9

Parametric Spectral Estimators

9.0 Introduction

In this chapter we discuss the basic theory behind parametric spectral density function (SDF) estimation. The main idea is simple. Suppose the discrete parameter stationary process $\{X_t\}$ has an SDF $S(\cdot)$ that is completely determined by K parameters $a_1, \ldots, a_K$:

$$S(f) = S(f; a_1, \ldots, a_K).$$

Using a time series that can be regarded as a realization of this process, suppose we can estimate the parameters of $S(\cdot)$ by, say, $\hat{a}_1, \ldots, \hat{a}_K$. If these parameter estimates are reasonable, then

$$\hat{S}(f) = S(f; \hat{a}_1, \ldots, \hat{a}_K)$$

should be a reasonable estimate of $S(f)$.

9.1 Notation

In what follows, it is important to keep in mind what basic assumptions are in effect. Accordingly, in this chapter we adopt these notational conventions for the following discrete parameter stationary processes, all of which are assumed to be real-valued and have zero mean:

[1] $\{X_t\}$ represents an arbitrary such process;
[2] $\{Y_t\}$, an *autoregressive* process of finite order p;
[3] $\{G_t\}$, a *Gaussian* process; and
[4] $\{H_t\}$, a *Gaussian autoregressive* process of finite order p.

For any of these four processes, we denote the autocovariance sequence (ACVS) by $\{s_\tau : \tau \in \mathbb{Z}\}$, and we assume there is a corresponding SDF $S(\cdot)$ with an associated Nyquist frequency $f_{\mathcal{N}} \stackrel{\text{def}}{=} 1/(2\,\Delta_{\text{t}})$, where Δ_{t} is the sampling interval between values in the process (hence $S(\cdot)$ is periodic with a period of $2f_{\mathcal{N}}$).

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

9.2 The Autoregressive Model

The most widely used form of parametric SDF estimation involves an autoregressive model of order p, denoted as AR(p), as the underlying functional form for $S(\cdot)$. Recall that a stationary AR(p) process $\{Y_t\}$ with zero mean satisfies the equation

$$Y_t = \phi_{p,1} Y_{t-1} + \cdots + \phi_{p,p} Y_{t-p} + \epsilon_t = \sum_{j=1}^{p} \phi_{p,j} Y_{t-j} + \epsilon_t, \tag{446a}$$

where $\phi_{p,1}, \ldots, \phi_{p,p}$ are p fixed coefficients, and $\{\epsilon_t\}$ is a white noise process with zero mean and finite variance $\sigma_p^2 \stackrel{\text{def}}{=} \text{var}\{\epsilon_t\}$ (we also assume $\sigma_p^2 > 0$). The process $\{\epsilon_t\}$ is often called the *innovation process* that is associated with the AR(p) process, and σ_p^2 is called the *innovation variance*. Since we will have to refer to AR(p) models of different orders, we include the order p of the process in the notation for its parameters. The parameterization of the AR model given in Equation (446a) is the same as that used by, e.g., Box et al. (2015), but the reader should be aware that there is another common way of writing an AR(p) model, namely,

$$Y_t + \varphi_{p,1} Y_{t-1} + \cdots + \varphi_{p,p} Y_{t-p} = \epsilon_t,$$

where $\varphi_{p,j} = -\phi_{p,j}$. This convention is used, for example, by Priestley (1981) and is prevalent in the engineering literature. Equation (446a) emphasizes the analogy of an AR(p) model to a multiple linear regression model, but it is only an analogy and not a simple correspondence: if we regard Equation (446a) as a regression model, the predictors $Y_{t-1}, \ldots, Y_{t-p}$ are just lagged copies of the response Y_t. This setup does not fit into the usual way of thinking about regression models (e.g., the mean function in such models is the expected value of the response conditional on the predictors, and hence, when modeling a time series of length N, certain (unconditioned) responses would also need to serve as conditioned predictors).

The SDF for a stationary AR(p) process is given by

$$S(f) = \frac{\sigma_p^2 \, \Delta_t}{\left| 1 - \sum_{j=1}^{p} \phi_{p,j} e^{-i2\pi f j \, \Delta_t} \right|^2} \tag{446b}$$

(cf. Equation (145a) with $\Delta_t = 1$; the above is a periodic function with a period of $2f_{\mathcal{N}}$). Here we have $p + 1$ parameters, namely, the $\phi_{p,j}$ coefficients and σ_p^2, all of which we must estimate to produce an AR(p) SDF estimate. The coefficients cannot be chosen arbitrarily if $\{Y_t\}$ is to be a stationary process. As we noted in Section 5.4, a necessary and sufficient condition for the existence of a stationary solution to $\{Y_t\}$ of Equation (446a) is that

$$G(f) = 1 - \sum_{j=1}^{p} \phi_{p,j} e^{-i2\pi f j} \neq 0 \text{ for any } f \in \mathbb{R} \tag{446c}$$

(see Equation (144d)). Another way to state this condition is that none of the solutions to the polynomial equation $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j} = 0$ lies *on* the unit circle in the complex plane (i.e., has an absolute value equal to unity). When $p = 1$, we can thus achieve stationarity as long as $\phi_{1,1} \neq \pm 1$. In addition to stationarity, an important property for $\{Y_t\}$ to possess is causality (see item [1] in the Comments and Extensions [C&Es] at the end of this section for a discussion of causality). The process $\{Y_t\}$ is causal if the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$ all lie *inside* the unit circle; if any are outside, it is acausal. Causality implies stationarity, but the converse is not true. An AR(1) process is causal (and hence stationary) if $|\phi_{1,1}| < 1$. For an

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

AR process of general order p, causality implies that $\mathrm{cov}\{\epsilon_s, Y_t\} = 0$ when $s > t$, which is a key property we put to good use in the next section.

The rationale for this particular class of parametric SDFs is six-fold.

[1] Any continuous SDF $S(\cdot)$ can be approximated arbitrarily well by an AR(p) SDF if p is chosen large enough (Anderson, 1971, p. 411). Thus the class of AR processes is rich enough to approximate a wide range of processes. Unfortunately "large enough" can well mean an order p that is too large compared to the amount of available data.

[2] There exist efficient algorithms for fitting AR(p) models to time series. This might seem like a strange justification, but, since the early days of spectral analysis, recommended methodology has often been governed by what can in practice be calculated with commonly available computers.

[3] For a Gaussian process $\{G_t\}$ with autocovariances known up to lag p, the maximum entropy spectrum is identical to that of an AR(p) process. We discuss the principle of maximum entropy and comment upon its applicability to spectral analysis in Section 9.6.

[4] A side effect of fitting an AR(p) process to a time series for the purpose of SDF estimation is that we have simultaneously estimated a linear predictor and potentially identified a linear prewhitening filter for the series (the role of prewhitening in SDF estimation is discussed in Section 6.5). In Section 9.10 we describe an overall approach to SDF estimation that views parametric SDF estimation as a good method for determining appropriate prewhitening filters.

[5] For certain phenomena a physical argument can be made that an AR model is appropriate. A leading example is the acoustic properties of human speech (Rabiner and Schafer, 1978, chapter 3).

[6] Pure sinusoidal variations can be expressed as an AR model with $\sigma_p^2 = 0$. This fact – and its implications – are discussed in more detail in Section 10.12.

We have already encountered two examples of AR(p) processes in Chapter 2: the AR(2) process of Equation (34) and the AR(4) process of Equation (35a). The SDF for the AR(2) process is plotted as thick curves in the upper two rows of Figures 172 and 173; the SDF for the AR(4) process, in the lower two rows. Four realizations of each process are shown in Figure 34.

To form an AR(p) SDF estimate from a given set of data, we face two problems: first, determination of the order p that is most appropriate for the data and, second, estimation of the parameters $\phi_{p,1}, \ldots, \phi_{p,p}$ and σ_p^2. Typically we determine p by examining how well a range of AR models fits our data (as judged by some criterion), so it is necessary first to assume that p is known and to learn how to estimate the parameters for an AR(p) model.

Comments and Extensions to Section 9.2

[1] As noted in this section, as long as $|\phi_{1,1}| \neq 1$, there is a unique stationary solution to the equation $Y_t = \phi_{1,1} Y_{t-1} + \epsilon_t$, $t \in \mathbb{Z}$, where $\{\epsilon_t\}$ is a zero mean white noise process with finite variance $\sigma_p^2 > 0$. The burden of Exercise [9.1a] is to show that these solutions are

$$Y_t = \sum_{j=0}^{\infty} \phi_{1,1}^j \epsilon_{t-j} \text{ when } |\phi_{1,1}| < 1 \text{ and } Y_t = -\sum_{j=1}^{\infty} \phi_{1,1}^{-j} \epsilon_{t+j} \text{ when } |\phi_{1,1}| > 1 \tag{447}$$

(note that Exercise [2.17a] eludes to the first solution). By definition the first solution is *causal* because Y_t only depends upon random variables (RVs) in the white noise process $\{\epsilon_s : s \in \mathbb{Z}\}$ such that $s \leq t$; by contrast, the second solution is *purely acausal* because Y_t depends solely upon RVs occurring *after*

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

index t. For AR(p) processes in general, the condition that all of the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$ lie inside the unit circle guarantees that there is a unique stationary solution to Equation (446a) of the form

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

where the ψ_j weights depend on $\phi_{p,1}, \ldots, \phi_{p,p}$ (in the AR(1) case, $\psi_j = \phi_{1,1}^j$). This causal solution implies that $\text{cov}\{\epsilon_s, Y_t\} = 0$ when $s > t$. When none of the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$ lies on the unit circle but one or more are outside of it, the unique stationary solution takes the form

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j},$$

where at least some of the ψ_j weights for $j < 0$ are nonzero (if all of the roots are outside of the unit circle, then $\psi_j = 0$ for all $j \geq 0$, and the solution is purely acausal). Acausality implies that we cannot guarantee $\text{cov}\{\epsilon_s, Y_t\}$ is zero when $s > t$.

When we have a time series that we regard as a realization of a portion $Y_0, Y_1, \ldots, Y_{N-1}$ of an AR(p) process, the question arises as to whether parameter estimates based upon this realization correspond to those for a causal (and hence stationary) AR process. Since causality implies stationarity (but not vice versa), causality is harder to satisfy than stationarity. Certain estimation procedures guarantee causality (e.g., the Yule–Walker estimator discussed in the next section – see Brockwell and Davis, 2016, section 5.1.1, which points to Brockwell and Davis, 1991, problem 8.3); others do not (e.g., least squares estimators – see Section 9.7). In cases where the parameter estimates correspond to an acausal stationary process, the estimated coefficients can still be substituted into Equation (446b) to produce an SDF estimator with all of the properties of a valid SDF (i.e., it is nonnegative everywhere, symmetric about the origin and integrates to a finite number); moreover, the argument that led us to deduce the form of an AR SDF assumes only stationarity and not causality (see the discussion following Equation (145a)). Additionally, for any SDF arising from an acausal stationary AR(p) process, there is a corresponding causal (and hence necessarily stationary) AR(p) process with *identically* the same SDF (see Exercise [9.1b] or Brockwell and Davis, 1991, section 4.4).

[2] We have stated the stationarity and causality conditions in terms of roots of the polynomial $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$, but an alternative formulation is to use $1 - \sum_{j=1}^{p} \phi_{p,j} z^{j}$ (see, e.g., Brockwell and Davis, 1991, section 4.4). In terms of this latter polynomial, the condition for causality is that all of its roots lie *outside* the unit circle; however, the stationarity condition is unchanged, namely, that none of the roots of the polynomial lies *on* the unit circle.

[3] Priestley (1981) relates the idea of causality to the concept of *asymptotic stationarity* (see his section 3.5.4). To understand the main idea behind this concept, consider an AR(1) process $Y_t = \phi_{1,1} Y_{t-1} + \epsilon_t$, for which the stationarity condition is $|\phi_{1,1}| \neq 1$. Given a realization of $\epsilon_0, \epsilon_1, \ldots$ of the white noise process $\{\epsilon_t\}$, suppose we define $\widetilde{Y}_0 = Z_0 + \epsilon_0$ and $\widetilde{Y}_t = \phi_{1,1} \widetilde{Y}_{t-1} + \epsilon_t$ for $t \geq 1$, where Z_0 is an RV with finite variance that is uncorrelated with $\{\epsilon_t\}$. The same argument leading to Equation (44a) says that

$$\widetilde{Y}_t = \phi_{1,1}^t Z_0 + \sum_{j=0}^{t-1} \phi_{1,1}^j \epsilon_{t-j}.$$

Since $\text{var}\{\widetilde{Y}_t\}$ depends on t, the process $\{\widetilde{Y}_t : t = 0, 1, \ldots\}$ is nonstationary; however, as t gets larger and larger, we can argue that, if $|\phi_{1,1}| < 1$ so that $\phi_{1,1}^t, \phi_{1,1}^{t+1}, \ldots$ become smaller and smaller, then $\widetilde{Y}_t$ resembles more and more

$$Y_t = \sum_{j=0}^{\infty} \phi_{1,1}^j \epsilon_{t-j},$$

which is the causal (and hence stationary) solution to $Y_t = \phi_{1,1} Y_{t-1} + \epsilon_t$; i.e., $\{\widetilde{Y}_t\}$ is asymptotically stationary. For an AR(1) process the condition for asymptotic stationarity is thus identical to the causality

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

condition, namely, $|\phi_{1,1}| < 1$. This correspondence continues to hold for other AR(p) processes: the condition for asymptotic stationarity is the same as the condition for causality, namely, that the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$ all lie inside the unit circle. Asymptotic stationarity justifies setting $Y_{-p} = Y_{-p+1} = \cdots = Y_{-1} = 0$ when using Equation (446a) to simulate an AR time series – after a suitable burn-in period of, say, length M, the influence of the initial settings for Y_t when $t < 0$ should be small, and, to a certain degree of accuracy, we can regard $Y_M, Y_{M+1}, \ldots, Y_{N+M-1}$ as a realization of length N from a causal (and hence stationary) AR process (see Section 11.1 for a method for simulating AR processes that avoids a burn-in period).

After his discussion of asymptotic stationarity, Priestley (1981) notes the existence of acausal stationary solutions, but then states that henceforth "... when we refer to the 'condition for stationarity' for an AR model, we shall mean the condition under which a stationary solution exists with Y_t expressed in terms of present and past ϵ_t's only" (see his p. 135). This redefinition of stationarity is widely used; however, its usage is potentially confusing because stationarity by this new definition excludes certain AR processes deemed to be stationary under the technically correct definition. Throughout this chapter, whenever we refer to a stationary AR model, we do *not* restrict ourselves to a causal model; when we do need to deal with just causal models, we will use the qualifier "causal (and hence stationary)".

[4] In Section 8.7 we defined the innovation variance for a purely nondeterministic stationary process $\{X_t\}$ to be $E\{(X_t - \widehat{X}_t)^2\}$, where $\widehat{X}_t$ is the best linear predictor of X_t given the infinite past X_{t-1}, $X_{t-2}, \ldots$ (see Equation (404a)). A causal (and hence stationary) AR(p) process $\{Y_t\}$ is purely nondeterministic, and the best linear predictor of Y_t given the infinite past is $\widehat{Y}_t = \sum_{j=1}^{p} \phi_{p,j} Y_{t-j}$; i.e., the predictor depends just on the p most recent RVs and not on the distant past $Y_{t-p-1}, Y_{t-p-2}, \ldots$. Since $E\{(Y_t - \widehat{Y}_t)^2\} = E\{\epsilon_t^2\} = \sigma_p^2$, the use of "innovation variance" here to describe σ_p^2 is consistent with its definition in Equation (404a).

[5] Following ideas presented in C&E [1] for Section 3.11, we can efficiently compute the AR(p) SDF of Equation (446b) over a grid of frequencies by using zero padding in conjunction with an FFT algorithm. To do so, let $N' \geq p+1$ be any integer that the algorithm deems to be acceptable, and define $\phi_{p,0} = -1$ and $\phi_{p,j} = 0$ for $j = p+1, \ldots, N'-1$; i.e., we pad the sequence $\phi_{p,0}, \phi_{p,1}, \ldots, \phi_{p,p}$ with $N' - (p+1)$ zeros. Ignoring Δ_t in Equation (91b), the DFT of the zero padded sequence is

$$G_k = \sum_{j=0}^{N'-1} \phi_{p,j} e^{-i2\pi kj/N'}, \quad k = 0, 1, \ldots, N'-1, \tag{449a}$$

and we have

$$\frac{\sigma_p^2 \, \Delta_t}{|G_k|^2} = \frac{\sigma_p^2 \, \Delta_t}{\left| 1 - \sum_{j=1}^{p} \phi_{p,j} e^{-i2\pi f'_k j \, \Delta_t} \right|^2} = S(f'_k), \tag{449b}$$

where $f'_k = k/(N' \, \Delta_t)$.

It should be noted that sharp peaks in the SDF might not be accurately represented if the grid of frequencies is not fine enough. C&E [3] for Section 10.12 discusses how to accurately locate such peaks, after which we can use Equation (446b) to check their accurate representation.

9.3 The Yule–Walker Equations

The oldest method of estimating the parameters for a causal (and hence stationary) AR(p) process $\{Y_t\}$ with zero mean and ACVS given by $s_\tau = E\{Y_{t+\tau} Y_t\}$ is based upon matching lagged moments, for which we need to express the parameters in terms of the ACVS. To do so, we first take Equation (446a) and multiply both sides of it by $Y_{t-\tau}$ to get

$$Y_{t-\tau} Y_t = \sum_{j=1}^{p} \phi_{p,j} Y_{t-\tau} Y_{t-j} + Y_{t-\tau} \epsilon_t. \tag{449c}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

If we take the expectation of both sides, we have, for $\tau > 0$,

$$s_\tau = \sum_{j=1}^{p} \phi_{p,j} s_{\tau-j}, \tag{450a}$$

where $E\{Y_{t-\tau}\epsilon_t\} = 0$ due to causality (as noted in the C&Es for the previous section, causality says that $Y_{t-\tau}$ can be written as an infinite linear combination of $\epsilon_{t-\tau}$, $\epsilon_{t-\tau-1}$, $\epsilon_{t-\tau-2}$, ..., but is uncorrelated with RVs in $\{\epsilon_t\}$ that occur after time $t-\tau$). Let $\tau = 1, 2, \ldots, p$ in Equation (450a) and recall that $s_{-j} = s_j$ to get the following p equations, known as the *Yule–Walker equations*:

$$\begin{aligned}
s_1 &= \phi_{p,1}s_0 + \phi_{p,2}s_1 + \cdots + \phi_{p,p}s_{p-1} \\
s_2 &= \phi_{p,1}s_1 + \phi_{p,2}s_0 + \cdots + \phi_{p,p}s_{p-2} \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\quad \vdots \\
s_p &= \phi_{p,1}s_{p-1} + \phi_{p,2}s_{p-2} + \cdots + \phi_{p,p}s_0
\end{aligned} \tag{450b}$$

or, in matrix notation,

$$\boldsymbol{\gamma}_p = \boldsymbol{\Gamma}_p \boldsymbol{\Phi}_p,$$

where $\boldsymbol{\gamma}_p \stackrel{\text{def}}{=} [s_1, s_2, \ldots, s_p]^T$; $\boldsymbol{\Phi}_p \stackrel{\text{def}}{=} [\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}]^T$; and

$$\boldsymbol{\Gamma}_p \stackrel{\text{def}}{=} \begin{bmatrix} s_0 & s_1 & \cdots & s_{p-1} \\ s_1 & s_0 & \cdots & s_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p-1} & s_{p-2} & \cdots & s_0 \end{bmatrix}. \tag{450c}$$

If the covariance matrix $\boldsymbol{\Gamma}_p$ is positive definite (as is always the case in practical applications), we can now solve for $\phi_{p,1}$, $\phi_{p,2}$, ..., $\phi_{p,p}$ in terms of the lag 0 to lag p values of the ACVS:

$$\boldsymbol{\Phi}_p = \boldsymbol{\Gamma}_p^{-1} \boldsymbol{\gamma}_p. \tag{450d}$$

Given a time series that is a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of *any* discrete parameter stationary process with zero mean and SDF $S(\cdot)$, one possible way to fit an AR(p) model to it is to replace s_τ in $\boldsymbol{\Gamma}_p$ and $\boldsymbol{\gamma}_p$ with the sample ACVS

$$\hat{s}_\tau^{(\mathrm{P})} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{t=0}^{N-|\tau|-1} X_{t+|\tau|} X_t$$

to produce estimates $\widetilde{\boldsymbol{\Gamma}}_p$ and $\widetilde{\boldsymbol{\gamma}}_p$. (How reasonable this procedure is for an arbitrary stationary process depends on how well it can be approximated by a stationary AR(p) process.) If the time series is not known to have zero mean (the usual case), we need to replace $\hat{s}_\tau^{(\mathrm{P})}$ with

$$\frac{1}{N} \sum_{t=0}^{N-|\tau|-1} (X_{t+|\tau|} - \overline{X})(X_t - \overline{X}),$$

where $\overline{X}$ is the sample mean. We noted in C&E [2] for Section 6.2 that, with the above form for $\hat{s}_\tau^{(\mathrm{P})}$ when $|\tau| \le N-1$ and with $\hat{s}_\tau^{(\mathrm{P})}$ set to zero when $|\tau| \ge N$, a realization of the sequence

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

$\{\hat{s}_\tau^{(\mathrm{P})}\}$ is positive definite if and only if the realizations of $X_0, \ldots, X_{N-1}$ are not all exactly the same (Newton, 1988, p. 165). This mild condition holds in all practical applications of interest. The positive definiteness of the sequence $\{\hat{s}_\tau^{(\mathrm{P})}\}$ in turn implies that the matrix $\widetilde{\boldsymbol{\Gamma}}_p$ is positive definite. Hence we can obtain estimates for AR(p) coefficients from

$$\widetilde{\boldsymbol{\Phi}}_p = \widetilde{\boldsymbol{\Gamma}}_p^{-1} \widetilde{\boldsymbol{\gamma}}_p. \tag{451a}$$

We are not quite done: we still need to estimate σ_p^2. To do so, let $\tau = 0$ in Equation (449c) and take expectations to get

$$s_0 = \sum_{j=1}^{p} \phi_{p,j} s_j + E\{Y_t \epsilon_t\}.$$

From the fact that $E\{Y_{t-j}\epsilon_t\} = 0$ for $j > 0$, it follows that

$$E\{Y_t \epsilon_t\} = E\left\{ \Big(\sum_{j=1}^{p} \phi_{p,j} Y_{t-j} + \epsilon_t \Big) \epsilon_t \right\} = \sigma_p^2$$

and hence

$$\sigma_p^2 = s_0 - \sum_{j=1}^{p} \phi_{p,j} s_j. \tag{451b}$$

This equation suggests that we estimate σ_p^2 by

$$\tilde{\sigma}_p^2 \stackrel{\mathrm{def}}{=} \hat{s}_0^{(\mathrm{P})} - \sum_{j=1}^{p} \tilde{\phi}_{p,j} \hat{s}_j^{(\mathrm{P})}. \tag{451c}$$

We call the estimators $\widetilde{\boldsymbol{\Phi}}_p$ and $\tilde{\sigma}_p^2$ the *Yule–Walker estimators* of the AR(p) parameters. These estimators depend only on the sample ACVS up to lag p and can be used to estimate the SDF of $\{X_t\}$ by

$$\hat{S}^{(\mathrm{YW})}(f) \stackrel{\mathrm{def}}{=} \frac{\tilde{\sigma}_p^2 \, \Delta_\mathrm{t}}{\left| 1 - \sum_{j=1}^{p} \tilde{\phi}_{p,j} \mathrm{e}^{-\mathrm{i}2\pi f j \, \Delta_\mathrm{t}} \right|^2}.$$

An important property of the Yule–Walker estimators is that the fitted AR(p) process has a theoretical ACVS that is *identical* to the sample ACVS up to lag p. This forced agreement implies that

$$\int_{-f_\mathcal{N}}^{f_\mathcal{N}} \hat{S}^{(\mathrm{YW})}(f) \mathrm{e}^{\mathrm{i}2\pi f \tau \, \Delta_\mathrm{t}} \, \mathrm{d}f = \hat{s}_\tau^{(\mathrm{P})}, \quad \tau = 0, 1, \ldots, p,$$

which is of questionable value: as we have seen in Chapter 7, the Fourier transform of $\{\hat{s}_\tau^{(\mathrm{P})}\}$ is the periodogram, which can suffer from severe bias (Sakai et al., 1979, discuss connections between $\hat{S}^{(\mathrm{YW})}(\cdot)$ and the periodogram). For $\tau = 0$, however, the above says that

$$\int_{-f_\mathcal{N}}^{f_\mathcal{N}} \hat{S}^{(\mathrm{YW})}(f) \, \mathrm{d}f = \hat{s}_0^{(\mathrm{P})}, \tag{451d}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

which is a useful property because it facilitates comparing the Yule–Walker SDF estimator with other estimators that also integrate to the sample variance. Under certain reasonable assumptions on a stationary process $\{X_t\}$, Berk (1974) has shown that $\hat{S}^{(\mathrm{YW})}(f)$ is a consistent estimator of $S(f)$ for all f if the order p of the approximating AR process is allowed to increase as the sample size N increases. Unfortunately, this result is not particularly informative if one has a time series of fixed length N.

Equations (450b) and (451b) can be combined to produce the so-called *augmented Yule–Walker equations*:

$$\begin{bmatrix} s_0 & s_1 & \dots & s_p \\ s_1 & s_0 & \dots & s_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_p & s_{p-1} & \dots & s_0 \end{bmatrix} \begin{bmatrix} 1 \\ -\phi_{p,1} \\ \vdots \\ -\phi_{p,p} \end{bmatrix} = \begin{bmatrix} \sigma_p^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The first row above follows from Equation (451b), while the remaining p rows are just transposed versions of Equations (450b). This formulation is sometimes useful for expressing the estimation problem so that all the AR(p) parameters can be found simultaneously using a routine designed for solving a Toeplitz system of equations (see C&E [3] for Section 9.4).

Given the ACVS of $\{X_t\}$ or its estimator up to lag p, Equations (450d) and (451a) formally require matrix inversions to obtain the values of the actual or estimated AR(p) coefficients. As we shall show in the next section, there is an interesting alternative way to relate these coefficients to the true or estimated ACVS that avoids matrix inversion by brute force and clarifies the relationship between AR SDF estimation and the so-called maximum entropy method (see Section 9.6).

Comments and Extensions to Section 9.3

[1] There is no reason to insist upon substituting the usual biased estimator of the ACVS into Equation (450d) to produce estimated values for the AR coefficients. Any other reasonable estimates for the ACVS can be used – the only requirement is that the estimated sequence be positive definite so that the matrix inversion in Equation (450d) can be done. For example, if the process mean is assumed to be 0, the direct spectral estimator $\hat{S}^{(\mathrm{D})}(\cdot)$ of Equation (186b) yields an estimate of the ACVS of the form

$$\hat{s}_\tau^{(\mathrm{D})} = \sum_{t=0}^{N-|\tau|-1} h_{t+|\tau|} X_{t+|\tau|} h_t X_t$$

(see Equation (188b)), where $\{h_t\}$ is the data taper used with $\hat{S}^{(\mathrm{D})}(\cdot)$. As is true for $\{\hat{s}_\tau^{(\mathrm{P})}\}$, realizations of the sequence $\{\hat{s}_\tau^{(\mathrm{D})}\}$ are always positive definite in practical applications. Since proper use of tapering ensures that the first-moment properties of $\hat{S}^{(\mathrm{D})}(\cdot)$ are better overall than those of the periodogram, it makes some sense to use $\{\hat{s}_\tau^{(\mathrm{D})}\}$ – the inverse Fourier transform of $\hat{S}^{(\mathrm{D})}(\cdot)$ – rather than the usual biased estimator $\{\hat{s}_\tau^{(\mathrm{P})}\}$ – the inverse Fourier transform of the periodogram. For an example, see C&E [2] for Section 9.4.

9.4 The Levinson–Durbin Recursions

The Levinson–Durbin recursions are an alternative way of solving for the AR coefficients in Equation (450b). Here we follow Papoulis (1985) and derive the equations that define the recursions by considering the following problem: given values of $X_{t-1}, \dots, X_{t-k}$ of a stationary process $\{X_t\}$ with zero mean, how can we predict the value of X_t? A mathematically tractable solution is to find that *linear* function of $X_{t-1}, X_{t-2}, \dots, X_{t-k}$, say

$$\overrightarrow{X}_t(k) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{k} \phi_{k,j} X_{t-j}, \tag{452}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

such that the *mean square linear prediction error*

$$\sigma_k^2 \stackrel{\text{def}}{=} E\left\{\left(X_t - \overrightarrow{X}_t(k)\right)^2\right\} = E\left\{\left(X_t - \sum_{j=1}^{k} \phi_{k,j} X_{t-j}\right)^2\right\} \tag{453a}$$

is minimized. We call $\overrightarrow{X}_t(k)$ the *best linear predictor* of X_t, given $X_{t-1}, \ldots, X_{t-k}$. We note the following.

[1] For reasons that will become clear in the next few paragraphs, we are purposely using the same symbols for denoting the coefficients in Equation (452) and the coefficients of an AR(k) process, which $\{X_t\}$ need *not* be. For the time being, the reader should regard Equation (452) as a new definition of $\phi_{k,j}$ – we will show that this agrees with our old definition if in fact $\{X_t\}$ is an AR(k) process (see the discussion concerning Equation (454b) and its connection to Equation (450a)).

[2] The quantity $\overrightarrow{X}_t(k)$ is a scalar – it is not a vector as the arrow over the X might suggest to readers familiar with textbooks on physics. The arrow is meant to connote "forward prediction."

[3] When the stationary process in question is an AR(p) process, then, following our conventions, we should replace X with Y on the right-hand side of Equation (453a). If we also replace k with p and then appeal to Equation (446a), the right-hand side of Equation (453a) becomes $\text{var}\{\epsilon_t\}$, and, just below Equation (446a), we defined σ_p^2 to be equal to this variance. Hence this previous definition for σ_p^2 is in agreement with the definition in Equation (453a).

[4] If we don't have any values prior to X_t that we can use to predict it, we take its prediction to be just its expected value, namely, $E\{X_t\} = 0$. Hence we augment Equations (452) and (453a), which assume $k \geq 1$, by taking

$$\overrightarrow{X}_t(0) \stackrel{\text{def}}{=} 0 \text{ and } \sigma_0^2 \stackrel{\text{def}}{=} E\left\{\left(X_t - \overrightarrow{X}_t(0)\right)^2\right\} = E\{X_t^2\} = s_0. \tag{453b}$$

If we denote the prediction error associated with the best linear predictor by

$$\overrightarrow{\epsilon}_t(k) \stackrel{\text{def}}{=} X_t - \overrightarrow{X}_t(k), \tag{453c}$$

we can then derive the following result. For any set of real-valued numbers $\psi_1, \ldots, \psi_k$, define

$$P_k(\psi_1, \ldots, \psi_k) = E\left\{\left(X_t - \sum_{l=1}^{k} \psi_l X_{t-l}\right)^2\right\}.$$

Since $P_k(\cdot)$ is a quadratic function of the ψ_l terms and since the best linear predictor is defined as that set of ψ_l terms that minimizes the above, we can find the $\phi_{k,l}$ terms by differentiating the above equation and setting the derivatives to zero:

$$\frac{\mathrm{d}P_k}{\mathrm{d}\psi_j} = -2E\left\{\left(X_t - \sum_{l=1}^{k} \psi_l X_{t-l}\right) X_{t-j}\right\} = 0, \qquad 1 \leq j \leq k. \tag{453d}$$

Since $\psi_j = \phi_{k,j}$ for all j at the solution and since

$$X_t - \sum_{l=1}^{k} \phi_{k,l} X_{t-l} = \overrightarrow{\epsilon}_t(k),$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Equation (453d) reduces to the so-called *orthogonality principle*, namely,

$$E\{\overrightarrow{\epsilon}_t(k)X_{t-j}\} = \operatorname{cov}\{\overrightarrow{\epsilon}_t(k), X_{t-j}\} = 0, \qquad 1 \le j \le k. \tag{454a}$$

In words, the orthogonality principle says that the prediction error is uncorrelated with all RVs utilized in the prediction. Note that, at the solution, $P_k(\phi_{k,1}, \ldots, \phi_{k,k}) = \sigma_k^2$.

If we rearrange Equations (453d), we are led to a series of equations that allows us to solve for $\phi_{k,l}$, namely,

$$E\{X_t X_{t-j}\} = \sum_{l=1}^{k} \phi_{k,l} E\{X_{t-l} X_{t-j}\}, \qquad 1 \le j \le k,$$

or, equivalently in terms of the ACVS,

$$s_j = \sum_{l=1}^{k} \phi_{k,l} s_{j-l}, \qquad 1 \le j \le k. \tag{454b}$$

Comparison of the above with Equation (450a), which generates the Yule–Walker equations, shows that they are identical! Thus, Equation (450d) here shows that, if the covariance matrix is positive definite, the $\phi_{k,l}$ terms are necessarily unique and hence $\overrightarrow{X}_t(k)$ is unique. This uniqueness leads to a useful corollary to the orthogonality principle, which is the subject of Exercise [9.5].

The reader should note that the $\phi_{k,l}$ terms of the best linear predictor are uniquely determined by the covariance properties of the process $\{X_t\}$ – we have not discussed so far the practical problem of estimating these coefficients for a process with an unknown covariance structure.

For the Yule–Walker equations for an AR(p) process, the innovation variance σ_p^2 can be related to values of the ACVS and the AR(p) coefficients (Equation (451b)). By an analogous argument, the mean square linear prediction error σ_k^2 can be expressed in terms of the ACVS and the $\phi_{k,j}$ terms:

$$\begin{aligned} \sigma_k^2 = E\{\overrightarrow{\epsilon}_t^2(k)\} &= E\Big\{\overrightarrow{\epsilon}_t(k)\Big(X_t - \overrightarrow{X}_t(k)\Big)\Big\} \\ &= E\{\overrightarrow{\epsilon}_t(k)X_t\} = E\Big\{\Big(X_t - \overrightarrow{X}_t(k)\Big)X_t\Big\} = s_0 - \sum_{j=1}^{k} \phi_{k,j} s_j, \end{aligned} \tag{454c}$$

where we have appealed to the orthogonality principle in order to go from the first line to the second. From the first expression on the second line, we note the important fact that, since both $\overrightarrow{\epsilon}_t(k)$ and X_t have zero mean,

$$\sigma_k^2 = E\{\overrightarrow{\epsilon}_t(k)X_t\} = \operatorname{cov}\{\overrightarrow{\epsilon}_t(k), X_t\}; \tag{454d}$$

i.e., the mean square linear prediction error is just the covariance between the prediction error $\overrightarrow{\epsilon}_t(k)$ and X_t, the quantity being predicted.

To summarize our discussion to this point, the Yule–Walker equations arise in two related problems:

[1] For a stationary AR(p) process $\{Y_t\}$ with zero mean, ACVS $\{s_\tau\}$ and coefficients $\phi_{p,1}$, …, $\phi_{p,p}$, the Yule–Walker equations relate the coefficients to the ACVS; the auxiliary

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Equation (451b) gives the innovation variance in terms of $\{s_\tau\}$ and the AR(p) coefficients. Sampling versions of these equations allow us to use an AR(p) process to approximate a time series that can be regarded as a portion of an arbitrary stationary process $\{X_t\}$.

[2] For a stationary process $\{X_t\}$ with zero mean and ACVS $\{s_\tau\}$, the Yule–Walker equations relate the coefficients of the best linear predictor of X_t, given the k most recent prior values, to the ACVS; The auxiliary Equation (454c) gives the mean square linear prediction error in terms of $\{s_\tau\}$ and the coefficients of the best linear predictor.

The fact that the two problems are related implies that, for the AR(p) process $\{Y_t\}$, the mean square linear prediction error of its pth-order linear predictor is identical to its innovation variance, both of which are denoted by σ_p^2.

Before we get to the heart of our derivation of the Levinson–Durbin recursions, we note the following seemingly trivial fact: if $\{X_t\}$ is a stationary process, then so is $\{X_{-t}\}$, the process with time reversed. Since

$$E\{X_{-t+\tau}X_{-t}\} = E\{X_{t+\tau}X_t\} = s_\tau,$$

both $\{X_t\}$ and $\{X_{-t}\}$ have the same ACVS. It follows from Equation (454b) that $\overleftarrow{X}_t(k)$, the best linear "predictor" of X_t given the next k future values, can be written as

$$\overleftarrow{X}_t(k) \stackrel{\text{def}}{=} \sum_{j=1}^{k} \phi_{k,j} X_{t+j},$$

where $\phi_{k,j}$ is *exactly* the same coefficient occurring in the best linear predictor of X_t, given the k most recent prior values. The orthogonality principle applied to the reversed process tells us that

$$E\{\overleftarrow{\epsilon}_t(k)X_{t+j}\} = \operatorname{cov}\{\overleftarrow{\epsilon}_t(k), X_{t+j}\} = 0, \;\; 1 \le j \le k, \;\; \text{where } \overleftarrow{\epsilon}_t(k) \stackrel{\text{def}}{=} X_t - \overleftarrow{X}_t(k).$$

From now on we refer to $\overrightarrow{X}_t(k)$ and $\overleftarrow{X}_t(k)$, respectively, as the *forward* and *backward* predictors of X_t of length k. We call $\overrightarrow{\epsilon}_t(k)$ and $\overleftarrow{\epsilon}_t(k)$ the corresponding *forward* and *backward prediction errors*. It follows from symmetry that

$$E\{\overleftarrow{\epsilon}_t^2(k)\} = E\{\overrightarrow{\epsilon}_t^2(k)\} = \sigma_k^2$$

and that the analog of Equation (454d) is

$$\sigma_k^2 = E\{\overleftarrow{\epsilon}_{t-k}(k)X_{t-k}\} = \operatorname{cov}\{\overleftarrow{\epsilon}_{t-k}(k), X_{t-k}\}. \tag{455a}$$

The Levinson–Durbin algorithm follows from an examination of the following equation:

$$\overrightarrow{\epsilon}_t(k) = \overrightarrow{\epsilon}_t(k-1) - \theta_k \overleftarrow{\epsilon}_{t-k}(k-1), \tag{455b}$$

where θ_k is a constant to be determined. This equation is by no means obvious, but it is clearly plausible: $\overrightarrow{\epsilon}_t(k)$ depends upon one more variable than $\overrightarrow{\epsilon}_t(k-1)$, namely, X_{t-k}, upon which $\overleftarrow{\epsilon}_{t-k}(k-1)$ obviously depends. To prove Equation (455b), we first note that, by the orthogonality principle,

$$E\{\overrightarrow{\epsilon}_t(k-1)X_{t-j}\} = 0 \;\text{ and }\; E\{\overleftarrow{\epsilon}_{t-k}(k-1)X_{t-j}\} = 0$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

for $j = 1, \ldots, k-1$, and, hence, for *any* constant θ_k,

$$E\left\{[\overrightarrow{\epsilon}_t(k-1) - \theta_k \overleftarrow{\epsilon}_{t-k}(k-1)]\, X_{t-j}\right\} = 0, \qquad j = 1, \ldots, k-1. \tag{456a}$$

The equation above will also hold for $j = k$ if we let

$$\theta_k = \frac{E\{\overrightarrow{\epsilon}_t(k-1)X_{t-k}\}}{E\{\overleftarrow{\epsilon}_{t-k}(k-1)X_{t-k}\}} = \frac{E\{\overrightarrow{\epsilon}_t(k-1)X_{t-k}\}}{\sigma^2_{k-1}} \tag{456b}$$

(this follows from Equation (455a)). With this choice of θ_k and with

$$d_t(k) \stackrel{\text{def}}{=} \overrightarrow{\epsilon}_t(k-1) - \theta_k \overleftarrow{\epsilon}_{t-k}(k-1),$$

we have

$$E\{d_t(k)X_{t-j}\} = 0, \qquad 1 \le j \le k.$$

The corollary to the orthogonality principle stated in Exercise [9.5] now tells us that $d_t(k)$ is in fact the same as $\overrightarrow{\epsilon}_t(k)$, thus completing the proof of Equation (455b).

For later use, we note the time-reversed version of Equation (455b), namely,

$$\overleftarrow{\epsilon}_{t-k}(k) = \overleftarrow{\epsilon}_{t-k}(k-1) - \theta_k \overrightarrow{\epsilon}_t(k-1). \tag{456c}$$

We are now ready to extract the Levinson–Durbin recursions. It readily follows from Equation (455b) that

$$\overrightarrow{X}_t(k) = \overrightarrow{X}_t(k-1) + \theta_k\Big(X_{t-k} - \overleftarrow{X}_{t-k}(k-1)\Big);$$

the definitions of $\overrightarrow{X}_t(k)$, $\overrightarrow{X}_t(k-1)$ and $\overleftarrow{X}_{t-k}(k-1)$ further yield

$$\sum_{j=1}^{k} \phi_{k,j} X_{t-j} = \sum_{j=1}^{k-1} \phi_{k-1,j} X_{t-j} + \theta_k\Big(X_{t-k} - \sum_{j=1}^{k-1} \phi_{k-1,j} X_{t-k+j}\Big).$$

Equating coefficients of X_{t-j} (and appealing to the fundamental theorem of algebra) yields

$$\phi_{k,j} = \phi_{k-1,j} - \theta_k \phi_{k-1,k-j}, \qquad 1 \le j \le k-1, \text{ and } \phi_{k,k} = \theta_k.$$

Given $\phi_{k-1,1}, \ldots, \phi_{k-1,k-1}$ and $\phi_{k,k}$, we can therefore compute $\phi_{k,1}, \ldots, \phi_{k,k-1}$. Moreover, $\phi_{k,k}$ can be expressed using $\phi_{k-1,1}, \ldots, \phi_{k-1,k-1}$ and the ACVS up to lag k. To see this, reconsider Equation (456b) with θ_k replaced by $\phi_{k,k}$:

$$\begin{aligned}\phi_{k,k} = \frac{E\{\overrightarrow{\epsilon}_t(k-1)X_{t-k}\}}{\sigma^2_{k-1}} &= \frac{E\left\{(X_t - \sum_{j=1}^{k-1} \phi_{k-1,j} X_{t-j})X_{t-k}\right\}}{\sigma^2_{k-1}} \\ &= \frac{s_k - \sum_{j=1}^{k-1} \phi_{k-1,j} s_{k-j}}{\sigma^2_{k-1}}.\end{aligned}$$

We have now completed our derivation of what is known in the literature as the *Levinson–Durbin recursions* (sometimes called the *Levinson recursions* or the *Durbin–Levinson recursions*), which we can summarize as follows. Suppose $\phi_{k-1,1}, \ldots, \phi_{k-1,k-1}$ and σ^2_{k-1} are known. We can calculate $\phi_{k,1}, \ldots, \phi_{k,k}$ and σ^2_k using the following three equations:

$$\phi_{k,k} = \frac{s_k - \sum_{j=1}^{k-1} \phi_{k-1,j} s_{k-j}}{\sigma^2_{k-1}}; \tag{456d}$$

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{k,k}\phi_{k-1,k-j}, \quad 1 \le j \le k-1; \tag{456e}$$

$$\sigma^2_k = s_0 - \sum_{j=1}^{k} \phi_{k,j} s_j. \tag{456f}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

We can initiate the recursions by solving Equations (454b) and (454c) explicitly for the case $k = 1$:

$$\phi_{1,1} = s_1/s_0 \text{ and } \sigma_1^2 = s_0 - \phi_{1,1}s_1. \tag{457a}$$

Since we defined σ_0^2 to be s_0 (see Equation (453b)), we can also think of $\phi_{1,1}$ as coming from Equation (456d) if we define the summation from $j = 1$ to 0 to be equal to zero.

There is an important variation on the Levinson–Durbin recursions (evidently due to Burg, 1975, p. 14). The difference between the two recursions is only in Equation (456f) for updating σ_k^2, but it is noteworthy for three reasons: first, it is a numerically better way of calculating σ_k^2 because it is not so sensitive to rounding errors; second, it requires fewer numerical operations and actually speeds up the recursions slightly; and third, it emphasizes the central role of $\phi_{k,k}$, the so-called *kth-order partial autocorrelation coefficient* (see C&E [4]). To derive the alternative to Equation (456f), we multiply both sides of Equation (455b) by X_t, recall that $\theta_k = \phi_{k,k}$, and take expectations to get

$$E\{\overrightarrow{\epsilon}_t(k)X_t\} = E\{\overrightarrow{\epsilon}_t(k-1)X_t\} - \phi_{k,k}E\{\overleftarrow{\epsilon}_{t-k}(k-1)X_t\}.$$

Using Equation (454d) we can rewrite the above as

$$\sigma_k^2 = \sigma_{k-1}^2 - \phi_{k,k}E\{\overleftarrow{\epsilon}_{t-k}(k-1)X_t\}.$$

Now

$$\begin{aligned} E\{\overleftarrow{\epsilon}_{t-k}(k-1)X_t\} &= E\Big\{\Big(X_{t-k} - \sum_{j=1}^{k-1}\phi_{k-1,j}X_{t-k+j}\Big)X_t\Big\} \\ &= s_k - \sum_{j=1}^{k-1}\phi_{k-1,j}s_{k-j} = \phi_{k,k}\sigma_{k-1}^2, \end{aligned}$$

with the last equality following from Equation (456d). An alternative to Equation (456f) is

$$\sigma_k^2 = \sigma_{k-1}^2(1 - \phi_{k,k}^2). \tag{457b}$$

The Levinson–Durbin recursions can thus be taken to be Equations (456d) and (456e) in combination with the above. Note that, since Equation (457a) says that $s_0 - \phi_{1,1}s_1 = s_0(1 - \phi_{1,1}^2)$, Equation (457b) for $k = 1$ is consistent with Equation (457a) when we make use of the definition $\sigma_0^2 = s_0$

Note that Equation (457b) tells us that, if $\sigma_{k-1}^2 > 0$, we must have $|\phi_{k,k}| \leq 1$ to ensure that σ_k^2 is nonnegative and, if in fact $|\phi_{k,k}| = 1$, the mean square linear prediction error σ_k^2 is 0. This implies that we could predict the process perfectly in the mean square sense with a kth-order linear predictor. Since a nontrivial linear combination of the RVs of the process $\{X_t\}$ thus has zero variance, the covariance matrix for $\{X_t\}$ is in fact positive semidefinite instead of positive definite (however, this cannot happen if $\{X_t\}$ is a purely continuous stationary process with nonzero variance; i.e., the derivative of its integrated spectrum (the SDF) exists – see Papoulis, 1985, for details).

We can now summarize explicitly the Levinson–Durbin recursive solution to the Yule–Walker equations for estimating the parameters of an AR(p) model from a sample ACVS. Although this method avoids the matrix inversion in Equation (451a), the two solutions are necessarily identical: the Levinson–Durbin recursions simply take advantage of the Toeplitz structure of $\widetilde{\boldsymbol{\Gamma}}_p$ (see the discussion following Equation (29a)) to solve the problem more efficiently than brute force matrix inversion can. On a digital computer, however, the two solutions might be annoyingly different due to the vagaries of rounding error. We begin by solving Equations (451a) and (451c) explicitly for an AR(1) model to get

$$\tilde{\phi}_{1,1} = \hat{s}_1^{(\mathrm{P})}/\hat{s}_0^{(\mathrm{P})} \text{ and } \tilde{\sigma}_1^2 = \hat{s}_0^{(\mathrm{P})} - \tilde{\phi}_{1,1}\hat{s}_1^{(\mathrm{P})} = \hat{s}_0^{(\mathrm{P})}(1 - \tilde{\phi}_{1,1}^2).$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

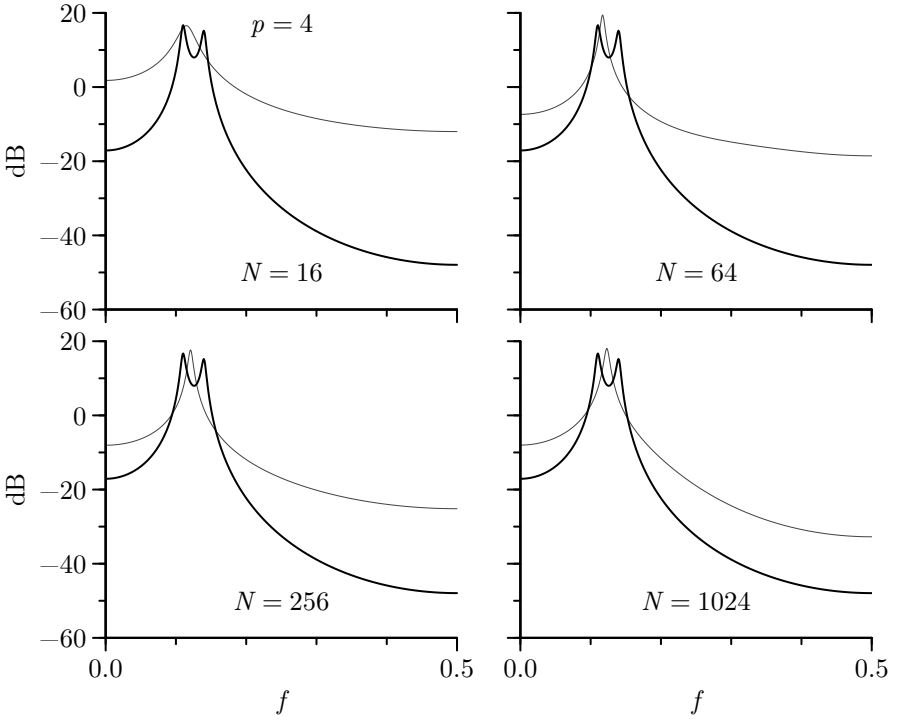

Figure 458 Yule–Walker AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e) (the process is defined in Equation (35a)). The thick curve on each plot is the true SDF.

For $k = 2, \ldots, p$, we then recursively evaluate

$$\tilde{\phi}_{k,k} = \frac{\hat{s}_k^{(\mathrm{P})} - \sum_{j=1}^{k-1} \tilde{\phi}_{k-1,j} \hat{s}_{k-j}^{(\mathrm{P})}}{\tilde{\sigma}_{k-1}^2}; \tag{458a}$$

$$\tilde{\phi}_{k,j} = \tilde{\phi}_{k-1,j} - \tilde{\phi}_{k,k} \tilde{\phi}_{k-1,k-j}, \quad 1 \le j \le k-1; \tag{458b}$$

$$\tilde{\sigma}_k^2 = \tilde{\sigma}_{k-1}^2 (1 - \tilde{\phi}_{k,k}^2) \tag{458c}$$

to obtain finally $\tilde{\phi}_{p,1}, \ldots, \tilde{\phi}_{p,p}$ and $\tilde{\sigma}_p^2$.

As an example of autoregressive SDF estimation, we reconsider the time series shown in Figure 34(e). This series is a realization of the AR(4) process defined in Equation (35a) – this process has been cited in the literature as posing a difficult case for SDF estimation. Figure 458 shows the result of using the Yule–Walker method to fit an AR(4) model to the first 16 values in this time series, the first 64, the first 256 and finally the entire series (1024 values). In each case we calculated the biased estimator of the ACVS from the appropriate segment of data and used it as input to the Levinson–Durbin recursions. In each of the plots the thin curve is the SDF corresponding to the fitted AR(4) model, whereas the thick curve is the true AR(4) SDF for the process from which the time series was drawn. We see that the SDF estimates improve with increasing sample size N, but that there is still significant deviation from the true SDF even for $N = 1024$ – particularly in the region of the twin peaks, which collapse incorrectly to a single peak in the estimated SDFs.

Figure 459 shows the effect on the spectral estimates of increasing the order of the fitted model to 8. Although the process that generated this time series is an AR(4) process,

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

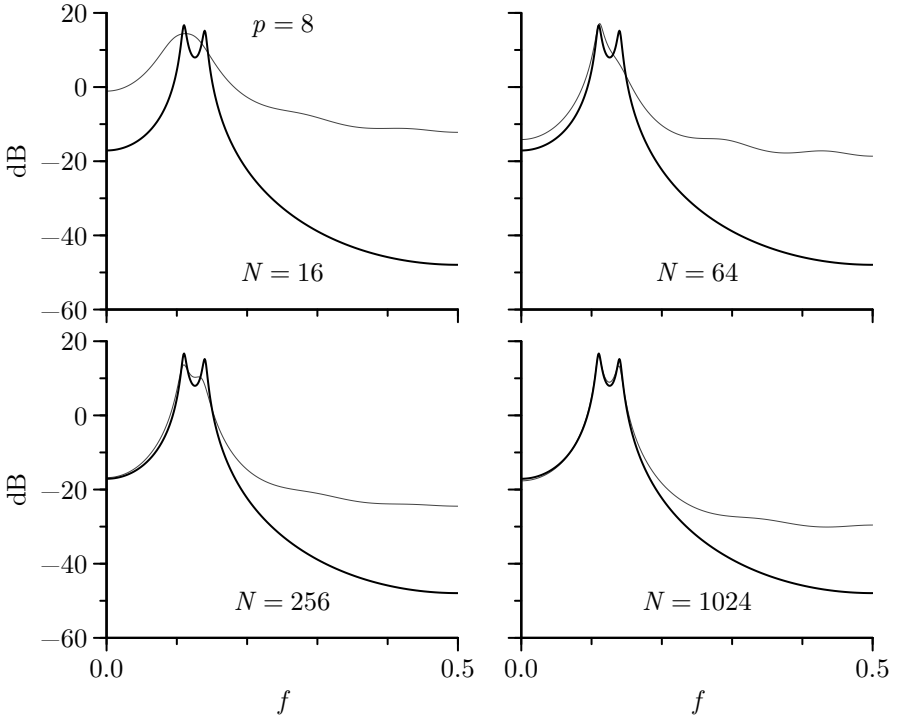

Figure 459 Yule–Walker AR(8) SDF estimates (see Figure 458).

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$\phi_{4,j}$	2.7607	−3.8106	2.6535	−0.9238	0	0	0	0
$\tilde{\phi}_{4,j}$	2.0218	−2.1202	1.0064	−0.2808	0	0	0	0
$\tilde{\phi}_{8,j}$	1.6144	−1.2381	−0.1504	0.1520	0.0224	−0.1086	−0.2018	0.0383
$\lvert\tilde{\phi}_{4,j} - \phi_{4,j}\rvert$	0.7389	1.6904	1.6471	0.6430	0	0	0	0
$\lvert\tilde{\phi}_{8,j} - \phi_{4,j}\rvert$	1.1463	2.5725	2.8039	1.0758	0.0224	0.1086	0.2018	0.0383

Table 459 Comparison of Yule–Walker AR(4) and AR(8) coefficient estimates based on realization of the AR(4) process in Figure 34(e) (the process is defined in Equation (35a)). Here all $N = 1024$ values of the time series are used. The top line shows the true AR coefficients; the next two lines, the AR(4) and AR(8) coefficient estimates; and the last two lines show the absolute differences between the estimates and the true coefficient values. The bottom right-hand plots of Figures 458 and 459 show the corresponding SDF estimates.

we generally get a much better fit to the true SDF by using a higher order model for this example – particularly in the low frequency portion of the SDF and around the twin peaks. Interestingly enough, the estimated coefficients for the AR(4) and AR(8) models superficially suggest otherwise. The top line of Table 459 shows the coefficients for the AR(4) process of Equation (35a) padded with four zeros to create an artificial AR(8) process. The next two lines show the Yule–Walker AR(4) and AR(8) coefficient estimates for the $N = 1024$ case (the corresponding SDF estimates are shown in the lower right-hand plots of Figures 458 and 459). The bottom two lines show the absolute differences between the estimated and true coefficients. These differences are uniformly *larger* for the fitted AR(8) coefficients than

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

for the AR(4), even though the AR(8) SDF estimate is visually superior to the AR(4)! The improvement in the SDF estimate is evidently not due to better estimates on a coefficient by coefficient basis, but rather is tied to interactions amongst the coefficients.

Comments and Extensions to Section 9.4

[1] The Levinson–Durbin recursions allow us to build up the coefficients for the one-step-ahead best linear predictor of order k in terms of the order $k-1$ coefficients (combined with the order $k-1$ mean square linear prediction error and the ACVS up to lag k). It is also possible to go in the other direction: given the order k coefficients, we can determine the corresponding order $k-1$ quantities. To do so, note that we can write Equation (456e) both as

$$\phi_{k-1,j} = \phi_{k,j} + \phi_{k,k}\phi_{k-1,k-j} \quad \text{and} \quad \phi_{k-1,k-j} = \phi_{k,k-j} + \phi_{k,k}\phi_{k-1,j}$$

for $1 \le j \le k-1$. If, for $\phi_{k-1,k-j}$ in the left-hand equation, we substitute its value in the right-hand equation and solve for $\phi_{k-1,j}$, we get the order $k-1$ coefficients in terms of the order k coefficients:

$$\phi_{k-1,j} = \frac{\phi_{k,j} + \phi_{k,k}\phi_{k,k-j}}{1-\phi_{k,k}^2}, \quad 1 \le j \le k-1. \tag{460a}$$

We can invert Equation (457b) to get the order $k-1$ mean square linear prediction error:

$$\sigma_{k-1}^2 = \frac{\sigma_k^2}{1-\phi_{k,k}^2}. \tag{460b}$$

One use for the step-down procedure is to generate the ACVS $\{s_\tau\}$ for an AR(p) process given its $p+1$ parameters $\phi_{p,1}, \ldots, \phi_{p,p}$ and σ_p^2. To do so, we apply the procedure starting with the $\phi_{p,j}$ coefficients to obtain, after $p-1$ iterations,

$$\begin{aligned} &\phi_{p-1,1}, \phi_{p-1,2}, \ldots, \phi_{p-1,p-1} \\ &\quad\vdots \\ &\phi_{2,1}, \phi_{2,2} \\ &\phi_{1,1}. \end{aligned} \tag{460c}$$

Next we take σ_p^2 and use Equation (460b) first with $\phi_{p,p}$ and then with the already obtained $\phi_{p-1,p-1}$, $\ldots$, $\phi_{2,2}$, $\phi_{1,1}$ to get $\sigma_{p-1}^2, \sigma_{p-2}^2, \ldots, \sigma_1^2, \sigma_0^2$. Noting that $s_0 = \sigma_0^2$, that $s_1 = \phi_{1,1}s_0$ (this comes from Equation (454b) upon setting $j = k = 1$) and that Equation (456d) can be rewritten as

$$s_k = \phi_{k,k}\sigma_{k-1}^2 + \sum_{j=1}^{k-1} \phi_{k-1,j}s_{k-j},$$

we have a recursive scheme for obtaining the ACVS:

$$\begin{aligned} s_0 &= \sigma_0^2 \\ s_1 &= \phi_{1,1}s_0 \\ s_2 &= \phi_{2,2}\sigma_1^2 + \phi_{1,1}s_1 \\ &\vdots \\ s_{p-1} &= \phi_{p-1,p-1}\sigma_{p-2}^2 + \sum_{j=1}^{p-2} \phi_{p-2,j}s_{p-1-j} \\ s_\tau &= \sum_{j=1}^{p} \phi_{p,j}s_{\tau-j}, \quad \tau = p, p+1, \ldots, \end{aligned} \tag{460d}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

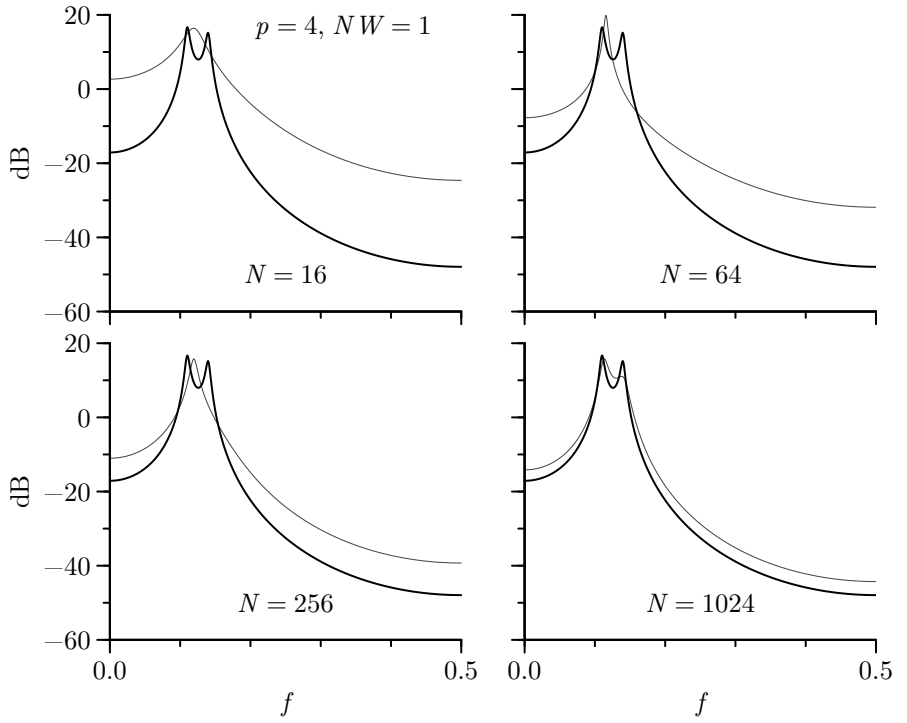

Figure 461 Yule–Walker AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e). The thick curves show the true SDF. Instead of the usual biased ACVS estimator, the ACVS estimator $\{\hat{s}_\tau^{(\mathrm{D})}\}$ corresponding to a direct spectral estimator with an $NW = 1$ Slepian data taper was used.

where the final equation above is the same as Equation (450a). Exercise [9.6] is to use the above to determine the ACVSs for the AR(2) process of Equation (34) and the AR(4) process of Equation (34).

Two other uses for the step-down procedure are to formulate an efficient algorithm for simulating Gaussian ARMA(p, q) processes (see Section 11.1) and to determine if the coefficients $\phi_{p,1}, \ldots, \phi_{p,p}$ for an AR(p) process correspond to those for a causal (and hence stationary) process (see C&E [5]).

[2] The Yule–Walker estimator as usually defined uses the biased ACVS estimator $\{\hat{s}_\tau^{(\mathrm{P})}\}$; however, as noted in the C&Es for Section 9.3, there is no reason why we cannot use other positive definite estimates of the ACVS such as $\{\hat{s}_\tau^{(\mathrm{D})}\}$. Figures 461 and 462 illustrate the possible benefits of doing so (see also Zhang, 1992). These show SDF estimates for the same data used in Figures 458 and 459, but now the Yule–Walker estimator uses $\{\hat{s}_\tau^{(\mathrm{D})}\}$ corresponding to a direct spectral estimator with $NW = 1$ (in Figure 461) and $NW = 2$ (in Figure 462) Slepian data tapers. The improvements over the usual Yule–Walker estimates are dramatic. (If we increase the degree of tapering to, say, an $NW = 4$ Slepian taper, the estimates deteriorate slightly, but the maximum difference over all frequencies between these estimates and the corresponding $NW = 2$ estimates in Figure 462 is less than 3 dB. An estimate with excessive tapering is thus still much better than the usual Yule–Walker estimate.)

[3] Note that we did *not* need to assume Gaussianity in order to derive the Levinson–Durbin recursions. The recursions are also *not* tied explicitly to AR(p) processes, since they can also be used to find the coefficients of the best linear predictor of X_t, given the p prior values of the process, when the only assumption on $\{X_t\}$ is that it is a discrete parameter stationary process with zero mean. The recursions are simply an efficient method of solving a system of equations that possesses a Toeplitz structure. The $p \times p$ matrix associated with such a system (for example, $\boldsymbol{\Gamma}_p$ in Equation (450c)) consists of at most p distinct elements $(s_0, \ldots, s_{p-1}$ in this example). A computational analysis of the recursions shows that they require $O(p^2)$ or fewer operations to solve the equations instead of the $O(p^3)$ operations required

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

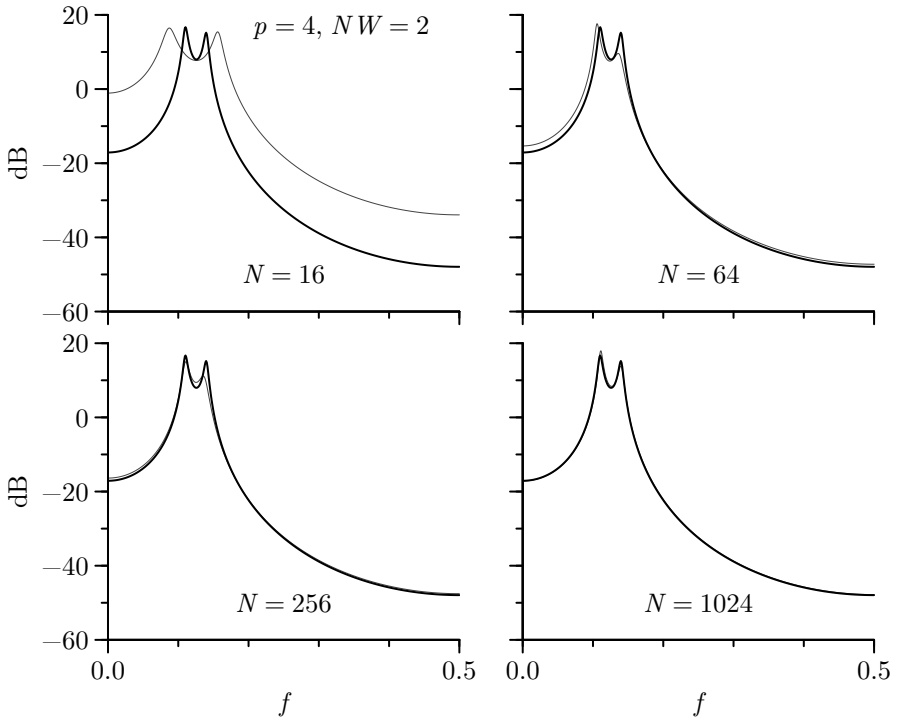

Figure 462 Yule–Walker AR(4) SDF estimates using the ACVS estimate $\{\hat{s}_\tau^{(\mathrm{D})}\}$ corresponding to a direct spectral estimator with an $NW = 2$ Slepian data taper.

by conventional methods using direct matrix inversion. The recursions also require less storage space on a computer (being proportional to p rather than p^2). Bunch (1985) gives an excellent discussion of Toeplitz matrices, their numerical properties and important refinements to the Levinson–Durbin recursions. Of particular interest, he discusses other algorithms for solving Toeplitz systems of equations that require just $O(p\ \log^2(p))$ operations. In practice, however, for p approximately less than 2000 (a rather high order for an AR model!), the Levinson–Durbin recursions are still the faster method because these alternatives require much more computer code. Morettin (1984) discusses the application of the recursions to other problems in time series analysis.

[4] The sequence $\{\phi_{k,k}\}$ indexed by k is known as the *partial autocorrelation sequence* (PACS) or the *reflection coefficient sequence*. The former terminology arises from the fact that $\phi_{k,k}$ is the correlation between X_t and X_{t-k} after they have been "adjusted" by the intervening $k-1$ values of the process. The adjustment takes the form of subtracting off predictions based upon $X_{t-k+1}, \ldots, X_{t-1}$; i.e., the adjusted values are

$$X_t - \overrightarrow{X}_t(k-1) = \overrightarrow{\epsilon}_t(k-1) \ \text{ and } \ X_{t-k} - \overleftarrow{X}_{t-k}(k-1) = \overleftarrow{\epsilon}_{t-k}(k-1) \tag{462a}$$

(for $k = 1$ we define $\overrightarrow{X}_t(0) = 0$ and $\overleftarrow{X}_{t-1}(0) = 0$). The claim is thus that

$$\frac{\operatorname{cov}\{\overrightarrow{\epsilon}_t(k-1), \overleftarrow{\epsilon}_{t-k}(k-1)\}}{\left(\operatorname{var}\{\overrightarrow{\epsilon}_t(k-1)\}\ \operatorname{var}\{\overleftarrow{\epsilon}_{t-k}(k-1)\}\right)^{1/2}} = \phi_{k,k} \tag{462b}$$

(Exercise [9.7] is to show that this is true).

Ramsey (1974) investigates properties of the PACS. He shows that, if $\phi_{p,p} \neq 0$ and $\phi_{k,k} = 0$ for all $k > p$ for a Gaussian stationary process, the process is necessarily an AR(p) process. Equation (457b) then tells us that

$$\sigma_{p-1}^2 > \sigma_p^2 = \sigma_{p+1}^2 = \cdots. \tag{462c}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

We use this fact later on to motivate a criterion for selecting the order of an AR model for a time series (see Equation (493a)).

[5] As noted in Section 9.2, the autoregressive process $\{Y_t\}$ of Equation (446a) is causal (and hence stationary) if the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{-j}$ all lie *inside* the unit circle (or, equivalently, if the roots of $1 - \sum_{j=1}^{p} \phi_{p,j} z^{j}$ all lie *outside* the unit circle). To assess causality for large p, we must compute the roots of a high-dimensional polynomial, which can be tricky due to the vagaries of finite-precision arithmetic. An alternative method for determining if an AR process is causal uses the step-down Levinson–Durbin recursions of Equation (460a). The idea is to use these numerically stable recursions to obtain what would constitute the partial autocorrelations $\phi_{p-1,p-1}$, $\phi_{p-2,p-2}$, …, $\phi_{1,1}$ for a causal process. If $|\phi_{k,k}| < 1$ for $k = 1, \ldots, p$, then the AR(p) process is causal (Papoulis, 1985; Newton, 1988); on the other hand, if either $|\phi_{p,p}| \geq 1$ or if the recursions yield $|\phi_{k,k}| \geq 1$ for some $k < p$, we are dealing with either a nonstationary process or an acausal stationary process. As examples, Exercise [9.9] lists six different AR(4) processes whose causality – or lack thereof – can be determined using the step-down recursions.

[6] A more general way to predict X_t given the k most recent prior values of $\{X_t\}$ is to find that function of $X_{t-1}, X_{t-2}, \ldots, X_{t-k}$ – call it $g(X_{t-1}, \ldots, X_{t-k})$ – such that the *mean square prediction error*

$$M_k \stackrel{\text{def}}{=} E\{[X_t - g(X_{t-1}, \ldots, X_{t-k})]^2\}$$

is minimized. The resulting form of $g(\cdot)$ is appealing (for a derivation of this result, see Priestley, 1981, p. 76):

$$g(X_{t-1}, \ldots, X_{t-k}) = E\{X_t | X_{t-1}, \ldots, X_{t-k}\};$$

i.e., under the mean square error criterion, the predictor of X_t, given $X_{t-1}, \ldots, X_{t-k}$, is simply the conditional mean of X_t, given $X_{t-1}, \ldots, X_{t-k}$. We call $g(\cdot)$ the *best predictor* of X_t, given $X_{t-1}, \ldots, X_{t-k}$. As simple and appealing as $g(\cdot)$ is, it is unfortunately mathematically intractable in many cases of interest. An important exception is the case of a stationary Gaussian process $\{G_t\}$, for which the best predictor is identical to the best linear predictor:

$$E\{G_t \mid G_{t-1}, \ldots, G_{t-k}\} = \overrightarrow{G}_t(k) \stackrel{\text{def}}{=} \sum_{j=1}^{k} \phi_{k,j} G_{t-j}.$$

For a non-Gaussian process, the best predictor need not be linear. Moreover, for these processes, the symmetry between the forward and backward best *linear* predictors of X_t does not necessarily carry through to the forward and backward best predictors. Rosenblatt (1985, p. 52) gives an example of a rather extreme case of asymmetry, in which the best forward predictor corresponds to the best linear predictor and has a prediction error variance of 4, whereas the best backward predictor is nonlinear and has a prediction error variance of 0 (i.e., it predicts perfectly with probability 1)!

[7] It is also possible to derive the Levinson–Durbin recursions using a direct linear algebra argument. As before, let us write the Yule–Walker equations in matrix form

$$\boldsymbol{\Gamma}_p \boldsymbol{\Phi}_p = \boldsymbol{\gamma}_p, \tag{463}$$

where $\boldsymbol{\gamma}_p = \left[s_1, s_2, \ldots, s_p\right]^T$; $\boldsymbol{\Phi}_p = \left[\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}\right]^T$; and $\boldsymbol{\Gamma}_p$ is as in Equation (450c). The trick is to solve for $\boldsymbol{\Phi}_{p+1}$ using the solution $\boldsymbol{\Phi}_p$. Suppose for the moment that it is possible to express $\phi_{p+1,j}$ for $j \leq p$ in the following manner:

$$\phi_{p+1,j} = \phi_{p,j} - \phi_{p+1,p+1} \theta_{p,j}.$$

Let $\boldsymbol{\Theta}_p = \left[\theta_{p,1}, \theta_{p,2}, \ldots, \theta_{p,p}\right]^T$ be a p-dimensional vector to be determined. With this recursion, Equation (463) becomes the following for order $p + 1$:

$$\begin{bmatrix} & & & s_p \\ & \boldsymbol{\Gamma}_p & & \vdots \\ & & & s_1 \\ s_p & \ldots & s_1 & s_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_p - \phi_{p+1,p+1} \boldsymbol{\Theta}_p \\ \phi_{p+1,p+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_p \\ s_{p+1} \end{bmatrix}.$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

If we separate the first p equations from the last one, we have

$$\boldsymbol{\Gamma}_p\boldsymbol{\Phi}_p - \phi_{p+1,p+1}\boldsymbol{\Gamma}_p\boldsymbol{\Theta}_p + \phi_{p+1,p+1}\begin{bmatrix} s_p \\ \vdots \\ s_1 \end{bmatrix} = \boldsymbol{\gamma}_p \tag{464a}$$

and

$$[\, s_p \quad \dots \quad s_1 \,]\left(\boldsymbol{\Phi}_p - \phi_{p+1,p+1}\boldsymbol{\Theta}_p\right) + s_0\phi_{p+1,p+1} = s_{p+1}. \tag{464b}$$

If we use the fact that $\boldsymbol{\Phi}_p$ satisfies Equation (463), we conclude that the following condition is sufficient for Equation (464a) to hold:

$$\boldsymbol{\Gamma}_p\boldsymbol{\Theta}_p = \begin{bmatrix} s_p \\ \vdots \\ s_1 \end{bmatrix}.$$

If we chose $\theta_{p,j} = \phi_{p,p+1-j}$, then the condition above is equivalent to

$$\begin{bmatrix} s_0 & s_1 & \dots & s_{p-1} \\ s_1 & s_0 & \dots & s_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p-1} & s_{p-2} & \dots & s_0 \end{bmatrix}\begin{bmatrix} \phi_{p,p} \\ \phi_{p,p-1} \\ \vdots \\ \phi_{p,1} \end{bmatrix} = \begin{bmatrix} s_p \\ s_{p-1} \\ \vdots \\ s_1 \end{bmatrix}.$$

If we reverse the order of the rows in the equation above, we have that it is equivalent to

$$\begin{bmatrix} s_{p-1} & s_{p-2} & \dots & s_0 \\ s_{p-2} & s_{p-3} & \dots & s_1 \\ \vdots & \vdots & \ddots & \vdots \\ s_0 & s_1 & \dots & s_{p-1} \end{bmatrix}\begin{bmatrix} \phi_{p,p} \\ \phi_{p,p-1} \\ \vdots \\ \phi_{p,1} \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_p \end{bmatrix}.$$

If we now reverse the order of the columns, we get Equation (463), which is assumed to hold. Hence our choice of $\boldsymbol{\Theta}_p$ satisfies Equation (464a). Equation (464b) can be satisfied by solving for $\phi_{p+1,p+1}$. In summary, if $\boldsymbol{\Phi}_p$ satisfies Equation (463),

$$\phi_{p+1,p+1} = \frac{s_{p+1} - \sum_{j=1}^{p}\phi_{p,j}s_{p+1-j}}{s_0 - \sum_{j=1}^{p}\phi_{p,j}s_j};$$

$$\phi_{p+1,j} = \phi_{p,j} - \phi_{p+1,p+1}\phi_{p,p+1-j}$$

for $j = 1, \dots, p$, then $\boldsymbol{\Gamma}_{p+1}\boldsymbol{\Phi}_{p+1} = \boldsymbol{\gamma}_{p+1}$.

[8] The Levinson–Durbin recursions are closely related to the *modified Cholesky decomposition* of a positive definite covariance matrix $\boldsymbol{\Gamma}_N$ for a portion $\boldsymbol{X} = [X_0, X_1, \dots, X_{N-1}]^T$ of a stationary process (Therrien, 1983; Newton, 1988, section A.1.1; Golub and Van Loan, 2013, theorem 4.1.3). Such a decomposition exists and is unique for any positive definite symmetric $N \times N$ matrix. In the case of $\boldsymbol{\Gamma}_N$ (defined as dictated by Equation (450c)), the decomposition states that we can write

$$\boldsymbol{\Gamma}_N = \boldsymbol{L}_N\boldsymbol{D}_N\boldsymbol{L}_N^T, \tag{464c}$$

where $\boldsymbol{L}_N$ is an $N \times N$ lower triangular matrix with 1 as each of its diagonal elements, and $\boldsymbol{D}_N$ is an $N \times N$ diagonal matrix, each of whose diagonal elements are positive. This decomposition is equivalent to both

$$\boldsymbol{\Gamma}_N^{-1} = \boldsymbol{L}_N^{-T}\boldsymbol{D}_N^{-1}\boldsymbol{L}_N^{-1} \text{ and } \boldsymbol{L}_N^{-1}\boldsymbol{\Gamma}_N\boldsymbol{L}_N^{-T} = \boldsymbol{D}_N, \text{ where } \boldsymbol{L}_N^{-T} \stackrel{\text{def}}{=} (\boldsymbol{L}_N^{-1})^T = (\boldsymbol{L}_N^T)^{-1}. \tag{464d}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Because the inverse of a lower triangular matrix with diagonal elements equal to 1 is also lower triangular with 1's along its diagonal (as can be shown using a proof by induction), the first equivalence is the modified Cholesky decomposition for $\boldsymbol{\Gamma}_N^{-1}$. Due to the special structure of $\boldsymbol{\Gamma}_N$, the matrix $\boldsymbol{L}_N^{-1}$ takes the form

$$\boldsymbol{L}_N^{-1} = \begin{bmatrix} 1 & & & & & & \\ -\phi_{1,1} & 1 & & & & & \\ -\phi_{2,2} & -\phi_{2,1} & 1 & & & & \\ \vdots & \vdots & \ddots & \ddots & & & \\ -\phi_{k,k} & -\phi_{k,k-1} & \cdots & -\phi_{k,1} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \\ -\phi_{N-1,N-1} & -\phi_{N-1,N-2} & \cdots & \cdots & -\phi_{N-1,1} & 1 \end{bmatrix},$$

where the Levinson–Durbin recursions yield the $\phi_{k,j}$ coefficients. To see why this is the correct form for $\boldsymbol{L}_N^{-1}$, note first that, with $\overrightarrow{e}_0(0) \stackrel{\text{def}}{=} X_0$, we have

$$\boldsymbol{L}_N^{-1}\boldsymbol{X} = \left[\overrightarrow{e}_0(0), \overrightarrow{e}_1(1), \overrightarrow{e}_2(2), \ldots, \overrightarrow{e}_k(k), \ldots, \overrightarrow{e}_{N-1}(N-1)\right]^T; \tag{465a}$$

i.e., the matrix $\boldsymbol{L}_N^{-1}$ transforms $\boldsymbol{X}$ into a vector of prediction errors. These prediction errors have two properties of immediate relevance. First, $\text{var}\{\overrightarrow{e}_k(k)\} = \sigma_k^2 > 0$, and these mean square prediction errors arise as part of the Levinson–Durbin recursions (Equation (453b) defines σ_0^2 to be s_0, the process variance). The second property is the subject of the following exercise.

▷ **Exercise [465]** Show that $\text{cov}\{\overrightarrow{e}_j(j), \overrightarrow{e}_k(k)\} = 0$ for $0 \le j < k \le N-1$. ◁

Since the prediction errors have variances dictated by σ_k^2 and are pairwise uncorrelated, it follows that the covariance matrix for $\boldsymbol{L}_N^{-1}\boldsymbol{X}$ is a diagonal matrix with $\sigma_0^2, \sigma_1^2, \ldots, \sigma_{N-1}^2$ as its diagonal elements – in fact this is the diagonal matrix $\boldsymbol{D}_N$ appearing in the modified Cholesky decomposition. A standard result from the theory of vectors of RVs says that, if $\boldsymbol{A}$ is a matrix with N columns, then the covariance matrix for $\boldsymbol{AX}$ is $\boldsymbol{A}\boldsymbol{\Gamma}_N\boldsymbol{A}^T$ (see, e.g., Brockwell and Davis, 1991, proposition 1.6.1). Thus the covariance matrix of $\boldsymbol{L}_N^{-1}\boldsymbol{X}$ is $\boldsymbol{L}_N^{-1}\boldsymbol{\Gamma}_N\boldsymbol{L}_N^{-T}$, which, upon equating to $\boldsymbol{D}_N$, yields the second equivalent formulation of the modified Cholesky decomposition stated in Equation (464d).

For the AR(p) process $\{Y_t\}$ of Equation (446a), the matrix $\boldsymbol{L}_N^{-1}$ specializes to

$$\boldsymbol{L}_N^{-1} = \begin{bmatrix} 1 & & & & & & \\ -\phi_{1,1} & 1 & & & & & \\ -\phi_{2,2} & -\phi_{2,1} & 1 & & & & \\ \vdots & \vdots & \ddots & & & & \\ -\phi_{p,p} & \cdots & -\phi_{p,1} & 1 & & & \\ 0 & -\phi_{p,p} & \cdots & -\phi_{p,1} & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \\ 0 & \cdots & 0 & -\phi_{p,p} & \cdots & -\phi_{p,1} & 1 \end{bmatrix}, \tag{465b}$$

while the upper p diagonal elements of $\boldsymbol{D}_N$ are $\sigma_0^2, \sigma_1^2, \ldots, \sigma_{p-1}^2$, and the lower $N-p$ elements are all equal to σ_p^2. With $\boldsymbol{Y} = \left[Y_0, Y_1, \ldots, Y_{N-1}\right]^T$, we have

$$\boldsymbol{L}_N^{-1}\boldsymbol{Y} = \left[\overrightarrow{e}_0(0), \overrightarrow{e}_1(1), \overrightarrow{e}_2(2), \ldots, \overrightarrow{e}_{p-1}(p-1), \epsilon_p, \epsilon_{p+1}, \ldots, \epsilon_{N-1}\right]^T. \tag{465c}$$

These facts will prove useful when we discuss maximum likelihood estimation in Section 9.8 and simulation of AR(p) processes in Section 11.1.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

9.5 Burg's Algorithm

The Yule–Walker estimators of the AR(p) parameters are only one of many proposed and studied in the literature. Since the late 1960s, an estimation technique based upon *Burg's algorithm* has been in wide use (particularly in engineering and geophysics). The popularity of this method is due to a number of reasons.

[1] Burg's algorithm is a variation on the solution of the Yule–Walker equations via the Levinson–Durbin recursions and, like them, is computationally efficient and recursive. If we recall that the AR(p) process that is fit to a time series via the Yule–Walker method has a theoretical ACVS that agrees *identically* with that of the sample ACVS (i.e., the biased estimator of the ACVS), we can regard Burg's algorithm as providing an alternative estimator of the ACVS. Burg's modification is intuitively reasonable and relies heavily on the relationship of the recursions to the prediction error problem discussed in Section 9.4.

[2] Burg's algorithm arose from his work with the maximum entropy principle in connection with SDF estimation. We describe and critically examine this principle in the next section, but it is important to realize that, as shown in the remainder of this section, Burg's algorithm can be justified without appealing to entropy arguments.

[3] While in theory Burg's algorithm can yield an estimated AR(p) process that is nonstationary, in practice its estimates – like those of the Yule–Walker method – correspond to a causal (and hence stationary) AR(p) process. Although nonstationarity is a possibility, the coefficients produced by Burg's algorithm cannot correspond to an acausal stationary process (as C&E [1] shows, the Burg estimated PACS must be such that each PACS estimate is bounded by unity in magnitude, and C&E [5] for Section 9.4 notes that $|\phi_{k,k}| < 1$ for all k corresponds to a causal AR process).

[4] Monte Carlo studies and experience with actual data indicate that, particularly for short time series, Burg's algorithm produces more reasonable estimates than the Yule–Walker method (see, for example, Lysne and Tjøstheim, 1987). As we note in Section 10.13, it is not, unfortunately, free of problems of its own.

The key to Burg's algorithm is the central role of $\phi_{k,k}$ (the kth-order partial autocorrelation coefficient) in the Levinson–Durbin recursions: if we have $\phi_{k-1,1}, \ldots, \phi_{k-1,k-1}$ and σ^2_{k-1}, Equations (456e) and (456f) tell us we need only determine $\phi_{k,k}$ to calculate the remaining order k parameters (namely, $\phi_{k,1}, \ldots, \phi_{k,k-1}$ and σ^2_k). In the Yule–Walker scheme, we estimate $\phi_{k,k}$ using Equation (458a), which is based on Equation (456d) in the Levinson–Durbin recursions. Note that it utilizes the estimated ACVS up to lag p. Burg's algorithm takes a different approach. It estimates $\phi_{k,k}$ by minimizing a certain sum of squares of *observed* forward and backward prediction errors. Given we have a time series of length N taken to be a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of a discrete parameter stationary process with zero mean, and given Burg's estimators $\bar{\phi}_{k-1,1}, \ldots, \bar{\phi}_{k-1,k-1}$ of the coefficients for a model of order $k-1$, the errors we need to consider are

$$\overrightarrow{e}_t(k-1) \stackrel{\text{def}}{=} X_t - \sum_{j=1}^{k-1} \bar{\phi}_{k-1,j} X_{t-j}, \qquad k \le t \le N-1,$$

and

$$\overleftarrow{e}_{t-k}(k-1) \stackrel{\text{def}}{=} X_{t-k} - \sum_{j=1}^{k-1} \bar{\phi}_{k-1,j} X_{t-k+j}, \qquad k \le t \le N-1.$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Suppose for the moment that $\bar{\phi}_{k,k}$ is *any* estimate of $\phi_{k,k}$ and that the remaining $\bar{\phi}_{k,j}$ terms are generated in a manner analogous to Equation (458b). We then have

$$\begin{aligned}
\overrightarrow{e}_t(k) &= X_t - \sum_{j=1}^{k} \bar{\phi}_{k,j} X_{t-j} \\
&= X_t - \sum_{j=1}^{k-1} \left(\bar{\phi}_{k-1,j} - \bar{\phi}_{k,k}\bar{\phi}_{k-1,k-j}\right) X_{t-j} - \bar{\phi}_{k,k} X_{t-k} \\
&= X_t - \sum_{j=1}^{k-1} \bar{\phi}_{k-1,j} X_{t-j} - \bar{\phi}_{k,k}\left(X_{t-k} - \sum_{j=1}^{k-1} \bar{\phi}_{k-1,k-j} X_{t-j}\right) \\
&= \overrightarrow{e}_t(k-1) - \bar{\phi}_{k,k}\left(X_{t-k} - \sum_{j=1}^{k-1} \bar{\phi}_{k-1,j} X_{t-k+j}\right) \\
&= \overrightarrow{e}_t(k-1) - \bar{\phi}_{k,k}\overleftarrow{e}_{t-k}(k-1), \quad k \le t \le N-1. \qquad (467a)
\end{aligned}$$

We recognize this as a sampling version of Equation (455b). In a similar way, we can derive a sampling version of Equation (456c):

$$\overleftarrow{e}_{t-k}(k) = \overleftarrow{e}_{t-k}(k-1) - \bar{\phi}_{k,k}\overrightarrow{e}_t(k-1), \quad k \le t \le N-1. \qquad (467b)$$

Equations (467a) and (467b) allow us to calculate the observed order k forward and backward prediction errors in terms of the order $k-1$ errors. These equations do not depend on *any* particular property of $\bar{\phi}_{k,k}$ and hence are valid no matter how we determine $\bar{\phi}_{k,k}$ as long as the remaining $\bar{\phi}_{k,j}$ terms are generated as dictated by Equation (458b). Burg's idea was to estimate $\phi_{k,k}$ such that the order k observed prediction errors are as small as possible by the appealing criterion that

$$\begin{aligned}
\mathrm{SS}_k(\bar{\phi}_{k,k}) \overset{\text{def}}{=} \sum_{t=k}^{N-1} \left\{\overrightarrow{e}_t^2(k) + \overleftarrow{e}_{t-k}^2(k)\right\} &= \sum_{t=k}^{N-1} \Big\{ \left[\overrightarrow{e}_t(k-1) - \bar{\phi}_{k,k}\overleftarrow{e}_{t-k}(k-1)\right]^2 \\
&\quad + \left[\overleftarrow{e}_{t-k}(k-1) - \bar{\phi}_{k,k}\overrightarrow{e}_t(k-1)\right]^2 \Big\} \\
&= A_k - 2\bar{\phi}_{k,k}B_k + A_k\bar{\phi}_{k,k}^2 \qquad (467c)
\end{aligned}$$

be as small as possible, where

$$A_k \overset{\text{def}}{=} \sum_{t=k}^{N-1} \left\{\overrightarrow{e}_t^2(k-1) + \overleftarrow{e}_{t-k}^2(k-1)\right\} \qquad (467d)$$

and

$$B_k \overset{\text{def}}{=} 2\sum_{t=k}^{N-1} \overrightarrow{e}_t(k-1)\overleftarrow{e}_{t-k}(k-1). \qquad (467e)$$

Since $\mathrm{SS}_k(\cdot)$ is a quadratic function of $\bar{\phi}_{k,k}$, we can differentiate it and set the resulting expression to 0 to find the desired value of $\bar{\phi}_{k,k}$, namely,

$$\bar{\phi}_{k,k} = B_k/A_k. \qquad (467f)$$

(Note that $\bar{\phi}_{k,k}$ has a natural interpretation as an estimator of the left-hand side of Equation (462b).) With $\bar{\phi}_{k,k}$ so determined, we can estimate the remaining $\bar{\phi}_{k,j}$ using an equation

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

analogous to Equation (458b) – the second of the usual Levinson–Durbin recursions – and the corresponding observed prediction errors using Equations (467a) and (467b).

Given how $\bar{\phi}_{k,k}$ arises, it should be obvious that, with the proper initialization, we can apply Burg's algorithm recursively to work our way up to estimates for the coefficients of an AR(p) model. In the same spirit as the recursive step, we can initialize the algorithm by finding that value of $\bar{\phi}_{1,1}$ that minimizes

$$\begin{aligned} \mathrm{SS}_1(\bar{\phi}_{1,1}) &= \sum_{t=1}^{N-1} \overrightarrow{e}_t^2(1) + \overleftarrow{e}_{t-1}^2(1) \\ &= \sum_{t=1}^{N-1} \left(X_t - \bar{\phi}_{1,1} X_{t-1} \right)^2 + \left(X_{t-1} - \bar{\phi}_{1,1} X_t \right)^2 . \end{aligned} \tag{468a}$$

This is equivalent to Equation (467c) with $k = 1$ if we adopt the conventions $\overrightarrow{e}_t(0) = X_t$ and $\overleftarrow{e}_{t-1}(0) = X_{t-1}$. We can interpret these as meaning that, with no prior (future) observations, we should predict (backward predict) X_t (X_{t-1}) by zero, the assumed known process mean.

Burg's algorithm also specifies a way of estimating the innovation variance σ_p^2. It is done recursively using an equation analogous to Equation (458c), namely,

$$\bar{\sigma}_k^2 = \bar{\sigma}_{k-1}^2 \left(1 - \bar{\phi}_{k,k}^2 \right) . \tag{468b}$$

This assumes that, at the kth step, $\bar{\sigma}_{k-1}^2$ is available. For $k = 1$ there is obviously a problem in that we have not defined what $\bar{\sigma}_0^2$ is. If we follow the logic of the previous paragraph and of Equation (453b) and consider the zeroth-order predictor of our time series to be zero (the known process mean), the obvious estimator of the zeroth-order mean square linear prediction error is just

$$\bar{\sigma}_0^2 \stackrel{\mathrm{def}}{=} \frac{1}{N} \sum_{t=0}^{N-1} X_t^2 = \hat{s}_0^{(\mathrm{P})},$$

the sample variance of the time series. If we let $\hat{S}^{(\mathrm{BURG})}(\cdot)$ denote the resulting Burg SDF estimator, the above stipulation implies that in practice

$$\int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \hat{S}^{(\mathrm{BURG})}(f) \, \mathrm{d}f = \hat{s}_0^{(\mathrm{P})}. \tag{468c}$$

The above follows from an argument similar to the one used to show that the Yule–Walker estimator integrates to the sample variance (see Equation (451d)); however, whereas the Yule–Walker ACVS estimator agrees with $\hat{s}_0^{(\mathrm{P})}, \hat{s}_1^{(\mathrm{P})}, \ldots, \hat{s}_p^{(\mathrm{P})}$, the same does *not* hold in general for the Burg estimator when $\tau \geq 1$.

As an example of the application of Burg's algorithm to autoregressive SDF estimation, Figure 469 shows the Burg AR(4) SDF estimates for the same series used to produce the Yule–Walker AR(4) and AR(8) SDF estimates of Figures 458 and 459. A comparison of these three figures shows that Burg's method is clearly superior to the Yule–Walker method (at least in the standard formulation of that method using the biased estimator of the ACVS and at least for this particular AR(4) time series).

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

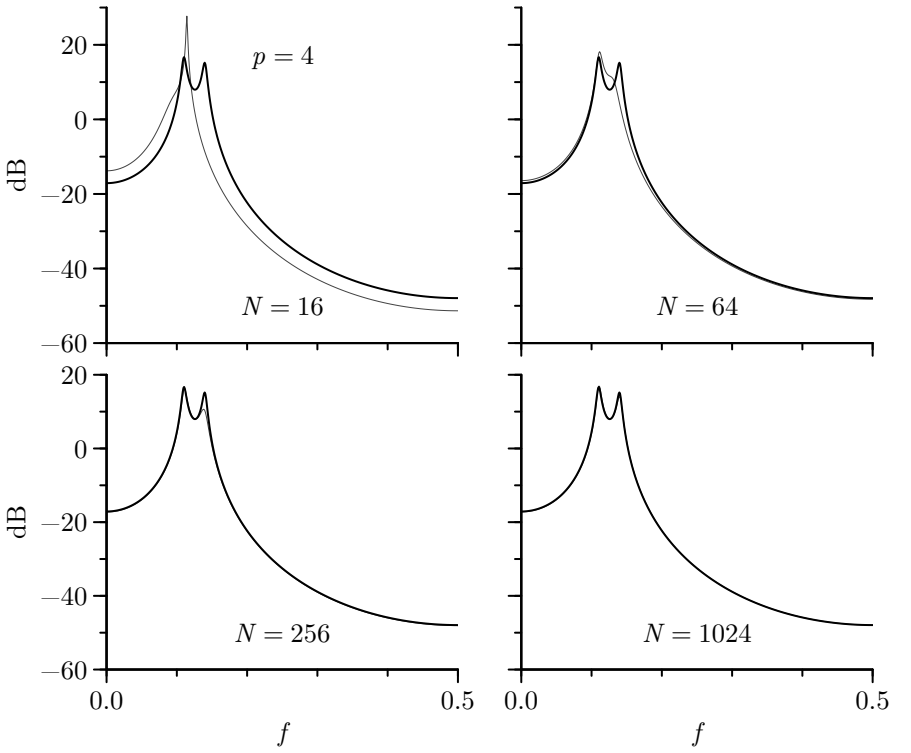

Figure 469 Burg AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e). The thick curve in each plot is the true SDF. This figure should be compared with Figures 458 and 459.

Comments and Extensions to Section 9.5

[1] We have previously noted that the *theoretical* kth-order partial autocorrelation coefficient $\phi_{k,k}$ can be interpreted as a correlation coefficient and thus must lie in the interval $[-1, 1]$. One important property of Burg's algorithm is that the corresponding *estimated* value $\bar{\phi}_{k,k}$ is also necessarily in that range. To see this, note that

$$0 \le \left(\overrightarrow{e}_t(k-1) \pm \overleftarrow{e}_{t-k}(k-1)\right)^2 = \overrightarrow{e}_t^2(k-1) \pm 2\overrightarrow{e}_t(k-1)\overleftarrow{e}_{t-k}(k-1) + \overleftarrow{e}_{t-k}^2(k-1)$$

implies that

$$|2\overrightarrow{e}_t(k-1)\overleftarrow{e}_{t-k}(k-1)| \le \overrightarrow{e}_t^2(k-1) + \overleftarrow{e}_{t-k}^2(k-1)$$

and hence that

$$|B_k| \stackrel{\text{def}}{=} \left| 2\sum_{t=k}^{N-1} \overrightarrow{e}_t(k-1)\overleftarrow{e}_{t-k}(k-1) \right| \le \sum_{t=k}^{N-1} \left[\overrightarrow{e}_t^2(k-1) + \overleftarrow{e}_{t-k}^2(k-1)\right] \stackrel{\text{def}}{=} A_k.$$

Since $A_k \ge 0$ (with equality rarely occurring in practical cases) and $\bar{\phi}_{k,k} = B_k \big/ A_k$, we have $|\bar{\phi}_{k,k}| \le 1$ as claimed.

[2] We have described the estimator of σ_k^2 commonly associated with Burg's algorithm. There is a second obvious estimator. Since σ_k^2 is the kth-order mean square linear prediction error, this estimator is

$$\bar{\nu}_k^2 \stackrel{\text{def}}{=} \frac{\text{SS}_k(\bar{\phi}_{k,k})}{2(N-k)}.$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

In the numerator the sum of squares of Equation (467c) is evaluated at its minimum value. Are $\bar{\nu}_k^2$ and $\bar{\sigma}_k^2$ in fact different? The answer is yes, but usually the difference is small. Exercise [9.13] says that

$$\bar{\nu}_k^2 = \left(1 - \bar{\phi}_{k,k}^2\right)\left(\bar{\nu}_{k-1}^2 + \frac{2\bar{\nu}_{k-1}^2 - \overrightarrow{e}_{k-1}^2(k-1) - \overleftarrow{e}_{N-k}^2(k-1)}{2(N-k)}\right). \tag{470a}$$

Comparison of this equation with the definition of $\bar{\sigma}_k^2$ in Equation (468b) shows that, even if $\bar{\nu}_{k-1}^2$ and $\bar{\sigma}_{k-1}^2$ are identical, $\bar{\nu}_k^2$ and $\bar{\sigma}_k^2$ need not be the same; however, since $\bar{\nu}_{k-1}^2$, $\overrightarrow{e}_{k-1}^2(k-1)$ and $\overleftarrow{e}_{N-k}^2(k-1)$ can all be regarded as estimators of the same parameter, σ_k^2, the last term in the parentheses should usually be small. For large N, we thus have

$$\bar{\nu}_k^2 \approx \bar{\nu}_{k-1}^2\left(1 - \bar{\phi}_{k,k}^2\right),$$

in agreement with the third equation of the Levinson–Durbin recursions. The relative merits of these two estimators are unknown, but, as far as SDF estimation is concerned, use of one or the other will only affect the level of the resulting estimate and not its shape. Use of $\bar{\sigma}_k^2$ ensures that the integral of the Burg-based SDF estimate is equal to the sample variance $\hat{s}_0^{(\mathrm{P})}$, which is a desirable property for an SDF estimator to have.

[3] There is a second – but equivalent – way of expressing Burg's algorithm that is both more succinct and also clarifies the relationship between this algorithm and the Yule–Walker method (Newton, 1988, section 3.4). Before presenting it, we establish some notation. Define a circular shift operator $\mathcal{L}$ and a subvector extraction operator $\mathcal{M}_{j,k}$ as follows: if $\boldsymbol{V}_0 = \left[v_0, v_1, \ldots, v_{M-2}, v_{M-1}\right]^T$ is any M-dimensional column vector of real numbers, then

$$\mathcal{L}\boldsymbol{V}_0 \stackrel{\text{def}}{=} \left[v_{M-1}, v_0, v_1, \ldots, v_{M-2}\right]^T \quad \text{and} \quad \mathcal{M}_{j,k}\boldsymbol{V}_0 \stackrel{\text{def}}{=} \left[v_j, v_{j+1}, \ldots, v_{k-1}, v_k\right]^T$$

(we assume that $0 \le j < k \le M-1$). If $\boldsymbol{V}_1$ is any vector whose dimension is the same as that of $\boldsymbol{V}_0$, we define $\langle \boldsymbol{V}_0, \boldsymbol{V}_1 \rangle = \boldsymbol{V}_0^T\boldsymbol{V}_1$ to be their inner product. We also define the squared norm of $\boldsymbol{V}_0$ to be $\|\boldsymbol{V}_0\|^2 = \langle \boldsymbol{V}_0, \boldsymbol{V}_0 \rangle = \boldsymbol{V}_0^T\boldsymbol{V}_0$.

To fit an AR(p) model to $X_0, \ldots, X_{N-1}$ using Burg's algorithm, we first set up the following vectors of length $M = N + p$:

$$\overrightarrow{\boldsymbol{e}}(0) \stackrel{\text{def}}{=} [X_0, X_1, \ldots, X_{N-1}, \underbrace{0, \ldots, 0}_{p \text{ of these}}]^T$$

and

$$\overleftarrow{\boldsymbol{e}}(0) \stackrel{\text{def}}{=} \mathcal{L}\overrightarrow{\boldsymbol{e}}(0) = [0, X_0, X_1, \ldots, X_{N-1}, \underbrace{0, \ldots, 0}_{p-1 \text{ of these}}]^T.$$

As before, we define $\bar{\sigma}_0^2 = \hat{s}_0^{(\mathrm{P})}$. For $k = 1, \ldots, p$, we then recursively compute

$$\begin{aligned}
\bar{\phi}_{k,k} &= \frac{2\langle \mathcal{M}_{k,N-1}\overrightarrow{\boldsymbol{e}}(k-1), \mathcal{M}_{k,N-1}\overleftarrow{\boldsymbol{e}}(k-1)\rangle}{\|\mathcal{M}_{k,N-1}\overrightarrow{\boldsymbol{e}}(k-1)\|^2 + \|\mathcal{M}_{k,N-1}\overleftarrow{\boldsymbol{e}}(k-1)\|^2} \\
\bar{\sigma}_k^2 &= \bar{\sigma}_{k-1}^2\left(1 - \bar{\phi}_{k,k}^2\right) \\
\overrightarrow{\boldsymbol{e}}(k) &= \overrightarrow{\boldsymbol{e}}(k-1) - \bar{\phi}_{k,k}\overleftarrow{\boldsymbol{e}}(k-1) \\
\overleftarrow{\boldsymbol{e}}(k) &= \mathcal{L}(\overleftarrow{\boldsymbol{e}}(k-1) - \bar{\phi}_{k,k}\overrightarrow{\boldsymbol{e}}(k-1)).
\end{aligned} \tag{470b}$$

This procedure yields the Burg estimators of $\phi_{k,k}$ and σ_k^2 for $k = 1, \ldots, p$ (if so desired, the remaining $\bar{\phi}_{k,j}$ can be generated using an equation similar to Equation (458b)). The $N - p$ forward prediction errors $\overrightarrow{e}_p(p), \ldots, \overrightarrow{e}_{N-1}(p)$ are the elements of $\mathcal{M}_{p,N-1}\overrightarrow{\boldsymbol{e}}(p)$; the backward prediction errors $\overleftarrow{e}_0(p)$, $\ldots$, $\overleftarrow{e}_{N-p-1}(p)$, those of $\mathcal{M}_{p+1,N}\overleftarrow{\boldsymbol{e}}(p)$.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

The Yule–Walker estimators $\tilde{\phi}_{k,k}$ and $\tilde{\sigma}_k^2$ can be generated by a scheme that is *identical* to the above, with one key modification: Equation (470b) becomes

$$\tilde{\phi}_{k,k} = \frac{2\langle \overrightarrow{\boldsymbol{e}}(k-1), \overleftarrow{\boldsymbol{e}}(k-1)\rangle}{\|\overrightarrow{\boldsymbol{e}}(k-1)\|^2 + \|\overleftarrow{\boldsymbol{e}}(k-1)\|^2} = \frac{\langle \overrightarrow{\boldsymbol{e}}(k-1), \overleftarrow{\boldsymbol{e}}(k-1)\rangle}{\|\overrightarrow{\boldsymbol{e}}(k-1)\|^2} \tag{471}$$

since $\|\overleftarrow{\boldsymbol{e}}(k-1)\|^2 = \|\overrightarrow{\boldsymbol{e}}(k-1)\|^2$ (all overbars in the three equations below Equation (470b) should also be changed to tildes). Note that, whereas Burg's algorithm uses an inner product tailored to involve just the actual data values, the inner product of the Yule–Walker estimator is influenced by the p zeros used to construct $\overrightarrow{\boldsymbol{e}}(0)$ and $\overleftarrow{\boldsymbol{e}}(0)$ (a similar interpretation of the Yule–Walker estimator appears in the context of least squares theory – see C&E [2] for Section 9.7). This finding supports Burg's contention (discussed in the next section) that the Yule–Walker method implicitly assumes $X_t = 0$ for $t < 0$ and $t \geq N$. Finally we note that the formulation involving Equation (471) allows us to compute the Yule–Walker estimators *directly* from a time series, thus bypassing the need to first compute $\{\hat{s}_\tau^{(\mathrm{P})}\}$ as required by Equation (458a).

9.6 The Maximum Entropy Argument

Burg's algorithm is an outgrowth of his work on maximum entropy spectral analysis (MESA). Burg (1967) criticizes the use of lag window spectral estimators of the form

$$\Delta_\mathrm{t} \sum_{\tau=-(N-1)}^{N-1} w_{m,\tau} \hat{s}_\tau^{(\mathrm{P})} \mathrm{e}^{-\mathrm{i}2\pi f \tau \Delta_\mathrm{t}}$$

because, first, we effectively force the sample ACVS to zero by multiplying it by a lag window $\{w_{m,\tau}\}$ and, second, we assume that $s_\tau = 0$ for $|\tau| \geq N$. To quote from Burg (1975),

> While window theory is interesting, it is actually a problem that has been artificially induced into the estimation problem by the assumption that $s_\tau = 0$ for $|\tau| \geq N$ and by the willingness to change perfectly good data by the weighting function. If one were not blinded by the mathematical elegance of the conventional approach, making unfounded assumptions as to the values of unmeasured data and changing the data values that one knows would be totally unacceptable from a common sense and, hopefully, from a scientific point of view. To overcome these problems, it is clear that a completely different philosophical approach to spectral analysis is required ....
>
> While readily understood, [the] conventional window function approach produces spectral estimates which are negative and/or spectral estimates which do not agree with their known autocorrelation values. These two affronts to common sense were the main reasons for the development of maximum entropy spectral analysis ....

Burg's "different philosophical approach" is to apply the principle of maximum entropy. In physics, entropy is a measure of disorder; for a stationary process, it can be usefully defined in terms of the predictability of the process. To be specific, suppose that, following our usual convention, $\{X_t\}$ is an arbitrary zero mean stationary process, but make the additional assumption that it is purely nondeterministic with SDF $S(\cdot)$ (as already mentioned in Section 8.7, purely nondeterministic processes play a role in the Wold decomposition theorem – see, e.g., Brockwell and Davis, 1991, for details). As in Section 8.7, consider the best linear predictor $\widehat{X}_t$ of X_t given the infinite past X_{t-1}, X_{t-2}, ..., and, in keeping with Equation (404a), denote the corresponding innovation variance by

$$E\{(X_t - \widehat{X}_t)^2\} \stackrel{\mathrm{def}}{=} \sigma_\infty^2.$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Equation (404b) states that

$$\sigma_\infty^2 = \frac{1}{\Delta_t} \exp\left(\Delta_t \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S(f)\right)\, \mathrm{d}f \right).$$

In what follows, we take the entropy of the SDF $S(\cdot)$ to be

$$H\{S(\cdot)\} = \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S(f)\right)\, \mathrm{d}f. \tag{472a}$$

Thus large entropy is equivalent to a large innovation variance, i.e., a large prediction error variance. (In fact, the concept of entropy is usually defined in such a way that the entropy measure above is only valid for the case of stationary *Gaussian* processes.)

We can now state Burg's maximum entropy argument. If we have *perfect* knowledge of the ACVS for a process just out to lag p, we know only that $S(\cdot)$ lies within a certain class of SDFs. One way to select a particular member of this class is to pick that SDF, call it $\widetilde{S}(\cdot)$, that maximizes the entropy subject to the constraints

$$\int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \tilde{S}(f) \mathrm{e}^{\mathrm{i}2\pi f \tau \Delta_t}\, \mathrm{d}f = s_\tau, \qquad 0 \le \tau \le p. \tag{472b}$$

To quote again from Burg (1975),

> Maximum entropy spectral analysis is based on choosing the spectrum which corresponds to the most random or the most unpredictable time series whose autocorrelation function agrees with the known values.

The solution to this constrained maximization problem is

$$\tilde{S}(f) = \frac{\sigma_p^2\, \Delta_t}{\left|1 - \sum_{k=1}^p \phi_{p,k} \mathrm{e}^{-\mathrm{i}2\pi f k \Delta_t}\right|^2},$$

where $\phi_{p,k}$ and σ_p^2 are the solutions to the augmented Yule–Walker equations of order p; i.e., the SDF is an AR(r) SDF with $r \le p$ (r is not necessarily equal to p since $\phi_{p,r+1}, \ldots, \phi_{p,p}$ could possibly be equal to 0). To see that this is true, let us consider the AR(p) process $\{Y_t\}$ with SDF $S_Y(\cdot)$ and innovation variance σ_p^2. For such a process the best linear predictor of Y_t given the infinite past is just

$$\widehat{Y}_t = \sum_{k=1}^p \phi_{p,k} Y_{t-k},$$

and the corresponding prediction error variance is just the innovation variance σ_p^2. Let $\{X_t\}$ be a stationary process whose ACVS agrees with that of $\{Y_t\}$ up to lag p, and let $S(\cdot)$ denote its SDF. We know that the best linear predictor of order p for X_t is

$$\sum_{k=1}^p \phi_{p,k} X_{t-k},$$

where the $\phi_{p,k}$ coefficients depend only on the ACVS of $\{X_t\}$ up to lag p; moreover, the associated prediction error variance must be σ_p^2. Because it is based on the infinite past, the

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

innovation variance for X_t cannot be larger than the prediction error variance for its pth-order predictor. Hence we have

$$\sigma_p^2 = \frac{1}{\Delta_t} \exp\left(\Delta_t \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S_Y(f)\right)\, \mathrm{d}f \right) \geq \frac{1}{\Delta_t} \exp\left(\Delta_t \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S(f)\right)\, \mathrm{d}f \right) = \sigma_\infty^2,$$

which implies that

$$H\{S_Y(\cdot)\} = \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S_Y(f)\right)\, \mathrm{d}f \geq \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(S(f)\right)\, \mathrm{d}f = H\{S(\cdot)\};$$

i.e., the entropy for a class of stationary processes – all of whose members are known to have the same ACVS out to lag p – is maximized by an AR(p) process with the specified ACVS out to lag p (Brockwell and Davis, 1991, section 10.6).

Burg's original idea for MESA (1967) was to *assume* that the biased *estimator* of the ACVS up to some lag p was in fact equal to the *true* ACVS. The maximum entropy principle under these assumptions leads to a procedure that is identical to the Yule–Walker estimation method. Burg (1968) abandoned this approach. He argued that it was not justifiable because, from his viewpoint, the usual biased estimator for the ACVS makes the implicit assumption that $X_t = 0$ for all $t < 0$ and $t \geq N$ (for example, with this assumption, the limits in Equation (170b) defining $\hat{s}_\tau^{(\mathrm{P})}$ can be modified to range from $t = -\infty$ to $t = \infty$). Burg's algorithm was his attempt to overcome this defect by estimating the ACVS in a different manner. There is still a gap in this logic, because true MESA requires *exact* knowledge of the ACVS up to a certain lag, whereas Burg's algorithm is just another way of *estimating* the ACVS from the data. Burg's criticism that window estimators effectively change "known" low-order values of the ACVS is offset by the fact that his procedure estimates lags beyond some arbitrary order p by an extension based upon only the first p values of the estimated ACVS and hence completely ignores what our time series might be telling us about the ACVS beyond lag p. (Another claim that has been made is that Burg's algorithm implicitly extends the time series – through predictions – outside the range $t = 0$ to $N - 1$ so that it is the "most random" one consistent with the observed data, i.e., closest to white noise. There are formidable problems in making this statement rigorous. A good discussion of other claims and counterclaims concerning maximum entropy can be found in a 1986 article by Makhoul with the intriguing title "Maximum Confusion Spectral Analysis.")

MESA is equivalent to AR spectral analysis when (a) the constraints on the entropy maximization are in terms of low-order values of the ACVS and (b) Equation (472a) is used as the measure of entropy. However, as Jaynes (1982) points out, the maximum entropy principle is more general so they need not be equivalent. If either the constraints (or "given data") are in terms of quantities other than just the ACVS or the measure of entropy is different from the one appropriate for stationary Gaussian processes, the resulting maximum entropy spectrum can have a different analytical form than the AR model. For example, recall from Section 7.9 that the cepstrum for an SDF $\tilde{S}(\cdot)$ is the sequence $\{c_\tau\}$ defined by

$$c_\tau = \int_{-f_{\mathcal{N}}}^{f_{\mathcal{N}}} \log\left(\tilde{S}(f)\right) \mathrm{e}^{\mathrm{i}2\pi f \tau \Delta_t}\, \mathrm{d}f \tag{473}$$

(Bogert et al., 1963). If we maximize the entropy (472a) over $\tilde{S}(\cdot)$ subject to the $p + 1$ constraints of Equation (472b) and q additional constraints that Equation (473) holds for known $c_1, \ldots, c_q$, then – subject to an existence condition – the maximum entropy principle yields an

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

ARMA(p, q) SDF (Lagunas-Hernández et al., 1984; Makhoul, 1986, section 8). Ihara (1984) and Franke (1985) discuss other sets of equivalent constraints that yield an ARMA spectrum as the maximum entropy spectrum.

Is the maximum entropy principle a good criterion for SDF estimation? Figure 475 shows an example adapted from an article entitled "Spectral Estimation: An Impossibility?" (Nitzberg, 1979). Here we assume that the ACVS of a certain process is known to be

$$s_\tau = 1 - \frac{|\tau|}{8} \text{ for } |\tau| \le 8 = p.$$

The left-hand column shows four possible extrapolations of the ACVS for $|\tau| > p$. The upper three plots assume that the ACVS is given by

$$s_\tau = \begin{cases} \alpha\left(1 - \left||\tau| - 16\right|/8\right), & 8 < |\tau| \le 24; \\ 0, & |\tau| > 24, \end{cases} \tag{474}$$

where, going down from the top, $\alpha = 0$, 1/2 and $-1/2$, respectively. The bottom-most plot shows the extension (out to $\tau = 32$) dictated by the maximum entropy principle. The right-hand column shows the low frequency portions of the four corresponding SDFs, which are rather different from each other (here we assume that the sampling interval Δ_t is unity so that the Nyquist frequency is 0.5). This illustrates the point that many different SDFs can have the same ACVS up to a certain lag. Even if we grant Burg's claim that we know the ACVS perfectly up to lag p (we never do with real data!), it is not clear why the maximum entropy extension of the ACVS has more credibility than the other three extensions. In particular, setting $\alpha = 0$ in Equation (474) yields an SDF with a monotonic decay over $f \in [0, 0.1]$, whereas the maximum entropy extension has an SDF associated with a broad peak centered at $f \doteq 0.034$. Given the prevalence of SDFs with monotonic decay in physical applications, the maximum entropy SDF has arguably less intuitive appeal than the $\alpha = 0$ SDF.

Figure 476 explores this example from another perspective. Suppose that the true ACVS comes from setting $\alpha = 0$ in Equation (474). This ACVS corresponds to an MA(7) process given by

$$X_t = \sum_{j=0}^{7} \epsilon_{t-j},$$

where $\{\epsilon_t\}$ is white noise process with zero mean and variance $\sigma_\epsilon^2 = 1/8$ (cf. Equation (32a)). The top row of plots in Figure 476 shows this ACVS for $\tau = 0, \ldots, 32$, and the corresponding SDF for $f \in [0, 0.1]$. This MA process is noninvertible; i.e., it cannot be expressed as an AR process of infinite order (Exercise [2.15] has an example of an invertible MA process). Noninvertibility suggests that AR approximations of finite orders p cannot be particularly good. The bottom three rows show AR(p) approximations of orders $p = 8$, 9 and 16 – these are the maximum entropy extensions given the ACVS out to lags 8, 9 or 16. These extensions result in unappealing peaks in the maximum entropy SDF where none exist in the true SDF and on specifying nonzero correlations at lags beyond ones at which the process is known to have reached decorrelation. Nonetheless, there are many processes with spectra that can be well approximated by an AR(p) process, and the maximum entropy method can be expected to work well for these.

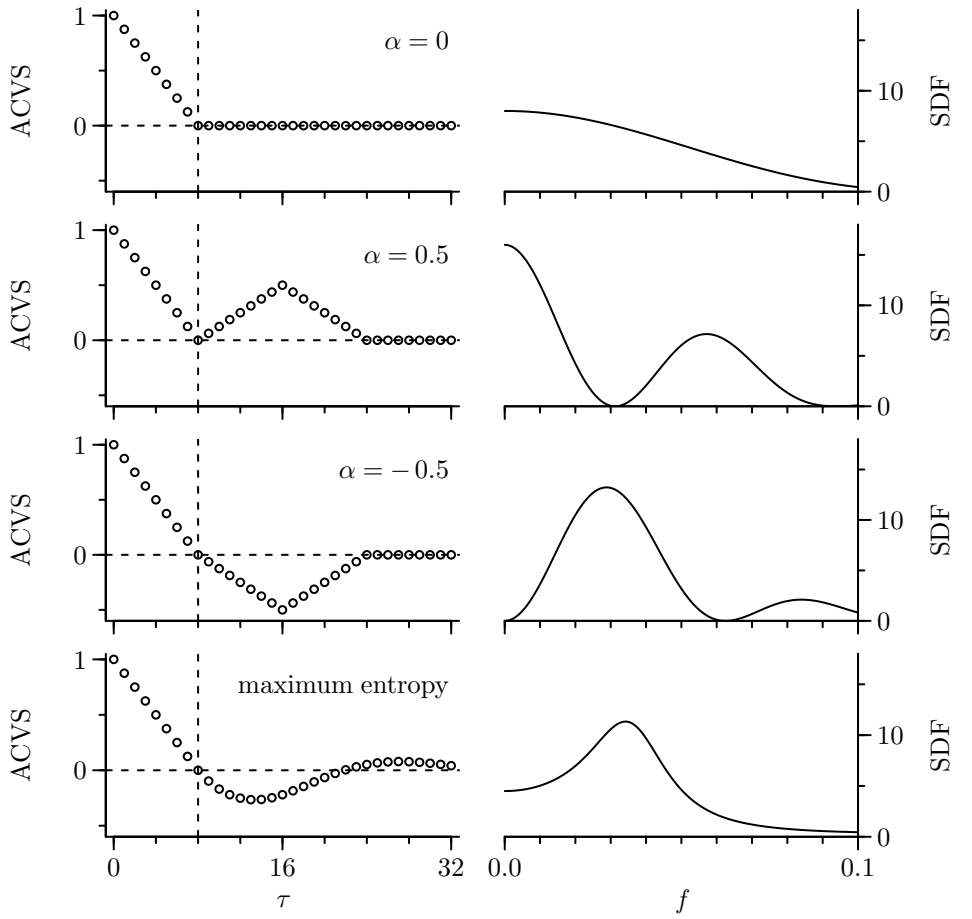

Figure 475 Example of four ACVSs (left-hand column) with identical values up to lag 8 and low frequency portions of corresponding SDFs (right-hand column – these are plotted on a linear/linear scale). The maximum entropy extension is shown on the bottom row.

9.7 Least Squares Estimators

In addition to the Yule–Walker estimator and Burg's algorithm, other ways for fitting an AR(p) model to a time series are least squares (LS) methods and the maximum likelihood method (discussed in the next section).

Suppose we have a time series of length N that can be regarded as a portion $Y_0, \ldots, Y_{N-1}$ of one realization of a stationary AR(p) process with zero mean. We can formulate an appropriate LS model in terms of our data as follows:

$$\boldsymbol{Y}_{(\mathrm{FLS})} = \mathcal{Y}_{(\mathrm{FLS})}\boldsymbol{\Phi} + \boldsymbol{\epsilon}_{(\mathrm{FLS})},$$

where $\boldsymbol{Y}_{(\mathrm{FLS})} = [Y_p, Y_{p+1}, \ldots, Y_{N-1}]^T$;

$$\mathcal{Y}_{(\mathrm{FLS})} = \begin{bmatrix} Y_{p-1} & Y_{p-2} & \cdots & Y_0 \\ Y_p & Y_{p-1} & \cdots & Y_1 \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N-2} & Y_{N-3} & \cdots & Y_{N-p-1} \end{bmatrix};$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

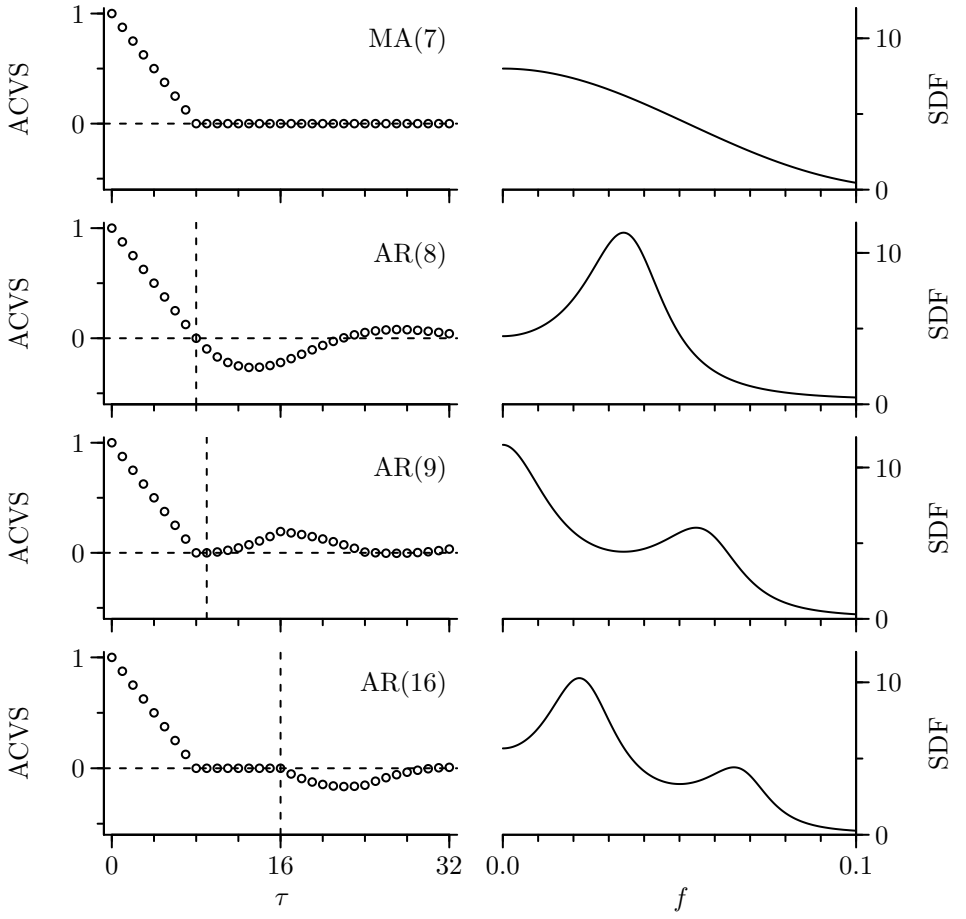

Figure 476 Four ACVSs (left-hand column) and low frequency portions of corresponding SDFs (right-hand column – these are plotted on a linear/linear scale). The top row of plots shows the ACVS for an MA(7) process at lags $\tau = 0, \ldots, 32$ and its corresponding SDF for $f \in [0, 0.1]$. The bottom three rows show corresponding maximum entropy extensions given knowledge of the MA(7) ACVS out to lags 8, 9 and 16 (marked by vertical dashed lines).

$\boldsymbol{\Phi} = [\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}]^T$; and $\boldsymbol{\epsilon}_{(\mathrm{FLS})} = [\epsilon_p, \epsilon_{p+1}, \ldots, \epsilon_{N-1}]^T$ (the rationale for FLS rather than just LS in the subscripts will become apparent in a moment). We can thus estimate $\boldsymbol{\Phi}$ by finding the $\boldsymbol{\Phi}$ for which

$$\mathrm{SS}_{(\mathrm{FLS})}(\boldsymbol{\Phi}) = \|\boldsymbol{Y}_{(\mathrm{FLS})} - \mathcal{Y}_{(\mathrm{FLS})}\boldsymbol{\Phi}\|^2 = \sum_{t=p}^{N-1} \left(Y_t - \sum_{k=1}^{p} \phi_{p,k} Y_{t-k} \right)^2 \qquad (476)$$

is minimized – as before, $\|\boldsymbol{V}\|^2 = \boldsymbol{V}^T\boldsymbol{V}$ for a vector $\boldsymbol{V}$. If we denote the vector that minimizes the above as $\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})}$, standard LS theory tells us that it is given by

$$\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})} = \left(\mathcal{Y}_{(\mathrm{FLS})}^T \mathcal{Y}_{(\mathrm{FLS})}\right)^{-1} \mathcal{Y}_{(\mathrm{FLS})}^T \boldsymbol{Y}_{(\mathrm{FLS})}$$

under the assumption that the matrix inverse exists (this assumption invariably holds in practice – if not, then $\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})}$ is a solution to $\mathcal{Y}_{(\mathrm{FLS})}^T \mathcal{Y}_{(\mathrm{FLS})} \boldsymbol{\Phi} = \mathcal{Y}_{(\mathrm{FLS})}^T \boldsymbol{Y}_{(\mathrm{FLS})}$). We can estimate the innovation variance σ_p^2 by the usual (approximately unbiased) least squares estimator of the

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

residual variation, namely,

$$\hat{\sigma}^2_{(\mathrm{FLS})} \stackrel{\mathrm{def}}{=} \frac{\mathrm{SS}_{(\mathrm{FLS})}(\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})})}{N-2p}, \tag{477a}$$

where the divisor arises because there are effectively $N-p$ observations and p parameters to be estimated. Alternatively, if we condition on $Y_0, Y_1, \ldots, Y_{p-1}$, then we can interpret $\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})}$ as a conditional maximum likelihood estimator, for which the corresponding innovation variance estimator is

$$\tilde{\sigma}^2_{(\mathrm{FLS})} \stackrel{\mathrm{def}}{=} \frac{\mathrm{SS}_{(\mathrm{FLS})}(\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})})}{N-p} \tag{477b}$$

(Priestley, 1981, p. 352; McQuarrie and Tsai, 1998, p. 90).

The estimator $\widehat{\boldsymbol{\Phi}}_{(\mathrm{FLS})}$ is known in the literature as the *forward least squares estimator* of $\boldsymbol{\Phi}$ to contrast it with two other closely related estimators. Motivated by the fact that the reversal of a stationary process is a stationary process with the same ACVS, we can reformulate the LS problem as

$$\boldsymbol{Y}_{(\mathrm{BLS})} = \mathcal{Y}_{(\mathrm{BLS})}\boldsymbol{\Phi} + \boldsymbol{\epsilon}_{(\mathrm{BLS})},$$

where $\boldsymbol{Y}_{(\mathrm{BLS})} = [Y_0, Y_1, \ldots, Y_{N-p-1}]^T$;

$$\mathcal{Y}_{(\mathrm{BLS})} = \begin{bmatrix} Y_1 & Y_2 & \ldots & Y_p \\ Y_2 & Y_3 & \ldots & Y_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N-p} & Y_{N-p+1} & \ldots & Y_{N-1} \end{bmatrix};$$

and $\boldsymbol{\epsilon}_{(\mathrm{BLS})}$ is a vector of uncorrelated errors. The function of $\boldsymbol{\Phi}$ to be minimized is now

$$\mathrm{SS}_{(\mathrm{BLS})}(\boldsymbol{\Phi}) = \|\boldsymbol{Y}_{(\mathrm{BLS})} - \mathcal{Y}_{(\mathrm{BLS})}\boldsymbol{\Phi}\|^2 = \sum_{t=0}^{N-p-1} \Big(Y_t - \sum_{k=1}^{p} \phi_{p,k} Y_{t+k}\Big)^2. \tag{477c}$$

The *backward least squares estimator* of $\boldsymbol{\Phi}$ is thus given by

$$\widehat{\boldsymbol{\Phi}}_{(\mathrm{BLS})} = \big(\mathcal{Y}^T_{(\mathrm{BLS})}\mathcal{Y}_{(\mathrm{BLS})}\big)^{-1}\mathcal{Y}^T_{(\mathrm{BLS})}\boldsymbol{Y}_{(\mathrm{BLS})},$$

and the appropriate estimate for the innovation variance is

$$\hat{\sigma}^2_{(\mathrm{BLS})} = \frac{\mathrm{SS}_{(\mathrm{BLS})}(\widehat{\boldsymbol{\Phi}}_{(\mathrm{BLS})})}{N-2p}.$$

Finally, in the same spirit as Burg's algorithm, we can define a *forward/backward least squares estimator* $\widehat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ of $\boldsymbol{\Phi}$ as that vector minimizing

$$\mathrm{SS}_{(\mathrm{FBLS})}(\boldsymbol{\Phi}) = \mathrm{SS}_{(\mathrm{FLS})}(\boldsymbol{\Phi}) + \mathrm{SS}_{(\mathrm{BLS})}(\boldsymbol{\Phi}). \tag{477d}$$

This estimator is associated with

$$\boldsymbol{Y}_{(\mathrm{FBLS})} = \begin{bmatrix} \boldsymbol{Y}_{(\mathrm{FLS})} \\ \boldsymbol{Y}_{(\mathrm{BLS})} \end{bmatrix}, \quad \mathcal{Y}_{(\mathrm{FBLS})} = \begin{bmatrix} \mathcal{Y}_{(\mathrm{FLS})} \\ \mathcal{Y}_{(\mathrm{BLS})} \end{bmatrix} \quad \text{and} \quad \hat{\sigma}^2_{(\mathrm{FBLS})} = \frac{\mathrm{SS}_{(\mathrm{FBLS})}(\widehat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})})}{2N-3p} \tag{477e}$$

(we use a divisor of $2N-3p$ in creating $\hat{\sigma}^2_{(\mathrm{FBLS})}$ because the length of $\boldsymbol{Y}_{(\mathrm{FBLS})}$ is $2N-2p$, and we have p coefficients to estimate). For a specified order p, we can compute the FBLS estimator as efficiently as the Burg estimator using a specialized algorithm due to Marple (1980);

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

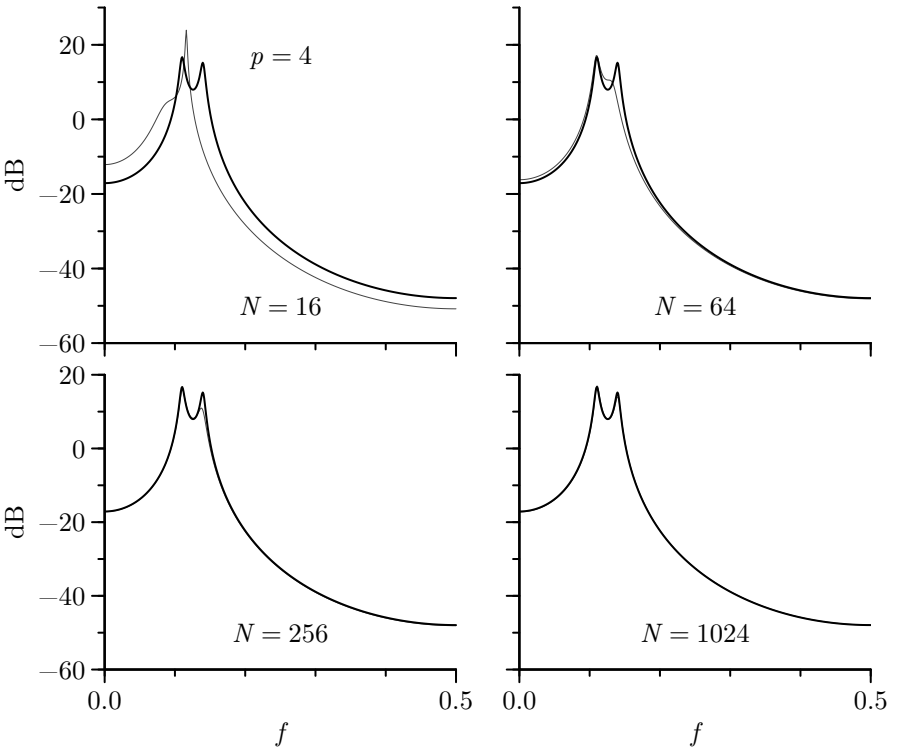

Figure 478 Forward/backward least squares AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e). The thick curve in each plot is the true SDF. This figure should be compared with Figures 458 and 469.

however, in contrast to Burg's algorithm, the FBLS algorithm does *not,* as a side product, yield fitted models of orders $1, \ldots, p-1$. These additional fits become useful when dealing with criteria for order selection (see Section 9.11). In terms of spectral estimation, Monte Carlo studies indicate that the FBLS estimator generally performs better than the Yule–Walker, Burg, FLS and BLS estimators (see Marple, 1987, and Kay, 1988, and references therein).

As an example of FBLS SDF estimation, Figure 478 shows these estimates for the same time series used to produce the Yule–Walker and Burg AR(4) SDF estimates of Figures 458 and 469. A comparison of these three figures shows that, for this example, Burg's method and the FBLS estimator lead to SDF estimates that are comparable with, or superior to, the estimates produced by the Yule–Walker method. (Exercise [9.14] invites the reader to create similar figures for the FLS and BLS methods and for three other AR(4) time series.)

Comments and Extensions to Section 9.7

[1] A cautionary note is in order about SDF estimators formed from least squares estimators of the AR parameters. In contrast to the Yule–Walker and Burg estimators, these SDF estimators need *not* integrate to the sample variance $\hat{s}_0^{(\mathrm{P})}$ for the time series (integration to $\hat{s}_0^{(\mathrm{P})}$ is a desirable property shared by certain nonparametric SDF estimators, including the periodogram – see Equation (171d)). If the least squares estimator corresponds to a causal AR process, we can correct this deficiency by replacing the estimator of the innovation variance suggested by standard least squares theory with a different estimator. To formulate this alternative estimator, let $\widehat{\boldsymbol{\Phi}}_{(\mathrm{LS})}$ and $\hat{\sigma}^2_{(\mathrm{LS})}$ denote the least squares parameter estimators produced by any one of the three least squares schemes. Using $\widehat{\boldsymbol{\Phi}}_{(\mathrm{LS})}$, we apply the step-down Levinson–Durbin recursions of Equation (460a) to obtain estimators, say $\hat{\phi}_{p-1,p-1}$, $\hat{\phi}_{p-2,p-2}$, $\ldots$, $\hat{\phi}_{1,1}$, of the

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

partial autocorrelations of all orders lower than p (the pth-order term is just the last element of $\widehat{\boldsymbol{\Phi}}_{(\mathrm{LS})}$ – we denote it as $\hat{\phi}_{p,p}$). The alternate to $\hat{\sigma}^2_{(\mathrm{LS})}$ is

$$\tilde{\sigma}^2_{(\mathrm{LS})} = \hat{s}^{(\mathrm{P})}_0 \prod_{j=1}^{p} \left(1 - \hat{\phi}^2_{j,j}\right) \tag{479}$$

(cf. Equation (509a), the subject of Exercise [9.8c]). The integral over $[-f_{\mathcal{N}}, f_{\mathcal{N}}]$ of the SDF estimator formed from $\widehat{\boldsymbol{\Phi}}_{(\mathrm{LS})}$ and $\tilde{\sigma}^2_{(\mathrm{LS})}$ is now guaranteed to be $\hat{s}^{(\mathrm{P})}_0$.

Unfortunately a least squares coefficient estimator $\widehat{\boldsymbol{\Phi}}_{(\mathrm{LS})}$ need *not* correspond to an estimated AR model that is causal (and hence stationary). Assuming that the estimated model is stationary (invariably the case in practical applications), the solution to Exercise [9.1b] points to a procedure for converting a stationary acausal AR model into a causal AR model, with the two models yielding exactly the same SDF estimators. In theory we could then take $\tilde{\sigma}^2_{(\mathrm{LS})}$ to be the innovation estimator for the causal model, which would ensure that the resulting SDF estimator integrates to $\hat{s}^{(\mathrm{P})}_0$. This corrective scheme is feasible for small p, but problems with numerical stability can arise as p increases because of the need to factor a polynomial of order p. (Exercise [9.15] in part concerns an FBLS estimate of a time series of length $N = 6$ and touches upon some of the issues raised here.)

[2] It should be noted that what we have defined as the Yule–Walker estimator in Section 9.3 is often called the least squares estimator in the statistical literature. To see the origin of this terminology, consider extending the time series $Y_0, Y_1, \ldots, Y_{N-1}$ by appending p dummy observations – each identically equal to zero – before and after it. The FLS model for this extended time series of length $N + 2p$ is

$$\boldsymbol{Y}_{(\mathrm{YW})} = \mathcal{Y}_{(\mathrm{YW})}\boldsymbol{\Phi} + \boldsymbol{\epsilon}_{(\mathrm{YW})},$$

where

$$\boldsymbol{Y}_{(\mathrm{YW})} \stackrel{\mathrm{def}}{=} [Y_0, Y_1, \ldots, Y_{N-1}, \underbrace{0, \ldots, 0}_{p \text{ of these}}]^T;$$

$$\mathcal{Y}_{(\mathrm{YW})} \stackrel{\mathrm{def}}{=} \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ Y_0 & 0 & \ldots & 0 & 0 \\ Y_1 & Y_0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{p-2} & Y_{p-3} & \ldots & Y_0 & 0 \\ Y_{p-1} & Y_{p-2} & \ldots & Y_1 & Y_0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{N-1} & Y_{N-2} & \ldots & Y_{N-p+1} & Y_{N-p} \\ 0 & Y_{N-1} & \ldots & Y_{N-p+2} & Y_{N-p+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & Y_{N-1} & Y_{N-2} \\ 0 & 0 & \ldots & 0 & Y_{N-1} \end{bmatrix};$$

and $\boldsymbol{\epsilon}_{(\mathrm{YW})} \stackrel{\mathrm{def}}{=} [\epsilon'_0, \ldots, \epsilon'_{p-1}, \epsilon_p, \epsilon_{p+1}, \ldots, \epsilon_{N-1}, \epsilon'_N, \ldots, \epsilon'_{N+p-1}]^T$ – here we need the primes because the first and last p elements of this vector are *not* members of the innovation process $\{\epsilon_t\}$. Since

$$\frac{1}{N}\mathcal{Y}^T_{(\mathrm{YW})}\mathcal{Y}_{(\mathrm{YW})} = \begin{bmatrix} \hat{s}^{(\mathrm{P})}_0 & \hat{s}^{(\mathrm{P})}_1 & \ldots & \hat{s}^{(\mathrm{P})}_{p-1} \\ \hat{s}^{(\mathrm{P})}_1 & \hat{s}^{(\mathrm{P})}_0 & \ldots & \hat{s}^{(\mathrm{P})}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{s}^{(\mathrm{P})}_{p-1} & \hat{s}^{(\mathrm{P})}_{p-2} & \ldots & \hat{s}^{(\mathrm{P})}_0 \end{bmatrix} = \widetilde{\boldsymbol{\Gamma}}_p$$

and $\mathcal{Y}^T_{(\mathrm{YW})}\boldsymbol{Y}_{(\mathrm{YW})}/N = [\hat{s}^{(\mathrm{P})}_1, \hat{s}^{(\mathrm{P})}_2, \ldots, \hat{s}^{(\mathrm{P})}_p]^T = \widetilde{\boldsymbol{\gamma}}_p$, it follows that the FLS estimator for this extended time series is

$$\left(\mathcal{Y}^T_{(\mathrm{YW})}\mathcal{Y}_{(\mathrm{YW})}\right)^{-1}\mathcal{Y}^T_{(\mathrm{YW})}\boldsymbol{Y}_{(\mathrm{YW})} = \widetilde{\boldsymbol{\Gamma}}_p^{-1}\widetilde{\boldsymbol{\gamma}}_p,$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

which is identical to the Yule–Walker estimator (cf. Equation (451a)).

[3] Some of the AR estimators that we have defined so far in this chapter are known in the engineering literature under a different name. For the record, we note the following correspondences (Kay and Marple, 1981):

$$\begin{aligned} \text{Yule–Walker} &\Longleftrightarrow \text{autocorrelation method} \\ \text{forward least squares} &\Longleftrightarrow \text{covariance method} \\ \text{forward/backward least squares} &\Longleftrightarrow \text{modified covariance method.} \end{aligned}$$

9.8 Maximum Likelihood Estimators

We now consider maximum likelihood estimation of $\boldsymbol{\Phi}_p = [\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}]^T$ and σ_p^2, the $p+1$ parameters in the AR(p) model. To do so, we assume that our observed time series is a realization of a portion $H_0, \ldots, H_{N-1}$ of a Gaussian AR(p) process with zero mean. Given these observations, the likelihood function for the unknown parameters is

$$L(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H}) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Gamma}_N|^{1/2}} \exp\left(-\boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H}/2\right), \tag{480a}$$

where $\boldsymbol{H} \stackrel{\text{def}}{=} [H_0, \ldots, H_{N-1}]^T$, $\boldsymbol{\Gamma}_N$ is the covariance matrix for $\boldsymbol{H}$ (i.e., its (j,k)th element is s_{j-k} – cf. Equation (450c)) and $|\boldsymbol{\Gamma}_N|$ is its determinant. For particular values of $\boldsymbol{\Phi}_p$ and σ_p^2, we can compute $L(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H})$ by assuming that $\{H_t\}$ has these as its true parameter values. The maximum likelihood estimators (MLEs) are defined to be those values of the parameters that maximize $L(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H})$ or, equivalently, that minimize the log likelihood function (after multiplication by -2 to eliminate a common factor):

$$l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H}) \stackrel{\text{def}}{=} -2 \log\left(L(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H})\right) = \log\left(|\boldsymbol{\Gamma}_N|\right) + \boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H} + N \log(2\pi). \tag{480b}$$

To obtain the MLEs, we must be able to compute $l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H})$ for particular choices of $\boldsymbol{\Phi}_p$ and σ_p^2, a task which at first glance appears to be formidable due to the necessity of finding the determinant for – and inverting – the $N \times N$-dimensional matrix $\boldsymbol{\Gamma}_N$. Fortunately, the Toeplitz structure of $\boldsymbol{\Gamma}_N$ allows us to simplify Equation (480b) considerably, as we now show (see, for example, Newton, 1988).

We first assume that $\boldsymbol{\Gamma}_N$ is positive definite so that we can make use of the modified Cholesky decomposition for it stated in Equation (464c), namely,

$$\boldsymbol{\Gamma}_N = \boldsymbol{L}_N \boldsymbol{D}_N \boldsymbol{L}_N^T,$$

where $\boldsymbol{L}_N$ is a lower triangular matrix, all of whose diagonal elements are 1, while $\boldsymbol{D}_N$ is a diagonal matrix. Because $\{H_t\}$ is assumed to be an AR(p) process, the first p diagonal elements of $\boldsymbol{D}_N$ are $\sigma_0^2, \sigma_1^2, \ldots, \sigma_{p-1}^2$, while the remaining $N-p$ elements are all equal to σ_p^2. Because the determinant of a product of square matrices is equal to the product of the individual determinants, and because the determinant of a triangular matrix is equal to the product of its diagonal elements, we now have

$$|\boldsymbol{\Gamma}_N| = |\boldsymbol{L}_N| \cdot |\boldsymbol{D}_N| \cdot |\boldsymbol{L}_N^T| = \sigma_p^{2(N-p)} \prod_{j=0}^{p-1} \sigma_j^2. \tag{480c}$$

The modified Cholesky decomposition immediately yields

$$\boldsymbol{\Gamma}_N^{-1} = \boldsymbol{L}_N^{-T} \boldsymbol{D}_N^{-1} \boldsymbol{L}_N^{-1},$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

where, for an AR(p) process, the matrix $\boldsymbol{L}_N^{-1}$ takes the form indicated by Equation (465b). Hence we have

$$\boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H} = \boldsymbol{H}^T \boldsymbol{L}_N^{-T} \boldsymbol{D}_N^{-1} \boldsymbol{L}_N^{-1} \boldsymbol{H} = (\boldsymbol{L}_N^{-1} \boldsymbol{H})^T \boldsymbol{D}_N^{-1} (\boldsymbol{L}_N^{-1} \boldsymbol{H}).$$

From Equation (465c) the first p elements of $\boldsymbol{L}_N^{-1}\boldsymbol{H}$ are simply $\overrightarrow{\epsilon}_0(0)$, $\overrightarrow{\epsilon}_1(1)$, …, $\overrightarrow{\epsilon}_{p-1}(p-1)$, where, in accordance with earlier definitions,

$$\overrightarrow{\epsilon}_0(0) = H_0 \text{ and } \overrightarrow{\epsilon}_t(t) = H_t - \sum_{j=1}^{t} \phi_{t,j} H_{t-j}, \qquad t = 1, \ldots, p-1.$$

The last $N - p$ elements follow from the assumed model for $\{H_t\}$ and are given by

$$\epsilon_t = H_t - \sum_{j=1}^{p} \phi_{p,j} H_{t-j}, \qquad t = p, \ldots, N-1.$$

Hence we can write

$$\boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H} = \sum_{t=0}^{p-1} \frac{\overrightarrow{\epsilon}_t^2(t)}{\sigma_t^2} + \sum_{t=p}^{N-1} \frac{\epsilon_t^2}{\sigma_p^2}. \tag{481a}$$

Combining this and Equation (480c) with Equation (480b) yields

$$l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H}) = \sum_{j=0}^{p-1} \log\left(\sigma_j^2\right) + (N-p)\, \log\left(\sigma_p^2\right) + \sum_{t=0}^{p-1} \frac{\overrightarrow{\epsilon}_t^2(t)}{\sigma_t^2} + \sum_{t=p}^{N-1} \frac{\epsilon_t^2}{\sigma_p^2} + N\, \log(2\pi)\,.$$

If we note that we can write

$$\sigma_j^2 = \sigma_p^2 \lambda_j \text{ with } \lambda_j \stackrel{\text{def}}{=} \left(\prod_{k=j+1}^{p} (1 - \phi_{k,k}^2) \right)^{-1}$$

for $j = 0, \ldots, p-1$ (this follows from Exercise [9.8c]), we now obtain a useful form for the log likelihood function:

$$l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H}) = \sum_{j=0}^{p-1} \log(\lambda_j) + N\, \log\left(\sigma_p^2\right) + \frac{\mathrm{SS}_{(\mathrm{ML})}(\boldsymbol{\Phi}_p)}{\sigma_p^2} + N\, \log(2\pi)\,, \tag{481b}$$

where

$$\mathrm{SS}_{(\mathrm{ML})}(\boldsymbol{\Phi}_p) \stackrel{\text{def}}{=} \sum_{t=0}^{p-1} \frac{\overrightarrow{\epsilon}_t^2(t)}{\lambda_t} + \sum_{t=p}^{N-1} \epsilon_t^2. \tag{481c}$$

We can obtain an expression for the MLE $\hat{\sigma}_{(\mathrm{ML})}^2$ of σ_p^2 by differentiating Equation (481b) with respect to σ_p^2 and setting it to zero. This shows that, no matter what the MLE $\widehat{\boldsymbol{\Phi}}_{(\mathrm{ML})}$ of $\boldsymbol{\Phi}_p$ turns out to be, the estimator $\hat{\sigma}_{(\mathrm{ML})}^2$ is given by

$$\hat{\sigma}_{(\mathrm{ML})}^2 \stackrel{\text{def}}{=} \mathrm{SS}_{(\mathrm{ML})}(\widehat{\boldsymbol{\Phi}}_{(\mathrm{ML})})/N. \tag{481d}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

The parameter σ_p^2 can thus be eliminated from Equation (481b), yielding what Brockwell and Davis (1991) refer to as the *reduced likelihood* (sometimes called the *profile likelihood*):

$$l(\boldsymbol{\Phi}_p \mid \boldsymbol{H}) \stackrel{\text{def}}{=} \sum_{j=0}^{p-1} \log(\lambda_j) + N \log(\text{SS}_{(\text{ML})}(\boldsymbol{\Phi}_p)/N) + N + N \log(2\pi). \tag{482a}$$

We can determine $\widehat{\boldsymbol{\Phi}}_{(\text{ML})}$ by finding the value of $\boldsymbol{\Phi}_p$ that minimizes the reduced likelihood.

Let us now specialize to the case $p = 1$. To determine $\phi_{1,1}$, we must minimize

$$l(\phi_{1,1} \mid \boldsymbol{H}) = -\log\left(1 - \phi_{1,1}^2\right) + N \log(\text{SS}_{(\text{ML})}(\phi_{1,1})/N) + N + N \log(2\pi), \tag{482b}$$

where

$$\text{SS}_{(\text{ML})}(\phi_{1,1}) = H_0^2(1 - \phi_{1,1}^2) + \sum_{t=1}^{N-1} (H_t - \phi_{1,1} H_{t-1})^2. \tag{482c}$$

Differentiating Equation (482b) with respect to $\phi_{1,1}$ and setting the result equal to zero yields

$$\frac{\phi_{1,1}\text{SS}_{(\text{ML})}(\phi_{1,1})}{N} - (1 - \phi_{1,1}^2)\left(\sum_{t=1}^{N-1} H_{t-1}H_t - \phi_{1,1}\sum_{t=1}^{N-2} H_t^2\right) = 0, \tag{482d}$$

where the second sum vanishes when $N = 2$. In general this is a cubic equation in $\phi_{1,1}$. The estimator $\hat{\phi}_{(\text{ML})}$ is thus equal to the root of this equation that minimizes $l(\phi_{1,1} \mid \boldsymbol{H})$.

For $p > 2$, we cannot obtain the MLEs so easily. We must resort to a nonlinear optimizer in order to numerically determine $\widehat{\boldsymbol{\Phi}}_{(\text{ML})}$. Jones (1980) describes in detail a scheme for computing $\widehat{\boldsymbol{\Phi}}_{(\text{ML})}$ that uses a transformation of variables to facilitate numerical optimization. An interesting feature of his scheme is the use of a Kalman filter to compute the reduced likelihood at each step in the numerical optimization. This approach has two important advantages: first, it can easily deal with time series for which some of the observations are missing; and, second, it can handle the case in which an AR(p) process is observed in the presence of additive noise (this is discussed in more detail in Section 10.12).

Maximum likelihood estimation has not seen much use in parametric spectral analysis for two reasons. First, in a pure parametric approach, we need to be able to routinely fit fairly high-order AR models, particularly in the initial stages of a data analysis (p in the range of 10 and higher is common). High-order models are often a necessity in order to adequately capture all of the important features of an SDF. The computational burden of the maximum likelihood method can become unbearable as p increases beyond even 4 or 5. Second, if we regard parametric spectral analysis as merely a way of designing low-order prewhitening filters (the approach we personally favor – see Section 9.10), then, even though the model order might now be small enough to allow use of the maximum likelihood method, other – more easily computed – estimators such as the Burg or FBLS estimators work perfectly well in practice. Any imperfections in the prewhitening filter can be compensated for in the subsequent nonparametric spectral analysis of the prewhitened series.

As an example, Figure 483 shows maximum likelihood SDF estimates for the same time series used to produce the Yule–Walker, Burg and FBLS AR(4) SDF estimates of Figures 458, 469 and 478. Visually the ML-based estimates correspond well to the Burg and FBLS AR(4) estimates. Table 483 tabulates the values of the reduced likelihood for the Yule-Walker, Burg, FBLS and ML estimates shown in Figures 458, 469, 478 and 483. As must be the case, the reduced likelihood is smallest for the ML estimates. For all sample sizes, the next smallest are for the FBLS estimates, with the Burg estimates being close behind (the Yule–Walker estimates are a distant fourth). These results are verification that the Burg and FBLS estimates are decent surrogates for the computationally more demanding ML estimates.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

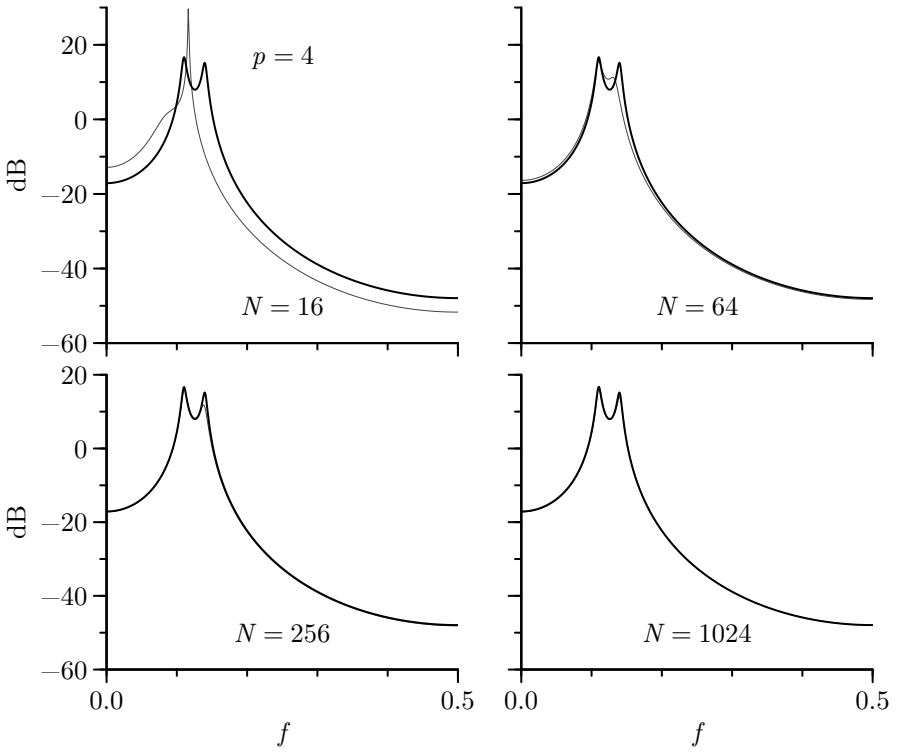

Figure 483 Maximum likelihood AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e). The thick curve in each plot is the true SDF. This figure should be compared with Figures 458, 469 and 478.

	$l(\widehat{\boldsymbol{\Phi}}_p \mid \boldsymbol{H})$			
N	Yule–Walker	Burg	FBLS	ML
16	10.59193	−46.77426	−47.24593	−48.14880
64	−47.21592	−206.0418	−206.3352	−206.5077
256	−314.3198	−864.5200	−864.6390	−864.7399
1024	−1873.498	−3446.536	−3446.557	−3446.560

Table 483 Reduced (profile) likelihoods for estimated AR(4) coefficients based on Yule–Walker, Burg, FBLS and ML methods. The time series are of varying lengths N, and the corresponding SDF estimates are shown in Figures 458, 469, 478 and 483. As must be the case, the ML estimates have the lowest profile likelihood values, but those for the Burg and FBLS methods are comparable (the values for the Yule–Walker method are much higher, and the SDF estimates for this method are visually the poorest of the four methods).

Comments and Extensions to Section 9.8

[1] The large-sample distribution of the MLEs for the parameters of an AR(p) process has been worked out (see, for example, Newton, 1988, or Brockwell and Davis, 1991). For large N, the estimator $\widehat{\boldsymbol{\Phi}}_{(\mathrm{ML})}$ is approximately distributed as a multivariate Gaussian RV with mean $\boldsymbol{\Phi}_p$ and covariance matrix $\sigma_p^2 \boldsymbol{\Gamma}_p^{-1}/N$, where $\boldsymbol{\Gamma}_p$ is the covariance matrix defined in Equation (450c); moreover, $\hat{\sigma}^2_{(\mathrm{ML})}$ approximately follows a Gaussian distribution with mean σ_p^2 and variance $2\sigma_p^4/N$ and is approximately

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

independent of $\widehat{\boldsymbol{\Phi}}_{(\mathrm{ML})}$. The matrix $\sigma_p^2 \boldsymbol{\Gamma}_p^{-1}$ (sometimes called the *Schur matrix*) can be computed using Equation (464d), with N set to p; however, it is also the case that

$$\sigma_p^2 \boldsymbol{\Gamma}_p^{-1} = \boldsymbol{A}^T \boldsymbol{A} - \boldsymbol{B}^T \boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^T - \boldsymbol{B}\boldsymbol{B}^T, \tag{484a}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the $p \times p$ lower triangular Toeplitz matrices whose first columns are, respectively, the vectors $[1, -\phi_{p,1}, \ldots, -\phi_{p,p-1}]^T$ and $[\phi_{p,p}, \phi_{p,p-1}, \ldots, \phi_{p,1}]^T$ (Pagano, 1973; Godolphin and Unwin, 1983; Newton, 1988). This formulation for $\sigma_p^2 \boldsymbol{\Gamma}_p^{-1}$ is more convenient to work with at times than that implied by Equation (464d).

Interestingly the Yule–Walker, Burg, FLS, BLS and FBLS estimators of the AR(p) parameters *all* share the *same* large-sample distribution as the MLEs. This result suggests that, given enough data, there is no difference in the performance of all the estimators we have studied; however, small sample studies have repeatedly shown, for example, that the Yule–Walker estimator can be quite poor compared to either the Burg or FBLS estimators and that, for certain narrow-band processes, the Burg estimator is inferior to the FBLS estimator. This discrepancy points out the limitations of large-sample results (see Lysne and Tjøstheim, 1987, for further discussion on these points and also on a second-order large-sample theory that explains some of the differences between the Yule–Walker and LS estimators).

[2] As we have already noted, the fact that we must resort to a nonlinear optimizer to use the maximum likelihood method for AR(p) parameter estimation limits its use in practice. Much effort therefore has gone into finding good approximations to MLEs that are easy to compute. These alternative estimators are useful not only by themselves, but also as the initial guesses at the AR parameters that are required by certain numerical optimization routines.

The FLS, BLS and FBLS estimators can all be regarded as approximate MLEs. To see this in the case of the FLS estimator, we first simplify the reduced likelihood of Equation (482a) by dropping the $\sum \log\left(\lambda_j\right)$ term to obtain

$$l(\boldsymbol{\Phi}_p \mid \boldsymbol{H}) \approx N \log\left(\mathrm{SS}_{(\mathrm{ML})}(\boldsymbol{\Phi}_p)/N\right) + N + N \log(2\pi).$$

A justification for this simplification is that the deleted term becomes negligible as N gets large. The value of $\boldsymbol{\Phi}_p$ that minimizes the right-hand side above is identical to the value that minimizes the sum of squares $\mathrm{SS}_{(\mathrm{ML})}(\boldsymbol{\Phi}_p)$. If we in turn approximate $\mathrm{SS}_{(\mathrm{ML})}(\boldsymbol{\Phi}_p)$ by dropping the summation of the first p squared prediction errors in Equation (481c) (again negligible as N get large), we obtain an approximate MLE given by the value of $\boldsymbol{\Phi}_p$ that minimizes $\sum_{t=p}^{N-1} \epsilon_t^2$. However, this latter sum of squares is identical to the sum of squares $\mathrm{SS}_{(\mathrm{FLS})}(\cdot)$ of Equation (476), so our approximate estimator turns out to be just the FLS estimator.

Turning now to the BLS estimator, note that the log likelihood function in Equation (480b) for $\boldsymbol{H} = [H_0, \ldots, H_{N-1}]^T$ is *identical* to that for the "time reversed" series $\widetilde{\boldsymbol{H}} \stackrel{\mathrm{def}}{=} [H_{N-1}, \ldots, H_0]^T$. This follows from the fact that

$$\widetilde{\boldsymbol{H}}^T \boldsymbol{\Gamma}_N^{-1} \widetilde{\boldsymbol{H}} = \boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H} \tag{484b}$$

(Exercise [9.17] is to prove the above). Repeating the argument of the previous paragraph for this "time reversed" series leads to the BLS estimator as an approximate MLE. Finally, if we rewrite Equation (480b) as

$$l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H}) = \log\left(|\boldsymbol{\Gamma}_N|\right) + \frac{\boldsymbol{H}^T \boldsymbol{\Gamma}_N^{-1} \boldsymbol{H} + \widetilde{\boldsymbol{H}}^T \boldsymbol{\Gamma}_N^{-1} \widetilde{\boldsymbol{H}}}{2} + N \log(2\pi), \tag{484c}$$

we can obtain the FBLS estimator as an approximate MLE (this is Exercise [9.18]). Due to the symmetric nature of the likelihood function, this particular estimator seems a more natural approximation for Gaussian processes than the other two estimators, both of which are direction dependent. (Here and elsewhere in this chapter, we have considered a time-reversed version $\{X_{-t}\}$ of a stationary process $\{X_t\}$, noting that both processes must have the same ACVS and that, in the Gaussian case, a time series and its reversal must have the same log likelihood function. Weiss, 1975, explores a technical concept called *time-reversibility* that involves invariance of the joint distribution of the RVs in a time series

rather than of just their second-order properties. Lawrance, 1991, notes that "... the key contribution in Weiss (1975) is the proof that ARMA processes with an autoregressive component are reversible if and only if they are Gaussian." This contribution implies that a non-Gaussian time series and its time-reversed version need *not* have the same log likelihood function and that a time series showing evidence of directionality cannot be fully modeled by a Gaussian AR process.)

[3] Kay (1983) proposes a recursive maximum likelihood estimator (RMLE) for the AR(p) parameters that is similar in spirit to Burg's algorithm. One interpretation of Burg's algorithm is that it recursively fits one AR(1) model after another by minimizing the sum of squares of forward and backward prediction errors at each stage. The RMLE also recursively fits AR(1) models but by minimizing a reduced AR(1) likelihood analogous to that of Equation (482b). Fitting an AR(p) model by this method thus requires successively finding the roots of p different cubic equations, which, in contrast to the MLE method, is a computational improvement because it avoids the need for a nonlinear optimizer when $p \geq 2$. Kay (1983) reports on a Monte Carlo study that compares the RMLE method with the FBLS method and indicates that, while the former yields more accurate parameter estimates, the latter gives better estimates of the locations of peaks in an SDF (see the discussion of Table 459 for another example of better parameter estimates failing to translate into a better SDF estimate).

9.9 Confidence Intervals Using AR Spectral Estimators

For each of the nonparametric SDF estimators considered in Chapters 6, 7 and 8, we described how to construct a confidence interval (CI) for the unknown SDF based upon the estimator. Here we consider creating CIs based upon the statistical properties of a parametric AR SDF estimator. As Kaveh and Cooper (1976) point out, since "the AR spectral estimate is obtained through nonlinear operations ..., analytical derivation of its statistical properties is, in general, formidable." Early work in this area include Kromer (1969), Berk (1974), Baggeroer (1976), Sakai (1979) and Reid (1979). Newton and Pagano (1984) use both inverse autocovariances and Scheffé projections to find simultaneous CIs, while Koslov and Jones (1985) and Burshtein and Weinstein (1987, 1988) develop similar approaches. We describe here an approach that is most similar to that of Burshtein and Weinstein, although differing in mathematical and statistical details. (The method described in this section for creating CIs is based on large-sample analytic theory. In Section 11.6 we present an entirely different approach using bootstrapping in our discussion of the ocean wave data.)

As we noted in C&E [1] for Section 9.8, all of the AR parameter estimators we have discussed in this chapter have the same large-sample distribution. Accordingly let $\widehat{\boldsymbol{\Phi}}_p = [\hat{\phi}_{p,1}, \hat{\phi}_{p,2}, \ldots, \hat{\phi}_{p,p}]^T$ and $\hat{\sigma}_p^2$ be any one of these estimators of $\boldsymbol{\Phi}_p = [\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}]^T$ and σ_p^2 but with the stipulation that they are based upon a sample H_0, H_1, ..., H_{N-1} from a Gaussian AR(p) process with zero mean. Recalling that $\stackrel{\mathrm{d}}{=}$ means "equal in distribution," we have, as $N \to \infty$,

$$\sqrt{N}(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p) \stackrel{\mathrm{d}}{=} \mathcal{N}(\mathbf{0}, \sigma_p^2 \boldsymbol{\Gamma}_p^{-1}) \text{ and } \sqrt{N}(\hat{\sigma}_p^2 - \sigma_p^2) \stackrel{\mathrm{d}}{=} \mathcal{N}(0, 2\sigma_p^4), \tag{485a}$$

where $\mathcal{N}(\mathbf{0}, \sigma_p^2 \boldsymbol{\Gamma}_p^{-1})$ denotes a p-dimensional vector of RVs obeying a p-dimensional Gaussian distribution with zero mean vector and covariance matrix $\sigma_p^2 \boldsymbol{\Gamma}_p^{-1}$, and $\boldsymbol{\Gamma}_p$ is the $p \times p$ covariance matrix of Equation (450c); moreover, $\hat{\sigma}_p^2$ is independent of all the elements of $\widehat{\boldsymbol{\Phi}}_p$ (Newton, 1988; Brockwell and Davis, 1991). If we let $\hat{s}_\tau$ stand for the estimator of s_τ derived from $\widehat{\boldsymbol{\Phi}}_p$ and $\hat{\sigma}_p^2$, we can obtain a consistent estimator $\widehat{\boldsymbol{\Gamma}}_p$ of $\boldsymbol{\Gamma}_p$ by replacing s_τ in $\boldsymbol{\Gamma}_p$ with $\hat{s}_\tau$. These results are the key to the methods of calculating asymptotically correct CIs for AR spectra that we now present.

Let us rewrite the expression in Equation (446b) for the SDF of a stationary AR(p) process as

$$S(f) = \frac{\sigma_p^2 \, \Delta_\mathrm{t}}{\left|1 - \boldsymbol{e}^H(f) \boldsymbol{\Phi}_p\right|^2}, \tag{485b}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

where H denotes complex-conjugate (Hermitian) transpose and

$$\boldsymbol{e}(f) \stackrel{\text{def}}{=} [\mathrm{e}^{\mathrm{i}2\pi f\,\Delta_\mathrm{t}}, \mathrm{e}^{\mathrm{i}4\pi f\,\Delta_\mathrm{t}}, \ldots, \mathrm{e}^{\mathrm{i}2p\pi f\,\Delta_\mathrm{t}}]^T.$$

Let $\boldsymbol{e}_\Re(f)$ and $\boldsymbol{e}_\Im(f)$ be vectors containing, respectively, the real and imaginary parts of $\boldsymbol{e}(f)$, and define the $p \times 2$ matrix $\boldsymbol{E}$ as

$$\boldsymbol{E}^T = [\boldsymbol{e}_\Re(f) \mid \boldsymbol{e}_\Im(f)]^T = \begin{bmatrix} \cos(2\pi f\,\Delta_\mathrm{t}) & \cdots & \cos(2p\pi f\,\Delta_\mathrm{t}) \\ \sin(2\pi f\,\Delta_\mathrm{t}) & \cdots & \sin(2p\pi f\,\Delta_\mathrm{t}) \end{bmatrix}.$$

We assume for now that $0 < |f| < f_\mathcal{N}$. It follows from Exercise [2.2] and Equation (485a) that the 2×1 vector

$$\boldsymbol{B} \stackrel{\text{def}}{=} \boldsymbol{E}^T(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p) \tag{486a}$$

has, for large N, a variance given by $\sigma_p^2 \boldsymbol{E}^T(\boldsymbol{\Gamma}_p^{-1}/N)\boldsymbol{E}$ and hence asymptotically

$$\boldsymbol{B} \stackrel{\text{d}}{=} \mathcal{N}(\boldsymbol{0}, \sigma_p^2 \mathcal{Q}), \quad \text{where } \mathcal{Q} \stackrel{\text{def}}{=} \boldsymbol{E}^T(\boldsymbol{\Gamma}_p^{-1}/N)\boldsymbol{E}. \tag{486b}$$

Under the conditions specified above, we know from the Mann–Wald theorem (Mann and Wald, 1943; see also Bruce and Martin, 1989) that, asymptotically,

$$\frac{\boldsymbol{B}^T \hat{\mathcal{Q}}^{-1} \boldsymbol{B}}{\hat{\sigma}_p^2} \stackrel{\text{d}}{=} \chi_2^2, \quad \text{where } \hat{\mathcal{Q}} \stackrel{\text{def}}{=} \boldsymbol{E}^T(\widehat{\boldsymbol{\Gamma}}_p^{-1}/N)\boldsymbol{E}. \tag{486c}$$

From the definition of $\boldsymbol{B}$ in Equation (486a), we can write

$$\frac{\boldsymbol{B}^T \hat{\mathcal{Q}}^{-1} \boldsymbol{B}}{\hat{\sigma}_p^2} = \frac{(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)^T \hat{\mathcal{Q}}^{-1}(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)}{\hat{\sigma}_p^2} \stackrel{\text{d}}{=} \chi_2^2$$

when $0 < |f| < f_\mathcal{N}$. Hence,

$$\mathbf{P}\left[\frac{(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)^T \hat{\mathcal{Q}}^{-1}(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)}{\hat{\sigma}_p^2} \le Q_2(1-\alpha)\right] = 1 - \alpha, \tag{486d}$$

where $Q_\nu(\alpha)$ is the $\alpha \times 100\%$ percentage point of the χ_ν^2 distribution. Let $\mathcal{A}_0$ represent the event displayed between the brackets above. An equivalent description for $\mathcal{A}_0$ is that the true value of the 2×1 vector $\boldsymbol{E}^T\boldsymbol{\Phi}_p$ lies inside the ellipsoid defined as the set of vectors, $\boldsymbol{E}^T\tilde{\boldsymbol{\Phi}}_p$ say, satisfying

$$(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\tilde{\boldsymbol{\Phi}}_p)^T \mathcal{M}(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\tilde{\boldsymbol{\Phi}}_p) \le 1,$$

where $\mathcal{M} \stackrel{\text{def}}{=} \hat{\mathcal{Q}}^{-1}/(\hat{\sigma}_p^2 Q_2(1-\alpha))$. Scheffé (1959, p. 407) shows that $\boldsymbol{E}^T\tilde{\boldsymbol{\Phi}}_p$ is in this ellipsoid if and only if

$$\left|\boldsymbol{a}^T(\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)\right|^2 \le \boldsymbol{a}^T\mathcal{M}^{-1}\boldsymbol{a} \tag{486e}$$

for *all* two-dimensional vectors $\boldsymbol{a}$, giving us another equivalent description for $\mathcal{A}_0$. Hence, specializing to two specific values for $\boldsymbol{a}$, the occurrence of the event $\mathcal{A}_0$ implies the occurrence of the following two events:

$$\left|[1,0](\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)\right|^2 = \left|\boldsymbol{e}_\Re^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \le [1,0]\mathcal{M}^{-1}[1,0]^T$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

and

$$\left|[0,1](\boldsymbol{E}^T\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{E}^T\boldsymbol{\Phi}_p)\right|^2 = \left|\boldsymbol{e}_{\Im}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \leq [0,1]\mathcal{M}^{-1}[0,1]^T.$$

These two events in turn imply the occurrence of the event $\mathcal{A}_1$, defined by

$$\left|\boldsymbol{e}_{\Re}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 + \left|\boldsymbol{e}_{\Im}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \leq [1,0]\mathcal{M}^{-1}[1,0]^T + [0,1]\mathcal{M}^{-1}[0,1]^T.$$

Let us rewrite both sides of this inequality. For the left-hand side, we can write

$$\begin{aligned}\left|\boldsymbol{e}_{\Re}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 + \left|\boldsymbol{e}_{\Im}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 &= \left|\boldsymbol{e}_{\Re}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p) - \mathrm{i}\boldsymbol{e}_{\Im}^T(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \\ &= \left|\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2.\end{aligned}$$

For the right-hand side, we let $\boldsymbol{d}^H = [1, -\mathrm{i}]$ and use the fact that the 2×2 matrix $\mathcal{M}$ is symmetric to obtain

$$[1,0]\mathcal{M}^{-1}[1,0]^T + [0,1]\mathcal{M}^{-1}[0,1]^T = \boldsymbol{d}^H\mathcal{M}^{-1}\boldsymbol{d}.$$

Hence the event $\mathcal{A}_1$ is equivalent to

$$\left|\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \leq \boldsymbol{d}^H\mathcal{M}^{-1}\boldsymbol{d}, \text{ or, } \frac{\left|\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2}{\boldsymbol{d}^H(\hat{\mathcal{Q}}\hat{\sigma}_p^2)\boldsymbol{d}} \leq Q_2(1-\alpha).$$

Because the occurrence of the event $\mathcal{A}_0$ implies the occurrence of the event $\mathcal{A}_1$, it follows that $\mathcal{A}_0$ is a subset of $\mathcal{A}_1$. We thus must have $\mathbf{P}[\mathcal{A}_0] \leq \mathbf{P}[\mathcal{A}_1]$. Since Equation (486d) states that $\mathbf{P}[\mathcal{A}_0] = 1 - \alpha$, it follows that $\mathbf{P}[\mathcal{A}_1] \geq 1 - \alpha$, i.e., that

$$\mathbf{P}\left[\left|\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 \leq Q_2(1-\alpha)\boldsymbol{d}^H(\hat{\mathcal{Q}}\hat{\sigma}_p^2)\boldsymbol{d}\right] \geq 1 - \alpha. \tag{487}$$

Let us define

$$G(f) = 1 - \boldsymbol{e}^H(f)\boldsymbol{\Phi}_p \text{ so that } |G(f)|^2 = \frac{\sigma_p^2\,\Delta_{\mathrm{t}}}{S(f)},$$

because of Equation (485b). With $\hat{G}(f) = 1 - \boldsymbol{e}^H(f)\widehat{\boldsymbol{\Phi}}_p$, we have

$$\left|\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Phi}}_p - \boldsymbol{\Phi}_p)\right|^2 = \left|\hat{G}(f) - G(f)\right|^2.$$

Recalling that $\boldsymbol{d}^H = [1, -\mathrm{i}]$ and using the definition for $\hat{\mathcal{Q}}$ in Equation (486c), we obtain

$$\boldsymbol{d}^H\hat{\mathcal{Q}}\boldsymbol{d} = \boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Gamma}}_p^{-1}/N)\boldsymbol{e}(f)$$

since $\boldsymbol{d}^H\boldsymbol{E}^T = \boldsymbol{e}^H(f)$. We can now rewrite Equation (487) as

$$\mathbf{P}\left[|\hat{G}(f) - G(f)|^2 \leq Q_2(1-\alpha)\hat{\sigma}_p^2\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Gamma}}_p^{-1}/N)\boldsymbol{e}(f)\right] \geq 1 - \alpha,$$

from which we conclude that asymptotically the transfer function $G(f)$ of the true AR filter resides within a circle of radius

$$\hat{r}_2(f) \stackrel{\text{def}}{=} \left[Q_2(1-\alpha)\hat{\sigma}_p^2\boldsymbol{e}^H(f)(\widehat{\boldsymbol{\Gamma}}_p^{-1}/N)\boldsymbol{e}(f)\right]^{1/2}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

with probability at least $1-\alpha$. Equivalently, as can readily be demonstrated geometrically,

$$\mathbf{P}\left[C_{\mathrm{L}} \le |G(f)|^{-2} \le C_{\mathrm{U}}\right] \ge 1-\alpha, \tag{488a}$$

where

$$C_{\mathrm{L}} \stackrel{\mathrm{def}}{=} \left[|\hat{G}(f)| + \hat{r}_2(f)\right]^{-2} \text{ and } C_{\mathrm{U}} \stackrel{\mathrm{def}}{=} \left[|\hat{G}(f)| - \hat{r}_2(f)\right]^{-2}.$$

Note that the quantity $\hat{\sigma}_p^2 \widehat{\boldsymbol{I}}_p^{-1}$ appearing in $\hat{r}_2(f)$ can be readily computed using the sampling version of Equation (484a) (other relevant computational details can be found in Burshtein and Weinstein, 1987).

Equation (488a) gives CIs for $|G(f)|^{-2} = S(f)/(\sigma_p^2\,\Delta_{\mathrm{t}})$, the normalized spectrum. These CIs come from an approach similar to that of Burshtein and Weinstein (1987), who note an adaptation to give asymptotically valid CIs for $S(f)$, i.e., without normalization. To see how, assume for the duration of this paragraph that $\hat{\sigma}_p^2$ is the MLE $\hat{\sigma}_{(\mathrm{ML})}^2$ of Equation (481d). Brockwell and Davis (1991) suggest approximating the distribution of $N\hat{\sigma}_p^2/\sigma_p^2$ by that of a χ_{N-p}^2 RV, which leads to

$$\mathbf{P}\left[D_{\mathrm{L}} \le \sigma_p^2 \le D_{\mathrm{U}}\right] = 1-\alpha, \text{ where } D_{\mathrm{L}} \stackrel{\mathrm{def}}{=} \frac{N\hat{\sigma}_p^2}{Q_{N-p}\left(1-\frac{\alpha}{2}\right)} \text{ and } D_{\mathrm{U}} \stackrel{\mathrm{def}}{=} \frac{N\hat{\sigma}_p^2}{Q_{N-p}\left(\frac{\alpha}{2}\right)}.$$

Let $\mathcal{C}$ and $\mathcal{D}$ denote the events in the brackets in, respectively, Equation (488a) and the above equation. Because $\hat{\sigma}_p^2$ is independent of all the elements of $\widehat{\boldsymbol{\Phi}}_p$, the events $\mathcal{C}$ and $\mathcal{D}$ are independent and hence

$$\mathbf{P}\left[\mathcal{C} \cap \mathcal{D}\right] = \mathbf{P}\left[\mathcal{C}\right] \times \mathbf{P}\left[\mathcal{D}\right] \ge (1-\alpha)^2.$$

The events $\mathcal{C}$ and $\mathcal{D}$ are equivalent to the events

$$\log\left(C_{\mathrm{L}}\right) \le \log\left(|G(f)|^{-2}\right) \le \log\left(C_{\mathrm{U}}\right) \text{ and } \log\left(D_{\mathrm{L}}\right) \le \log\left(\sigma_p^2\right) \le \log\left(D_{\mathrm{U}}\right).$$

The occurrence of both $\mathcal{C}$ and $\mathcal{D}$ implies the occurrence of

$$\log\left(C_{\mathrm{L}} D_{\mathrm{L}}\,\Delta_{\mathrm{t}}\right) \le \log\left(\frac{\sigma_p^2\,\Delta_{\mathrm{t}}}{|G(f)|^2}\right) = \log\left(S(f)\right) \le \log\left(C_{\mathrm{U}} D_{\mathrm{U}}\,\Delta_{\mathrm{t}}\right),$$

which we denote as event $\mathcal{B}$ and which is equivalent to $C_{\mathrm{L}} D_{\mathrm{L}}\,\Delta_{\mathrm{t}} \le S(f) \le C_{\mathrm{U}} D_{\mathrm{U}}\,\Delta_{\mathrm{t}}$. Since $\mathcal{C} \cap \mathcal{D}$ is a subset of event $\mathcal{B}$, we have $\mathbf{P}\left[\mathcal{B}\right] \ge \mathbf{P}\left[\mathcal{C} \cap \mathcal{D}\right] \ge (1-\alpha)^2$, which establishes

$$\mathbf{P}\left[\frac{\left[N\hat{\sigma}_p^2\,\Delta_{\mathrm{t}}/Q_{N-p}\left(1-\frac{\alpha}{2}\right)\right]}{\left[|\hat{G}(f)| + \hat{r}_2(f)\right]^2} \le S(f) \le \frac{\left[N\hat{\sigma}_p^2\,\Delta_{\mathrm{t}}/Q_{N-p}\left(\frac{\alpha}{2}\right)\right]}{\left[|\hat{G}(f)| - \hat{r}_2(f)\right]^2}\right] \ge (1-\alpha)^2. \tag{488b}$$

Note that the numerators in the event expressed above reflect the uncertainty due to estimation of the innovation variance, while the denominators reflect that of the AR coefficients. Setting $\alpha = 1-\sqrt{0.95} \doteq 0.0253$ yields a CI for $S(f)$ with an associated probability of at least 0.95. (All the estimators of σ_p^2 we have discussed in this chapter are asymptotically equivalent to the MLE $\hat{\sigma}_{(\mathrm{ML})}^2$, so Equation (488b) also holds when adapted to make use of them.)

Figures 489 and 490 show examples of pointwise CIs created using Equation (488b) with α set to $1-\sqrt{0.95}$ to yield a probability of at least 0.95 (see the analysis of the ocean wave data in Section 9.12 for an additional example). The thin solid curve in each plot shows an MLE-based AR(p) SDF estimate using a time series generated by an AR(p) process; the solid

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

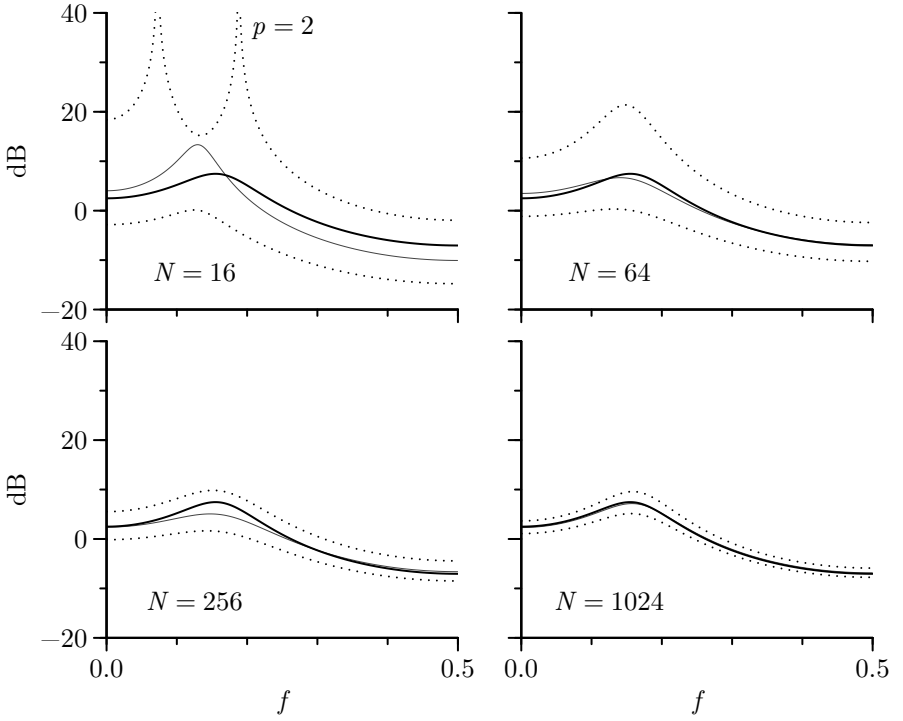

Figure 489 Maximum likelihood AR(2) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(2) process shown in Figure 34(a) (the process is defined in Equation (34)). The thick curve on each plot is the true SDF. The dotted curves show the lower and upper limits of pointwise confidence intervals with an associated probability of at least 0.95 based upon Equation (488b).

curve is the corresponding true AR(p) SDF; and the two dotted curves depict the lower and upper limits of the CIs. Figure 489 makes use of the AR(2) time series shown in Figure 34(a), with four AR(2) SDF estimates formed using the first $N = 16$, $N = 64$ and $N = 256$ values of the series and then all $N = 1024$ values. As is intuitively reasonable, the lower CI limits get progressively closer to the true AR(2) SDF as N increases. The upper limits do also *if* we ignore the $N = 16$ case. The latter has two peaks associated with two frequencies at which $|\hat{G}(f)| = \hat{r}_2(f)$, thus causing the upper limit in Equation (488b) to be infinite. While an infinite upper limit is not inconsistent with the inequality $\mathbf{P}[\mathcal{B}] \geq 0.95$ behind the CIs, this limit is not particularly informative and is an indication that the asymptotic theory behind Equation (488b) does not yield a decent approximation for such a small time series ($N = 16$). This deficiency in the upper limits for small sample sizes is also apparent in Figure 490, which makes use of the AR(4) time series of Figure 34(e). The $N = 16$ case is particularly unsettling: in addition to two frequencies at which the upper limits are infinite, there is an interval surrounding $f = 0.12$ over which the AR(4) SDF point estimates are more than an order of magnitude *outside* of the CIs! The $N = 64$ case also has point estimates outside of the CIs, but this strange pattern does not occur for the $N = 256$ and $N = 1024$ cases. The upper limits are all finite when the entire AR(4) time series is utilized ($N = 1024$), but these limits are infinite for the $N = 64$ case at two frequencies and for $N = 256$ at four frequencies (in the bizarre pattern of double twin peaks). Evidently we need a larger sample size in the AR(4) case than in the AR(2) for the asymptotic theory to offer a decent approximation, which is not surprising. For $N = 1024$ the widths of the CIs at frequencies just above 0 and just below 0.5 in the AR(2) case are, respectively, 2.6 dB and 1.9 dB; counterintuitively, they are somewhat

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

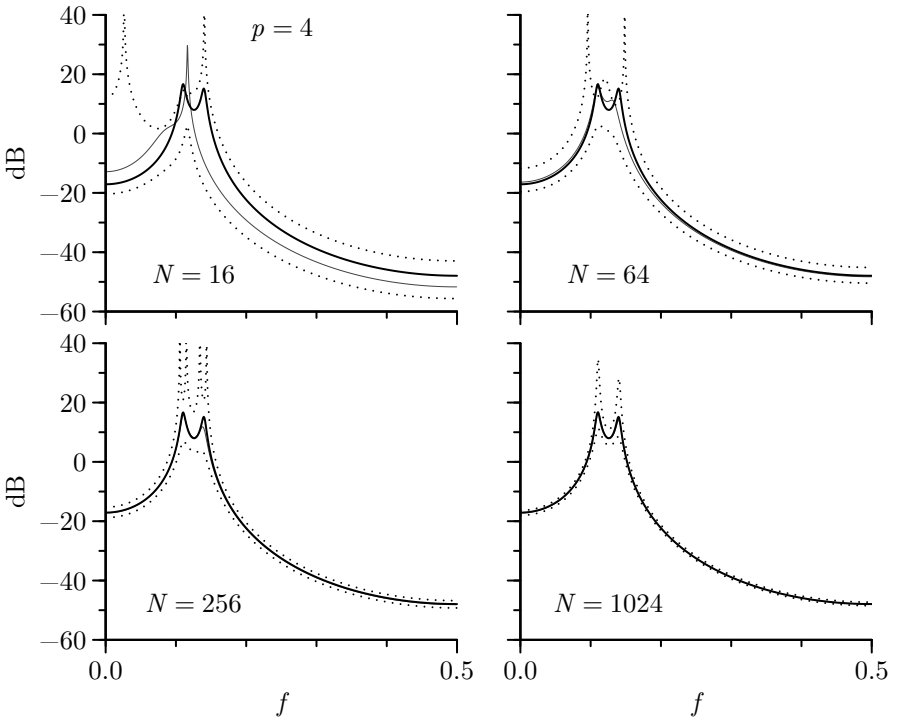

Figure 490 Maximum likelihood AR(4) SDF estimates (thin curves) for portions of lengths 16, 64, 256 and 1024 of the realization of the AR(4) process shown in Figure 34(e) (the process is defined in Equation (35a), and the estimates are also shown in Figure 483). The thick curve on each plot is the true SDF. The dotted curves show the lower and upper limits of pointwise confidence intervals with an associated probability of at least 0.95 based upon Equation (488b).

smaller in the more challenging AR(4) case (1.7 dB and 1.2 dB; for a possible explanation, see the second paragraph in C&E [2], noting that the spectral levels for the AR(4) process near $f = 0$ and $f = 0.5$ are considerably lower than those in the AR(2) case).

Comments and Extensions to Section 9.9

[1] In the theory presented in this section, we assumed that the frequency of interest f satisfies $0 < |f| < f_{\mathcal{N}}$. If in fact f is equal to zero or the Nyquist frequency, then $\boldsymbol{e}_{\Im}(f) = \mathbf{0}$, and the rank of the quadratic form $\boldsymbol{B}^T \hat{\mathcal{Q}}^{-1} \boldsymbol{B} / \hat{\sigma}_p^2$ is one (rather than two as previously), so that now

$$\frac{\boldsymbol{B}^T \hat{\mathcal{Q}}^{-1} \boldsymbol{B}}{\hat{\sigma}_p^2} \stackrel{\mathrm{d}}{=} \chi_1^2, \qquad f = 0 \text{ or } \pm f_{\mathcal{N}}.$$

The only adjustment that is needed in our development is to replace $Q_2(1-\alpha)$ with $Q_1(1-\alpha)$, i.e., use the $(1-\alpha) \times 100\%$ percentage point of the χ_1^2 distribution.

[2] For a Gaussian stationary process $\{G_t\}$ with an SDF $S_G(\cdot)$ that is strictly positive, bounded and sufficiently smooth, Kromer (1969) shows that, provided that $N \to \infty$ and $p \to \infty$ in such a manner that $N/p \to \infty$, then

$$\frac{\hat{S}_G^{(\mathrm{YW})}(f) - S_G(f)}{(2p/N)^{1/2} S_G(f)} \stackrel{\mathrm{d}}{=} \begin{cases} \mathcal{N}(0,1), & 0 < |f| < f_{\mathcal{N}}; \\ \mathcal{N}(0,2), & f = 0 \text{ or } \pm f_{\mathcal{N}}. \end{cases} \tag{490}$$

Comparing the above with the asymptotic distribution of lag window spectral estimators (Section 7.4) reveals that, for $f \neq 0$ or $\pm f_{\mathcal{N}}$, the value N/p in AR spectral estimation plays the role of ν, the equivalent number of degrees of freedom, in lag window spectral estimation.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

A comparison of Equation (488b) with Kromer's result of Equation (490) is not particularly easy. The reader is referred to Burshtein and Weinstein (1987, p. 508) for one possible approach, which leads to the conclusion that in both cases the width of the CI at f is proportional to the spectral level at the frequency; however, the constant of proportionality is somewhat different in the two cases.

9.10 Prewhitened Spectral Estimators

In the previous sections we have concentrated on estimating the parameters of an AR(p) process. We can in turn use these estimators to estimate the SDF for a time series by taking a *pure parametric approach*, in which we simply substitute the estimated AR parameters for the corresponding theoretical parameters in Equation (446b). Alternatively, we can employ a *prewhitening approach*, in which we use the estimates of the AR coefficients $\phi_{p,j}$ to form a prewhitening filter. We introduced the general idea behind prewhitening in Section 6.5. Here we give details on how to implement prewhitening using an estimated AR model and then comment on the relative merits of the pure parametric and prewhitening approaches.

Given a time series that is a realization of a portion $X_0, \ldots, X_{N-1}$ of a stationary process with zero mean and SDF $S_X(\cdot)$, the AR-based prewhitening approach consists of the following steps.

[1] We begin by fitting an AR(p) model to the time series using, say, Burg's algorithm to obtain $\bar{\phi}_{p,1}, \ldots, \bar{\phi}_{p,p}$ and $\bar{\sigma}_p^2$ (as noted by an example in Section 9.12, the particular estimator we use can be important; we have found the Burg method to work well in most practical situations). We set the order p to be just large enough to capture the general structure of the SDF $S_X(\cdot)$ (see Section 9.11 for a discussion of subjective and objective methods for selecting p). In any case, p should not be large compared to the sample size N.

[2] We then use the fitted AR(p) model as a prewhitening filter. The output from this filter is a time series of length $N - p$ given by

$$\overrightarrow{e}_t(p) = X_t - \bar{\phi}_{p,1} X_{t-1} - \cdots - \bar{\phi}_{p,p} X_{t-p}, \qquad t = p, \ldots, N-1,$$

where, as before, $\overrightarrow{e}_t(p)$ is the forward prediction error observed at time t. If we let $S_e(\cdot)$ denote the SDF for $\{\overrightarrow{e}_t(p)\}$, linear filtering theory tells us that the relationship between $S_e(\cdot)$ and $S_X(\cdot)$ is given by

$$S_e(f) = \left| 1 - \sum_{j=1}^{p} \bar{\phi}_{p,j} \mathrm{e}^{-\mathrm{i}2\pi f j \, \Delta_\mathrm{t}} \right|^2 S_X(f). \tag{491a}$$

[3] If the prewhitening filter has been appropriately chosen, the SDF $S_e(\cdot)$ for the observed forward prediction errors will have a smaller dynamic range than $S_X(\cdot)$. We can thus produce a direct spectral estimate $\hat{S}_e^{(\mathrm{D})}(\cdot)$ for $S_e(\cdot)$ with good bias properties using little or no tapering. If we let $\{\hat{s}_{e,\tau}^{(\mathrm{D})}\}$ represent the corresponding ACVS estimate, we can then produce an estimate of $S_e(\cdot)$ with decreased variability by applying a lag window $\{w_{m,\tau}\}$ to obtain

$$\hat{S}_{e,m}^{(\mathrm{LW})}(f) = \Delta_\mathrm{t} \sum_{\tau=-(N-p-1)}^{N-p-1} w_{m,\tau} \hat{s}_{e,\tau}^{(\mathrm{D})} \mathrm{e}^{-\mathrm{i}2\pi f \tau \, \Delta_\mathrm{t}}.$$

[4] Finally we estimate $S_X(\cdot)$ by "postcoloring" $\hat{S}_{e,m}^{(\mathrm{LW})}(\cdot)$ to obtain

$$\hat{S}_X^{(\mathrm{PC})}(f) = \frac{\hat{S}_{e,m}^{(\mathrm{LW})}(f)}{\left| 1 - \sum_{j=1}^{p} \bar{\phi}_{p,j} \mathrm{e}^{-\mathrm{i}2\pi f j \, \Delta_\mathrm{t}} \right|^2}. \tag{491b}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

There are two obvious variations on this technique that have seen some use. First, we can make use of the observed backward prediction errors to produce a second estimator similar to $\hat{S}^{(\mathrm{LW})}_{e,m}(\cdot)$. These two estimators can be averaged together prior to postcoloring in step [4]. Second, we can obtain the first p values of the ACVS corresponding to $\hat{S}^{(\mathrm{PC})}_X(\cdot)$ and use them in the Yule–Walker method to produce a refined prewhitening filter for iterative use in step [2].

Here are five notes about this combined parametric/nonparametric approach to spectral estimation.

[1] The chief difference between this approach and a pure parametric approach is that we no longer regard the observed prediction errors $\{\overrightarrow{e}_t(p)\}$ as white noise, but rather we use a nonparametric approach to estimate their SDF.

[2] The problem of selecting p is lessened: any imperfections in the prewhitening filter can be compensated for in the nonparametric portion of this combined approach.

[3] Since all lag window spectral estimators correspond to an ACVS estimator that is identically zero after a finite lag q (see Equation (248b)), the numerator of $\hat{S}^{(\mathrm{PC})}_X(\cdot)$ has the form of an MA(q) SDF with $q \leq N - p - 1$. Hence the estimator $\hat{S}^{(\mathrm{PC})}_X(\cdot)$ is the SDF for some ARMA(p, q) model. Our combined approach is thus a way of implicitly fitting an ARMA model to a time series (see the end of Section 9.14 for an additional comment).

[4] The combined approach only makes sense in situations where tapering is normally required; i.e., the SDF $S_X(\cdot)$ has a high dynamic range (see the discussion in Section 6.5).

[5] Even if we use a pure parametric approach, it is useful to carefully examine the observed prediction errors $\{\overrightarrow{e}_t(p)\}$ because this is a valuable way to detect outliers in a time series (Martin and Thomson, 1982).

See Section 9.12 for an example of the prewhitening and pure parametric approaches to SDF estimation.

9.11 Order Selection for AR(p) Processes

In discussing the various estimators of the parameters for an AR(p) process, we have implicitly assumed that the model order p is known in advance, an assumption that is rarely valid in practice. To make a reasonable choice of p, various order selection criteria have been proposed and studied in the literature. We describe briefly here a few commonly used ones, but these are by no means the only ones that have been proposed or are in extensive use. The reader is referred to the articles by de Gooijer et al. (1985), Broersen (2000, 2002) and Stoica and Selén (2004) and to the books by Choi (1992), Stoica and Moses (2005) and Broersen (2006) for more comprehensive discussions.

In what follows, we first consider two subjective criteria (direct spectral estimate and partial autocorrelation) and then several objective criteria. Examples of their use are given in Section 9.12. In describing the criteria, we take $\hat{\phi}_{k,j}$ and $\hat{\sigma}^2_k$ to be any of the estimators of $\phi_{k,j}$ and σ^2_k we have discussed so far in this chapter.

A comment is in order before we proceed. Any order selection method we use should be appropriate for what we intend to do with the fitted AR model. We usually fit AR models in the context of spectral analysis either to directly obtain an estimate of the SDF (a pure parametric approach) or to produce a prewhitening filter (a combined parametric/nonparametric approach). Some commonly used order selection criteria are geared toward selecting a low-order AR model that does well for one-step-ahead predictions. These criteria thus seem to be more appropriate for producing a prewhitening filter than for pure parametric spectral estimation. Unfortunately, selection of an appropriate order is vital for pure parametric spectral estimation: if p is too large, the resulting SDF estimate tends to exhibit spurious peaks; if p is too small, structure in the SDF can be smoothed over.

Direct Spectral Estimate Criterion
If we are primarily interested in using a fitted AR model to produce a prewhitening filter, the following subjective criterion is viable. First, we compute a direct spectral estimate $\hat{S}^{(\mathrm{D})}(\cdot)$ using a data taper that provides good leakage protection (see Section 6.4). We then fit a sequence of relatively low-order AR models (starting with, say, $k = 2$ or 4) to our time series, compute the corresponding AR SDF estimates via Equation (446b), and compare these estimates with $\hat{S}^{(\mathrm{D})}(\cdot)$. We select our model order p as the smallest value k such that the corresponding AR SDF estimate generally captures the overall shape of $\hat{S}^{(\mathrm{D})}(\cdot)$.

Partial Autocorrelation Criterion
For an AR process of order p, the PACS $\{\phi_{k,k} : k = 1, 2, \ldots\}$ is nonzero for $k = p$ and zero for $k > p$. In other words, the PACS of a pth-order AR process has a cutoff after lag p. It is known that, for a Gaussian AR(p) process, the $\hat{\phi}_{k,k}$ terms for $k > p$ are approximately independently distributed with zero mean and a variance of approximately $1/N$ (see Kay and Makhoul, 1983, and references therein). Thus a rough procedure for testing $\phi_{k,k} = 0$ is to examine whether $\hat{\phi}_{k,k}$ lies between $\pm 2/\sqrt{N}$. By plotting $\hat{\phi}_{k,k}$ versus k, we can thus set p to a value beyond which $\phi_{k,k}$ can be regarded as being zero.

Final Prediction Error (FPE) Criterion
Equation (462c) notes that, for an AR(p) process, the mean square linear prediction errors of orders $p - 1$ and higher obey the pattern

$$\sigma_{p-1}^2 > \sigma_p^2 = \sigma_{p+1}^2 = \sigma_{p+2}^2 = \cdots, \tag{493a}$$

where σ_k^2 for $k > p$ is defined via the augmented Yule–Walker equations of order k. This suggests a criterion for selecting p that consists of plotting $\hat{\sigma}_k^2$ versus k and setting p equal to that value of k such that

$$\hat{\sigma}_{k-1}^2 > \hat{\sigma}_k^2 \approx \hat{\sigma}_{k+1}^2 \approx \hat{\sigma}_{k+2}^2 \approx \cdots. \tag{493b}$$

Recall, however, that, for the Yule–Walker method and Burg's algorithm,

$$\hat{\sigma}_k^2 = \hat{\sigma}_{k-1}^2 \left(1 - \hat{\phi}_{k,k}^2\right),$$

showing that $\hat{\sigma}_k^2 < \hat{\sigma}_{k-1}^2$ except in the unlikely event that either $\hat{\phi}_{k,k}$ or $\hat{\sigma}_{k-1}^2$ happen to be identically zero. The sequence $\{\hat{\sigma}_k^2\}$ is thus nonincreasing, making it problematic at times to determine where the pattern of Equation (493b) occurs. The underlying problem is that $\hat{\sigma}_k^2$ tends to underestimate σ_k^2, i.e., to be biased toward zero. In the context of the Yule–Walker estimator, Akaike (1969, 1970) proposed the FPE estimator that in effect corrects for this bias. Theoretical justification for this estimator presumes the existence of two time series that are independent of each other, but drawn from the same AR process. The first series is used to estimate the AR coefficients, and these coefficients are in turn used to estimate the mean square linear prediction error – with this estimator being called the FPE – by forming one-step-ahead predictions through the second series. After appealing to the large-sample properties of the coefficient estimators, the FPE estimator for a kth-order AR model is taken to be

$$\mathrm{FPE}(k) = \left(\frac{N + k + 1}{N - k - 1}\right) \hat{\sigma}_k^2 \tag{493c}$$

(Akaike, 1970, equations (4.7) and (7.3)). This form of the FPE assumes that the process mean is *unknown* (the usual case in practical applications) and that we have centered our time

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

series by subtracting the sample mean prior to the estimation of the AR parameters (if the process mean is known, the number of unknown parameters decreases from $k+1$ to k, and the term in parentheses becomes $(N+k)/(N-k)$). Note that, as required, FPE(k) is an inflated version of $\hat{\sigma}_k^2$. The FPE order selection criterion is to set p equal to the value of k that minimizes FPE(k).

How well does the FPE criterion work in practice? Using simulated data, Landers and Lacoss (1977) found that the FPE – used with Burg's algorithm – selected a model order that was insufficient to resolve spectral details in an SDF with sharp peaks. By increasing the model order by a factor 3, they obtained adequate spectral estimates. Ulrych and Bishop (1975) and Jones (1976b) found that, when used with Burg's algorithm, the FPE tends to pick out spuriously high-order models. The same does not happen with the Yule–Walker estimator, illustrating that a particular order selection criterion need not perform equally well for different AR parameter estimators. Ulrych and Clayton (1976) found that the FPE does not work well with short time series, prompting Ulrych and Ooe (1983) to recommend that p be chosen between $N/3$ and $N/2$ for such series. Kay and Marple (1981) report that the results from using the FPE criteria have been mixed, particularly with actual time series rather than simulated AR processes.

Likelihood-Based Criteria

In the context of the maximum likelihood estimator, Akaike (1974) proposed another order selection criterion, known as *Akaike's information criterion* (AIC). For a kth-order AR process with an unknown process mean, the AIC is defined as

$$\text{AIC}(k) = -2\,\log\{\text{maximized likelihood}\} + 2(k+1); \tag{494a}$$

if the process mean is known, the above becomes

$$\text{AIC}(k) = -2\,\log\{\text{maximized likelihood}\} + 2k. \tag{494b}$$

The AIC, which is based on cross-entropy ideas, is very general and applicable in more than just a time series context. For a Gaussian AR(k) process $\{H_t\}$ with known mean, using the MLEs $\widehat{\boldsymbol{\Phi}}_{(\text{ML})}$ and $\hat{\sigma}^2_{(\text{ML})}$ in the right-hand side of Equation (481b) tells us – in conjunction with Equations (480b) and (481d) – that we can write

$$-2\,\log\{\text{maximized likelihood}\} = \sum_{j=0}^{k-1} \log(\hat{\lambda}_j) + N\,\log(\hat{\sigma}^2_{(\text{ML})}) + N + N\,\log(2\pi),$$

where $\hat{\sigma}^2_{(\text{ML})}$ is the MLE of σ_k^2, while $\hat{\lambda}_j$ depends on the MLEs of the $\phi_{k,j}$ terms and not on $\hat{\sigma}^2_{(\text{ML})}$. The first term on the right-hand side becomes negligible for large N, while the last two terms are just constants. If we drop these three terms and allow the use of other estimators $\hat{\sigma}_k^2$ of σ_k^2 besides the MLE, we obtain the usual approximation to Equation (494b) that is quoted in the literature as the AIC for AR processes (see, for example, de Gooijer et al., 1985; Rosenblatt, 1985; or Kay, 1988):

$$\text{AIC}(k) = N\,\log\left(\hat{\sigma}_k^2\right) + 2k \tag{494c}$$

(alternatively, we can justify this formulation by using the conditional likelihood of H_0, …, H_{N-1}, given $H_{-1} = H_{-2} = \cdots = H_{-k} = 0$). The AIC order selection criterion is to set p equal to the value of k that minimizes AIC(k). If the process mean is unknown, the appropriate equation is now

$$\text{AIC}(k) = N\,\log\left(\hat{\sigma}_k^2\right) + 2(k+1), \tag{494d}$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

which amounts to adding two to the right-hand side of Equation (494c) – hence the minimizers of Equations (494c) and (494d) are identical.

How well does the AIC criterion work in practice? Landers and Lacoss (1977), Ulrych and Clayton (1976) and Kay and Marple (1981) report results as mixed as those for the FPE criterion. Hurvich and Tsai (1989) note that the FPE and AIC both tend to select inappropriately large model orders. They argue that there is a bias in the AIC and that correcting for this bias yields a criterion that avoids the tendency to overfit in sample sizes of interest to practitioners. In the practical case of an unknown mean, correcting Equation (494d) for this bias yields the following criterion:

$$\text{AICC}(k) = N \log\left(\hat{\sigma}_k^2\right) + 2\frac{N}{N-(k+1)-1}(k+1) \tag{495a}$$

(if the mean is known, replacing $k+1$ in the right-most term above with k yields the correction for Equation (494c)). As $N \to \infty$, the ratio in the above converges to unity, so asymptotically the AICC and the AIC of Equation (494d) are equivalent.

Comments and Extensions to Section 9.11

[1] When using MLEs for a Gaussian AR(k) process $\{H_t\}$ with an unknown mean μ, $N \log\left(\hat{\sigma}_k^2\right)$ in Equations (494d) and (495a) for the AIC and the AICC is a stand-in for $-2 \log\{\text{maximized likelihood}\}$. The likelihood in question depends upon μ. Rather than actually finding the MLE for μ, a common practice in time series analysis is to use the sample mean to estimate μ. The time series is then centered by subtracting off its sample mean, and the centered time series is treated as if it were a realization of a process with a known mean of zero. After centering, a surrogate more accurate than $N \log\left(\hat{\sigma}_k^2\right)$ would be $-2 \log\left(L(\widehat{\boldsymbol{\Phi}}_{(\text{ML})}, \hat{\sigma}^2_{(\text{ML})} \mid \boldsymbol{H})\right)$, where $L(\widehat{\boldsymbol{\Phi}}_{(\text{ML})}, \hat{\sigma}^2_{(\text{ML})} \mid \boldsymbol{H})$ is the maximum value of $L(\boldsymbol{\Phi}_k, \sigma_k^2 \mid \boldsymbol{H})$ – as defined by Equation (480a) – with $\boldsymbol{H} = \left[H_0, \ldots, H_{N-1}\right]^T$ taken to be the centered time series.

Now suppose we use the Yule–Walker estimators $\widetilde{\boldsymbol{\Phi}}_k$ and $\tilde{\sigma}_k^2$ with the centered series. Rather than using $N \log\left(\tilde{\sigma}_k^2\right)$ in Equations (494d) and (495a) for the AIC and the AICC, we could use $-2 \log\left(L(\widetilde{\boldsymbol{\Phi}}_k, \tilde{\sigma}_k^2 \mid \boldsymbol{H})\right)$, which involves the likelihood function computed under the assumption that its true parameter values are the Yule–Walker estimators. We could also do the same for the Burg estimators $\overline{\boldsymbol{\Phi}}_k$ and $\bar{\sigma}_k^2$. Letting $\widehat{\boldsymbol{\Phi}}_k$ and $\hat{\sigma}_k^2$ now stand for either the MLEs, the Yule–Walker estimators or the Burg estimators, the more accurate AIC and the AICC for a centered time series are

$$\text{AIC}(k) = -2 \log\left(L(\widehat{\boldsymbol{\Phi}}_k, \hat{\sigma}_k^2 \mid \boldsymbol{H})\right) + 2(k+1) \tag{495b}$$

and

$$\text{AICC}(k) = -2 \log\left(L(\widehat{\boldsymbol{\Phi}}_k, \hat{\sigma}_k^2 \mid \boldsymbol{H})\right) + 2\frac{N}{N-(k+1)-1}(k+1). \tag{495c}$$

Since the MLEs are the minimizers of $-2 \log\left(L(\boldsymbol{\Phi}_k, \sigma_k^2 \mid \boldsymbol{H})\right)$, the smaller of the Yule–Walker-based $-2 \log\left(L(\widetilde{\boldsymbol{\Phi}}_k, \tilde{\sigma}_k^2 \mid \boldsymbol{H})\right)$ and of the Burg-based $-2 \log\left(L(\overline{\boldsymbol{\Phi}}_k, \bar{\sigma}_k^2 \mid \boldsymbol{H})\right)$ tells us which of these two estimators yields a likelihood closer to the MLE minimizer.

Can Equations (495b) and (495c) be used with the FLS, BLS and FBLS estimators? As noted in C&E [1] for Section 9.7, the FLS estimator of $\boldsymbol{\Phi}_k$ need not always correspond to a causal AR process (the same is true for BLS and FBLS estimators). The procedure that we have outlined in Section 9.8 for evaluating the likelihood function implicitly assumes causality and hence cannot be used when causality fails to hold. Evaluation of the likelihood function for acausal AR models is a topic in need of further exploration.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

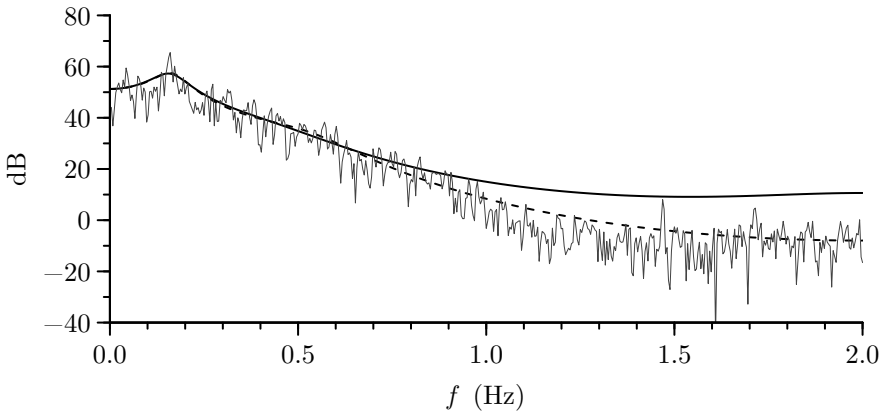

Figure 496 Determination of AR prewhitening filter and illustration of effect of different choices for parameter estimation. The jagged curve shows a Slepian-based direct spectral estimate of the ocean wave data (shown also in the plots of Figure 319). The dashed curve shows the estimated SDF corresponding to an AR(5) model fit to the data using Burg's algorithm. The solid smooth curve shows the estimated SDF for the same model, but now with parameters estimated using the Yule–Walker method.

9.12 Examples of Parametric Spectral Estimators

Ocean Wave Data

Here we return to the ocean wave data $\{X_t\}$ examined previously in Sections 6.8, 7.12 and 8.9. We consider both a prewhitening approach and a pure parametric approach (as in our previous analyses, we first center the time series by subtracting off its sample mean $\overline{X} \doteq 209.1$). For the prewhitening approach, we begin by finding a low-order AR model that can serve as a prewhitening filter. The jagged curve in Figure 496 shows a direct spectral estimate $\hat{S}_X^{(\mathrm{D})}(\cdot)$ for the ocean wave data using an $NW = 2/\Delta_{\mathrm{t}}$ Slepian data taper (this curve is also shown in all four plots of Figure 319 – recall that, based upon a study of Figure 226, we argued that this estimate does not suffer unduly from leakage). Experimentation indicates that an AR(5) model with parameters estimated using Burg's algorithm captures the overall spectral structure indicated by $\hat{S}_X^{(\mathrm{D})}(\cdot)$. The SDF for this model is shown by the dashed curve in Figure 496. For comparison, the solid smooth curve in this plot shows the AR(5) SDF estimate based upon the Yule–Walker method. Note that, whereas the Burg estimate generally does a good job of tracking $\hat{S}_X^{(\mathrm{D})}(\cdot)$ across all frequencies, the same cannot be said for the Yule–Walker estimate – it overestimates the power in the high frequency region by as much as two orders of magnitude (20 dB).

Figure 497(a) shows the forward prediction errors $\{\overrightarrow{e}_t(5)\}$ obtained when we use the fitted AR(5) model to form a prewhitening filter. Since there were $N = 1024$ data points in the original ocean wave time series, there are $N - p = 1019$ points in the forward prediction error series. The jagged curve in Figure 497(b) shows the periodogram $\hat{S}_e^{(\mathrm{P})}(\cdot)$ for $\{\overrightarrow{e}_t(5)\}$ plotted versus the Fourier frequencies. A comparison of $\hat{S}_e^{(\mathrm{P})}(\cdot)$ with direct spectral estimates using a variety of Slepian data tapers indicates that the periodogram is evidently leakage-free, so we can take it to be an approximately unbiased estimate of the SDF $S_e(\cdot)$ associated with $\{\overrightarrow{e}_t(5)\}$ (Equation (491a) relates this SDF to the SDF $S_X(\cdot)$ for the ocean wave series). We now smooth $\hat{S}_e^{(\mathrm{P})}(\cdot)$ using a Parzen lag window $\{w_{m,\tau}\}$ with $m = 55$ to match what we did in Figure 319(a). The smooth curve in Figure 497(b) shows the Parzen lag window estimate $\hat{S}_{e,m}^{(\mathrm{LW})}(\cdot)$. The crisscross in this figure is analogous to the one in Figure 319(a). Its width and height here are based upon the statistical properties of a Parzen lag window with $m = 55$ applied to a time series of length 1019 with no tapering (see Table 279). This yields a lag window spectral estimate with $\nu \doteq 68.7$ equivalent degrees of freedom, whereas we found

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

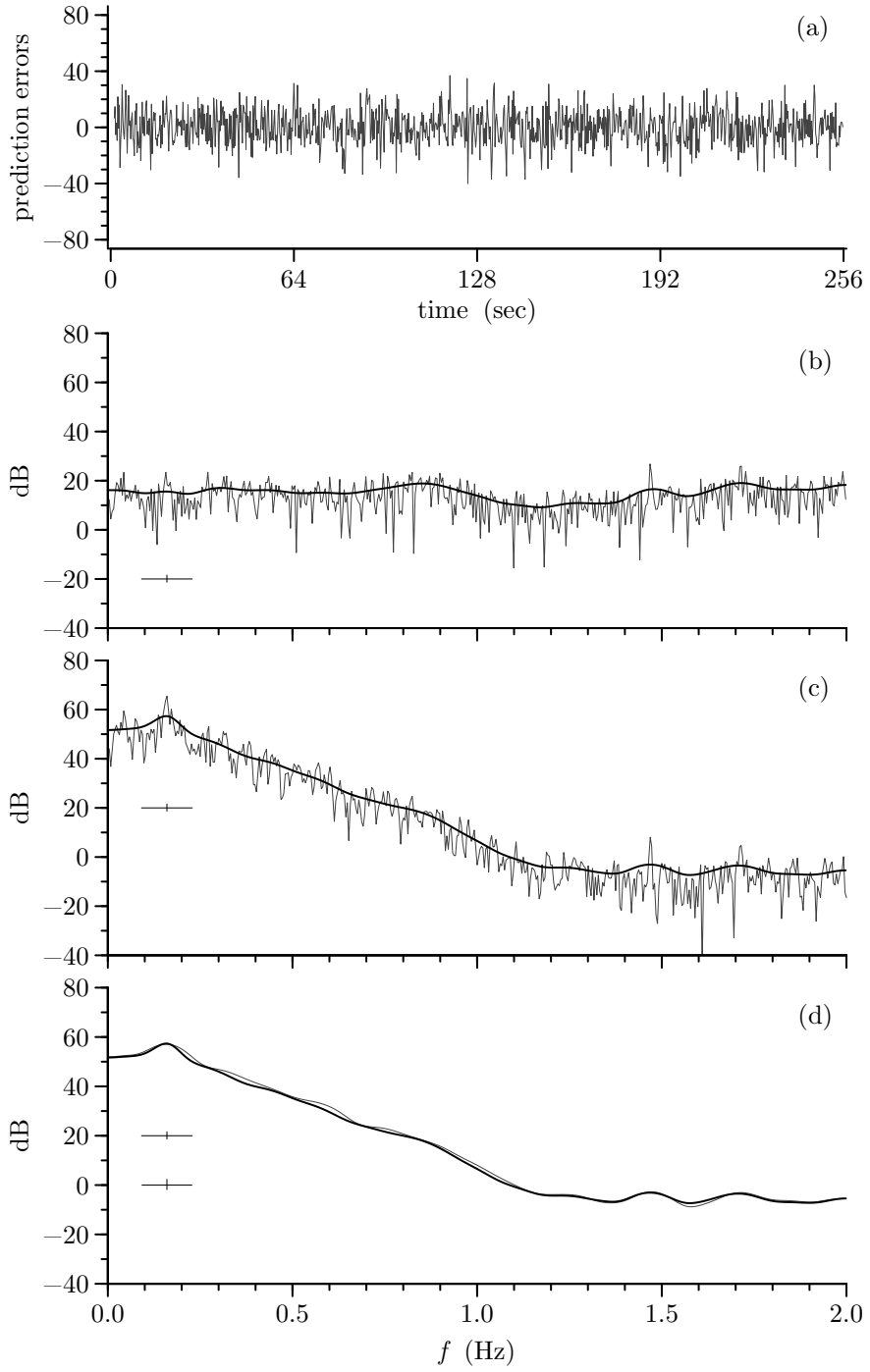

Figure 497 Illustration of prewhitening applied to ocean wave data (see main text for details).

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

$\nu \doteq 35.2$ for the estimate in Figure 319(a), which involved a Slepian data taper. The increase in the number of degrees of freedom occurs because prewhitening has obviated the need for tapering. This increase translates into a decrease in the length of a 95% CI, so that the height of the crisscross in Figure 497(b) is noticeably smaller than the one in Figure 319(a) (the heights are 2.9 dB and 4.1 dB, respectively; these crisscrosses also are replicated in Figure 497(d) – the top one is the same as in Figure 497(b), and the bottom one, as in Figure 319(a)).

The thick smooth curve in Figure 497(c) shows the postcolored lag window estimate $\hat{S}_X^{(\mathrm{PC})}(\cdot)$, which is formed from $\hat{S}_{e,m}^{(\mathrm{LW})}(\cdot)$ as per Equation (491b). The jagged curve is the same as in Figure 496 and shows that $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ can be regarded as a smoothed version of the leakage-free direct spectral estimate $\hat{S}_X^{(\mathrm{D})}(\cdot)$ of the original ocean wave series. The amount of smoothing is appropriate given that our main objective is to determine the rate at which the SDF decreases over the frequency range 0.2 to 1.0 Hz. It is of interest to compare $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ with the $m = 55$ Parzen lag window estimate $\hat{S}_{X,m}^{(\mathrm{LW})}(\cdot)$ shown in Figure 319(a). As we noted in our discussion of that figure, this estimate suffers from smoothing window leakage at the high frequencies. By contrast, there is no evidence of such leakage in $\hat{S}_X^{(\mathrm{PC})}(\cdot)$, illustrating an additional advantage to the combined parametric/nonparametric approach. We also noted that the $m = 23.666$ Gaussian lag window estimate plotted in Figure 319(d) shows no evidence of smoothing window leakage. This estimate is replicated as the thin curve in Figure 497(d) for comparison with $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ (the thick curve). The two estimates are quite similar, differing by no more than 2.3 dB. The top crisscross is for $\hat{S}_X^{(\mathrm{PC})}(\cdot)$, while the bottom one is for the Gaussian lag window estimate. The prewhitening approach yields an estimate with the same bandwidth as the Gaussian estimate, but its 95% CI is noticeably tighter. (Exercise [9.19] invites the reader to redo this analysis using a prewhitening filter dictated by the Yule–Walker method rather than the Burg-based filter used here.)

Let us now consider a pure parametric approach for estimating $S_X(\cdot)$, for which we need to determine an appropriate model order p. As a first step, we examine the sample PACS for the ocean wave data. With K taken to be the highest order of potential interest, the standard definition of the sample PACS is $\{\tilde{\phi}_{k,k} : k = 1, \ldots, K\}$, which is the result of recursively manipulating the usual biased estimator $\{\hat{s}_\tau^{(\mathrm{P})}\}$ of the ACVS (see Equation (458a); these manipulations also yield Yule–Walker parameter estimates for the models AR(1) up to AR(K). As discussed in Section 9.11, the PACS criterion compares the estimated PACS to the approximate 95% confidence bounds $\pm 2/\sqrt{N}$ and sets p to the order beyond which the bulk of the estimates fall within these bounds. The top plot of Figure 499 shows $\{\tilde{\phi}_{k,k} : k = 1, \ldots, 40\}$. The horizontal dashed lines indicate the bounds $\pm 2/\sqrt{N} = \pm 0.0625$. Beyond $k = 6$ only two $\tilde{\phi}_{k,k}$ estimates exceed these limits ($k = 15$ and 26). Since we would expect roughly 5% of the deviates to exceed the limits just due to sampling variability, it is tempting to chalk two rather small exceedances out of 34 possibilities as being due to chance ($2/34 \doteq 6\%$). Indeed minimization of FPE(k), AIC(k) and AICC(k) in Equations (493c), (494a) and (495a) all pick out $k = 6$ as the appropriate model order. The corresponding Yule–Walker AR(6) SDF estimate differs by no more than 1 dB from the AR(5) estimate shown as the solid line in Figure 496. The rate at which this AR(6) SDF estimate decreases from 0.8 to 1.0 Hz differs markedly from the lag windows estimates shown in the bottom three plots of Figure 319.

Rather than using the Yule–Walker-based sample PACS, we can instead use estimates $\{\bar{\phi}_{k,k}\}$ associated with Burg's algorithm (Equation (467f)). The bottom plot of Figure 499 shows these estimates, which give an impression different from the standard estimates $\{\tilde{\phi}_{k,k}\}$. Here $\bar{\phi}_{k,k}$ exceeds the $\pm 2/\sqrt{N}$ bounds at lags 1–5, 7, 9–13, 16, 18, 20, 22 and 24–25. The Burg-based PACS criterion thus suggests picking $p = 25$. Minimization of Burg-based FPE(k), AIC(k) and AICC(k) picks orders of 27, 27 and 25. Since the AICC is the most

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

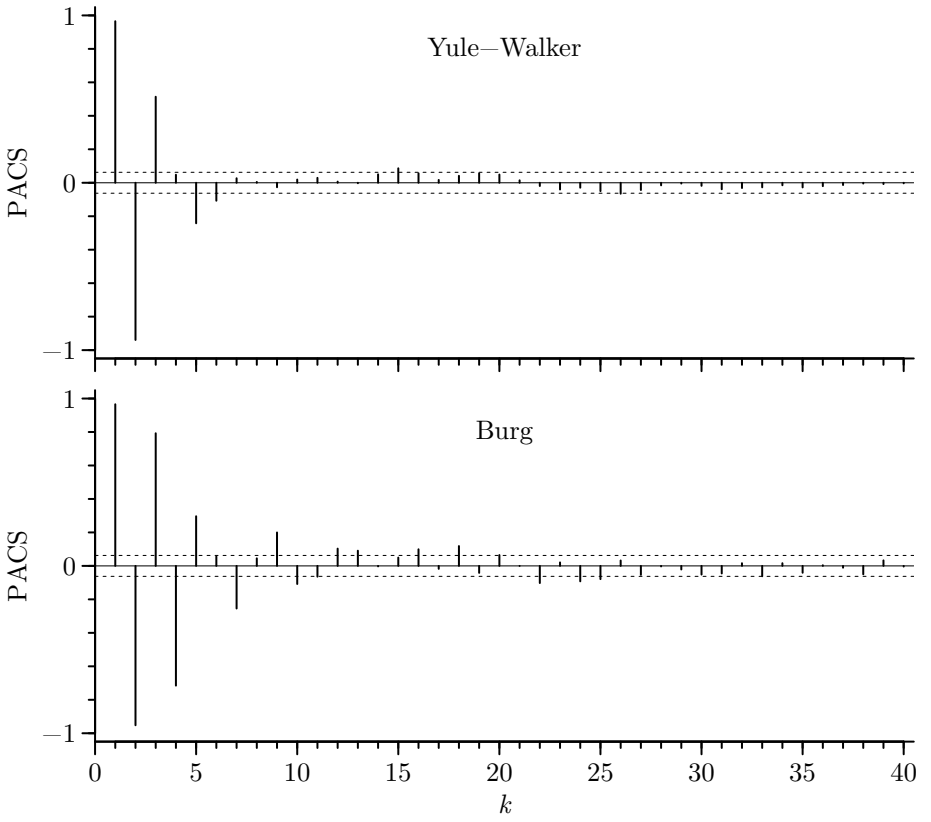

Figure 499 Partial autocorrelation sequences estimates at lags $k = 1, \ldots, 40$ for the ocean wave data. The upper plot shows estimates $\{\tilde{\phi}_{k,k}\}$ based on the Yule–Walker method (i.e., the standard sample PACS); the bottom shows estimates $\{\bar{\phi}_{k,k}\}$ based on Burg's algorithm.

sophisticated of these criteria, we select $p = 25$, in agreement with the PACS criterion. The AR(25) Burg SDF estimate is shown as the thick curves in Figure 500(a) and (b) at all Fourier frequencies satisfying $0 < f_j < f_{\mathcal{N}}$. The two thin curves above and below the thick curve in (a) represent no less than 95% CIs for a hypothesized true AR(25) SDF (see Equation (488b); the plot clips the upper CI, which goes up to 125 dB at $f_{41} \doteq 0.160$ Hz). The thin curve in (b) is the Burg-based AR(5) postcolored Parzen lag window estimate $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ (also shown as the thick curves in Figures 497(c) and (d)). The two estimates agree well for $f \geq 0.4$ Hz, over which they differ by no more than 1.5 dB. The AR(25) estimate portrays more structure over $f < 0.4$ Hz and, in particular, captures the peak centered at $f = 0.16$ Hz nicely, but the postcolored lag window estimate is better at portraying a hypothesized monotonic decay over 0.2 to 1.0 Hz, a primary interest of the investigators.

Atomic Clock Data

Here we revisit the atomic clock fractional frequency deviates $\{Y_t\}$, which are shown in Figure 326(b). In Section 8.9 we considered multitaper-based estimates of the innovation variance both for the deviates and for the first difference of the deviates $\{Y_t - Y_{t-1}\}$, the results of which are summarized in Table 432. Here we compare this nonparametric approach with an AR-based parametric approach. Using both the Yule–Walker method and Burg's algorithm, we fit AR models of orders $k = 1, 2, \ldots, 100$ to the deviates and also to their first differences (fitting was done after centering each series by subtracting off its sample mean). This gives us

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

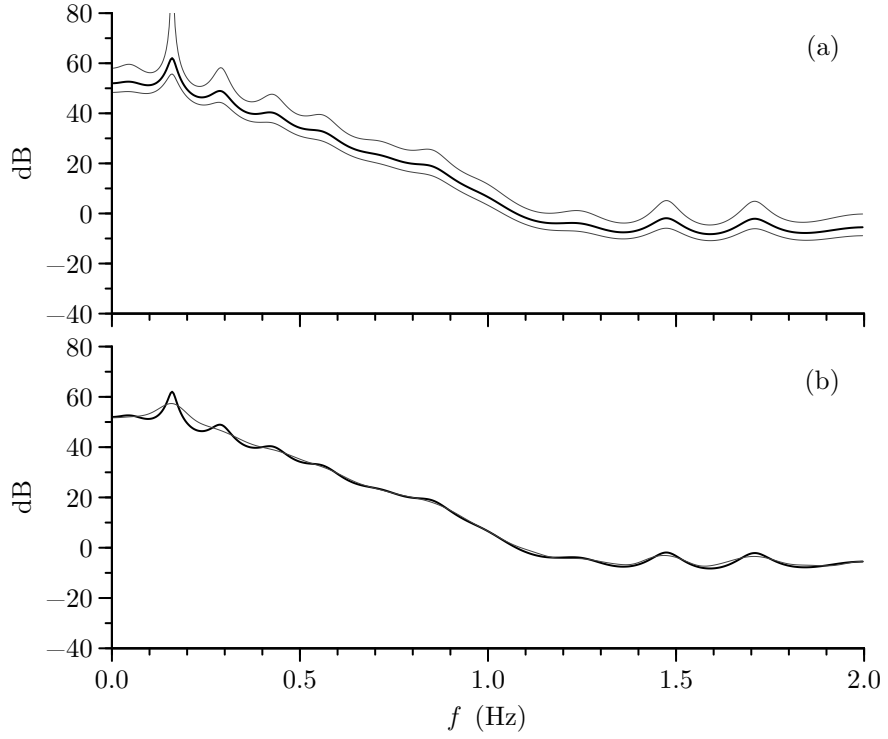

Figure 500 Burg AR(25) SDF estimate for ocean wave data (thick curves in both plots), along with – in upper plot – no less than 95% CIs (two thin curves above and below the thick curve) and – in lower plot – the Burg-based AR(5) postcolored Parzen lag window estimate.

	Yule–Walker				Burg			
Time Series	p	$\tilde{\sigma}_p^2$	sd $\{\tilde{\sigma}_p^2\}$	FPE(p)	p	$\bar{\sigma}_p^2$	sd $\{\bar{\sigma}_p^2\}$	FPE(p)
$Y_t - Y_{t-1}$	41	0.02259	0.00051	0.02307	41	0.02230	0.00050	0.02277
Y_t	42	0.02222	0.00050	0.02270	42	0.02216	0.00050	0.02264

Table 500 AR-based estimates of innovation variance for first difference $Y_t - Y_{t-1}$ of fractional frequency deviates (top row) and for fractional frequency deviates Y_t themselves (bottom).

100 different innovation variance estimates $\tilde{\sigma}_k^2$ for each of the two estimation methods. The multitaper estimates of the innovation variance depend on the parameter K, but not markedly so; on the other hand, the AR estimates depend critically on k, so we look at minimization of FPE(k), AIC(k) and AICC(k) in Equations (493c), (494a) and (495a) for guidance as to its choice. For both the Yule–Walker method and Burg's algorithm, all three criteria pick the rather large model orders of $p = 42$ for $\{Y_t\}$ and of $p = 41$ for $\{Y_t - Y_{t-1}\}$. The top and bottom rows of Table 500 show the Yule–Walker and Burg estimates of σ_p^2 for $\{Y_t\}$ and $\{Y_t - Y_{t-1}\}$, respectively. All four estimate are close to each other and to the ten multitaper-based estimates shown in Table 432 (the ratio of the largest to the smallest of the fourteen estimates is 1.02). Large-sample statistical theory suggests that the variance of of AR-based estimators of the innovation variance is approximately $2\sigma_p^4/N$ (see C&E [1] for Section 9.8). Substituting estimates for σ_p^4 and taking square roots give us an estimated standard deviation

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

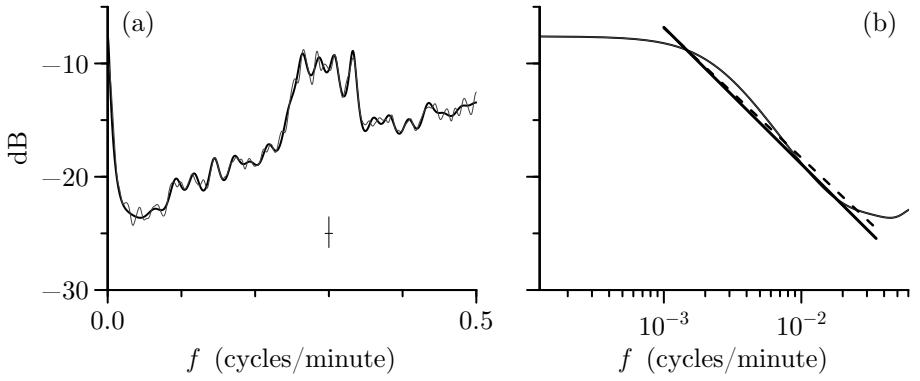

Figure 501 Burg AR(42) SDF estimate for fractional frequency deviates $\{Y_t\}$ (thick curve in both plots). Plot (a) shows the estimate over $f \in (0, f_{\mathcal{N}}]$ on a linear/dB scale while (b) shows the estimate at low frequencies on a log/dB scale. The thin curve in (a) is the same $m = 80$ Gaussian lag window estimate as is displayed in Figure 328(a) (the crisscross is copied from that plot). There are two lines in plot (b). The solid line is copied from Figure 431(b) and is a simple linear regression fit in log/dB space to a $K = 7$ sinusoidal multitaper SDF estimate over Fourier frequencies ranging from $f_4 = 4/(N\,\Delta_t) \doteq 0.010$ cycles/minute to $f_{140} = 140/(N\,\Delta_t) \doteq 0.035$ cycles/minute. The dashed line is a similar fit to the AR SDF estimate.

for the estimates – these are noted in Table 500 (the difference between the largest and smallest of the fourteen estimates is smaller than these standard deviations). Finally we note that the idea behind the FPE is to correct estimates of the innovation variance for a downward bias. This correction is noted in Table 500 (the standard deviations associated with the FPE estimates are only slightly bigger than the ones for the estimates of σ_p^2, increasing by no more than 0.00002).

Figure 501(a) shows the Burg AR(42) SDF estimate (thick curve; the Yule–Walker estimate is virtually the same, differing from the Burg estimate by no more than 0.2 dB). Of the estimates considered previously, the AR estimate most resembles the $m = 80$ Gaussian lag window estimate shown in Figure 328(a) and replicated here as the thin curve (the crisscross is also a replicate from that figure). While the two estimates differ in detail, they both give the same overall impression and have roughly the same bumpiness. Figure 501(b) zooms into the same low-frequency region as was considered in Figure 431(b) for the $K = 7$ sinusoidal multitaper estimate, for which we did a simple linear regression fit in log/dB space over a subset of Fourier frequencies in the low-frequency region – the resulting fitted line was shown in Figure 431(b) and is replicated in Figure 501(b) as the solid line. If we do a similar regression fit to the AR(42) estimate, we obtain the dashed line in Figure 501(b). The two lines are in reasonable agreement. The slopes of these lines can be translated into estimates of the exponent α for an SDF obeying a power law over the subsetted low-frequency region. The multitaper based estimate is $\hat{\alpha}^{(\mathrm{MT})} \doteq -1.21$, while the Burg-based estimate is $\hat{\alpha}^{(\mathrm{AR})} \doteq -1.15$. Given that we estimated $\sqrt{\mathrm{var}\,\{\hat{\alpha}^{(\mathrm{MT})}\}}$ to be 0.11, the two estimates of α are indeed in reasonable agreement (theory is lacking as to how to quantify the variance of $\hat{\alpha}^{(\mathrm{AR})}$).

In conclusion, the AR-based parametric estimates of the innovation variance and of the SDF for the atomic clock data agree well with previously considered nonparametric estimates. Also, in contrast to the ocean wave example, here the Yule–Walker and Burg SDF estimates give comparable results.

9.13 Comments on Complex-Valued Time Series

We have assumed throughout this chapter that our time series of interest is real-valued. If we wish to fit an AR(p) model to a complex-valued series that is a portion of a realization of a

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

complex-valued stationary process $\{Z_t\}$ with zero mean, we need to make some modifications in the definitions of our various AR parameter estimators (the form of the AR(p) model itself is the same as indicated by Equation (446a), except that $\phi_{p,j}$ is now in general complex-valued). Most of the changes are due to the following fundamental difference between real-valued and complex-valued processes. For a real-valued stationary process $\{X_t\}$ with ACVS $\{s_{X,\tau}\}$, the time reversed process $\{X_{-t}\}$ also has ACVS $\{s_{X,\tau}\}$, whereas, for a complex-valued process $\{Z_t\}$ with ACVS $\{s_{Z,\tau}\}$, the *complex conjugate* time reversed process $\{Z^*_{-t}\}$ has ACVS $\{s_{Z,\tau}\}$ (the proof is an easy exercise). One immediate implication is that, if the best linear predictor of Z_t, given $Z_{t-1}, \ldots, Z_{t-k}$, is

$$\overrightarrow{Z}_t(k) = \sum_{j=1}^{k} \phi_{k,j} Z_{t-j},$$

then the best linear "predictor" of Z_t, given $Z_{t+1}, \ldots, Z_{t+k}$, has the form

$$\overleftarrow{Z}_t(k) = \sum_{j=1}^{k} \phi^*_{k,j} Z_{t+j}. \tag{502}$$

Given these facts, a rederivation of the Levinson–Durbin recursions shows that we must make the following minor changes to obtain the proper Yule–Walker and Burg parameter estimators for complex-valued time series. For the Yule–Walker estimator, Equations (458b) and (458c) now become

$$\begin{aligned} \tilde{\phi}_{k,j} &= \tilde{\phi}_{k-1,j} - \tilde{\phi}_{k,k} \tilde{\phi}^*_{k-1,k-j}, \quad 1 \le j \le k-1 \\ \tilde{\sigma}^2_k &= \tilde{\sigma}^2_{k-1}(1 - |\tilde{\phi}_{k,k}|^2) \end{aligned}$$

(Equation (458a) remains the same). As before, we initialize the recursions using $\tilde{\phi}_{1,1} = \hat{s}^{(\mathrm{P})}_1 / \hat{s}^{(\mathrm{P})}_0$, but now we set $\tilde{\sigma}^2_1 = \hat{s}^{(\mathrm{P})}_0 (1 - |\tilde{\phi}_{1,1}|^2)$. In lieu of Equation (467f), the Burg estimator of the kth-order partial autocorrelation coefficient now takes the form

$$\bar{\phi}_{k,k} = \frac{2 \sum_{t=k}^{N-1} \overrightarrow{e}_t(k-1) \overleftarrow{e}^*_{t-k}(k-1)}{\sum_{t=k}^{N-1} |\overrightarrow{e}_t(k-1)|^2 + |\overleftarrow{e}_{t-k}(k-1)|^2}.$$

Equation (467a) still gives the recursive formula for the forward prediction errors, but the formula for the backward prediction errors is now given by

$$\overleftarrow{e}_{t-k}(k) = \overleftarrow{e}_{t-k}(k-1) - \bar{\phi}^*_{k,k} \overrightarrow{e}_t(k-1), \quad k \le t \le N-1,$$

instead of by Equation (467b).

Finally we note that

[1] the various least squares estimators are now obtained by minimizing an appropriate sum of squared moduli of prediction errors rather than sum of squares; and

[2] once we have obtained the AR parameter estimates, the parametric spectral estimate is obtained in the same manner as in the real-valued case, namely, by substituting the estimated parameters into Equation (446b).

For further details on parametric spectral estimation using complex-valued time series, see Kay (1988).

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

9.14 Use of Other Models for Parametric SDF Estimation

So far we have concentrated entirely on the use of AR(p) models for parametric SDF estimation. There are, however, other classes of models that could be used. We discuss briefly one such class here – namely, the class of moving average processes of order q (MA(q)) – because it provides an interesting contrast to AR(p) processes.

An MA(q) process $\{X_t\}$ with zero mean is defined by

$$X_t = \epsilon_t - \theta_{q,1}\epsilon_{t-1} - \theta_{q,2}\epsilon_{t-2} - \cdots - \theta_{q,q}\epsilon_{t-q},$$

where $\theta_{q,1}, \theta_{q,2}, \ldots, \theta_{q,q}$ are q fixed coefficients, and $\{\epsilon_t\}$ is a white noise process with zero mean and variance σ_ϵ^2 (see Equation (32a)). The above is stationary no matter what values the coefficients assume. Its SDF is given by

$$S(f) = \sigma_\epsilon^2 \, \Delta_t \left| 1 - \sum_{j=1}^{q} \theta_{q,j} \mathrm{e}^{-\mathrm{i}2\pi f j \, \Delta_t} \right|^2 .$$

There are $q+1$ parameters (the q coefficients $\theta_{q,j}$ and the variance σ_ϵ^2) that we need to estimate to produce an MA(q) parametric SDF estimate.

It is interesting to reconsider the six-fold rationale for using AR(p) models discussed in Section 9.2 and to contrast it with the MA(q) case.

[1] As in the case of AR(p) processes, any continuous SDF $S(\cdot)$ can be approximated arbitrarily well by an MA(q) SDF if q is chosen large enough (Anderson, 1971, p. 411). In this sense the class of MA processes is as rich as that of AR processes, but in another sense it is *not* as rich (see the discussion surrounding Equation (504a)).

[2] Algorithms for explicitly fitting MA(q) models to time series tend to be computationally intensive compared with those needed to fit AR(p) models. However, if our interest is primarily in spectral estimation so that we do not require explicit estimates of $\theta_{q,j}$, there is an analog to the Yule–Walker approach that is computationally efficient but unfortunately has poor statistical properties (see the discussion surrounding Equation (504b)).

[3] From the discussion following Equation (473), we know that, for a Gaussian process with known variance s_0 and known cepstral values $c_1, \ldots, c_q$, the maximum entropy spectrum is identical to that of an MA(q) process (there is always a Gaussian process satisfying these $q + 1$ constraints).

[4] Fitting an MA(q) model to a time series leads to an IIR filter, which – because of the problem of turn-on transients – is not as easy to work with as a FIR filter for prewhitening a time series of finite length.

[5] Physical arguments sometimes suggest that an MA model is appropriate for a particular time series. For example, we have already noted in Section 2.6 that Spencer-Smith and Todd (1941) argued thus for the thickness of textile slivers as a function of displacement along a sliver.

[6] Pure sinusoidal variations cannot be expressed as an MA(q) model.

Figure 33 shows eight examples of realizations of MA(1) processes.

As in the case of AR(p) processes, we need to estimate each $\theta_{q,j}$ and the variance σ_ϵ^2 from available data. If we attempt to use the method of matching lagged moments that led to the Yule–Walker equations in the AR(p) case, we find that

$$s_\tau = \sigma_\epsilon^2 \sum_{j=0}^{q-\tau} \theta_{q,j+\tau}\theta_{q,j}, \quad \tau = 0, 1, \ldots, q,$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

where we define $\theta_{q,0} = -1$ (cf. Equation (32b)). These are nonlinear in the MA coefficients, but an additional difficulty is that a valid solution need not exist. As an example, for an MA(1) process, we have

$$s_0 = \sigma_\epsilon^2 \left(1 + \theta_{1,1}^2\right) \quad \text{and} \quad s_1 = -\sigma_\epsilon^2 \theta_{1,1}, \tag{504a}$$

which can only be solved for a valid σ_ϵ^2 and $\theta_{1,1}$ if $|s_1/s_0| \leq 1/2$. Thus, whereas there is always an AR(p) process whose theoretical ACVS agrees out to lag p with $\hat{s}_0^{(\mathrm{P})}, \ldots, \hat{s}_p^{(\mathrm{P})}$ computed for any time series, the same is not true of MA processes. In this sense we can claim that the class of AR processes is "richer" than the class of MA processes.

There is, however, a second way of looking at the Yule–Walker estimation procedure for AR(p) processes as far as SDF estimation is concerned. If we fit an AR(p) process to a time series by that procedure, the ACVS of the fitted model necessarily agrees *exactly* up to lag p with the estimated ACVS. The values of the ACVS after lag p are generated recursively in accordance with an AR(p) model, namely,

$$s_\tau = \sum_{j=1}^{p} \phi_{p,j} s_{\tau - j}, \quad \tau = p+1, p+2, \ldots$$

(see Equation (450a)). The estimated SDF can thus be regarded as given by

$$\hat{S}^{(\mathrm{YW})}(f) = \Delta_\mathrm{t} \sum_{\tau=-\infty}^{\infty} \tilde{s}_\tau \mathrm{e}^{-\mathrm{i}2\pi f \tau \Delta_\mathrm{t}}, \quad \text{where} \quad \tilde{s}_\tau = \begin{cases} \hat{s}_\tau^{(\mathrm{P})}, & |\tau| \leq p; \\ \sum_{j=1}^{p} \tilde{\phi}_{p,j} \tilde{s}_{|\tau|-j}, & |\tau| > p, \end{cases}$$

with $\tilde{\phi}_{p,j}$ defined to be the Yule–Walker estimator of $\phi_{p,j}$. Thus the Yule–Walker SDF estimator can be regarded as accepting $\hat{s}_\tau^{(\mathrm{P})}$ up to lag p and then estimating the ACVS beyond that lag based upon the AR(p) model. If we follow this procedure for an MA(q) model, we would set the ACVS beyond lag q equal to 0 (see Equation (32b)). Our SDF estimate would thus be just

$$\hat{S}(f) = \Delta_\mathrm{t} \sum_{\tau=-q}^{q} \hat{s}_\tau^{(\mathrm{P})} \mathrm{e}^{-\mathrm{i}2\pi f \tau \Delta_\mathrm{t}}. \tag{504b}$$

This is identical to the truncated periodogram of Equation (343b). Although this estimator was an early example of a lag window estimator, its smoothing window is inferior to that for other estimators of the lag window class (such as those utilizing the Parzen or Papoulis lag windows). In particular, the smoothing window is such that $\hat{S}(f)$ need *not* be nonnegative at all frequencies, thus potentially yielding an estimate that violates a fundamental property of an actual SDF (see Exercise [7.7]).

To conclude, the choice of the class of models to be used in parametric spectral analysis is obviously quite important. The class of AR models has a number of advantages when compared to MA models, not the least of which is ease of computation. Unfortunately, algorithms for explicitly fitting ARMA models are much more involved than those for fitting AR models; on the other hand, we have noted that the spectral estimator of Equation (491b) is in fact an ARMA SDF, so it is certainly possible to obtain *implicit* ARMA spectral estimates in a computationally efficient manner.

9.15 Summary of Parametric Spectral Estimators

We assume that $X_0, X_1, \ldots, X_{N-1}$ is a sample of length N from a real-valued stationary process $\{X_t\}$ with zero mean and unknown SDF $S_X(\cdot)$ defined over the interval $[-f_{\mathcal{N}}, f_{\mathcal{N}}]$, where $f_{\mathcal{N}} \stackrel{\text{def}}{=} 1/(2\,\Delta_t)$ is the Nyquist frequency and Δ_t is the sampling interval between observations. (If $\{X_t\}$ has an unknown mean, then we need to replace X_t with $X'_t = X_t - \overline{X}$ in all computational formulae, where $\overline{X} \stackrel{\text{def}}{=} \sum_{t=0}^{N-1} X_t/N$ is the sample mean.) The first step in parametric spectral analysis is to estimate the $p+1$ parameters – namely, $\phi_{p,1}, \ldots, \phi_{p,p}$ and σ_p^2 – of an AR(p) model (see Equation (446a)) using one of the following estimators (Section 9.11 discusses choosing the order p).

[1] *Yule–Walker estimators $\tilde{\phi}_{p,1}, \ldots, \tilde{\phi}_{p,p}$ and $\tilde{\sigma}_p^2$*
We first form the usual biased estimator $\{\hat{s}_\tau^{(\mathrm{P})}\}$ of the ACVS using $X_0, \ldots, X_{N-1}$ (see Equation (170b)). With $\tilde{\sigma}_0^2 \stackrel{\text{def}}{=} \hat{s}_0^{(\mathrm{P})}$, we obtain the Yule–Walker estimators by recursively computing, for $k = 1, \ldots, p$,

$$\tilde{\phi}_{k,k} = \frac{\hat{s}_k^{(\mathrm{P})} - \sum_{j=1}^{k-1} \tilde{\phi}_{k-1,j} \hat{s}_{k-j}^{(\mathrm{P})}}{\tilde{\sigma}_{k-1}^2}; \qquad \text{(see (458a))}$$

$$\tilde{\phi}_{k,j} = \tilde{\phi}_{k-1,j} - \tilde{\phi}_{k,k} \tilde{\phi}_{k-1,k-j}, \quad 1 \le j \le k-1; \qquad \text{(see (458b))}$$

$$\tilde{\sigma}_k^2 = \tilde{\sigma}_{k-1}^2 (1 - \tilde{\phi}_{k,k}^2) \qquad \text{(see (458c))}$$

(for $k = 1$, we obtain $\tilde{\phi}_{1,1} = \hat{s}_1^{(\mathrm{P})} \big/ \hat{s}_0^{(\mathrm{P})}$ and $\tilde{\sigma}_1^2 = \hat{s}_0^{(\mathrm{P})}(1 - \tilde{\phi}_{1,1}^2)$). An equivalent formulation that makes direct use of $X_0, \ldots, X_{N-1}$ rather than $\{\hat{s}_k^{(\mathrm{P})}\}$ is outlined in the discussion surrounding Equation (471). A variation on the Yule–Walker method is to use the ACVS estimator corresponding to a direct spectral estimator $\hat{S}^{(\mathrm{D})}(\cdot)$ other than the periodogram (see the discussion concerning Figures 461 and 462). The Yule–Walker parameter estimators are guaranteed to correspond to a causal – and hence stationary – AR process.

[2] *Burg estimators $\bar{\phi}_{p,1}, \ldots, \bar{\phi}_{p,p}$ and $\bar{\sigma}_p^2$*
With $\overrightarrow{e}_t(0) \stackrel{\text{def}}{=} X_t$, $\overleftarrow{e}_t(0) \stackrel{\text{def}}{=} X_t$ and $\bar{\sigma}_0^2 \stackrel{\text{def}}{=} \hat{s}_0^{(\mathrm{P})}$, we obtain the Burg estimators by recursively computing, for $k = 1, \ldots, p$,

$$A_k = \sum_{t=k}^{N-1} \overrightarrow{e}_t^2(k-1) + \overleftarrow{e}_{t-k}^2(k-1) \qquad \text{(see (467d))}$$

$$B_k = 2 \sum_{t=k}^{N-1} \overrightarrow{e}_t(k-1) \overleftarrow{e}_{t-k}(k-1) \qquad \text{(see (467e))}$$

$$\bar{\phi}_{k,k} = B_k \big/ A_k \qquad \text{(see (467f))}$$

$$\overrightarrow{e}_t(k) = \overrightarrow{e}_t(k-1) - \bar{\phi}_{k,k} \overleftarrow{e}_{t-k}(k-1), \quad k \le t \le N-1; \qquad \text{(see (467a))}$$

$$\overleftarrow{e}_{t-k}(k) = \overleftarrow{e}_{t-k}(k-1) - \bar{\phi}_{k,k} \overrightarrow{e}_t(k-1), \quad k \le t \le N-1; \qquad \text{(see (467b))}$$

$$\bar{\phi}_{k,j} = \bar{\phi}_{k-1,j} - \bar{\phi}_{k,k} \bar{\phi}_{k-1,k-j}, \quad 1 \le j \le k-1;$$

$$\bar{\sigma}_k^2 = \bar{\sigma}_{k-1}^2 (1 - \bar{\phi}_{k,k}^2). \qquad \text{(see (468b))}$$

The term A_k can be recursively computed using Equation (509b). The Burg parameter estimators in practice correspond to a causal – and hence stationary – AR process and have been found to be superior to the Yule–Walker estimators in numerous Monte Carlo experiments.

[3] *Least squares (LS) estimators*
Here we use the resemblance between an AR(p) model and a regression model to formulate LS parameter estimators. Three different formulations are detailed in Section 9.7, namely, the forward LS, backward LS and forward/backward LS estimators (FLS, BLS and FBLS, respectively). Calculation of these estimators can be done using standard least squares techniques (Golub and Van

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

Loan, 2013; Press et al., 2007) or specialized algorithms that exploit the structure of the AR model (Marple, 1987; Kay, 1988). None of the three LS parameter estimators is guaranteed to correspond to a causal AR process. In terms of spectrum estimation, Monte Carlo studies show that the FBLS estimator is generally superior to the Yule–Walker estimator and outperforms the Burg estimator in some cases.

[4] *Maximum likelihood estimators (MLEs)*
Under the assumption of a Gaussian process, we derived MLEs for the AR parameters in Section 9.8. Unfortunately, these estimators require the use of a nonlinear optimization routine and hence are computationally intensive compared to Yule–Walker, Burg and LS estimators (particularly for large model orders p). These computationally more efficient estimators can all be regarded as approximate MLEs.

Let $\hat{\phi}_{p,1}, \ldots, \hat{\phi}_{p,p}$ and $\hat{\sigma}_p^2$ represent any of the estimators just mentioned (Yule–Walker, Burg, LS or MLEs). There are two ways we can use these estimators to produce an estimator of the SDF $S_X(\cdot)$.

[1] *Pure parametric approach*
Here we merely substitute the AR parameter estimators for the corresponding theoretical parameters in Equation (446b) to obtain the SDF estimator

$$\hat{S}_X(f) = \frac{\hat{\sigma}_p^2 \, \Delta_t}{\left| 1 - \sum_{j=1}^{p} \hat{\phi}_{p,j} e^{-i2\pi f j \, \Delta_t} \right|^2},$$

which is a periodic function of f with a period of $2f_{\mathcal{N}}$. Section 9.9 gives an approach for obtaining confidence intervals for the normalized SDF based upon this estimator.

[2] *Prewhitening approach*
Here we use the estimated coefficients $\hat{\phi}_{p,1}, \ldots, \hat{\phi}_{p,p}$ to create a prewhitening filter. Section 9.10 discusses this combined parametric/nonparametric approach in detail – its features are generally more appealing than the pure parametric approach.

9.16 Exercises

[9.1] (a) Using Equation (145b), verify the two stationary solutions stated in Equation (447) for AR(1) processes.

(b) Show that, if $\widetilde{Y}_t = \sum_{j=1}^{p} \varphi_{p,j} \widetilde{Y}_{t-j} + \varepsilon_t$ is a stationary acausal AR(p) process with SDF

$$S_{\widetilde{Y}}(f) = \frac{\sigma_\varepsilon^2}{\left| 1 - \sum_{j=1}^{p} \varphi_{p,j} e^{-i2\pi f j} \right|^2},$$

there exists a causal AR(p) process, say, $Y_t = \sum_{j=1}^{p} \phi_{p,j} Y_{t-j} + \epsilon_t$, whose SDF $S_Y(\cdot)$ is given by the above upon replacing $\varphi_{p,j}$ with $\phi_{p,j}$, and $S_Y(\cdot)$ is *identically* the same as $S_{\widetilde{Y}}(\cdot)$. Here $\{\varepsilon_t\}$ and $\{\epsilon_t\}$ are both zero-mean white noise processes with finite variances given by, respectively, $\sigma_\varepsilon^2 > 0$ and $\sigma_\epsilon^2 > 0$. Hint: for a complex-valued variable z, consider the polynomial $\widetilde{\mathcal{G}}(z) = 1 - \sum_{j=1}^{p} \varphi_{p,j} z^{-j}$, rewrite it as

$$\widetilde{\mathcal{G}}(z) = \prod_{j=1}^{p} \left(1 - \frac{r_j}{z} \right),$$

where $r_1, \ldots, r_p$ are the p roots of the polynomial (at least one of which is such that $|r_j| > 1$), and then consider $z = e^{i2\pi f}$ and $|\widetilde{\mathcal{G}}(z)|^2$.

(c) For a real-valued ϕ such that $0 < |\phi| < 1$, consider the AR(2) process

$$\widetilde{Y}_t = \left(\phi + \frac{1}{\phi} \right) \widetilde{Y}_{t-1} - \widetilde{Y}_{t-2} + \varepsilon_t,$$

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

where $\{\varepsilon_t\}$ is a zero-mean white noise process with variance $\sigma_\varepsilon^2 > 0$. By expressing $\widetilde{Y}_t$ as a linear combination of RVs in $\{\varepsilon_t\}$, show that $\{\widetilde{Y}_t\}$ has a stationary – but acausal – solution. Use this expression to determine var $\{\widetilde{Y}_t\}$.

(d) Determine the causal AR(2) process $Y_t = \phi_{2,1} Y_{t-1} + \phi_{2,2} Y_{t-2} + \epsilon_t$ that has the same SDF as $\{\widetilde{Y}_t\}$ of part (c) (here $\{\epsilon_t\}$ is a zero-mean white noise process with variance $\sigma_\epsilon^2 > 0$). Express the stationary solution for Y_t as a linear combination of RVs in $\{\epsilon_t\}$. Use this expression to determine var $\{Y_t\}$. Show that var $\{Y_t\}$ is related to var $\{\widetilde{Y}_t\}$ of part (c) in a manner consistent with the result of part (b).

(e) Use the step-down Levinson–Durbin procedure stated in Equations (460a) and (460b) to verify the expression for var $\{Y_t\}$ obtained in part (d). Is it possible to do the same for var $\{\widetilde{Y}_t\}$ obtained in part (c)?

[9.2] In what follows, take $Y_t = \phi Y_{t-1} + \epsilon_t$ to be a causal (and hence stationary) AR(1) process with mean zero, where $\{\epsilon_t\}$ is a white noise process with mean zero and variance σ_1^2. Let $S(\cdot)$ and Δ_t denote the SDF and sampling interval for $\{Y_t\}$.

(a) Using equations developed in Section 9.3, show that the ACVS for $\{Y_t\}$ is

$$s_\tau = \frac{\phi^{|\tau|}\sigma_1^2}{1-\phi^2}, \quad \tau \in \mathbb{Z} \tag{507a}$$

(cf. Exercise [2.17a], which tackles the above using a different approach).

(b) Use Equation (165a) to show that the variance of the sample mean $\overline{Y}_N$ of $Y_0, Y_1, \ldots, Y_{N-1}$ is

$$\text{var}\left\{\overline{Y}_N\right\} = \frac{s_0}{N}\left(\frac{1+\phi}{1-\phi} - \frac{2\phi(1-\phi^N)}{N(1-\phi)^2}\right). \tag{507b}$$

Equation (165c) suggests the approximation $S(0)/(N\,\Delta_t)$. Show this yields

$$\text{var}\left\{\overline{Y}_N\right\} \approx \frac{s_0}{N}\left(\frac{1+\phi}{1-\phi}\right). \tag{507c}$$

(c) For $N = 100$, plot the log of the ratio var $\{\overline{Y}_N\}/s_0$ versus $\phi = -0.99, -0.98, \ldots, 0.98, 0.99$, and then superimpose a similar plot, but with the exact variance replaced by the approximation of Equation (507c). Plot the ratio of the approximate variance of Equation (507c) to the exact variance of Equation (507b) versus ϕ. Comment on how well the approximation does.

(d) For $\phi = -0.9$, $\phi = 0$ and finally $\phi = 0.9$, plot $\log_{10}\left(\text{var}\{\overline{Y}_N\}/s_0\right)$ versus $\log_{10}(N)$ for $N = 1, 2, \ldots, 1000$. Comment briefly on the appearance of the three plots (in particular, how they differ and how they are similar).

(e) Equation (298c) expresses the degrees of freedom η in a time series of length N in terms of its ACVS. Derive an expression for η when the series is a Gaussian AR(1) process. For $N = 100$, plot η versus versus $\phi = -0.99, -0.98, \ldots, 0.98, 0.99$. Comment upon your findings.

(f) Suppose that $\{Y(t) : t \in \mathbb{R}\}$ is a continuous parameter stationary process with a Lorenzian SDF and ACVS given by

$$S_{Y(t)}(f) = \frac{2L\sigma^2}{1+(2\pi f L)^2} \quad \text{and} \quad s_{Y(t)}(\tau) = \sigma^2 \mathrm{e}^{-|\tau|/L}, \quad \text{for } f, \tau \in \mathbb{R},$$

where $\sigma^2 > 0$ and $L > 0$ are finite and represent the variance of $\{Y(t)\}$ and its correlation length (see Equations (131b) and (131c)). Show that the discrete parameter stationary process defined by $Y_t = Y(t\,\Delta_t)$ is an AR(1) process, i.e., has an SDF and ACVS in agreement with that of $\{Y_t : t \in \mathbb{Z}\}$ used in previous parts of this exercise. How does the AR(1) coefficient ϕ depend on σ^2, L and Δ_t, and what range of values can ϕ assume?

[9.3] Obtain Yule–Walker estimates of the parameters $\phi_{2,1}$, $\phi_{2,2}$ and σ_2^2 for an AR(2) process using the following biased estimates of the ACVS: $\hat{s}_0^{(\mathrm{P})} = 1$, $\hat{s}_1^{(\mathrm{P})} = -1/2$ and $\hat{s}_2^{(\mathrm{P})} = -1/5$. Assuming

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

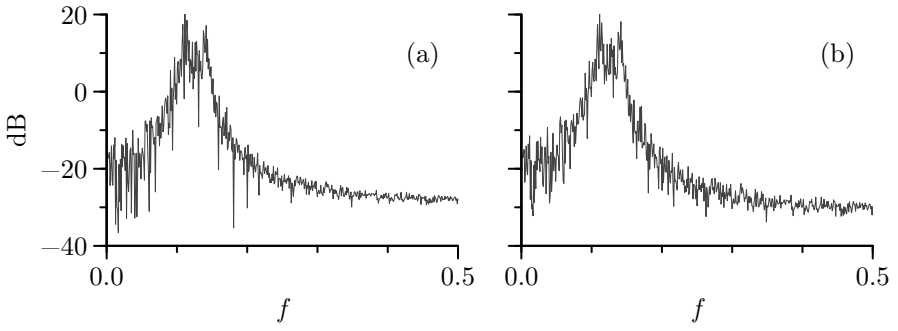

Figure 508 Periodogram (left-hand plot) and Yule–Walker SDF estimate of order $p = 1023$ (right-hand) for AR(4) time series shown in Figure 34(e) – see Exercise [9.4] for details.

$\Delta_t = 1$, state the corresponding estimated SDF in a reduced form (i.e, cosines rather than complex exponentials), and compute its value at frequencies $f = 0$, $1/4$ and $1/2$. Plot the SDF on a decibel scale versus frequency over a fine grid of frequencies, e.g., $f_k = k/256$ for $k = 0, 1, \ldots, 128$. Finally, determine the corresponding estimated ACVS at lags $\tau = 0, 1, \ldots 5$.

[9.4] Figure 508(a) shows the periodogram $\hat{S}^{(\mathrm{P})}(f_k)$ versus Fourier frequencies $f_k = k/N$, $k = 0,1, \ldots, N/2$, for the AR(4) time series of length $N = 1024$ displayed in Figure 34(e). Figure 508(b) shows a Yule–Walker SDF estimate $\hat{S}^{(\mathrm{YW})}(\cdot)$ of order $p = 1023$ over the same grid of frequencies; i.e., we have set $p = N - 1$, which is the maximum lag for which we can compute an estimate of the ACVS from $X_0, \ldots, X_{N-1}$. As noted in Section 9.3, an important property of a Yule–Walker estimator is that its corresponding ACVS is identical to the sample ACVS $\{\hat{s}^{(\mathrm{P})}\}$ up to lag p. Since the periodogram is the Fourier transform of $\{\hat{s}^{(\mathrm{P})}\}$, for this example we thus have, for $\tau = 0, \ldots, 1023$,

$$\int_{-1/2}^{1/2} \hat{S}^{(\mathrm{YW})}(f) \mathrm{e}^{\mathrm{i}2\pi f\tau} \, \mathrm{d}f = \hat{s}_\tau^{(\mathrm{P})} \quad \text{and} \quad \int_{-1/2}^{1/2} \hat{S}^{(\mathrm{P})}(f) \mathrm{e}^{\mathrm{i}2\pi f\tau} \, \mathrm{d}f = \hat{s}_\tau^{(\mathrm{P})}$$

(cf. Equation (171c)). Figure 508 indicates that the periodogram and Yule-Walker estimate resemble one another closely. Explain why they aren't identical.

[9.5] Here we prove a useful corollary to the orthogonality principle. Let $\{X_t\}$ be any stationary process with zero mean and with an ACVS $\{s_k\}$ that is positive definite. Define

$$d_t(k) = X_t - \sum_{l=1}^{k} \psi_l X_{t-l},$$

where the ψ_l terms are any set of constants. Suppose that $E\{d_t(k)X_{t-j}\} = 0$, $1 \le j \le k$; i.e., the $d_t(k)$ terms mimic the property stated in Equation (454a) for $\overrightarrow{\epsilon}_t(k)$. Show that we must have $d_t(k) = \overrightarrow{\epsilon}_t(k)$. Hint: note that the system of equations here is identical to that in Equation (454b).

[9.6] Equation (460d) gives a procedure for determining the ACVS for an AR(p) process given its coefficients $\phi_{p,1}, \ldots, \phi_{p,p}$ and innovation variance σ_p^2. Use this procedure to show that ACVS for the AR(2) process of Equation (34) is

$$s_0 = \tfrac{16}{9}, \quad s_1 = \tfrac{8}{9} \quad \text{and} \quad s_\tau = \tfrac{3}{4}s_{\tau-1} - \tfrac{1}{2}s_{\tau-2}, \quad \tau = 2, 3, \ldots \tag{508a}$$

and that the ACVS for the AR(4) process of Equation (35a) is

$$s_0 \doteq 1.523\,434\,580, \quad s_1 \doteq 1.091\,506\,153, \quad s_2 \doteq 0.054\,284\,646, \quad s_3 \doteq -0.975\,329\,449 \tag{508b}$$

and $s_\tau = 2.7607s_{\tau-1} - 3.8106s_{\tau-2} + 2.6535s_{\tau-3} - 0.9238s_{\tau-4}$, $\tau = 4, 5, \ldots$. Use the above to check that the top row of plots in Figure 168 is correct.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

[9.7] Verify Equation (462b), which states that $\phi_{k,k}$ can be interpreted as a correlation coefficient. Hint: express $\overleftarrow{\epsilon}_{t-k}(k-1)$ in terms of $\{X_t\}$; evoke the orthogonality principle (Equation (454a)); and use Equation (456d).

[9.8] (a) Use Equations (457a) and (457b) to show that $\sigma_0^2 = s_0$; i.e., the zeroth-order mean square linear prediction error is just the process variance. Is this a reasonable interpretation for σ_0^2?

(b) Show how the Levinson–Durbin recursions can be initialized starting with $k = 0$ rather than $k = 1$ as is done in Equation (457a).

(c) Use part (a) and Equation (457b) to show that

$$\sigma_k^2 = s_0 \prod_{j=1}^{k} (1 - \phi_{j,j}^2). \tag{509a}$$

[9.9] Consider the following six AR(4) processes:

(a) $Y_t = 0.8296Y_{t-1} + 0.8742Y_{t-2} - 0.4277Y_{t-3} + 0.6609Y_{t-4} + \epsilon_t$;
(b) $Y_t = 0.2835Y_{t-1} + 0.0382Y_{t-2} + 0.4732Y_{t-3} - 0.7307Y_{t-4} + \epsilon_t$;
(c) $Y_t = 0.3140Y_{t-1} + 0.4101Y_{t-2} - 0.0845Y_{t-3} + 0.4382Y_{t-4} + \epsilon_t$;
(d) $Y_t = 0.8693Y_{t-1} - 0.4891Y_{t-2} - 0.0754Y_{t-3} + 0.8800Y_{t-4} + \epsilon_t$;
(e) $Y_t = 0.9565Y_{t-1} - 0.7650Y_{t-2} - 0.0500Y_{t-3} + 0.1207Y_{t-4} + \epsilon_t$;
(f) $Y_t = 0.8081Y_{t-1} - 0.7226Y_{t-2} + 0.9778Y_{t-3} + 0.8933Y_{t-4} + \epsilon_t$,

where, as usual, $\{\epsilon_t\}$ is a white noise process with zero mean and finite nonzero variance. Use the step-down Levinson–Durbin recursions of Equation (460a) to determine which (if any) of these processes are causal (and hence stationary).

[9.10] Suppose that we are given the following ACVS values: $s_0 = 3$, $s_1 = 2$ and $s_2 = 1$. Use the Levinson–Durbin recursions to show that $\phi_{1,1} = 2/3$, $\phi_{2,1} = 4/5$, $\phi_{2,2} = -1/5$, $\sigma_1^2 = 5/3$ and $\sigma_2^2 = 8/5$. Construct $\boldsymbol{L}_3^{-1}$ and $\boldsymbol{D}_3$, and verify that $\boldsymbol{L}_3^{-1}\boldsymbol{\Gamma}_3\boldsymbol{L}_3^{-T} = \boldsymbol{D}_3$ (see Equation (464d)). Determine $\boldsymbol{L}_3$ also, verifying that it is lower triangular. (This example is due to Therrien, 1983.)

[9.11] Any one of the following five sets of $p + 1$ quantities completely characterizes the second-order properties of the AR(p) process of Equation (446a):

(i) $\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}$ and σ_p^2; (ii) $\phi_{p,1}, \phi_{p,2}, \ldots, \phi_{p,p}$ and s_0;
(iii) $\phi_{1,1}, \phi_{2,2}, \ldots, \phi_{p,p}$ and s_0; (iv) $\phi_{1,1}, \phi_{2,2}, \ldots, \phi_{p,p}$ and σ_p^2; (v) $s_0, s_1, \ldots, s_p$.

Show these are equivalent by demonstrating how, given any one of them, we can get the other four.

[9.12] (a) Figure 458 shows Yule–Walker AR(4) SDF estimates based on the realization of the AR(4) process shown in Figure 34(e). The four estimates shown in the figure make use of the first 16 values of the time series, the first 64 values, the first 256 values and, finally, the entire series of 1024 values. Figure 469 shows corresponding Burg AR(4) SDF estimates. Create similar figures of Yule–Walker and Burg estimates for the time series shown in Figure 34(f), (g) and (h), which are downloadable from the website for this book.

(b) Repeat part (a), but now using the four AR(2) time series shown in Figures 34(a) to (d).

[9.13] (a) Show that Equation (470a) holds (hint: use Equation (467c)).

(b) Show that A_k of Equation (467d) obeys the recursive relationship

$$A_k = \left(1 - \bar{\phi}_{k-1,k-1}^2\right) A_{k-1} - \overrightarrow{e}_{k-1}^2(k-1) - \overleftarrow{e}_{N-k}^2(k-1), \tag{509b}$$

which is helpful in computing $\bar{\phi}_{k,k}$ in Equation (467f) (Andersen, 1978).

[9.14] (a) Figure 478 shows FBLS SDF estimates based on the realization of the AR(4) process shown in Figure 34(e). The four estimates shown in the figure make use of the first 16 values of the time series, the first 64 values, the first 256 values and, finally, the entire series of 1024 values. Create similar figures for the FLS estimates and the BLS estimates.

(b) Repeat part (a), but now using the three AR(4) time series shown in Figures 34(f), (g) and (h) and also including the FBLS estimator.

[9.15] (a) Use Burg's algorithm to fit an AR(2) model to the six-point time series $\{100, 10, 1, -1, -10, -100\}$. Report the Burg estimates $\bar{\phi}_{2,1}$, $\bar{\phi}_{2,2}$ and $\bar{\sigma}_2^2$. Assuming $\Delta_t = 1$, state the corresponding estimated SDF in a reduced form (i.e, cosines rather than complex exponentials), and plot the SDF on a decibel scale over the frequencies $f'_k = k/256$, $k = 0, 1, \ldots, 128$.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press

(b) In the AR(2) case, the FBLS estimator of $\boldsymbol{\Phi} = [\phi_{2,1}, \phi_{2,2}]^T$ minimizes the sum of squares $\mathrm{SS}_{(\mathrm{FBLS})}(\boldsymbol{\Phi})$ of Equation (477d). Show that this estimator $\hat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ satisfies the equation

$$\begin{bmatrix} 2\sum_{t=1}^{N-2} X_t^2 & A \\ A & \sum_{t=0}^{N-3} X_t^2 + \sum_{t=2}^{N-1} X_t^2 \end{bmatrix} \boldsymbol{\Phi} = \begin{bmatrix} A \\ 2\sum_{t=0}^{N-3} X_t X_{t+2} \end{bmatrix},$$

where you are to determine what A is. For the six-point time series, find and state both elements of $\hat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ along with the estimate of the innovation variance given by Equation (477e). As done for the Burg estimate, create and plot the FBLS AR(2) SDF on a decibel scale versus the frequencies f'_k. How well do the Burg and FBLS SDF estimates agree?

(c) Compute and plot the periodogram for the six-point time series over a fine grid of frequencies. How well does the periodogram agree with the Burg and FBLS AR(2) SDF estimates?

(d) Show that the estimated coefficients $\hat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ obtained in part (b) correspond to an acausal (but stationary) AR(2) process.

(e) (Note: this part assumes familiarity with the solution to Exercise [9.1b].) Find a causal AR(2) process with parameters, say, $\check{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ and $\check{\sigma}^2_{(\mathrm{FBLS})}$ whose SDF is in exact agreement with the one based on $\hat{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ and $\hat{\sigma}^2_{(\mathrm{FBLS})}$. Use $\check{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ in conjunction with Equation (479) to compute an alternative estimate, say $\tilde{\sigma}^2_{(\mathrm{FBLS})}$, for the innovation variance. Compute and plot the AR(2) SDF estimate based on $\check{\boldsymbol{\Phi}}_{(\mathrm{FBLS})}$ and $\tilde{\sigma}^2_{(\mathrm{FBLS})}$ versus f'_k. How well does this estimate agree with the Burg estimate of part (a)?

[9.16] Suppose that we wish to fit an AR(1) model to a time series that is a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of a stationary process with zero mean.

(a) Show that the Yule–Walker, Burg, FLS, and BLS estimators of the coefficient $\phi_{1,1}$ are given by, respectively,

$$\frac{\sum_{t=1}^{N-1} X_t X_{t-1}}{\sum_{t=0}^{N-1} X_t^2}, \quad \frac{\sum_{t=1}^{N-1} X_t X_{t-1}}{\frac{1}{2}X_0^2 + \sum_{t=1}^{N-2} X_t^2 + \frac{1}{2}X_{N-1}^2}, \quad \frac{\sum_{t=1}^{N-1} X_t X_{t-1}}{\sum_{t=0}^{N-2} X_t^2} \text{ and } \frac{\sum_{t=1}^{N-1} X_t X_{t-1}}{\sum_{t=1}^{N-1} X_t^2}.$$

(b) Show that the Yule–Walker estimator of $\phi_{1,1}$ must be less than or equal to the Burg estimator in magnitude. Assuming $\{X_t\}$ is multivariate Gaussian, what is the probability that the Yule–Walker and Burg estimators are equal?

(c) Show that the FBLS estimator of $\phi_{1,1}$ is identical to that of the Burg estimator (this relationship does not hold in general when $p > 1$).

(d) Consider the case $N = 2$ (Jones, 1985, p. 226). Show that, whereas the Yule–Walker, Burg and FBLS estimators are bounded in magnitude, the FLS estimator and the BLS estimator need not be. Determine the MLE when $N = 2$ (see Equation (482d)). Assuming a bivariate Gaussian distribution for X_0 and X_1, discuss briefly which of the six estimators can or cannot yield an estimated AR(1) model that is causal (and hence stationary).

[9.17] Show that Equation (484b) is true.

[9.18] Assuming that Equation (484b) is true (the burden of Exercise [9.17]), show how the FBLS estimator of the AR(p) coefficients can be regarded as an approximation to the MLE. Hint: consider the relationship between $l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \boldsymbol{H})$ and $l(\boldsymbol{\Phi}_p, \sigma_p^2 \mid \widetilde{\boldsymbol{H}})$ and how it impacts Equation (484c); derive an equivalent of Equation (481a) that uses $\widetilde{\boldsymbol{H}}$ rather than $\boldsymbol{H}$; produce the analog of Equation (482a); and then use approximations similar to those that yielded the FLS estimator as an approximation to the MLE.

[9.19] Figure 496 shows two different AR(5) SDF estimates for the ocean wave time series (available from the website for this book): the thick smooth curve is the Yule–Walker estimate, and the dashed curve, the Burg estimate. We used the Burg estimates $\bar{\phi}_{5,1}, \ldots, \bar{\phi}_{5,5}$ to create a prewhitening filter, leading to the prediction errors shown in Figure 497(a) and the postcolored $m = 55$ Parzen lag window estimate $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ shown as the thick smooth curve in Figure 497(c). To investigate how sensitive $\hat{S}_X^{(\mathrm{PC})}(\cdot)$ is to the chosen prewhitening filter, use the Yule–Walker estimates $\tilde{\phi}_{5,1}, \ldots, \tilde{\phi}_{5,5}$ rather than the Burg estimates to create the prewhitening filter, and produce plots similar to Figure 497(a), (b) and (c). Comment briefly on how the Yule–Walker-based postcolored $m = 55$ Parzen lag window estimate compares with the corresponding Burg-based estimate.

https://doi.org/10.1017/9781139235723.010 Published online by Cambridge University Press