RAQUEL PRADO[*] and MIKE WEST[†]

TIME SERIES MODELLING, INFERENCE AND FORECASTING

[*] AMS, University of California, Santa Cruz
[†] ISDS, Duke University

Contents

Raquel Prado and Mike West

Part I

UNIVARIATE TIME SERIES

1 Notation, Definitions and Basic Inference

Problem Areas, Graphical Displays and Objectives

1.1 The expression *time series data*, or *time series*, usually refers to a set of observations collected sequentially in time. These observations could have been collected at equally-spaced time points. In this case we use the notation y_t with $(t = \ldots, -1, 0, 1, 2, \ldots)$, i.e., the set of observations is indexed by t, the time at which each observation was taken. If the observations were not taken at equally-spaced points then we use the notation y_{t_i}, with $i = 1, 2, \ldots$, and so, $(t_i - t_{i-1})$ is not necessarily equal to one.

A *time series process* is a stochastic process or a collection of random variables y_t indexed in time. Note that y_t will be used throughout the book to denote a random variable or an actual realisation of the time series process at time t. We use the notation $\{y_t, t \in \mathcal{T}\}$, or simply $\{y_t\}$, to refer to the time series process. If $\mathcal{T}$ is of the form $\{t_i, i \in N\}$, then the process is a discrete-time random process and if $\mathcal{T}$ is an interval in the real line, or a collection of intervals in the real line, then the process is a continuous-time random process. In this framework, a time series data set y_t, $(t = 1, \ldots, n)$, also denoted by $y_{1:n}$, is just a collection of n equally-spaced realisations of some time series process.

In many statistical models the assumption that the observations are realisations of independent random variables is key. In contrast, time series analysis is concerned with describing the dependence among the elements of a sequence of random variables.

At each time t, y_t can be a scalar quantity, such as the total amount of rainfall collected at a certain location in a given day t, or it can be a k-dimensional vector collecting k scalar quantities that were recorded simultaneously. For instance, if the total amount of rainfall and the average temperature at a given location are measured in day t, we have $k = 2$ scalar quantities $y_{1,t}$ and $y_{2,t}$ and so, at time t we have a 2-dimensional vector of observations $\mathbf{y}_t = (y_{1,t}, y_{2,t})'$. In general, for k scalar quantities recorded simultaneously at time t we have a realisation $\mathbf{y}_t$ of a vector process $\{\mathbf{y}_t, t \in \mathcal{T}\}$, with $\mathbf{y}_t = (y_{1,t}, \ldots, y_{k,t})'$.

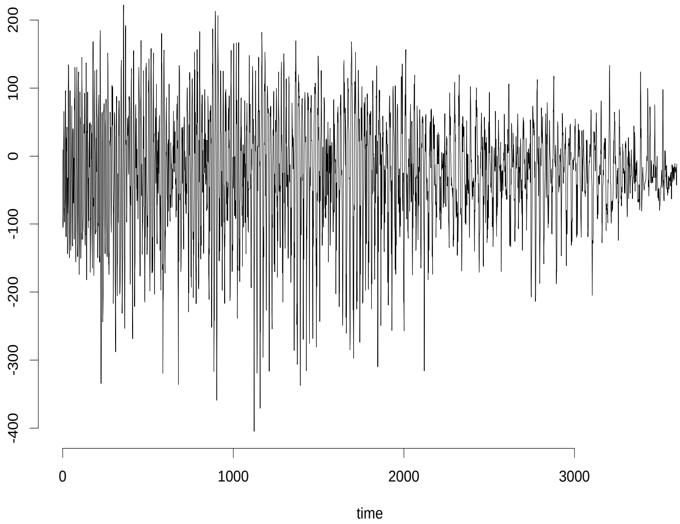

Fig. 1.1 EEG series (units in millivolts)

1.2 Figure 1.1 displays a portion of a human electroencephalogram or EEG, recorded on a patient's scalp under certain electroconvulsive therapy (ECT) conditions. ECT is an effective treatment for patients under major clinical depression (Krystal *et al.*, 1999). When ECT is applied to a patient, seizure activity appears and can be recorded via electroencephalograms. The series corresponds to one of 19 EEG channels recorded simultaneously at different locations over the scalp. The main objective in analysing this signal is the characterisation of the clinical efficacy of ECT in terms of particular features that can be inferred from the recorded EEG traces. The data are fluctuations in electrical potential taken at time intervals of roughly one fortieth of a second (more precisely 256 Hz). For a more detailed description of these data and a full statistical analysis see West *et al.* (1999); Krystal *et al.* (1999) and Prado *et al.* (2001). From the time series analysis viewpoint, the objective here is modelling the data in order to provide useful insight into the underlying processes driving the multiple series during a seizure episode. Studying the differences and commonalities among the 19 EEG channels is also key. Univariate time series models for each individual EEG series could be explored and used to investigate relationships across the 19 channels. Multivariate time series analyses — in which the observed series, $\mathbf{y}_t$, is a 19-dimensional vector whose elements are the observed voltage levels measured simultaneously at the 19 scalp locations at each time t — can also be considered.

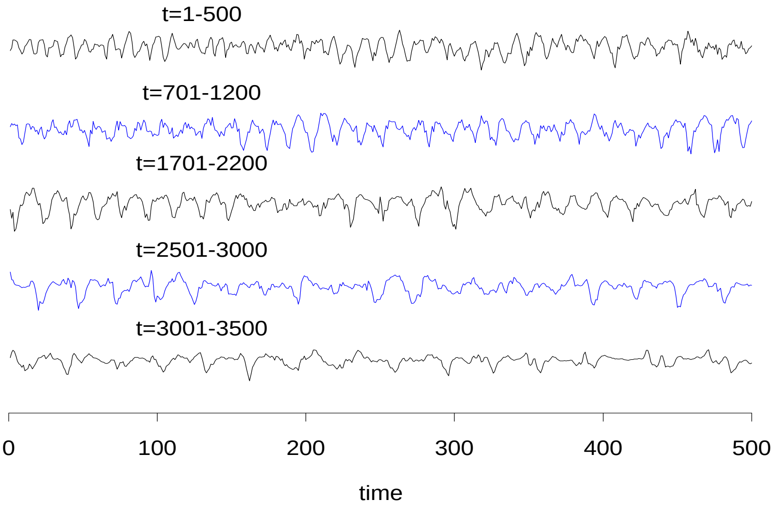

Fig. 1.2 Sections of the EEG trace displayed in Figure 1.1.

These EEG series display a quasi-periodic behaviour that changes dynamically in time, as shown in Figure 1.2, where different portions of the EEG trace shown in Figure 1.1 are displayed. In particular, it is clear that the relatively high frequency components that appear initially are slowly decreasing towards the end of the series. Any time series model used to describe these data should take into account their non-stationary and quasi-periodic structure. We discuss various modelling alternatives for these data in the subsequent chapters, including the class of time-varying autoregressions or TVAR models and other multi-channel models.

1.3 Figure 1.3 shows the annual per capita GDP (gross domestic product) time series for Austria, Canada, France, Germany, Greece, Italy, Sweden, UK and USA during 1950 and 1983. Interest lies in the study of "periodic" behaviour of such series in connection with understanding business cycles. Other goals of the analysis include forecasting turning points and comparing characteristics of the series across the national economies.

One of the main differences between any time series analysis of the GDP series and any time series analysis of the EEG series, regardless of the type of models used in such analyses, lies in the objectives. One of the goals in analysing the GDP data is forecasting future outcomes of the series for the several countries given the observed values. In the EEG study there is no interest in forecasting future values of the series given the observed traces, instead the objective is finding an appropriate model

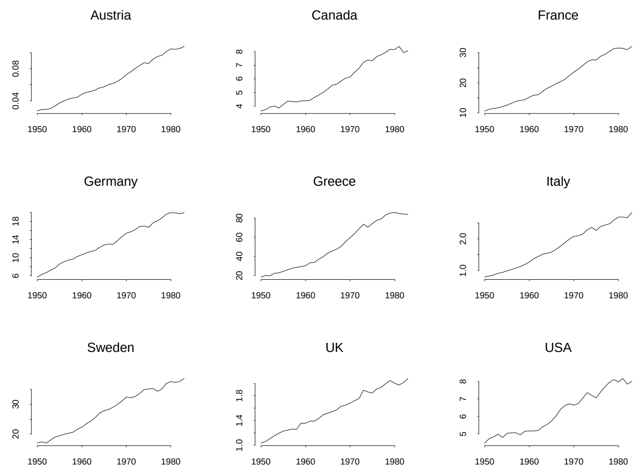

Fig. 1.3 International annual GDP time series

that determines the structure of the series and its latent components. Univariate and multivariate analyses of the GDP data can be considered.

1.4 Other objectives of time series analysis include monitoring a time series in order to detect possible “on-line” changes. This is important for control purposes in engineering, industrial and medical applications. For instance, consider a time series generated from the process $\{y_t\}$ with

$$y_t = \begin{cases} 0.9y_{t-1} & +\epsilon_t^{(1)}, \quad y_{t-1} > 1.5 \quad (M_1) \\ -0.3y_{t-1} & +\epsilon_t^{(2)}, \quad y_{t-1} \leq -1.5 \quad (M_2), \end{cases} \tag{1.1}$$

where $\epsilon_t^{(1)} \sim N(0, v_1)$, $\epsilon_t^{(2)} \sim N(0, v_2)$ and $v_1 = v_2 = 1$. Figure 1.4 (a) shows a time series plot of 1,500 observations simulated according to (1.1). Figure 1.4 (b) displays the values of an indicator variable, δ_t, such that $\delta_t = 1$ if y_t was generated from M_1 and $\delta_t = 2$ if y_t was generated from M_2. The model (1.1) belongs to the class of so called threshold autoregressive (TAR) models, initially developed by H. Tong (Tong, 1983; Tong, 1990). In particular, (1.1) is a TAR model with two regimes, and so, it can be written in the following, more general, form

$$y_t = \begin{cases} \phi^{(1)}y_{t-1} & +\epsilon_t^{(1)}, \quad \theta + y_{t-d} > 0 \quad (M_1) \\ \phi^{(2)}y_{t-1} & +\epsilon_t^{(2)}, \quad \theta + y_{t-d} \leq 0 \quad (M_2), \end{cases} \tag{1.2}$$

with $\epsilon_t^{(1)} \sim N(0, v_1)$ and $\epsilon_t^{(2)} \sim N(0, v_2)$. These are non-linear models and the interest lies in making inference about d, θ and the parameters $\phi^{(1)}, \phi^{(2)}, v_1$ and v_2.

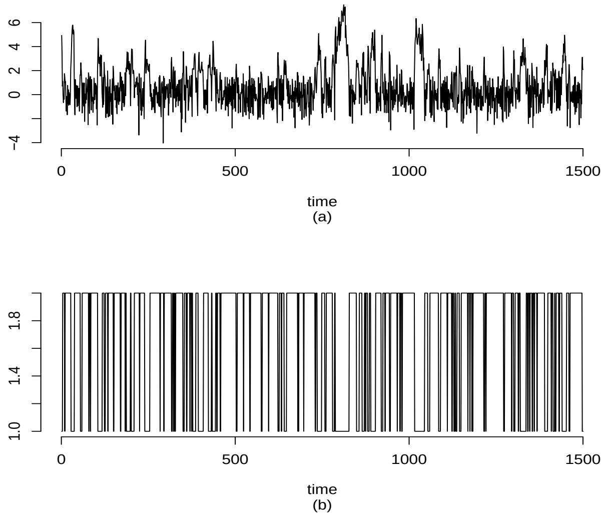

Fig. 1.4 (a): Simulated time series y_t; (b) Indicator variable δ_t such that $\delta_t = 1$ if y_t was a sampled from M_1 and $\delta_t = 2$ if y_t was sampled from M_2.

Model (1.2) serves the purpose of illustrating, at least for a very simple case, a situation that arises in many engineering applications, particularly in the area of control theory. From a control theory viewpoint we can think of model (1.2) as a bimodal process in which two scenarios of operation are handled by two control modes (M_1 and M_2). In each mode the evolution is governed by a stochastic process. Autoregressions of order one, or AR(1) models (a formal definition of this type of processes is given later in this chapter), were chosen in this example, but more sophisticated structures can be considered. The transitions between the modes occur when the series crosses a specific threshold and so, we can talk about an internally-triggered mode switch. In an externally-triggered mode switch the moves are defined by external variables.

In terms of the goals of a time series analysis we can consider two possible scenarios. In many control settings where the transitions between modes occur in response to controller's actions, the current state is always known. In this setting we can split the learning process in two: learning the stochastic models that control each mode conditional on the fact that we know in which mode we are, i.e., inferring $\phi^{(1)}, \phi^{(2)}, v_1$ and v_2, and learning the transition rule, that is, making inferences about d and θ assuming we know the values $\delta_{1:n}$. In other control settings, where the mode transitions do not occur in response to the controller's actions, it is necessary to simultaneously infer the parameters associated to the stochastic models that describe each mode and the transition rule. In this case we want to estimate $\phi^{(1)}, \phi^{(2)}, v_1, v_2, \theta$

and d conditioning only on the observed data $y_{1:n}$. Depending on the application it may also be necessary to do the learning from the time series sequentially in time (see Chapter 5).

1.5 Finally, we may use time series techniques to model serial dependences between parameters of a given model with additional structure. For example, we could have a linear regression model of the form $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$, for which ϵ_t does not exhibit the usual independent structure $\epsilon_t \sim N(0, v)$ for all t but instead, the probability distribution of ϵ_t depends on $\epsilon_{t-1}, \ldots, \epsilon_{t-k}$.

Stochastic Processes and Stationarity

Many time series models are based on the assumption of stationarity. Intuitively, a stationary time series process is a process whose behaviour does not depend on when we start to observe it. In other words, different sections of the series will look roughly the same at intervals of the same length. Here we provide two widely used definitions of stationarity.

1.6 A time series process $\{y_t, t \in \mathcal{T}\}$ is *completely* or *strongly stationary* if, for any sequence of times $t_1, t_2, \ldots, t_n$, and any lag k, the probability distribution of $(y_{t_1}, \ldots, y_{t_n})'$ is identical to the probability distribution of $(y_{t_1+k}, \ldots, y_{t_n+k})'$.

1.7 In practice it is very difficult to verify that a process is strongly stationary and so, the notion of *weak* or *second order stationarity* arises. A process is said to be weakly stationary, or second order stationary if, for any sequence of times $t_1, \ldots, t_n$, and any lag k, all the first and second joint moments of $(y_{t_1}, \ldots, y_{t_n})'$ exist and are equal to the first and second joint moments of $(y_{t_1+k}, \ldots, y_{t_n+k})'$. If $\{y_t\}$ is second order stationary we have that

$$E(y_t) = \mu, \quad Var(y_t) = v, \quad Cov(y_t, y_s) = \gamma(s-t), \tag{1.3}$$

where μ, v are constant, independent of t and $\gamma(s-t)$ is also independent of t and s, since it only depends on the length of the interval between time points. It is also possible to define stationarity up to order m in terms of the m joint moments, see for example Priestley (1994).

If the first two moments exist, complete stationarity implies second order stationarity, but the converse is not necessarily true. If $\{y_t\}$ is a Gaussian process, i.e., if for any sequence of time points $t_1, \ldots, t_n$ the vector $(y_{t_1}, \ldots, y_{t_n})'$ follows a multivariate normal distribution, strong and weak stationarity are equivalent.

Exploratory Analysis: Auto-Correlation and Cross-Correlation

The first step in any statistical analysis usually consists on performing a descriptive study of the data in order to summarise their main characteristic features. One of the most widely used descriptive techniques in time series data analysis is that of exploring the correlation patterns displayed by a series, or a couple of series, at different time points. This is done by plotting the sample auto-correlation and cross-correlation values, which are estimates of the auto-correlation and cross-correlation functions.

1.8 We begin by defining the concepts of auto-covariance, auto-correlation and cross-correlation functions. We then show how to estimate these functions from data. Let $\{y_t, t \in \mathcal{T}\}$ be a time series process. The auto-covariance function of $\{y_t\}$ is defined as follows

$$\gamma(s,t) = Cov\{y_t, y_s\} = E\{(y_t - \mu_t)(y_s - \mu_s)\}, \tag{1.4}$$

for all s, t, with $\mu_t = E(y_t)$. For stationary processes $\mu_t = \mu$ for all t and the covariance function depends on $|s - t|$ only. In this case we can write the auto-covariance as a function of a particular time lag k

$$\gamma(k) = Cov\{y_t, y_{t-k}\}. \tag{1.5}$$

The auto-correlation function (ACF) is then given by

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{1.6}$$

For stationary processes, the ACF can be written in terms of a lag k

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}. \tag{1.7}$$

The auto-correlation function inherits the properties of any correlation function. In this particular case the ACF is a measure of the linear dependence between a value of the time series process at time t and past or future values of such process. $\rho(k)$ always takes values in the interval $[-1, 1]$. In addition, $\rho(k) = \rho(-k)$ and if y_t and y_{t-k} are independent $\rho(k) = 0$.

It is also possible to define the cross-covariance and cross-correlation functions of two univariate time series. If $\{y_t\}$ and $\{z_t\}$ are two time series processes, the cross-covariance is defined as

$$\gamma_{y,z}(s,t) = E\{(y_t - \mu_{y_t})(z_s - \mu_{z_s})\}, \tag{1.8}$$

for all s, t and the cross-correlation is then given by

$$\rho_{y,z}(s,t) = \frac{\gamma_{y,z}(s,t)}{\sqrt{\gamma_y(s,s)\gamma_z(t,t)}}. \tag{1.9}$$

If both processes are stationary we can again write the cross-covariance and cross-correlation functions in terms of a lag value k. This is

$$\gamma_{y,z}(k) = E\{(y_t - \mu_y)(z_{t-k} - \mu_z)\}, \tag{1.10}$$

and

$$\rho_{y,z}(k) = \frac{\gamma_{y,z}(k)}{\sqrt{\gamma_y(0)\gamma_z(0)}}. \tag{1.11}$$

Example 1.8.1 *White Noise.*

Consider a time series process such that $y_t \sim N(0, v)$ for all t. In this case $\gamma(0) = v$, $\gamma(k) = 0$ for all $k \neq 0$, $\rho(0) = 1$ and $\rho(k) = 0$ for all $k \neq 0$.

Example 1.8.2 *First order autoregression or AR(1).*

In Chapter 2 we formally define and study the properties of general autoregressions of order p, or AR(p) processes. Here, we illustrate some properties of the simplest AR process, the AR(1). Consider a time series process such that $y_t = \phi y_{t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, v)$ for all t. It is possible to show that $\gamma(k) = \phi^{|k|}\gamma(0)$ for $k = 0, \pm 1, \pm 2, \ldots$, with $\gamma(0) = \frac{v}{(1-\phi^2)}$, and $\rho(k) = \phi^{|k|}$ for $k = 0, \pm 1, \pm 2, \ldots$ Figure 1.5 displays the auto-correlation functions for AR(1) processes with parameters $\phi = 0.9, \phi = -0.9$ and $\phi = 0.3$, for lag values $0, 1, \ldots, 50$. For negative values of ϕ the ACF has an oscillatory behaviour. In addition, the rate of decay of the ACF is a function of ϕ. The closer $|\phi|$ gets to the unity the lower the rate of decay is (e.g., compare the ACFs for $\phi = 0.9$ and $\phi = 0.3$). It is also obvious from the form of the ACF that this is an explosive function when $|\phi| > 1$ and is equal to unity for all k when $\phi = 1$. This is related to the characterisation of stationarity in AR(1) processes. An AR(1) process is stationary if and only if $|\phi| < 1$. The stationary condition can also be written as a function of the characteristic root of the process. An AR(1) is stationary if and only if the root of the characteristic polynomial u, such that $\Phi(u) = 0$, with $\Phi(u) = 1 - \phi u$, lies outside the unit circle, and this happens if and only if $|\phi| < 1$.

1.9 We now show how to estimate the auto-covariance, auto-correlation, cross-covariance and cross-correlation functions from data. Assume we have data $y_{1:n}$. The usual estimate of the auto-covariance function is the sample auto-covariance, which, for $k > 0$, is given by

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y}), \tag{1.12}$$

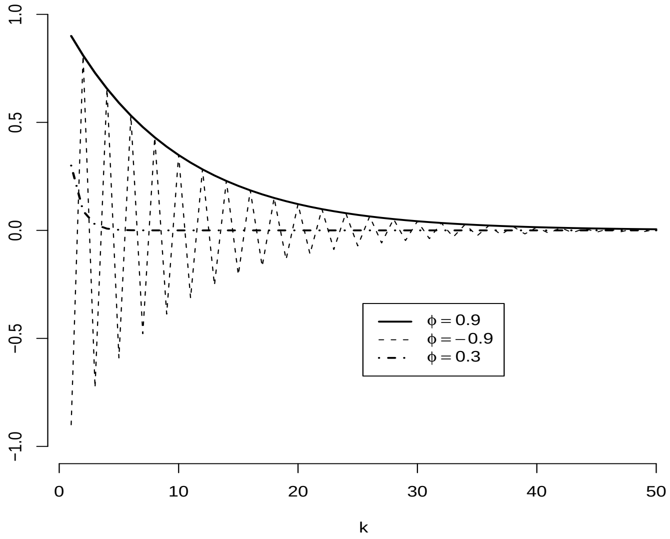

Fig. 1.5 Auto-correlation functions for AR processes with parameters 0.9, -0.9 and 0.3

where $\bar{y} = \sum_{t=1}^{n} y_t/n$ is the sample mean. We can then obtain the estimates of the auto-correlation function as $\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$, for $k = 0, 1, \dots$

Similarly, estimates of the cross-covariance and cross-correlation functions can be obtained. The sample cross-covariance is given by

$$\hat{\gamma}_{y,z}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(z_{t+k} - \bar{z}), \tag{1.13}$$

and so, the sample cross-correlation is obtained as $\hat{\rho}_{y,z}(k) = \hat{\gamma}_{y,z}(k)/\sqrt{\hat{\gamma}_y(0)\hat{\gamma}_z(0)}$.

Figure 1.6 displays the sample auto-correlation functions of simulated AR(1) processes with parameters $\phi = 0.9$, $\phi = -0.9$ and $\phi = 0.3$, respectively. The sample ACFs were computed based on a sample of $n = 200$ data points. For $\phi = 0.9$ and $\phi = 0.3$ the corresponding sample ACFs decay as a function of the lag. The oscillatory form of the ACF for the process with $\phi = -0.9$ is captured by the corresponding sample ACF.

The estimates given in (1.12) and (1.13) are not unbiased estimates of the auto-covariance and cross-covariance functions. For results related to the sample distribution of the sample auto-correlation and sample cross-correlation functions see for example Shumway and Stoffer (2000).

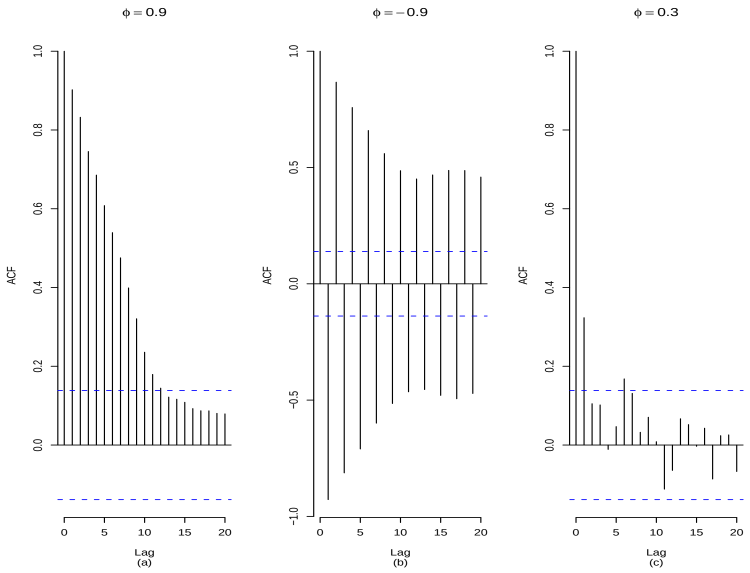

Fig. 1.6 Sample auto-correlation functions for AR processes with parameters 0.9, -0.9 and 0.3 (graphs (a), (b) and (c), respectively)

Exploratory Analysis: Smoothing and Differencing

As mentioned before, many time series models are built under the stationarity assumption. Several descriptive techniques have been developed to study the stationary properties of a time series so that an appropriate model can then be applied to the data. For instance, looking at the sample auto-correlation function may be helpful in identifying some features of the data. However, in many practical scenarios the data are realisations from one or several non-stationary processes. In this case, methods that aim to eliminate the non-stationary components are often used. The idea is to separate the non-stationary components from the stationary ones so that the later can be carefully studied via traditional time series models such as, for example, the ARMA (autoregressive-moving-average) models that will be discussed in subsequent chapters.

In this Section we enumerate some commonly used methods for extracting the non-stationary components of a time series. We do not attempt to provide a comprehensive list of methods, since this would be a nearly impossible task beyond the scope of this book. Instead, we just list and summarise a few of them. We view these techniques as purely descriptive. We believe that if the data display non-stationary components, such components should be explicitly included in any proposed model.

Several descriptive time series methods are based on the notion of smoothing the data, this is, decomposing the series as a sum of two components: a so called "smooth" component, plus another component that includes all the features of the data that are left unexplained by the smooth component. This is similar to the “signal plus noise” concept used in signal processing. The main difficulty with this approach lies in deciding which features of the data are part of the signal or the smooth component and which ones are part of the noise.

1.10 One way to do smoothing is via moving averages (see Kendall *et al.*, 1983; Kendall and Ord, 1990; Chatfield, 1996 and Diggle, 1990 for detailed discussions and examples). If we have data $y_{1:n}$, we can smooth them by applying an operation of the form

$$z_t = \sum_{j=-q}^{p} a_j y_{t+j}, \quad t = q+1, \ldots, n-p, \tag{1.14}$$

where the a_j's are weights such that $\sum_{j=-q}^{p} a_j = 1$. It is generally assumed that $p = q$, $a_j \geq 0$ for all j and $a_j = a_{-j}$. The order of the moving average in this case is $2p + 1$. The first question that arises when applying a moving average to a series is how to choose p and the weights. The simplest alternative is choosing a low value of p and equal weights. The higher the value of p, the smoother z_t is going to be. Other alternatives include successively applying a simple moving average with equal weights or choosing the weights in such a way that a particular feature of the data is highlighted. So, for example, if a given time series recorded monthly displays a trend plus a yearly cycle, choosing a moving average with $p = 6$, $a_6 = a_{-6} = 1/24$ and $a_j = 1/12$ for $j = 0, \pm 1, \ldots, \pm 5$ would diminish the impact of the periodic component and therefore, emphasising the trend (see Diggle, 1990 for an example).

Figure 1.7 (a) shows monthly values of the Souther Oscillation Index (SOI) during 1950-1995. This series consists of 540 observations of the SOI, computed as the difference of the departure from the long-term monthly mean sea level pressures at Tahiti in the South Pacific and Darwin in Northern Australia. The index is one measure of the so called "El Niño-Southern Oscillation" – an event of critical importance and interest in climatological studies in recent decades. The fact that most of the observations in the last part of the series take negative values is related to a recent warming in the tropical Pacific. A key question of interest is to determine just how unusual this event is, and if it can reasonably be explained by standard "stationary" time series models, or requires models that include drifts/trends that may be related to global climatic change. Figures 1.7 (b) and (c) show two smoothed series obtained via moving averages of orders 3 and 9, respectively, with equal weights. As explained before, we can see that the higher the order of the moving average the smoother is the resulting series.

1.11 Other ways to smooth a time series include fitting a linear regression to remove a trend or, more generally, fitting a polynomial regression; fitting a harmonic regression to remove periodic components and performing kernel smoothing or spline smoothing.

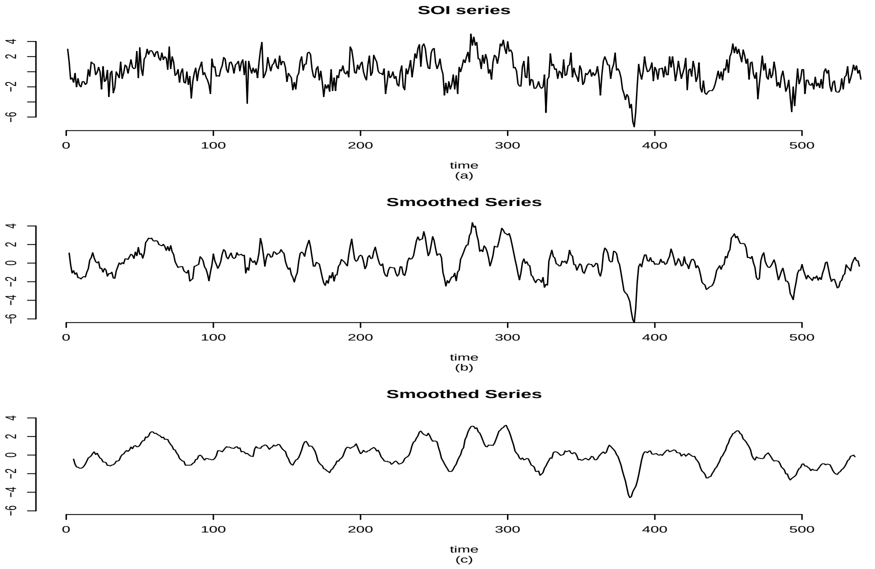

Fig. 1.7 (a): Southern oscillation index (SOI) time series; (b): Smoothed series obtained using a moving average of order 3 with equal weights; (c): Smoothed series obtained using a moving average of order 9 with equal weights

Smoothing by polynomial regression consists on fitting a polynomial to the series. In other words, we want to estimate the parameters of the model

$$y_t = \beta_0 + \beta_1 t + \ldots + \beta_p t^p + \epsilon_t,$$

where ϵ_t is usually assumed as a sequence of zero mean, independent Gaussian random variables. Similarly, fitting harmonic regressions provides a way to remove cycles from a time series. So, if we want to remove periodic components with frequencies $w_1, \ldots, w_p$, we need to estimate $a_1, b_1, \ldots, a_p, b_p$ in the model

$$\begin{aligned} y_t &= a_1 \cos(2\pi w_1 t) + b_1 \sin(2\pi w_1 t) + \ldots + \\ & \quad a_p \cos(2\pi w_p t) + b_p \cos(2\pi w_p t) + \epsilon_t. \end{aligned}$$

In both cases the smoothed series would then be obtained as $z_t = y_t - \hat{y}_t$, with $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \ldots + \hat{\beta}_p t^p$, and $\hat{y}_t = \hat{a}_1 \cos(2\pi w_1 t) + \hat{b}_1 \sin(2\pi w_1 t) + \ldots + \hat{a}_p \cos(2\pi w_p t) + \hat{b}_p \cos(2\pi w_p t)$, respectively, where $\hat{\beta}_i$, $\hat{a}_i$ and $\hat{b}_i$ are point estimates of the parameters. Usually $\hat{\beta}_i$ and $\hat{a}_i, \hat{b}_i$ are obtained by least squares estimation.

In kernel smoothing a smoothed version z_t of the original series y_t is obtained as follows

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

where $K(\cdot)$ is a kernel function, such as a normal kernel. The parameter b is a bandwidth. The larger the value of b, the smoother z_t is.

Cubic splines and smoothing splines are also commonly used smoothing techniques. See Shumway and Stoffer (2000) for details and illustrations on kernel and spline smoothing.

1.12 Another way to smooth a time series is by taking its differences. Differencing provides a way to remove trends. The first difference of a series y_t is defined in terms of an operator D that produces the transformation $Dy_t = y_t - y_{t-1}$. Higher order differences are defined by successively applying the operator D. Differences can also be defined in terms of the back shift operator B, with $By_t = y_{t-1}$ and so, $Dy_t = (1 - B)y_t$. Higher order differences can be written as $D^d y_t = (1 - B)^d y_t$.

1.13 In connection with the methods presented in this Section, it is worth mentioning that wavelet decompositions have been widely used in recent years for smoothing time series. Vidakovic (1999) presents a statistical approach to modelling by wavelets. Wavelets are bases functions that are used to represent other functions. They are the analogous to the sines and cosines in the Fourier transformation. One of the advantages of using wavelets basis, as opposed to Fourier representations, is that they are localised in frequency and time, and so, they are suitable for dealing with non-stationary signals that display jumps and other abrupt changes.

A Primer on Likelihood and Bayesian Inference

Assume that we have collected n observations, $y_{1:n}$, of a scalar quantity over time. Suppose that for each observation y_t we have a probability distribution that can be written as a function of some parameter, or collection of parameters, namely $\boldsymbol{\theta}$, in such a way that the dependence of y_t on $\boldsymbol{\theta}$ is described in terms of a probability density function $p(y_t|\boldsymbol{\theta})$. If we think of $p(y_t|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, rather than a function of y_t, we refer to it as the likelihood function. Using Bayes' theorem it is possible to obtain the posterior density function of $\boldsymbol{\theta}$ given the observation y_t, $p(\boldsymbol{\theta}|y_t)$, as the product of the likelihood and the prior density $p(\boldsymbol{\theta})$, i.e.,

$$p(\boldsymbol{\theta}|y_t) = \frac{p(\boldsymbol{\theta})p(y_t|\boldsymbol{\theta})}{p(y_t)}, \tag{1.15}$$

with $p(y_t) = \int p(\boldsymbol{\theta})p(y_t|\boldsymbol{\theta})d\boldsymbol{\theta}$. $p(y_t)$ defines the so called predictive density function. The prior distribution offers a way to incorporate our prior beliefs about $\boldsymbol{\theta}$ and Bayes' theorem provides the way to update such beliefs after observing the data.

Bayes' theorem can also be used in a sequential way. So, before collecting any data, the prior beliefs about $\boldsymbol{\theta}$ are expressed in a probabilistic form via $p(\boldsymbol{\theta})$. Assume that we then collect our first observation at time $t = 1, y_1$, and we obtain $p(\boldsymbol{\theta}|y_1)$ using Bayes' theorem. Once y_2 is observed we can obtain $p(\boldsymbol{\theta}|y_{1:2})$ using Bayes' theorem as $p(\boldsymbol{\theta}|y_{1:2}) \propto p(\boldsymbol{\theta})p(y_{1:2}|\boldsymbol{\theta})$. Now, if y_1 and y_2 are conditionally independent on $\boldsymbol{\theta}$ we can write $p(\boldsymbol{\theta}|y_{1:2}) \propto p(\boldsymbol{\theta}|y_1)p(y_2|\boldsymbol{\theta})$, i.e., the posterior of $\boldsymbol{\theta}$ given y_1 becomes a prior distribution before observing y_2. Similarly, $p(\boldsymbol{\theta}|y_{1:n})$ can be obtained in a sequential way, if all the observations are independent. However, in time series analysis the observations are not independent. For example, a common assumption is that each observation at time t depends only on $\boldsymbol{\theta}$ and the observation taken at time $t - 1$. In this case we have

$$p(\boldsymbol{\theta}|y_{1:n}) \propto p(\boldsymbol{\theta})p(y_1|\boldsymbol{\theta}) \prod_{t=2}^{n} p(y_t|y_{t-1}, \boldsymbol{\theta}). \tag{1.16}$$

General models in which y_t depends on an arbitrary number of past observations will be studied in subsequent chapters. We now consider an example in which the posterior distribution has the form (1.16).

Example 1.13.1 *The AR*(1) *model.*

We consider again the AR(1) process. The model parameters in this case are given by $\boldsymbol{\theta} = (\phi, v)'$. Now, for each time $t > 1$, the conditional likelihood is $p(y_t|y_{t-1}, \boldsymbol{\theta}) = N(y_t|\phi y_{t-1}, v)$. In addition, it can be shown that $y_1 \sim N(0, v/(1 - \phi^2))$ (see Problem (1) in Chapter 2) and so, the likelihood in this case is

$$p(y_{1:n}|\boldsymbol{\theta}) = \frac{(1-\phi^2)^{1/2}}{(2\pi v)^{n/2}} \exp\left\{-\frac{Q^*(\phi)}{2v}\right\}, \tag{1.17}$$

with

$$Q^*(\phi) = y_1^2(1-\phi^2) + \sum_{t=2}^{n}(y_t - \phi y_{t-1})^2. \tag{1.18}$$

The posterior density is obtained via Bayes' rule and so

$$p(\boldsymbol{\theta}|y_{1:n}) \quad \propto \quad p(\boldsymbol{\theta})\frac{(1-\phi^2)^{1/2}}{(2\pi v)^{n/2}} \exp\left\{\frac{-Q^*(\phi)}{2v}\right\}.$$

We can also use the conditional likelihood $p(y_{2:n}|\boldsymbol{\theta}, y_1)$ as an approximation to the likelihood (see Box *et al.*, 1994 A7.4 for a justification), which leads to the following posterior density

$$p(\boldsymbol{\theta}|y_{1:n}) \quad \propto \quad p(\boldsymbol{\theta}) v^{-(n-1)/2} \text{exp}\left\{ \frac{-Q(\phi)}{2v} \right\}, \tag{1.19}$$

with $Q(\phi) = \sum_{t=2}^{n} (y_t - \phi y_{t-1})^2$. Several choices of $p(\boldsymbol{\theta})$ can be considered and will be discussed later. In particular, it is common to assume a prior structure such that $p(\boldsymbol{\theta}) = p(v)p(\phi|v)$, or $p(\boldsymbol{\theta}) = p(v)p(\phi)$.

Another important class of time series models is one in which parameters are indexed in time. In this case each observation is related to a parameter, or a set of parameters, say $\boldsymbol{\theta}_t$, that evolves over time. The so called class of Dynamic Linear Models (DLMs) considered in Chapter 4 deals with models of this type. In this framework it is necessary to define a process that describes the evolution of $\boldsymbol{\theta}_t$ over time. As an example, consider the time-varying AR model of order one, or TVAR(1), given by

$$\begin{aligned} y_t &= \phi_t y_{t-1} + \epsilon_t, \\ \phi_t &= \phi_{t-1} + \nu_t, \end{aligned}$$

with ϵ_t and ν_t independent in time and mutually independent and with $\epsilon_t \sim N(0, v)$ and $\nu_t \sim N(0, w)$. Some distributions of interest are the posteriors at time t, $p(\phi_t|y_{1:t})$ and $p(v|y_{1:t})$, the filtering or smoothing distributions $p(\phi_t|y_{1:n})$, and the m-steps ahead forecast distribution $p(y_{t+m}|y_{1:t})$. Details on how to find these distributions for rather general DLMs are given in Chapter 4.

1.14 ML, MAP and LS estimation. It is possible to obtain point estimates of the model parameters by maximising the likelihood function or the full posterior distribution. A variety of methods and algorithms have been developed to achieve this goal. We briefly discuss some of these methods. In addition, we illustrate how these methods work in the simple AR(1) case.

A point estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, can be obtained by maximising the likelihood function $p(y_{1:n}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In this case we use the notation $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{ML}}$. Similarly, if instead of maximising the likelihood function we maximise the posterior distribution $p(\boldsymbol{\theta}|y_{1:n})$, we obtain the maximum a posteriori estimate for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{MAP}}$.

Usually, the likelihood function and the posterior distribution are complicated non-linear functions of $\boldsymbol{\theta}$ and so, it is necessary to use methods such as the Newton-Raphson algorithm or the scoring method to obtain the maximum likelihood estimator (MLE) or the maximum a posteriori (MAP) estimator. In general, the Newton-Raphson algorithm can be summarised as follows. Let $g(\boldsymbol{\theta})$ be the function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ that we want to maximise and $\hat{\boldsymbol{\theta}}$ be the maximum. At iteration m

of the Newton-Raphson algorithm we obtain $\boldsymbol{\theta}^{(m)}$, an approximation to $\hat{\boldsymbol{\theta}}$, as follows

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - \left[g''(\boldsymbol{\theta}^{(m-1)})\right]^{-1} \times \left[g'(\boldsymbol{\theta}^{(m-1)})\right], \tag{1.20}$$

where $g'(\boldsymbol{\theta})$ and $g''(\boldsymbol{\theta})$ denote the first and second order partial derivatives of the function g, i.e. $g'(\boldsymbol{\theta})$ is a k-dimensional vector $g'(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k}\right)'$, and $g''(\boldsymbol{\theta})$ is a $k \times k$ matrix of second order partial derivatives whose ij-th element is given by $\left[\frac{\partial g^2(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]$, for $i, j = 1, \ldots, k$. Under certain conditions this algorithm produces a sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots,$ that will converge to $\hat{\boldsymbol{\theta}}$. In particular, it is important to begin with a good starting value $\boldsymbol{\theta}^{(0)}$, since the algorithm does not necessarily converge for values in regions where $-g''(\cdot)$ is not positive definite. An alternative method is the scoring method, which involves replacing $g''(\boldsymbol{\theta})$ in (1.20) by the matrix of expected values $E(g''(\boldsymbol{\theta}))$.

In many practical scenarios, specially when dealing with models that have very many parameters, it is not useful to summarise the inferences in terms of the joint posterior mode. In such cases it is often interesting and appropriate to summarise posterior inference in terms of the marginal posterior modes, this is, the posterior modes for subsets of model parameters. Let us say that we can partition our model parameters in two sets, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and assume we are interested in $p(\boldsymbol{\theta}_2|y_{1:n})$. The EM (Expectation-Maximisation) algorithm proposed in Dempster *et al.* (1977) is useful when dealing with models for which $p(\boldsymbol{\theta}_2|y_{1:n})$ is hard to maximise directly but it is relatively easy to work with $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y_{1:n})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, y_{1:n})$. The EM algorithm can be described as follows

1. Start with some initial value $\boldsymbol{\theta}_2^{(0)}$.
2. For i=1,2,...
   - Compute $E^{(i-1)}[\log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_{1:n})]$ given by the expression

     $$\int \log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_{1:n}) p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(i-1)}, y_{1:n}) d\boldsymbol{\theta}_1. \tag{1.21}$$

     This is the E-step.
   - Set $\boldsymbol{\theta}_2^{(i)}$ to the value that maximises (1.21). This is the M-step.

At each iteration of the EM algorithm $p(\boldsymbol{\theta}_2|y_{1:n})$ should increase and so, the algorithm should converge to the mode. Some extensions of the EM algorithm include the ECM (expectation-conditional-maximisation) algorithm, ECME (variant of the ECM in which either the log-posterior density or the expected log-posterior density is maximised) and SEM (sumplemented EM) algorithms (see Gelman *et al.*, 2004 and references therein) and stochastic versions of the EM algorithm such as the MCEM (Wei and Tanner, 1990).

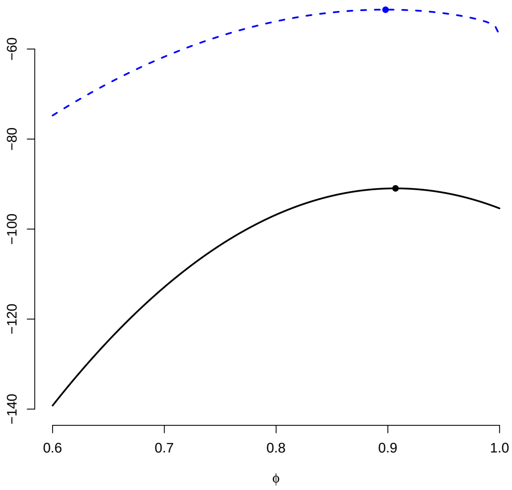

Fig. 1.8 Conditional and unconditional likelihoods (solid and dotted lines respectively) for 100 simulated observations

Example 1.14.1 *ML, MAP and LS estimators for the AR(1) model.*
Consider an AR(1) such that $y_t = \phi y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, 1)$. In this case $v = 1$ and $\theta = \phi$. The conditional MLE is found by maximising the function $\exp\{-\frac{Q(\phi)}{2}\}$ or equivalently, by minimising $Q(\phi)$. Therefore, we obtain $\hat{\phi} = \phi_{\text{ML}} = \sum_{t=2}^{n} y_t y_{t-1} / \sum_{t=2}^{n} y_{t-1}^2$. Similarly, the MLE for the unconditional likelihood function is obtained by maximising $p(y_{1:n}|\phi)$, or equivalently, by minimising the expression

$$-0.5[\log(1 - \phi^2) + Q^*(\phi)].$$

Thus, the Newton-Raphson or scoring methods can be used to find $\hat{\phi}$. As an illustration, the conditional and unconditional ML estimators were found for 100 samples from an AR(1) with $\phi = 0.9$. Figure 1.8 shows a graph with the conditional and unconditional log-likelihood functions (solid and dotted lines respectively). The points correspond to the maximum likelihood estimators with $\hat{\phi} = 0.9069$ and $\hat{\phi} = 0.8979$ being the MLEs for the conditional and unconditional likelihoods, respectively. For the unconditional case, a Newton-Raphson algorithm was used to find the maximum. The algorithm converged after 5 iterations with a starting value of 0.1.

Figure 1.9 shows the form of the log-posterior distribution of ϕ under Gaussian priors of the form $\phi \sim N(\mu, c)$, for $\mu = 0$, $c = 1.0$ (left panel) and $c = 0.01$ (right panel). Note that this prior does not impose any restriction on ϕ and so, it gives non-negative probability to values of ϕ that lie in the non-stationary region. It is possible to choose priors on ϕ whose support is the stationary region. This will be considered in Chapter 2. Figure 1.9 illustrates the effect of the prior on the MAP

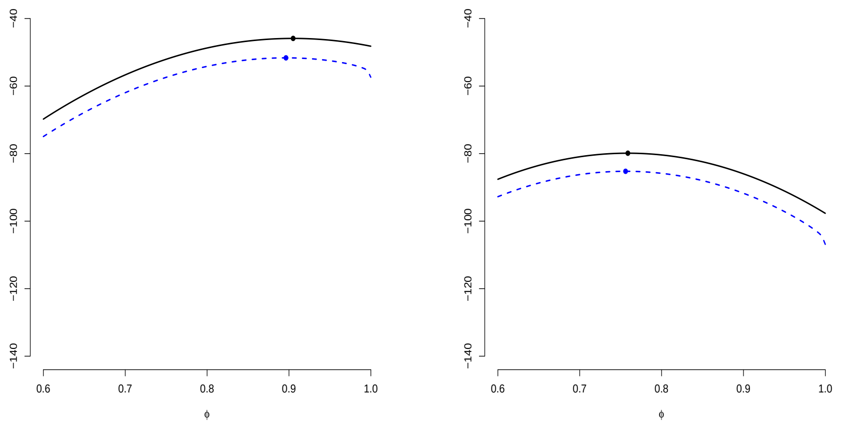

Fig. 1.9 Conditional and unconditional posteriors (solid and dotted lines respectively) with priors of the form $\phi \sim N(0, c)$, for $c = 1$ (left panel) and $c = 0.01$ (right panel).

estimators. For a prior $\phi \sim N(0, 1)$, the MAP estimators are $\hat{\phi}_{\text{MAP}} = 0.9051$ and $\hat{\phi}_{\text{MAP}} = 0.8963$ for the conditional and unconditional likelihoods, respectively. When a smaller value of c is considered, or in other words, when the prior distribution is more concentrated around zero, then the MAP estimates shift towards the prior mean. For a prior $\phi \sim N(0, 0.01)$, the MAP estimators are $\hat{\phi}_{\text{MAP}} = 0.7588$ and $\hat{\phi}_{\text{MAP}} = 0.7550$ for the conditional and unconditional likelihoods, respectively. Again, the MAP estimators for the unconditional likelihoods were found using a Newton-Raphson algorithm.

It would have also been possible to obtain the least squares estimators for the conditional and unconditional likelihoods. For the conditional case, the least squares estimator, or LSE, is obtained by minimising the conditional sum of squares $Q(\phi)$, and so, in this case $\phi_{\text{MLE}} = \phi_{\text{LSE}}$. In the unconditional case the LSE is found by minimising the unconditional sum of squares $Q^*(\phi)$ and so, the LSE and the MLE do not coincide.

1.15 Traditional Least Squares. Likelihood and Bayesian methods for fitting linear autoregressions rely on very standard methods of linear regression analysis therefore, some review of the central ideas and results in regression is in order and given here. This introduces notation and terminology that will be used throughout the book.

A linear model with a univariate response variable and $p > 0$ regressor variables (otherwise predictors or covariates) has the form

$$y_i = \mathbf{f}_i' \boldsymbol{\beta} + \epsilon_i$$

for $i = 1, 2, \ldots,$ where y_i is the i^{th} observation on the response variable, and has corresponding values of the regressors in the design vector $\mathbf{f}_i = (f_{i1}, \ldots, f_{ip})'$. The design vectors are assumed known and fixed prior to observing corresponding responses. The error terms ϵ_i are assumed independent and normal, distributed as $N(\epsilon_i|0, v)$ with some variance v. The regression parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is to be estimated, along with the error variance. The model for an observed set of n responses $\mathbf{y} = (y_1, \ldots, y_n)'$ is

$$\mathbf{y} = \mathbf{F}'\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{F}$ is the known $p \times n$ design matrix with i^{th} column $\mathbf{f}_i$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|0, v\mathbf{I}_n)$, with $\mathbf{I}_n$ the $n \times n$ identity matrix. This defines the sampling distribution

$$p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v) = \prod_{i=1}^{n} N(y_i|\mathbf{f}_i'\boldsymbol{\beta}, v) = (2\pi v)^{-n/2}\text{exp}(\,-Q(\mathbf{y}, \boldsymbol{\beta})/2v),$$

where $Q(\mathbf{y}, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{F}'\boldsymbol{\beta})'(\mathbf{y} - \mathbf{F}'\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{f}_i'\boldsymbol{\beta})^2$. Observing $\mathbf{y}$ this gives a likelihood function for $(\boldsymbol{\beta}, v)$. We can write

$$Q(\mathbf{y}, \boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{F}\mathbf{F}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + R$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}\mathbf{F}')^{-1}\mathbf{F}\mathbf{y}$ and $R = (\mathbf{y} - \mathbf{F}'\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{F}'\hat{\boldsymbol{\beta}})$. This assumes that $\mathbf{F}$ is of full rank p, otherwise an appropriate linear transformation of the design vectors will reduce to a full rank matrix and the model simply reduces in dimension. Here $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and the residual sum of squares R gives the MLE of v as R/n; a more usual estimate of v is $s^2 = R/(n-p)$, with $n-p$ being the associated degrees of freedom.

1.16 Reference Bayesian Analysis. Reference Bayesian analysis is based on the traditional reference (improper) prior $p(\boldsymbol{\beta}, v) \propto 1/v$. The corresponding posterior density is $p(\boldsymbol{\beta}, v|\mathbf{y}, \mathbf{F}) \propto p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v)/v$ and has the following features.

- The marginal posterior for $\boldsymbol{\beta}$ is multivariate T with $n-p$ degrees of freedom, has mode $\hat{\boldsymbol{\beta}}$ and density

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}) = c(n,p)|\mathbf{F}\mathbf{F}'|^{1/2}\{1 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{F}\mathbf{F}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/(n-p)s^2\}^{-n/2}$$

with $c(n,p) = \Gamma(n/2)/[\Gamma((n-p)/2)(s\pi(n-p))^{p/2}]$. For a large n, the posterior is roughly $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, s^2(\mathbf{F}\mathbf{F}')^{-1})$. Note also that, given any assumed value of v, the conditional posterior for $\boldsymbol{\beta}$ is exactly normal, namely $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, v(\mathbf{F}\mathbf{F}')^{-1})$.

- The total sum of squares of the responses $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} y_i^2$ factorises as $\mathbf{y}'\mathbf{y} = R + \hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}$. The sum of squares explained by the regression is $\mathbf{y}'\mathbf{y} - R = \hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}$; this is also called the fitted sum of squares, and a larger value implies a smaller residual sum of squares and, in this sense, a closer fit to the data.

- In connection with model comparisons and related issues, a key quantity is the value of the marginal density of the response data (conditional on the model form, **F** and the adopted reference prior) at the observed value of **y**, namely

$$p(\mathbf{y}|\mathbf{F}) = \int p(\mathbf{y}|\hat{\boldsymbol{\beta}}, v)/v \; d\boldsymbol{\beta} dv = c\frac{\Gamma((n-p)/2)}{\pi^{(n-p)/2}}|\mathbf{FF}'|^{-1/2}R^{-(n-p)/2},$$

for some constant c that does not depend on **F** or p. This can also be written as

$$p(\mathbf{y}|\mathbf{F}) \propto \frac{\Gamma((n-p)/2)}{\pi^{(n-p)/2}}|\mathbf{FF}'|^{-1/2}\{1-\hat{\boldsymbol{\beta}}'\mathbf{FF}'\hat{\boldsymbol{\beta}}/(\mathbf{y}'\mathbf{y})\}^{(p-n)/2}.$$

For large n, the term $\{1-\hat{\boldsymbol{\beta}}'\mathbf{FF}'\hat{\boldsymbol{\beta}}/(\mathbf{y}'\mathbf{y})\}^{(p-n)/2}$ in the above expression is approximately $\exp(\hat{\boldsymbol{\beta}}'\mathbf{FF}'\boldsymbol{\beta}/2r)$ where $r = \mathbf{y}'\mathbf{y}/(n-p)$.

Some additional comments:

- For models with the same number of parameters that differ only through **F**, the corresponding observed data densities will tend to be larger for those models with larger values of the explained sum of squares $\hat{\boldsymbol{\beta}}'\mathbf{FF}'\hat{\boldsymbol{\beta}}$ (though the determinant term plays a role too). Otherwise, $p(\mathbf{y}|\mathbf{F})$ also depends on the parameter dimension p.
- Treating **F** as a "parameter" of the model, and making this explicit in the model, we see that $p(\mathbf{y}|\mathbf{F})$ is the likelihood function for **F** from the data (in this reference analysis).
- *Orthogonal regression.* If $\mathbf{FF}' = k\mathbf{I}_p$ for some k, then everything simplifies. Write $\mathbf{f}_j^*$ for the j^{th} column of $\mathbf{F}'$, and β_j for the corresponding component of the parameter vector $\boldsymbol{\beta}$. Then $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ where each $\hat{\beta}_j$ is the individual MLE from a model on $\mathbf{f}_j^*$ alone, i.e. $\mathbf{y} = \mathbf{f}_j^*\beta_j + \boldsymbol{\epsilon}$, and the elements of $\boldsymbol{\beta}$ are uncorrelated under the posterior T distribution. The explained sum of squares partitions into a sum of individual pieces too, namely $\hat{\boldsymbol{\beta}}'\mathbf{FF}'\hat{\boldsymbol{\beta}} = \sum_{j=1}^{p}\mathbf{f}_j^{*'}\mathbf{f}_j^*\hat{\beta}_j^2$, and so calculations as well as interpretations are easy.

Example 1.16.1 *Reference analysis in the AR(1) model.*

For the conditional likelihood, the reference prior is given by $p(\phi, v) \propto 1/v$. The MLE for ϕ is $\phi_{\text{ML}} = \sum_{t=2}^{n} y_{t-1}y_t/\sum_{t=1}^{n-1} y_t^2$. Under the reference prior $\phi_{\text{MAP}} = \phi_{\text{ML}}$. The residual sum of squares is given by

$$R = \sum_{t=2}^{n} y_t^2 - \frac{(\sum_{t=2}^{n} y_t y_{t-1})^2}{\sum_{t=1}^{n-1} y_t^2},$$

and so, $s^2 = R/(n-2)$ estimates v. The marginal posterior distribution of ϕ is a univariate t distribution with $n-2$ degrees of freedom, centered at ϕ_{ML} with scale

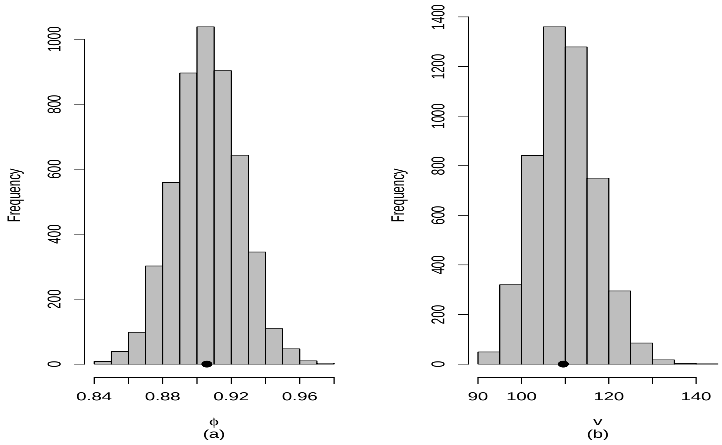

Fig. 1.10 (a) $p(\phi|y)$; (b) $p(v|y)$.

$s^2(\mathbf{FF}')^{-1}$, i.e.,

$$(\phi|\mathbf{y},\mathbf{F}) \sim t_{(n-2)}\left(\frac{\sum_{t=2}^{n} y_{t-1}y_t}{\sum_{t=1}^{n-1} y_t^2}, \frac{\sum_{t=2}^{n} y_t^2 \sum_{t=2}^{n} y_{t-1}^2 - (\sum_{t=2}^{n} y_t y_{t-1})^2}{\left(\sum_{t=1}^{n-1} y_t^2\right)^2 (n-2)}\right).$$

Finally, the posterior for v is a scaled inverse chi-squared with $n-2$ degrees of freedom and scale s^2, $Inv-\chi^2(v|n-2,s^2)$, or equivalently, an inverse gamma with parameters $(n-2)/2$ and $(n-2)s^2/2$, i.e. $IG(v|(n-2)/2,(n-2)s^2/2)$.

As an illustration, a reference analysis was performed for a time series of 500 points simulated from an AR(1) model with $\phi = 0.9$ and $v = 100$. Figures 1.10 (a) and (b) display the marginal posterior densities of $(\phi|\mathbf{y})$ and $(v|\mathbf{y})$ based on a sample of 5,000 observations from the joint posterior. The circles in the histogram indicate ϕ_{ML} and s^2 respectively.

1.17 Conjugate Bayesian Analysis. Let $p(y_t|\boldsymbol{\theta})$ be a likelihood function. A class Π of prior distributions forms a *conjugate family* if the posterior $p(\boldsymbol{\theta}|y_t)$ belongs to the class Π for every prior $p(\boldsymbol{\theta})$ in Π.

Consider again the model $\mathbf{y} = \mathbf{F}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{F}$ a known $p \times n$ design matrix and $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|0, v\mathbf{I}_n)$. In a conjugate Bayesian analysis for this model priors of the form

$$p(\boldsymbol{\beta}, v) = p(\boldsymbol{\beta}|v)p(v) = N(\boldsymbol{\beta}|\mathbf{m}_0, v\mathbf{C}_0) \times IG(v|n_0/2, d_0/2), \qquad (1.22)$$

are taken, with $\mathbf{m}_0$ a vector of dimension p and $\mathbf{C}_0$ a $p \times p$ matrix. Both, $\mathbf{m}_0$ and $\mathbf{C}_0$ are known quantities. The corresponding posterior distribution has the following form

$$\begin{aligned} p(\boldsymbol{\beta}, v|\mathbf{F}, \mathbf{y}) &\propto v^{-[(p+n+n_0)/2+1]} \times \\ &\quad \exp\left\{-\frac{(\boldsymbol{\beta}-\mathbf{m}_0)'\mathbf{C}_0^{-1}(\boldsymbol{\beta}-\mathbf{m}_0) + (y-\mathbf{F}'\boldsymbol{\beta})'(y-\mathbf{F}'\boldsymbol{\beta}) + d_0}{2v}\right\}. \end{aligned}$$

This posterior distribution has the following features:

- $(\mathbf{y}|\mathbf{F}, v) \sim N(\mathbf{F}'\mathbf{m}_0, v(\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n))$.
- The posterior for $\boldsymbol{\beta}$ conditional on v is Gaussian, $(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}, v) \sim N(\mathbf{m}, v\mathbf{C})$, with

$$\begin{aligned} \mathbf{m} &= \mathbf{m}_0 + \mathbf{C}_0\mathbf{F}[\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n]^{-1}(\mathbf{y} - \mathbf{F}'\mathbf{m}_0) \\ \mathbf{C} &= \mathbf{C}_0 - \mathbf{C}_0\mathbf{F}[\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n]^{-1}\mathbf{F}'\mathbf{C}_0, \end{aligned}$$

or, defining $\mathbf{e} = \mathbf{y} - \mathbf{F}'\mathbf{m}_0$, $\mathbf{Q} = \mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n$ and $\mathbf{A} = \mathbf{C}_0\mathbf{F}\mathbf{Q}^{-1}$ we have, $\mathbf{m} = \mathbf{m}_0 + \mathbf{A}\mathbf{e}$ and $\mathbf{C} = \mathbf{C}_0 - \mathbf{A}\mathbf{Q}\mathbf{A}'$.

- $(v|\mathbf{F}, \mathbf{y}) \sim IG(n^*/2, d^*/2)$, with $n^* = n + n_0$ and

$$d^* = (\mathbf{y} - \mathbf{F}'\mathbf{m}_0)'\mathbf{Q}^{-1}(\mathbf{y} - \mathbf{F}'\mathbf{m}_0) + d_0.$$

- $(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}) \sim T_{n^*}[\mathbf{m}, d^*\mathbf{C}/n^*]$

Example 1.17.1 *Conjugate analysis in the AR(1) model.*
Assume we choose a prior of the form $\phi|v \sim N(0, v)$ and $v \sim IG(n_0/2, d_0/2)$, with n_0 and d_0 known. Then, $p(\phi|\mathbf{F}, \mathbf{y}, v) \sim N(m, vC)$ with

$$m = \frac{\sum_{t=1}^{n-1} y_t y_{t+1}}{\sum_{t=1}^{n-1} y_t^2 + 1}, \quad C = \frac{1}{1 + \sum_{t=1}^{n-1} y_t^2},$$

$(v|F, \mathbf{y}) \sim IG(n^*/2, d^*/2)$ with $n^* = n + n_0 - 1$ and

$$d^* = \sum_{t=2}^{n} y_t^2 - \frac{\left(\sum_{t=1}^{n-1} y_t y_{t+1}\right)^2}{\sum_{t=1}^{n-1} y_t^2 + 1} + d_0.$$

1.18 Non-conjugate Bayesian analysis. For the general regression model the reference and conjugate priors produce joint posterior distributions that have closed analytical forms. However, in many scenarios it is either not possible or not desirable to work with a conjugate prior or with a prior that leads to a posterior distribution that can be written in analytical form. In these cases it might be possible to use analytical or numerical approximations to the posterior. Another alternative consists on summarising the inference by obtaining random draws from the posterior distribution. Sometimes it is possible to obtain such draws by direct simulation, but often this is not the case and so, methods such as Markov chain Monte Carlo (MCMC) are used.

Consider for example the AR(1) model under the full likelihood (1.17). No conjugate prior is available in this case. Furthermore, a prior of the form $p(\phi, v) \propto$

$1/v$ does not produce a posterior distribution in closed form. In fact, the joint posterior distribution is such that

$$p(\phi, v|y_{1:n}) \propto v^{-(n/2+1)}(1-\phi^2)^{1/2}\exp\left\{\frac{-Q^*(\phi)}{2v}\right\}. \tag{1.23}$$

Several approaches could be considered to summarise this posterior distribution. For instance, we could take a normal approximation to $p(\phi, v|y_{1:n})$ centered at the ML or MAP estimates of (ϕ, v). In general, the normal approximation to a posterior distribution $p(\boldsymbol{\theta}|y_{1:n})$ is given by

$$p(\boldsymbol{\theta}|y_{1:n}) \approx N(\hat{\boldsymbol{\theta}}, v(\hat{\boldsymbol{\theta}})), \tag{1.24}$$

with $\hat{\boldsymbol{\theta}} = \theta_{\text{MAP}}$ and $v(\boldsymbol{\theta}) = [-\log p''(\boldsymbol{\theta}|y_{1:n})]^{-1}$.

Alternatively, it is possible to use iterative MCMC methods to obtain samples from $p(\phi, v|y_{1:n})$. We summarise two of the most widely used MCMC methods below: the Metropolis algorithm and the Gibbs sampler. For full consideration of MCMC methods see for example Gamerman (1997).

Posterior Sampling

1.19 The Metropolis-Hastings algorithm. Assume that our target posterior distribution, $p(\boldsymbol{\theta}|y_{1:n})$, can be computed up to a normalising constant. The Metropolis algorithm (Metropolis *et al.*, 1953; Hastings, 1970) creates a sequence of random draws $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots,$ whose distributions converge to the target distribution. Each sequence can be considered as a random walk whose stationary distribution is $p(\boldsymbol{\theta}|y_{1:n})$. The sampling algorithm can be summarised as follows:

- Draw a starting point $\boldsymbol{\theta}^0$ with $p(\boldsymbol{\theta}^0|y_{1:n}) > 0$ from a starting distribution $p_0(\boldsymbol{\theta})$.
- For $i = 1, 2, \ldots$

1. Sample a candidate $\boldsymbol{\theta}^*$ from a jumping distribution $J_i(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})$. If the distribution J_i is symmetric, i.e., if $J_i(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = J_i(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ for all $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ and i, then we refer to the algorithm as the Metropolis algorithm. If J_i is not symmetric we refer to the algorithm as the Metropolis-Hastings algorithm.

2. Compute the importance ratio

$$r = \frac{p(\boldsymbol{\theta}^*|y_{1:n})/J_i(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})}{p(\boldsymbol{\theta}^{i-1}|y_{1:n})/J_i(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)}.$$

3. Set

$$\boldsymbol{\theta}^i = \begin{cases} \boldsymbol{\theta}^* & \text{with probability} = \min(r,1) \\ \boldsymbol{\theta}^{i-1} & \text{otherwise.} \end{cases}$$

An ideal jumping distribution is one that is easy to sample from and makes the evaluation of the importance ratio easy. In addition, the jumping distributions $J_i(\cdot|\cdot)$ should be such that each jump moves a reasonable distance in the parameter space so that the random walk is not too slow, and also, the jumps should not be rejected too often.

1.20 Gibbs sampling. Assume $\boldsymbol{\theta}$ has k components, i.e. $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_k)'$. The Gibbs sampler (Geman and Geman, 1984) can be viewed as a special case of the Metropolis-Hastings algorithm for which the jumping distribution at each iteration i is a function of the conditional posterior density $p(\boldsymbol{\theta}^*_j|\boldsymbol{\theta}^{i-1}_{-j}, y_{1:n})$, where $\boldsymbol{\theta}_{-j}$ denotes a vector with all the components of $\boldsymbol{\theta}$ except for component $\boldsymbol{\theta}_j$. In other words, for each component of $\boldsymbol{\theta}$ we do a Metropolis step for which the jumping distribution is given by

$$J_{j,i}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1}) = \begin{cases} p(\boldsymbol{\theta}^*_j|\boldsymbol{\theta}^{i-1}_{-j}, y_{1:n}) & \text{if } \boldsymbol{\theta}^*_{-j} = \boldsymbol{\theta}^{i-1}_{-j} \\ 0 & \text{otherwise,} \end{cases}$$

and so, $r = 1$ and every jump is accepted.

If it is not possible to sample from $p(\boldsymbol{\theta}^*_j|\boldsymbol{\theta}^i_{-j}, y_{1:n})$, then an approximation $g(\boldsymbol{\theta}^*_j|\boldsymbol{\theta}^{i-1}_{-j})$ can be considered. However, in this case it is necessary to compute the Metropolis acceptance ratio r.

1.21 Convergence. In theory, a value from $p(\boldsymbol{\theta}|y_{1:n})$ is obtained by MCMC when the number of iterations of the chain approaches infinity. In practice, a value obtained after a sufficiently large number of iterations is taken as a value from $p(\boldsymbol{\theta}|y_{1:n})$. How can we determine how many MCMC iterations are enough to obtain convergence? As pointed out in Gamerman (1997), there are two general approaches to the study of convergence. One is probabilistic and tries to measure distances and bounds on distribution functions generated from a chain. So, for example, it is possible to measure the total variation distance between the distribution of the chain at iteration i and the target distribution $p(\boldsymbol{\theta}|y_{1:n})$. An alternative approach consists on studying the convergence of the chain from a statistical perspective. This approach is easier and more practical than the probabilistic one, however, it cannot guarantee convergence.

There are several ways of monitoring convergence from a statistical viewpoint, ranging from graphical displays of the MCMC traces for all or some of the model parameters or functions of such parameters, to sophisticated statistical tests. As mentioned before, one of the two main problems with simulation-based iterative methods is deciding whether the chain has reached convergence, i.e., if the number of iterations is large enough to guarantee that the available samples are draws from the target posterior distribution. In addition, once the chain has reached convergence it is important to obtain uncorrelated draws from the posterior distribution. Some well known statistical tests to assess convergence are implemented in freely available software such as Bayesian Output Analysis (BOA) (currently available at `www.public-health.uiowa.edu/boa/`, Smith 2004). Specifically, BOA includes the following convergence diagnostics: the Brooks, Gelman and Rubin convergence diagnostics for a list of MCMC sequences (Brooks and Gelman, 1998; Gelman and Rubin, 1992), which monitors the mixing of the simulated sequences by comparing the within and between variance of the sequences; the Geweke (Geweke, 1992) and Heidelberger and Welch (Welch, 1983) diagnostics, which are based on sequential testing of portions of the simulated chains to determine if they correspond to samples from the same distribution; and the Raftery and Lewis method (Raftery and Lewis, 1992), which considers the problem of how many iterations are needed to estimate a particular posterior quantile from a single MCMC chain. BOA also provides the user with some descriptive plots of the chains —auto-correlations, density, means and trace plots— as well as plots of some of the convergence diagnostics.

Example 1.21.1 *A Metropolis-Hastings for an AR(1) model.*

Consider again the AR(1) model with the unconditional likelihood (1.17) and a prior of the form $p(\phi, v) \propto 1/v$. An MCMC to obtain samples from the posterior distribution is described below. For each iteration $i = 1, 2, \ldots$

- Sample v^i from $(v|\phi, y_{1:n}) \sim IG(n/2, Q^*(\phi)/2)$. Note that this is a Gibbs step and so every draw will be accepted.
- Sample ϕ^i using a Metropolis step with a Gaussian jumping distribution. Therefore, at iteration i we draw a candidate sample ϕ^* from a Gaussian distribution centered at ϕ^{i-1}, this is

$$\phi^* \sim N\left(\phi^{i-1}, cv\right),$$

with c a constant. The value of c controls the acceptance rate of the algorithm. In practice, target acceptance rates usually go from 25% to 40%. See for instance Gelman *et al.* (2004), Chapter 11 for a discussion on how to set the value of c.

In order to illustrate the MCMC methodology, we considered 500 observations generated from an AR(1) model with coefficient $\phi = 0.9$ and variance $v = 1.0$. The MCMC scheme above was implemented in order to obtain posterior estimation of the model parameters based on the 500 synthetic observations. Figures 1.11 (a) and (b) display the traces of the model parameters for two chains of 1,000 MCMC samples using $c = 2$. Several values of c were considered and the value $c = 2$ was chose since it led to a Metropolis acceptance rate of approximately 37%. The starting values for the chains were set at $v^0 = 0.1$, $\phi^0 = 0.5$ and $v^0 = 3$, $\phi^0 = 0.0$. It is clear from the pictures that there seems to be no convergence problems. Figures 1.11 (c) and (d) show the posterior distributions for ϕ and v based on 450 samples of one of the MCMC chains taken every other iteration after a burn-in period of 100 iterations. The early iterations of a MCMC output are usually discarded in order to eliminate, or diminish as much as possible, the effect of the starting distribution. This is referred to

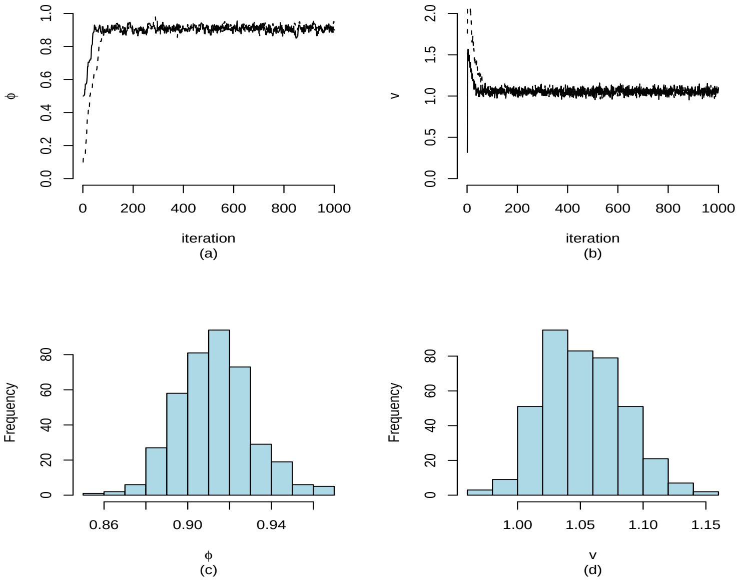

Fig. 1.11 (a) and (b) Traces of 1,000 MCMC samples of the parameters ϕ and v respectively. The draws from two chains are displayed. The solid lines correspond to traces from a chain with starting values of $(\phi^0, v^0) = (0.5, 0.1)$ and the dotted lines correspond to traces with starting values of $(\phi^0, v^0) = (0, 3)$. Panels (c) and (d) show histograms of 450 samples from the marginal posteriors of ϕ and v. The samples were taken every other MCMC iteration after a burn-in period of 100 iterations.

as burn-in. The length of the burn-in period varies greatly depending on the context and the complexity of the MCMC sampler.

Discussion and Further Topics

Appendix

1.22 The uniform distribution. A random variable x follows a uniform distribution in the interval (a, b), with $a < b$, $x \sim U(a, b)$, or $p(x) = U(x|a, b)$, if its density function is given by

$$p(x) = \frac{1}{(b-a)}, \quad x \in [a, b].$$

1.23 The univariate normal distribution. A random variable x follows normal distribution with mean μ and variance v, if its density is given by

$$p(x) = \frac{1}{\sqrt{2\pi v}} \mathrm{exp}\left(-\frac{(x-\mu)^2}{2v}\right).$$

We use $x \sim N(\mu, v)$, or $p(x) = N(x|\mu, v)$, to denote that x follows a univariate normal distribution. If $\mu = 0$ and $\sigma = 1$ we say that x follows a standard normal distribution.

1.24 The multivariate normal distribution. A random vector of dimension k $\mathbf{x} = (x_1, \ldots, x_k)'$, that follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ, $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$, or $p(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, has a density function given by

$$p(\mathbf{x}) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \mathrm{exp}\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

1.25 The gamma and inverse-gamma distributions. A random variable x that follows a gamma distribution with shape parameter α and inverse scale parameter β, $x \sim G(\alpha, \beta)$, or $p(x) = G(x|\alpha, \beta)$, has a density of the form

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

where $\Gamma(\cdot)$ is the gamma function. If $\frac{1}{x} \sim G(\alpha, \beta)$, then x follows an inverse-gamma distribution $x \sim IG(\alpha, \beta)$, or $p(x) = IG(x|\alpha, \beta)$ with

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}, \quad x > 0.$$

1.26 The chi-square distribution. x follows a chi-square distribution with ν degrees of freedom if its density is given by

$$p(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \quad x > 0.$$

This distribution is the same as the $G(x|\nu/2, 1/2)$.

1.27 The inverse-chi-square and the scaled inverse chi-square distributions. If $x \sim Inv - \chi^2(\nu)$, or $p(x) = Inv - \chi^2(\nu)$, then $x \sim IG(\nu/2, 1/2)$. Also, if x follows a scaled inverse-chi-squared with ν degrees of freedom and scale s, i.e., $x \sim Inv - \chi^2(\nu, s^2)$, then $x \sim IG(\nu/2, \nu s^2/2)$.

1.28 The univariate Student-t distribution. If x follows a Student-t distribution with ν degrees of freedom, location μ and scale σ, $x \sim t_\nu(\mu, \sigma^2)$, if its density is

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}.$$

1.29 The multivariate Student-t distribution. A random vector $\mathbf{x}$ of dimension k follows a multivariate Student-t distribution with ν degrees of freedom, location $\boldsymbol{\mu}$ and scale matrix Σ, $\mathbf{x} \sim T_\nu(\boldsymbol{\mu}, \Sigma)$ if its density is given by

$$p(\mathbf{x}) = \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2}} |\Sigma|^{-1/2} \left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+k)/2}.$$

Problems

1. Consider the AR(1) model $y_t = \phi y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, v)$.

   (a) Find the MLE of (ϕ, v) for the conditional likelihood.

   (b) Find the MLE of (ϕ, v) for the unconditional likelihood (1.17).

   (c) Assume that v is known. Find the MAP estimator of ϕ under a uniform prior $p(\phi) = U(\phi|0, 1)$ for the conditional and unconditional likelihoods.

2. Show that the distributions of $(\phi|\mathbf{y}, \mathbf{F})$ and $(v|\mathbf{y}, \mathbf{F})$ obtained for the AR(1) reference analysis are the ones given in example 1.16.1.

3. Show that the distributions of $(\phi|\mathbf{y}, \mathbf{F})$ and $(v|\mathbf{y}, \mathbf{F})$ obtained for the AR(1) conjugate analysis are the ones given in example 1.17.1.

4. Consider the following models:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \quad (1) \\ y_t &= a\cos(2\pi\omega_0 t) + b\sin(2\pi\omega_0 t) + \epsilon_t \quad (2), \end{aligned}$$

with $\epsilon_t \sim N(0, v)$.

(a) Sample 200 observations from each model using your favorite choice of the parameters. Make sure your choice for (ϕ_1, ϕ_2) in model (1) lies in the stationary region. This is, choose (ϕ_1, ϕ_2) such that $-2 < \phi_1 < 2$, $\phi_1 < 1 - \phi_2$ and $\phi_1 > \phi_2 - 1$.

(b) Find the MLE of the models parameters. Use the conditional likelihood for model (1).

(c) Find the MAP estimators of the model parameters under the reference prior. Again, use the conditional likelihood for model (1).

(d) Sketch the marginal posterior distributions $p(\phi_1, \phi_2|y_{1:n})$ and $p(v|y_{1:n})$ for model (1).

(e) Sketch the marginal posterior distributions $p(a, b|y_{1:n})$ and $p(v|y_{1:n})$.

(f) Perform a conjugate Bayesian analysis, i.e., repeat (c) to (e) assuming conjugate prior distributions in both models. Study the sensitivity of the posterior distributions to the choice of the hyperparameters in the prior.

5. Refer to the conjugate analysis of the AR(1) model in example 1.17.1. Using the fact that $\phi|\mathbf{y}, \mathbf{F}, v \sim N(m, vC)$, find the posterior mode of v using the EM algorithm.

6. Sample 1,000 observations from the model (1.1). Using a prior distribution of the form $p(\phi_1^{(i)}) = p(\phi_2^{(i)}) = N(0, c)$, for some c and $i = 1, 2$, $p(\theta) = U(\theta| - a, a)$ and $p(v) = IG(\alpha_0, \beta_0)$, obtain samples from the joint posterior distribution by implementing a Metropolis-Hastings algorithm.

2 Traditional Time Series Models

Autoregressive time series models are central to modern stationary time series data analysis and, as components of larger models or in suitably modified and generalised forms, underlie non-stationary time-varying models. The concepts and structure of linear autoregressive models also provide important background material for appreciation of non-linear models. This chapter discusses model forms and inference for AR models, and related topics. This is followed by discussion of the class of stationary autoregressive, moving average models, one which a large area of traditional linear time series analysis is predicated.

Structure of Autoregressions

2.1 Consider the time series of equally-spaced quantities $y_t,$ for $t = 1, 2, \ldots,$ arising from the model

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t, \tag{2.1}$$

where ϵ_t is a sequence of uncorrelated error terms and the ϕ_j are constant parameters. This is a sequentially defined model; y_t is generated as a function of past values, parameters and errors. The ϵ_t are termed innovations, and are assumed to be conditionally independent of the past values of the series. They are also often assumed normally distributed, $N(\epsilon_t|0, v)$, and so they are independent. This is a standard autoregressive model framework, AR(p) for short; p is the order of the autoregression.

AR models may be viewed from a purely empirical standpoint; the data are assumed related over time and the AR form is about the simplest class of empirical models for exploring dependencies. A more formal motivation is, of course, based on the genesis in stationary stochastic process theory. Here we proceed to inference in the model class.

The sequential definition of the model and its Markovian nature imply a sequential structuring of the data density

$$p(y_{1:T}) = p(y_{1:p}) \prod_{t=p+1}^{T} p(y_t|y_{(t-p):(t-1)}) \tag{2.2}$$

for any $T > p$. The leading term is the joint density of the p initial values of the series, as yet undefined. Here the densities are conditional on $(\phi_1, \dots, \phi_p, v)$; though this is not made explicit in the notation. If the first p values of the series are known and viewed as fixed constants, and $T = n + p$ for some $n > 1$, then the conditional density of $\mathbf{y} = (y_T, y_{T-1}, \dots, y_{p+1})'$ given the first p values is

$$\begin{aligned} p(\mathbf{y}|y_{1:p}) &= \prod_{t=p+1}^{T} p(y_t|y_{(t-p):(t-1)}) \\ &= \prod_{t=p+1}^{T} N(y_t|\mathbf{f}_t'\boldsymbol{\phi}, v) = N(\mathbf{y}|\mathbf{F}'\boldsymbol{\phi}, v\mathbf{I}_n), \end{aligned} \tag{2.3}$$

where $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$, $\mathbf{f}_t = (y_{t-1}, \dots, y_{t-p})'$ and $\mathbf{F}$ is a $p \times n$ matrix given by $\mathbf{F} = [\mathbf{f}_T, \dots, \mathbf{f}_{p+1}]$. This has a linear model form and so, the standard estimation methods discussed in Chapter 1 apply.

Practically useful extensions of the model (2.3) include models with additional regression terms for the effects of independent regressor variables on the series, differing variances for the ϵ_t over time, and non-normal error distributions. This standard normal linear model is a very special case of autoregressions which, generally, define models via sequences of conditional distributions for $(y_t|\mathbf{f}_t)$ over time.

2.2 Stationary AR Processes. The series y_t, assumed (at least in principle) to extend over all time $t = 0, \pm 1, \pm 2, \dots$, follows a stationary autoregressive model of order p, if the stationarity conditions are satisfied. With the innovations independent $N(\epsilon_t|0, v)$, the stationary distribution of each y_t, and of any set of $k > 1$ of the y_t, is zero-mean normal. Extending the model to include a non-zero mean μ for each y_t gives $y_t = \mu + (\mathbf{f}_t - \mu\mathbf{l})'\boldsymbol{\phi} + \epsilon_t$ where $\mathbf{l} = (1, \dots, 1)'$, or $y_t = \beta + \mathbf{f}_t'\boldsymbol{\phi} + \epsilon_t$ where $\beta = (1 - \mathbf{l}'\boldsymbol{\phi})\mu$. The special case of $p = 2$ is discussed in detail in the following Section.

As mentioned in Example 1.8.1, when $p = 1$, the AR process is stationary for $-1 < \phi_1 < 1$ when the stationary distribution of each of the y_t is $N(y_t|0, v/(1-\phi_1^2))$. At the boundary $\phi_1 = 1$ the model becomes a non-stationary random walk. The bivariate stationary distribution of $(y_t, y_{t-1})'$ is normal with correlation $\rho(1) = \phi_1$; that of $(y_t, y_{t-k})'$ for any k is $\rho(k) = \phi_1^k$. A positive autoregressive parameter ϕ_1 leads to a process that wanders away from the stationary mean of the series, with such excursions being more extensive when ϕ_1 is closer to unity; $\phi_1 < 0$ leads to more oscillatory behaviour about the mean.

With $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$, the process is stationary for parameter values lying in the region $-2 < \phi_1 < 2$, $\phi_1 < 1 - \phi_2$ and $\phi_1 > \phi_2 - 1$. Further discussion appears in the following Section.

In the case of general order p models, the stationarity condition imposes a set of restrictions on the coefficients $\boldsymbol{\phi}$ best represented in terms of the roots of the autoregressive polynomial $\Phi(u) = 1 - \sum_{j=1}^{p} \phi_j u^j$, for $|u| \leq 1$. This arises through the representation of (2.1) as

$$\Phi(B) y_t = \epsilon_t,$$

using the back-shift operator B, with $By_t = y_{t-1}$. The process is stationary if, and only if, the inversion of this equation, namely

$$y_t = \Phi(B)^{-1} \epsilon_t = \sum_{j=0}^{\infty} \pi_j \epsilon_{t-j}$$

exists and converges, and this is true only if the roots of the polynomial $\Phi(u)$ have moduli greater than unity. Write $\Phi(u) = \prod_{j=1}^{p} (1 - \alpha_j u)$ so that the roots are the reciprocals of the α_j. Generally, the α_j may be real-valued or may appear as pairs of complex conjugates. Either way, the process is stationary if $|\alpha_j| < 1$ for all j, and non-stationary otherwise.

2.3 State-Space Representation of an AR(p). The state-space representation of an AR(p) model has utility in both, exploring mathematical structure and, as we shall see later, in inference and data analysis. One version of this representation of (2.1) is simply

$$y_t = \mathbf{F}'\mathbf{x}_t \tag{2.4}$$

$$\mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \boldsymbol{\omega}_t, \tag{2.5}$$

where $\mathbf{x}_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$, the state vector at time t. The innovation at time t appears in the error vector $\boldsymbol{\omega}_t = (\epsilon_t, 0, \ldots, 0)'$. In addition, $\mathbf{F} = (1, 0, \ldots, 0)'$ and

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}. \tag{2.6}$$

The expected behaviour of the future of the process may be exhibited through the forecast function $f_t(k) = E(y_{t+k}|y_{1:t})$ as a function of integers $k > 0$ for any fixed "origin" $t \geq p$, conditional on the most recent p values of the series in the current state vector $\mathbf{x}_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$. We have $f_t(k) = \mathbf{F}'\mathbf{G}^k\mathbf{x}_t$. The form is most easily appreciated in cases when the matrix $\mathbf{G}$ has distinct eigenvalues, real and/or complex. It easily follows that these eigenvalues are precisely the reciprocals roots of the autoregressive polynomial equation $\Phi(u) = 0$, namely the α_j above. Then

$$f_t(k) = \sum_{j=1}^{p} c_{tj}\alpha_j^k, \tag{2.7}$$

where the c_{tj} are (possibly complex valued) constants depending on $\boldsymbol{\phi}$ and the current state $\mathbf{x}_t$, and the α_j's are the p distinct eigenvalues/reciprocal roots. The c_{tj} coefficients are given by $c_{tj} = d_j e_{tj}$. The d_j and e_{tj} values are the elements of the $p-$vectors $\mathbf{d} = \mathbf{E}'\mathbf{F}$ and $\mathbf{e}_t = \mathbf{E}^{-1}\mathbf{x}_t$, where $\mathbf{E}$ is the eigenmatrix of $\mathbf{G}$, i.e., $\mathbf{E}$ is the $p \times p$ matrix whose columns are the eigenvectors in order corresponding to the eigenvalues α_j.

The form of the forecast function depends on the combination of real and complex eigenvalues of $\mathbf{G}$. Suppose α_j, for example, is real and positive; the contribution to the forecast function is then $c_{tj}\alpha_j^k$. If the process is stationary $|\alpha_i| < 1$ for all i so that this function of k decays exponentially to zero, monotonically if $\alpha_j > 0$, otherwise oscillating between consecutive positive and negative values. If $|\alpha_j| \geq 1$ the process is non-stationary and the forecast function is explosive. The relative contribution to the overall forecast function is measured by the decay rate and the initial amplitude c_{tj}, the latter depending explicitly on the current state, and therefore having different impact at different times as the state varies in response to the innovations sequence.

In the case of complex eigenvalues, the fact that $\mathbf{G}$ is real-valued implies that any complex eigenvalues appear in pairs of complex conjugates. Suppose, for example, that α_1 and α_2 are complex conjugates $\alpha_1 = r\exp(i\omega)$ and $\alpha_2 = r\exp(-i\omega)$ with modulus r and argument ω. In this case, the corresponding complex factors c_{t1} and c_{t2} are conjugate, $a_t\exp(\pm i b_t)$, and the resulting contribution to $f_t(k)$, which must be real-valued, is

$$c_{t1}\alpha_1^k + c_{t2}\alpha_2^k = 2a_t r^k \cos(\omega k + b_t).$$

Hence, ω determines the constant frequency of a sinusoidal oscillation in the forecast function, the corresponding wavelength or period being $\lambda = 2\pi/\omega$. In a stationary model $|r| < 1$, and so, the sinusoidal oscillations over times $t + k$ with $k > 0$ are subject to exponentially decay through the damping factor r^k, with additional oscillatory effects if $r < 0$. In non-stationary cases the sinusoidal variation explodes in amplitude as $|r|^k$ increases. The factors a_t and b_t determine the relative amplitude and phase of the component. The amplitude factor $2a_t$ measures the initial magnitude of the contribution of this term to the forecast function, quite separately from the decay factor r. At a future time epoch $s > t$, the new state vector $\mathbf{x}_s$ will define an updated forecast function $f_s(k)$ with the same form as (2.7) but with updated coefficients c_{sj} depending on $\mathbf{x}_s$, and so affecting the factors a_s and b_s. Therefore, as time evolves, the relative amplitudes and phases of the individual components vary according to the changes in state induced by the sequence of innovations.

Generally, the forecast function (2.7) is a linear combination of exponentially decaying or exploding terms, and decaying or exploding factors multiplying sinusoids of differing periods. Returning to the model (2.1), this basic expected behaviour translates into a process that has the same form but in which, at each time point, the innovation ϵ_t provides a random shock to the current state of the process. This describes a process that exhibits such exponentially damped or exploding behaviour, possibly with periodic components, but in which the amplitudes and phases of the components are randomly varying over time in response to the innovations.

2.4 Characterisation of AR(2) Processes. The special case of $p = 2$ is illuminating and of practical importance in its own right. The process is stationary if $-2 < \phi_1 < 2$, $\phi_1 < 1 - \phi_2$ and $\phi_1 > \phi_2 - 1$. In such cases, the quadratic characteristic polynomial $\Phi(u) = 0$ has reciprocal roots α_i lying within the unit circle, and these define:

- Two real roots when $\phi_1^2 + 4\phi_2 \geq 0$, in which case the forecast function decays exponentially;

- A pair of complex conjugate roots $r\exp(\pm i\omega)$ when $\phi_1^2 + 4\phi_2 < 0$. The roots have modulus $r = \sqrt{-\phi_2}$ and argument given by $\cos(\omega) = |\phi_1|/2r$. The forecast function behaves as an exponentially damped cosine.

We already know that $-2 < \phi_1 < 2$ for stationarity; for complex roots, we have the additional restriction to $-1 < \phi_2 < -\phi_1^2/4$. So, in these cases, the model $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$ represents a quasi-cyclical process, behaving as a damped sine wave of fixed period $2\pi/\omega$, but with amplitude and phase characteristics randomly varying over time in response to the innovations ϵ_t. A large innovations variance v induces greater degrees of variation in this dynamic, quasi-cyclical process. If the innovation variance is very small, or were to become zero at some point, the process would decay to zero in amplitude due to the damping factor. On the boundary of this region at $\phi_2 = -1$, the modulus is $r = 1$ and the forecast function is sinusoidal with no damping; in this case, $\phi_1 = 2\cos(\omega)$. So, for $|\phi_1| < 2$, the model $y_t = \phi_1 y_{t-1} - y_{t-2} + \epsilon_t$ is the one of a sinusoid with randomly varying amplitude and phase; with a small or zero innovation variance v the sinusoidal form sustains, representing essentially a fixed sine wave of constant amplitude and phase. It is easily seen that the difference equation $y_t = 2\cos(\omega)y_{t-1} - y_{t-2}$ defines, for given initial values, a sine wave of period $2\pi/\omega$.

2.5 Autocorrelation Structure of an AR(p). The autocorrelation structure of an AR(p) is given in terms of the solution of a homogeneous difference equation

$$\rho(k) - \phi_1\rho(k-1) - \ldots - \phi_p\rho(k-p) = 0, \quad k \geq p. \tag{2.8}$$

In general, if $\alpha_1, \ldots, \alpha_r$ denote the reciprocal roots of the characteristic polynomial $\Phi(u)$, where each root has multiplicity $m_1, \ldots, m_r$ and $\sum_{i=1}^r m_i = p$, then, the general solution to (2.8) is

$$\rho(k) = \alpha_1^k p_1(k) + \alpha_2^k p_2(k) + \ldots + \alpha_r^k p_r(k), \quad k \geq p, \tag{2.9}$$

where $p_j(k)$ is a polynomial of degree $m_j - 1$.

For instance, in the AR(2) case we have the following scenarios:

• The characteristic polynomial has two different real roots, each one with multiplicity $m_1 = m_2 = 1$. Then, the autocorrelation function has the form

$$\rho(k) = a\alpha_1^k + b\alpha_2^k, \quad k \geq 2,$$

where a and b are constants and α_1, α_2 are the reciprocal roots. Under stationarity this autocorrelation function decays exponentially as k goes to infinity and, as we saw before, this behaviour is shared by the forecast function. The constants a and b are determined by specifying two initial conditions such as $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1 - \phi_2)$.

• The characteristic polynomial has one real root with multiplicity $m_1 = 2$ and so, the autocorrelation function is given by

$$\rho(k) = (a + bk)\alpha_1^k, \quad k \geq 2,$$

where a and b are constants and α_1 is the reciprocal root. Under stationarity this autocorrelation function also decays exponentially as k goes to infinity.

• The characteristic polynomial has two complex conjugate roots. In this case the reciprocal roots can be written as $\alpha_1 = r\exp(i\omega)$ and $\alpha_2 = r\exp(-i\omega)$ and so, the autocorrelation function is

$$\rho(k) = ar^k \cos(k\omega + b) \quad , k \geq 2,$$

where a and b are constants. Under stationarity the autocorrelation and forecast functions behave as an exponentially damped cosine.

2.6 The Partial Autocorrelation Function. The autocorrelation and forecast functions summarise important features of autoregressive processes. We now introduce another function that will provide additional information about autoregressions: the partial autocorrelation function or PACF. We start by defining the general form of the PACF and we then see that the partial autocorrelation coefficients of a stationary AR(p) process are zero after lag p. This fact has important consequences in estimating the order of an autoregression, at least informally. In practice, it is possible to decide if an autoregression may be a suitable model for a given time series by looking at the estimated PACF plot. If the series was originally generated by an AR(p) model then its estimated partial autocorrelation coefficients should not be significant after the p-th lag.

The partial autocorrelation function or PACF of a process is defined in terms of the partial autocorrelation coefficients at lag k, denoted by $\phi(k,k)$. The PACF coefficient at lag k is a function of the so called best linear predictor of y_k given $y_{k-1}, \dots, y_1$. Specifically, this best linear predictor, denoted by y_k^{k-1}, has the form $y_k^{k-1} = \beta_1 y_{k-1} + \ldots + \beta_{k-1} y_1$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{k-1})'$ is chosen to minimise the mean square linear prediction error, $E(y_k - y_k^{k-1})^2$. If y_0^{k-1} is the minimum mean square linear predictor of y_0 based on $y_1, \dots, y_{k-1}$ and the process is stationary, then it can be shown that y_0^{k-1} is given by $y_0^{k-1} = \beta_1 y_1 + \ldots + \beta_{k-1} y_{k-1}$. The PACF is then defined in terms of the partial correlation coefficients $\phi(k,k)$, for $k = 1, 2, \ldots$, given by

$$\phi(k,k) = \begin{cases} \rho(y_1, y_0) = \rho(1) & k = 1 \\ \rho(y_k - y_k^{k-1}, y_0 - y_0^{k-1}) & k > 1, \end{cases} \tag{2.10}$$

where $\rho(y_i, y_j)$ denotes the correlation between y_i and y_j.

If $\{y_t\}$ follows an AR(p) it is possible to show that $\phi(k,k) = 0$ for all $k > p$ (for a proof see for example Shumway and Stoffer, 2000, Chapter 2). Using some properties of the best linear predictors it is also possible to show that the autocorrelation coefficients satisfy the following equation,

$$\Gamma_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n, \tag{2.11}$$

where Γ_n is an $n \times n$ matrix whose elements are $\{\gamma(j-k)\}_{j,k=1}^n$, and $\boldsymbol{\phi}_n$, $\boldsymbol{\gamma}_n$ are n-dimensional vectors given by $\boldsymbol{\phi}_n = (\phi(n,1), \ldots, \phi(n,n))'$ and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))'$. If Γ_n is non-singular then we can write $\boldsymbol{\phi}_n = \Gamma_n^{-1} \boldsymbol{\gamma}_n$. Alternatively, when dealing with stationary processes it is possible to find $\boldsymbol{\phi}_n$ using the Durbin-Levinson recursion (Levinson, 1947; Durbin, 1960) as follows. For $n = 0$, $\phi(0,0) = 0$. Then, for $n \geq 1$,

$$\phi(n,n) = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi(n-1,k)\rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi(n-1,k)\rho(k)},$$

with

$$\phi(n,k) = \phi(n-1,k) - \phi(n,n)\phi(n-1,n-k),$$

for $n \geq 2$ and $k = 1, \ldots, n-1$.

The sample PACF can be obtained by substituting the autocovariances in (2.11), or the autocorrelations in the Durbin-Levinson recursion by the sample autocovariances and the sample autocorrelations $\hat{\gamma}(\cdot)$ and $\hat{\rho}(\cdot)$. The sample PACF coefficients are denoted by $\hat{\phi}(k,k)$.

Forecasting

2.7 In traditional time series analysis, the one-step-ahead prediction of y_{t+1}, i.e., the forecast of y_{t+1} given $y_{1:t}$ is given by

$$y_{t+1}^t = \phi(t,1)y_t + \phi(t,2)y_{t-1} + \ldots + \phi(t,t)y_1, \tag{2.12}$$

with $\boldsymbol{\phi}_t = (\phi(t,1),\ldots,\phi(t,t))'$ the solution of (2.11) at $n = t$. The mean square error of the one-step-ahead prediction is given by

$$MSE_{t+1}^t = E(y_{t+1} - y_{t+1}^t)^2 = \gamma(0) - \boldsymbol{\gamma}_t'\Gamma_t^{-1}\boldsymbol{\gamma}_t, \tag{2.13}$$

or, using the Durbin-Levinson recursion this can be recursively computed as,

$$MSE_{t+1}^t = MSE_t^{t-1}(1 - \phi(t,t)^2),$$

with $MSE_1^0 = \gamma(0)$.

2.8 Similarly, the k-step ahead prediction of y_{t+k} based on $y_{1:t}$ is given by

$$y_{t+k}^t = \phi^{(k)}(t,1)y_t + \ldots + \phi^{(k)}(t,t)y_1, \tag{2.14}$$

with $\boldsymbol{\phi}_t^{(k)} = (\phi^{(k)}(t,1),\ldots,\phi^{(k)}(t,t))'$ the solution of $\Gamma_t\boldsymbol{\phi}_t^{(k)} = \boldsymbol{\gamma}_t^{(k)}$, where $\boldsymbol{\gamma}_t^{(k)} = (\gamma(k),\gamma(k+1),\ldots,\gamma(t+k-1))'$. The mean square error associated with the k-step-ahead prediction is given by

$$MSE_{t+k}^t = E(y_{t+k} - y_{t+k}^t)^2 = \gamma(0) - \boldsymbol{\gamma}_t^{'(k)}\Gamma_t^{-1}\boldsymbol{\gamma}_t^{(k)}. \tag{2.15}$$

It is also possible to compute the forecasts and the associated mean square errors using the innovations algorithm proposed by Brockwell and Davis (1991) as follows.

The one-step-ahead predictor and its associated mean squared error can be computed iteratively via

$$\begin{aligned} y_{t+1}^t &= \sum_{j=1}^{t} b_{t,j}(y_{t+1-j} - y_{t-j+1}^{t-j}), & (2.16)\\ MSE_{t+1}^t &= \gamma(0) - \sum_{j=0}^{t-1} b_{t,t-j}^2 MSE_{j+1}^j, & (2.17)\end{aligned}$$

for $t = 1,2,\ldots$, where for $j = 0,1,\ldots,t-1$,

$$b_{t,t-j} = \frac{\gamma(t-j) - \sum_{l=0}^{j-1} b_{l,j-l}b_{t,t-l}MSE_{l+1}^l}{MSE_{j+1}^j}.$$

Similarly, the k-steps ahead prediction and the corresponding mean squared error are given by

$$y_{t+k}^t = \sum_{j=k}^{t+k-1} b_{t+k-1,j}(y_{t+k-j} - y_{t+k-j}^{t+k-j-1}), \tag{2.18}$$

$$MSE^t_{t+k} \quad = \quad \gamma(0) - \sum_{j=k}^{t+k-1} b^2_{t+k-1,j} MSE^t_{t+k-j}. \tag{2.19}$$

For AR(p) models with $t > p$, the previous equations provide the exact one-step-ahead and k-steps-ahead predictions. In particular, it is possible to see that is y_t follows a stationary AR(p) process, then

$$y^t_{t+1} = \phi_1 y_t + \phi_2 y_{t-1} + \ldots + \phi_p y_{t-p+1}. \tag{2.20}$$

So far we have written the forecasting equations assuming that the parameters are known. If the parameters are unknown and need to be estimated, which is usually the case in practice, then it is necessary to substitute the parameter values by the estimated values in the previous equations.

When a Bayesian analysis of the time series model is performed, the forecasts are obtained directly from the model equations. So, for instance, if we are dealing with an AR(p), the k-step-ahead predictions can be computed using either posterior estimates for the model parameters or samples from the posterior distributions of the parameters. This will be discussed in detail in the next section.

Estimation in AR Models

2.9 Yule-Walker and Maximum Likelihood. Writing a set of diference equations of the form (2.8), in which the autocorrelations are substituted by the estimated autocorrelations, together with the corresponding set of initial conditions leads to the Yule-Walker estimates $\hat{\boldsymbol{\phi}}$ and $\hat{v}$, such that

$$\hat{\mathbf{R}}_p \hat{\boldsymbol{\phi}} = \hat{\rho}_p, \quad \hat{v} = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}' \hat{\mathbf{R}}_p^{-1} \hat{\boldsymbol{\phi}}, \tag{2.21}$$

where $\mathbf{R}_p$ is a $p \times p$ matrix with elements $\hat{\rho}(k-j), j, k = 1, \ldots, p$. These estimators can also be computed via the Durbin-Levinson recursion (see Brockwell and Davis, 1991 for details). It is possible to show that in the case of stationary AR processes, the Yule-Walker estimators are such that $\sqrt{T}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \approx N(\mathbf{0}, v\Gamma_p^{-1})$ and that $\hat{v}$ is close to v when the sample size T is large. These results can be used to obtain confidence regions about $\hat{\boldsymbol{\phi}}$.

Maximum likelihood estimation in AR(p) models can be achieved by maximising the conditional likelihood given in (2.3). It is also possible to work with the unconditional likelihood. This will be discussed later when the ML estimation method for general ARMA models are described.

2.10 Basic Bayesian Inference for AR models. Return to the basic model (2.1) and the conditional sampling density (2.3), and suppose that the data $y_{(p+1):T}$ are observed. Now make the parameters $(\boldsymbol{\phi}, v)$ explicit in the notation, so that (2.3) is

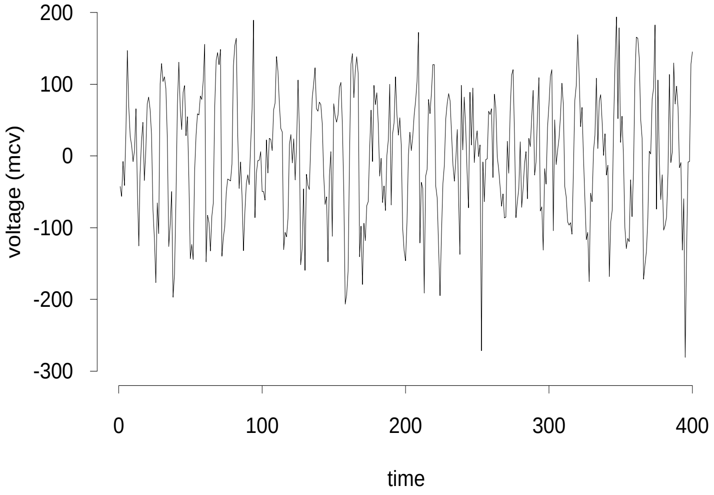

Fig. 2.1 A section of an EEG trace

formally $p(\mathbf{y}|\boldsymbol{\phi}, v, y_{1:p})$. Equation (2.3) defines the resulting likelihood function of $(\boldsymbol{\phi}, v)$. This is a conditional likelihood function, conditional on the assumed initial values $y_{1:p}$, so that resulting inferences, reference posterior inferences or otherwise, are also explicitly conditional on these initial values. More on dealing with this later. For now, we have a linear model $p(\mathbf{y}|\boldsymbol{\phi}, v, y_{1:p}) = N(\mathbf{y}|\mathbf{F}'\boldsymbol{\phi}, v\mathbf{I}_n)$ and we can apply standard theory. In particular, the reference posterior analysis described in Chapter 1 can be applied to obtain baseline inferences for $(\boldsymbol{\phi}, v)$.

Example 2.10.1 *EEG data analysis.*

Figure 2.1 displays recordings of an electro-encephalogram (EEG). The data displayed represent variations in scalp potentials in micro-volts during a seizure, the time intervals being just less than one fortieth of a second. The original data were sampled at 256 observations per second, and the 400 points in the Figure were obtained by selecting every sixth observation from a mid-seizure section.

The sample autocorrelations (not shown) have an apparent damped sinusoial form, indicative of the periodic behaviour evident from the data plot, with a period around 12-14 time units. The damping towards zero evident in the sample autocorrelations is consistent with stationary autoregressive components with complex roots. The sample partial autocorrelations are evidently strongly negative at lags bewteen 2 and 7 or 8, but appear to drop off thereafter, suggesting an autoregression of order $p = 7$ or $p = 8$.

An AR(8) model is explored as an initial model for these data; $p = 8$ and $y_{9:400}$ represent the final $n = 392$ observations, the first 8 being conditioned upon for initial values. The posterior multivariate Student-T distribution has 384 degrees of freedom and so, it is practically indistinguishable from a normal; it has mean

$$\hat{\boldsymbol{\phi}} = (0.27, 0.07, -0.13, -0.15, -0.11, -0.15, -0.23, -0.14)'$$

and approximately common standard deviations at 0.05. This illustrates quite typical variation and, to some degree, decay of coefficients with increasing lag. The innovations standard deviation has posterior estimate $s = 61.52$.

We fix $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$ to explore the model based on this point estimate of the parameter vector. The corresponding autoregressive equation $\Phi(u) = 0$ has four pairs of complex conjugate roots; the corresponding moduli and wavelength pairs (r_j, λ_j) are (in order of decreasing modulus)

$$(0.97, 12.73); \quad (0.81, 5.10); \quad (0.72, 2.99); \quad (0.66, 2.23).$$

The first term here represents the apparent cyclical pattern of wavelength around $12 - 13$ time units, and has a damping factor close to unity, indicating a rather persistent waveform; the half-life is about $k = 23$, i.e. 0.97^k decays to about 0.5 at $k = 23$, so that, with zero future innovations, the amplitude of this waveform is expected to decay to half a starting level in about two full cycles. By comparison, the three other, higher frequency components have much faster decay rates. The pattern here is quite typical of quasi-cyclical series. The high frequency terms, close to the Nyquist frequency limit, represent terms capturing very short run oscillations in the data of very low magnitude, essentially tailoring the model to low level noise features in the data rather than representing meaningful cyclical components in the series.

At time $T = 400$, or $t = n = 392$, the current state vector $\mathbf{x}_t$ together with the estimated parameter $\hat{\boldsymbol{\phi}}$ implies a forecast function of the form given in (2.7) in which the four component, damped sinusoids have relative amplitudes $2a_{tj}$ of approximately 157.0, 6.9, 18.0 and 7.0. So the first component of wavelength around 12.73 is quite dominant at this time epoch (as it is over the full span of the data), both in terms of the initial amplitude and in terms of a much lower decay rate. Thus the description of the series as close to a time-varying sine wave is reinforced.

Figure 2.2 displays the data and the forecast function from the end of the series over the next $k = 200$ time epochs based on the estimated value $\hat{\boldsymbol{\phi}}$. Figure 2.3 represents more useful extrapolation, displaying a single 'sampled future' based on estimated parameter values. This is generated simply by successively simulating future values $y_{T+k} = \sum_{j=1}^{p} \hat{\phi}_j y_{T+k-j} + \epsilon_{T+k}$ over $k = 1, 2, \ldots,$ etc., where the ϵ_{T+k} are drawn from $N(\cdot|0, s^2)$, and substituting sampled values as regressors for the future. This gives some flavour of likely development and the form is apparently similar to that of the historical data, suggesting a reasonable model description. These forecasts do not account for uncertainties about the estimated parameters $(\hat{\boldsymbol{\phi}}, s^2)$ so, they do not represent formal predictive distributions though are quite close approximations. This point is explored further below. Further insight into the nature of the likely

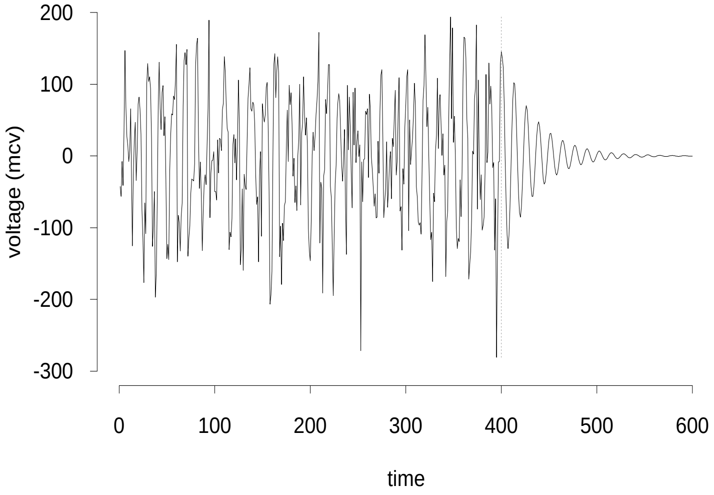

Fig. 2.2 EEG trace and forecast function from end of series

development, and also of aspects of model fit, are often gleaned by repeating this exercise, generating and comparing small sets of possible futures.

2.11 Simulation of Posterior Distributions. Inferences for other functions of model parameters and formal forecast distributions may be explored via simulation. Suppose interest lies in more formal inference about, for example, the period λ_1 of the dominant cyclical component in the above analysis of the EEG series, and other features of the structure of the roots of the AR polynomial. Though the posterior for $(\boldsymbol{\phi}, v)$ is analytically manageable, that for the α_i is not; posterior simulation may be used to explore these analytically intractable distributions. Similarly, sampled futures incorporating posterior uncertainties about $(\boldsymbol{\phi}, v)$ may be easily computed.

Example 2.11.1 *EEG data analysis (continued).*

A total number of 5,000 draws were made from the full normal/inverse-gamma posterior distribution for $(\boldsymbol{\phi}, v)$. For each such draw, a sampled future $y_{T+1}, \dots, y_{T+k}$, for any horizon k, was sampled as before, but now based on the simulated values $(\boldsymbol{\phi}, v)$ at each sample, rather than the estimates $(\hat{\boldsymbol{\phi}}, s^2)$. This delivers a sample of size 5,000 from the full joint posterior predictive distribution for $(y_{T+1}, \dots, y_{T+k})$. Averaging values across samples provides a Monte Carlo approximation to the forecast function. Exploring sampled futures provides Figures like 2.4, where additional

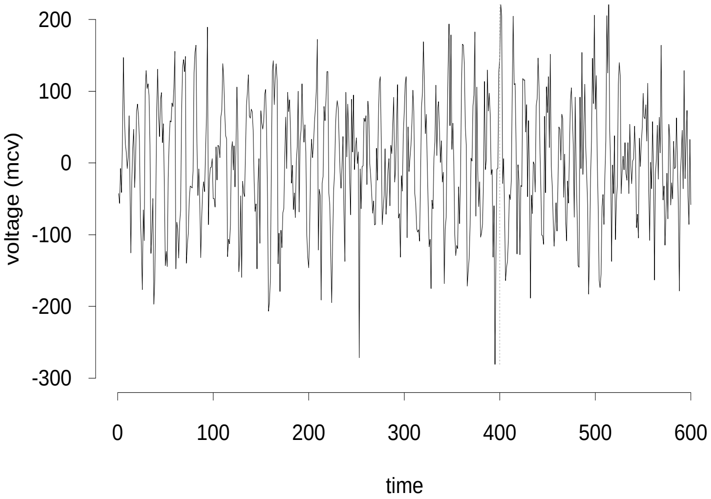

Fig. 2.3 EEG trace and sampled future conditional on parameter estimates $(\hat{\phi}, s^2)$

parameter uncertainties are incorporated. In this analysis, the additional uncertainties are small and have slight effects; other applications may be different.

Turn now to inference on the AR polynomial roots $\boldsymbol{\alpha}$. Each posterior draw $(\boldsymbol{\phi}, v)$ delivers a corresponding root vector $\boldsymbol{\alpha}$ which represents a random sample from the full posterior $p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}_p)$. Various features of this posterior sample for $\boldsymbol{\alpha}$ may be summarised. Note first the inherent identification issue, that the roots are unidentifiable as the AR model is unchanged under permutations of the subscripts on the α_i. One way around this difficulty is to consider inference on roots ordered by modulus or frequency (note the case of real roots formally corresponds to zero frequency). For example, the dominant component of the EEG model has been identified as that corresponding to the complex conjugate roots with the largest period, around $12 - 13$ time units. Ordering the complex values of each sampled set of roots leads to those with the largest period representing a sample from the posterior distribution for the period of the dominant component, and similarly for the samples of the corresponding modulus. The left and right panels of Figure 2.5 display the corresponding histograms in this analysis.

Note that no mention of stationarity has been made in this analysis. The reference posterior for $\boldsymbol{\phi}$, a multivariate Student-T distribution, is unconstrained and does not theoretically respect a constraint such as stationarity. In some applications, it may be physically meaningful and desirable to impose such an assumption and the analysis

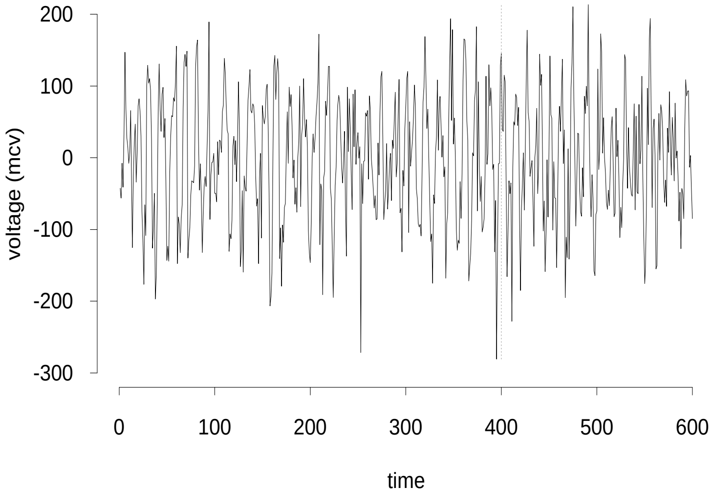

Fig. 2.4 EEG trace and sampled future from full posterior predictive distribution

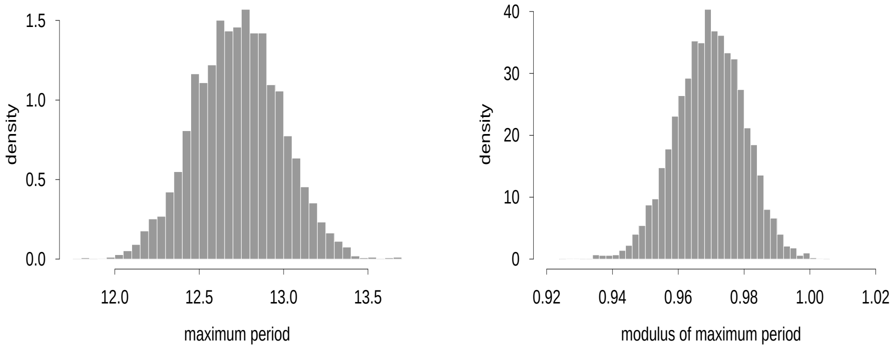

Fig. 2.5 Posterior for maximum period of sinusoidal components of the EEG series (left panel) and posterior for modulus of the damped sinusoidal component of maximum period in the EEG analysis (right panel)

should then be modified; theoretically, the prior for $(\boldsymbol{\phi}, v)$ should be defined as zero outside the stationarity region, whatever the form inside. The simplest approach is to proceed as in the unconstrained analysis, but to simply reject sampled $(\boldsymbol{\phi}, v)$ values if the $\boldsymbol{\phi}$ vector lies outside the stationarity region, a condition that is trivially checked by evaluating the roots of the implied AR polynomial. In cases where the data/model match really supports a stationary series, the rejection rate will be low and this provides a reasonable and efficient approximation to the analysis imposing the stationarity constraint through the prior. In other cases, evidence of non-stationary features may lead to higher rejection rates and an inefficient analysis; other methods are then needed. Some references below indicate work along these lines. Of course, an over-riding consideration is the suitability of a strict stationarity assumption to begin with; if the series, conditional on the appropriateness of the assumed model, is really consistent with stationarity, this should be evidenced automatically in the posterior for the AR parameters, whose mass should be concentrated on values consistent with stationarity. This is, in fact, true in the unconstrained EEG data analysis. Here the estimated AR polynomial root structure (at the reference posterior mean $\hat{\boldsymbol{\phi}}$) has all reciprocal roots with moduli less than unity, suggesting stationarity. In addition, the 5,000 samples from the posterior can be checked similarly; in fact, the actual sample drawn has no values with roots violating stationarity, indicating high posterior probability (probability one on the Monte Carlo posterior sample) on stationarity. In other applications, sampling the posterior may give some values outside the stationary region; whatever the values, this provides a Monte Carlo approach to evaluating the posterior probability of a stationary series (conditional on the assumed AR model form).

2.12 Order Assessment. Analysis may be repeated for different values of model order $p,$ it being useful and traditional to explore variations in inferences and predictions across a range of increasing values. Larger values of p are limited by the sample size, of course, and fitting high order models to only moderate data sets produces meaningless reference posterior inferences; large number of parameters, relative to sample size, can be entertained only with informed and proper prior distributions for those parameters, such as smoothness priors and others mentioned below. Otherwise, increasing p runs into the usual regression problems of over-fitting and collinearity.

Simply proceeding to sequentially increase p and exploring fitted residuals, changes in posterior parameter estimates and so forth, is a very valuable exercise. Various numerical summaries may be easily computed as adjunct to this, the two most widely known and used being the so-called Akaike information criterion, or AIC and the Bayesian information criterion or BIC (Akaike, 1969; Akaike, 1974; Schwarz, 1978). The AIC and BIC are now described together with a more formal, reference Bayesian measure of model fit. As we are comparing models with differing numbers of parameters, we do so based on a common sample size; thus, we fix a maximum order p^* and, when comparing models of various orders $p \leq p^*$, we do so in conditional reference analyses using the latter $n = T - p^*$ of the full T observations in the series.

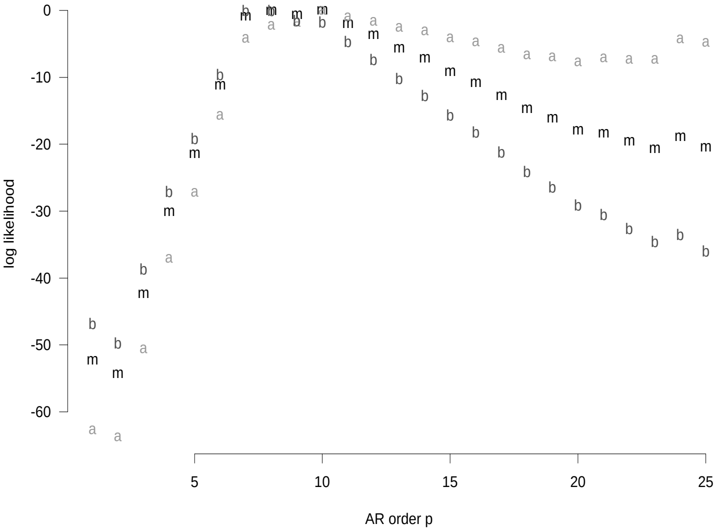

Fig. 2.6 Log-likelihood function for AR model order, computed from marginal data densities (labelled M), together with negated AIC criterion (labelled A) and BIC criterion (labelled B)

For a chosen model order p, explicit dependence on p is made by writing $\hat{\boldsymbol{\phi}}_p$ for the MLE of the AR parameters, and s_p^2 for the corresponding posterior estimate of innovations variance, i.e. the residual sum of squares divided by $n - p$. For our purposes, the AIC measure of model fit is taken as $2p + n\log(s_p^2)$, while the BIC is taken as $\log(n)p + n\log(s_p^2)$. Values of p leading to small AIC and BIC values are taken as indicative of relatively good model fits, within the class of AR models so explored (they may, of course, be poor models compared with other classes). Larger values of p will tend to give smaller variance estimates which decreases the second term in both expressions here, but this decrease is penalised for parameter dimension by the first term. BIC tends to choose simpler models than AIC. For the EEG series, negated AIC and BIC values, normalised to zero at the maximum, appear in Figure 2.6, based on $p^* = 25$. Also displayed there is a plot of the corresponding log-likelihood function for model order, computed as follows.

In a formal Bayesian analysis, the order p is viewed as an uncertain parameter and so any prior over p is updated via a likelihood function proportional to the marginal data density $p(y_{(p+1):T}|\mathbf{x}_p) = \int p(y_{(p+1):T}|\boldsymbol{\phi}, v, \mathbf{x}_p)dp(\boldsymbol{\phi}, v)$, where $p(\boldsymbol{\phi}, v)$ is the prior under the AR(p) model and it should be remembered that the dimension of $\boldsymbol{\phi}$ depends on p. Given proper priors $p(\boldsymbol{\phi}, v)$ across the interesting range of order values $p \leq p^*$, a direct numerical measure of relative fit is available through this collection of marginal densities which defines a valid likelihood function for the model order. To do this, however, requires a proper prior $p(\boldsymbol{\phi}, v)$ that naturally depends on the parameter dimension p and this dependency is important in determining the resulting likelihood function. The use of the traditional reference prior invalidates these calculations due to impropriety. Alternative approaches to constructing proper but, in some senses, uninformative priors may be pursued (see later references) but the critical need for priors to be consistent as model dimension varies remains. Nevertheless, under the assumedly common reference prior $p(\boldsymbol{\phi}, v) \propto 1/v$, the marginal data densities are defined up to a proportionality constant and follow directly from the reference Bayesian analysis of the linear regression model in Chapter 1. The maginal density values are closely related to the AIC and BIC values. The reference log-likelihood function so computed for the EEG series, with $p^* = 25$, appears in figure 2.6. Apparently, both this reference log-likelihood function and the usual AIC and BIC criteria suggest orders between 8 and 10 as preferable, hence the earlier analysis was based on $p = 8$.

Various alternatives based on different priors give similar results, at least in terms of identifying $p = 8$ or 9 as most appropriate. We note also that formal computation of, for example, predictive inferences involving averaging over p with respect to computed posterior probabilities on model order is possible, in contexts where proper priors for $(\boldsymbol{\phi}, v)$ are defined across models.

2.13 Analytic Considerations: Initial Values and Missing Data. The above analysis partitions the full data series $y_{1:T}$ into the p initial values $y_{1:p}$ and the final $n = T - p$ values $y_{(p+1):T}$ and is then conditional on $y_{1:p}$. Turn now to the unconditional analysis, in which the full likelihood function for $(\boldsymbol{\phi}, v)$ is

$$\begin{aligned} p(y_{1:T}|\boldsymbol{\phi}, v) &= p(y_{(p+1):T}|\boldsymbol{\phi}, v, y_{1:p})p(y_{1:p}|\boldsymbol{\phi}, v) \\ &= p(\mathbf{y}|\boldsymbol{\phi}, v, \mathbf{x}_p)p(\mathbf{x}_p|\boldsymbol{\phi}, v). \end{aligned} \tag{2.22}$$

The conditional analysis simply ignores the second component in (2.23). Apparently, whether or not this is justifiable or sensible depends on context, as follows.

In some applications, it is appropriate to assume some form of distribution for the initial values $\mathbf{x}_p$ that does not, in fact, depend on $(\boldsymbol{\phi}, v)$ at all. For example, it is perfectly reasonable to specify a model in which, say, the distribution $N(\mathbf{x}_p|0, \mathbf{A})$ is assumed, for some specified variance matrix **A**. In such cases, (2.23) reduces to the first component alone, and the conditional analysis is exact.

Otherwise, when $p(\mathbf{x}_p|\boldsymbol{\phi}, v)$ actually depends on $(\boldsymbol{\phi}, v)$, there will be a contribution to the likelihood from the initial values, and the conditional analysis is only approximate. Note, however, that, as the series length T increases, the first term of the likelihood, based on $n = T - p$ observations, becomes more and more dominant; the effect of the initial values in the second likelihood factor is fixed based on these values, and does not change with n. On a log-likelihood scale, the first factor behaves in expectation as $o(n)$, and so the conditional and unconditional analyses are asymptotically the same. In real problems with finite n, but in which p is usually low compared to n, experience indicates that the agreement is typically close even with rather moderate sample sizes. It is therefore common practice, and completely justifiable in applications with reasonable data sample sizes, to adopt the conditional analysis.

The situation has been much studied under a stationarity assumption, and a variation of the reference Bayesian analysis is explored here. Under stationarity, any subset of the data will have a marginal multivariate normal distribution, with zero mean and a variance matrix whose elements are determined by the model parameters. In particular, the initial values follow $N(\mathbf{x}_p|0, v\mathbf{A}(\boldsymbol{\phi}))$ where the $p \times p$ matrix $\mathbf{A}(\boldsymbol{\phi})$ depends (only) on $\boldsymbol{\phi}$ through the defining equations for autocorrelations in AR models. So (2.23), as a function of $(\boldsymbol{\phi}, v)$, is

$$p(y_{1:T}|\boldsymbol{\phi}, v) \propto v^{-T/2}|\mathbf{A}(\boldsymbol{\phi})|^{-1/2}\text{exp}(-Q(y_{1:T}, \boldsymbol{\phi})/2v), \tag{2.23}$$

where $Q(y_{1:T}, \boldsymbol{\phi}) = \sum_{t=p+1}^{T}(y_t - \mathbf{f}_t'\boldsymbol{\phi})^2 + \mathbf{x}_p'\mathbf{A}(\boldsymbol{\phi})^{-1}\mathbf{x}_p$. As developed in Box *et al.* (1994, Chapter 7), this reduces to a quadratic form $Q(y_{1:T}, \boldsymbol{\phi}) = a - 2\mathbf{b}'\boldsymbol{\phi} + \boldsymbol{\phi}'\mathbf{C}\boldsymbol{\phi}$, where the quantities a, **b**, **C** are easily calculable, as follows. Define the the symmetric $(p+1) \times (p+1)$ matrix $\mathbf{D} = \{D_{ij}\}$ by elements $D_{ij} = \sum_{r=0}^{T+1-j-i} y_{i+r}y_{j+r}$; then **D** is partitioned as

$$\mathbf{D} = \begin{pmatrix} a & -\mathbf{b}' \\ -\mathbf{b} & \mathbf{C} \end{pmatrix}.$$

One immediate consequence of this is that, if we ignore the determinant factor $|A(\boldsymbol{\phi})|$, the likelihood function is of standard linear model form. The traditional reference prior $p(\boldsymbol{\phi}, v) \propto v^{-1}$ induces a normal/inverse gamma posterior, for example; other normal/inverse gamma priors might be used similarly. In the reference case, full details of the posterior analysis can be worked through by the reader. The posterior mode for $\boldsymbol{\phi}$ is now $\hat{\boldsymbol{\phi}}^* = \mathbf{C}^{-1}\mathbf{b}$. For the EEG series, the calculations lead to

$$\hat{\boldsymbol{\phi}}^* = (0.273, 0.064, -0.128, -0.149, -0.109, -0.149, -0.229, -0.138)'$$

to three decimal places. The approximate value based on the conditional analysis is

$$\hat{\boldsymbol{\phi}} = (0.272, 0.068, -0.130, -0.148, -0.108, -0.148, -0.226, -0.136)',$$

earlier quoted to only two decimal places in light of the corresponding posterior standard deviations around 0.05 in each case. The differences, in the third decimal place in each case, are negligible, entirely so in the context of spread of the posterior. Here we are in the (common) context where T is large enough compared to p, and so, the effect of the initial values in (2.23) is really negligible. Repeating the analysis with just the first $T = 100$ EEG observations, the elements of $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\phi}}^*$ differ by only about 0.01, whereas the associated posterior standard errors are around 0.1; the effects become more marked with smaller sample sizes, though are still well within the limits of posterior standard deviations with much smaller values of T. In other applications the effects may be more substantial.

Ignoring the determinant factor can be justified by the same, asymptotic reasoning. Another justification is based on the use of an alternative reference prior: that based on Jeffrey's rule. In this case, as shown in Box *et al.*(1994), the Jeffrey's prior is approximately $p(\boldsymbol{\phi}, v) \propto |\mathbf{A}(\boldsymbol{\phi})|^{1/2}v^{-1/2}$; this results in cancellation of the determinant factor so the above analysis is exact.

Otherwise, under different prior distributions, the exact posterior involves the factor $|\mathbf{A}(\boldsymbol{\phi})|$, a complicated polynomial function of $\boldsymbol{\phi}$. However, $|\mathbf{A}(\boldsymbol{\phi})|$ can be evaluated at any specified $\boldsymbol{\phi}$ value, and numerical methods can be used to analyse the complete posterior. Numerical evaluation of the exact MLE is now a standard feature in some software packages, for example. Bayesian analysis using Monte Carlo methods is also easy to implement in this framework.

2.13.1 Initial Values Revisited via Simulation. Introduce the truly uncertain initial values $\mathbf{x}_0 = (y_0, y_{-1}, \ldots, y_{-(p-1)})'$. Adjust the earlier conditional analysis to be based on all T observations $y_{1:T}$ and now to be conditional on these (imaginary) initial values $\mathbf{x}_0$. Then, whatever the prior, we have the posterior $p(\boldsymbol{\phi}, v|y_{1:T}, \mathbf{x}_0)$. In the reference analysis, we have a normal/inverse gamma posterior now based on all T observations rather than just the last $n = T - p$, with obvious modifications. Note that this posterior can be simulated, to deliver draws for $(\boldsymbol{\phi}, v)$ conditional on any specific initial vector $\mathbf{x}_0$. This can be embedded in an iterative simulation of the full joint posterior $p(\boldsymbol{\phi}, v, \mathbf{x}_0|y_{1:T})$ if, in addition, we can sample $\mathbf{x}_0$ vectors from the conditional posterior $p(\mathbf{x}_0|\boldsymbol{\phi}, v, y_{1:T})$ for any specified $(\boldsymbol{\phi}, v)$ parameters.

In the case of a stationary series, stationarity and the linear model form imply reversibility with respect to time; that is, the basic AR model holds backwards, as well as forwards, in time. Hence, conditional on $(\boldsymbol{\phi}, v)$ and future series values $y_{t+1}, y_{t+2}, \ldots,$ the current value y_t follows the distribution $N(y_t|\mathbf{g}_t'\boldsymbol{\phi}, v)$ where $\mathbf{g}_t = rev(\mathbf{x}_{t+p}) = (y_{t+1}, \ldots, y_{t+p})'$; here the operator $rev(\cdot)$ simply reverses the elements of its vector argument. Applying this to the initial values at $t = 0, -1, \ldots,$ leads to

$$p(\mathbf{x}_0|\boldsymbol{\phi}, v, y_{1:T}) = \prod_{t=0}^{-(p-1)} N(y_t|\mathbf{g}_t'\boldsymbol{\phi}, v).$$

Hence, given $(\boldsymbol{\phi}, v)$, a vector $\mathbf{x}_0$ is simulated by sequentially sampling the individual component normal distributions in this product: first draw y_0 given the known data $\mathbf{x}_p$ and the parameters; then substitute the sampled value y_0 as the first element of the otherwise know data vector $\mathbf{x}_{p-1}$, and draw y_1; continue this way down to $y_{-(p-1)}$. This is technically similar to the process of simulating a future of the series illustrated earlier; now we are simulating the past.

In the modern, computational world of applied statistics, this approach is both trivially implemented and practically satisfying as it provides, modulo the Monte Carlo simulation, exact analysis. Further, extensions of basic AR models to incorporate various practically relevant additional features, naturally lead to Markov chain simulations as natural, and typically necessary, approaches to analysis, so that dealing with the starting value issue in this framework makes good sense.

It should also be clear that the same principle applies to problems of missing data. For any set of indices t such that the values y_t are missing (at random, that is, the reasons for missing data do not have a bearing on the values of the model parameters), then iterative simulation analysis can be extended and modified to incorporate the missing values as additional uncertain quantities to be estimated. Further details can be worked out in the framework here, as with the missing initial values above, and details are left to the reader. We revisit missing values later in the context of general state space models, and state space representations of autoregressions in particular.

Further Issues on Bayesian Inference for AR Models

2.14 Sensitivity to the choice of prior distributions. Additional analyses explore inferences based on longer order AR models with various proper priors for the AR coefficients. One interest is in exploring the sensitivity of the earlier, reference inferences under ranges of proper and perhaps more plausible prior assumptions. In each case the model is based on (a maximum lag) $p = 25,$ assuming that higher order models would have negligible additional coefficients and that, in any case, the higher order coefficients in the model are likely to decay. The two priors for $\boldsymbol{\phi}$ are centred around zero and so induce shrinkage of the posteriors towards the prior means of zero for all parameters. In each case, the first p values of $y_{1:T}$ are fixed to provide conditional analyses comparable to that earlier discussed at length.

2.14.1 Analysis based on normal priors. A first analysis assumes a traditional prior with the coefficients i.i.d. normal; the joint prior is $N(\boldsymbol{\phi}|0, w\mathbf{I}_p),$ for some scalar variance w and so, it induces shrinkage of the posterior towards the prior mean of zero for all parameters. The hyperparameter w will be estimated together with the primary parameters $(\boldsymbol{\phi}, v)$ via Gibbs sampling to simulate the full posterior for $(\boldsymbol{\phi}, v, w).$ We assume prior independence of v and w and adopt uniform priors, so $p(v)$ and $p(w)$ are constant over a wide range; in each analysis we assume this range is large enough so that the corresponding conditional posteriors are effectively proportional to the appropriate conditional likelihood functions, i.e., the truncation implied under the prior has little effect. Posterior simulations draw sequentially from the following three conditional posteriors, easily deduced from the model form and general normal linear model theory reviewed in Chapter 1.

- Given $(v, w),$ posterior for $\boldsymbol{\phi}$ is $N(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}}, \mathbf{B})$ where $\mathbf{B}^{-1} = w^{-1}\mathbf{I}_p + v^{-1}\mathbf{F}\mathbf{F}'$ and $\hat{\boldsymbol{\phi}} = \mathbf{B}v^{-1}\mathbf{F}\mathbf{y}.$
- Given $(\boldsymbol{\phi}, w),$ posterior for v^{-1} is $Ga(v^{-1}|n/2, \mathbf{e}'\mathbf{e}/2)$ based on residual vector $\mathbf{e} = \mathbf{y} - \mathbf{F}'\boldsymbol{\phi}.$
- Given $(\boldsymbol{\phi}, v),$ posterior for w^{-1} is $Ga(w^{-1}|p/2, \boldsymbol{\phi}'\boldsymbol{\phi}/2).$

For the EEG series, Figure 2.7 graphs the approximate posterior means of the ϕ_j's based on a Monte Carlo sample of size 5,000 from the simulation analysis so specified. This sample is saved following burn-in of 500 iterations. Also plotted

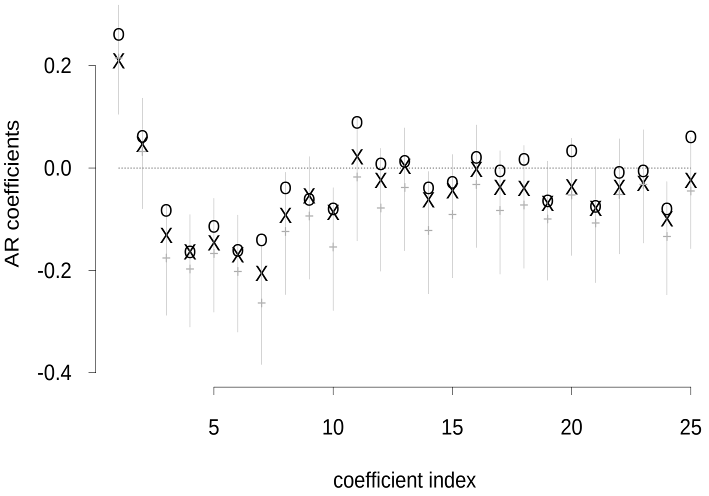

Fig. 2.7 Estimates of ϕ in EEG analyses. The vertical bars indicate approximate 95% posterior intervals for the ϕ_j from the reference analysis, centred about reference posterior means. The symbols X indicate approximate posterior means from the analysis based on independent normal priors. Symbols O indicate approximate posterior means from the analysis based on the two-component, normal mixture priors

are the reference posterior means with two posterior standard deviation intervals, for comparison. Some shrinkage of the coefficients is evident, though apparently not dramatic in extent, and the posterior means are not incomparable with the reference values, indicating some robustness to prior specification. Inferences and forecasts based on the normal prior will not differ substantially from those based on the reference prior. In this analysis, the posterior for the shrinkage parameter $\sqrt{w}$ is apparently unimodal, centred around 0.12 with mass predominantly concentrated in the range 0.08-0.16.

2.14.2 Discrete Normal Mixture Prior and Subset Models. A further analysis illustrates priors inducing differential shrinkage effects across the ϕ_j parameters; some of the ϕ_j may indeed be close to zero, others quite clearly distinct from zero, and a prior view that this may be the case can be embodied in standard modifications of the above analysis. One such approach uses independent priors conditional on individual scale factors, namely $N(\phi_j|0, w/\delta_j),$ where the weights δ_j are random quantities to be estimated. For example, a model in which only one or two of the ϕ_j are really significant is induced by weights δ_j close to unity for those parameters, the other weights being relatively large resulting in priors and posteriors concentrated around zero for the negligible weights. This links to the concept of subset autoregressions, in which only a few parameters at specific lags are really relevant, the others, at possibly intervening lags, being zero or close to zero. A class of priors for $\boldsymbol{\phi}$ that embody this kind of qualitative view provides for automatic inference on relevant subsets of non-negligible parameters and, effectively, addresses the variable selection question.

Probably the simplest approach extends the case of independent normal priors above, in which each $\delta_j = 1,$ to the case of independent priors that are two-component normals, namely

$$\pi N(\phi_j|0, w) + (1-\pi)N(\phi_j|0, w/L)$$

where π is a probability and L a specified precision factor. If $L >> 1,$ the second normal component is very concentrated around zero, so this mixture prior effectively states that each ϕ_j is close to zero, with probability $1-\pi,$ and is otherwise drawn from the earlier normal with variance $w.$

Assume L is specified. Introduce indicators u_j such that $u_j = 1$ or 0 according to whether ϕ_j is drawn from the first or the second of the normal mixture components. These u_j are latent variables that may be introduced to enable the simulation analysis. Write $u = (u_1, \ldots, u_p)$ and, for any set of values $u,$ write $\delta_j = u_j + (1-u_j)L,$ so that $\delta_j = 1$ or L; also, define the matrix $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_p).$ Further, write $k = \sum_{j=1}^{p} u_j$ for the number of coefficients drawn from the first normal component; k can be viewed as the number of non-neglible coefficients, the others being close to zero. Note that, given $\pi,$ k has a prior binomial distribution with success probability $\pi.$

For completeness and robustness, π is usually viewed as uncertain too; in the analysis below, π is assigned a beta prior, $Be(\pi|a, b),$ independently of the other random quantities in the model. This implies, among other things, a beta-binomial marginal prior for the number k of significant coefficients, namely

$$p(k) = \binom{n}{k} \frac{\beta(a,b)}{\beta(a+k, b+p-k)},$$

over $k = 0, \ldots, p$, where $\beta(\cdot, \cdot)$ is the beta function.

Under this model and prior specification, the various conditional posterior distributions to be used in Gibbs sampling of the full posterior for $(\boldsymbol{\phi}, v, w, u, \pi)$ are as follows.

- Given (v, w, u, π), posterior for $\boldsymbol{\phi}$ is $N(\boldsymbol{\phi}|\mathbf{b}, \mathbf{B})$ where $\mathbf{B}^{-1} = w^{-1}\Delta + v^{-1}\mathbf{F}\mathbf{F}'$ and $\mathbf{b} = \mathbf{B}v^{-1}\mathbf{F}\mathbf{y}$.
- Given $(\boldsymbol{\phi}, w, u, \pi)$, posterior for v^{-1} is $Ga(v^{-1}|n/2, \mathbf{e}'\mathbf{e}/2)$ based on residual vector $\mathbf{e} = \mathbf{y} - \mathbf{F}'\boldsymbol{\phi}$.
- Given $(\boldsymbol{\phi}, v, u, \pi)$, posterior for w^{-1} is $Ga(w^{-1}|p/2, q/2)$ with scale factor defined by $q = \sum_{j=1}^{p} \phi_j^2 \delta_j$.
- Given $(\boldsymbol{\phi}, v, w, \pi)$, the u_j are independent with conditional posterior probabilities $\pi_j = Pr(u_i = 0|\boldsymbol{\phi}, v, w, \pi)$ given, in odds form, by

$$\frac{\pi_j}{1-\pi_j} = \frac{\pi}{1-\pi} \exp(-(L-1)\phi_j^2/2w)/\sqrt{L}.$$

- Given $(\boldsymbol{\phi}, v, w, u)$, posterior for π is beta, namely $Be(\pi|a+k, b+p-k)$ where $k = \sum_{j=1}^{p} u_j$.

Iterative sampling of these conditional distributions provides samples of $\boldsymbol{\phi}$, v, w, u, and π for inference. The additional symbols in Figure 2.7 indicate the posterior means for the ϕ_j from such an analysis, again based on a simulation sample size of 5,000 from the full posterior; the analysis adopts $a = 1, b = 4$ and $L = 25$. We note little difference in posterior means relative to the earlier analyses, again indicating robustness to prior specifications as there is a good deal of data here.

The implied beta-binomial prior for k appears in Figure 2.8, indicating mild support for smaller values consistent with the view that, though there is much prior uncertainty, several or many of the AR coefficients are likely to be negligible. The posterior simulation analysis provides posterior samples of k, and the relative frequencies estimate the posterior distribution, as plotted in Figure 2.8. This indicates a shift to favouring values in the 5–15 ranges based on the data analysis under this specific prior structure; there is much uncertainty about k represented under this posterior, though the indication of a evidence for more than just a few coefficients is strong. Additional information is available in the full posterior sample; it carries, for instance, Monte Carlo estimates of the posterior probabilities that individual coefficients ϕ_j are drawn from the first or second mixture component, simply the approximate posterior means of the corresponding indicators u_j. This information can be used to assess subsets of significant coefficients, as adjunct to exploring posterior estimates and uncertainties about the coefficients, as in Figure 2.7.

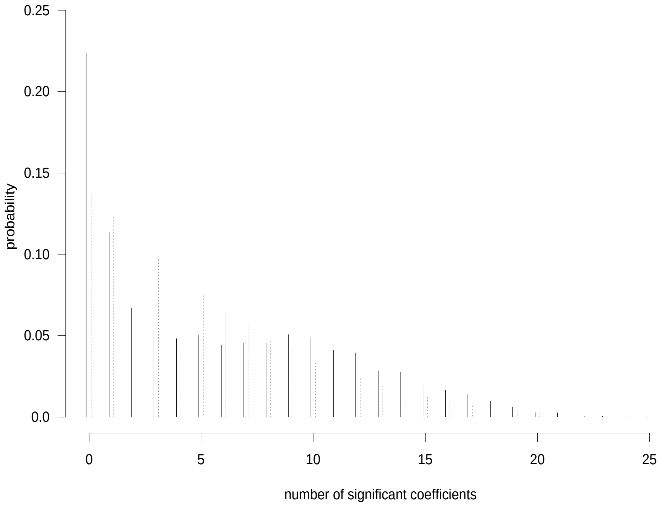

Fig. 2.8 Prior and approximate posterior distribution for the number of non-negligible AR coefficients, out of the total $p = 25$, in the EEG analysis under the two-component mixture prior

2.15 Alternative Prior Distributions.

2.15.1 Scale-mixtures and Smoothness Priors. Analyses based on alternative priors may be similarly explored; some examples are mentioned here, and may be explored by the reader. For instance, the second analysis is an example of a prior constructed via scale-mixtures of a basic normal prior for the individual coefficients. The mixing distribution in that case is discrete, placing mass of π at $\delta_j = 1$ and $\delta_j = 25$. Other mixing distributions are common in applied Bayesian work, a key example being the class of gamma distributions. For instance, take the weights δ_j to be independently drawn from a gamma distribution with shape and scale equal to $k/2$ for some $k > 0$; this implies that the resulting marginal prior for each ϕ_j is a Student-t distribution with k degrees of freedom, mode at zero and scale factor $\sqrt{w}$. This is, in some senses, a natural heavy-tailed alternative to the normal prior, assigning greater prior probabilities to ϕ_j values further from the prior location at zero. This can result in differential shrinkage, as in the case of the discrete normal mixture in the example.

A further class of priors incorporate the view that AR coefficients are unlikely to be large at higher lags, and ultimately decay towards zero. This kind of qualitative information may be important in contexts where p is large relative to expected sample sizes. This can be incorporated in the earlier normal prior framework, for example, by generalising to independent priors $N(\phi_j|0, w/\delta_j)$ where the weights are now fixed constants that concentrate the priors around zero for larger lags j; an example would be $\delta_j = j^2$. Note that this may be combined with additional, random weights to develop decaying effects within a normal mixture prior, and is trivially implemented.

Traditional smoothness priors operate on differences of parameters at successive lags, so that priors for $|\phi_{j+1} - \phi_j|$ are also centred around zero to induce a smooth form of behaviour of ϕ_j as a function of lag j, a traditional 'distributed lag' concept; a smooth form of decay of the effects of lagged values of the series is often naturally anticipated. This is again a useful concept in contexts where long order models are being used. One example of a smoothness prior is given by generalising the normal prior structure as follows. Take the normal margin $N(\phi_1|0, w/\delta_1)$ and, for $j > 1$, assume conditional priors $N(\phi_j|\phi_{j-1}, w/\delta_j)$; here the δ_j weights are assumed to increase with lag j to help induce smoothness at higher lags. This specification induces a multivariate normal prior (conditional on the δ_j and w), $p(\boldsymbol{\phi}) = p(\phi_1)\prod_{j=2}^{p} p(\phi_j|\phi_{j-1}) = N(\boldsymbol{\phi}|0, \mathbf{A}^{-1}w)$, where the precision matrix $\mathbf{A} = \mathbf{H}'\Delta\mathbf{H}$ is defined by $\Delta = \text{diag}(\delta_1, \ldots, \delta_p)$ and

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Again, the δ_j weights may be either specified or random, or a mix of the two. Posterior inferences follow easily using iterative simulation, via straightforward modifications of the analyses above.

2.15.2 Priors based on AR latent structure. Consider again the AR(p) model whose characteristic polynomial is given by $\Phi(u) = 1 - \phi_1 u - \ldots - \phi_p u^p$. The process is stationary if the reciprocal roots of this polynomial have moduli less than unity. Now, consider the case in which there is a maximum number of C pairs of complex valued reciprocal roots and a maximum number of R real valued reciprocal roots with $p = 2C + R$. The complex roots appear in pairs of complex conjugates, each pair having modulus r_j and wavelength λ_j —or equivalently, frequency $\omega_j = 2\pi/\lambda_j$ — for $j = 1, \ldots, C$. Each real reciprocal root has modulus r_j, for $j = C+1, \ldots, C+R$. Following Huerta and West (1999), the prior structure given below can be assumed on the real reciprocal roots

$$\begin{aligned} r_j \quad \sim \quad & \pi_{r,-1} I_{(-1)}(r_j) + \pi_{c,0} I_0(r_j) + \pi_{r,1} I_1(r_j) \\ & + (1 - \pi_{r,0} - \pi_{r,-1} - \pi_{r,1}) g_r(r_j), \end{aligned} \tag{2.24}$$

where $I(\cdot)$ denotes the indicator function, $g_r(\cdot)$ is a continuous density over $(-1, 1)$ and $\pi_{r,\cdot}$ are prior probabilities. The point masses at $r_j = \pm 1$ allow us to consider non-stationary unit roots. The point mass at $r_j = 0$ handles the uncertainty in the number of real roots, since this number may reduce below the prespecified maximum R. The default option for $g_r(\cdot)$ is the uniform $g_r(\cdot) = U(\cdot| - 1, 1)$, i.e., the reference prior for a component $\mathrm{AR}(1)$ coefficient r_j truncated to the stationary region. Similarly, for the complex reciprocal roots the following prior can be assumed

$$\begin{aligned} r_j &\sim \pi_{c,0} I_0(r_j) + \pi_{c,1} I_1(r_j) + (1 - \pi_{c,1} - \pi_{c,0}) g_c(r_j), \\ \lambda_j &\sim h(\lambda_j), \end{aligned} \tag{2.25}$$

with $g_c(\cdot)$ a continuous distribution on $0 < r_j < 1$ and $h(\lambda_j)$ a continuous distribution on $2 < \lambda_j < \lambda_u$, for $j = 1, \ldots, C$. The value of λ_u is fixed and by default it could be set to $n/2$. In addition, a so called "component reference prior" (Huerta and West, 1999) is induced by assuming a uniform prior on the implied $\mathrm{AR}(2)$ coefficients $2r_j \cos(2\pi/\lambda_j)$ and $-r_j^2$, but restricted to the finite support of λ_j for propriety. This is defined by $g_c(r_j) \propto r_j^2$, so that the marginal for r_j is $Be(\cdot|3, 1)$, and $h(\lambda_j) \propto \sin(2\pi/\lambda_j)/\lambda_j^2$ on $2 < \lambda_j < \lambda_u$. The probabilities $\pi_{c,0}$ and $\pi_{c,1}$ handle the uncertainty in the number of complex components and non-stationary unit roots, respectively. Finally, uniform Dirichlet distributions are the default choice for the probabilities $\pi_{r,\cdot}$ and $\pi_{c,\cdot}$, this is

$$Dir(\pi_{r,-1}, \pi_{r,0}, \pi_{r,1}|1, 1, 1), \quad Dir(\pi_{c,0}, \pi_{c,1}|1, 1),$$

and an inverse-Gamma prior is assumed for v, $IG(v|a, b)$.

A MCMC sampling scheme can be implemented to obtain samples from the posterior distribution of the model parameters

$$\boldsymbol{\theta} = \{(r_1, \lambda_1), \ldots, (r_C, \lambda_C), r_{C+1}, \ldots, r_{C+R}, \pi_{r,-1}, \pi_{r,0}, \pi_{r,1}, \pi_{c,0}, \pi_{c,1}, v, \mathbf{x}_0\},$$

with $\mathbf{x}_0 = (y_0, \ldots, y_{-(p-1)})'$, the p initial values. Specifically, if for any subset $\boldsymbol{\theta}^*$ of elements of $\boldsymbol{\theta}$, $\boldsymbol{\theta} \backslash \boldsymbol{\theta}^*$ denotes all the elements of $\boldsymbol{\theta}$ with the subset $\boldsymbol{\theta}^*$ removed, the MCMC algorithm can be summarised as follows.

• For each $j = C + 1, \ldots, C + R$, sample the real roots from the conditional marginal posterior $p(r_j|\boldsymbol{\theta} \backslash r_j, \mathbf{x}_0, y_{1:n})$. As detailed in Huerta and West (1999), the conditional likelihood function for r_j provides a normal kernel in r_j and so, obtaining draws for each r_j reduces to sampling from a mixture posterior with four components, which can be easily done.

• For each $j = 1, \ldots, C$, sample the complex roots from the conditional marginal posterior $p(r_j, \lambda_j|\boldsymbol{\theta} \backslash (r_j, \lambda_j), \mathbf{x}_0, y_{1:n})$. Sampling from this conditional posterior directly is difficult and so, a reversible jump Markov chain Monte Carlo step is necessary. The reversible jump MCMC (RJMCMC) method introduced in Green (1995), permits jumps between parameter subspaces of different dimensions at each iteration. The method consists on creating a random sweep Metropolis-Hastings algorithm adapted for changes in dimensionality. The RJMCMC algorithm is described in the Appendix.

- Sample $(\pi_{r,-1}, \pi_{r,0}, \pi_{r,1})$ and $(\pi_{c,0}, \pi_{c,1})$ from conditionally independent Dirichlet posteriors as detailed in Huerta and West (1999).
- Sample v from an Inverse-Gamma distribution.
- Sample $\mathbf{x}_0$. Huerta and West (1999) shows the time reversibility property for AR models with unit roots, and so, it is possible to sample the initial values $\mathbf{x}_0$ in similar way to the one described in section 2.13.1.

Example 2.15.1 *A RJMCMC for an AR(4) model with structured priors.*

We consider the analysis of 100 observations simulated from an AR(2) process with a single pair of complex roots with modulus $r = 0.9$ and wavelength $\lambda = 8$. We fit an AR(4) to these data using the structured priors previously described. We set $C = 2$ and $R = 0$ and so, two RJMCMC steps are needed to sample (r_1, λ_1) and (r_2, λ_2). Each RJMCMC step has a certain number of moves. For instance, if the chain is currently at $r_j = 0$, the following moves can be considered, each with probability 1/3

- Remain at the origin.
- Jump at new values of the form $(1, \omega_j^*)$.
- Jump at new values of the form (r_j^*, ω_j^*).

Details about the RJMCMC algorithm for the general AR(p) case are discussed in Huerta (1998). Free software is available to perform posterior inference for AR models with structured priors. The software is called `ARcomp` and can be downloaded from `www.isds.duke.edu/isds-info/software.html`. `ARcomp` was used to fit an AR(4) with structured priors to the simulated data. Figure 2.9 shows the posterior distribution for the model order p, and the posteriors for the number of complex pairs C and the number of real characteristic roots R. Note that in this example R was set to zero a priori and so, the posterior gives probability one to $R = 0$. From these graphs it is clear that the model is adequately capturing the AR(2) structure in the simulated data, as $Pr(p = 2|y_{1:n}) > 0.8$ and $Pr(C = 1|y_{1:n}) > 0.8$.

Figure 2.10 displays the posterior distribution of (r_1, λ_1) (bottom panels) and (r_2, λ_2) (top panels). We obtain $Pr(r_1 = 0|y_{1:n}) = 0.98$ and $Pr(r_2 = 0|y_{1:n}) = 0$, which are consistent with the fact that the data were simulated from an AR(2) process. In addition, the marginal posteriors for r_2 and λ_2 are concentrated around the true values $r = 0.9$ and $\lambda = 8$.

Example 2.15.2 *Analysis of the EEG data with structured priors.*

We now consider an analysis of the EEG data shown in Figure 2.1 using structured priors. In this example we set $C = R = 6$ and so, the maximum model order is $p_{\max} = 2 * 6 + 6 = 18$. Figure 2.11 shows the posterior distributions of p, C and R. This analysis gives highest posterior probability to a model with 4 pairs of characteristic roots and 3 real roots, or equivalently a model with $p = 11$. However, there is considerable uncertainty in the number of real and complex roots and so, models with $10 \leq p \leq 16$ get significant posterior probabilities. Figure 2.12 displays

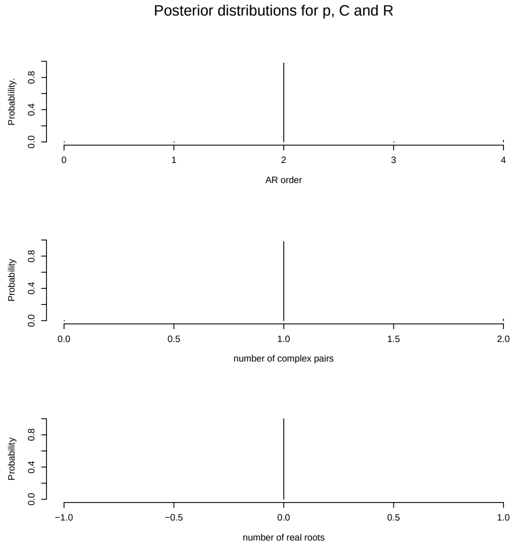

Fig. 2.9 Posterior distibutions of the model order, the number of complex pairs and the number of real components for the simulated data

the marginal posterior distributions of r_1 and λ_1, i.e., the marginals for the modulus and wavelength of the component with the highest modulus. Note that these pictures are consistent with the results obtained for the reference analysis of an AR(8) model presented previously.

Autoregressive, Moving Average (ARMA) Models

2.16 Structure of ARMA Models. Consider a time series y_t, for $t = 1, 2, \ldots,$ arising from the model

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{2.26}$$

with $\epsilon_t \sim N(0, v)$. Then, $\{y_t\}$ follows an autoregressive moving average model, or ARMA(p, q), where p and q are the orders of the autoregressive and moving average parts, respectively.

Example 2.16.1 *MA*(1) *process.*

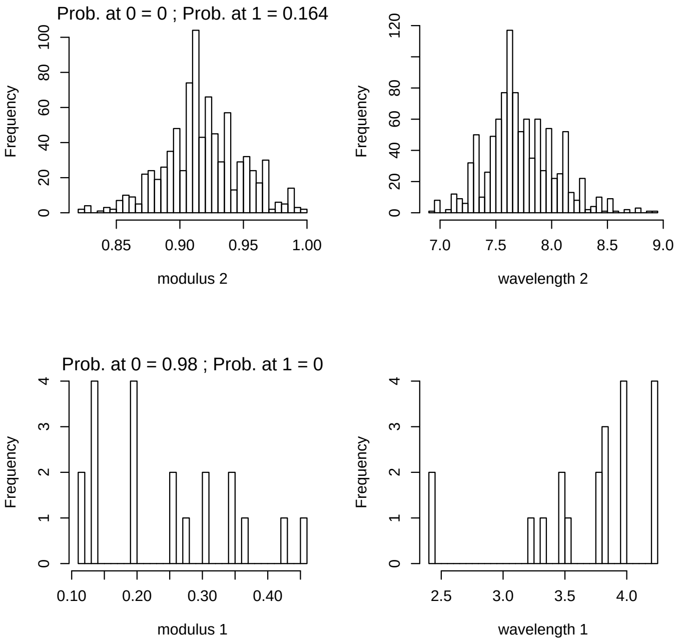

Fig. 2.10 Posterior distributions of (r_1, λ_1) and (r_2, λ_2) for the simulated data

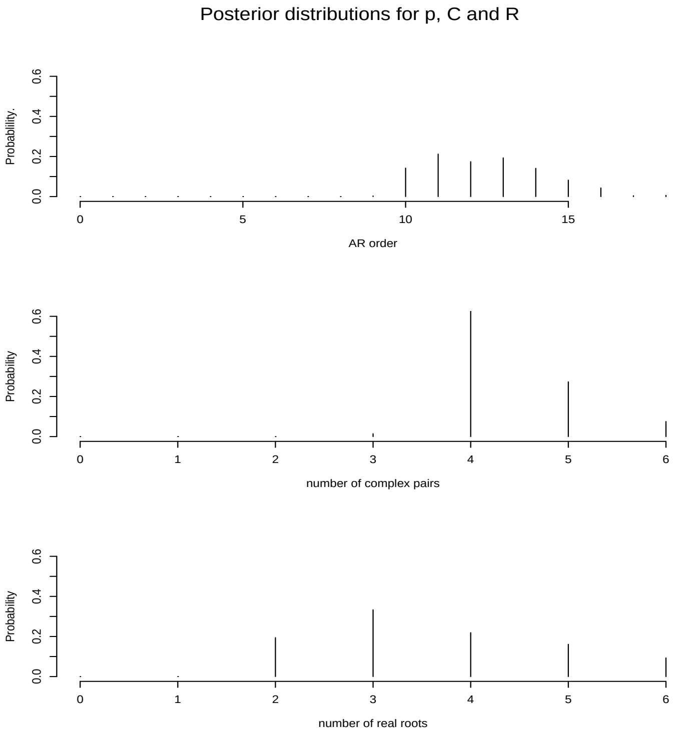

Fig. 2.11 Posterior distributions of the model order, C and R for the EEG data

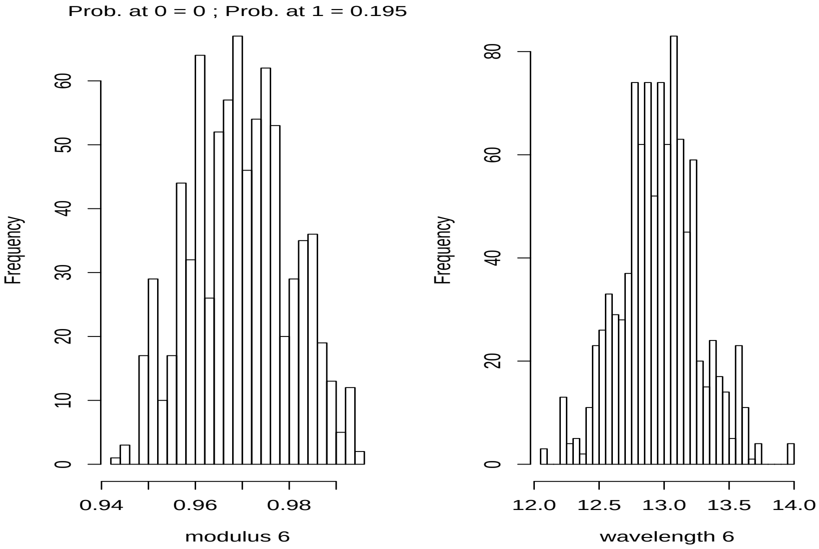

Fig. 2.12 Posterior distributions of (r_1, λ_1) for the EEG data

If y_t follows a MA(1) process, $y_t = \theta\epsilon_{t-1} + \epsilon_t$, the process is stationary for all the values of θ. In addition, it is easy to see that the autocorrelation function has the following form

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \frac{\theta}{(1+\theta^2)} & k = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Now, if we consider a MA(1) process with coefficient $\frac{1}{\theta}$ instead of θ, we would obtain the same autocorrelation function and so, it would be impossible to determine which of the two processes generated the data. Therefore, it is necessary to impose identifiability conditions on θ. In particular, $\frac{1}{\theta} > 1$ is the identifiability condition for a MA(1), which is also known as the invertibility condition, since it implies that the MA process can be "inverted" into an infinite order AR process.

In general, for a MA(q), the process is identifiable or invertible only when the roots of the MA characteristic polynomial $\Theta(u) = 1 + \theta_1 u + \ldots + \theta_q u^q$ lie outside the unit circle. In this case it is possible to write the MA process as an infinite order AR process. For an ARMA(p,q) process, the stationarity conditions are written in terms of the AR coefficients, i.e., the process is stationary only when the roots of the AR characteristic polynomial $\Phi(u) = 1 - \phi_1 u - \ldots - \phi_p u^p$ lie outside the unit circle. The ARMA process is invertible only when the roots of the MA characteristic polynomial lie outside the unit circle. So, if the ARMA process is stationary and invertible, it can be written either as a purely AR process of infinite order, or as a purely MA process of infinite order.

If y_t follows an ARMA(p,q) we can write $\Phi(B)y_t = \Theta(B)\epsilon_t$, with

$$\Phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p \quad \text{and} \quad \Theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q,$$

where B is the backshift operator. If the process is stationary then we can write it as a purely MA process of infinite order

$$y_t = \Phi^{-1}(B)\Theta(B)\epsilon_t = \Psi(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

with $\Psi(B)$ such that $\Phi(B)\Psi(B) = \Theta(B)$. The ψ_j values can be found by solving the homogeneous difference equations given by

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = 0, \quad j \geq \max(p, q+1), \tag{2.27}$$

with initial conditions

$$\psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, \quad 0 \leq j \leq \max(p, q+1), \tag{2.28}$$

and $\theta_0 = 1$. The general solution to the equations (2.27) and (2.28), is given by

$$\psi_j = \alpha_1^j p_1(j) + \ldots + \alpha_r^j p_r(j), \tag{2.29}$$

where $\alpha_1, \ldots, \alpha_r$ are the reciprocal roots of the characteristic polynomial $\Phi(u) = 0$, with multiplicities $m_1, \ldots, m_r$ respectively, and each $p_i(j)$ is a polynomial of degree $m_i - 1$.

2.17 Auto-Correlation and Partial-Autocorrelation Functions. If y_t follows a MA(q) process, it is possible to show that the ACF is given by

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \frac{\sum_{j=0}^{q-k} \theta_j \theta_{j+k}}{1+\sum_{j=1}^{q} \theta_j^2} & k = 1, \ldots, q \\ 0 & k > q, \end{cases} \tag{2.30}$$

and so, from a practical viewpoint it is possible to identify purely MA processes by looking at sample ACF plots, since the estimated ACF coefficients should drop after the q-th lag.

For general ARMA processes the autocovariance function can be written in terms of the general homogeneous equations

$$\gamma(k) - \phi_1 \gamma(k-1) - \ldots - \phi_p \gamma(k-p) = 0, \quad k \geq \max(p, q+1), \tag{2.31}$$

with initial conditions given by

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = v \sum_{j=k}^{q} \theta_j \psi_{j-k}, \quad 0 \leq k < \max(p, q+1). \tag{2.32}$$

The ACF of an ARMA is obtained dividing (2.31) and (2.32) by $\gamma(0)$.

The PACF can be obtained using any of the methods described in Section 2.6. The partial autocorrelation coefficients of a MA(q) process are never zero, as opposed to the partial autocorrelation coefficients of an AR(p) process which are zero after lag p. Similarly, for an invertible ARMA model, the partial autocorrelation coefficients will never drop to zero since the process can be written as an infinite order AR.

2.18 Inversion of AR Components. In contexts where data series are of reasonable length, we can fit longer order AR models rather than ARMA or other, more complex forms. One key reason is that the statistical analysis, at least conditional analyses based on assumedly fixed initial values, is much easier. The reference analysis for AR(p) processes described previously, for example, is essentially trivial compared with the numerical analysis required to produce samples from posterior distributions in ARMA models (see next sections). Another driving motivation is that longer order AR models will closely approximate ARMA forms. The proliferation of parameters is an issue, though with longer series and possible use of smoothness priors or other constraints, such as in using subset AR models, this is not an over-riding consideration.

If this view is adopted in a given problem, it may be useful and informative to use the results of an AR analysis to explore possible MA component structure using the device of inversion, or partial inversion, of the AR model. This is described here. Assume that y_t follows an AR(p) model with parameter vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$, so we can write

$$\Phi(B)y_t = \prod_{i=1}^{p}(1 - \alpha_i B)y_t = \epsilon_t,$$

where the α_i are the autoregressive characteristic roots. Often there will be subsets of pairs of complex conjugate roots corresponding to quasi-periodic components, perhaps with several real roots. Stationary components are implied by roots with moduli less than unity.

For some positive integer $r < p$, suppose that the final $p - r$ roots are identified as having moduli less than unity; some or all of the first r roots may also represent stationary components, though that is not necessary for the following development. Then, we can rewrite the model as

$$\prod_{i=1}^{r}(1 - \alpha_i B)y_t = \prod_{i=r+1}^{p}(1 - \alpha_i B)^{-1}\epsilon_t = \Psi^*(B)\epsilon_t,$$

where the (implicitly) infinite order MA component has the coefficients of the infinite order polynomial $\Psi^*(u) = 1 + \sum_{j=1}^{\infty}\psi_j^* u^j$, defined by

$$1 = \Psi^*(u)\prod_{j=r+1}^{p}(1 - \alpha_i u).$$

So we have the representation

$$y_t = \sum_{j=1}^{r}\phi_j^* y_{t-j} + \epsilon_t + \sum_{j=1}^{\infty}\psi_j^*\epsilon_{t-j},$$

where the r new AR coefficients ϕ_j^* are defined by the characteristic equation $\Phi^*(u) = \prod_{i=1}^{r}(1 - \alpha_i u) = 0$. The MA terms ψ_j^* can be easily calculated recursively, up to some appropriate upper bound on their number, say q. Explicitly, they are recursively computed as follows:

- for $i = 1, \ldots, q$, take $\psi_i^* = 0$ for $i = 1, 2, \ldots, q$; then,

- for $i = r + 1, \ldots, p$,
  - update $\psi_1^* = \psi_1^* + \alpha_i$, and then,
    - for $j = 2, \ldots, q$, update $\psi_j^* = \psi_j^* + \alpha_i \psi_{j-1}^*$.

Suppose $\boldsymbol{\phi}$ is set at some estimate, such as a posterior mean, in the AR(p) model analysis. The above calculations can be performed for any specified value of r to compute the corresponding MA coefficients in an inversion to the approximating ARMA(r, q) model. If the posterior for $\boldsymbol{\phi}$ is sampled in the AR analysis, the above computations can be performed repeatedly for all sampled $\boldsymbol{\phi}$ vectors, so producing corresponding samples of the ARMA parameters $\boldsymbol{\phi}^*$ and $\boldsymbol{\psi}^*$. Thus, inference in various relevant ARMA models can be directly, and quite easily, deduced by inversion of longer order AR models. Typically, various values of r will be explored. Guidance is derived from the estimated amplitudes and, in the case of complex roots, periods of the roots of the AR model. Analyses indicating some components that are persistent, i.e. that have moduli close to unity and, in the cases of complex roots, longer periods, suggest that these components be retained in the AR description. The remaining roots, corresponding to high frequency characteristics in the data with lower moduli and, if complex, high frequency oscillations, are then the candidates for inversion to what will often be a low order MA component. The calculations can be repeated, sequentially increasing q and exploring inferences about the MA parameters, to assess a relevant approximating order.

Example 2.18.1 *Exploring ARMA structure in the EEG data.*

It is of interest to enquire as to whether or not the residual noise structure in the EEG series may be adequately described by alternative moving average structure with, perhaps, fewer parameters than the above 8 or more in the AR description. This can be initiated directly from the AR analysis by exploring inversions of components of the auto-regressive characteristic polynomial, as follows.

For any AR parameter vector $\boldsymbol{\phi}$, we have the model

$$\phi(B)y_t = \prod_{i=1}^{8}(1 - \alpha_i B)y_t = \epsilon_t$$

where, by convention, the roots in order of decreasing moduli. In our AR(8) analysis there is a key and dominant component describing the major cyclical features that has modulus close to unity; the first two roots are complex conjugates corresponding to this component, the reference estimate of $\boldsymbol{\phi}$ produces an estimated modulus of 0.97 and frequency of 0.494. Identifying this as the key determinant of the AR structure, we can write the model as

$$\prod_{i=1}^{2}(1 - \alpha_i B)y_t = \prod_{i=3}^{8}(1 - \alpha_i B)^{-1}\epsilon_t = \Psi^*(B)\epsilon_t,$$

where the infinite order MA component is defined via

$$1 = \Psi^*(u) \prod_{j=3}^{8} (1 - \alpha_i u).$$

So we have the representation

$$y_t = \phi_1^* y_{t-1} + \phi_2^* y_{t-2} + \epsilon_t + \sum_{j=1}^{\infty} \psi_j^* \epsilon_{t-j},$$

where $\phi_1^* = 2r_1 \cos(\omega_1)$ and $\phi_2^* = -r_1^2$, with (r_1, ω_1) being the modulus and amplitude of the dominant cycle; in our case, the reference posterior mean from the fitted AR(8) model indicates values close to $\phi_1^* = 1.71$ and $\phi_2^* = -0.94$. The MA terms ψ_j^* can be easily calculated recursively, as detailed above.

This can be done for any specified AR(8) vector $\boldsymbol{\phi}$. Note that the roots typically are complex, though the resulting ψ_j^* must be real-valued. Note also that the ψ_j^* will decay rapidly so that q in the recursive algorithm is often rather moderate. Figure 2.13 displays a summary of such calculations based on the existing AR(8) analysis. Here $q = 8$ is chosen, so that the approximating ARMA model is ARMA$(2, 8)$, but with the view that the MA term is almost necessarily over-fitting. The above computations are performed in parallel for each of the 5,000 $\boldsymbol{\phi}$ vectors sampled from the reference posterior. This provides a Monte Carlo sample of size 5,000 from the posterior for the MA parameters obtained via this inversion techique. For each j, the sample distribution of values of ψ_j^* is summarised in Figure 2.13 by the vertical bar and points denoting approximate 90% intervals, 50% intervals and median. Note the expected feature that only rather few, in this case really only 2, of the MA coefficients are non-negligible; as a result, the inversion methods suggests that the longer order AR model is an approximation to a perhaps more parsimonious ARMA$(2, 2)$ form with AR parameters near 1.71 and -0.94, and with MA parameters around 1.4 and -0.6.

This analysis is supported by an exploratory search across ARMA(p, q) models for p and q taking values between 1 and 8. This can be done simply to produce rough guidelines as to model order using the conditional and approximate log-likelihood and AIC computations, for example. Conditioning on the first 16 observations in each case, the AIC values so computed are actually minimised at $p = q = 2$, so supporting the approach above. This model very significantly dominates others with $p \leq 2$, with AIC values differing by at least 5 units. The differences are far less for higher order models, and indeed a range of models with $p = 3$ or 4 come close on the AIC scale, with the ARMA$(4, 7)$ being the closest, less than one unit away on the AIC scale.

The approximate MLEs of the ARMA$(2, 2)$ parameters, based on this conditional analysis in R (R Development Core Team, 2004), are 1.70 (0.03) and -0.92 (0.03) for the AR component, and 1.37 (0.06) and -0.51 (0.06) for the MA. These agree well with the inversion of our Bayesian AR(8) analysis. Note that the inversion

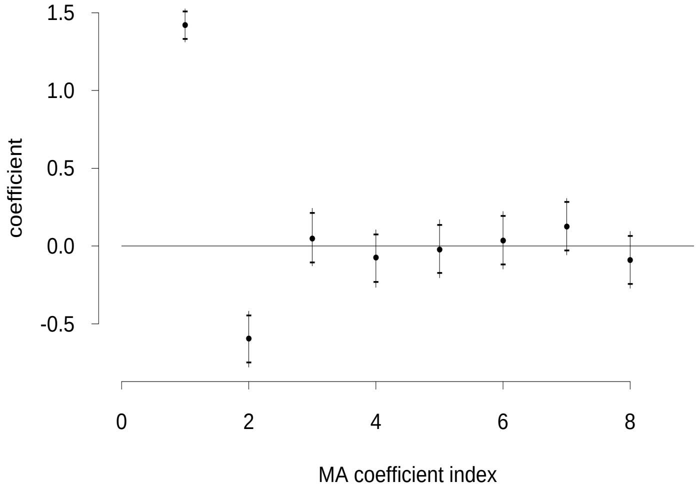

Fig. 2.13 Approximate posterior intervals for the first 8 MA coefficients from a partial inversion of the reference AR(8) analysis of the EEG series. Vertical bars display approximate 90% highest posterior density intervals, the marks denote 50% intervals and the dots denote posterior medians

approach directly supplies full posterior inferences, through easily implemented posterior simulations, in contrast to likelihood approaches. Note that this analysis could be repeated for higher order AR models. Proceeding to AR(10) or AR(12) produces models more taylored to minor noises features of the data. Subsequent inversion suggests possible higher order refinements, e.g., an ARMA$(3,3)$ model, though the global improvements in data fit and description are minor. Overall, though some additional insights are gleaned from exploring the MA structure, this particular segment of the EEG series is best described by the AR(8) and further analysis should be based on that. In other contexts, however, an ARMA structure may often be preferred.

2.19 Forecasting and Estimation ARMA processes.

2.19.1 Forecasting ARMA models. Consider a stationary and invertible ARMA process with parameters $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$. Given the stationarity and invertibility conditions, it is possible to write the process as a purely AR process of infinite order and so

$$y_{t+k} = \sum_{j=1}^{\infty} \phi_j^* y_{t+k-j} + \epsilon_{t+k}, \tag{2.33}$$

or as an infinite order MA process

$$y_{t+k} = \sum_{j=1}^{\infty} \theta_j^* \epsilon_{t+k-j} + \epsilon_{t+k}. \tag{2.34}$$

Let $y_{t+k}^{-\infty}$ be the minimum mean square predictor of y_{t+k} based on $y_t, y_{t-1}, \ldots, y_1, y_0$, $y_{-1}, \ldots$, which we denote as $y_{-\infty:t}$. In other words, $y_{t+k}^{-\infty} = E(y_{t+k}|y_{-\infty:t})$. Then, it is possible to show that (see problem 4)

$$y_{t+k} - y_{t+k}^{-\infty} = \sum_{j=0}^{k-1} \theta_j^* \epsilon_{t+k-j}, \tag{2.35}$$

with $\theta_0^* = 1$ and so, the mean square prediction error is given by

$$\text{MSE}_{t+k}^{-\infty} = E(y_{t+k} - y_{t+k}^{-\infty})^2 = v \sum_{j=0}^{k-1} (\theta_j^*)^2. \tag{2.36}$$

For a given sample size T, only the observations $y_1, \ldots, y_T$ are available, and so, we consider the following truncated predictor as an approximation

$$y_{T+k}^{-\infty,T} = \sum_{j=1}^{k-1} \phi_j^* y_{T+k-j}^{-\infty,T} + \sum_{j=k}^{T+k-1} \phi_j^* y_{T+k-j}. \tag{2.37}$$

This predictor is computed recursively for $k = 1, 2, \ldots,$ and the mean square prediction error is given approximately by (2.36).

In the AR(p) case, if $T > p$, the predictor y_{T+1}^T computed as in (2.12), given by

$$y_{T+1}^T = \phi_1 y_T + \phi_2 y_{T-1} + \ldots + \phi_p y_{T-p+1}, \tag{2.38}$$

yields to the exact predictor. This is true in general for any k, in other words, $y_{T+k}^T = y_{T+k}^{-\infty} = y_{T+k}^{-\infty,T}$, and so, there is no need for approximations. For general ARMA(p, q) models, the truncated predictor in (2.37) is

$$y_{T+k}^{-\infty,T} = \sum_{j=1}^{p} \phi_j y_{T+k-j}^{-\infty,T} + \sum_{j=1}^{q} \theta_j \epsilon_{T+k-j}^T, \tag{2.39}$$

where $y_t^{-\infty,T} = y_t$ for $1 \leq t \leq T$, $y_t^{-\infty,T} = 0$ for $t \leq 0$, and the truncated prediction errors are given by $\epsilon_t^T = 0$ for $t \leq 0$ or $t > T$ and

$$\epsilon_t^T = \boldsymbol{\phi}(B) y_t^{-\infty,T} - \theta_1 \epsilon_{t-1}^T - \ldots - \theta_q \epsilon_{t-q}^T$$

for $1 \leq t \leq T$.

2.19.2 MLE and least squares estimation. For an ARMA(p, q) model we need to estimate the parameters $\boldsymbol{\beta}$ and v where $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$. The likelihood function can be written as follows

$$p(y_{1:T}|\boldsymbol{\beta}, v) = \prod_{t=1}^{T} p(y_t|y_{1:(t-1)}, \boldsymbol{\beta}, v). \tag{2.40}$$

Assuming that the conditional distribution of y_t given $y_{1:(t-1)}$ is Gaussian with mean y_t^{t-1} and variance $V_t^{t-1} = v r_t^{t-1}$, we can write

$$-2 \log\left[p(y_{1:T}|\boldsymbol{\beta}, v)\right] = T \log(2\pi v) + \sum_{t=1}^{T} \left[\log(r_t^{t-1}) + \frac{(y_t - y_t^{t-1})^2}{r_t^{t-1}}\right], \tag{2.41}$$

where y_t^{t-1} and r_t^{t-1} are functions of $\boldsymbol{\beta}$ and so, the maximum likelihood estimates of $\boldsymbol{\beta}$ and v are computed by minimizing the expression (2.41) with respect to $\boldsymbol{\beta}$ and v. The equation (2.41) is usually a non-linear function of the parameters and so, the minimization has to be done using a non-linear optimization algorithm such as the Newton-Raphson algorithm described in Chapter 1.

Least squares (LS) estimation can be performed by minimising the expression

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{T} \frac{(y_t - y_t^{t-1})^2}{r_t^{t-1}},$$

with respect to $\boldsymbol{\beta}$. Similarly, conditional least squares estimation is performed by conditioning on the first p values of the series $y_{1:p}$ and assuming that $\epsilon_p = \epsilon_{p-1} = \cdots = \epsilon_{1-q} = 0$. In this case we can minimize the conditional sum of squares given by

$$S_c(\boldsymbol{\beta}) = \sum_{t=p+1}^{T} \epsilon_t(\boldsymbol{\beta})^2, \tag{2.42}$$

where $\epsilon_t(\boldsymbol{\beta}) = y_t - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j}(\boldsymbol{\beta})$. When $q = 0$ this reduces to a linear regression problem and so, no numerical minimisation technique is required. When the number of observations T is not very large conditioning on the first initial values will have an influence on the parameter estimates. In such cases working with the unconditional sum of squares might be preferable. Several methodologies have been proposed to handle unconditional least squares estimation. In particular, Box *et al.* (1994, Appendix A7.3) showed that an approximation to the unconditional sum of squares $S(\boldsymbol{\beta})$ is

$$S(\boldsymbol{\beta}) = \sum_{t=-M}^{T} \hat{\epsilon}_t^2(\boldsymbol{\beta}), \tag{2.43}$$

with $\hat{\epsilon}_t(\boldsymbol{\beta}) = E(\epsilon_t | y_{1:n})$ and if $t \leq 0$ these values are obtained by backcasting. Here M is chosen to be such that $\sum_{t=-\infty}^{-M} \hat{\epsilon}_t^2(\boldsymbol{\beta}) \approx 0$.

A Gauss-Newton procedure (see Shumway and Stoffer, 2000, Section 2.6 and references therein) can be used to obtain an estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, that minimises $S(\boldsymbol{\beta})$ or $S_c(\boldsymbol{\beta})$. For instance, in order to find an estimate of $\boldsymbol{\beta}$ that minimises the conditional sum of squares in (2.42), the following algorithm is repeated by computing $\boldsymbol{\beta}^{(j)}$ at each iteration $j = 1, 2, \ldots,$ until convergence is reached

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

where

$$\Delta(\boldsymbol{\beta}) = \frac{\sum_{t=p+1}^{T} \mathbf{z}_t(\boldsymbol{\beta})\epsilon_t(\boldsymbol{\beta})}{\sum_{t=p+1}^{T} \mathbf{z}_t'(\boldsymbol{\beta})\mathbf{z}_t(\boldsymbol{\beta})}$$

and

$$\mathbf{z}_t(\boldsymbol{\beta}) = \left(-\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, -\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_{p+q}} \right)'. \tag{2.44}$$

Convergence is considered to be achieved when $|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}| < \delta_{\boldsymbol{\beta}}$, or when $|Q_c(\boldsymbol{\beta}^{(j+1)}) - Q_c(\boldsymbol{\beta}^{(j)})| < \delta_Q$, where $\delta_{\boldsymbol{\beta}}$ and δ_Q are set to some fixed small values. Here, $Q_c(\boldsymbol{\beta})$ is a linear approximation of $S_c(\boldsymbol{\beta})$ given by

$$Q_c(\boldsymbol{\beta}) = \sum_{t=p+1}^{T} \left[\epsilon_t(\boldsymbol{\beta}^{(0)}) - (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})'\mathbf{z}_t(\boldsymbol{\beta}^{(0)})\right]^2$$

and $\boldsymbol{\beta}^{(0)}$ is an initial estimate of $\boldsymbol{\beta}$.

Example 2.19.1 *Conditional LS estimation of the parameters of an ARMA*$(1,1)$. Consider an stationary and invertible ARMA$(1,1)$ process described by

$$y_t = \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t,$$

with $\epsilon_t \sim N(0, v)$. Then, we can write $\epsilon_t(\boldsymbol{\beta}) = y_t - \phi_1 y_{t-1} - \theta_1 \epsilon_{t-1}(\boldsymbol{\beta})$, with $\boldsymbol{\beta} = (\phi_1, \theta_1)'$. Additionally, we condition on $\epsilon_0(\boldsymbol{\beta}) = 0$ and y_1. Now, using the expression (2.44) we have that $\mathbf{z}_t = (z_{t,1}, z_{t,2})'$ with $z_{t,1} = y_{t-1} + \theta_1 z_{t-1,1}$ and $z_{t,2} = \epsilon_{t-1} + \theta_1 z_{t-1,2}$, where $\mathbf{z}_0 = \mathbf{0}$. The Gauss-Newton algorithm starts with some initial value of $\boldsymbol{\beta}^{(0)} = (\phi_1^{(0)}, \theta_1^{(0)})'$ and then, at each iteration $j = 1, 2, \ldots,$ we have

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \frac{\sum_{t=2}^{T} \mathbf{z}_t(\boldsymbol{\beta})\epsilon_t(\boldsymbol{\beta})}{\sum_{t=2}^{T} \mathbf{z}_t'(\boldsymbol{\beta})\mathbf{z}_t(\boldsymbol{\beta})}.$$

2.19.3 State-Space representation and Kalman-Filter estimation. Due to the computational burden of maximising the exact likelihood given in (2.40), many of the existing methods for parameter estimation in the ARMA modelling framework consider approximations to the exact likelihood, such as the backcasting method of Box *et al.* (1994). There are also approaches that allow the computation of the exact likelihood function. Some of these approaches involve rewriting the ARMA model in state-space or dynamic linear model (DLM) form, and then applying the Kalman filter to achieve parameter estimation (see for example Kohn and Ansley, 1985; Harvey, 1981 and Harvey, 1991).

A state-space or DLM model is usually defined in terms two equations, one that describes the evolution of the time series at the observational level, and another equation that describes the evolution of the system over time. One of the most useful ways of representing the ARMA(p,q) model given in (2.26) is by writing it in the state-space or DLM form given by the following equations

$$\begin{aligned} y_t &= \mathbf{E}_m' \boldsymbol{\theta}_t \\ \boldsymbol{\theta}_t &= \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \end{aligned} \tag{2.45}$$

where $\mathbf{E}_m = (1, 0, \ldots, 0)'$ is a vector of dimension m, with $m = \max(p, q+1)$, $\boldsymbol{\omega}_t$ is also a vector of dimension m with $\boldsymbol{\omega}_t = (1, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m-1})'\epsilon_t$ and $\mathbf{G}$ is an $m \times m$

matrix given by

$$\mathbf{G} = \begin{pmatrix} \phi_1 & 1 & 0 & \dots & 0 \\ \phi_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1} & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Here $\phi_r = 0$ for all $r > p$ and $\theta_r = 0$ for all $r > q$. The evolution noise has a variance-covariance matrix **U** given by $\mathbf{U} = \sigma^2(1, \theta_1, \dots, \theta_{m-1})'(1, \theta_1, \dots, \theta_{m-1})$.

Using this representation it is possible to perform parameter estimation for general ARMA(p, q) models. We will revisit this topic after developing the theory of DLMs in Chapter 4.

2.19.4 Bayesian Estimation of ARMA processes. There are several approaches to Bayesian estimation of general ARMA models, e.g., Zellner (1996), Box *et al.* (1994), Monahan (1983), Marriott and Smith (1992), Marriott *et al.* (1996), Chib and Greenberg (1994) and Barnett *et al.* (1997).

We briefly outline the approach proposed in Marriott *et al.* (1996) and discuss some aspects related to alternative ways of performing Bayesian estimation in ARMA models. Such approach leads to parameter estimation of ARMA(p, q) models via Markov chain Monte Carlo by reparameterising the ARMA parameters in terms of partial autocorrelation coefficients. Specifically, let $f(y_{1:T}|\boldsymbol{\psi}^*)$ be the likelihood for the T observations given the vector of parameters $\boldsymbol{\psi}^* = (\boldsymbol{\phi}', \boldsymbol{\theta}', \sigma^2, \mathbf{x}_0', \boldsymbol{\epsilon}_0')$, with $\boldsymbol{\epsilon}_0 = (\epsilon_0, \epsilon_{-1}, \dots, \epsilon_{1-q})'$. This likelihood function is given by

$$f(y_{1:T}|\boldsymbol{\psi}^*) = (2\pi\sigma^2)^{-T/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \mu_t)^2 \right\}, \tag{2.46}$$

where,

$$\begin{aligned}
\mu_1 &= \sum_{i=1}^{p} \phi_i y_{1-i} + \sum_{i=1}^{q} \theta_i \epsilon_{1-i}, \\
\mu_t &= \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{t-1} \theta_i (y_{t-i} - \mu_{t-i}) + \sum_{i=t}^{q} \theta_i \epsilon_{t-i}, \quad t = 2, \dots, q, \\
\mu_t &= \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i (y_{t-i} - \mu_{t-i}), \quad t = q+1, \dots, T.
\end{aligned}$$

The prior specification is as follows

$$\pi(\boldsymbol{\psi}^*) = \pi(\mathbf{x}_0, \boldsymbol{\epsilon}_0 | \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) \pi(\sigma^2) \pi(\boldsymbol{\phi}, \boldsymbol{\theta}),$$

with $\pi(\mathbf{x}_0, \boldsymbol{\epsilon}_0|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = N(\mathbf{0}, \sigma^2\Omega)$, $\pi(\sigma^2) \propto \sigma^{-2}$ and $\pi(\boldsymbol{\phi}, \boldsymbol{\theta})$ a uniform distribution in the stationary and invertibility regions of the ARMA process denoted by $\mathcal{C}_p$ and $\mathcal{C}_q$, respectively. Therefore, the joint posterior for $\boldsymbol{\psi}^*$ is given by

$$\begin{aligned} \pi(\boldsymbol{\psi}^*|y_{1:T}) \quad &\propto \quad (\sigma^2)^{-(T+2)/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \mu_t)^2\right\} \times \qquad (2.47) \\ & \qquad N((\mathbf{x}_0', \boldsymbol{\epsilon}_0')'|\mathbf{0}, \sigma^2\Omega) \qquad (2.48) \end{aligned}$$

The MCMC algorithm to perform parameter estimation can be summarized in terms of the following steps:

- Sample $(\sigma^2|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}_0, \boldsymbol{\epsilon}_0)$. This is done by sampling σ^2 from the inverse-Gamma full conditional distribution with the following form

$$IG\left(\frac{T+p+q}{2}, \frac{1}{2}\left[\begin{pmatrix} \mathbf{x}_0 \\ \boldsymbol{\epsilon}_0 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} \mathbf{x}_0 \\ \boldsymbol{\epsilon}_0 \end{pmatrix} + \sum_{t=1}^{T}(y_t - \mu_t)^2\right]\right)$$

- Sample $(\mathbf{x}_0, \boldsymbol{\epsilon}_0|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$. The full conditional distribution of $(\mathbf{x}_0', \boldsymbol{\epsilon}_0')$ is a multivariate normal, however, it is computationally simpler to use Metropolis steps with Gaussian proposal distributions.

- Sample $(\boldsymbol{\phi}, \boldsymbol{\theta}|\sigma^2, \mathbf{x}_0, \boldsymbol{\epsilon}_0)$. In order to sample $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, successive transformations for $\mathcal{C}_p$ and $\mathcal{C}_q$ to p-dimensional and q-dimensional hypercubes and then to R^p and R^q, respectively, are considered. The transformations of $\mathcal{C}_p$ and $\mathcal{C}_q$ to the p-dimensional and q-dimensional hypercubes were proposed by Monnahan (1984), extending the work of Barndorff-Nielsen and Schou (1973). Specifically, the transformation for the AR parameters is given by

  $$\phi(i,k) = \phi(i,k-1) - \phi(k,k)\phi(k-i,k-1), \quad i = 1, \ldots, k-1,$$

  where $\phi(k,k)$ is the partial autocorrelation coefficient and $\phi(j,p) = \phi_j$, the j-th coefficient from the AR(p) process defined by the characteristic polynomial $\Phi(u) = 1 - \phi_1 u - \ldots - \phi_p u^p$. The inverse transformation in iterative form is given by

  $$\phi(i,k-1) = [\phi(i,k) + \phi(k,k)\phi(k,k-i)]/[1 - \phi^2(k,k)],$$

  and the Jacobian of the transformation is

  $$J = \prod_{i=1}^{p}(1 - \phi(k,k)^2)^{[(k-1)/2]} \prod_{j=1}^{[p/2]}(1 - \phi(2j,2j)).$$

  Now, the stationarity condition on $\boldsymbol{\phi}$ can be written in terms of the partial autocorrelation coefficients as $|\phi(k,k)| < 1$ for all $k = 1, \ldots, p$. Marriott *et al.* (1996) then propose a transformation from $\mathbf{r}_\phi = (\phi(1,1), \ldots, \phi(p,p))'$ to

$\mathbf{r}^*_\phi = (\phi^*(1,1), \ldots, \phi^*(p,p))'$, with $\mathbf{r}^*_\phi \in R^p$. The $\phi^*(j,j)$ elements are given by

$$\phi^*(j,j) = \log\left(\frac{1+\phi(j,j)}{1-\phi(j,j)}\right).$$

Similarly, a transformation from $\boldsymbol{\theta}$ to $\mathbf{r}^*_\theta \in R^q$ can be defined using the previous two steps replacing $\boldsymbol{\phi}$ by $\boldsymbol{\theta}$.

Then, instead of sampling $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ from the constrained full conditional distributions, we can sample unconstrained full conditional distributions for $\mathbf{r}^*_\phi$ and $\mathbf{r}^*_\theta$ on R^p and R^q, respectively. Marriott *et al.* (1996) suggest using a Metropolis step as follows. First, compute MLE estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, say $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, with its asymptotic variance covariance matrix $\Sigma_{(\hat{\phi},\hat{\theta})}$. Use the transformations described above to obtain $(\hat{\mathbf{r}}^*_{\hat{\phi}}, \hat{\mathbf{r}}^*_{\hat{\theta}})$ and a corresponding variance covariance matrix Σ^* (computed via the delta method). Let $g_{p+q}(\mathbf{r}^*_\phi, \mathbf{r}^*_\theta)$ be the $p+q$-dimensional multivariate normal distribution with mean $(\hat{\mathbf{r}}^*_{\hat{\phi}}, \hat{\mathbf{r}}^*_{\hat{\theta}})$ and variance covariance matrix Σ^*. Take g_{p+q} to be the proposal density in the Metropolis step build to sample $\mathbf{r}^*_\phi$ and $\mathbf{r}^*_\theta$.

Example 2.19.2 *Bayesian estimation in an ARMA(1,1).*

Consider an ARMA(1,1) model described by $y_t = \phi y_{t-1} + \theta y_{t-1} + \epsilon_t$, with $N(\epsilon_t|0,\sigma^2)$. In this case $\mathbf{x}_0 = y_0$, $\boldsymbol{\epsilon}_0 = \epsilon_0$, $\mathbf{r}_\phi = \phi$, $\mathbf{r}_\theta = \theta$, $\mathbf{r}^*_\phi = \phi^*$, $\mathbf{r}^*_\theta = \theta^*$,

$$\Omega = \begin{pmatrix} 1 & 1 \\ 1 & \frac{(1+\theta^2+2\phi\theta)}{(1-\phi^2)} \end{pmatrix}, \quad \phi^* = \log\left(\frac{1+\phi}{1-\phi}\right), \quad \theta^* = \log\left(\frac{1+\theta}{1-\theta}\right),$$

and the inverse of the determinant of the Jacobian of the transformation is given by $(1-\phi^2)(1-\theta^2)/4$.

Discussion and Further Topics

2.20 ARIMA Models

2.21 ARFIMA Models

Appendix

The Reversible Jump MCMC algorithm. In general, the RJMCMC method can be described as follows. Assume that $\boldsymbol{\theta}$ is a vector of parameters to be estimated and $\pi(d\boldsymbol{\theta})$ is the target probability measure, which often is a mixture of densities, or a mixture with continuous and discrete parts. Suppose that $m = 1, 2, \ldots,$ indexes all the possible dimensions of the model. If the current state of the Markov chain is $\boldsymbol{\theta}$ and a move of type m and destination $\boldsymbol{\theta}^*$ is proposed from a proposal measure $q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}^*)$, the move is accepted probability

$$\alpha_m(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(d\boldsymbol{\theta}^*)q_m(\boldsymbol{\theta}^*, d\boldsymbol{\theta})}{\pi(d\boldsymbol{\theta})q_m(\boldsymbol{\theta}, d\boldsymbol{\theta}^*)}\right\}.$$

For cases in which the move type does not change the dimension of the parameter, the expression above reduces to the Metropolis-Hastings acceptance probability,

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|y_{1:n})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}{p(\boldsymbol{\theta}|y_{1:n})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}\right\},$$

where $p(\cdot|y_{1:n})$ denotes the target density of posterior density in our case. If $\boldsymbol{\theta}$ is a parameter vector of dimension m_1 and $\boldsymbol{\theta}^*$ a parameter vector of dimension m_2, with $m_1 \neq m_2$, the transition between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ is done by generating u_1 of dimension n_1 from a density $q_1(u_1|\boldsymbol{\theta})$, and u_2 of dimension n_2 from a density $q_2(u_2|\boldsymbol{\theta}^*)$, such that $m_1 + n_1 = m_2 + n_2$. Now, if $J(m, m^*)$ denotes the probability of a move of type m^* given that the chain is at m, the acceptance probability is

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*, m_2|y_{1:n})J(m_1, m_2)q_2(u_2|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}, m_1|y_{1:n})J(m_2, m_1)q_1(u_1|\boldsymbol{\theta})}\left|\frac{\partial(\boldsymbol{\theta}^*, u_2)}{\partial(\boldsymbol{\theta}, u_1)}\right|\right\}.$$

Problems

1. Consider the AR(1) process. If $|\phi| < 1$ the process is stationary and it is possible to write $y_t = \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j}$. Use this fact to prove that $y_1 \sim N(0, v/(1-\phi^2))$ and so, the likelihood function has the form (1.17).

2. Consider an AR(2) process with AR coefficients $\phi = (\phi_1, \phi_2)'$. Show that the process is stationary for parameter values lying in the region $-2 < \phi_1 < 2$, $\phi_1 < 1 - \phi_2$ and $\phi_1 > \phi_2 - 1$.

3. Show that the general solution of a homogeneous difference equation of the form (2.8) has the form (2.9).

4. Show that equations (2.35) and (2.36) hold by taking expected values in (2.33) and (2.34) with respect to the whole past history $y_{-\infty,t}$.

5. Consider the ARMA(1,1) model described by

$$y_t = 0.95y_{t-1} + 0.8\epsilon_{t-1} + \epsilon_t,$$

with $\epsilon_t \sim N(0,1)$ for all t.

(a) Show that the one-step-ahead truncated forecast is given by $y_{t+1}^{t,-\infty} = 0.95y_t + 0.8\epsilon_t^{t,-\infty}$, with $\epsilon_t^{t,-\infty}$ computed recursively via $\epsilon_j^{t,-\infty} = y_j - 0.95y_{j-1} - 0.8\epsilon_{j-1}^{t,-\infty}$, for $j = 1, \ldots, t$ with $\epsilon_0^{t,-\infty} = 0$ and $y_0 = 0$.

(b) Show that the approximate mean square prediction error is

$$MSE_t^{t,-\infty} = v\left[1 + \frac{(\phi+\theta)^2(1-\phi^{2(k-1)})}{(1-\phi^2)}\right]$$

3 The Frequency Domain

4 Dynamic Linear Models

5 TVAR Models

6 Other Univariate Models and Methods

Part II

MULTIVARIATE TIME SERIES

7 Multiple Time Series

8 Multivariate Models

9 Multivariate Models in Finance

10 Other Multivariate Models and Methods

REFERENCES

Akaike, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematical*, **21,** 243–247.

Akaike, H. (1974) A new look to statistical model identification. *IEEE Trans. Automat. Contr.* A, **C-19,** 716–723.

Barndorff-Nielsen, O.E. and Schou, G. (1973) On the reparameterization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, **3,** 408–419.

Barnett, G., Kohn, R. and Sheather, S. (1997) Robust Bayesian estimation of autoregressive-moving-average models. *Journal of Time Series Analysis*, **18,** 11–28.

Box, G., Jenkins, G.M. and Reinsel, G. (1994) *Time series Analysis Forecasting and Control*, Third edn. Pearson Education.

Brockwell, P.J. and Davis, R.A. (1991) *Time series: Theory and Methods*, Second edn. New York: Springer-Verlag.

Brooks, S. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7(4),** 434–55.

Chatfield, C. (1996) *The Analysis of Time Series: An Introduction*, Fifth edn. London: Chapman and Hall.

Chib, S. and Greenberg, E. (1994) Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics*, **64,** 183–206.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B-Methodological*, **39(1),** 1–18.

Diggle, P. (1990) *Time Series: A Biostatistical Introduction*. Oxford.

Durbin, J. (1960) Estimation of parameters in time series regression models. *Journal of the Royal Statistical Society, Series* B, **22,** 139–153.

Gamerman, D. (1997) *Markov Chain Monte Carlo*. London: Chapman & Hall.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, Second edn. Chapman & Hall/CRC.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7,** 457–72.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Clarendon Press, Oxford, UK.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82,** 711–732.

Harvey, A. (1981) *Time Series Models*. London: Philip Allan.

Harvey, A.C. (1991) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and its applications. *Biometrika*, **57,** 97–109.

Huerta, G. (1998) Bayesian analysis of latent structure in time series models. Ph.D. Thesis. Duke University, Durham, NC.

Huerta, G. and West, M. (1999) Priors and component structures in autoregressive time series models. *J. R. Statist. Soc. B.*, **61,** 881–899.

Kendall, M.G. and Ord, J.K. (1990) *Time Series*, 3rd edn. Sevenoaks: Edward Arnold.

Kendall, M.G., Stuart, A. and J.K., J.K. Ord (1983) *The Advance Theory of Statistics*, 4th edn, vol. 3. London: Griffin.

Kohn, R. and Ansley, C. F. (1985) Efficient estimation and prediction in time series regression models. *Biometrika*, **72,** 694–697.

Krystal, A.D., Prado, R. and West, M. (1999) New methods of time series analysis of non-stationary EEG data: eigenstructure decompositions of time-varying autoregressions. *Clinical Neurophysiology*, **110,** 2197–2206.

Levinson, N. (1947) The Wiener (root mean square) error criterion in filter and design prediction. *J. Math. Phys.*, **25,** 262–278.

Marriott, J., Ravishanker, N., Gelfand, A. and Pai, J. (1996) Bayesian analysis of ARMA processes: complete sampling-based inference under exact likelihoods. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (eds K. M. Chaloner D. A. Berry and J. K. Geweke), pp. 243–256.

Marriott, J.M. and Smith, A.F.M. (1992) Reparametrization aspects of numerical Bayesian methodology for autoregressive moving-average models. *Journal of Time Series Analysis*, **13,** no. 4, 327–343.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21,** 1087–1091.

Monahan, J. F. (1983) Fully bayesian analysis of ARMA time series models. *Journal of Econometrics,* **21,** 307–331.

Monnahan, J.F. (1984) A note on enforcing stationarity in autoregressive moving average models. *Biometrika,* **71,** 403–404.

Prado, R., West, M. and Krystal, A.D. (2001) Multi-channel EEG analyses via dynamic regression models with time-varying lag/lead structure. *Journal of the Royal Statistical Society, Series C (Applied Statistics),* **50,** 95–109.

Priestley, M. B. (1994) *Spectral Analysis and Time Series,* Eighth edn. Academic Press.

R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Raftery, A. L. and Lewis, S. (1992) How many iterations in the gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 763–74. Oxford University Press.

Schwarz, G. (1978) Estimating the dimension of the model. *Annals of Statistics,* **6,** 461–464.

Shumway, R.H. and Stoffer, D.S. (2000) *Time Series Analysis and Its Applications*. New York: Springer-Verlag.

Smith, Brian J. (2004) *Bayesian Output Analysis program (BOA) version 1.1 user's manual*.

Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag.

Tong, H. (1990) *Non-linear Time Series: A Dynamical Systems Approach*. Oxford: Oxford University Press.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. Wiley Texts in Statistics.

Wei, G.C.G and Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association,* **85,** 699–704.

Welch, P. Heidelberger P. (1983) Simulation run length control in the presence of an initial transient. *Operations Research*, **31,** 1109–44.

West, M., Prado, R. and Krystal, A.D. (1999) Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association,* **94,** 1083–1095.

Zellner, A. (1996) *An introduction to Bayesian inference in econometrics*. New York: Wiley.