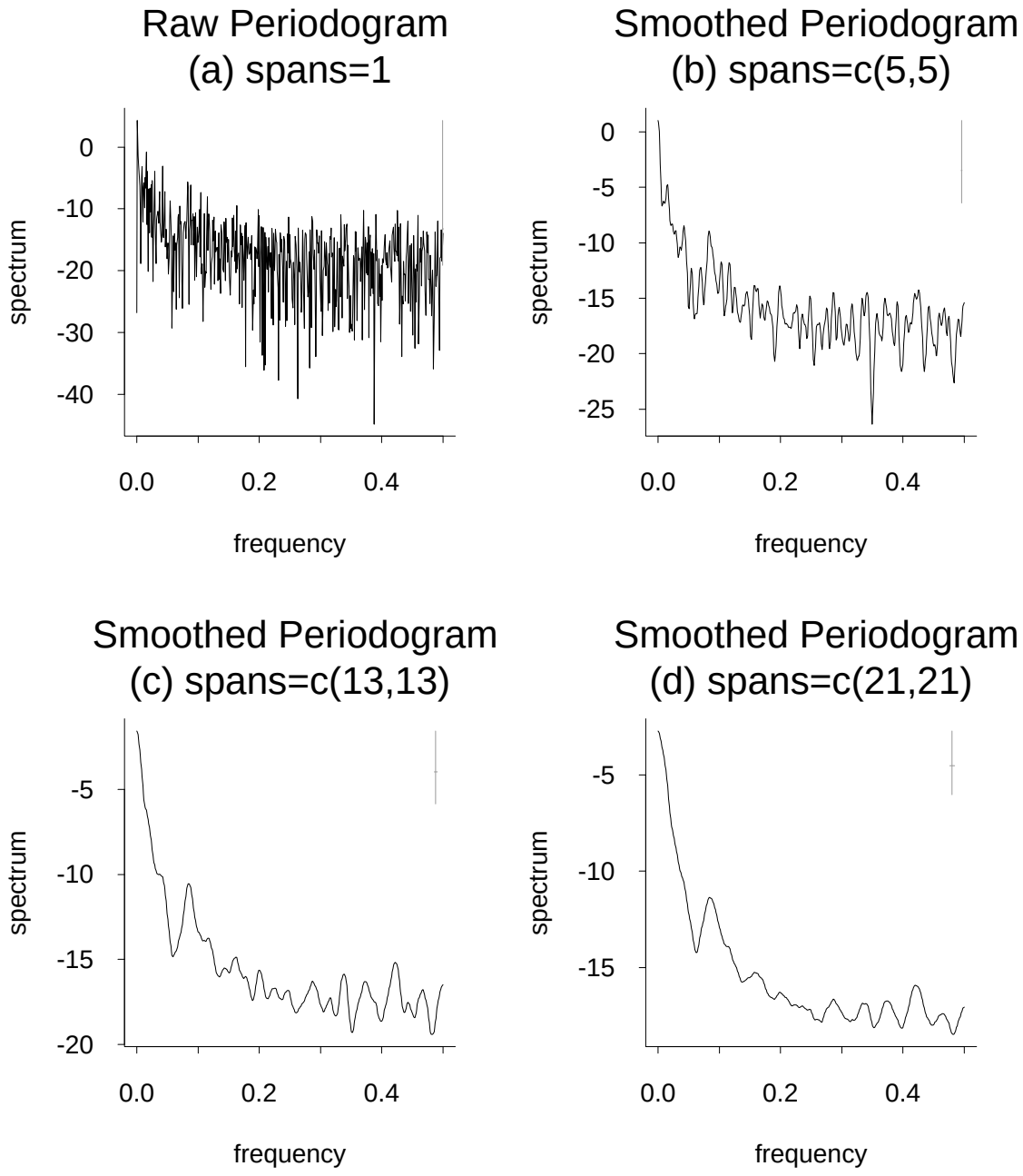


The top plot in Fig. 5.6 shows the full data series, but it is obvious that in the early years of the data, the variance is much higher than later, a consequence of the sparseness of available data in the nineteenth century. The bottom plot shows just the last 1024 observations (correponding roughly to years 1905-1989) and this is much smoother; henceforth we use just the bottom series.

In this case there seems to be visual evidence of a trend, which indeed has been the focal point of much of the current debate on global warming. Once again, residuals from a linear trend are used as the basis for a spectral plot, and Fig. 5.7 shows four such plots with different degrees of smoothing. In this case, the longer length of the series (compared with Mauna Loa) might lead one to expect to use a spectral estimate that smoothes over a larger number of frequencies, and this is confirmed by visual inspection of the plots which makes the fluctuations in (b) seem clearly insignificant – the ideal here seems to be somewhere between (c) and (d).

Two features of the spectral plots seem apparent. First, there is again a peak near 0 – which may be due to additional trend terms that have not been allowed for, but in this case there is no clear evidence for that, and the plots have alternatively been interpreted as indicating long-range dependence in the data. The second feature is that, in spite of deseasonalizing the data through the calculation of anomalies, there is still a spectral peak near frequency 1/12. This seems most likely to be the result of a differential warming between the summer and winter months.

Fig. 5.7: Temperature Spectra



6. EXAMPLES OF TIME SERIES MODELS

6.1. EEG data analysis

This section is based on data provided by Mike West and Andrew Krystal of Duke University.

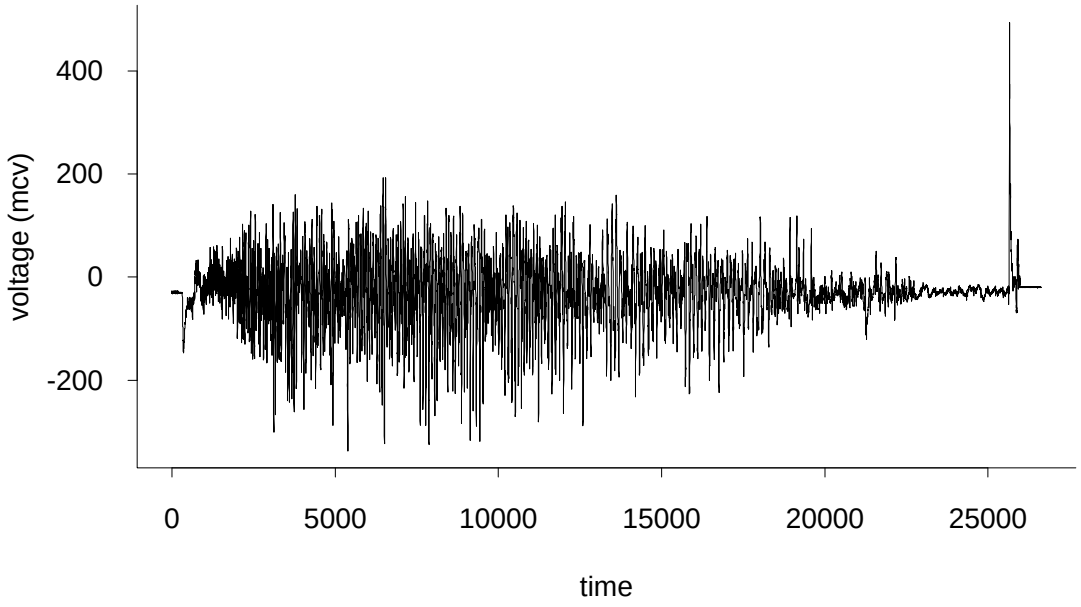


Fig. 6.1. A section of an EEG trace.

Fig. 6.1 displays recordings of an electro-encephalogram (EEG) trace on a patient in electro-convulsive therapy. This is a segment of a long EEG trace arising in a study of waveform characteristics in multi-channel EEG signal analyses. These studies are germane to assessments of differing ECT protocols. Comparison of two or more such time series underlies part of the study, and appropriate modelling of individual time series represents a starting position for comparative analyses. The data displayed represent variations in scalp potentials in micro-volts during a seizure, the frequency of data collection being 256 data points per second; thus the sampled trace represents about 100 seconds' worth of real-time data.

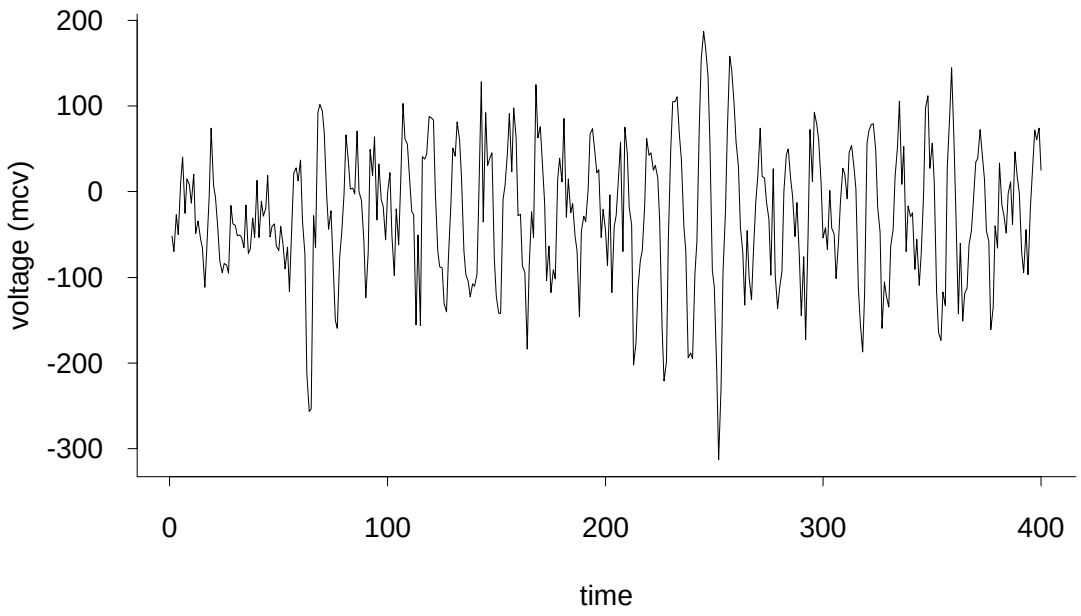


Fig. 6.2. A subset of the same data.

It is immediately obvious that the data cannot be modelled as a stationary time series. If nothing else, there are clearly wide variations in the variance of the time series. Moreover, it is extremely unlikely that such a simple device as differencing would achieve anything at all.

One approach to such a data set is to look at short sections of it to ascertain whether there are interesting features that might then be generalizable to the full data set. Another feature is that the sampling frequency is very high, and the task of analysis can be simplified somewhat by only using every k'th observation, for some $k > 1$. For the following analysis we have somewhat arbitrarily selected $k = 6$.

Fig. 6.2 shows a much shorter data set, of 400 observations, abstracted from Fig. 6.1 by picking out every sixth observation starting at observation 5,000. In this case the assumption of a stationary series does seem reasonable, so our next step is to plot the acf and pcf functions. This is done in Fig. 6.3.

The acf shows a strong sinusoidal pattern with a frequency of about 12. The pcf shows a significant positive pcf at lag 1, and a group of significant negative pcfs at lags 2–7. Thereafter the pcfs are smaller in magnitude, though it is difficult to assert that they are negligible.

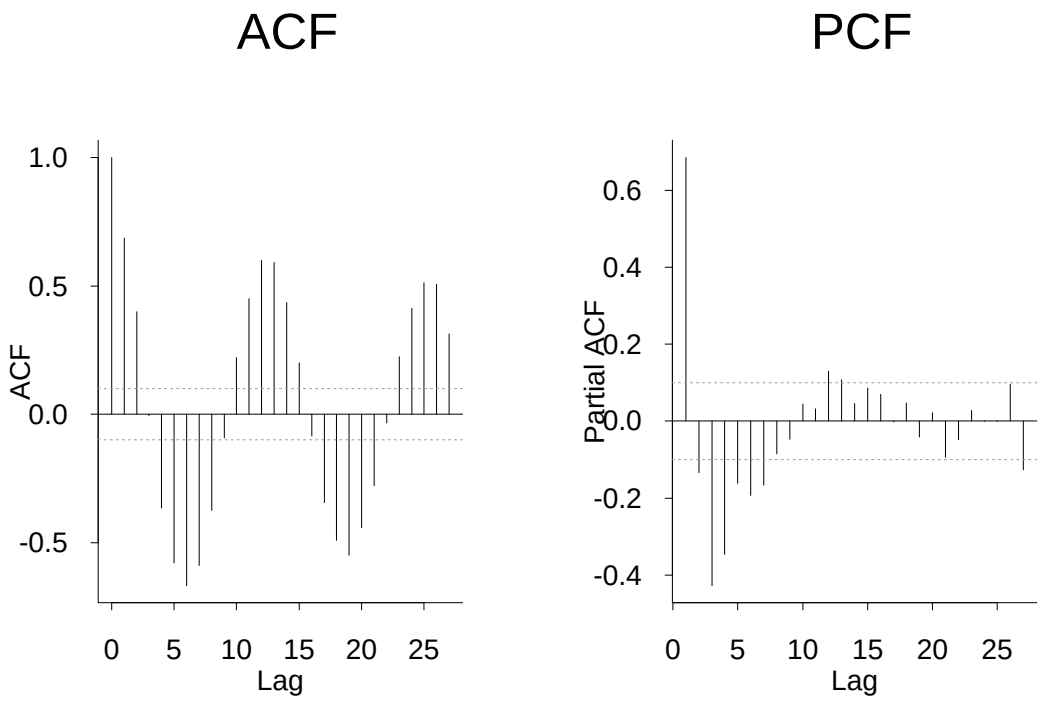


Fig. 6.3. ACF and PCF plots of the data.

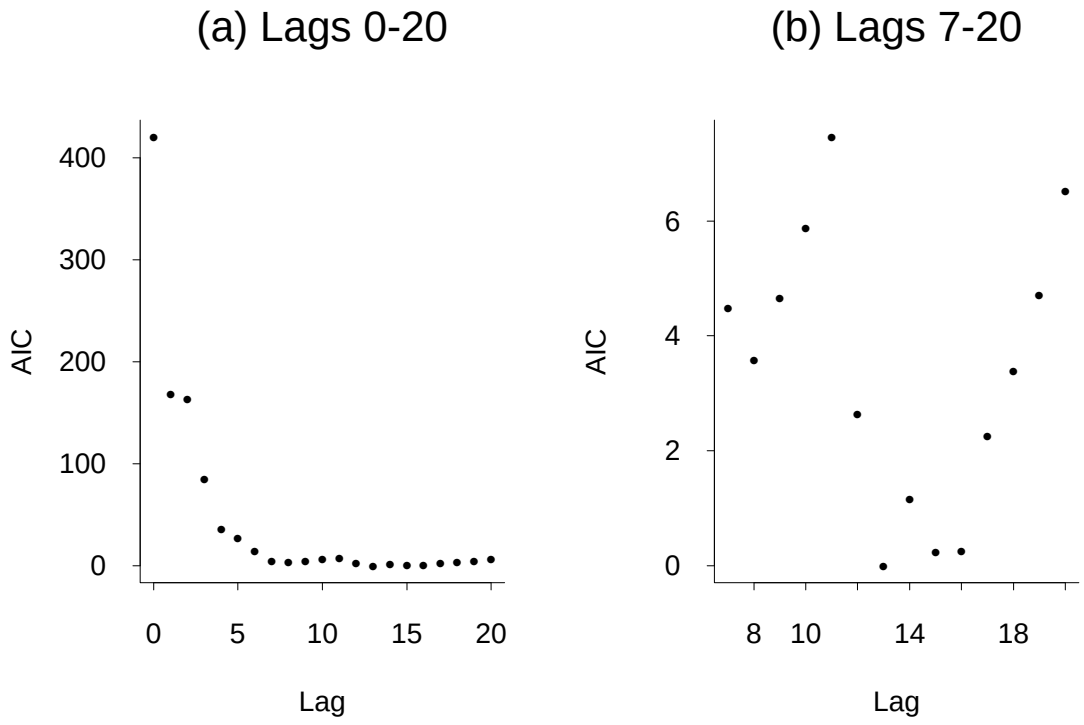


Fig. 6.4. AIC plots for AR models.

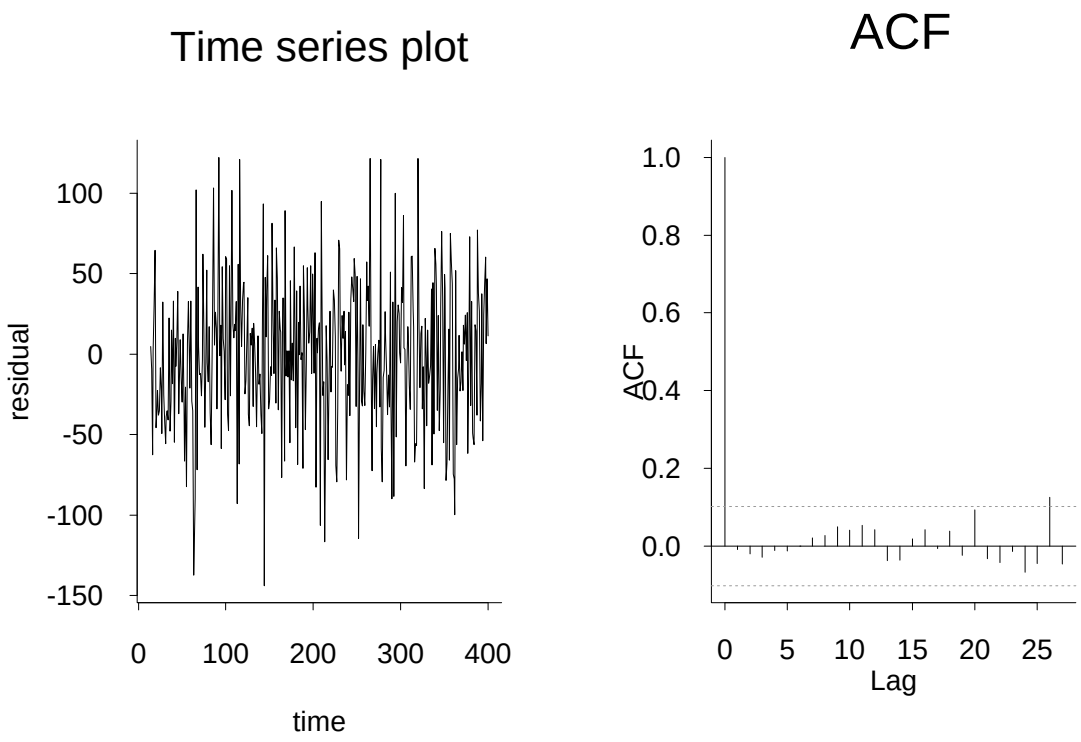


Fig. 6.5. Residuals from AR(16) model.

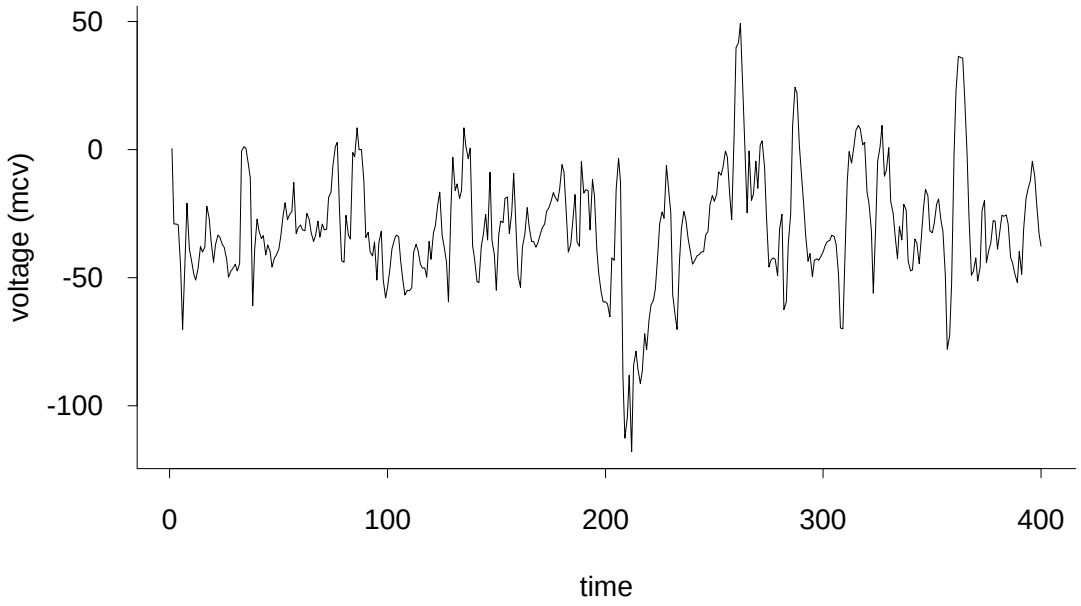


Fig. 6.6. Another subset of the same series.

The preliminary plots suggest that a high-order autoregressive process might be a reasonable fit to the data, though it is clear that we must take the order p to be at least

8. To test this assertion further, a series of AR fits was made using the `ar.yw` command in SPlus, with model selection according to the AIC criterion.

Fig. 6.4 shows the AIC values (relative to minimum value set to 0) for models with $p = 0, 1, 2, ..., 20$. However, it is clear from Fig. 6.4 (a) that the first few lags, for which the AIC is of the order of hundreds, are not reasonable models, so we concentrate on lags 7–20 as in Fig. 6.4 (b). This picks out lag 16 as the one with smallest AIC, though lag 13 is only infinitesimally worse, and any of the lags 12–17 would seem to give reasonable models.

Accordingly the model was set at AR(16). The estimated coefficients are the same as the first 16 partial autocorrelations from Fig. 6.3 (b), and the residual standard deviation is 45.4. Residuals from this model were calculated, along with their ACF, and are plotted in Fig. 6.5. The main feature of the residuals is a significant ACF at lag 26 which may be of concern.

Fig. 6.6 shows another subset of the data in Fig. 6.1, constructed in exactly the same way but this time starting from observation 20,000. There is none of the apparent periodicity of Fig. 6.2, and plots of the ACF and PCF (Fig. 6.7) do not show periodic behaviour either. There is a *slight* hint of periodicity in the ACF plot, but this is not reflected in the PCF plot, which looks very consistent with an AR(2) model. In fact the AIC criterion does pick out $p = 2$ as the appropriate order in this model, and the residuals (not shown) again look completely random. The residual standard deviation in this case is 12.4.

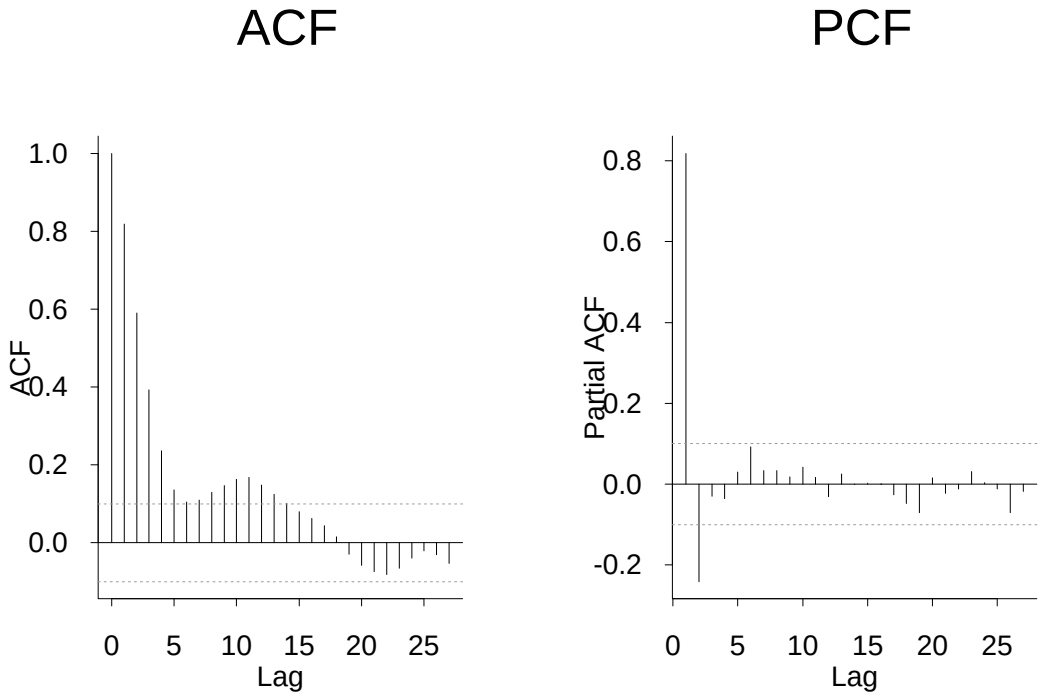


Fig. 6.7. ACF and PCF of the series in Fig. 6.6.

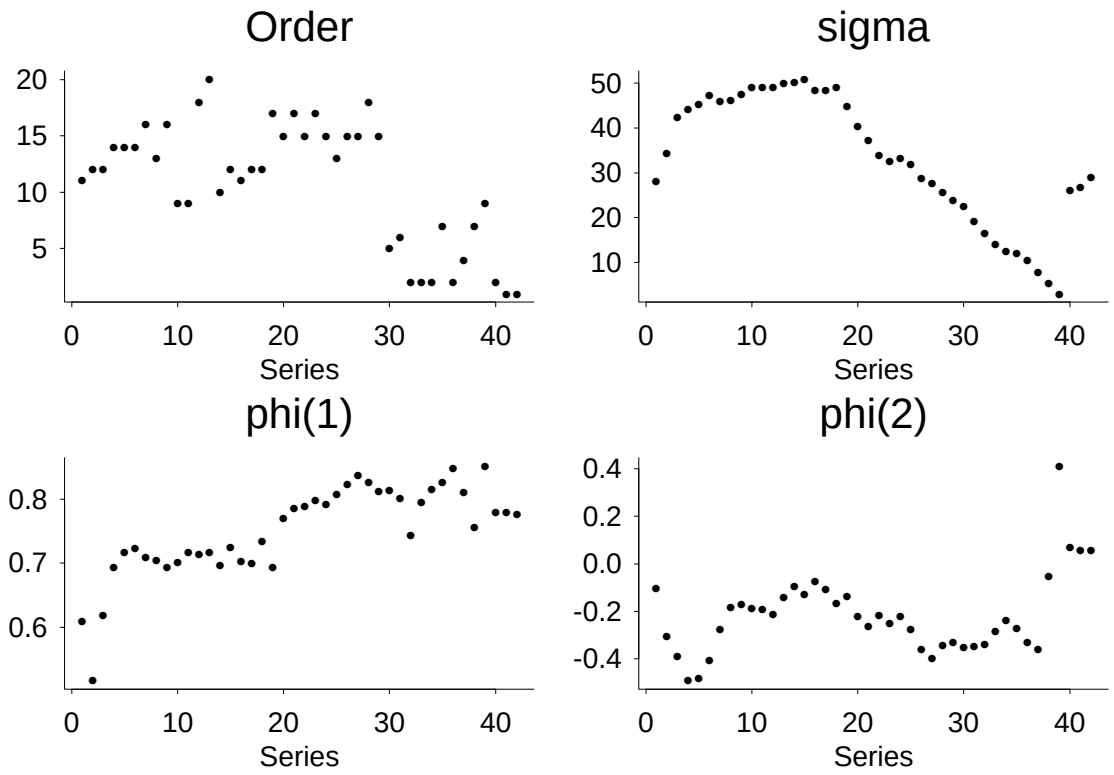


Fig. 6.8. Statistics for a sequence of subseries.

This shows a number of distinctions between the two subseries. Not only is the residual standard deviation much smaller in the second series — that much at least was evident from Fig. 6.1 — but the structure of the series is different (much lower order in the second series) and there is no evidence of any periodic behaviour.

To take the comparison further, the same analysis was repeated for 41 subseries, each constructed in exactly the same way, starting with observations 1, 601, 1201,..., (in other words, the starting point is advanced along the thinned series in lags of 100). Fig. 6.8 plots the order of the series as identified by AIC, the residual standard deviation σ, and the first two autoregressive coefficients ϕ_1 and ϕ_2, for each of the 41 subseries. The order is consistently in the range 10—15 for about the first 30 subseries, but then drops into the 2—5 range. Each of the three parameters plotted shows a systematic drift over the full length of the series expect possibly at the very end (note the impulse at the very end of the series in Fig 6.1, which explains why σ suddenly jumps up at the end in Fig. 6.8).

In conclusion, except at the very end, the series seems consistent with an autoregressive model with a steady drift of the main coefficients.

6.2. Temperature trends in Amherst, MA

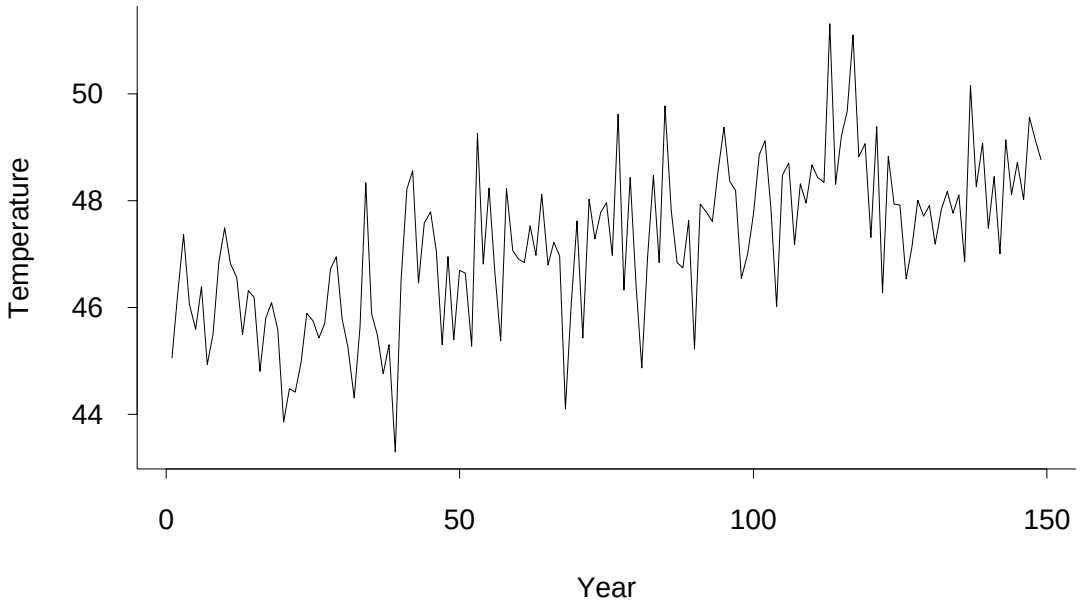


Fig. 6.9. Annual mean temperature in Amherst, MA.

The data consist of 149 years' of temperatures in Amherst, Massachusetts. The original data available consisted of monthly averages but initially we consider only annual averages, which are plotted in Fig. 6.9.

It is immediately apparent that there is a strong upward trend — unusually so for a series sampled at a single station, since most examples used to illustrate "global warming" are based on temperature averages over the whole world or at least major portions of it.

In Fig. 6.10, the ACF and PCF of the raw data series are plotted. There is no sign that the ACF is converging to zero – a typical sign that the series in nonstationary – and the PCF is also hard to characterize.

One way to deal with nonstationary series is to difference, and Fig. 6.11 shows the ACF and PCF of the differenced series. The ACF now looks much nicer, but somewhat worryingly, the PCF still shows relatively large values at large lags, which suggests that perhaps the differenced series is not very stationary either.

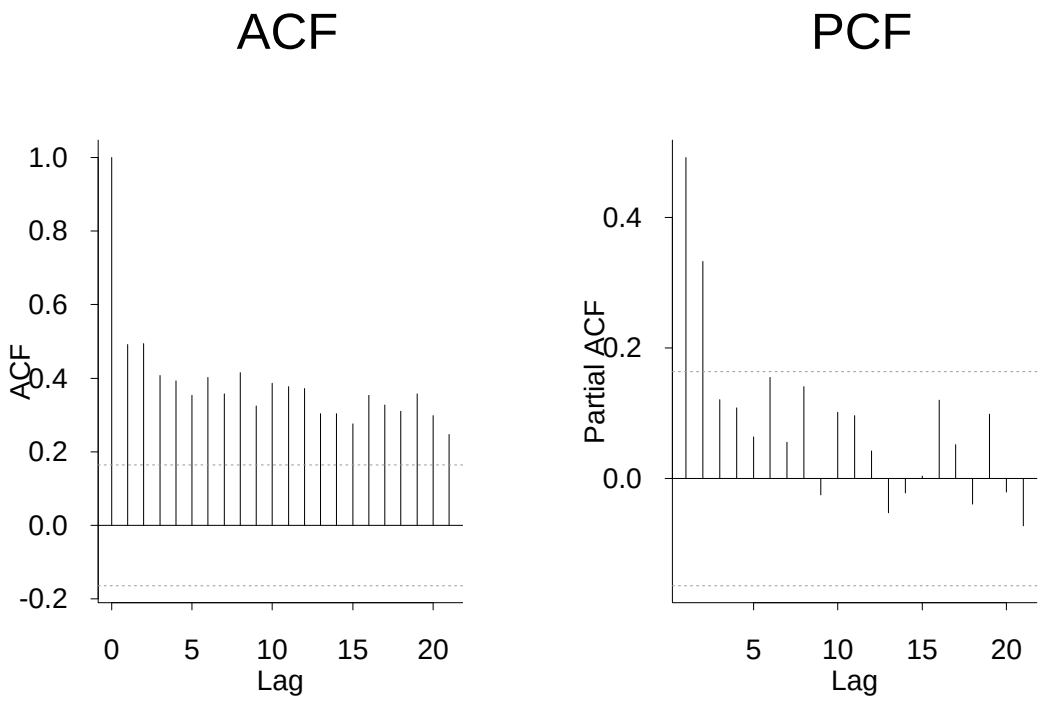


Fig. 6.10. ACF and PCF of annual Amherst data.

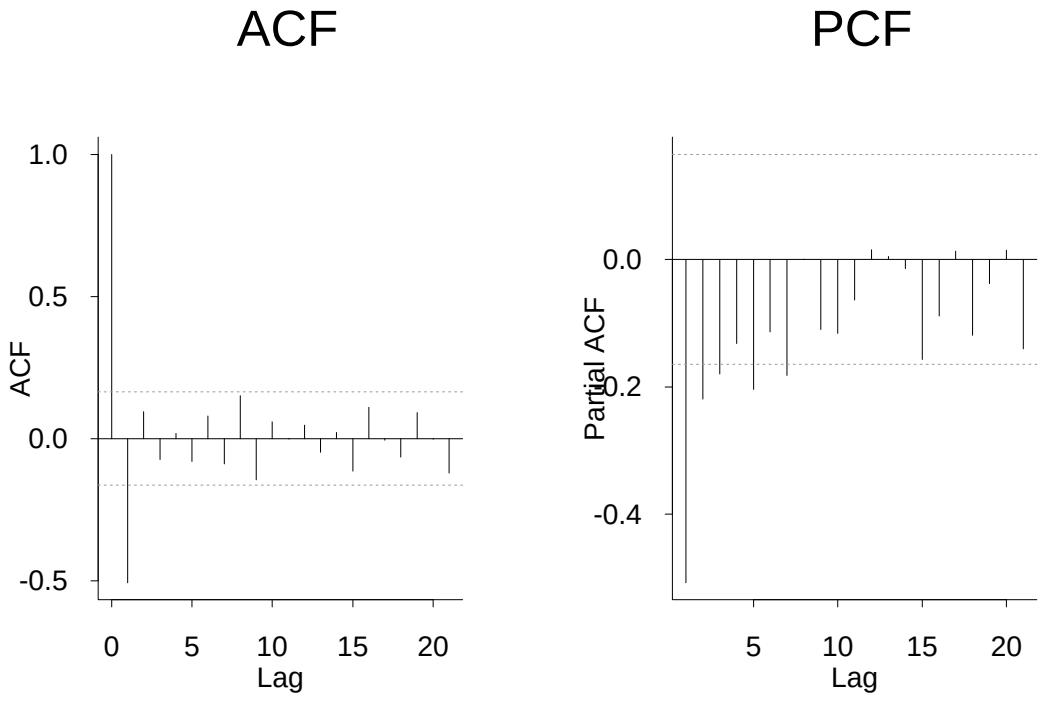


Fig. 6.11. ACF and PCF of differenced Amherst data.

An alternative method of dealing with the nonstationarity in this series is to fit a

linear trend to the observed series $\{y_t\}$ in the form

$$y_t = \alpha + \beta u_t + x_t, \quad t = 1, 2, ..., T, \tag{6.1}$$

where α and β are constants, $\{u_t\}$ is a set of fixed covariates and $\{x_t\}$ is a zero-mean residual process, which we take to be itself a time series model. In the discussion which follows we take $u_t = t - \frac{T+1}{2}$ (centred about 0).

It is possible to estimate α and β by least squares in (6.1), and then fit a time series model to the residuals $\{x_t\}$. However, a slightly more efficient approach is to estimate α, β and the time series parameters in a single shot. The `arima.mle` programme within Splus allows one to do this by the addition of a parameter `xreg`.

We need to note some caveats about `arima.mle`. The implementation of this in Splus *assumes the mean is 0.* Thus by using the arima model with $d = 0$, it is necessary *first* to subtract the sample mean from every y_t value — a point which has been noted by some authors (Venables and Ripley 1994, p. 364) but is not stressed in the SPlus manual.

This gets more complicated if one fits ARIMA(p, d, q) with $d > 0$ because SPlus *still* assumes that the mean of the differenced series is 0. To get around this with, say, $d = 1$, it is necessary to replace each y_t with $y_t - \bar{y} - \bar{\beta} u_t$ where $\bar{y} = \frac{1}{T}\sum y_t$, $u_t = t - \frac{T+1}{2}$ as above and

$$\bar{\beta} = \frac{y_T - y_1}{T - 1}. \tag{6.2}$$

In this way, as may be checked directly, the sample mean of $\{y_t - y_{t-1},\ 2 \le t \le T\}$ is adjusted to 0. However, this adjustment should not be made when fitting with $d = 0$ as then removing the slope has a real effect on the model! Because of this, the application of `arima.mle` with different d is a more complicated business than it may appear at first sight.

A typical format of the `arima.mle` command is as follows:

```
z<-arima.mle(y,n.cond=5,model=list(order=c(1,0,2)))
```

Here `y` is the data series to which the model is fitted, `n.cond` is the number of initial values on which the likelihood fit is conditioned (m in our earlier notation) and the order of the model is specified by (1,0,2) representing p, d and q respectively — thus, the command above would fit an ARMA(1,2) model. The output `z` contains all the details of the fitted model — for example, `z$model` lists all the parameters, `z$var.coef` the estimated variance-covariance matrix, `z$loglik` is twice the negative log likelihood, `z$AIC=z$loglik+2(p+q)` is the AIC value, and so on. Another useful variable to check is `z$converged` — beware if this is `F`!

To modify this to fit the model (6.1), one could use one of the commands

```
z<-arima.mle(y,n.cond=5,model=list(order=c(2,0,0)),xreg=u
```

or

```
z<-arima.mle(y,n.cond=5,model=list(order=c(2,0,0)),xreg=cbind(1,u))
```

where `u` is the $\{u_t\}$ variable mentioned above — in this case, we are assuming the structure of $\{x_t\}$ is AR(2) but of course the p and q parameters could be modified to allow other models. The first of these commands fits just one regressor u_t while the second also treats the constant 1 as a regressor. The second of these forms is the most comprehensive model of all the ones considered in this section, but to allow direct comparability with the earlier fits (where we simply estimated the mean by $\bar{y}$), the first form (with `xreg=u`) is the one adopted here. This issue is of no practical importance for the results obtained.

For the data set in question, 15 models were tried: five with $d = 0$ and no `xreg`, five with $d = 1$ (incorporating the adjustment noted above) and five with $d = 0$ and `xreg`. The results were as follows:

(p, d, q)	xreg?	converged?	Loglik	AIC
(1,0,0)	No	T	479.5823	481.5823
(0,0,1)	No	T	497.3821	499.3821
(2,0,0)	No	T	460.7184	464.7184
(1,0,1)	No	T	450.5513	454.5513
(1,0,2)	No	T	478.7780	482.7780
(1,1,0)	No	T	476.8604	478.8604
(0,1,1)	No	T	449.6190	451.6190
(2,1,0)	No	T	469.6354	473.6354
(1,1,1)	No	T	447.8834	451.8834
(0,1,2)	No	T	447.8317	451.8317
(1,0,0)	Yes	T	441.4168	445.4168
(0,0,1)	Yes	T	442.2024	446.2024
(2,0,0)	Yes	T	437.8472	443.8472
(1,0,1)	Yes	T	439.4146	445.4146
(0,0,2)	Yes	T	437.9397	443.9397

Table 6.1 Summary of models for annual Amherst data

A first glance at this table indicates that, judged by AIC, the models incorporating a linear trend are the best; for those without a trend, the models with $d = 1$ improve on those with $d = 0$ but are still considerably worse than those with a trend.

In a little more detail, the best of the $d = 0$ models without a trend is the ARMA(1,1) model, but the coefficients for this are very close to the boundary of the stationarity region:

$\hat{\phi}_1 = 0.983$ with standard error 0.018 ($= \sqrt{0.00034}$, the latter being the leading entry of the `z$var.coef` matrix), and $\hat{\theta}_1 = 0.833$ (standard error also 0.018). In particular, $\hat{\phi}_1$ is not significantly different from the boundary value 1. Amongst the $d = 1$ models, the best appears the ARIMA(0,1,1) model but this is not as good as judged by log likelihood or AIC as the best of the trend-based models, which appears to be the AR(2). This model has AR coefficients $\hat{\phi}_1 = 0.124$ and $\hat{\phi}_2 = 0.156$, each with standard error 0.082. The residual standard deviation is $\hat{\sigma} = 1.22$, which improves on the $\hat{\sigma} = 1.32$ for the ARIMA(0,1,1) model (this is one concrete indication of the superior fit of the trend+AR(2) model). Finally, application of the `arima.diag(z)` command produces the diagnostic plots shown in Fig. 6.12.

This plot shows the standardized residuals, the ACF, and the P-value of the Box-Pierce statistic, computed for various values of K (the total number of lags used in computing the statistic). All the P-values are quite close to one (whereas it is small values, $< .05$ for instance, which would give us cause for concern). Thus the evidence is that the model fits the data very well.

6.2.1. Estimating the trend

At this stage we have three estimators for the trend coefficient β: the ordinary least squares estimator ($\tilde{\beta}$ say), $\bar{\beta}$ given by (6.2), and the MLE $\hat{\beta}$. In this example their values are $\tilde{\beta} = .00219$, $\bar{\beta} = .00251$, $\hat{\beta} = .00226$. There is not much of an argument for using $\bar{\beta}$ as an estimator of β, although we should point out that if we follow the recommendations of some authors and use the mean of the differenced series as an estimator of trend, then $\bar{\beta}$ is the estimate obtained. As far as the comparison between $\tilde{\beta}$ and $\hat{\beta}$ is concerned, classical studies going back to papers by Grenander in the 1950s shows that there is very little difference between them as point estimators, as indeed appears to be the case in this example. However, where it is necessary to be careful is in estimating the standard error of this parameter. The classical formulae for linear regression give $\tilde{\beta}$ a standard error of .0022 (a t-value of 10.2) but this of course ignores the correlation — one would expect this standard error to be too small and hence the t-value too large.

Unfortunately when `arima.mle` is applied using the `xreg` option, the programme does not give the standard error of $\hat{\beta}$. One can deduce the t-value in an indirect way, however, as follows. Consider the AR(2) model with trend and compare it to the corresponding model without trend — line 3 in Table 6.1. The difference in Loglik values is 460.7184–437.8472=22.87 so under a standard likelihood ratio test, recalling that the SPlus value of `loglik` is actually –2 log L, this is to be compared with the 95% or 99% point of a χ^2_1 distribution (and so is, of course, overwhelmingly significant). However, in this context, it follows from standard likelihood theory that the χ^2_1 value for a one-parameter likelihood ratio test is approximately the square of the t value for a Wald test — in other words, the latter would be approximately $\sqrt{22.87}$ or 4.78. By this indirect argument we deduce that the true t-value for the β parameter in this model is about 4.8, and not 10.2 as quoted above. In this particular example, the correction does not affect our conclusion that β is

statistically significant, but in many similar examples the difference between modelling the correlations and ignoring them is critical to the final results!

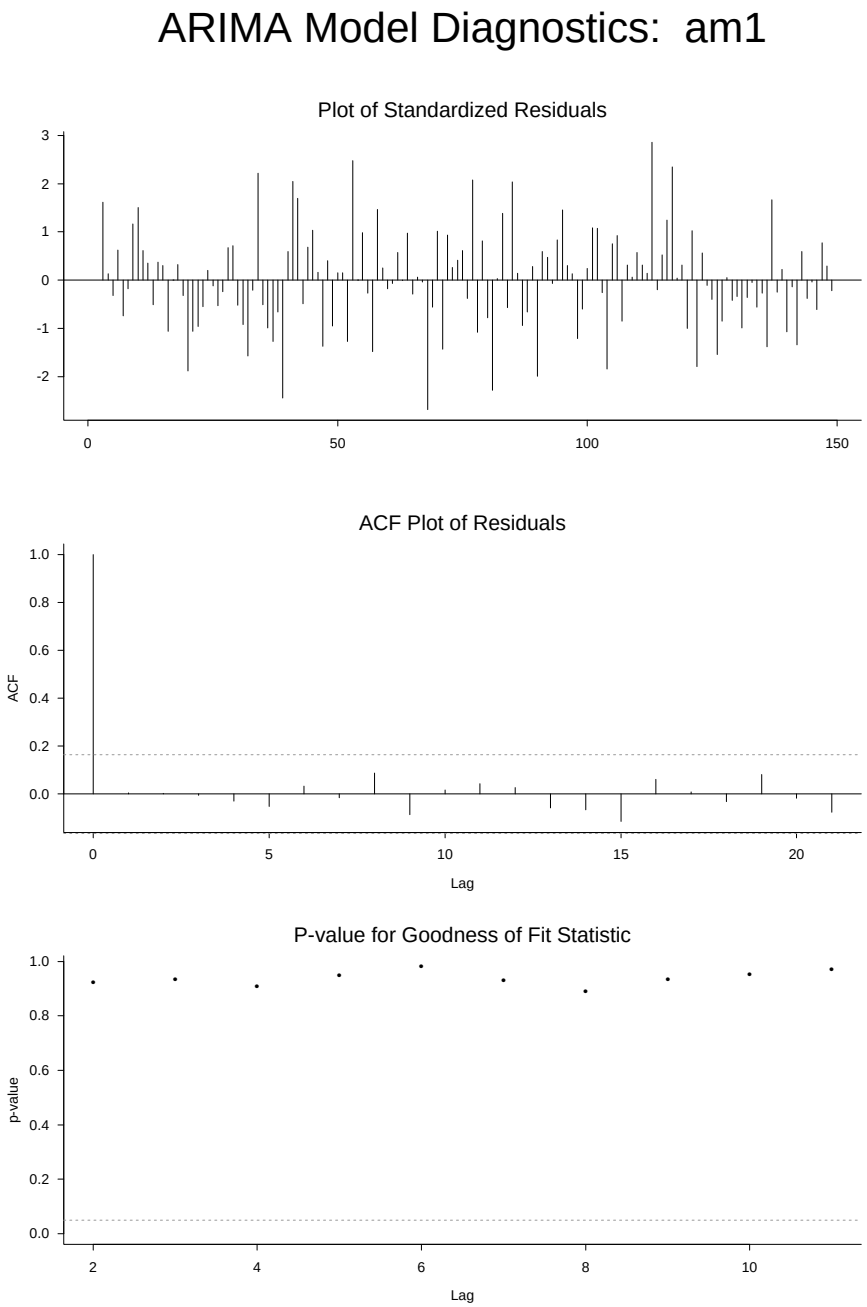


Fig. 6.12. Diagnostics from final model.

There are more direct ways of computing standard errors, especially for the least squares estimator $\tilde{\beta}$. For instance, from the formula

$$\tilde{\beta} = \frac{\sum y_t u_t}{\sum u_t^2}$$

(we are assuming $\sum u_t = 0$), we deduce

$$\mathrm{Var}\{\tilde{\beta}\} = \frac{\sum_{s=1}^{T} \sum_{t=1}^{T} u_s u_t \gamma_{|s-t|}}{\{\sum u_t^2\}^2} \tag{6.3}$$

where $\{\gamma_k\}$ denotes the autocovariance function.

One could try to evaluate (6.3) using the sample autocovariances, but this is generally considered to be a bad idea because the sample autocovariances have rather poor sampling properties. A better idea is to substitute the theoretical γ_k's based on the fitted model. In this case the fitted model is AR(2) with $\phi_1 = .12424, \phi_2 = .15643$ and residual standard deviation $\sigma_\epsilon^2 = 1.22474$. In this case it may be checked (exercise for the reader!) that $\gamma_k = .77813 \times .46248^k + .50516 \times (-.33824)^k$ for all $k \geq 0$. Direct evaluation of (6.3) then yields that the estimated standard deviation of $\tilde{\beta}$ is .0029. Note that this is substantially smaller than the above-quoted approximate standard error of .0047 for $\hat{\beta}$, which seems to point to a failure of the asymptotics somewhere along the line, since from a theoretical point of view, the standard deviation of $\hat{\beta}$ should be smaller than that of $\tilde{\beta}$.

However, this is not the only way to evaluate (6.3). From the formula

$$\gamma_k = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{i\lambda k} f(\lambda) d\lambda,$$

where $f(\lambda)$ is the spectral density, we see that

$$\begin{aligned} \frac{\sum_{s=1}^{T} \sum_{t=1}^{T} u_s u_t \gamma_{s-t}}{\{\sum u_t^2\}^2} &= \frac{\sum_{s=1}^{T} \sum_{t=1}^{T} u_s u_t \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{i\lambda(s-t)} f(\lambda) d\lambda}{\{\sum u_t^2\}^2} \\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} U(\lambda) f(\lambda) d\lambda \end{aligned} \tag{6.4}$$

where

$$U(\lambda) = \frac{\left|\sum_{t=1}^{T} u_t e^{i\lambda t}\right|^2}{\{\sum u_t^2\}^2}. \tag{6.5}$$

Since for the AR(2) model,

$$f(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \left| \frac{1}{1 - \phi_1 e^{i\lambda} - \phi_2 e^{2i\lambda}} \right|^2,$$

the integrand in (6.4) is easily evaluated and the integral itself may then be found by numerical integration. The functions $U(\lambda)$ and $f(\lambda)$ are depicted in Fig. 6.13.

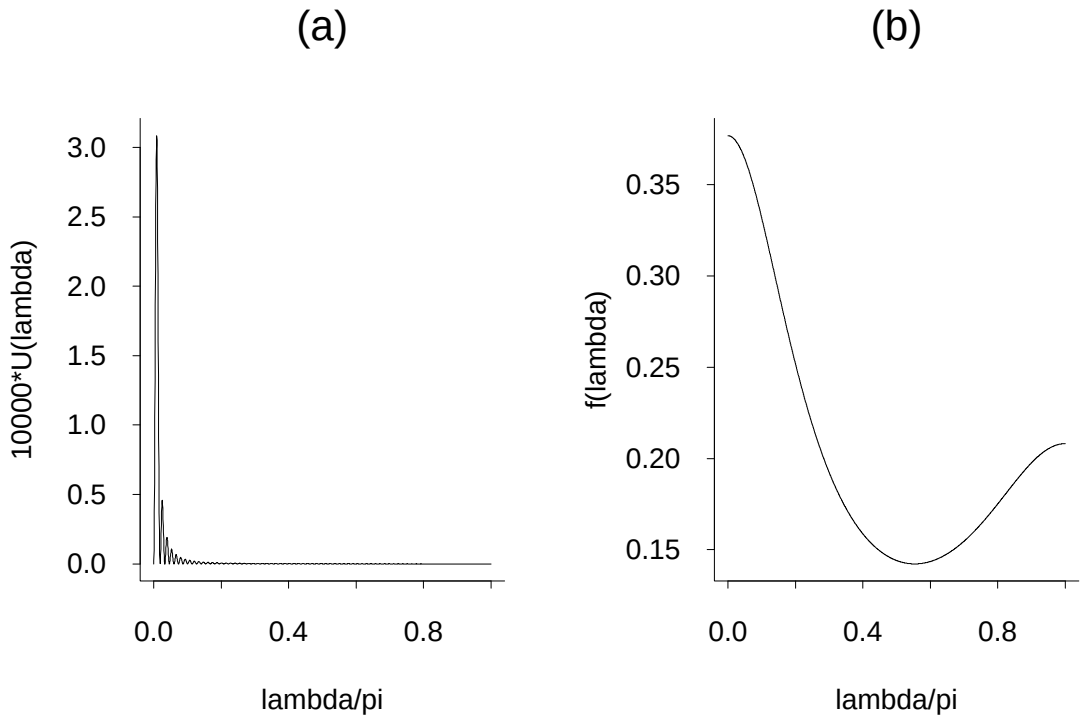


Fig. 6.13. Illustration of (6.4): (a) $U(\lambda)$, (b) $f(\lambda)$, using fitted AR(2) model, for the annual Amherst data.

For high-order ARMA models, it is easier to calculate the spectral density than the autocorrelations. This makes (6.4) easier to use in practice than (6.3). For example, sometimes spectral densities are estimated by fitting a high-order AR model to the data and plotting the theoretical spectral density of the fitted model. This is done in Fig. 6.14(a) below. Alternatively, of course, one may use nonparametric spectral estimates as in Chapter 5. One such estimate is plotted in Fig. 6.14(b). Both of these plots were based on the detrended series in which both the mean and the linear trend had been removed from the data. In this case, it is not at all clear that there is any reason to depart from the parametric AR(2) fit, and indeed, the dissimilarity between the spectral densities in Fig. 6.14(a) and (b) serves as a warning of overfitting. The resulting standard errors of $\tilde{\beta}$, however, are very similar — .0029 using the AR(2) model, .0030 using AR(8), .0026 using the nonparametric spectral density estimate.

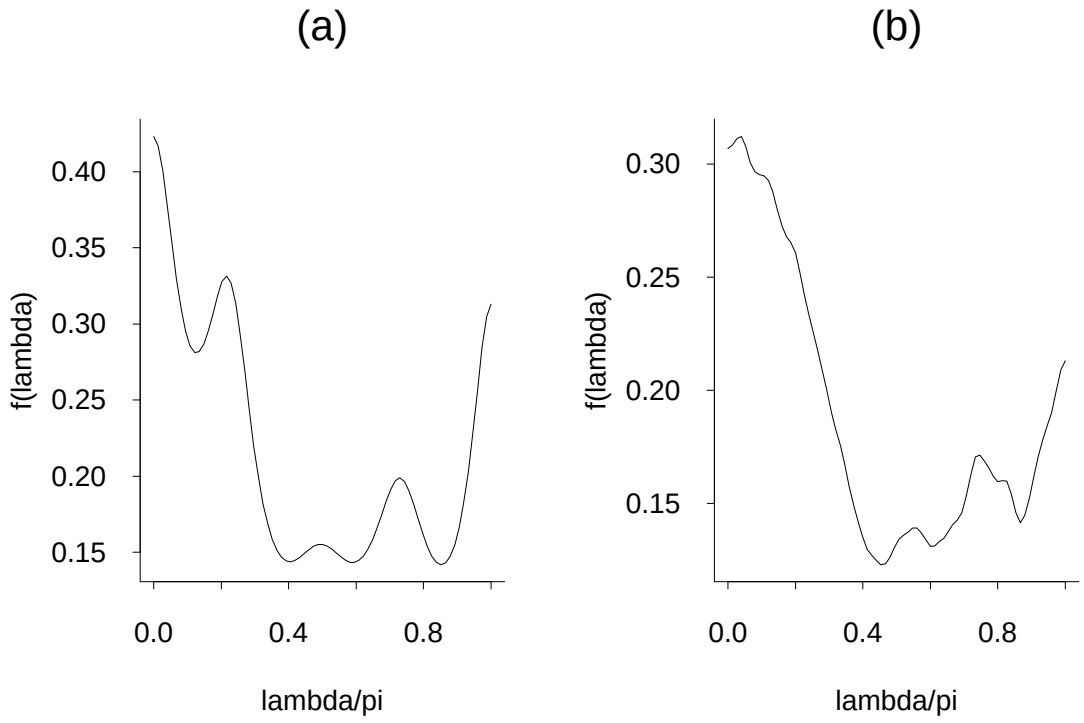


Fig. 6.14. Estimated spectral densities under (a) AR(8) fit, (b) Nonparametric smoothed estimate.

6.3 Seasonal analysis of the Amherst temperature data

As mentioned in the previous section, the original data set from which Fig. 6.9 was plotted consist of monthly averages, a total of 1788 observations. It is therefore natural to use the whole of this data set and to fit a seasonal model.

The series was standardized by removing the overall mean and a linear trend (since we already know the latter is present) and the ACF and PCF computed (Fig. 6.15). These plots actually look a lot like Fig. 6.3 for the EEG data, which suggests we should again use a high-order AR(p) model. In fact AIC selects $p = 26$ but this approach does not remove the difficulty which we are about to identify in the automatic fitting of ARMA models to this series.

In this case, a sequence of seasonal ARIMA models was fitted. A typical command in Splus might be

```
> z<-arima.mle(x,n.cond=48,model=list(list(order=
+ c(0,0,2)),list(order=c(1,0,0),period=12)))
```

This fits the model which, in the notation of (4.7) of the main notes, has $\theta(B) = 1 + \theta_1 B + \theta_2 B^2$, $\Theta(B) = 1$, $\phi(B) = 1$, $\Phi(B) = 1 - \Phi_1 B^{12}$. This particular model was

successfully fitted (`z$converged=T`) with a `loglik` value of 10031.66 and AIC of 10037.66. The luxury of a relatively large `n.cond` seems affordable in view of the total length of the series.

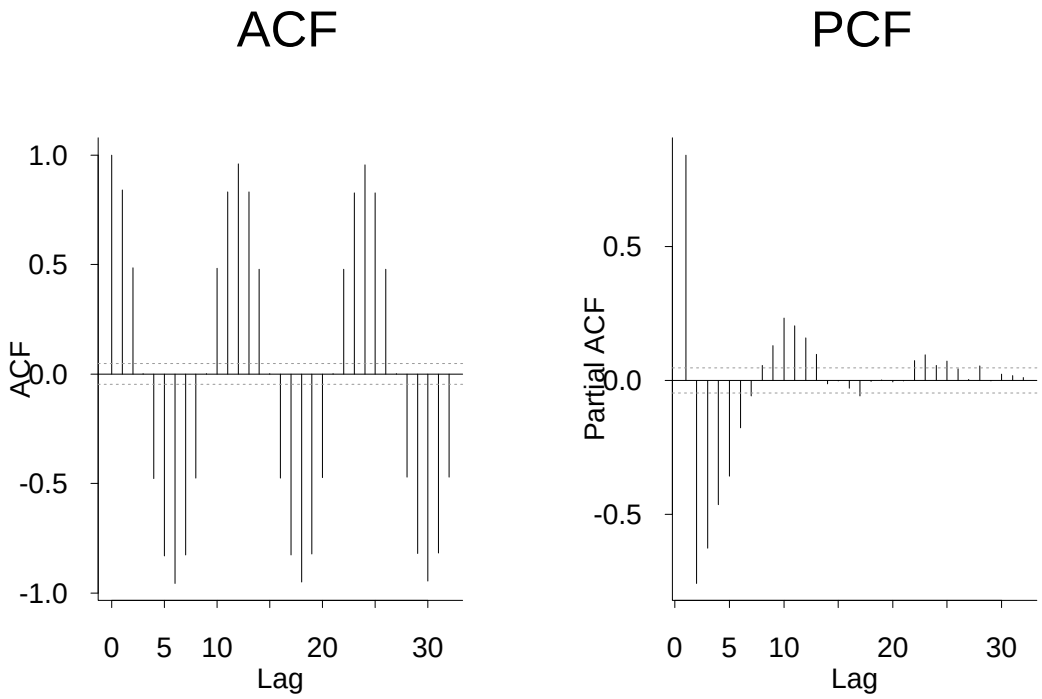


Fig. 6.15. ACF and PCF for monthly data.

Nonseasonal Component	Seasonal Component			
	None	(1,0,0)	(0,0,1)	(2,0,0)
(1,0,0)	12687.53	10040.27	11716.97	9559.978
(0,0,1)	13241.32	10048.26	12227.67	9562.820*
(2,0,0)	11137.40	10033.46	11096.45*	9558.905*
(1,0,1)	12081.16	10033.31	11536.65*	9557.585*
(0,0,2)	12307.51	10037.66	11655.63*	9561.175*

Table 6.2 AIC for seasonal ARIMA models (data with linear trend removed). *=no convergence as represented by the `$converged` variable.

In fact for the nonseasonal (p, d, q)(=(0,0,2) in the example just given), every combination of (1,0,0), (0,0,1), (2,0,0), (1,0,1) and (0,0,2) was tried. For the seasonal (P, D, Q), the combinations were (1,0,0), (0,0,1), and (2,0,0), as well as nonseasonal models in which this component was absent. Thus twenty models in all were tried, and the results are

shown in Table 6.2. Unfortunately, quite a few of these model fits (shown by asterisks in the table) resulted in `z$converged=F`. The best `loglik` or AIC models resulted from $(P, D, Q) = (2, 0, 0)$, but in this case the only one of the (p, d, q) combinations to give `z$converged=T` was (1,0,0). For this model the AIC value was 9560.0, which is clearly better than the above, but unfortunately this model still does not fit the data. The problem is shown by the `arima.diag` diagnostics, plotted in Fig. 6.16.

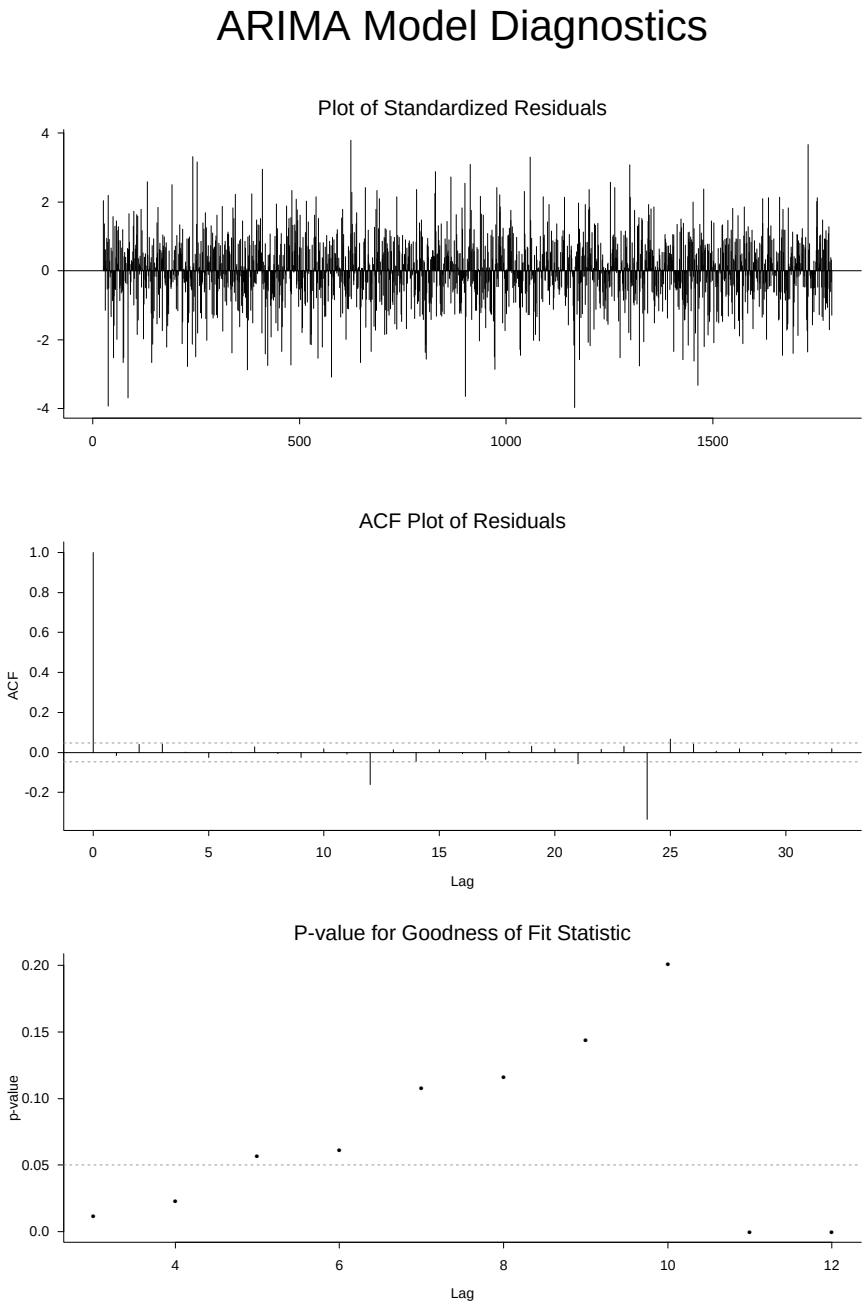


Fig. 6.16. Diagnostics from seasonal ARMA model.

The ACF plot of the residuals shows strong negative autocorrelations at lags 12 and 24, indicating that the seasonal effect has not been removed. The (Box-Pierce) goodness of fit statistic confirms this, with P values well below 0.05 at $K = 11$ or 12. The PCF plot (not shown) is even worse, with strongly significant negative correlations at lags 12, 24, 36, 48... Moreover the same problem was found with every other seasonal ARIMA model tried. At this stage, the seasonal ARIMA idea does not seem to be working.

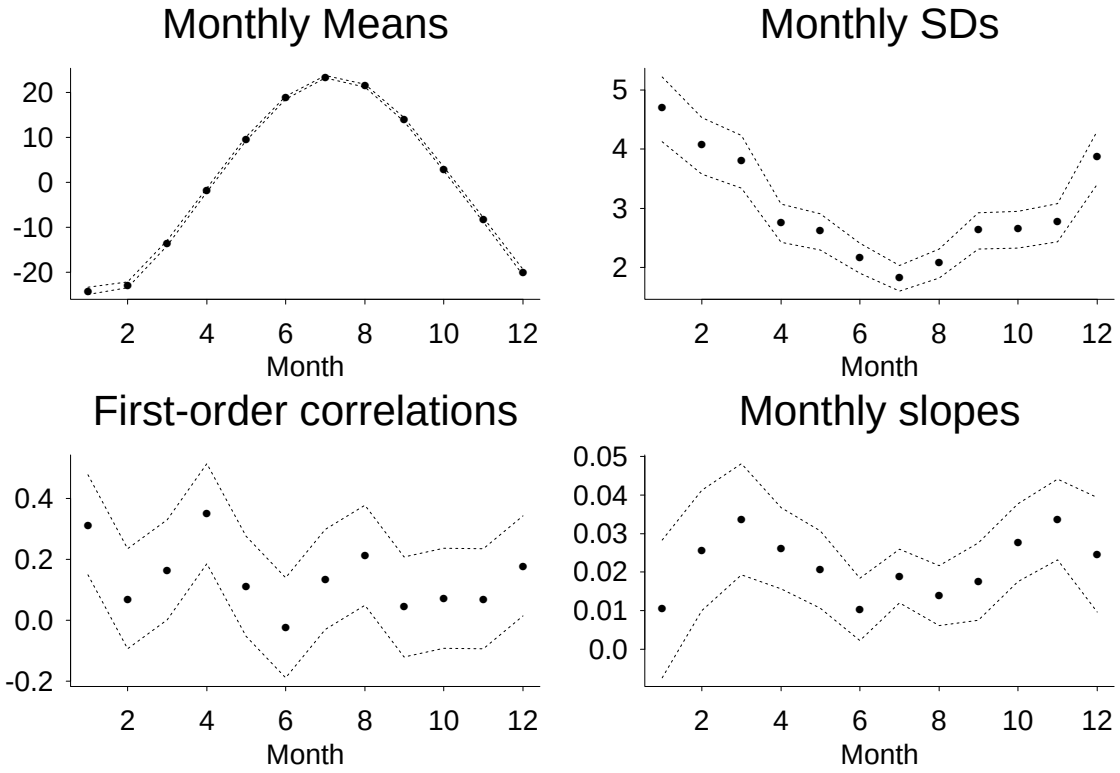


Fig. 6.17. Monthly summary statistics.

To investigate the source of this problem further, some monthly summary statistics were computed (Fig 6.17). All of these except the "Monthly Slopes" plot were obtained from the detrended data in which the overall mean and linear trend had been removed. For each of the twelve months, all data values for that month were extracted, and the mean, standard deviation and correlation with the previous month were computed, along with standard errors (the standard errors being computed with the assumption that all values along the series were independent, but this suffices for an exploratory analysis). The standard deviations and correlations were computed with the idea that the true model might be of PAR(1) form, as in (4.8) earlier. In this model, $\phi^{(m)}$ represents a monthly autoregressive coefficient (for m =1,...,12), and $\sigma^{(m)}$ a monthly standard deviation. Thus if the monthly standard deviations were significantly different, that would point towards a model with different values of $\sigma^{(m)}$ for each month, while if the monthly correlations were significantly different, that would indicate that we need the separate $\phi^{(m)}$ terms as well. The monthly slopes were computed because of the suggestion at the end of Chapter 5 of the main notes, that the warming effect might be different for different months of the year.

In fact, what the plots in Fig. 6.17 show is that the monthly means are very highly significantly different, as of course was bound to happen, *but also* that there is significant variation in the standard deviations, these being much lower in the summer than the winter. Any analysis which does not take this into account is bound to be deficient. It is not so clear whether we should also take into account the seasonal variations in the first-order correlations and in the coefficient of the linear trend. At least, there is some justification for ignoring these variations, but we cannot ignore the variation in monthly standard deviation.

This suggests that we should standardize the series with respect to both the monthly means and standard deviations (as well as the overall linear trend) before fitting an ARMA model. After doing this, the same 20 models were fitted to the detrended data (Table 6.3) and the AIC criterion picked out the ARMA(1,1) model (with no seasonal component) as the most suitable. Note that the AIC values shown here are not directly comparable with those in Table 6.2, as they are not adjusted to allow for rescaling. The ARIMA diagnostics (not shown) are fully consistent with this model.

Nonseasonal Component	Seasonal Component			
	None	(1,0,0)	(0,0,1)	(2,0,0)
(1,0,0)	4889.461	4891.458	4891.458	4893.326
(0,0,1)	4894.467	4896.460	4896.460	4898.411
(2,0,0)	4881.163	4883.148	4883.148	4884.736
(1,0,1)	4878.088	4880.084	4880.084	4881.688*
(0,0,2)	4884.985	4886.971	4886.971	4888.681

Table 6.3 AIC for seasonal ARIMA models (data with linear trend removed standardized by monthly means and standard deviations). *=no convergence as represented by the `$converged` variable.

6.3.1 Forecasting

Forecasts are generated in SPlus, after a model has been fitted, using the `arima.forecast` function. The first 24 values in Fig. 6.18 represent the observed last 24 values of the time series. The next 24 are forecast values generated from the `arima.forecast` function, adjusted to reflect the original means, standard deviations and long-term linear trend. The dotted lines around the last 24 observations represent 95% probability limits for the forecasts, again adjusted to allow for the variable monthly standard deviation.

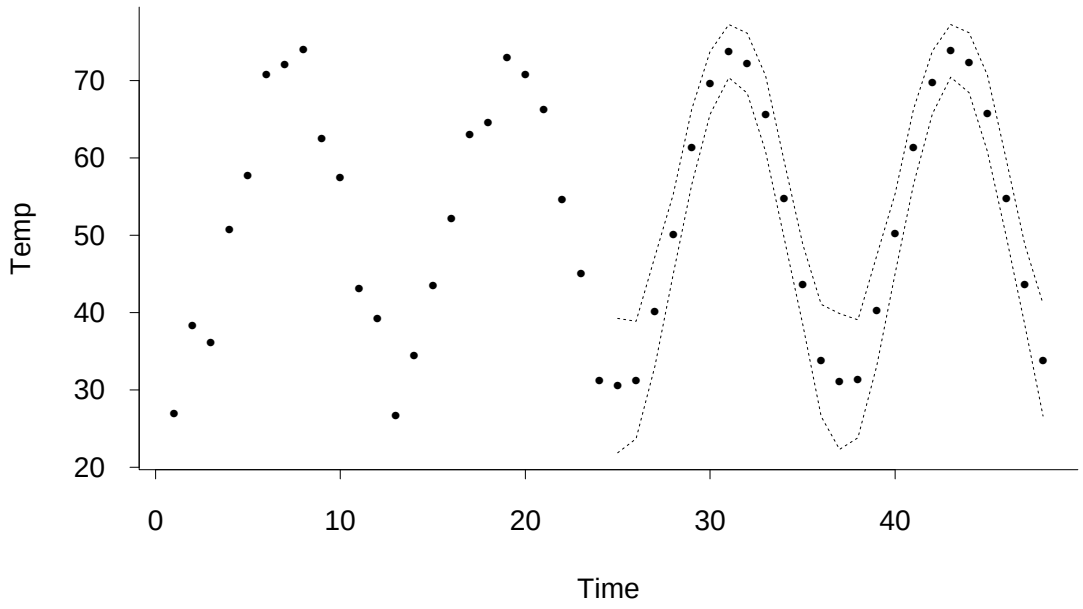


Fig. 6.18. Illustration of ARMA forecasting.

6.3.2 PC processes

A generalization of ARMA processes is to *periodically correlated* (or PC) processes in which both the variances and autocovariances are, potentially, seasonally dependent.

Definition. A time series $\{x_t\}$ is periodically correlated with period M if

$$\mathrm{Cov}\{x_{kM+m}, x_{kM+m+r}\} = \gamma_r^{(m)}, \tag{6.6}$$

independently of k, whenever $k \geq 0$, $1 \leq m \leq M$, $r \geq 0$. The case $M = 1$ is the usual definition of a (second-order) stationary process, but for $M > 1$, the implication is that the variances and autocovariances cycle with period M. As an example, the "monthly S.D.s" and "First-order correlations" plots of Fig. 6.17 represent estimates of $\gamma_r^{(m)}$ ($1 \leq m \leq M = 12$) for $r = 0$ and 1 respectively. Note also that this in principle a more general concept than seasonal ARMA models — the latter may always be rewritten as ordinary ARMA models and so are stationary processes in the usual sense, whereas PC processes are nonstationary except when $M = 1$.

A number of models have been proposed for PC processes. Probably the easiest one to handle is the PAR(1) model:

$$x_{kM+m} = \phi^{(m)} x_{kM+m-1} + \sigma^{(m)} z_{kM+m}, \quad 1 \leq m \leq M,\ k \geq 0, \tag{6.7}$$

again defined for $k \geq 0$, $1 \leq m \leq M$. Here $\phi^{(m)}$ and $\sigma^{(m)}$ are seasonal autoregressive components and standard deviations, and the residual process $\{z_t\}$ is a (non-seasonal) ARMA process. The stationarity condition is $\prod_m |\phi^{(m)}| < 1$.

For the present example we must also allow for different means in each month, as well as an overall trend term. Therefore we expand the model (6.7) to

$$y_{kM+m} = \mu^{(m)} + \beta u_{kM+m} + x_{kM+m} \quad 1 \leq m \leq M,\ k \geq 0, \tag{6.8}$$

where $\{y_t\}$ is the observed time series, β is the overall linear trend, $\{\mu_m,\ 1 \leq m \leq M\}$ are the residual monthly means and $\{x_t\}$ is a zero-mean PAR time series as in (6.7). The possibility of different trends in each month is not considered in this analysis.

Lund, Hurd, Bloomfield and Smith (*Journal of Climate* **11**, pp. 2789–2809, 1995) fitted a number of models of the form (6.7)-(6.8) to the Amherst data. In the case where $\phi^{(m)} = 0$ for all m, the series is not really PC since the series becomes an ordinary ARMA process after standardizing (or adjusting) with respect to the monthly standard deviations; Lund *et al.* called this kind of series *seasonally adjusted* or SA. Using a maximum likelihood fit, they selected an AR(2) process for $\{z_t\}$ and based on fitting all 27 parameters $(\beta, \mu^{(1)}, ..., \mu^{(12)}, \sigma^{(1)}, ..., \sigma^{(12)}, \phi_1, \phi_2)$ by joint maximum likelihood, they obtained a loglik (i.e. $-2 \log L$) value of 8802.2. For the corresponding PAR(1) model, the residual process $\{z_t\}$ was identified as ARMA(2,1) and they found loglik=8775.6 based on 40 parameters. As judged by a χ^2 test, the difference between the two fits ($T = 26.6$ with 13 d.f.) is significant at the level $P = .014$. Using the PAR(1) model, they estimated β (in ^{o}F per year, for comparability with the results quoted in Section 6.2) to be 0.019 with a standard error of 0.0025, a t ratio of 7.6.

There is some possibility of reducing the number of parameters in this model by representing the coefficients $\mu^{(m)}, \sigma^{(m)}$ and $\phi^{(m)}$ as Fourier series. For instance,

$$\sigma^{(m)} = D_0 + \sum_{j=1}^{S} D_j \cos\{2\pi j(m - \rho_j)/12\} \tag{6.9}$$

reduces $\{\sigma^{(m)}\}$ to an S-term Fourier series with $2S+1$ parameters $D_0, ..., D_S, \rho_1, ..., \rho_S$ still needing to be estimated. It was found that $S = 3$ was adequate to represent the variations in standard deviation. Similar analysis for the PAR(1) model reduced $\{\phi^{(m)}\}$ to a 2-term Fourier series; no such reduction was possible for $\{\mu^{(m)}\}$. With these simplifications, the loglik for the SA model became 8808.8 with 22 parameters, that for the PAR(1) model was 8788.7 with 28 parameters. The evidence in favour of the PAR model is again very strong ($T = 20.1$ with 6 d.f.; $P = .003$).

6.4 Volatility and the stock market

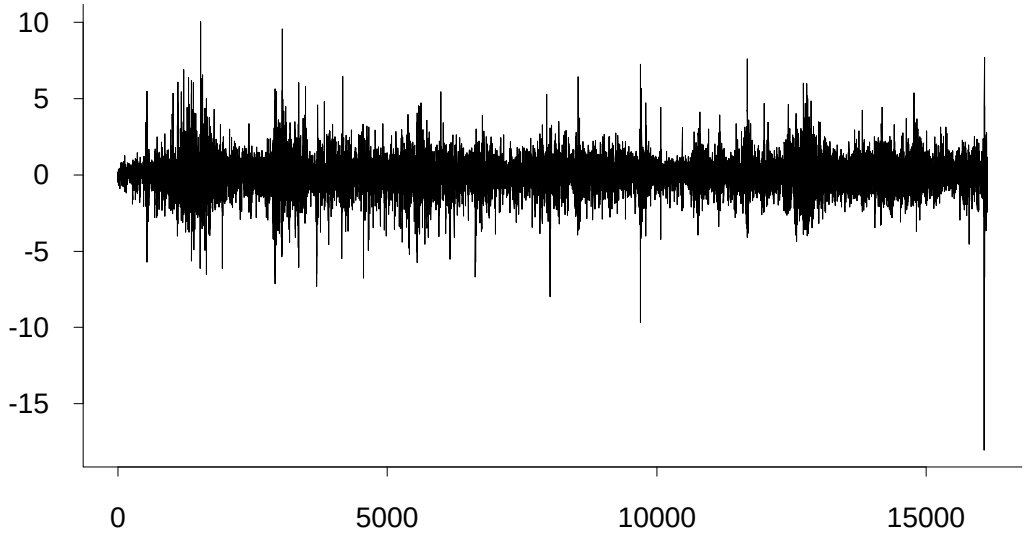


Fig. 6.19. Plot of the S&P series.

Fig. 6.19 shows 16,127 values for the daily logarithmic price change, $\Delta p_t = 100\{\log p_t - \log p_{t-1}\}$, where p_t is the Standard and Poor's Composite Price Index, from 1928–1987. The series has been adjusted to remove systematic calendar and trend effects. Henceforth, this will be called the S&P series.

A key feature of most financial time series is "volatility", or the tendency of such series to display sharp changes in the variance over time. As an example, Fig. 6.20 shows three relatively short sub-series, together with PACF plots. All three subseries are not very far from white noise as judged by the PACF plot, but the standard deviations of the three are very different from one another.

A first possible analysis of the series is simply to fit a high-order AR model to the whole series. The AIC criterion selects $p = 11$ as the appropriate order of the series, but the residual standard deviation is 1.142, compared with 1.154 from the original series – only about a 1% reduction! Of course it would be unrealistic to expect a series of this nature to yield to some simple time series analysis, but a more specific criticism is that the AR analysis takes no account of volatility.

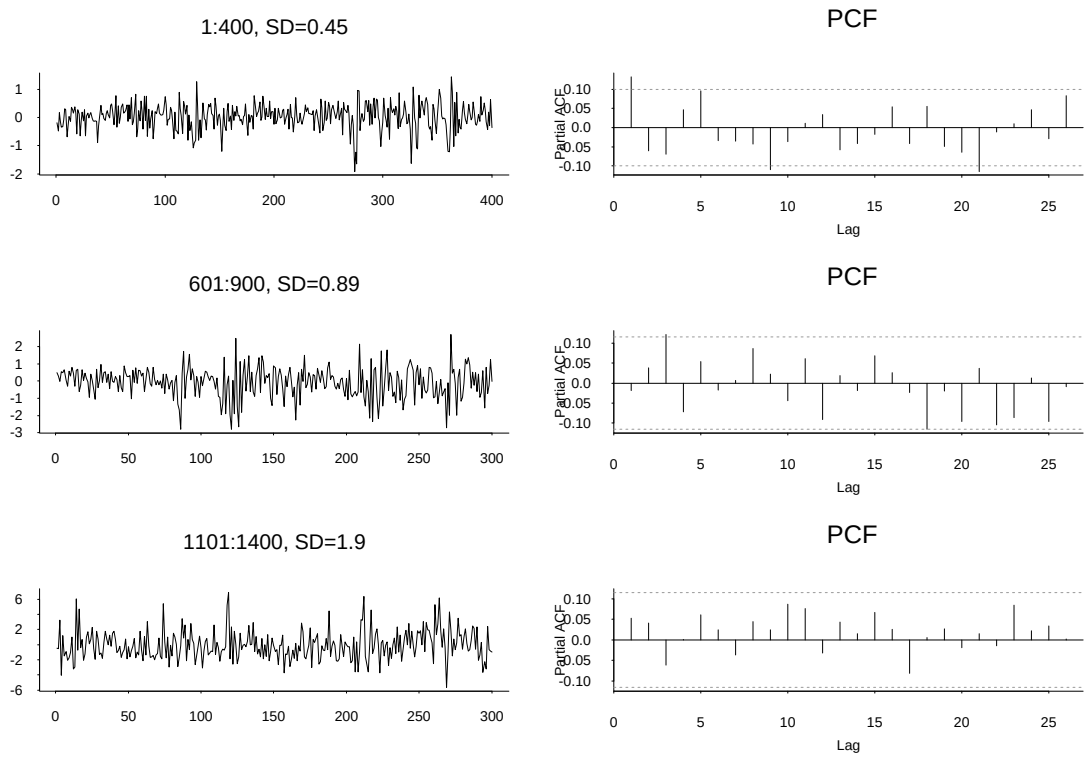


Fig. 6.20. Three subseries and PACF plots.

One possible way to deal with this is to mimic the analysis shown earlier for the EEG data, i.e. fitting AR models to successive subsets of the data and tracing the variability of the coefficients. This is done in Fig. 6.21, which has been computed in the same way as Fig. 6.8 for EEG. In this case, 79 subseries of length 400 were extracted, starting at $n = 1, 201, 401, ..., 15601$. However, the results are not so satisfactory, since the parameters in Fig. 6.21 do not vary nearly as smoothly as those in Fig. 6.8. Moreover, it is not clear that we want to analyse the series this way anyway. In Fig. 6.19, unlike Fig. 6.1, there is no reason to disbelieve the notion that the series is stationary — the difficulty is finding a model which accounts for the changing volatility.

One class of models that has been widely studied in the econometrics literature is the family of ARCH (for *AutoRegressive Conditional Heteroscedastic*) models. *Heteroscedasticity* refers, of course, to the property of nonconstant variance; *conditional* heteroscedasticity implies that this variance can be modelled conditionally on past data, and this is *autoregressive* if the form of the conditioning is based on linear functions. The ARCH model of order 1 for a zero-mean time series $\{x_t\}$ is defined by

$$\begin{aligned} x_t|x_s,\ s<t &\sim N(0,\sigma_t^2), \\ \sigma_t^2 &= \alpha_0 + \alpha_1 x_{t-1}^2. \end{aligned} \tag{6.10}$$

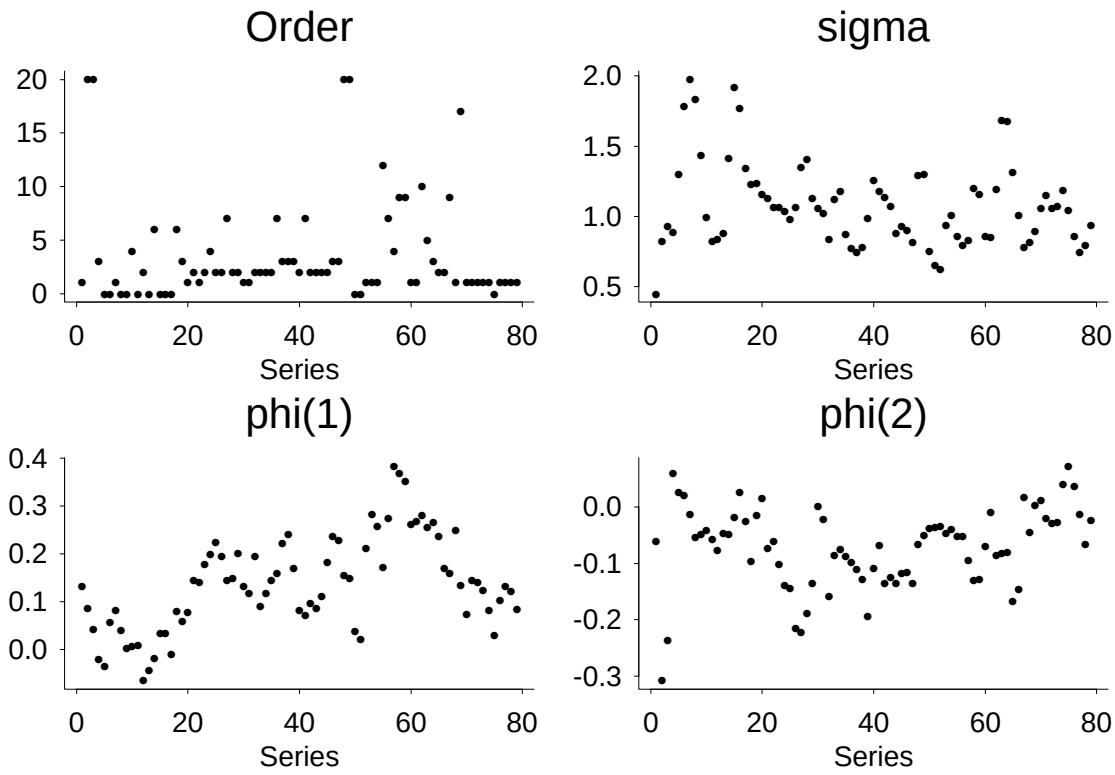


Fig. 6.21. AR models fitted to 79 subseries of the S&P data.

A generalization of (6.10) is to the model

$$\begin{aligned} x_t|x_s,\ s<t &\sim N(0,\sigma_t^2), \\ \sigma_t^2 &= \alpha_0+\alpha_1 x_{t-1}^2+\beta_1\sigma_{t-1}^2 \end{aligned} \tag{6.11}$$

known as the first-order GARCH model. Both models may be generalized in the obvious way to higher-order autoregressions. Then we may extend them to include an ordinary autoregressive term. This leads us to suggest the following GARCH(p,q,r) model for an observed series $\{y_t\}$:

$$\begin{aligned} y_t &= \sum_{j=1}^{p}\phi_j y_{t-j}+x_t, \\ x_t|x_s,\ s<t &\sim N(0,\sigma_t^2), \\ \sigma_t^2 &= \alpha_0+\sum_{j=1}^{q}\alpha_j x_{t-j}^2+\sum_{j=1}^{r}\beta_j\sigma_{t-j}^2. \end{aligned} \tag{6.12}$$

As a first check to see whether this kind of model might be reasonable, Fig. 6.22 shows a scatter plot of y_{t+1}^2 against y_t^2, together with a superimposed smoothed curve obtained using the `lowess` function in Splus. Some outliers outside the bounds of the plotting region are omitted from the plot. The smoothed curve suggests that, at least for relatively small y_t^2 (but covering most of the data), a linear fit is reasonable.

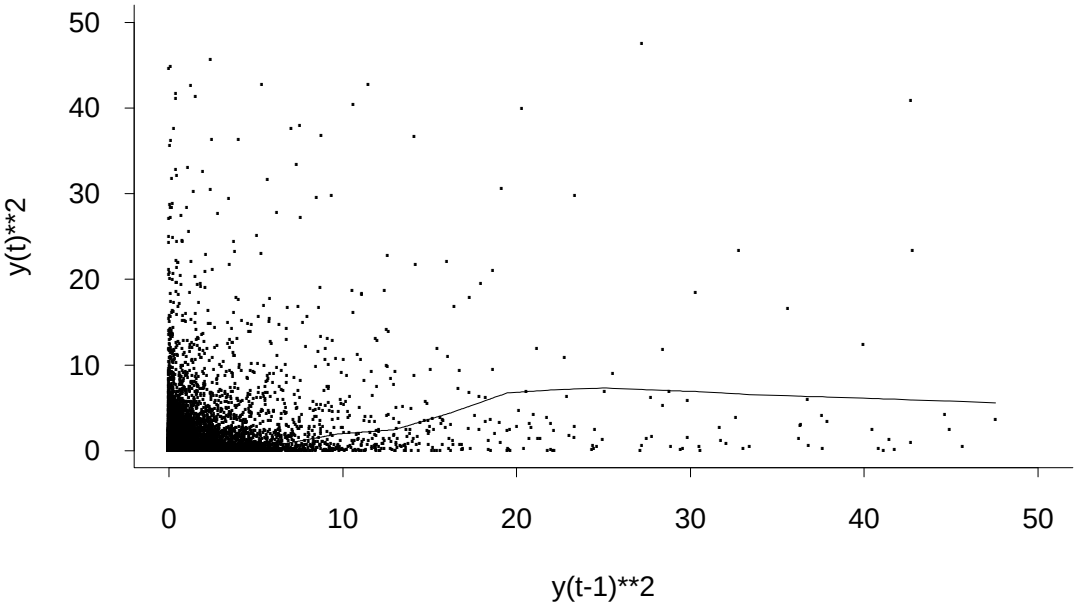


Fig. 6.22. Plot of y_t^2 against y_{t-1}^2, and a smoothed curve.

A sequence of GARCH models were fitted by a numerical maximum likelihood procedure, where the likelihood was computed conditionally on the first 20 data values. The values of loglik (=–2 log L, to be consistent with earlier usage) are given in the following table:

p	q	r	Loglik
1	1	0	48538.2
1	1	1	46174.6
2	1	1	46138.6
2	2	1	46097.3
3	2	1	46086.6
3	3	1	46082.9

Table 6.4 Summary of GARCH models for S&P data

Each of these models is a significant improvement on its predecessor, but attempting to add further terms either resulted in no significant improvement or else ran into boundary problems. The latter arise because the conditional variances must always be positive and this is a meaningful constraint on the parameters! The final fitted model has $\hat{\phi}_1 = 0.169$ (standard error .009), $\hat{\phi}_2 = -.056$ (.009), $\hat{\phi}_3 = .028$ (.008), $\hat{\alpha}_0 = .012$ (.002), $\hat{\alpha}_1 = .145$ (.010), $\hat{\alpha}_2 = -.059$ (.013), $\hat{\alpha}_3 = -.019$ (.010), $\hat{\beta}_1 = .926$ (.007).

In summary, the GARCH models improve significantly (as judged by log likelihoods) on simple AR models, but it is open to debate whether they are really satisfactory models. For example, there is still a danger that one might find $\sigma_t^2 < 0$ when one tries to predict from the model! The positive feature is that they capture two features that are clearly present in financial time series — nonlinearity and heteroscedasticity — but whether they do this in the best way is still very much open to debate.

7. STATE SPACE MODELS AND THE KALMAN FILTER

State space models are an alternative formulation of time series which have a number of positive features, including:

• All ARMA processes may be reformulated as state space models, and this allows a superior treatment of some problems associated with ARMA models, such as exact likelihoods and predictive distributions in short time series, or for a more elegant handling of missing data problems.

• Extension to nonstationary models — for example, a common model in applications of time series analysis is an ARMA with time-varying coefficients, and this situation is handled straightforwardly within a state space formulation.

• Multivariate time series — the standard state space model formulation is multivariate and so automatically suggests models for multivariate time series, which may be easier to handle than multivariate ARMA models (though, of course, such things do exist and there is an extensive theory about them; see Lütkepohl 1993, for example).

• Bayesian approach — the state space model is naturally treated from a Bayesian point of view and so allows one to take advantage of the more general and flexible approach to inference that Bayesian theory provides.

The general model which we shall consider is of the form

$$\begin{aligned} X_t &= F_t S_t + v_t, \\ S_t &= G_t S_{t-1} + w_t, \\ v_t &\sim N(0, V_t), \\ w_t &\sim N(0, W_t), \end{aligned} \tag{7.1}$$

in which X_t represents an observed data vector at time t, S_t is an unobserved "state" of the underlying system, $\{v_t,\ w_t\}$ is an unobserved system of mutually independent errors, and the matrices $F_t,\ G_t,\ V_t$ and W_t are assumed known. In many cases they are independent of t but this is not required. Later on we shall also consider cases in which these matrices contain unknown parameters which are estimated from the data.

Within the framework of (7.1), the main problem is the estimation or prediction of the unobserved sequence of states $\{S_t\}$ in terms of the observed data points $\{X_t\}$. The *Kalman Filter* is a recursive algorithm, first devised by R.E. Kalman in 1960, designed to solve this problem. It is the centerpiece of all statistical analysis based on state space models.

A number of books have presented time series analysis from a state space point of view. In particular, we refer to Harvey (1989), West and Harrison (1996) and Pole, West and Harrison (1994).

7.1 Examples of State Space Models

1. Here is an elementary example to start things off:

$$\begin{aligned} x_t &= s_t + v_t, \\ s_t &= \phi s_{t-1} + w_t, \end{aligned} \tag{7.2}$$

in which all variables are scalar. It follows that

$$x_t - \phi x_{t-1} = v_t - \phi v_{t-1} + w_t.$$

The right hand side has all correlations 0 at lags greater than 1. It is therefore equivalent to an MA(1) process, and this shows that (7.2) is equivalent to an ARMA(1,1) model for $\{x_t\}$. This is the simplest example of an ARMA model being recast in state space form.

2. A slightly more complicated model is represented by the equations

$$\begin{aligned} x_t &= s_{1,t} + v_t, \\ s_{1,t} &= s_{2,t} + w_{1,t}, \\ s_{2,t} &= s_{2,t-1} + w_{2,t}, \end{aligned}$$

so that

$$s_{1,t} - s_{1,t-1} = w_{2,t} + w_{1,t} - w_{1,t-1}$$

and

$$s_{1,T} - s_{1,0} = \sum_{t=1}^{T} w_{2,t} + w_{1,T} - w_{1,0}.$$

Thus the $\{w_{2,t}\}$ variables have the interpretation of a random drift. This is the simplest example of a state space model being used for a nonstationary process, essentially via a mechanism whereby a key model parameter ($s_{1,t}$) gets randomly updated in time.

3. A more complicated model, typical of those in Harvey (1989) or West and Harrison (1996), is to decompose the time series into trend, seasonal and noise components via

$$x_t = m_t + r_t + u_t \tag{7.3}$$

with the individual components updated as follows:

$$m_t = m_{t-1} + \beta_{t-1} + \eta_t, \tag{7.4}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \tag{7.5}$$

$$r_t = -\sum_{j=1}^{M-1} r_{t-j} + \omega_t, \tag{7.6}$$

$$u_t = \sum_{j=1}^{p} \phi_i u_{t-i} + \epsilon_t, \tag{7.7}$$

with $\eta_t, \zeta_t, \omega_t$ and ϵ_t representing mutually independent random errors with mean 0. The interpretation of these equations is as follows: (7.4) and (7.5) represent a trend component subject to a random drift $\{\beta_t\}$, (7.6) represents a slowly varying seasonal component and (7.7) shows that the $\{u_t\}$ sequence is being modeled as AR(p). If $\omega_t \equiv 0$, then (7.6) would imply $\sum_{j=0}^{M-1} r_{t-j} = 0$ for all t, which is a deterministic periodic component of period M. Thus, allowing for small errors $\{\omega_t\}$ means that the seasonal components are changing slowly with time.

To represent (7.3)–(7.8) as a state space model, define

$$
\begin{aligned}
S_t &= (\, m_t \quad \beta_t \quad r_t \quad r_{t-1} \quad \ldots \quad r_{t-M+2} \quad u_t \quad u_{t-1} \quad \ldots \quad u_{t-p+1} \,)^T, \\
w_t &= (\, \eta_t \quad \zeta_t \quad \omega_t \quad 0 \quad \ldots \quad 0 \quad \epsilon_t \quad 0 \quad \ldots \quad 0 \,)^T, \\
F_t &= (\, 1 \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0 \,), \\
v_t &= (\, 0 \quad 0 \quad 0 \quad 0 \quad \ldots \quad 0 \quad 0 \quad 0 \quad \ldots \quad 0 \,)^T,
\end{aligned}
$$

all of these being vectors with $M + 1 + p$ components, and define the updating matrix to be

$$
G_t = \begin{pmatrix}
1 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & -1 & -1 & \ldots & -1 & 0 & 0 & \ldots & 0 \\
0 & 0 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1 & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & 0 & \ldots & 0 & \phi_1 & \phi_2 & \ldots & \phi_p \\
0 & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 1 & 0
\end{pmatrix}.
$$

With these definitions, the general model (7.1) reduces to (7.3), as required.

Remark. It is also possible to write a general ARIMA process as a state space model, using ideas similar to the last example. See section 7.4 for details of this.

7.2 The Kalman Filter

Consider the general model (7.1). Suppose we have observed $X_1, ..., X_t$. Our objective is to obtain the conditional distribution of S_t, given $X_1, ..., X_t$. This problem may be viewed from either a Bayesian perspective, in which we start off with a prior distibution for S_0 and calculate the successive posterior distributions of $S_1, ..., S_t$ as data become available, or from a classical viewpoint, in which the objective is simply to calculate a conditional distribution within the multivariate normal framework. Both viewpoints are encompassed by the recursive algorithm known as the Kalman filter, the difference between them being primarily in the way the recursion is started.

Our derivation follows Meinhold and Singpurwalla (1983). Before giving this, we note the following well-known facts about the multivariate normal distribution (see e.g. Anderson, 1984):

Suppose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim MVN \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]. \tag{7.9}$$

Here Y_1 and Y_2 are vectors of arbitrary dimension, and MVN denotes the multivariate nornmal distribution.

Then the conditional distribution of Y_1 given $Y_2 = y_2$ is

$$MVN \left[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right] \tag{7.10}$$

.

Conversely, if $Y_2 \sim MVN[\mu_2,\ \Sigma_{22}]$ and (7.10) holds, then so does (7.9).

We now use these properties of the multivariate normal distribution to derive the Kalman filter. Suppose the conditional distribution of S_{t-1} given $\mathcal{X}^{t-1} = \{X_1, ..., X_{t-1}\}$ is $MVN[\hat{S}_{t-1}, P_{t-1}]$.

In view of (7.1) we have,

$$(S_t | \mathcal{X}^{t-1}) \sim MVN[G_t \hat{S}_{t-1},\ R_t]$$

where

$$R_t = G_t P_{t-1} G_t^T + W_t. \tag{7.11}$$

We also have $\mathrm{E}\{X_t|S_t\} = F_t S_t$, $\mathrm{Var}\{X_t|S_t\} = V_t$. Applying (7.9)–(7.10) where we identify

$$\begin{aligned} Y_1 &\equiv X_t, \\ Y_2 &\equiv S_t, \\ \mu_2 &\equiv G_t \hat{S}_{t-1}, \\ \Sigma_{22} &\equiv R_t, \\ \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) &\equiv F_t S_t, \\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} &\equiv V_t, \end{aligned}$$

and making all calculations conditional on $\mathcal{X}^{t-1}$, we find that

$$\left\{ \begin{pmatrix} X_t \\ S_t \end{pmatrix} \;\middle|\; \mathcal{X}^{t-1} \right\} \sim MVN \left[\begin{pmatrix} F_t G_t \hat{S}_{t-1} \\ G_t \hat{S}_{t-1} \end{pmatrix}, \begin{pmatrix} V_t + F_t R_t F_t^T & F_t R_t \\ R_t F_t^T & R_t \end{pmatrix} \right]. \tag{7.12}$$

Now we apply (7.9)–(7.10) a second time, reversing the roles of Y_1 and Y_2, to compute the conditional distribution of S_t given $\mathcal{X}^{t-1} \cup X_t = \mathcal{X}^t$ as $MVN[\hat{S}_t, P_t]$ where

$$\begin{aligned} \hat{S}_t &= G_t \hat{S}_{t-1} + R_t F_t^T (V_t + F_t R_t F_t^T)^{-1} (X_t - F_t G_t \hat{S}_{t-1}), \\ P_t &= R_t - R_t F_t^T (V_t + F_t R_t F_t^T)^{-1} F_t R_t. \end{aligned} \tag{7.13}$$

Equations (7.11) and (7.13) together are the Kalman filter updating equations.

To start this recursion, i.e. to determine $\hat{S}_0$ and P_0, there are at least three commonly used approaches,

(i) Bayesian, i.e. use "prior distributions" that reflect a reasonable state of knowledge at the start of the observation period,

(ii) In the case where the matrices F_t, G_t, V_t and W_t are independent of t and the whole system is stationary, we might use the stationary distribution of S_t to start the recursion,

(iii) Set $S_0 = 0,\ P_0 = kI$ where I is the identity matrix and k is a very large positive number; this may be regarded as a reasonable approximation to a prior state of ignorance.

7.3 Prediction and Smoothing

So far, we have only considered the case where, after observing $X_1, ..., X_T$, we want to estimate the final state S_T. The Kalman filter equations provide a recursive formula for calculating $\hat{S}_T$ and P_T, respectively the conditional mean and the conditional covariance matrix of S_T given $\mathcal{X}^T = \{X_1, ..., X_T\}$.

However, there is no reason why we should restrict our objectives to estimating S_T. We may well want to estimate S_t for some other value of t, given $\mathcal{X}^T$. If $t > T$ this is the *prediction* problem, while if $1 \leq t < T$ it is usually known as the *smoothing* or *interpolation* problem. These problems have the same kind of structure as that for estimating S_T, in the sense that the joint distribution of the observed and unobserved quantities is multivariate normal, and therefore the problem reduces to obtaining recursive formulae for the conditional mean and the conditional covariance matrix of the unobserved quantities given the observed data — once the conditional means and conditional covariance matrices are known, it will follow from the multivariate normal theory that all the conditional distributions are also multivariate normal.

Accordingly, let $\hat{S}_{T,t}$ and $P_{T,t}$ denote, respectively, the conditional mean and the conditional covariance matrix of S_t, given $\mathcal{X}^T$. For the case when $t = T$, we shall continue to write $\hat{S}_{T,T} = \hat{S}_T$ and $P_{T,T} = P_T$, as previously. The derivations are quite different in the cases $t \geq T$ and $t \leq T$, so we consider them separately.

Throughout this section, we continue to assume that the matrices $F_t,\ G_t,\ V_t$ and W_t are known for all t (including, for the prediction problem, $t > T$). In many cases, these

matrices will be independent of t, so this does not represent any extension of the preceding assumptions. In any case, if these matrices are not pre-specified, the problem becomes one of estimating unknown parameters, which is the subject of section 7.4.

$t \geq T$: The Prediction Problem

This is done by induction on t, noting that the case $t = T$ is already solved.

Suppose $t > T$ and we know $\hat{S}_{T,t-1}$ and $P_{T,t-1}$. The equation

$$S_t = G_t S_{t-1} + w_t$$

leads to the calculations

$$\begin{aligned} \hat{S}_{T,t} &= G_t \hat{S}_{T,t-1}, \\ P_{T,t} &= G_t P_{T,t-1} G_t^T + W_t. \end{aligned} \tag{7.14}$$

(7.14) is a recursive formula for expressing $\hat{S}_{T,t}$ and $P_{T,t}$ in terms of $\hat{S}_{T,t-1}$ and $P_{T,t-1}$, and therefore allows us to compute these quantities for all $t > T$.

As an extension of this calculation, suppose our actual interest is in predicting X_t rather than S_t. By (7.14) and the formula

$$X_t = F_t S_t + v_t,$$

we are led to conditional distribution

$$(X_t | \mathcal{X}^T) \sim MVN[F_t \hat{S}_{T,t},\ F_t P_{T,t} F_t^T + V_t]. \tag{7.15}$$

In particular, if $t = T + 1$ then $P_{T,t} = R_{T+1}$ by (7.11), and in this case (7.15) reduces to

$$(X_{T+1} | \mathcal{X}^T) \sim MVN[F_{T+1} G_{T+1} \hat{S}_T,\ F_{T+1} R_{T+1} F_{T+1}^T + V_{T+1}]. \tag{7.16}$$

Equation (7.16) is of particular importance because this is the basis of the *prediction error decomposition* for writing down the joint density of $(X_1, ..., X_T)$ for any T, which is important in parameter estimation (section 7.4).

$1 \leq t \leq T$: The Smoothing Problem

Now suppose we are interested in going back through the existing data to compute updated estimates of S_t for $1 \leq t \leq T$, using the full data set $\mathcal{X}^T$. Our approach in this case is backwards induction: the case $t = T$ is known already, and we successively compute $\hat{S}_{T,t}$ and $P_{T,t}$ in terms of $\hat{S}_{T,t+1}$ and $P_{T,t+1}$, for each $t = T - 1,\ T - 2$ and so on down to $t = 1$. The precise formulae are

$$\begin{aligned} \hat{S}_{T,t} &= \hat{S}_t + P_t^* (\hat{S}_{T,t+1} - G_{t+1} \hat{S}_t), \\ P_{T,t} &= P_t + P_t^* (P_{T,t+1} - R_{t+1}) P_t^{*T}, \end{aligned} \tag{7.17}$$

where $\hat{S}_t = \hat{S}_{t,t}$ and $P_t = P_{t,t}$ are the conditional mean and covariance matrix from the "forward" part of the Kalman filter algorithm, and

$$P_t^* = P_t G_{t+1} R_{t+1}^{-1}. \tag{7.18}$$

Proof of (7.17)

The proof is in two parts. First, we show how to compute the conditional mean and covariance matrix of S_t, assuming that S_{t+1} is known (as well as $\mathcal{X}^t$).

This is another application of (7.9) and (7.10). Conditionally on $\mathcal{X}^t$, and using the formula $S_{t+1} = G_{t+1} S_t + w_{t+1}$, we have

$$\begin{pmatrix} S_t \\ S_{t+1} \end{pmatrix} \sim MVN \left[\begin{pmatrix} \hat{S}_t \\ G_{t+1}\hat{S}_t \end{pmatrix}, \begin{pmatrix} P_t & P_t G_{t+1}^T \\ G_{t+1} P_t & R_{t+1} \end{pmatrix} \right].$$

Applying (7.10), we deduce that the conditional distribution of S_t given $\mathcal{X}^t$ and S_{t+1} is

$$\begin{aligned} &MVN[\hat{S}_t + P_t G_{t+1}^T R_{t+1}^{-1}(S_{t+1} - G_{t+1}\hat{S}_t), P_t - P_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} P_t] \\ =&MVN[\hat{S}_t + P_t^*(S_{t+1} - G_{t+1}\hat{S}_t), P_t - P_t^* R_{t+1} P_t^{*T}], \end{aligned} \tag{7.19}$$

using (7.18).

Remark 1. The conditional distribution is the same if we condition on $\mathcal{X}^T$ and S_{t+1}. The reason is that S_t and $\{X_{t+1}, ..., X_T\}$ are conditionally independent given S_{t+1}: in other words, there is no additional information about S_t if we know the whole of $S_{t+1}, X_{t+1}, ..., X_T$ instead of just S_{t+1} (as well as $\mathcal{X}^t$).

Remark 2. Suppose we are interested in generating a Monte Carlo simulation of the entire sequence $(S_1, ..., S_T)$ conditionally on the observed data $\mathcal{X}^T$. This is important for simulation-based Bayesian procedures (section 7.5). Equation (7.19) shows how to do it: first generate $S_T \sim MVN[\hat{S}_T, P_T]$, then successively generate S_{T-1}, $S_{T-2}, ..., S_1$ from the conditional distribution of S_t given S_{t+1} and $\mathcal{X}^T$, which by Remark 1 is the same as (7.19).

Now we come to the second part of the derivation of (7.17). This time, the entire calculation is conditional on $\mathcal{X}^T$. We use the converse part of (7.9)–(7.10), identifying Y_1 with S_t and Y_2 with S_{t+1}. In the notation of (7.9), we identify

$$\mu_1 \equiv \hat{S}_{T,t}, \tag{7.20a}$$
$$\Sigma_{11} \equiv P_{T,t}, \tag{7.20b}$$
$$\mu_2 \equiv \hat{S}_{T,t+1}, \tag{7.20c}$$
$$\Sigma_{22} \equiv P_{T,t+1}, \tag{7.20d}$$
$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) \equiv \hat{S}_t + P_t^*(S_{t+1} - G_{t+1}\hat{S}_t), \tag{7.20e}$$
$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \equiv P_t - P_t^* R_{t+1} P_t^{*T}. \tag{7.20f}$$

From this we deduce, successively, from (7.20e) that $\Sigma_{12}\Sigma_{22}^{-1} \equiv P_t^*$, from (7.20b), (7.20d) and (7.20f) that

$$P_{T,t} - P_t^* P_{T,t+1} {P_t^*}^T = P_t - P_t^* R_{t+1} {P_t^*}^T,$$

which gives the second part of (7.17), and finally from (7.20a), (7.20c) and (7.20e) again that

$$\hat{S}_{T,t} - P_t^* \hat{S}_{T,t+1} = \hat{S}_t - P_t^* G_{t+1} \hat{S}_t,$$

which leads to the first half of (7.17), so concluding the proof.

7.4 Estimation of Unknown Parameters

In practice, the matrices F_t, G_t, V_t and W_t may not be known in advance. It is possible that they depend on other parameters, which we may need to estimate. For example, when taking the ARMA(p, q) model and rewriting it in state space form, the ARMA parameters $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ appear as part of these matrices. The general structure would be to write

$$\begin{aligned} F_t &= F_t(\psi) \\ G_t &= G_t(\psi) \\ V_t &= V_t(\psi) \\ W_t &= W_t(\psi) \end{aligned} \tag{7.21}$$

in terms of a finite-dimensional vector of parameter ψ, and to incorporate maximum likelihood or Bayesian estimates of ψ into the Kalman filter procedure.

As we have already seen in Chapter 4, there is a general technique to do this based on the *prediction error decomposition* summarized in equation (4.5). This depends on being able to calculate the conditional mean and variance of X_t given $X_1, ..., X_{t-1}$ for any $t > 1$. However, writing t in place of $T + 1$, (7.16) gives the answer to that, and so may be fed directly into (4.5). As we saw in Chapter 4, this method is often used to estimate the parameters of ARMA processes. Harvey (1989) has much further detail on this.

Once we have the likelihood function given by (4.5), it is possible to proceed by maximum likelihood estimation, this being the approach that Harvey (1989) recommmends. In that case we simply write (4.5) as a function of ψ and minimize numerically, using any of the standard numerical procedures for unconstrained function minimization. The alternative is a Bayesian approach, as advocated by West and Harrison (1996) amongst others, but in this case also, the first step is to obtain the likelihood function.

As with the Kalman filter itself, there remains the problem of how to start the Kalman recursion. In this context any of the three solutions (i)–(iii), at the end of section 7.2, may be used. From a non-Bayesian perspective, the most satisfactory solution is (ii), i.e. use the stationary distribution of the process as a start-up distribution. However, this may require some calculation, and does not work at all when the process is not stationary. In that case, some variant of (iii) is usually the preferred solution.

7.4.1. ARIMA models with missing data : the Kohn-Ansley approach

For the remainder of this section, we describe in some detail the approach of Kohn and Ansley (1986), which is probably the most comprehensive treatment of parameter estimation for ARMA and ARIMA models, which has the additional benefit of allowing missing data to be treated as part of the same general structure. The essential idea is to show how the Kalman filtering equations can be used to calculate an exact likelihood function for any ARIMA model subject to the usual stationarity conditions. Note, however, the phrase "an exact likelihood" rather than "the exact likelihood" — one of the points of their paper is that the likelihood function is not uniquely defined in the case of a nonstationary process, but in this case they proposed a solution which is equivalent to one adopted on similar contexts in such diverse areas as random effects modeling and spatial statistics, and which also has the advantage of an appealing Bayesian interpretation.

The basic model adopted by Kohn and Ansley is of the form

$$\Phi(B^s)\phi(B)\nabla_s^{d_s}\nabla^d Y_t = \Theta(B^s)\theta(B)\epsilon_t, \tag{7.22}$$

where s is the period of the seasonal effect and

$$\begin{aligned}
\Phi(B) &= 1 - \sum_{j=1}^{P} \Phi_j B^j, \\
\phi(B) &= 1 - \sum_{j=1}^{p} \phi_j B^j, \\
\Theta(B) &= 1 - \sum_{j=1}^{Q} \Theta_j B^j, \\
\theta(B) &= 1 - \sum_{j=1}^{q} \theta_j B^j, \\
\nabla &= 1 - B, \\
\nabla_s &= 1 - B^s.
\end{aligned} \tag{7.23}$$

With slightly different notation, this is equivalent to the general seasonal ARIMA model considered in chapter 4. The assumptions are:

1. The roots of Φ and ϕ lie outside the unit circle in the complex plane.

2. $\epsilon_t \sim N(0, \sigma^2)$ independently for each t.

3. We observe Y_t at time points $t = t_1 < t_2 < ... < t_N = T$.

Let $U_t = \nabla_s^{d_s} \nabla^d Y_t$, so that $\{U_t\}$ is a stationary seasonal ARMA process of nonseasonal orders (p, q) and seasonal orders (P, Q), and define also $D = d + sd_s$, $r = p + sP + D$. Then it is possible to define operators δ, ν and ψ, and associated coefficients $\{\delta_j\}, \{\nu_j\}, \{\psi_j\}$, by

$$\begin{aligned} \delta(B) &= \nabla_s^{d_s} \nabla^d = 1 - \sum_{j=1}^{D} \delta_j B^j, \\ \nu(B) &= \Phi(B^s)\phi(B)\delta(B) = 1 - \sum_{j=1}^{r} \nu_j B^j, \\ \psi(B) &= \Theta(B^s)\theta(B) = \sum_{j=0}^{q+sQ} \psi_j B^j, \end{aligned} \tag{7.24}$$

so that we can write Y_t in two equivalent ways as

$$Y_t = \sum_{j=1}^{D} \delta_j Y_{t-j} + U_t \tag{7.25}$$

or

$$Y_t = \sum_{j=1}^{r} \nu_j Y_{t-j} + \sum_{j=1}^{q+sQ} \psi_j \epsilon_{t-j} \quad (\psi_0 = 1). \tag{7.26}$$

The unknown parameters may be represented as σ^2 and α, where

$$\alpha = (\Phi_1, ..., \Phi_P, \phi_1, ..., \phi_p, \Theta_1, ..., \Theta_Q, \theta_1, ..., \theta_q). \tag{7.27}$$

The initial discussion uses (7.25) to represent the variables $\{Y_t\}$ in terms of a stationary sequence $\{U_t\}$ for which the joint distributions are well defined. Let $\eta = (Y_{1-D}, Y_{2-D}, ..., Y_0)^T$. Then by successively applying (7.25) to $t = 1$, $t = 2$,..., we can write

$$Y_t = a_t^T \eta + Z_t, \quad t \geq 1, \tag{7.28}$$

where a_t is a vector of constant coefficients and Z_t is some linear combination of $\{U_s,\ 1 \leq s \leq t\}$. Defining $Y = (Y_1, ..., Y_T)^T$, $Z = (Z_1, ..., Z_T)^T$ and A to be the matrix with rows a_t^T, $1 \leq t \leq T$, we may also write (7.28) in vector-matrix notation as

$$Y = A\eta + Z, \quad Z \sim N(0, \sigma^2 V(\alpha)), \tag{7.29}$$

where A is a $T \times D$ matrix of known coefficients, η is a $D \times 1$ vector of unknown observations and $V(\alpha)$ is some known (in principle, explicitly calculable) $T \times T$ covariance matrix whose entries depend on the unknown parameters α.

The discussion so far implicitly assumes that there are no missing observations. However, even in the case with missing observations it is still possible to write each observed

value Y_{t_j} as a linear combination of $\{U_s,\ 1 \le s \le t_j\}$ together with a linear function of η. Therefore, the representation (7.29) still holds, the only difference being that the dimension of Y is now N, the actual number of observed values, rather than the nominal length of the series T. We then have that A is a $N \times D$ matrix; it is possible that A is of rank $D' < D$, but in that case, we eliminate unwanted components of η and write the model again in the form (7.29), but with η now of dimension D' and A a $N \times D'$ matrix of rank D'.

7.4.2. A digression: Marginal likelihood, restricted likelihood, integrated likelihood and Bayesian statistics.

The model (7.29) is of a structure that arises in numerous different statistical contexts. For example, in spatial statistics it is sometimes known as the *universal kriging* problem (Cressie 1993) — here $A\eta$ represents the deterministic component and Z the random component in some spatially distributed set of observations, but our interest is very often in the parameters σ^2 and α which govern the random component. Many models for random effects in linear models also reduce to equations of the form (7.29) where the parameters σ^2 and α need to be estimated. In most of these context, η represents a set of unknown nuisance parameters whereas in the present context, it is really a set of nuisance *observations* rather than parameters, but since we have no well-defined model for those observations, there is no practical distinction between the two problems.

One solution to (7.29) is by maximum likelihood. If we abbreviate $V(\alpha)$ to V and write the joint density of Y in the form

$$(2\pi\sigma^2)^{-N/2}|V|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(Y-A\eta)^T V^{-1}(Y-A\eta)\right\}, \tag{7.30}$$

and then maximize successively with respect to η, σ^2 and α, we find

$$\hat{\eta} = (A^T V^{-1} A)^{-1} A^T V^{-1} Y, \tag{7.31a}$$

$$\hat{\sigma}^2 = \frac{(Y-A\hat{\eta})^T V^{-1}(Y-A\hat{\eta})}{N}, \tag{7.31a}$$

and then choose α to maximize the resulting expression (7.30) when $\hat{\eta}$ and $\hat{\sigma}^2$ are substituted for η and σ^2.

Unfortunately, there are various disadvantages of this solution, the most obvious of which is apparent when α is known: in that case (7.31b) gives an estimate of σ^2 which is known to be biased, the correct divisor being $N - D'$ rather than N.

A second solution is the *restricted maximum likelihood* (also called *reduced maximum likelihood*) solution, usually abbreviated to REML. This was introduced by Patterson and Thompson (1971) in the context of random effects linear models. We assume the matrix

A is of rank D': suppose it is possible to decompose Y into two components, a D'-dimensional component in the column space of A, and an orthogonal component BY, where B is $(N-D') \times N$ and $BA = 0$. Then the distribution of $BY = BZ$ is independent of η, and may be used to define a likelihood of (σ^2, α) which may be maximized directly to obtained the so-called REML estimates. In a different context this is also known as the marginal likelihood solution (Kalbfleish and Sprott 1970) and it is from the latter point of view that Kohn and Ansley introduce the problem.

Before stating the REML solution, we introduce a third solution, which makes clear the connection with Bayesian statistics. Suppose we introduce a flat prior density for η, of the form $\pi(\eta) = 1$ for all η. The objective is then to integrate out (7.30) with respect to η, leaving a function of σ^2 and α alone. If we again define $\hat{\eta}$ by (7.31a), we have the sum of squares decomposition

$$(Y - A\eta)^T V^{-1} (Y - A\eta) = (Y - A\hat{\eta})^T V^{-1} (Y - A\hat{\eta}) + (\eta - \hat{\eta})^T A^T V^{-1} A (\eta - \hat{\eta}),$$

and we also have

$$\int \exp\left\{ -\frac{1}{2\sigma^2} (\eta - \hat{\eta})^T A^T V^{-1} A (\eta - \hat{\eta}) \right\} d\eta = (2\pi\sigma^2)^{D'/2} |A^T V^{-1} A|^{1/2}.$$

Putting the pieces together, we get a likelihood for (σ^2, α) of the form

$$(2\pi\sigma^2)^{-(N-D')/2} |V|^{-1/2} |A^T V^{-1} A|^{1/2} \exp\left\{ -\frac{1}{2\sigma^2} (Y - A\hat{\eta})^T V^{-1} (Y - A\hat{\eta}) \right\}. \qquad (7.32)$$

Note that if α is treated as known and (7.32) maximized for σ^2, we get the correct unbiased estimator, with divisor $N - D'$ instead of N.

As an alternative form of this calculation, suppose we take prior distribution $\eta \sim N(0, kI_{D'})$, where $I_{D'}$ is the $D' \times D'$ identity matrix and k is some large constant. Ignoring a constant component of the density, this is $\pi(\eta) = \exp\{-\eta^T \eta/(2k)\}$. Then the same calculation leads to a marginal density of Y (after multiplying the density of Y by $\pi(\eta)$ and integrating out η) which may be written as

$$\begin{aligned} &(2\pi\sigma^2)^{-(N-D')/2} |V|^{-1/2} \left| A^T V^{-1} A + \frac{\sigma^2 I_d}{k} \right|^{1/2} \cdot \\ &\cdot \exp\left\{ -\frac{(Y - A\tilde{\eta})^T V^{-1} (Y - A\tilde{\eta})}{2\sigma^2} - \frac{\tilde{\eta}^T \tilde{\eta}}{2k} \right\}. \end{aligned} \qquad (7.33)$$

Here $\tilde{\eta}$ is the "Bayes estimator" of η, which is

$$\tilde{\eta} = \left(\frac{A^T V^{-1} A}{\sigma^2} + \frac{I_D}{k} \right)^{-1} \frac{A^T V^{-1} Y}{\sigma^2}.$$

The proof of (7.33) relies on the algebraic identity

$$\begin{aligned}&\frac{(Y-A\eta)^T V^{-1}(Y-A\eta)}{\sigma^2}+\frac{\eta^T\eta}{k}\\&=\frac{(Y-A\tilde{\eta})^T V^{-1}(Y-A\tilde{\eta})}{\sigma^2}+\frac{\tilde{\eta}^T\tilde{\eta}}{k}+(\eta-\tilde{\eta})^T\left(\frac{A^T V^{-1}A}{\sigma^2}+\frac{I_D}{k}\right)^{-1}(\eta-\tilde{\eta}).\end{aligned}\tag{7.34}$$

The formula (7.34) may easily be checked by expanding out as a function of η.

As $k\to\infty$, we have $\tilde{\eta}-\hat{\eta}=O(1/k)$, therefore the solution (7.33) differs from (7.32) by $O(1/k)$, and in particular, (7.32) arises as a limiting form of (7.33) as $k\to\infty$.

Strictly speaking, either (7.32) or (7.33) should be regarded as an *integrated likelihood* solution rather than a formal Bayesian solution, because we have not introduced any prior distribution for (σ^2,α). However, if we were to introduce a prior distribution for (σ^2,α), independent of that for η, then a suitable solution would be to define (7.32) or (7.33) as an marginal likelihood for Y after removing the dependence on η, which may then be treated as a Bayesian inference problem for σ^2 and α alone — in other words, we multiply the marginal likelihood by the prior density of (σ^2,α), and then normalize that expression to obtain the posterior density of (σ^2,α).

Now we return to the REML solution. Recall that this is based on the density of BY, where B is a $(N-D')\times N$ matrix such that $BA=0$. The key result is due to Harville (1974), and states: modulo a normalizing constant, this density is given by (7.32). This result is invariant to the precise specification of B, which only affects the normalizing constant, and which is of no relevance for either maximum likelihood or Bayesian calculations.

This result therefore shows that the marginal likelihood or REML solution, and the integrated likelihood or Bayesian solution with $\pi(\eta)=1$, all lead to the same solution (7.32). The proper prior $\eta\sim N(0,kI_D)$ also leads to a solution which differs from (7.32) by $O(1/k)$, and therefore is effectively equivalent if k is chosen large enough.

Kohn and Ansley (1986) did not refer directly to REML estimation or Harville's result, but they developed the mathematics of their model in an earlier paper of theirs (Ansley and Kohn 1985), which includes a proof that the integrated likelihood solution based on $\eta\sim N(0,kI_D)$ differs by $O(1/k)$ from the marginal likelihood solution.

7.4.3. Representing an ARIMA model as a state space model.

There are numerous variations on the basic method of representing ARMA or ARIMA models in state space forms — for example, the representation due to Kohn and Ansley (1986) is different from that of Harvey and Pierse (1984). We follow Kohn and Ansley here.

The starting point is equation (7.26). Let $f = \max(r, q + sQ + 1)$ and let $S_t = (S_t^{(1)}, ..., S_t^{(f)})$ be the $f \times 1$ state vector whose components are

$$\begin{aligned} S_t^{(1)} &= Y_t, \\ S_t^{(j)} &= \sum_{i=j}^{r} \nu_i Y_{t-1+j-i} + \sum_{i=j-1}^{q+sQ} \psi_i \epsilon_{t-1+j-i}, \end{aligned}$$

it being understood that all coefficients ν_i and ψ_i outside the range of their definition are 0.

Define an $f \times f$ matrix F and $f \times 1$ vectors g and h by

$$F = \begin{pmatrix} \nu_1 & 1 & 0 & 0 & \dots & 0 \\ \nu_2 & 0 & 1 & 0 & \dots & 0 \\ \nu_3 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \nu_{f-1} & 0 & 0 & 0 & \dots & 1 \\ \nu_f & 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \qquad g = \begin{pmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{f-2} \\ \psi_{f-1} \end{pmatrix}, \qquad h = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Then

$$\begin{aligned} Y_t &= h^T S_t, \\ S_{t+1} &= F S_t + g \epsilon_{t+1}. \end{aligned} \tag{7.35}$$

Proof of (7.35)

The first equation is obvious, so we concentrate on the second. For $j = 1$,

$$\begin{aligned} S_{t+1}^{(1)} &= \nu_1 Y_t + \sum_{i=2}^{r} \nu_i Y_{t+1-i} + \sum_{i=1}^{q+sQ} \psi_i \epsilon_{t+1-i} + \epsilon_{t+1} \\ &= \sum_{i=1}^{r} \nu_i Y_{t+1-i} + \sum_{i=0}^{q+sQ} \psi_i \epsilon_{t+1-i} \\ &= Y_{t+1}, \end{aligned}$$

as required.

Now consider the j'th component, $j > 1$:

$$\begin{aligned} S_{t+1}^{(j)} &= \nu_j S_t^{(1)} + S_t^{(j+1)} \\ &= \nu_j Y_t + \sum_{i=j+1}^{r} \nu_i Y_{t+j-i} + \sum_{i=j}^{q+sQ} \psi_i \epsilon_{t+j-i} \\ &= \sum_{i=j}^{r} \nu_i Y_{t+j-i} + \sum_{i=j}^{q+sQ} \psi_i \epsilon_{t+j-i}, \end{aligned}$$

as required. These equations include the case $j = f$ if we note that $\nu_i = 0$ for $i > f$ and $\psi_i = 0$ for $i \geq f$. With that, the proof of (7.35) is complete.

7.4.4. Limiting forms of the prediction error decomposition and the Kalman filter as $k \to \infty$.

We continue to follow Kohn and Ansley (1986). Assuming the prior distribution $\eta \sim N(0, kI_{D'})$, define innovations $\{\epsilon(j;k),\ j = 1, ..., N\}$ by

$$\begin{aligned}\epsilon(1;k) &= Y_{t_1}, \\ \epsilon(j;k) &= Y_{t_j} - \mathrm{E}\{Y_{t_j} \mid Y_{t_1}, ..., Y_{t_{j-1}}\},\end{aligned}$$

and also let

$$R(j;k) = \mathrm{Var}\{\epsilon(j;k)\}.$$

Kohn and Ansley show that there exist limiting quantities $\epsilon_j^{(0)}$, $R_j^{(0)}$, $R_j^{(1)}$ such that

$$\begin{aligned}\epsilon(j;k) &= \epsilon_j^{(0)} + O\left(\frac{1}{k}\right), \\ R(j;k) &= kR_j^{(1)} + \sigma^2 R_j^{(0)} + O\left(\frac{1}{k}\right),\end{aligned}$$

and, moreover, there are only D' values of j for which $R_j^{(1)} \neq 0$. The limiting marginal likelihood as $k \to \infty$ is then given, modulo a constant, by

$$\left\{\sigma^{2(N-D')} \prod_{j=1}^{N}{}' R_j^{(0)}\right\}^{-1/2} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^{N}{}' \frac{\epsilon_j^{(0)\,2}}{R_j^{(0)}}\right\},$$

where $\prod'$ and $\sum'$ mean respectively product and sum over all j values for which $R_j^{(1)} = 0$.

We now define the modified Kalman filter algorithm, whose specific purpose is to handle directly the limiting case as $k \to \infty$ in the ordinary Kalman filter. Let $\hat{S}_{T;t}(k)$ and $P_{T;t}(k)$ denote the conditional mean and conditional variance of Y_t given $\{Y_{t_j},\ t_j \leq T\}$, analogously to section 7.3 except that the dependence on k is made explicit. It is possible to write the initial state of the system in the form

$$S_0 = M\eta + \xi, \tag{7.36}$$

in which M is some $f \times D'$ matrix and $\xi \sim N_f(0, \sigma^2 V_\xi(\alpha))$ for some $f \times f$ matrix $V_\xi(\alpha)$. The initial conditions are of the form

$$\hat{S}_{0;0} = 0, \quad P_{0,0}(k) = kMM^T + \sigma^2 V_\xi,$$

and there is a representation of the form

$$\hat{S}_{T;t}(k) = \hat{S}^{(0)}_{T;t} + O\left(\frac{1}{k}\right),$$
$$P_{T;t}(k) = kP^{(1)}_{T;t} + \sigma^2 P^{(0)}_{T;t} + O\left(\frac{1}{k}\right),$$

where $\hat{S}^{(0)}_{T;t}$, $P^{(1)}_{T;t}$ and $P^{(0)}_{T;t}$ do not depend on k. The *modified Kalman filter* is as follows:

Step 0 (initialization). Set

$$\hat{S}^{(0)}_{0;0} = 0, \qquad P^{(0)}_{0;0} = MM^T, \qquad P^{(1)}_{0;0} = V_\xi.$$

Steps 1–5 are repeated for $t = 0, ..., T-1$.

Step 1. Set

$$\hat{S}^{(0)}_{t;t+1} = F\hat{S}^{(0)}_{t;t},$$
$$P^{(1)}_{t;t+1} = FP^{(1)}_{t;t}F^T,$$
$$P^{(0)}_{t;t+1} = FP^{(0)}_{t;t}F^T + gg^T.$$

If Y_{t+1} is missing, then perform Step 2. Otherwise go to Step 3.

Step 2. Set

$$\hat{S}^{(0)}_{t+1;t+1} = \hat{S}^{(0)}_{t;t+1},$$
$$P^{(1)}_{t+1;t+1} = P^{(1)}_{t;t+1},$$
$$P^{(0)}_{t+1;t+1} = P^{(0)}_{t;t+1}.$$

Return to Step 1.

Step 3. Since Y_{t+1} is observed, $t+1 = t_j$ for some j.

$$\epsilon^{(0)}_j = Y_{t_j} - h^T\hat{S}^{(0)}_{t;t+1},$$
$$R^{(1)}_j = h^T P^{(1)}_{t;t+1}h,$$
$$R^{(0)}_j = h^T P^{(0)}_{t;t+1}h.$$

If $R^{(1)}_j > 0$ go to Step 4, otherwise go to Step 5.

Step 4.

$$\hat{S}^{(0)}_{t+1;t+1} = \hat{S}^{(0)}_{t;t+1} + \frac{P^{(1)}_{t;t+1} h \epsilon^{(0)}_j}{R^{(1)}_j},$$

$$P^{(1)}_{t+1;t+1} = P^{(1)}_{t;t+1} - \frac{P^{(1)}_{t;t+1} h h^T P^{(1)}_{t;t+1}}{R^{(1)}_j},$$

$$P^{(0)}_{t+1;t+1} = P^{(0)}_{t;t+1} + \frac{P^{(1)}_{t;t+1} h h^T P^{(1)}_{t;t+1} R^{(0)}_j}{{R^{(1)}_j}^2} - \frac{P^{(1)}_{t;t+1} h h^T P^{(0)}_{t;t+1}}{R^{(1)}_j} - \frac{P^{(0)}_{t;t+1} h h^T P^{(1)}_{t;t+1}}{R^{(1)}_j}.$$

Return to Step 1.

Step 5.

$$\hat{S}^{(0)}_{t+1;t+1} = \hat{S}^{(0)}_{t;t+1} + \frac{P^{(0)}_{t;t+1} h \epsilon^{(0)}_j}{R^{(0)}_j},$$

$$P^{(1)}_{t+1;t+1} = P^{(1)}_{t;t+1},$$

$$P^{(0)}_{t+1;t+1} = P^{(0)}_{t;t+1} - \frac{P^{(0)}_{t;t+1} h h^T P^{(0)}_{t;t+1}}{R^{(0)}_j}.$$

Return to Step 1.

In most cases, $P^{(1)}_{t;t}$ becomes 0 for a fairly small value of t; thereafter, it remains 0, and the modified Kalman filter is identical to the ordinary Kalman filter.

The paper by Kohn and Ansley also describes modifications of the prediction and smoothing procedures of section 7.3, to allow for direct calculation of the limiting distributions as $k \to \infty$.

7.5 Modern Bayesian Approaches to State Space Modeling

7.5.1. The Gibbs sampler and Hastings-Metropolis algorithms

Much modern Bayesian statistics relies on *Markov chain Monte Carlo* (MCMC) algorithms to obtain simulations from the posterior distribution of a statistical inference problem. The idea is that, in the absence of analytic solutions to Bayesian problems, we will try to construct a numerical solution by Monte Carlo simulation. Markov chains enter the discussion because, in many cases, the most convenient way to simulate a posterior distribution is to construct a Markov chain whose stationary distribution can be proved to equal the desired posterior distribution. Once such a chain is constructed, it is run for a sufficiently large number of iterations that its marginal distributions can be assumed to approximate the posterior distribution required. What constitutes a "sufficiently large

number of iterations" is a difficult question which we shall not get into here, but we shall describe the two most commonly used forms of MCMC algorithm, the *Gibbs sampler* and the *Hastings-Metropolis* algorithm. For further reading on MCMC methods, the books of Gilks *et al.* (1996) and Gamerman (1997) are recommended, or the recent review paper by Brooks (1998).

The Gibbs Sampler.

The algorithm now known as the *Gibbs sampler* was first proposed in its present form by Geman and Geman (1984), though there were a number of precedents for the basic idea. Suppose we want to simulate a K-dimensional random variable $(X_1, ..., X_K)$ with density $g(x_1, ..., x_K)$. Define the conditional density of X_k given $\{X_j,\ j \neq k\}$,

$$g_k(x_k \mid x_j,\ j \neq k) = \frac{g(x_1, ..., x_{k-1}, x_k, x_{k+1}, ..., x_K)}{\int g(x_1, ..., x_{k-1}, x'_k, x_{k+1}, ..., x_K) dx'_k}.$$

The situation when the Gibbs sampler works particularly well is when each of the marginal densities g_k is easy to draw a Monte Carlo sample from (for example, because it is a standard density such as normal, gamma or beta), but the joint density of all K random variables is not. In that case, we proceed as follows:

Step 1: Define arbitrary $X_1^{(0)}, ..., X_K^{(0)}$ and set $n = 0$.

Step 2: Generate successive random variables

$$\begin{aligned}
X_1^{(n+1)} &\sim g_1(\cdot \mid X_2^{(n)}, X_3^{(n)}, ..., X_K^{(n)}), \\
X_2^{(n+1)} &\sim g_2(\cdot \mid X_1^{(n+1)}, X_3^{(n)}, ..., X_K^{(n)}), \\
X_3^{(n+1)} &\sim g_3(\cdot \mid X_1^{(n+1)}, X_2^{(n+1)}, ..., X_K^{(n)}), \\
&\vdots \\
X_K^{(n+1)} &\sim g_K(\cdot \mid X_1^{(n+1)}, X_2^{(n+1)}, ..., X_{K-1}^{(n+1)}).
\end{aligned}$$

Step 3: Set $n = n + 1$ and return to Step 2.

The iteration proceeds until n is considered large enough both to achieve convergence and to generate a large enough sample of values. Usually some initial number n_0 iterations are treated as a warm-up or burn-in sample and discarded, and then the process is repeated for a further n_1 iterations, which are treated as a sample of vectors from the joint density f.

It should be noted that the algorithm does not in any way depend on the assumption that each X_k be scalar — the key point is that there should be a partition of the vector X into subvectors X_k such that each X_k may be sampled from directly.

A typical application of the Gibbs sampler in Bayesian statistics would arise if we observe $Y \sim f(y; \theta_1, ..., \theta_K)$ where $\theta_1, ..., \theta_K$ are parameters, and we have a prior density $\pi(\theta_1, ..., \theta_K)$. Then the posterior density is of the form

$$\pi(\theta_1, ..., \theta_K \mid Y) = C\pi(\theta_1, ..., \theta_K) f(Y; \theta_1, ..., \theta_K), \tag{7.37}$$

where the normalizing constant C is chosen to make the joint density (7.37) integrate to 1. Except in cases admitting a conjugate prior, exact calculation of C may require complicated numerical integration beyond the scope of most computer packages.

In many cases, however, it is possible to define the prior density in a hierarchical way so that the prior for each parameter θ_k, conditionally on all the other parameters $\{\theta_j,\ j \neq k\}$, is of a conjugate form. In this case, the Gibbs sampler is a very natural solution: we sample successively from each θ_k, using the conjugate form of posterior distribution, and repeat the whole process for enough iterations to guarantee sufficient coverage of the full joint posterior density.

The Hastings-Metropolis algorithm

An alternative MCMC algorithm applies in cases where it is not possible to apply the Gibbs sampler because there is no way to break up the random vector X into subcomponents for which exact conditional sampling is possible. The original version of the algorithm was given by Metropolis *et al.* (1953), and was reworked into its present form by Hastings (1970).

In the following discussion, we shall suppose that X is an arbitrary random vector with density $g(x)$. As with the Gibbs sampler, the application in Bayesian statistics arises when X is identified with some parameter vector θ and g is the posterior density of θ given some observations Y. Thus it is typically the case that the analytic form of g is known up to an unknown normalizing constant.

As with the Gibbs sampler, the algorithm typically starts from some arbitrary initial value $X^{(0)}$ and proceeds by iteration. Suppose after n iterations we have a current value $X^{(n)}$. We proceed as follows: generate a "trial value" X' from some density $q(X^{(n)}, x')$ such that $\int q(x, x')dx' = 1$ for all x'. The choice of trial density q is almost arbitrary. Once X' is defined, however, we perform a second independent randomization to decide whether to accept or reject it. The probability of acceptance is set to be

$$\min\left\{1, \frac{g(X')q(X', X^{(n)})}{g(X^{(n)})q(X^{(n)}, X')}\right\}. \tag{7.38}$$

If we accept, then set $X^{(n+1)} = X'$, otherwise $X^{(n+1)} = X^{(n)}$. A key point of the algorithm is that (7.38) can be calculated even though the density g is known only up to a normalizing constant, because the constant cancels from the numerator and denominator of the ratio $g(X')/g(X^{(n)})$.

Sketch of proof.

A complete proof of the Hastings-Metropolis algorithm involves the theory of continuous-state Markov chains, which lies beyond the scope of the present discussion (see Meyn and Tweedie 1993 or Tierney 1994 for details). We can, however, give the main idea, which is to show that if the Markov chain $\{X^{(n)}\}$ is defined by the algorithm, g is the density of an invariant measure (in other words, if g is the density of $X^{(n)}$ then it is also the density of $X^{(n+1)}$). The remainder of the proof, which we shall not give, involves checking continuous-state versions of the irreducibility and aperiodicity conditions which are used in proving convergence of discrete-state Markov chains.

Suppose, then, that $X^{(n)}$ has the density g. For any x in the domain of g, let

$$A_x = \left\{ y : \ \frac{g(y)q(y,x)}{g(x)q(x,y)} < 1. \right\}$$

and write A_x^c for the complement of A_x. The density of $X^{(n+1)}$, evaluated at g, is then

$$\begin{aligned} & g(x) \int_{A_x} q(x,y) \cdot \left\{ 1 - \frac{g(y)q(y,x)}{g(x)q(x,y)} \right\} dy \\ & + \int_{A_x} g(y)q(y,x)dy + \int_{A_x^c} g(y)q(y,x) \cdot \frac{g(x)q(x,y)}{g(y)q(y,x)} dy, \end{aligned} \tag{7.39}$$

in which the first line represents the probability that $X^{(n)} = x$ and any proposed move is rejected, and the second that $X^{(n)} = y$ but the chain moves from y to x at time $n+1$. However, on rearrangement, (7.39) becomes

$$g(x) \int_{A_x} q(x,y)dy + g(x) \int_{A_x^c} q(x,y)dy = g(x),$$

as required.

So far, we have given no guidance to the choice of a suitable trial density q. Very many possibilities could be considered, but among the most frequently used in practice are

(i) The *independence sampler*, in which $q(x,y)$ is some density $q_0(y)$ independent of x. In Bayesian analysis, this is sometimes applied with a normal approximation to the posterior density used to determine q_0.

(ii) The *random walk* sampler, in which $q(x,y)$ is of the form $q_0(y-x)$ for some density q_0 — in this case, successive steps of the sampler follow a random walk where the (vector) step length has density q_0. For example, one possibility is to sample uniformly over a box of the form $|y_i - x_i| < h_i$ for all i, where y_i and x_i are the ith components of the vectors y and x and h_i is a bound on the maximum step length. This still leaves open the choice of the h_i parameters, but they are very often chosen to satisfy a rule of

thumb that the overall acceptance rate should be between 15% and 50% (Gilks *et al.* 1996, page 55). A further simplification in the case of a random walk sampler with q_0 symmetric about 0 is that $q(x,y) = q(y,x)$ for all x, y, and in that case the qs cancel in (7.38). This case is sometime called *the Metropolis algorithm* to distinguish it from the slightly more complicated asymmetric case introduced by Hastings.

In practice, it is quite common to combine the Gibbs sampler and Hastings-Metropolis algorithms, possibly with other Monte Carlo generation procedures, into a single algorithm. Tierney (1994) gives extensive coverage of such *hybrid* procedures. One idea, for instance, is to break down the vector X into subcomponents $X_1, ..., X_K$, using Gibbs sampling to update one component at a time, but when exact analytic procedures are not available for the subcomponents, to update them using some version of the Hastings-Metropolis algorithm.

7.5.2. The inverse Wishart prior

One specific idea used in Bayesian inference about covariance matrices is the inverse Wishart distribution, which is the (matrix) inverse of the Wishart distribution, which is a multivariate generalization of the χ^2 distribution. We follow West and Harrison (1996), p. 601.

If V is a symmetric $p \times p$ random positive definite matrix, then we say that V has an inverse Wishart distribution with m degrees of freedom and $p \times p$ symmetric positive definite centering matrix A (notation: $V \sim IW[m, A]$) if the density of V is of the form

$$f(V) \propto |V|^{-p-m/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(mAV^{-1})\right\}. \tag{7.40}$$

Some properties of the inverse Wishart distribution are:

1. $\mathrm{E}\{V^{-1}\} = A^{-1}$.

2. If $m > 2$ then $\mathrm{E}\{V\} = mA/(m-2)$.

3. If v_j is the jth diagonal entry of V, then $v_j^{-1} \sim Gam(m/2, ma_j/2)$ when $Gam(\alpha, \beta)$ denote the gamma distribution with density proportional to $x^{\alpha-1}e^{-\beta a}$.

4. If m is integer, then V^{-1} may be given by $\frac{1}{m}\sum_{j=1}^{m} Z_j Z_j^T$ where $Z_1, ..., Z_m$ are independent with common distribution $MVN[0, A^{-1}]$.

Suppose $X_1, ..., X_n$ are independent p-dimensional vectors with common distribution $MVN[0, V]$. (The whole analysis is easily extended to the case where the observations

have a common unknown mean μ, but we shall not need that case here.) If V has the prior distribution $IW[m, A]$, then the joint density of V and $X_1, ..., X_n$ is proportional to

$$\begin{aligned} &|V|^{-p-m/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(mAV^{-1})\right\} \cdot |V|^{-n/2} \exp\left\{-\frac{1}{2}\sum_i X_i^T V^{-1} X_i\right\} \\ &\quad = |V|^{-p-(m+n)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\{(mA + \sum_i X_i X_i^T)V^{-1}\}\right] \end{aligned} \tag{7.41}$$

which uses the identity $X_i^T V^{-1} X_i = \mathrm{tr}(X_i^T V^{-1} X_i) = \mathrm{tr}(X_i X_i^T V^{-1})$ which in turn follows from the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, which is true whenever the matrices AB and BA are both defined. Comparing (7.41) and (7.40), we see that the posterior distribution of V given $X_1, ..., X_n$ is of the form

$$IW\left[m + n, \frac{mA + \sum X_i X_I^T}{m + n}\right]. \tag{7.42}$$

7.5.3. Bayesian analysis for the state space model.

Suppose (7.1) holds with $F_t = F$, $G_t = G$, $V_t = V$ and $W_t = W$, all independent of t, and continue to assume for the moment that F and G are known. (The case where they are unknown is considered at the end of this subsection.) The problem is too complicated for a direct conjugate-prior Bayesian analysis, but by combining the inverse Wishart prior with the Gibbs sampler, we can construct a Bayesian analysis, as follows.

Suppose $V \sim IW[m_v, A_v]$, $W \sim IW[m_w, A_w]$, independently of each other and of all the observations. We apply a Gibbs sampler in which each of V, W and $S = (S_1, ..., S_T)$ is updated conditionally on the other two. In more detail, this means iterating among the following three steps:

Step 1. Suppose V and W are known. We want to generate a Monte Carlo sample from S, conditionally on V, W and the observed data $\mathcal{X}^T$. However, we can do this using Remark 2 from section 7.3: first generate S_T using the Kalman filter equations, then generate S_{T-1} from the conditional distribution using $\mathcal{X}^T$ and S_T, and so on back to S_1.

Step 2. Suppose S is known and we want to update W. We can compute the values of $w_2 = S_2 - GS_1$, $w_3 = S_3 - GS_2$,..., and so on to $w_T = S_T - GS_{T-1}$, and then compute the posterior distribution of W as

$$IW\left[m_w + T - 1, \frac{m_w A_w + \sum_2^T w_j w_j^T}{m_w + T - 1}\right].$$

Step 3. Suppose S is known and we want to update V. We can compute the values of $v_1 = X_1 - FS_1$, $v_2 = X_2 - FS_2$,..., and so on to $v_T = X_T - FS_T$, and then compute the posterior distribution of V as

$$IW\left[m_v + T, \frac{m_v A_v + \sum_1^T v_j v_j^T}{m_v + T}\right].$$

Steps 1–3 are repeated as many times as needed. At the end, we compute posterior distributions by averaging over the simulations (after possibly discarding some initial warm-up sample). As an example, suppose we want the final posterior distribution of S_T, taking into account the uncertainty of V and W. If we denote the conditional expectation and covariance matrix of S_T, given $\mathcal{X}^T$, V and W, by $\hat{S}_T(V,W)$ and $P(V,W)$, then the conditional density of S_T is the normal density with mean $\hat{S}_T(V,W)$ and covariance matrix $P(V,W)$ — the overall posterior density is obtained by *averaging* this conditional density with respect to V and W. In practice this is usually performed by simply averaging over the sample values of V and W from the Monte Carlo simulation.

Now let us consider the case in which $F_t \equiv F$ and $G_t \equiv G$ are also unknown in (7.1) — assume they may be written $F(\psi)$ and $G(\psi)$ where ψ is some finite-dimensional parameter vector. In this case we again update S, V and W successively, but add a Step 4 in which we also update ψ. Assuming a prior density $\pi(\psi)$, independent of the prior distributions for V and W and of the observations, this fourth step consists of generating an updated value ψ from the conditional density given $(S, V, W, \mathcal{X}^T)$, which may be expressed, up to a constant of proportionality, as

$$\pi(\psi)\prod_{t=1}^{T} p_v(X_t - F(\psi)S_t; V)\cdot\prod p_w(S_t - G(\psi)S_{t-1}; W), \tag{7.43}$$

$p_v(\cdot; V)$ and $p_w(\cdot; W)$ denoting the densities of v_t and w_t respectively as functions of covariance matrices V and W.

In most cases of practical interest, such as those involving ARIMA models using the representations in section 7.4, the functions $F(\psi)$ and $G(\psi)$ will be too complicated for the density (7.43) to be sampled directly. However, this is where the Metropolis-Hastings sampler comes in: in place of an exact draw from the conditional density of ψ, we update ψ using one or several iterations of the Metropolis-Hastings sampler.

7.6 Conditionally Gaussian Dynamic Models

In recent years, there have been many extensions of time series analysis to allow for such features as nonlinearity, nonstationarity and non-Gaussianity. Although many different approaches have been proposed, some of the most powerful are based on the ideas of state space models combined with Monte Carlo analysis to facilitate the handling of

nonstandard distributions. *Conditionally Gaussian dynamic models* are models in which a state space process evolves under Gaussian assumptions as in the Kalman filter, but there are additional components to the model which give rise to nonlinear or nonstationary features. A recent paper by Cargnoni, Müller and West (1997) has outlined many of these ideas, and we follow that paper here.

Although the application described by Cargnoni *et al.* is somewhat specialized, we describe it here as an illustration of the kind of problem this methodology can handle. It is concerned with data from the Italian school system, in which the focus of interest is the proportion of students in each grade who either repeat that grade, advance to the next grade, or leave the system entirely. Data are collected over several grades for a number of years, so the basic problem is one of time series in which the outcomes are trinomial variables and with the extra complication of dependence among several parallel time series corresponding to the different grades.

To introduce some notation, let n_{it} denote the number of students in grade i in year t, and let y_{i1t} denote the number who remain in the same grade into year $t+1$, y_{i2t} the number who advance to the next grade, and y_{i3t} the number who leave the system. The vector $(y_{i1t}, y_{i2t}, y_{i3t})$ is denoted y_{it}. The natural model is multinomial:

$$y_{it} \mid n_{it} \sim Mu(n_{it}, \pi_{it}),$$

where $\pi_{it} = (\pi_{i1t}, \pi_{i2t}, \pi_{i3t})$ is a vector of probabilities subject to $\sum_j \pi_{ijt} = 1$ for each i and t. In general, we consider a multinomial distribution with $r+1$ cells in which the vector of cell probabilities is $\pi_{it} = (\pi_{i1t}, ..., \pi_{i,r+1,t})$ — we shall formulate a probability model for $(\pi_{i1t}, ..., \pi_{irt})$ leaving $\pi_{i,r+1,t}$ to be defined as $1 - \sum_{j=1}^{r} \pi_{ijt}$.

The first step in formulating a model is to define some transformation h from $[0,1]$ to the real line — examples are $h(\pi) = \log \pi/(1-\pi)$ or $h(\pi) = 2 \arcsin \sqrt{\pi}$. Define $\eta_{it} = (\eta_{i1t}, ..., \eta_{1rt})$ where $\eta_{ijt} = h(\pi_{ijt})$, $1 \leq j \leq r$. Assume the vectors y_{it} are conditionally independent for each i and t given η_{it}. We assume all the $\{\eta_{it}\}$ for each t are combined into a single vector η_t which satisfies the state space model

$$\begin{aligned} \eta_t &= F_t \theta_t + v_t, \\ \theta_t &= H_t \theta_{t-1} + w_t, \\ v_t &\sim N[0, V] \\ w_t &\sim N[0, W] \end{aligned} \tag{7.44}$$

with F_t and H_t known, v_t and w_t independent random errors and unknown covariance matrices V and W. Also assume some prior density exists for (θ_0, V, W).

The Bayesian analysis of this model consists primarily of drawing a sample from the posterior distribution of the parameters (η, θ, V, W) conditionally on the data y. As in the Gibbs sampler, this is broken down into several stages and a sample drawn at each stage conditionally on all the variables in the other stages:

(i) Updating θ given (η, V, W): this uses the Kalman filter, with the backwards sequential generation method employed in section 7.5.

(ii) Updating V and W given (η, θ): if we assume inverse Wishart priors for V and W (independent of each other and of everything else), then the posterior distributions of V and W, given η and θ, are also of inverse Wishart form. This is analogous to Steps 2 and 3 of section 7.5.

(iii) Updating η given (y, θ, V, W): the key point here is that the values of η_t for different t depend on each other only through the values of θ_t, and therefore, if we condition on the entire sequence of $\{\theta_t\}$, the posterior distributions of the individual η_t, for $t = 1, 2, ..., T$, are independent. In other words, it suffices to update each η_t, one at a time, without having to consider the joint distribution of all the $\{\eta_t\}$ simultaneously — this was a disadvantage of some earlier attempts to analyze this kind of model by Monte Carlo sampling. As for the details of the sampling, any form of Hastings-Metropolis algorithm would presumably suffice for the job: Cargnoni *et al.* in fact used an independence Metropolis sampler in which they used a normal approximation to the posterior distribution as the trial distribution for the sampler.

The brief description given here does not reflect all the subtleties of the Cargnoni *et al.* procedure. They actually assumed that the values of η_{it}, as i varies over the different grades, are conditionally independent given the corresponding θ_{it} values, and this allows for some improvement of the algorithm by separate updating of subcomponents, but the preceding discussion gives the essential ideas behind their method.

In the actual analysis of the Italian school data, they used the transformation $h(\pi) = 2 \arcsin \sqrt{\pi}$ and adopted a model for the $\{\eta_{it}\}$ of the form

$$\eta_{it} = \mu_t + \gamma_{it} + v_{it},$$

where μ_t is an overall trend (common to all grades), γ_{it} is a grade-specific component and v_{it} is a random component. They assumed, in effect, a cubic trend in time for μ_t and a linear trend for each γ_{it}, updated through the equations

$$\begin{aligned} \mu_t &= \mu_{t-1} + \delta_{t-1} + \epsilon_{t-1} + w_{\mu t}, \\ \delta_t &= \delta_{t-1} + \epsilon_{t-1} + w_{\delta t}, \\ \epsilon_t &= \epsilon_{t-1} + w_{\epsilon t}, \\ \gamma_{it} &= \gamma_{i,t-1} + w_{\gamma it}, \end{aligned} \tag{7.45}$$

with independent normal assumptions for the error terms $w_{\mu t}, w_{\delta t}, w_{\epsilon t}, w_{\gamma it}$. For identifiability purposes, with I grades being considered together, the parameters $\{\gamma_{it}\}$ were updated according to (7.45) only for $i = 1, ..., I-1$, γ_{It} being defined as $-\sum_1^{I-1} \gamma_{it}$. A particular feature of their algorithm was the ability to track trends across time of proportions of students in each of the three categories (repeated a grade, moved to next grade,

left system) either within each grade or aggregated across all grades. However, it seems to be a whole new set of questions to decide how to verify a model of the structure defined here.

7.7. Models for financial time series.

Financial time series are series of stock prices, stock price indices such as the S&P or Dow Jones, spot interest rates, currency exchange rates, etc. All of these types of series exhibit certain characteristic features which distinguish them from other kinds of time series and which have led to a tremendous outgrowth of specialized models and methods. The most popular models subdivide into two classes. One of these leads to models analogous to the ARMA structure of traditional time series models but with the autoregressive and moving average components acting on the variances of the process, as well as or instead of the means. These models are grouped together under the heading ARCH (for *autoregressive conditionally heteroscedatic*) models. The other kinds of models are *stochastic volatility* or SV models in which there is some unobserved process known as the volatility which directly influences the variance of the observed series. SV models share many of the characteristics of state space models, hence their inclusion in the present chapter.

The following discussion is based largely on the superb review chapter by Shephard (1996), to which we refer for a more detailed account.

7.7.1. Basic facts about financial time series

Nearly always, we analyze the daily returns $y_t = 100 \log(x_t/x_{t-1})$ where x_t is the price on day t. A number of empirical observations (known as "stylized facts") seem common to all series of this type:

- Symmetric distribution about the mean
- Little autocorrelation among the values of y_t
- Strong autocorrelation among the values of y_t^2 or $|y_t|$
- Long-tailed distributions
- Variable volatility: in other words, the local variance of the process changes substantially across time.

Much of the interest in financial time series stems from the trade in *options* of various types. The best known example of an option is the *European call option* which gives its holder the right to buy an asset at a given price K on a given date $T + v$, where T is today's date and v is some specified period into the future. The famous theory of Black

and Scholes (1973) showed how to price an option under the assumption that an underlying asset S follows a stochastic differential equation of the form

$$dS = \mu S dt + \sigma S dz, \tag{7.46}$$

where μ is an instantaneous mean return and σ is the volatility. By constructing an imaginary *riskless portfolio* based on continuously trading between the asset S and an alternative asset which can be borrowed at a fixed interest rate, they contructed a differential equation for the value of an option as a function of the time to maturity of that option. The *Black-Scholes option price formula* shows that the price of an option depends on σ but not on μ. This leads to one method of estimating σ: one simply observes the actual option price and solves the Black-Scholes equation to determine σ. This is known as the *implied volatility* estimate. However, a more sophisticated approach assumes that the volatility is not constant and estimates it using various models, and our main purpose here is to study some of these models.

Most models of financial time series are of the general structure of

$$y_t \mid z_t \sim N[\mu_t, \sigma_t^2], \tag{7.47}$$

where z_t is some set of conditioning random variables (possibly, but not necessarily, lagged values of $\{y_t\}$) and μ_t and σ_t^2 are some function of z_t. As canonical examples of each of the two main types, we mention the ARCH model in which z_t is equated with $\mathcal{Y}^{t-1}$, the observed series up to time $t-1$, and

$$\sigma^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + ... + \alpha_p y_{t-p}^2. \tag{7.48}$$

The simplest SV model is of the form

$$y_t \mid h_t \sim N[0, e^{h_t}], \qquad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t, \qquad \eta_t \sim N[0, \sigma_\eta^2], \tag{7.49}$$

all the normal random variables being conditionally independent given the underlying means and variances. The model of (7.49) has a state space form reminiscent of many of the models considered in this chapter, but with the critical difference that the dependence of the state variable h_t on the observation y_t lies in the variance of y_t rather than the mean. The distinction immediately wipes out any direct attempt to analyze the model through the Kalman filter which has been our main focus in this chapter.

7.7.2. ARCH models.

The simplest ARCH model, ARCH(1), is of the form

$$y_t = \epsilon_t \sigma_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2, \qquad t = 1, ..., T, \tag{7.50}$$

with $\epsilon_t \sim N[0,1]$ (mutually independent). We need α_0 and α_1 to be positive to ensure $\sigma_t^2 > 0$. If $3\alpha_1^2 < 1$ then y_t^2 is covariance stationary with autocorrelation function $\rho_{y_t^2}(s) =$

α_1^s, and in this case y_t is leptokurtic (tails longer than normal). The condition $3\alpha_1^2 < 1$ is needed primarily to ensure that the variance of y_t^2 is finite. Precise conditions for $\{y_t\}$ to be strictly stationary seem to be harder to pin down but a sufficient condition is $\alpha_1 < 3.5622$.

One of the key features of the ARCH(1) and indeed of all ARCH models is that the conditional distribution of y_t given $\mathcal{Y}^{t-1}$ is easily written down explicitly and this allows us to formulate directly the prediction error decomposition for such a process. This makes maximum likelihood estimation straightforward and also implies a direct approach to the forecasting of future values. Pinning down the properties of maximum likelihood estimators is not so easy — for example, asymptotic normality of the estimators is not easy to prove though it is known to be true for a very wide class of processes. A more practical difficulty is that the likelihood surface tends to be flat so that even in this simplest form of the model, the maximum likelihood estimates of α_0 and α_1 can be quite imprecise.

The ARCH model can be thought of as an autoregressive model in y_t^2. An obvious extension of this idea is to consider adding moving average ocmponents as well, and this leads to the generalized ARCH or GARCH class. The simplest such model is the GARCH(1,1), defined by

$$y_t = \epsilon_t \sigma_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \qquad t = 1, ..., T. \tag{7.51}$$

The series $\{y_t\}$ is covariance stationary if $\alpha_1 + \beta_1 < 1$. However it is also well defined (and stationary) if $\alpha_1 + \beta_1 = 1$ and this is an important special case known as integrated GARCH or IGARCH. It corresponds to having persistent shocks in the system.

The simplest estimation scheme for the GARCH(1,1) model uses some initial sample of observations (say 20) to come up with a crude initial estimate of σ_t^2, and then use maximum likelihood estimation based on the prediction error decomposition.

A generalization of the GARCH(1,1) model is the GARCH(p, q) model, in which (7.51) is extended to lagged terms in $y_{t-1}^2, ..., y_{t-p}^2$ and $\sigma_{t-1}^2, ..., \sigma_{t-q}^2$.

A number of other forms of GARCH model are mentioned by Shephard (1996). These include:

Log GARCH:

$$y_t^2 = \epsilon_t^2 e^{h_t}, \qquad h_t = \gamma_0 + \gamma_1 \log y_{t-1}^2.$$

A disadvantage of this model is that it may result in y_t close to 0 or even exactly 0. A possible resolution of this problem is to replace $\gamma_1 \log y_{t-1}^2$ by $\gamma_1 \log\{\max y_{t-1}^2, c\}$ for some suitable $c > 0$, but it seems that this idea has not caught on in the finance literature.

Exponential GARCH:

This is a variant on the log GARCH model in which the equation for h_t is replaced by

$$h_t = \gamma_0 + \gamma_1 h_{t-1} + g(\epsilon_{t-1}), \qquad g(x) = wx + \lambda(|x| - E|x|).$$

An advantage of this model over some of the others is that it responds asymmetrically to shocks. This corresponds to the empirically observed phenomenon that for many assets, the volatility responds more rapidly to sudden drops in the price than to sudden rises. The model is stationary if and only if $|\gamma_1| < 1$. One distinction from traditional ARCH models is that it is possible for $\rho_{y_t^2}(s)$ to be negative for some lags s. Like the other forms of ARCH model, the likelihood is easily computed via a prediction error decomposition. Properties of the maximum likelihood estimates are not easy to establish, but it seems clear that asymptotic normality holds if $|\gamma_1| < 1$.

Decomposing IGARCH models

This idea works by decomposing both the conditional mean and conditional variance of y_t so allowing both persistent and transitory effects to be modeled (the conditional mean terms being the persistent effects). A typical model is

$$\sigma_t^2 = \mu_t + \alpha_1(y_{t-1}^2 - \mu_t) + \beta_1(\sigma_{t-1}^2 - \mu_t),$$
$$\mu_t = w + \mu_{t-2} + \phi(y_{t-1}^2 - \sigma_{t-1}^2).$$

Absolute residuals ARCH

$$\sigma_t = \alpha_0 + \alpha_1|y_{t-1}|$$

Nonlinear ARCH (NARCH)

$$\sigma_t^2 = \alpha_0 + \alpha_1|y_{t-1} - k|^\gamma,$$

$k = 0$ being the symmetric case, $k \neq 0$ asymmetric.

Partially nonparametric ARCH:

This allows for a nonlinear and empirically determined relationship between σ_t^2 and y_{t-1}, a simple version being a linear spline representation,

$$\sigma_t^2 = \alpha_0 + \sum_{j=0}^{m^+} \alpha_1^{+j} I(y_{t-1} - \tau_j > 0)(y_{t-1} - \tau_j) + \sum_{j=0}^{m^-} \alpha_1^{-j} I(y_{t-1} - \tau_j < 0)(y_{t-1} - \tau_j),$$

in which $I(\cdot)$ are indicator functions and $(\tau_{-m}, ..., \tau_m)$ is a sequence of knots typically set as $\tau_j = j\sqrt{Var(y_t)}$.

Quadratic ARCH or QARCH

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \alpha_1^* y_{t-1}$$

with contraints on the coefficients to ensure $\sigma_t^2 > 0$. This is another model used to capture asymmetry.

Threshold ARCH

Taking the asymmetry theme further,

$$\sigma_t^2 = \alpha_0 + \alpha_1^+ I(y_{t-1} > 0) y_{t-1}^2 + \alpha_1^- I(y_{t-1} < 0) y_{t-1}^2.$$

ARCH in mean (ARCH-M)

$$y_t = g(\sigma_t^2, \theta) + \epsilon_t \sigma_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 \{y_{t-1} - g(\sigma_{t-1}^2, \theta)\}^2,$$

with, for example, $g(\sigma^2, \theta) = \mu_o + \mu_1 \sigma_t^2$. This is intended to reflect a direct relationship between the level of returns of a risky asset and the level of volatility.

Empirical evaluation of ARCH models

Shephard (1996) illustrated several ARCH models by evaluating them on four financial time series, two of them consisting of currency exchange rates (yen/pound and Deutsch Mark/pound), two of stock prices incides (Nikkei 500 and FTSE 100), each running from 1986–1994. He fitted GARCH(1,1) and EGARCH models both assuming the innovations $\{\epsilon_t\}$ are normal, and assuming they have a student's t distribution with degrees of freedom ν to be estimated. He also compared them with "benchmark" models of i.i.d. normal and i.i.d. t_ν variables, the motivation behind the latter being that it would allow for the long-tailed distributions which are observed in real data, but not for time-series dependence.

The two "benchmark" models were definitively rejected by a Box-Ljung test applied to the *squares* of the observations (recall Chapter 4 for a description of the Box-Ljung and the older Box-Pierce tests). The ordinary GARCH model (with normal errors) passed this test, but failed one based on the kurtosis of the transformed innovations (a test of long-tailedness). Only by assuming a GARCH model with t-distributed innovations, with ν typically about 4, was reasonable fit obtained by both tests. Similar results were obtained for the EGARCH model, in which reasonable fit was obtained only by allowing the innovations to have a t distribution with small ν. Comparing GARCH and EGARCH, perhaps the most interesting comparison is in terms of the negative log likelihoods achieved by both models:

Series	GARCH NLLH	EGARCH NLLH
Nikkei 500	2836	2795
FTSE 100	2595	2593
DM/pound	945.3	942
Yen/pound	1879	1879

Table 7.1. Evaluation of the negative log likelihood for GARCH and EGARCH models applied to four financial time series, each including t-distributed innovations. Results from Shephard (1996).

In effect, what the EGARCH model is doing that the GARCH is not is allowing for asymmetry in the response to shocks, and Shephard commented that this seems to be a significant effect for all series, commenting that "this ... is a standard result for equities... (but) it is non-standard for currencies where the asymmetry effects are usually not significant". In fact, based on Table 7.1, I have a different interpretation from Shephard: in the precise form considered by Shephard, the EGARCH model has two more parameters than the GARCH model, and taking this into account, only the Nikkei 500 series shows a large improvement in the log likelihood. It would be interesting to establish to what extent this effect was due entirely to the data a few days either side of October 19, 1987.

7.7.3. Stochastic volatility

The most popular form of SV model, the *log-normal SV model*, is based on rewriting (7.49) in the form

$$y_t = \epsilon_t \exp(h_t/2), \qquad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t, \tag{7.52}$$

in which $\epsilon_t \sim N[0,1]$, $\eta_t \sim N[0,\sigma_\eta^2]$ are mutually independent for all t. Some elementary properties are easy to derive:

(i) $\{h_t\}$ is strongly and weakly stationary if and only if $|\gamma_1| < 1$ and in that case,

$$\mu_h = \mathrm{E}\{h_t\} = \frac{\gamma_0}{1-\gamma_1}, \qquad \sigma_h^2 = \mathrm{Var}\{h_t\} = \frac{\sigma_\eta^2}{1-\gamma_1^2}. \tag{7.53}$$

(ii) If $\{h_t\}$ is stationary then $\{y_t\}$, being the product of two stationary processes, is too, and we can compute, for r even,

$$\begin{aligned} \mathrm{E}\{y_t^r\} &= \mathrm{E}\{\epsilon_t^r\}\mathrm{E}\left\{\exp\left(\frac{r h_t}{2}\right)\right\} \\ &= \frac{r!}{2^{r/2}(r/2)!}\exp\left(\frac{r\mu_h}{2} + \frac{r^2\sigma_h^2}{8}\right), \end{aligned} \tag{7.54}$$

the odd-numbered moments all being 0. In particular, it follows from (7.54) that the kurtosis $\mathrm{E}\{y_t^4\}/(\mathrm{E}\{y_t^2\})^2$ is $3\exp(\sigma_h^2) \geq 3$, proving that the SV model has fatter tails than a normal distribution.

(iii) The autocorrelation function of y_t^2 can be computed as

$$\mathrm{Corr}\{y_t^2, y_{t-s}^2\} = \frac{\exp(\sigma_h^2\gamma_1^s) - 1}{3\exp(\sigma_h^2) - 1}, \tag{7.55}$$

which for small σ_h^2 is approximately proportional to γ_1^s, as in the ARMA(1,1) model. Thus the SV model behaves similarly to the GARCH(1,1) model. It is also possible to calculate

$$\mathrm{Corr}\{\log y_t^2, \log y_{t-s}^2\} = \frac{\gamma_1^s}{1 + 4.93/\sigma_h^2}. \tag{7.56}$$

The derivation of (7.56) uses the representation

$$\log y_t^2 = h_t + \log \epsilon_t^2, \qquad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t. \tag{7.57}$$

We now consider several methods of estimating the log-normal SV model:

(i) Generalized method of moments (GMM)

Clearly it would be possible to combine (7.53), (7.54) for $r = 2$ and 4, and one of (7.55) or (7.56), to find three equations for the three unknowns γ_0, γ_1 and σ_η^2, and then to solve those equations using sample values of the mean, variance and first-order autocorrelation of either y_t^2 or $\log\, y_t^2$, so obtaining method-of-moments estimators. There are, however, many more variables whose theoretical expectations are computable, so the question arises of which three to use. GMM methods use a larger number of moment equations than there are parameters to estimate, typically using a least-squares type of criterion to determine the best fit between theoretical and empirical moments. However there also seem to be a number of disadvantages to GMM methods when applied to SV models – in particular, Shephard remarks that when γ_1 is close to 1, as can be expected to happen for many financial series with high persistence, the GMM estimates will behave poorly.

(ii) Quasi-likelihood

If we ignore the fact that $\log \epsilon_t^2$ does not have a normal distribution, (7.57) has the form of a standard Kalman filter model and hence can be solved (including the prediction error decomposition) to form a "likelihood function" for the unknown parameters. This is known as a *quasi-likelihood* (QL) because it does not use the correct distribution for the log ϵ_t^2 variables, but nevertheless, estimation based on maximizing the QL has many desirable properties, including consistency, asymptotic normality and in many cases an asymptotic efficiency that is close to 1 relative to the true MLE. The main practical difficulty in applying QL methods is that the usual asymptotic form for the covariance matrix of the estimators is not valid and some adjustment must be made. The details of this were worked out by Harvey, Ruiz and Shephard (1994).

(iii) Estimating the mode of the distribution of $\{h_t\}$

Clearly, the main difficulty in applying maximum likelihood methods for the SV model is that the volatility process $\{h_t\}$ is not observed. The QL method gets around this problem by using the Kalman filter to compute an approximation to $\mathrm{E}\{h_t \mid \mathcal{Y}^T\}$, but this is not a true conditional expectation because it is not based on the true distribution of the $\{\epsilon_t\}$ variables. Durbin and Koopman (1997) developed an ingenious alternative approach to this problem based on a linear approximation to $(\partial/\partial h_t) \log f(y_t \mid h_t)$, using the Kalman filter to solve the resulting approximate linear system. The method leads, in effect, to the conditional mode of h_t given $\mathcal{Y}^T$, which may then be used to calculate the likelihood function. Thus one use for this approach is to give another way of computing approximate

maximum likelihood estimates, but this is not the only use, because independently of which method is used to estimate the model parameters, having a good reconstruction of the $\{h_t\}$ process is a useful product in its own right.

(iv) Monte Carlo approaches

Despite the existence of these various approximate methods, most of the recent research has been in and around some form of Monte Carlo estimation technique. All of these methods are ultimately built around the equation

$$p(y_1, ..., y_T; \gamma_0, \gamma_1, \sigma_\eta^2) = \int p(h_1, ..., h_T; \gamma_0, \gamma_1, \sigma_\eta^2) \prod_{t=1}^{T} p(y_t \mid h_t), \tag{7.58}$$

p denoting a generic probability density, but direct evaluation of (7.58) using a Monte Carlo generation of $(h_1, ..., h_T)$ would be much too inefficient to be applied directly. One idea discussed by Shephard (1996) is to use importance sampling (Ripley, 1987) as a variance reduction technique, but this seems to have fallen out of favor in comparison with MCMC techniques. The key issue in this case is to find efficient algorithms for updating the sequence $\{h_1, ..., h_T\}$, conditionally on all the other variables in the model, which are both easy to implement and which will provide a reasonable rate of convergence to the true conditional distribution.

One idea mentioned by Shephard (1996) and described in more detail by Shephard and Pitt (1997) is to use an approximate Gaussian density to obtain a good trial distribution for the Metropolis-Hastings sampler. First, let us rewrite (7.52) in yet another way as

$$y_t = \epsilon_t \beta \exp\left(\frac{\alpha_t}{2}\right), \qquad \alpha_{t+1} = \phi\alpha_t + \eta_t, \qquad \epsilon_t \sim N[0,1], \quad \eta_t \sim N[0, \sigma_\eta^2], \tag{7.59}$$

and let us compute the conditional distribution of α_t given α_{t-1} and α_{t+1}. The joint density of $(\alpha_{t-1}, \alpha_t, \alpha_{t+1})$ is proportional to

$$\exp\left\{-\frac{(\alpha_t - \phi\alpha_{t-1})^2}{2\sigma_\eta^2} - \frac{(\alpha_{t+1} - \phi\alpha_t)^2}{2\sigma_\eta^2}\right\} \tag{7.60}$$

and the key point is to complete the square with respect to α_t in the exponent of (7.60) so that is the same, up to a constant of proportionality, as

$$\exp\left\{-\frac{(\alpha_t - \mu_t)^2}{2\sigma_t^2}\right\}$$

for suitable μ_t amd σ_t^2. It is readily verified that this is achieved by

$$\mu_t = \frac{\phi}{1+\phi^2}(\alpha_{t-1} + \alpha_{t+1}), \qquad \sigma_t^2 = \frac{\sigma_\eta^2}{1+\phi^2}.$$

Next, let us write the joint density of (α_t, y_t) given $(\alpha_{t-1}, \alpha_{t+1})$, up to proportionality, as

$$\exp\left\{-\frac{(\alpha_t-\mu_t)^2}{2\sigma_t^2}\right\}\cdot\exp\left(-\frac{\alpha_t}{2}\right)\cdot\exp\left(-\frac{y_t^2e^{-\alpha_t}}{2\beta^2}\right)$$

and rewrite the exponent of this, after Taylor expanding $e^{-\alpha_t}$ around $e^{-\mu_t}$, in the form

$$-\frac{(\alpha_t-\mu_t)^2}{2\sigma_t^2}-\frac{\alpha_t}{2}-\frac{y_t^2}{2\beta^2}e^{-\mu_t}\left\{1+(\alpha_t-\mu_t)+\frac{1}{2}(\alpha_t-\mu_t)^2+O(|\alpha_t-\mu_t|^3)\right\},$$

and if we complete the square with respect to α_t in this expression, ignoring the $O(|\alpha_t - \mu_t|^3)$ term, we obtain the approximation

$$\alpha_t \mid \alpha_{t-1}, \alpha_{t+1}, y_t \sim N[\mu_t^*, \sigma_t^{*2}], \tag{7.61}$$

where

$$\frac{1}{\sigma_t^{*2}} = \frac{1}{\sigma_t{}^2}+\frac{y_t^2e^{-\mu_t}}{2\beta^2},$$
$$\mu_t^* = \sigma_t^{*2}\left\{\frac{\mu_t}{\sigma_t^2}-\frac{1}{2}+\frac{y_t^2(1+\mu_t)e^{-\mu_t}}{2\beta^2}\right\}.$$

The idea is then to use the normal approximation (7.61) to generate a trial step in the Metropolis-Hastings algorithm, but to use the exact conditional density $p(\alpha_t \mid \alpha_{t-1}, \alpha_{t+1}, y_t)$ in deciding whether to accept or reject the new value. Thus, although the approximation (7.61) is a key step in defining an efficient sampler, the actual algorithm is still faithful to the exact conditional density.

The difficulty with this kind of procedure lies not in the quality of the approximation (7.61), but in the whole idea of updating the $\{\alpha_t\}$ vector one variable at a time. In many practical situations ϕ is close to 1; the entire sequence $\{\alpha_t\}$ is highly correlated, and it is going to take a very large number of single-variable updates before the entire vector achieves reasonable coverage of its true conditional distribution. Two procedures which have been proposed for dealing with this difficulty are

(a) *Multimove samplers.* Return to equation (7.57), and suppose we approximate the distribution of $\log \epsilon_t^2$ by a mixture of J normal variables with mean-variance parameters $\{(m_j, s_j^2),\ 1 \le j \le J\}$. Let w_t be an indicator varable taking values $1, 2, ..., J$, so that $w_t = j$ means that ϵ_t is drawn from $N[m_j, s_j^2]$. Conditionally on $\{w_t\}$, the sequence of $\{\epsilon_t\}$ is normal, and so its conditional distribution given $\{y_t\}$ can be computed using the Kalman filter algorithm. A secondary randomization then allows the $\{w_t\}$ to be updated, *but* the structure of the model is such that each w_t is conditionally independent of all w_s, $s \neq t$, given ϵ_t, so the w_t variables can be updated one at a time without any worries about their being highly correlated. This is similar to the main idea behind conditionally Gaussian dynamic models, mentioned in section 7.6. The difficulty with this method is that we do not know just how adequate the mixture-of-Gaussians approximation is to the true distribution of $\log \epsilon_t^2$.

(b) *Block samplers.* The main new idea introduced by Shephard and Pitt (1997) is that instead of using the approximation (7.61) to update α_t one variable at a time, it is possible to use a similarly motivated multivariate normal approximation to update a whole sequence of $\alpha_t, \alpha_{t+1}, ..., \alpha_{t+k}$ given $\alpha_{t-1}, \alpha_{t+k+1}$ and $y_t, ..., y_{t+k}$. The endpoints α_{t-1} and α_{t+k+1} they call *knots*, and these are chosen stochastically from one run to the next. A key decision is the mean number of knots K. At one extreme we could take $K = 1$, meaning updating the entire sequence $\alpha_1, ..., \alpha_T$ in a single step, using a multivariate normal approximation to the joint conditional distribution given $y_1, ..., y_T$. However it is unlikely that the normal approximation will work well in very high dimensions, so this algorithm is likely to result in far too high a rejection rate at the accept-reject step. The other extreme is represented by $K = T$, in other words, update one variable at a time, but as already noted, this is likely to be inefficient because of high correlations among the individual α_t variables. Identifying an "optimal" K seems likely to be a tough theoretical problem, but Shephard and Pitt suggest that it may not be too important in practice. They argue that good efficiency gains are obtained across a wide range of K values and recommend block sizes between 50 and 500 in practice. They also emphasize the main advantage of *random* block sizes — because there will sometimes be short blocks, there will be some Metropolis steps for which the acceptance probability is high, and this help to avoid the algorithm getting stuck.

Estimating unknown parameters

The preceding discussion has concentrated on either estimating or reconstructing the conditional distribution of $\{h_t,\ t = 1, ..., T\}$ given $\{y_t,\ t = 1, ..., T\}$. The MCMC methods are easily incorporated into a Bayesian algorithm for estimating unknown parameter vector θ through the strategy of alternately updating the $\{h_t\}$ variables and θ as part of a Gibbs sampler. Computing maximum likelihood estimates is a little harder since we need a simulated version of the likelihood function; Shephard (1996) describes a way of doing this via a simulated version of the EM algorithm and Shephard and Pitt (1997) describe an importance sampling algorithm, but we shall not go into the details of these.

Extensions of the SV model

It is logical to try to extend the SV model by including other components in h_t, for example, to a structure of the form

$$y_t = \epsilon_t \exp(z_t^T h_t/2), \qquad h_{t+1} = T_t h_t + \eta_t, \qquad \eta_t \sim N[0, H_t],$$

with h_t a d-dimensional vector, T_t and H_t given $d \times d$ covariance matrices and z_t a given $d \times 1$ vector. Indeed the main model considered by Shephard and Pitt (1997) is even more general than this. One specific possibility is with $d = 2$,

$$z_t = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad T_t = \begin{pmatrix} \gamma_1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad H_t = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix},$$

so that the stochastic volatility is a sum of two components, one an AR(1) process as in the standard log-normal SV model, and the other a random walk. Thus, this potentially allows for much more persistent volatility than the standard SV model (but presumably with $\sigma_2^2 << \sigma_1^2$ so that the persistent component does not dominate all the other sources of variation). Clearly, there are many other possibilities along these lines.

Empirical studies

Shephard (1996) compared the GARCH models with both normal and t-distributed innovations with a simple log-normal SV model. For estimation purposes he used a simulated maximum likelihood approach in order to compare the resulting log likelihood directly with that of the GARCH model; he also computed a Box-Ljung statistic based on estimated innovations as a test of the overall fit of the model. For all four series discussed in section 7.7.2, Shephard found that the SV model was a better fit than the normal GARCH model but not as good as the t GARCH model. This is true judging both by the value of the log likelihood and the Box-Ljung statistic. The conclusion is that the SV model goes some way towards explaining the empirical observation of long-tailedness, but still does not do as well as the GARCH model with t-distributed innovations. A possible further extension of the SV model would be to allow either the ϵ_t or the η_t variables to have t distributions. This could be fitted within the conditionally Gaussian framework by noting that (for example) if $\epsilon_t \sim t_\nu$ then we can write $\epsilon_t = \xi_t/\omega_t$ with ξ_t, ω_t independent, $\xi_t \sim N[0,1]$, $\nu\omega_t^2 \sim \chi_\nu^2$, and alternating between successively updating the $\{\xi_t\}$ and the $\{\omega_t\}$. This idea is mentioned by Shephard and Pitt (1999).

Another empirical study is in the paper by Shephard and Pitt (1999) — they fitted univariate SV models (with Gaussian innovations) to five series of currency exchange rates against the US dollar, the five currencies being the British pound, French franc, Swiss franc, Deutsche Mark and Japanese yen. The lowest value of γ_1 was 0.84 for the yen series, but there was also empirical evidence of long-tailedness. For the four other series, γ_1 was in the range 0.94–0.97.

7.7.4. Multivariate models

In principle we would like to be able to extend all of these models to multivariate time series, since in any problem concerned with the construction of a portfolio, correlations between time series are of considerable importance. This has been considered by Shephard (1996) and by Shephard and Pitt (1999).

Although there have been direct attempts to generalize the ARCH structure to multivariate series, the difficulty is that the models tend to be highly unparsimonious — for example, Shephard (1996, page 42) remarks that for a 5-dimensional model, 465 parameters have to be estimated (yet even 5 dimensions are very few compared with the number of assets held in a typical portfolio). The difficulty is how to reduce the parametrization of the model without imposing unreasonabe restrictions. The most popular models at the present time seem to be those motivated by factor analysis, in which the obsrerved series

are expressed as linear combinations of a much smaller number of unobserved series which act as factors, with some residual noise thrown in.

As an example, consider the main model of Shephard and Pitt (1999). With some change of notation compared with Shephard and Pitt, let us write the model (7.52) as $\mathrm{SV}(\gamma_0, \gamma_1, \sigma_\eta)$. Shephard and Pitt consider a model of the form

$$y_t = \beta f_t + w_t \tag{7.62}$$

in which y_t is $N \times 1$, the factor series f_t is $K \times 1$, and β is a $N \times K$ matrix of *factor loadings*. Here w_t is a set of N independent series which Shephard and Pitt call *idiosyncracies*, relecting features specific to each of the series. The K factor series $f_t^{(k)}$, $1 \le k \le K$ and the N idiosyncracy series $w_t^{(n)}$, $1 \le n \le N$, are assumed to be mutually independent SV processes with

$$w_t^{(j)} \sim SV(\gamma_{0j}, \gamma_{1j}, \sigma_j), \qquad f_t^{(k)} \sim SV(0, \gamma_{1k}^*, \sigma_k^*).$$

Further there are some restrictions on the β matrix, mainly to ease some of the identifiability conditions associated with this kind of model. The main restriction imposed is $\beta_{ii} = 1$ for $1 \le i \le K$.

The fitting of this model is based on a block-updating MCMC scheme similar to Shephard and Pitt (1997) but generalized to the multivariate case. A Bayesian approach is taken with proper priors for all the model parameters. As an example, they fitted a factor model with a single principal factor ($K = 1$) to the five series of currency exchange rates against the US dollar. The main factor was found to explain respectively 57%, 99%, 35%, 84% and 92% of the variability for the pound, DM, yen, SF and FF respectively. This is consistent with the interpretation that the yen moves largely independent of the European currencies, because of the obvious influence of Asian factors which do not affect the other currencies.

REFERENCES

Anderson, T.W. (1984), *An Introduction to Multivariate Analysis* (second edition). Wiley, New York.

Ansley, C.F. (1979), An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika* **66**, 59–65.

Ansley, C.F. and Kohn, R. (1985), Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Ann. Statist.* **13**, 1286–1316.

Ansley, C.F. and Kohn, R. (1986), Prediction mean square error for state space models with estimated parameters. *Biometrika* **73**, 467–474.

Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637–654.

Bloomfield, P. (1976), *Fourier Analysis of Time Series: An Introduction.* John Wiley, New York.

Bloomfield, P., Hurd, H. and Lund, R. (1994), Periodic correlation in stratospheric ozone data. *Journal of Time Series Analysis* **15**, 127–150.

Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis, Forecasting and Control* (second edition). Holden Day, San Francisco.

Brockwell, P.J. and Davis, R.A. (1991), *Time Series: Theory and Methods* (second edition). Springer-Verlag, New York.

Brooks, S.P. (1998), The Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69–100.

Cargnoni, C., Müller, P. and West, M. (1997), Bayesian forecasting of multinominal time series through conditionally Guassian dynamic models. *J. Amer. Statist. Assoc.* **92**, 640–647.

Cressie, N. (1993), *Statistics for Spatial Data.* Second edition, John Wiley, New York.

Durbin, J. and Koopman, S.J. (1997), Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669–684.

Gamerman, D. (1997), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman and Hall, London.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-721.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996), *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Harvey, A.C. and Phillips, G.D.A. (1979), The estimation of regression models with autoregressive-moving average disturbances. *Biometrika* **66**, 49–58.

Harvey, A.C. and Pierse, R.G. (1984), Estimating missing observations in economic time series. *J. Amer. Statist. Assoc.* **79**, 125–131.

Harvey, A.C., Ruiz, E. and Shephard, N. (1994), Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–264.

Harville, D.A. (1974), Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Kalbfleisch, J.D. and Sprott, D.A. (1970), Application of likelihood methods to problems involving large numbers of parameters (with discussion). *J.R. Statist. Soc. B* **32**, 175–208.

Kohn, R. and Ansley, C.F. (1985), Efficient estimation and prediction in time series regresion models. *Biometrika* **72**, 694–697.

Kohn, R. and Ansley, C.F. (1986), Estimation, interpolation and prediction for ARIMA models with missing data. *J. Amer. Statist. Assoc.* **81**, 751–761.

Lund, R., Hurd, H., Bloomfield, P. and Smith, R.L. (1995), Climatological time series with periodic correlation. *Journal of Climate* **8**, 2787–2809, 1995.

Lütkepohl, H. (1993), *Introduction to Multiple Time Series Analysis* (Second Edition). Springer Verlag, Berlin.

Meinhold, R.J. and Singpurwalla, N.D. (1983), Understanding the Kalman filter. *The American Statistician* **37**, 123–127.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1092.

Meyn, S.P. and Tweedie, R.L. (1993), *Markov Chains and Stochastic Stability.* Springer Verlag, New York.

Patterson, H.D. and Thompson, R. (1971), Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Percival, D.B. and Walden, A.T. (1993), *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge University Press, Cambridge.

Pole, A., West, M. and Harrison, P.J. (1994), *Applied Bayesian Forecasting and Time Series Analysis.* Chapman and Hall, New York.

Priestley, M.B. (1981), *Spectral Analysis and Time Series. Volume 1: Univariate Series. Volume 2: Multivariate Series, Prediction and Control.* Academic Press, London.

Ripley, B.D. (1987), *Stochastic Simulation.* John Wiley, New York.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility. In *Time Series Models: In econometrics, finance and other fields.* Edited by D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen. Chapman and Hall, London, pp. 1–67.

Shephard, N. and Pitt, M.K. (1997), Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.

Shephard, N. and Pitt, M.K. (1999), Analysis of time varying covariances: a factor stochstic volatility approach. To appear in *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press.

Tierney, L. (1994), Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.

West, M. and Harrison, P.J. (1997), *Bayesian Forecasting and Dynamic Models* (second edition). Springer Verlag, New York.