9 State-Space Models

In recent years state-space representations and the associated Kalman recursions have had a profound impact on time series analysis and many related areas. The techniques were originally developed in connection with the control of linear systems (for accounts of this subject see Davis and Vinter 1985; Hannan and Deistler 1988). An extremely rich class of models for time series, including and going well beyond the linear ARIMA and classical decomposition models considered so far in this book, can be formulated as special cases of the general state-space model defined below in Section 9.1. In econometrics the structural time series models developed by Harvey (1990) are formulated (like the classical decomposition model) directly in terms of components of interest such as trend, seasonal component, and noise. However, the rigidity of the classical decomposition model is avoided by allowing the trend and seasonal components to evolve randomly rather than deterministically. An introduction to these structural models is given in Section 9.2, and a state-space representation is developed for a general ARIMA process in Section 9.3. The Kalman recursions, which play a key role in the analysis of state-space models, are derived in Section 9.4. These recursions allow a unified approach to prediction and estimation for all processes that can be given a state-space representation. Following the development of the Kalman recursions we discuss estimation with structural models (Section 9.5) and the formulation of state-space models to deal with missing values (Section 9.6). In Section 9.7 we introduce the EM algorithm, an iterative procedure for maximizing the

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2_9

likelihood when only a subset of the complete data set is available. The EM algorithm is particularly well suited for estimation problems in the state-space framework. Generalized state-space models are introduced in Section 9.8. These are Bayesian models that can be used to represent time series of many different types, as demonstrated by two applications to time series of count data. Throughout the chapter we shall use the notation

$$\{\mathbf{W}_t\} \sim \mathrm{WN}(\mathbf{0}, \{R_t\})$$

to indicate that the random vectors $\mathbf{W}_t$ have mean $\mathbf{0}$ and that

$$E\left(\mathbf{W}_s\mathbf{W}_t'\right) = \begin{cases} R_t, & \text{if } s = t,\\ 0, & \text{otherwise.}\end{cases}$$

9.1 State-Space Representations

A state-space model for a (possibly multivariate) time series $\{\mathbf{Y}_t, t = 1, 2, \ldots\}$ consists of two equations. The first, known as the **observation equation**, expresses the w-dimensional observation $\mathbf{Y}_t$ as a linear function of a v-dimensional state variable $\mathbf{X}_t$ plus noise. Thus

$$\mathbf{Y}_t = G_t\mathbf{X}_t + \mathbf{W}_t, \quad t = 1, 2, \ldots, \tag{9.1.1}$$

where $\{\mathbf{W}_t\} \sim \mathrm{WN}(\mathbf{0}, \{R_t\})$ and $\{G_t\}$ is a sequence of $w \times v$ matrices. The second equation, called the **state equation**, determines the state $\mathbf{X}_{t+1}$ at time $t+1$ in terms of the previous state $\mathbf{X}_t$ and a noise term. The state equation is

$$\mathbf{X}_{t+1} = F_t\mathbf{X}_t + \mathbf{V}_t, \quad t = 1, 2, \ldots, \tag{9.1.2}$$

where $\{F_t\}$ is a sequence of $v \times v$ matrices, $\{\mathbf{V}_t\} \sim \mathrm{WN}(\mathbf{0}, \{Q_t\})$, and $\{\mathbf{V}_t\}$ is uncorrelated with $\{\mathbf{W}_t\}$ (i.e., $E(\mathbf{W}_t\mathbf{V}_s') = 0$ for all s and t). To complete the specification, it is assumed that the initial state $\mathbf{X}_1$ is uncorrelated with all of the noise terms $\{\mathbf{V}_t\}$ and $\{\mathbf{W}_t\}$.

Remark 1. A more general form of the state-space model allows for correlation between $\mathbf{V}_t$ and $\mathbf{W}_t$ (see Brockwell and Davis (1991), Chapter 12) and for the addition of a control term $H_t\mathbf{u}_t$ in the state equation. In control theory, $H_t\mathbf{u}_t$ represents the effect of applying a "control" $\mathbf{u}_t$ at time t for the purpose of influencing $\mathbf{X}_{t+1}$. However, the system defined by (9.1.1) and (9.1.2) with $E\left(\mathbf{W}_t\mathbf{V}_s'\right) = 0$ for all s and t will be adequate for our purposes. □

Remark 2. In many important special cases, the matrices F_t, G_t, Q_t, and R_t will be independent of t, in which case the subscripts will be suppressed. □

Remark 3. It follows from the observation equation (9.1.1) and the state equation (9.1.2) that $\mathbf{X}_t$ and $\mathbf{Y}_t$ have the functional forms, for $t = 2, 3, \ldots$,

$$\begin{aligned}\mathbf{X}_t &= F_{t-1}\mathbf{X}_{t-1} + \mathbf{V}_{t-1}\\ &= F_{t-1}(F_{t-2}\mathbf{X}_{t-2} + \mathbf{V}_{t-2}) + \mathbf{V}_{t-1}\\ &\ \ \vdots\\ &= (F_{t-1}\cdots F_1)\mathbf{X}_1 + (F_{t-1}\cdots F_2)\mathbf{V}_1 + \cdots + F_{t-1}\mathbf{V}_{t-2} + \mathbf{V}_{t-1}\\ &= f_t(\mathbf{X}_1, \mathbf{V}_1, \ldots, \mathbf{V}_{t-1})\end{aligned} \tag{9.1.3}$$

and

$$\mathbf{Y}_t = g_t(\mathbf{X}_1, \mathbf{V}_1, \ldots, \mathbf{V}_{t-1}, \mathbf{W}_t). \qquad \square \qquad (9.1.4)$$

Remark 4. From Remark 3 and the assumptions on the noise terms, it is clear that

$$E\left(\mathbf{V}_t\mathbf{X}_s'\right) = 0, \qquad E\left(\mathbf{V}_t\mathbf{Y}_s'\right) = 0, \quad 1 \le s \le t,$$

and

$$E\left(\mathbf{W}_t\mathbf{X}_s'\right) = 0, \quad 1 \le s \le t, \qquad E(\mathbf{W}_t\mathbf{Y}_s') = 0, \quad 1 \le s < t. \qquad \square$$

Definition 9.1.1 A time series $\{\mathbf{Y}_t\}$ has a **state-space representation** if there exists a state-space model for $\{\mathbf{Y}_t\}$ as specified by equations (9.1.1) and (9.1.2).

As already indicated, it is possible to find a state-space representation for a large number of time-series (and other) models. It is clear also from the definition that neither $\{\mathbf{X}_t\}$ nor $\{\mathbf{Y}_t\}$ is necessarily stationary. The beauty of a state-space representation, when one can be found, lies in the simple structure of the state equation (9.1.2), which permits relatively simple analysis of the process $\{\mathbf{X}_t\}$. The behavior of $\{\mathbf{Y}_t\}$ is then easy to determine from that of $\{\mathbf{X}_t\}$ using the observation equation (9.1.1). If the sequence $\{\mathbf{X}_1, \mathbf{V}_1, \mathbf{V}_2, \ldots\}$ is independent, then $\{\mathbf{X}_t\}$ has the Markov property; i.e., the distribution of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t, \ldots, \mathbf{X}_1$ is the same as the distribution of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t$. This is a property possessed by many physical systems, provided that we include sufficiently many components in the specification of the state $\mathbf{X}_t$ (for example, we may choose the state vector in such a way that $\mathbf{X}_t$ includes components of $\mathbf{X}_{t-1}$ for each t).

Example 9.1.1 An AR(1) Process

Let $\{Y_t\}$ be the causal AR(1) process given by

$$Y_t = \phi Y_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right). \qquad (9.1.5)$$

In this case, a state-space representation for $\{Y_t\}$ is easy to construct. We can, for example, define a sequence of state variables X_t by

$$X_{t+1} = \phi X_t + V_t, \quad t = 1, 2, \ldots, \qquad (9.1.6)$$

where $X_1 = Y_1 = \sum_{j=0}^{\infty} \phi^j Z_{1-j}$ and $V_t = Z_{t+1}$. The process $\{Y_t\}$ then satisfies the observation equation

$$Y_t = X_t,$$

which has the form (9.1.1) with $G_t = 1$ and $W_t = 0$.

$\square$

Example 9.1.2 An ARMA(1,1) Process

Let $\{Y_t\}$ be the causal and invertible ARMA(1,1) process satisfying the equations

$$Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right). \qquad (9.1.7)$$

Although the existence of a state-space representation for $\{Y_t\}$ is not obvious, we can find one by observing that

$$Y_t = \theta(B)X_t = \begin{bmatrix} \theta & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix}, \qquad (9.1.8)$$

where $\{X_t\}$ is the causal AR(1) process satisfying

$$\phi(B)X_t = Z_t,$$

or the equivalent equation

$$\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix} + \begin{bmatrix} 0 \\ Z_{t+1}. \end{bmatrix}. \tag{9.1.9}$$

Noting that $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$, we see that equations (9.1.8) and (9.1.9) for $t = 1, 2, \ldots$ furnish a state-space representation of $\{Y_t\}$ with

$$\mathbf{X}_t = \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix} \text{ and } \mathbf{X}_1 = \begin{bmatrix} \sum\limits_{j=0}^{\infty} \phi^j Z_{-j} \\ \sum\limits_{j=0}^{\infty} \phi^j Z_{1-j} \end{bmatrix}.$$

The extension of this state-space representation to general ARMA and ARIMA processes is given in Section 9.3.

□

In subsequent sections we shall give examples that illustrate the versatility of state-space models. (More examples can be found in Aoki 1987; Hannan and Deistler 1988; Harvey 1990; West and Harrison 1989.) Before considering these, we need a slight modification of (9.1.1) and (9.1.2), which allows for series in which the time index runs from $-\infty$ to ∞. This is a more natural formulation for many time series models.

9.1.1 State-Space Models with $t \in \{0, \pm 1, \ldots\}$

Consider the observation and state equations

$$\mathbf{Y}_t = G\mathbf{X}_t + \mathbf{W}_t, \qquad t = 0, \pm 1, \ldots, \tag{9.1.10}$$

$$\mathbf{X}_{t+1} = F\mathbf{X}_t + \mathbf{V}_t, \qquad t = 0, \pm 1, \ldots, \tag{9.1.11}$$

where F and G are $v \times v$ and $w \times v$ matrices, respectively, $\{\mathbf{V}_t\} \sim \text{WN}(\mathbf{0}, Q)$, $\{\mathbf{W}_t\} \sim \text{WN}(\mathbf{0}, R)$, and $E(\mathbf{V}_s\mathbf{W}_t') = 0$ for all s, and t.

The state equation (9.1.11) is said to be **stable** if the matrix F has all its eigenvalues in the interior of the unit circle, or equivalently if $\det(I - Fz) \neq 0$ for all z complex such that $|z| \leq 1$. The matrix F is then also said to be stable.

In the stable case equation (9.1.11) has the unique stationary solution (Problem 9.1) given by

$$\mathbf{X}_t = \sum_{j=0}^{\infty} F^j \mathbf{V}_{t-j-1}.$$

The corresponding sequence of observations

$$\mathbf{Y}_t = \mathbf{W}_t + \sum_{j=0}^{\infty} GF^j \mathbf{V}_{t-j-1}$$

is also stationary.

9.2 The Basic Structural Model

A structural time series model, like the classical decomposition model defined by (1.5.1), is specified in terms of components such as trend, seasonality, and noise, which are of direct interest in themselves. The deterministic nature of the trend and seasonal components in the classical decomposition model, however, limits its applicability. A natural way in which to overcome this deficiency is to permit random variation in these components. This can be very conveniently done in the framework of a state-space representation, and the resulting rather flexible model is called a structural model. Estimation and forecasting with this model can be encompassed in the general procedure for state-space models made possible by the Kalman recursions of Section 9.4.

Example 9.2.1 The Random Walk Plus Noise Model

One of the simplest structural models is obtained by adding noise to a random walk. It is suggested by the nonseasonal classical decomposition model

$$Y_t = M_t + W_t, \quad \text{where } \{W_t\} \sim \text{WN}\left(0, \sigma_w^2\right), \tag{9.2.1}$$

and $M_t = m_t$, the deterministic "level" or "signal" at time t. We now introduce randomness into the level by supposing that M_t is a random walk satisfying

$$M_{t+1} = M_t + V_t, \quad \text{and} \quad \{V_t\} \sim \text{WN}\left(0, \sigma_v^2\right), \tag{9.2.2}$$

with initial value $M_1 = m_1$. Equations (9.2.1) and (9.2.2) constitute the "local level" or "random walk plus noise" model. Figure 9-1 shows a realization of length 100 of this model with $M_1 = 0$, $\sigma_v^2 = 4$, and $\sigma_w^2 = 8$. (The realized values m_t of M_t are plotted as a solid line, and the observed data are plotted as square boxes.) The differenced data

$$D_t := \nabla Y_t = Y_t - Y_{t-1} = V_{t-1} + W_t - W_{t-1}, \quad t \ge 2,$$

constitute a stationary time series with mean 0 and ACF

$$\rho_D(h) = \begin{cases} \dfrac{-\sigma_w^2}{2\sigma_w^2 + \sigma_v^2}, & \text{if } |h| = 1, \\ 0, & \text{if } |h| > 1. \end{cases}$$

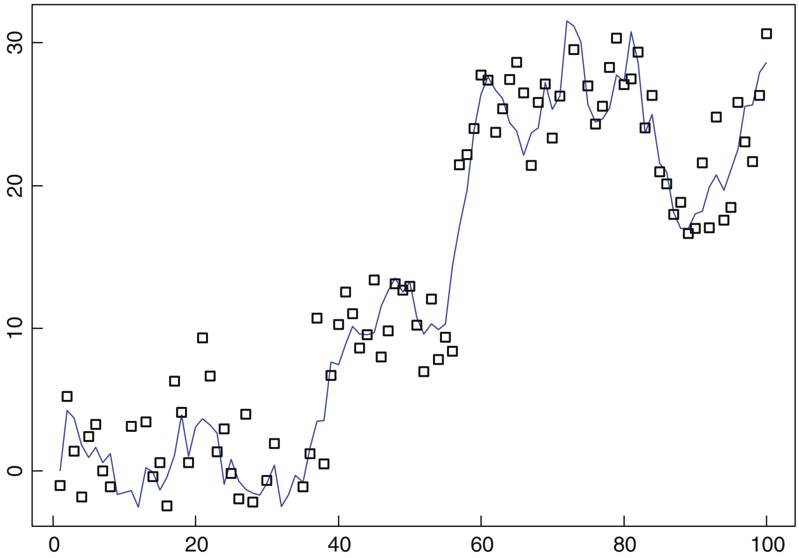

Figure 9-1 Realization from a random walk plus noise model. The random walk is represented by the *solid line* and the data are represented by *boxes*

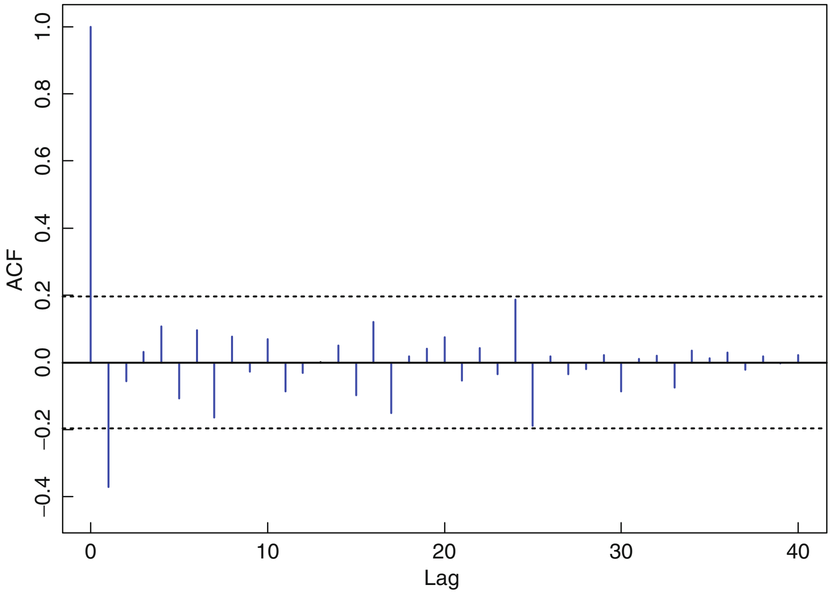

Figure 9-2 Sample ACF of the series obtained by differencing the data in Figure 9-1

Since $\{D_t\}$ is 1-correlated, we conclude from Proposition 2.1.1 that $\{D_t\}$ is an MA(1) process and hence that $\{Y_t\}$ is an ARIMA(0,1,1) process. More specifically,

$$D_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right), \tag{9.2.3}$$

where θ and σ^2 are found by solving the equations

$$\frac{\theta}{1+\theta^2} = \frac{-\sigma_w^2}{2\sigma_w^2 + \sigma_v^2} \quad \text{and} \quad \theta\sigma^2 = -\sigma_w^2.$$

For the process $\{Y_t\}$ generating the data in Figure 9-1, the parameters θ and σ^2 of the differenced series $\{D_t\}$ satisfy $\theta/(1+\theta^2) = -0.4$ and $\theta\sigma^2 = -8$. Solving these equations for θ and σ^2, we find that $\theta = -0.5$ and $\sigma^2 = 16$ (or $\theta = -2$ and $\sigma^2 = 4$). The sample ACF of the observed differences D_t of the realization of $\{Y_t\}$ in Figure 9-1 is shown in Figure 9-2.

The local level model is often used to represent a measured characteristic of the output of an industrial process for which the unobserved process level $\{M_t\}$ is intended to be within specified limits (to meet the design specifications of the manufactured product). To decide whether or not the process requires corrective attention, it is important to be able to test the hypothesis that the process level $\{M_t\}$ is constant. From the state equation, we see that $\{M_t\}$ is constant (and equal to m_1) when $V_t = 0$ or equivalently when $\sigma_v^2 = 0$. This in turn is equivalent to the moving-average model (9.2.3) for $\{D_t\}$ being noninvertible with $\theta = -1$ (see Problem 8.2). Tests of the unit root hypothesis $\theta = -1$ were discussed in Section 6.3.2.

□

The local level model can easily be extended to incorporate a locally linear trend with slope β_t at time t. Equation (9.2.2) is replaced by

$$M_t = M_{t-1} + B_{t-1} + V_{t-1}, \tag{9.2.4}$$

where $B_{t-1} = \beta_{t-1}$. Now if we introduce randomness into the slope by replacing it with the random walk

$$B_t = B_{t-1} + U_{t-1}, \quad \text{where } \{U_t\} \sim \text{WN}\left(0, \sigma_u^2\right), \tag{9.2.5}$$

we obtain the “local linear trend” model.

To express the local linear trend model in state-space form we introduce the state vector

$$\mathbf{X}_t = (M_t, B_t)'.$$

Then (9.2.4) and (9.2.5) can be written in the equivalent form

$$\mathbf{X}_{t+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{X}_t + \mathbf{V}_t, \quad t = 1, 2, \ldots, \tag{9.2.6}$$

where $\mathbf{V}_t = (V_t, U_t)'$. The process $\{Y_t\}$ is then determined by the observation equation

$$Y_t = [1 \quad 0]\, \mathbf{X}_t + W_t. \tag{9.2.7}$$

If $\{\mathbf{X}_1, U_1, V_1, W_1, U_2, V_2, W_2, \ldots\}$ is an uncorrelated sequence, then equations (9.2.6) and (9.2.7) constitute a state-space representation of the process $\{Y_t\}$, which is a model for data with randomly varying trend and added noise. For this model we have $v = 2$, $w = 1$,

$$F = \begin{bmatrix} 1 & 1 \\ 0 & 1, \end{bmatrix} \quad G = [1 \quad 0], \quad Q = \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}, \quad \text{and } R = \sigma_w^2.$$

Example 9.2.2 A Seasonal Series with Noise

The classical decomposition (1.5.11) expressed the time series $\{X_t\}$ as a sum of trend, seasonal, and noise components. The seasonal component (with period d) was a sequence $\{s_t\}$ with the properties $s_{t+d} = s_t$ and $\sum_{t=1}^{d} s_t = 0$. Such a sequence can be generated, for *any* values of $s_1, s_0, \ldots, s_{-d+3}$, by means of the recursions

$$s_{t+1} = -s_t - \cdots - s_{t-d+2}, \quad t = 1, 2, \ldots. \tag{9.2.8}$$

A somewhat more general seasonal component $\{Y_t\}$, allowing for random deviations from strict periodicity, is obtained by adding a term S_t to the right side of (9.2.8), where $\{V_t\}$ is white noise with mean zero. This leads to the recursion relations

$$Y_{t+1} = -Y_t - \cdots - Y_{t-d+2} + S_t, \quad t = 1, 2, \ldots. \tag{9.2.9}$$

To find a state-space representation for $\{Y_t\}$ we introduce the $(d-1)$-dimensional state vector

$$\mathbf{X}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+2})'.$$

The series $\{Y_t\}$ is then given by the observation equation

$$Y_t = [1 \quad 0 \quad 0 \cdots 0]\, \mathbf{X}_t, \quad t = 1, 2, \ldots, \tag{9.2.10}$$

where $\{\mathbf{X}_t\}$ satisfies the state equation

$$\mathbf{X}_{t+1} = F\mathbf{X}_t + \mathbf{V}_t, \quad t = 1, 2 \ldots, \tag{9.2.11}$$

$\mathbf{V}_t = (S_t, 0, \ldots, 0)'$, and

$$F = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{9.2.12}$$

□

Example 9.2.3 A Randomly Varying Trend with Random Seasonality and Noise

A series with randomly varying trend, random seasonality and noise can be constructed by adding the two series in Examples 9.2.1 and 9.2.2. (Addition of series with state-space representations is in fact always possible by means of the following construction. See Problem 9.9.) We introduce the state vector

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{X}_t^1 \\ \mathbf{X}_t^2 \end{bmatrix},$$

where $\mathbf{X}_t^1$ and $\mathbf{X}_t^2$ are the state vectors in (9.2.6) and (9.2.11). We then have the following representation for $\{Y_t\}$, the sum of the two series whose state-space representations were given in (9.2.6)–(9.2.7) and (9.2.10)–(9.2.11). The state equation is

$$\mathbf{X}_{t+1} = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix} \mathbf{X}_t + \begin{bmatrix} \mathbf{V}_t^1 \\ \mathbf{V}_t^2 \end{bmatrix}, \tag{9.2.13}$$

where F_1, F_2 are the coefficient matrices and $\{\mathbf{V}_t^1\}$, $\{\mathbf{V}_t^2\}$ are the noise vectors in the state equations (9.2.6) and (9.2.11), respectively. The observation equation is

$$Y_t = [1 \quad 0 \quad 1 \quad 0 \cdots 0]\, \mathbf{X}_t + W_t, \tag{9.2.14}$$

where $\{W_t\}$ is the noise sequence in (9.2.7). If the sequence of random vectors $\{\mathbf{X}_1, \mathbf{V}_1^1, \mathbf{V}_1^2, W_1, \mathbf{V}_2^1, \mathbf{V}_2^2, W_2, \ldots\}$ is uncorrelated, then equations (9.2.13) and (9.2.14) constitute a state-space representation for $\{Y_t\}$.

□

9.3 State-Space Representation of ARIMA Models

We begin by establishing a state-space representation for the causal AR(p) process and then build on this example to find representations for the general ARMA and ARIMA processes.

Example 9.3.1 State-Space Representation of a Causal AR(p) Process

Consider the AR(p) process defined by

$$Y_{t+1} = \phi_1 Y_t + \phi_2 Y_{t-1} + \cdots + \phi_p Y_{t-p+1} + Z_{t+1}, \quad t = 0, \pm 1, \ldots, \tag{9.3.1}$$

where $\{Z_t\} \sim \text{WN}\big(0, \sigma^2\big)$, and $\phi(z) := 1 - \phi_1 z - \cdots - \phi_p z^p$ is nonzero for $|z| \le 1$. To express $\{Y_t\}$ in state-space form we simply introduce the state vectors

$$\mathbf{X}_t = \begin{bmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_t, \end{bmatrix}, \quad t = 0, \pm 1, \ldots. \tag{9.3.2}$$

From (9.3.1) and (9.3.2) the observation equation is

$$Y_t = [0 \quad 0 \quad 0 \cdots 1]\mathbf{X}_t, \quad t = 0, \pm 1, \ldots, \tag{9.3.3}$$

while the state equation is given by

$$\mathbf{X}_{t+1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_p & \phi_{p-1} & \phi_{p-2} & \cdots & \phi_1 \end{bmatrix} \mathbf{X}_t + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} Z_{t+1}, \quad t = 0, \pm 1, \ldots. \tag{9.3.4}$$

These equations have the required forms (9.1.10) and (9.1.11) with $\mathbf{W}_t = \mathbf{0}$ and $\mathbf{V}_t = (0, 0, \ldots, Z_{t+1})'$, $t = 0, \pm 1, \ldots$.

□

Remark 1. In Example 9.3.1 the causality condition $\phi(z) \neq 0$ for $|z| \leq 1$ is equivalent to the condition that the state equation (9.3.4) is stable, since the eigenvalues of the coefficient matrix in (9.3.4) are simply the reciprocals of the zeros of $\phi(z)$ (Problem 9.3). □

Remark 2. If equations (9.3.3) and (9.3.4) are postulated to hold only for $t = 1, 2, \ldots$, and if $\mathbf{X}_1$ is a random vector such that $\{\mathbf{X}_1, Z_1, Z_2, \ldots\}$ is an uncorrelated sequence, then we have a state-space representation for $\{Y_t\}$ of the type defined earlier by (9.1.1) and (9.1.2). The resulting process $\{Y_t\}$ is well-defined, regardless of whether or not the state equation is stable, but it will not in general be stationary. It will be stationary if the state equation is stable and if $\mathbf{X}_1$ is defined by (9.3.2) with $Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, $t = 1, 0, \ldots, 2-p$, and $\psi(z) = 1/\phi(z)$, $|z| \leq 1$. □

Example 9.3.2 State-Space Form of a Causal ARMA(p, q) Process

State-space representations are not unique. Here we shall give one of the (infinitely many) possible representations of a causal ARMA(p,q) process that can easily be derived from Example 9.3.1. Consider the ARMA(p,q) process defined by

$$\phi(B)Y_t = \theta(B)Z_t, \quad t = 0, \pm 1, \ldots, \tag{9.3.5}$$

where $\{Z_t\} \sim \mathrm{WN}\big(0, \sigma^2\big)$ and $\phi(z) \neq 0$ for $|z| \leq 1$. Let

$$r = \max(p, q+1), \quad \phi_j = 0 \quad \text{for } j > p, \quad \theta_j = 0 \quad \text{for } j > q, \quad \text{and} \quad \theta_0 = 1.$$

If $\{U_t\}$ is the causal AR(p) process satisfying

$$\phi(B)U_t = Z_t, \tag{9.3.6}$$

then $Y_t = \theta(B)U_t$, since

$$\phi(B)Y_t = \phi(B)\theta(B)U_t = \theta(B)\phi(B)U_t = \theta(B)Z_t.$$

Consequently,

$$Y_t = [\theta_{r-1} \quad \theta_{r-2} \cdots \theta_0]\mathbf{X}_t, \tag{9.3.7}$$

where

$$\mathbf{X}_t = \begin{bmatrix} U_{t-r+1} \\ U_{t-r+2} \\ \vdots \\ U_t \end{bmatrix}. \tag{9.3.8}$$

But from Example 9.3.1 we can write

$$\mathbf{X}_{t+1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_r & \phi_{r-1} & \phi_{r-2} & \cdots & \phi_1 \end{bmatrix} \mathbf{X}_t + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} Z_{t+1}, \quad t = 0, \pm 1, \ldots. \tag{9.3.9}$$

Equations (9.3.7) and (9.3.9) are the required observation and state equations. As in Example 9.3.1, the observation and state noise vectors are again $\mathbf{W}_t = \mathbf{0}$ and $\mathbf{V}_t = (0, 0, \ldots, Z_{t+1})'$, $t = 0, \pm 1, \ldots$.

□

Example 9.3.3 State-Space Representation of an ARIMA(p, d, q) Process

If $\{Y_t\}$ is an ARIMA(p, d, q) process with $\{\nabla^d Y_t\}$ satisfying (9.3.5), then by the preceding example $\{\nabla^d Y_t\}$ has the representation

$$\nabla^d Y_t = G\mathbf{X}_t, \quad t = 0, \pm 1, \ldots, \tag{9.3.10}$$

where $\{\mathbf{X}_t\}$ is the unique stationary solution of the state equation

$$\mathbf{X}_{t+1} = F\mathbf{X}_t + \mathbf{V}_t,$$

F and G are the coefficients of $\mathbf{X}_t$ in (9.3.9) and (9.3.7), respectively, and $\mathbf{V}_t = (0, 0, \ldots, Z_{t+1})'$. Let A and B be the $d \times 1$ and $d \times d$ matrices defined by $A = B = 1$ if $d = 1$ and

$$A = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ (-1)^{d+1}\binom{d}{d} & (-1)^{d}\binom{d}{d-1} & (-1)^{d-1}\binom{d}{d-2} & \cdots & d \end{bmatrix}$$

if $d > 1$. Then since

$$Y_t = \nabla^d Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j Y_{t-j}, \tag{9.3.11}$$

the vector

$$\mathbf{Y}_{t-1} := (Y_{t-d}, \ldots, Y_{t-1})'$$

satisfies the equation

$$\mathbf{Y}_t = A\nabla^d Y_t + B\mathbf{Y}_{t-1} = AG\mathbf{X}_t + B\mathbf{Y}_{t-1}.$$

Defining a new state vector $\mathbf{T}_t$ by stacking $\mathbf{X}_t$ and $\mathbf{Y}_{t-1}$, we therefore obtain the state equation

$$\mathbf{T}_{t+1} := \begin{bmatrix} \mathbf{X}_{t+1} \\ \mathbf{Y}_t \end{bmatrix} = \begin{bmatrix} F & 0 \\ AG & B \end{bmatrix} \mathbf{T}_t + \begin{bmatrix} \mathbf{V}_t \\ \mathbf{0} \end{bmatrix}, \quad t = 1, 2, \ldots, \tag{9.3.12}$$

and the observation equation, from (9.3.10) and (9.3.11),

$$Y_t = \left[G \; (-1)^{d+1}\binom{d}{d} \; (-1)^{d}\binom{d}{d-1} \; (-1)^{d-1}\binom{d}{d-2} \; \cdots \; d \right] \begin{bmatrix} \mathbf{X}_t \\ \mathbf{Y}_{t-1} \end{bmatrix}, \qquad t = 1, 2, \ldots, \tag{9.3.13}$$

with initial condition

$$\mathbf{T}_1 = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{Y}_0 \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{\infty} F^j \mathbf{V}_{-j} \\ \mathbf{Y}_0 \end{bmatrix}, \tag{9.3.14}$$

and the assumption

$$E(\mathbf{Y}_0 Z_t') = 0, \quad t = 0, \pm 1, \ldots, \tag{9.3.15}$$

where $\mathbf{Y}_0 = (Y_{1-d}, Y_{2-d}, \ldots, Y_0)'$. The conditions (9.3.15), which are satisfied in particular if $\mathbf{Y}_0$ is considered to be nonrandom and equal to the vector of *observed* values $(y_{1-d}, y_{2-d}, \ldots, y_0)'$, are imposed to ensure that the assumptions of a state-space model given in Section 9.1 are satisfied. They also imply that $E\left(\mathbf{X}_1 \mathbf{Y}_0'\right) = 0$ and $E(\mathbf{Y}_0 \nabla^d Y_t') = 0$, $t \ge 1$, as required earlier in Section 6.4 for prediction of ARIMA processes.

State-space models for more general ARIMA processes (e.g., $\{Y_t\}$ such that $\{\nabla\nabla_{12} Y_t\}$ is an ARMA(p, q) process) can be constructed in the same way. See Problem 9.4.

□

For the ARIMA(1, 1, 1) process defined by

$$(1 - \phi B)(1 - B)Y_t = (1 + \theta B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}\left(0, \sigma^2\right),$$

the vectors $\mathbf{X}_t$ and $\mathbf{Y}_{t-1}$ reduce to $\mathbf{X}_t = (X_{t-1}, X_t)'$ and $\mathbf{Y}_{t-1} = Y_{t-1}$. From (9.3.12) and (9.3.13) the state-space representation is therefore (Problem 9.8)

$$Y_t = \begin{bmatrix} \theta & 1 & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \\ Y_{t-1} \end{bmatrix}, \tag{9.3.16}$$

where

$$\begin{bmatrix} X_t \\ X_{t+1} \\ Y_t \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \phi & 0 \\ \theta & 1 & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ Z_{t+1} \\ 0 \end{bmatrix}, \quad t = 1, 2, \ldots, \tag{9.3.17}$$

and

$$\begin{bmatrix} X_0 \\ X_1 \\ Y_0 \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{\infty} \phi^j Z_{-j} \\ \sum_{j=0}^{\infty} \phi^j Z_{1-j} \\ Y_0 \end{bmatrix}. \tag{9.3.18}$$

9.4 The Kalman Recursions

In this section we shall consider three fundamental problems associated with the state-space model defined by (9.1.1) and (9.1.2) in Section 9.1. These are all concerned with finding best (in the sense of minimum mean square error) linear estimates of the state-vector $\mathbf{X}_t$ in terms of the observations $\mathbf{Y}_1, \mathbf{Y}_2, \dots,$ and a random vector $\mathbf{Y}_0$ that is orthogonal to $\mathbf{V}_t$ and $\mathbf{W}_t$ for all $t \geq 1$. In many cases $\mathbf{Y}_0$ will be the constant vector $(1, 1, \dots, 1)'$. Estimation of $\mathbf{X}_t$ in terms of:

a. $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$ defines the **prediction problem**,
b. $\mathbf{Y}_0, \dots, \mathbf{Y}_t$ defines the **filtering problem**,
c. $\mathbf{Y}_0, \dots, \mathbf{Y}_n$ $(n > t)$ defines the **smoothing problem**.

Each of these problems can be solved recursively using an appropriate set of Kalman recursions, which will be established in this section.

In the following definition of best linear predictor (and throughout this chapter) it should be noted that we do not automatically include the constant 1 among the predictor variables as we did in Sections 2.5 and 8.5. (It can, however, be included by choosing $\mathbf{Y}_0 = (1, 1, \dots, 1)'$.)

Definition 9.4.1

For the random vector $\mathbf{X} = (X_1, \dots, X_v)'$,

$$P_t(\mathbf{X}) := (P_t(X_1), \dots, P_t(X_v))',$$

where $P_t(X_i) := P(X_i|\mathbf{Y}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_t)$, is the best linear predictor of X_i in terms of all components of $\mathbf{Y}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_t$.

Remark 1. By the definition of the best predictor of each component X_i of $\mathbf{X}$, $P_t(\mathbf{X})$ is the unique random vector of the form

$$P_t(\mathbf{X}) = A_0\mathbf{Y}_0 + \cdots + A_t\mathbf{Y}_t$$

with $v \times w$ matrices $A_0, \dots, A_t$ such that

$$[\mathbf{X} - P_t(\mathbf{X})] \perp \mathbf{Y}_s, \quad s = 0, \dots, t$$

[cf. (8.5.2) and (8.5.3)]. Recall that two random vectors $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal (written $\mathbf{X} \perp \mathbf{Y}$) if $E(\mathbf{X}\mathbf{Y}')$ is a matrix of zeros. □

Remark 2. If all the components of $\mathbf{X}, \mathbf{Y}_1, \dots, \mathbf{Y}_t$ are jointly normally distributed and $\mathbf{Y}_0 = (1, \dots, 1)'$, then

$$P_t(\mathbf{X}) = E(\mathbf{X}|\mathbf{Y}_1, \dots, \mathbf{Y}_t), \quad t \geq 1.$$ □

Remark 3. P_t is linear in the sense that if A is any $k \times v$ matrix and $\mathbf{X}, \mathbf{V}$ are two v-variate random vectors with finite second moments, then (Problem 9.10)

$$P_t(A\mathbf{X}) = AP_t(\mathbf{X})$$

and

$$P_t(\mathbf{X} + \mathbf{V}) = P_t(\mathbf{X}) + P_t(\mathbf{V}).$$

□

Remark 4. If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with v and w components, respectively, each with finite second moments, then

$$P(\mathbf{X}|\mathbf{Y}) = M\mathbf{Y},$$

where M is a $v \times w$ matrix, $M = E(\mathbf{X}\mathbf{Y}')[E(\mathbf{Y}\mathbf{Y}')]^{-1}$ with $[E(\mathbf{Y}\mathbf{Y}')]^{-1}$ any generalized inverse of $E(\mathbf{Y}\mathbf{Y}')$. (A generalized inverse of a matrix S is a matrix S^{-1} such that $SS^{-1}S = S$. Every matrix has at least one. See Problem 9.11.)

In the notation just developed, the prediction, filtering, and smoothing problems (a), (b), and (c) formulated above reduce to the determination of $P_{t-1}(\mathbf{X}_t)$, $P_t(\mathbf{X}_t)$, and $P_n(\mathbf{X}_t)$ $(n > t)$, respectively. We deal first with the prediction problem. □

Kalman Prediction:
For the state-space model (9.1.1)–(9.1.2), the one-step predictors $\hat{\mathbf{X}}_t := P_{t-1}(\mathbf{X}_t)$ and their error covariance matrices $\Omega_t = E\left[\left(\mathbf{X}_t - \hat{\mathbf{X}}_t\right)\left(\mathbf{X}_t - \hat{\mathbf{X}}_t\right)'\right]$ are uniquely determined by the initial conditions

$$\hat{\mathbf{X}}_1 = P(\mathbf{X}_1|\mathbf{Y}_0), \qquad \Omega_1 = E\left[\left(\mathbf{X}_1 - \hat{\mathbf{X}}_1\right)\left(\mathbf{X}_1 - \hat{\mathbf{X}}_1\right)'\right]$$

and the recursions, for $t = 1, \ldots,$

$$\hat{\mathbf{X}}_{t+1} = F_t\hat{\mathbf{X}}_t + \Theta_t\Delta_t^{-1}\left(\mathbf{Y}_t - G_t\hat{\mathbf{X}}_t\right), \tag{9.4.1}$$

$$\Omega_{t+1} = F_t\Omega_t F_t' + Q_t - \Theta_t\Delta_t^{-1}\Theta_t', \tag{9.4.2}$$

where

$$\Delta_t = G_t\Omega_t G_t' + R_t,$$

$$\Theta_t = F_t\Omega_t G_t',$$

and Δ_t^{-1} is any generalized inverse of Δ_t.

Proof. We shall make use of the **innovations** $\mathbf{I}_t$ defined by $\mathbf{I}_0 = \mathbf{Y}_0$ and

$$\mathbf{I}_t = \mathbf{Y}_t - P_{t-1}\mathbf{Y}_t = \mathbf{Y}_t - G_t\hat{\mathbf{X}}_t = G_t\left(\mathbf{X}_t - \hat{\mathbf{X}}_t\right) + \mathbf{W}_t, \quad t = 1, 2, \ldots.$$

The sequence $\{\mathbf{I}_t\}$ is orthogonal by Remark 1. Using Remarks 3 and 4 and the relation

$$P_t(\cdot) = P_{t-1}(\cdot) + P(\cdot|\mathbf{I}_t) \tag{9.4.3}$$

(see Problem 9.12), we find that

$$\begin{aligned}\hat{\mathbf{X}}_{t+1} &= P_{t-1}(\mathbf{X}_{t+1}) + P(\mathbf{X}_{t+1}|\mathbf{I}_t) = P_{t-1}(F_t\mathbf{X}_t + \mathbf{V}_t) + \Theta_t\Delta_t^{-1}\mathbf{I}_t \\ &= F_t\hat{\mathbf{X}}_t + \Theta_t\Delta_t^{-1}\mathbf{I}_t,\end{aligned} \tag{9.4.4}$$

where

$$\Delta_t = E(\mathbf{I}_t\,\mathbf{I}_t') = G_t\Omega_t G_t' + R_t,$$

$$\Theta_t = E(\mathbf{X}_{t+1}\mathbf{I}_t') = E\left[\left(F_t\mathbf{X}_t + \mathbf{V}_t\right)\left(\left[\mathbf{X}_t - \hat{\mathbf{X}}_t\right]'G_t' + \mathbf{W}_t'\right)\right] \quad = F_t\Omega_t G_t'.$$

To verify (9.4.2), we observe from the definition of Ω_{t+1} that

$$\Omega_{t+1} = E\left(\mathbf{X}_{t+1}\mathbf{X}_{t+1}'\right) - E\left(\hat{\mathbf{X}}_{t+1}\hat{\mathbf{X}}_{t+1}'\right).$$

With (9.1.2) and (9.4.4) this gives

$$\begin{aligned}\Omega_{t+1} &= F_t E(\mathbf{X}_t\mathbf{X}_t')F_t' + Q_t - F_t E\left(\hat{\mathbf{X}}_t\hat{\mathbf{X}}_t'\right)F_t' - \Theta_t\Delta_t^{-1}\Theta_t' \\ &= F_t\Omega_t F_t' + Q_t - \Theta_t\Delta_t^{-1}\Theta_t'. \end{aligned}$$ ■

9.4.1 *h*-Step Prediction of $\{\mathbf{Y}_t\}$ Using the Kalman Recursions

The Kalman prediction equations lead to a very simple algorithm for recursive calculation of the best linear mean square predictors $P_t\mathbf{Y}_{t+h}$, $h = 1, 2, \ldots$. From (9.4.4), (9.1.1), (9.1.2), and Remark 3 in Section 9.1, we find that

$$P_t\mathbf{X}_{t+1} = F_tP_{t-1}\mathbf{X}_t + \Theta_t\Delta_t^{-1}(\mathbf{Y}_t - P_{t-1}\mathbf{Y}_t), \tag{9.4.5}$$

$$\begin{aligned}P_t\mathbf{X}_{t+h} &= F_{t+h-1}P_t\mathbf{X}_{t+h-1} \\ &\vdots \\ &= (F_{t+h-1}F_{t+h-2}\cdots F_{t+1})\,P_t\mathbf{X}_{t+1}, \quad h = 2, 3, \ldots,\end{aligned} \tag{9.4.6}$$

and

$$P_t\mathbf{Y}_{t+h} = G_{t+h}P_t\mathbf{X}_{t+h}, \quad h = 1, 2, \ldots. \tag{9.4.7}$$

From the relation

$$\mathbf{X}_{t+h} - P_t\mathbf{X}_{t+h} = F_{t+h-1}(\mathbf{X}_{t+h-1} - P_t\mathbf{X}_{t+h-1}) + \mathbf{V}_{t+h-1}, \quad h = 2, 3, \ldots,$$

we find that $\Omega_t^{(h)} := E[(\mathbf{X}_{t+h} - P_t\mathbf{X}_{t+h})(\mathbf{X}_{t+h} - P_t\mathbf{X}_{t+h})']$ satisfies the recursions

$$\Omega_t^{(h)} = F_{t+h-1}\Omega_t^{(h-1)}F_{t+h-1}' + Q_{t+h-1}, \quad h = 2, 3, \ldots, \tag{9.4.8}$$

with $\Omega_t^{(1)} = \Omega_{t+1}$. Then from (9.1.1) and (9.4.7), $\Delta_t^{(h)} := E[(\mathbf{Y}_{t+h} - P_t\mathbf{Y}_{t+h})(\mathbf{Y}_{t+h} - P_t\mathbf{Y}_{t+h})']$ is given by

$$\Delta_t^{(h)} = G_{t+h}\Omega_t^{(h)}G_{t+h}' + R_{t+h}, \quad h = 1, 2, \ldots. \tag{9.4.9}$$

Example 9.4.1. Consider the random walk plus noise model of Example 9.2.1 defined by

$$Y_t = X_t + W_t, \quad \{W_t\} \sim \mathrm{WN}\left(0, \sigma_w^2\right),$$

where the local level X_t follows the random walk

$$X_{t+1} = X_t + V_t, \quad \{V_t\} \sim \mathrm{WN}\left(0, \sigma_v^2\right).$$

Applying the Kalman prediction equations with $Y_0 := 1$, $R = \sigma_w^2$, and $Q = \sigma_v^2$, we obtain

$$\begin{aligned}\hat{Y}_{t+1} = P_tY_{t+1} &= \hat{X}_t + \frac{\Theta_t}{\Delta_t}\left(Y_t - \hat{Y}_t\right) \\ &= (1 - a_t)\hat{Y}_t + a_tY_t\end{aligned}$$

where

$$a_t = \frac{\Theta_t}{\Delta_t} = \frac{\Omega_t}{\Omega_t + \sigma_w^2}.$$

For a state-space model (like this one) with time-independent parameters, the solution of the Kalman recursions (9.4.2) is called a **steady-state solution** if Ω_t is independent of t. If $\Omega_t = \Omega$ for all t, then from (9.4.2)

$$\Omega_{t+1} = \Omega = \Omega + \sigma_v^2 - \frac{\Omega^2}{\Omega + \sigma_w^2} = \frac{\Omega\sigma_w^2}{\Omega + \sigma_w^2} + \sigma_v^2.$$

Solving this quadratic equation for Ω and noting that $\Omega \geq 0$, we find that

$$\Omega = \frac{1}{2}\left(\sigma_v^2 + \sqrt{\sigma_v^4 + 4\sigma_v^2\sigma_w^2}\right)$$

Since $\Omega_{t+1} - \Omega_t$ is a continuous function of Ω_t on $\Omega_t \geq 0$, positive at $\Omega_t = 0$, negative for large Ω_t, and zero only at $\Omega_t = \Omega$, it is clear that $\Omega_{t+1} - \Omega_t$ is negative for $\Omega_t > \Omega$ and positive for $\Omega_t < \Omega$. A similar argument shows (Problem 9.14) that $(\Omega_{t+1} - \Omega)(\Omega_t - \Omega) \geq 0$ for all $\Omega_t \geq 0$. These observations imply that Ω_{t+1} always falls between Ω and Ω_t. Consequently, regardless of the value of Ω_1, Ω_t converges to Ω, the unique solution of $\Omega_{t+1} = \Omega_t$. For *any* initial predictors $\hat{Y}_1 = \hat{X}_1$ and any initial mean squared error $\Omega_1 = E\left(X_1 - \hat{X}_1\right)^2$, the coefficients $a_t := \Omega_t / \left(\Omega_t + \sigma_w^2\right)$ converge to

$$a = \frac{\Omega}{\Omega + \sigma_w^2},$$

and the mean squared errors of the predictors defined by

$$\hat{Y}_{t+1} = (1 - a_t)\hat{Y}_t + a_t Y_t$$

converge to $\Omega + \sigma_w^2$.

If, as is often the case, we do not know Ω_1, then we cannot determine the sequence $\{a_t\}$. It is natural, therefore, to consider the behavior of the predictors defined by

$$\hat{Y}_{t+1} = (1 - a)\hat{Y}_t + aY_t$$

with a as above and arbitrary $\hat{Y}_1$. It can be shown (Problem 9.16) that this sequence of predictors is also asymptotically optimal in the sense that the mean squared error converges to $\Omega + \sigma_w^2$ as $t \to \infty$.

As shown in Example 9.2.1, the differenced process $D_t = Y_t - Y_{t-1}$ is the MA(1) process

$$D_t = Z_t + \theta Z_{t-1},\ \left\{Z_t\right\} \sim \text{WN}\left(0, \sigma^2\right),$$

where $\theta / \left(1 + \theta^2\right) = -\sigma_w^2 / \left(2\sigma_w^2 + \sigma_v^2\right)$. Solving this equation for θ (Problem 9.15), we find that

$$\theta = -\frac{1}{2\sigma_w^2}\left(2\sigma_w^2 + \sigma_v^2 - \sqrt{\sigma_v^4 + 4\sigma_v^2\sigma_w^2}\right)$$

and that $\theta = a - 1$.

It is instructive to derive the exponential smoothing formula for $\hat{Y}_t$ directly from the ARIMA(0,1,1) structure of $\{Y_t\}$. For $t \geq 2$, we have from Section 6.5 that

$$\hat{Y}_{t+1} = Y_t + \theta_{t1}(Y_t - \hat{Y}_t) = -\theta_{t1}\hat{Y}_t + (1 + \theta_{t1})Y_t$$

for $t \geq 2$, where θ_{t1} is found by application of the innovations algorithm to an MA(1) process with coefficient θ. It follows that $1 - a_t = -\theta_{t1}$, and since $\theta_{t1} \to \theta$

(see Remark 1 of Section 3.3) and a_t converges to the steady-state solution a, we conclude that

$$1 - a = \lim_{t\to\infty}(1 - a_t) = -\lim_{t\to\infty}\theta_{t1} = -\theta.$$

□

Example 9.4.2. The lognormal stochastic volatility model

We can rewrite the defining equations (7.4.2) and (7.4.3) of the lognormal SV process $\{Z_t\}$ in the following state-space form

$$X_t = \gamma_1 X_{t-1} + \eta_t, \tag{9.4.10}$$

and

$$Y_t = X_t + \varepsilon_t, \tag{9.4.11}$$

where the (one-dimensional) state and observation vectors are

$$X_t = \ell_t - \frac{\gamma_0}{1-\gamma_1}, \tag{9.4.12}$$

and

$$Y_t = \ln Z_t^2 + 1.27 - \frac{\gamma_0}{2(1-\gamma_1)} \tag{9.4.13}$$

respectively. The independent white-noise sequences $\{\eta_t\}$ and $\{\varepsilon_t\}$ have zero means and variances σ^2 and 4.93 respectively.

Taking

$$\hat{X}_0 = EX_0 = 0 \tag{9.4.14}$$

and

$$\hat{\Omega}_0 = \mathrm{Var}(X_0) = \sigma^2/(1-\gamma_1^2), \tag{9.4.15}$$

and we can directly apply the Kalman prediction recursions (9.4.1), (9.4.2), (9.4.6) and (9.4.8), to compute recursively the best linear predictor of X_{t+h} in terms of $\{Y_s, s \le t\}$, or equivalently of the log volatility ℓ_{t+h} in terms of the observations $\{\ln Z_s^2, s \le t\}$.

□

Kalman Filtering:

The filtered estimates $\mathbf{X}_{t|t} = P_t(\mathbf{X}_t)$ and their error covariance matrices $\Omega_{t|t} = E[(\mathbf{X}_t - \mathbf{X}_{t|t})(\mathbf{X}_t - \mathbf{X}_{t|t})']$ are determined by the relations

$$P_t\mathbf{X}_t = P_{t-1}\mathbf{X}_t + \Omega_t G_t'\Delta_t^{-1}\left(\mathbf{Y}_t - G_t\hat{\mathbf{X}}_t\right) \tag{9.4.16}$$

and

$$\Omega_{t|t} = \Omega_t - \Omega_t G_t'\Delta_t^{-1}G_t\Omega_t'. \tag{9.4.17}$$

Proof. From (9.4.3) it follows that

$$P_t\mathbf{X}_t = P_{t-1}\mathbf{X}_t + M\mathbf{I}_t,$$

where

$$M = E(\mathbf{X}_t\,\mathbf{I}_t')[E(\mathbf{I}_t\,\mathbf{I}_t')]^{-1} = E\big[\mathbf{X}_t(G_t(\mathbf{X}_t - \hat{\mathbf{X}}_t) + W_t)'\big]\Delta_t^{-1} = \Omega_t G_t'\Delta_t^{-1}. \tag{9.4.18}$$

To establish (9.4.17) we write

$$\mathbf{X}_t - P_{t-1}\mathbf{X}_t = \mathbf{X}_t - P_t\mathbf{X}_t + P_t\mathbf{X}_t - P_{t-1}\mathbf{X}_t = \mathbf{X}_t - P_t\mathbf{X}_t + M\mathbf{I}_t.$$

Using (9.4.18) and the orthogonality of $\mathbf{X}_t - P_t\mathbf{X}_t$ and $M\mathbf{I}_t$, we find from the last equation that

$$\Omega_t = \Omega_{t|t} + \Omega_t G_t' \Delta_t^{-1} G_t \Omega_t',$$

as required. ■

Kalman Fixed-Point Smoothing:
The smoothed estimates $\mathbf{X}_{t|n} = P_n\mathbf{X}_t$ and the error covariance matrices $\Omega_{t|n} = E[(\mathbf{X}_t - \mathbf{X}_{t|n})(\mathbf{X}_t - \mathbf{X}_{t|n})']$ are determined for fixed t by the following recursions, which can be solved successively for $n = t, t+1, \ldots$:

$$P_n\mathbf{X}_t = P_{n-1}\mathbf{X}_t + \Omega_{t,n}G_n'\Delta_n^{-1}\left(\mathbf{Y}_n - G_n\hat{\mathbf{X}}_n\right), \tag{9.4.19}$$

$$\Omega_{t,n+1} = \Omega_{t,n}[F_n - \Theta_n\Delta_n^{-1}G_n]', \tag{9.4.20}$$

$$\Omega_{t|n} = \Omega_{t|n-1} - \Omega_{t,n}G_n'\Delta_n^{-1}G_n\Omega_{t,n}', \tag{9.4.21}$$

with initial conditions $P_{t-1}\mathbf{X}_t = \hat{\mathbf{X}}_t$ and $\Omega_{t,t} = \Omega_{t|t-1} = \Omega_t$ (found from Kalman prediction).

Proof. Using (9.4.3) we can write $P_n\mathbf{X}_t = P_{n-1}\mathbf{X}_t + C\mathbf{I}_n$, where $\mathbf{I}_n = G_n\big(\mathbf{X}_n - \hat{\mathbf{X}}_n\big) + \mathbf{W}_n$. By Remark 4 above,

$$C = E\left[\mathbf{X}_t\left(G_n(\mathbf{X}_n - \hat{\mathbf{X}}_n) + \mathbf{W}_n\right)'\right]\left[E\left(\mathbf{I}_n\mathbf{I}_n'\right)\right]^{-1} = \Omega_{t,n}G_n'\Delta_n^{-1}, \tag{9.4.22}$$

where $\Omega_{t,n} := E\big[\big(\mathbf{X}_t - \hat{\mathbf{X}}_t\big)\big(\mathbf{X}_n - \hat{\mathbf{X}}_n\big)'\big]$. It follows now from (9.1.2), (9.4.5), the orthogonality of $\mathbf{V}_n$ and $\mathbf{W}_n$ with $\mathbf{X}_t - \hat{\mathbf{X}}_t$, and the definition of $\Omega_{t,n}$ that

$$\Omega_{t,n+1} = E\left[\left(\mathbf{X}_t - \hat{\mathbf{X}}_t\right)\left(\mathbf{X}_n - \hat{\mathbf{X}}_n\right)'\left(F_n - \Theta_n\Delta_n^{-1}G_n\right)'\right] = \Omega_{t,n}\left[F_n - \Theta_n\Delta_n^{-1}G_n\right]',$$

thus establishing (9.4.20). To establish (9.4.21) we write

$$\mathbf{X}_t - P_n\mathbf{X}_t = \mathbf{X}_t - P_{n-1}\mathbf{X}_t - C\mathbf{I}_n.$$

Using (9.4.22) and the orthogonality of $\mathbf{X}_t - P_n\mathbf{X}_t$ and $\mathbf{I}_n$, the last equation then gives

$$\Omega_{t|n} = \Omega_{t|n-1} - \Omega_{t,n}G_n'\Delta_n^{-1}G_n\Omega_{t,n}', \quad n = t, t+1, \ldots,$$

as required. ■

9.5 Estimation for State-Space Models

Consider the state-space model defined by equations (9.1.1) and (9.1.2) and suppose that the model is completely parameterized by the components of the vector $\boldsymbol{\theta}$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is found by maximizing the likelihood of the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ with respect to the components of the vector $\boldsymbol{\theta}$. If the conditional probability density of $\mathbf{Y}_t$ given $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \ldots, \mathbf{Y}_0 = \mathbf{y}_0$ is $f_t(\cdot|\mathbf{y}_{t-1}, \ldots, \mathbf{y}_0)$, then the likelihood of $\mathbf{Y}_t$, $t = 1, \ldots, n$ (conditional on $\mathbf{Y}_0$), can immediately be written as

$$L(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \prod_{t=1}^{n} f_t(\mathbf{Y}_t|\mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0). \tag{9.5.1}$$

The calculation of the likelihood for any fixed numerical value of $\boldsymbol{\theta}$ is extremely complicated in general, but is greatly simplified if $\mathbf{Y}_0, \mathbf{X}_1$ and $\mathbf{W}_t, \mathbf{V}_t, t = 1, 2, \ldots,$ are assumed to be jointly Gaussian. The resulting likelihood is called the Gaussian likelihood and is widely used in time series analysis (cf. Section 5.2) whether the time series is truly Gaussian or not. As before, we shall continue to use the term *likelihood* to mean Gaussian likelihood.

If $\mathbf{Y}_0, \mathbf{X}_1$ and $\mathbf{W}_t, \mathbf{V}_t, t = 1, 2, \ldots,$ are jointly Gaussian, then the conditional densities in (9.5.1) are given by

$$f_t(\mathbf{Y}_t|\mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_0) = (2\pi)^{-w/2} (\det \Delta_t)^{-1/2} \exp\left[-\frac{1}{2}\mathbf{I}_t' \Delta_t^{-1}\mathbf{I}_t\right],$$

where $\mathbf{I}_t = \mathbf{Y}_t - P_{t-1}\mathbf{Y}_t = \mathbf{Y}_t - G\hat{\mathbf{X}}_t$, $P_{t-1}\mathbf{Y}_t$, and Δ_t, $t \geq 1$, are the one-step predictors and error covariance matrices found from the Kalman prediction recursions. The likelihood of the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ (conditional on $\mathbf{Y}_0$) can therefore be expressed as

$$L(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = (2\pi)^{-nw/2} \left(\prod_{j=1}^{n} \det \Delta_j\right)^{-1/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n} \mathbf{I}_j' \Delta_j^{-1}\mathbf{I}_j\right]. \tag{9.5.2}$$

Given the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, the distribution of $\mathbf{Y}_0$ (see Section 9.4), and a particular parameter value $\boldsymbol{\theta}$, the numerical value of the likelihood L can be computed from the previous equation with the aid of the Kalman recursions of Section 9.4. To find maximum likelihood estimates of the components of $\boldsymbol{\theta}$, a nonlinear optimization algorithm must be used to search for the value of $\boldsymbol{\theta}$ that maximizes the value of L.

Having estimated the parameter vector $\boldsymbol{\theta}$, we can compute forecasts based on the fitted state-space model and estimated mean squared errors by direct application of equations (9.4.7) and (9.4.9).

9.5.1 Application to Structural Models

The general structural model for a univariate time series $\{Y_t\}$ of which we gave examples in Section 9.2 has the form

$$Y_t = G\mathbf{X}_t + W_t, \quad \{W_t\} \sim \mathrm{WN}\left(0, \sigma_w^2\right), \tag{9.5.3}$$

$$\mathbf{X}_{t+1} = F\mathbf{X}_t + \mathbf{V}_t, \quad \{\mathbf{V}_t\} \sim \mathrm{WN}(0, Q), \tag{9.5.4}$$

for $t = 1, 2, \ldots,$ where F and G are assumed known. We set $Y_0 = 1$ in order to include constant terms in our predictors and complete the specification of the model by prescribing the mean and covariance matrix of the initial state $\mathbf{X}_1$. A simple and convenient assumption is that $\mathbf{X}_1$ is equal to a deterministic but unknown parameter $\boldsymbol{\mu}$ and that $\hat{\mathbf{X}}_1 = \boldsymbol{\mu}$, so that $\Omega_1 = 0$. The parameters of the model are then $\boldsymbol{\mu}$, Q, and σ_w^2.

Direct maximization of the likelihood (9.5.2) is difficult if the dimension of the state vector is large. The maximization can, however, be simplified by the following stepwise procedure. For fixed Q we find $\hat{\boldsymbol{\mu}}(Q)$ and $\sigma_w^2(Q)$ that maximize the likelihood $L\left(\boldsymbol{\mu}, Q, \sigma_w^2\right)$. We then maximize the "reduced likelihood" $L\left(\hat{\boldsymbol{\mu}}(Q), Q, \hat{\sigma}_w^2(Q)\right)$ with respect to Q.

To achieve this we define the mean-corrected state vectors, $\mathbf{X}_t^* = \mathbf{X}_t - F^{t-1}\boldsymbol{\mu}$, and apply the Kalman prediction recursions to $\{\mathbf{X}_t^*\}$ with initial condition $\mathbf{X}_1^* = \mathbf{0}$. This gives, from (9.4.1),

$$\hat{\mathbf{X}}_{t+1}^* = F\hat{\mathbf{X}}_t^* + \Theta_t\Delta_t^{-1}\left(Y_t - G\hat{\mathbf{X}}_t^*\right), \quad t = 1, 2, \ldots, \tag{9.5.5}$$

with $\hat{\mathbf{X}}_1^* = \mathbf{0}$. Since $\hat{\mathbf{X}}_t$ also satisfies (9.5.5), but with initial condition $\hat{\mathbf{X}}_t = \boldsymbol{\mu}$, it follows that

$$\hat{\mathbf{X}}_t = \hat{\mathbf{X}}_t^* + C_t\boldsymbol{\mu} \tag{9.5.6}$$

for some $v \times v$ matrices C_t. (Note that although $\hat{\mathbf{X}}_t = P(\mathbf{X}_t|Y_0, Y_1, \ldots, Y_t)$, the quantity $\hat{\mathbf{X}}_t^*$ is not the corresponding predictor of $\mathbf{X}_t^*$.) The matrices C_t can be determined recursively from (9.5.5), (9.5.6), and (9.4.1). Substituting (9.5.6) into (9.5.5) and using (9.4.1), we have

$$\begin{aligned}
\hat{\mathbf{X}}_{t+1}^* &= F\left(\hat{\mathbf{X}}_t - C_t\boldsymbol{\mu}\right) + \Theta_t\Delta_t^{-1}\left(Y_t - G\left(\hat{\mathbf{X}}_t - C_t\boldsymbol{\mu}\right)\right)\\
&= F\hat{\mathbf{X}}_t + \Theta_t\Delta_t^{-1}\left(Y_t - G\hat{\mathbf{X}}_t\right) - \left(F - \Theta_t\Delta_t^{-1}G\right)C_t\boldsymbol{\mu}\\
&= \hat{\mathbf{X}}_{t+1} - \left(F - \Theta_t\Delta_t^{-1}G\right)C_t\boldsymbol{\mu},
\end{aligned}$$

so that

$$C_{t+1} = \left(F - \Theta_t\Delta_t^{-1}G\right)C_t \tag{9.5.7}$$

with C_1 equal to the identity matrix. The quadratic form in the likelihood (9.5.2) is therefore

$$S(\boldsymbol{\mu}, Q, \sigma_w^2) = \sum_{t=1}^{n}\frac{\left(Y_t - G\hat{\mathbf{X}}_t\right)^2}{\Delta_t} \tag{9.5.8}$$

$$= \sum_{t=1}^{n}\frac{\left(Y_t - G\hat{\mathbf{X}}_t^* - GC_t\boldsymbol{\mu}\right)^2}{\Delta_t}. \tag{9.5.9}$$

Now let $Q^* := \sigma_w^{-2}Q$ and define L^* to be the likelihood function with this new parameterization, i.e., $L^*\left(\boldsymbol{\mu}, Q^*, \sigma_w^2\right) = L\left(\boldsymbol{\mu}, \sigma_w^2Q^*, \sigma_w^2\right)$. Writing $\Delta_t^* = \sigma_w^{-2}\Delta_t$ and $\Omega_t^* = \sigma_w^{-2}\Omega_t$, we see that the predictors $\hat{X}_t^*$ and the matrices C_t in (9.5.7) depend on the parameters only through Q^*. Thus,

$$S\left(\boldsymbol{\mu}, Q, \sigma_w^2\right) = \sigma_w^{-2}S\left(\boldsymbol{\mu}, Q^*, 1\right),$$

so that

$$\begin{aligned}
-2\ln L^*\left(\boldsymbol{\mu}, Q^*, \sigma_w^2\right) &= n\ln(2\pi) + \sum_{t=1}^{n}\ln\Delta_t + \sigma_w^{-2}S\left(\boldsymbol{\mu}, Q^*, 1\right)\\
&= n\ln(2\pi) + \sum_{t=1}^{n}\ln\Delta_t^* + n\ln\sigma_w^2 + \sigma_w^{-2}S\left(\boldsymbol{\mu}, Q^*, 1\right).
\end{aligned}$$

For Q^* fixed, it is easy to show (see Problem 9.18) that this function is minimized when

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}\left(Q^*\right) = \left[\sum_{t=1}^{n}\frac{C_t'G'GC_t}{\Delta_t^*}\right]^{-1}\sum_{t=1}^{n}\frac{C_t'G'\left(Y_t - G\hat{\mathbf{X}}_t^*\right)}{\Delta_t^*} \tag{9.5.10}$$

and

$$\hat{\sigma}_w^2 = \hat{\sigma}_w^2\left(Q^*\right) = n^{-1}\sum_{t=1}^{n}\frac{\left(Y_t - G\hat{\mathbf{X}}_t^* - GC_t\hat{\boldsymbol{\mu}}\right)^2}{\Delta_t^*}. \tag{9.5.11}$$

Replacing $\boldsymbol{\mu}$ and σ_w^2 by these values in $-2\ln L^*$ and ignoring constants, the reduced likelihood becomes

$$\ell\left(Q^*\right) = \ln\left(n^{-1}\sum_{t=1}^{n}\frac{\left(Y_t - G\hat{\mathbf{X}}_t^* - GC_t\hat{\boldsymbol{\mu}}\right)^2}{\Delta_t^*}\right) + n^{-1}\sum_{t=1}^{n}\ln\left(\det\Delta_t^*\right). \tag{9.5.12}$$

If $\hat{Q}^*$ denotes the minimizer of (9.5.12), then the maximum likelihood estimator of the parameters $\boldsymbol{\mu}$, Q, σ_w^2 are $\hat{\boldsymbol{\mu}}$, $\hat{\sigma}_w^2\hat{Q}^*$, $\hat{\sigma}_w^2$, where $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}_w^2$ are computed from (9.5.10) and (9.5.11) with Q^* replaced by $\hat{Q}^*$.

We can now summarize the steps required for computing the maximum likelihood estimators of $\boldsymbol{\mu}$, Q, and σ_w^2 for the model (9.5.3)–(9.5.4).

1. For a fixed Q^*, apply the Kalman prediction recursions with $\hat{\mathbf{X}}_1^* = \mathbf{0}$, $\Omega_1 = 0$, $Q = Q^*$, and $\sigma_w^2 = 1$ to obtain the *predictors* $\hat{\mathbf{X}}_t^*$. Let Δ_t^* denote the one-step prediction error produced by these recursions.
2. Set $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(Q^*) = \left[\sum_{t=1}^{n} C_t'G'GC_t/\Delta_t\right]^{-1}\sum_{t=1}^{n} C_t'G'(Y_t - G\hat{\mathbf{X}}_t^*)/\Delta_t^*$.
3. Let $\hat{Q}^*$ be the minimizer of (9.5.12).
4. The maximum likelihood estimators of $\boldsymbol{\mu}$, Q, and σ_w^2 are then given by $\hat{\boldsymbol{\mu}}$, $\hat{\sigma}_w^2\hat{Q}^*$, and $\hat{\sigma}_w^2$, respectively, where $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}_w^2$ are found from (9.5.10) and (9.5.11) evaluated at $\hat{Q}^*$.

Example 9.5.1. Random Walk Plus Noise Model

In Example 9.2.1, 100 observations were generated from the structural model

$$\begin{aligned} Y_t &= M_t + W_t, \quad \{W_t\} \sim \mathrm{WN}\left(0, \sigma_w^2\right), \\ M_{t+1} &= M_t + V_t, \quad \{V_t\} \sim \mathrm{WN}\left(0, \sigma_v^2\right), \end{aligned}$$

with initial values $\mu = M_1 = 0$, $\sigma_w^2 = 8$, and $\sigma_v^2 = 4$. The maximum likelihood estimates of the parameters are found by first minimizing (9.5.12) with $\hat{\mu}$ given by (9.5.10). Substituting these values into (9.5.11) gives $\hat{\sigma}_w^2$. The resulting estimates are $\hat{\mu} = 0.906$, $\hat{\sigma}_v^2 = 5.351$, and $\hat{\sigma}_w^2 = 8.233$, which are in reasonably close agreement with the true values.

□

Example 9.5.2. International Airline Passengers, 1949–1960; AIRPASS.TSM

The monthly totals of international airline passengers from January 1949 to December 1960 (Box and Jenkins 1976) are displayed in Figure 9-3. The data exhibit both a strong seasonal pattern and a nearly linear trend. Since the variability of the data $Y_1, \ldots, Y_{144}$ increases for larger values of Y_t, it may be appropriate to consider a logarithmic transformation of the data. For the purpose of this illustration, however, we will fit a structural model incorporating a randomly varying trend and seasonal and noise components (see Example 9.2.3) to the raw data. This model has the form

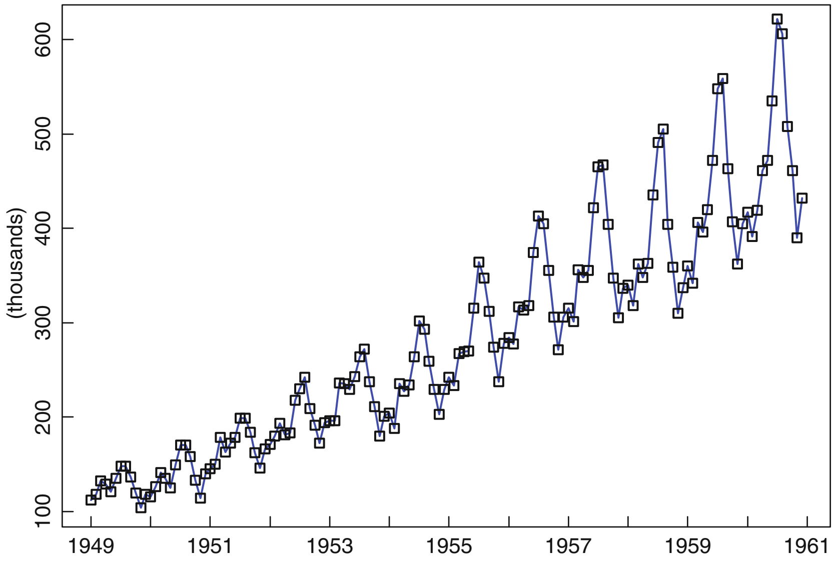

Figure 9-3 International airline passengers; monthly totals from January 1949 to December 1960

$$Y_t = G\mathbf{X}_t + W_t, \quad \{W_t\} \sim \mathrm{WN}\left(0,\ \sigma_w^2\right),$$

$$\mathbf{X}_{t+1} = F\mathbf{X}_t + \mathbf{V}_t, \quad \{\mathbf{V}_t\} \sim \mathrm{WN}(0,\ Q),$$

where $\mathbf{X}_t$ is a 13-dimensional state-vector,

$$F = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} 1\ 0\ 1\ 0\ \cdots\ 0 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The parameters of the model are $\boldsymbol{\mu}, \sigma_1^2, \sigma_2^2, \sigma_3^2$, and σ_w^2, where $\boldsymbol{\mu} = \mathbf{X}_1$. Minimizing (9.5.12) with respect to Q^* we find from (9.5.11) and (9.5.12) that

$$\left(\hat\sigma_1^2, \hat\sigma_2^2, \hat\sigma_3^2, \hat\sigma_w^2\right) = (170.63,\ .00000,\ 11.338,\ .014179)$$

and from (9.5.10) that $\hat{\boldsymbol{\mu}} = (146.9,\ 2.171,\ -34.92,\ -34.12,\ -47.00,\ -16.98,\ 22.99,\ 53.99,\ 58.34,\ 33.65,\ 2.204,\ -4.053,\ -6.894)'$. The first component, X_{t1}, of the state vector corresponds to the local linear trend with slope X_{t2}. Since $\hat\sigma_2^2 = 0$, the slope at time t, which satisfies

$$X_{t2} = X_{t-1,2} + V_{t2},$$

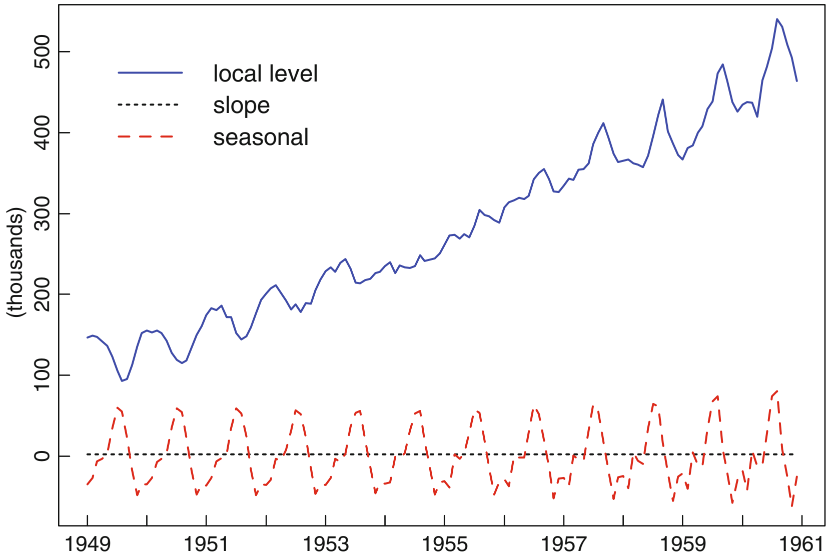

Figure 9-4 The one-step predictors $(\hat{X}_{t1}, \hat{X}_{t2}, \hat{X}_{t3})'$ for the airline passenger data in Example 9.5.2

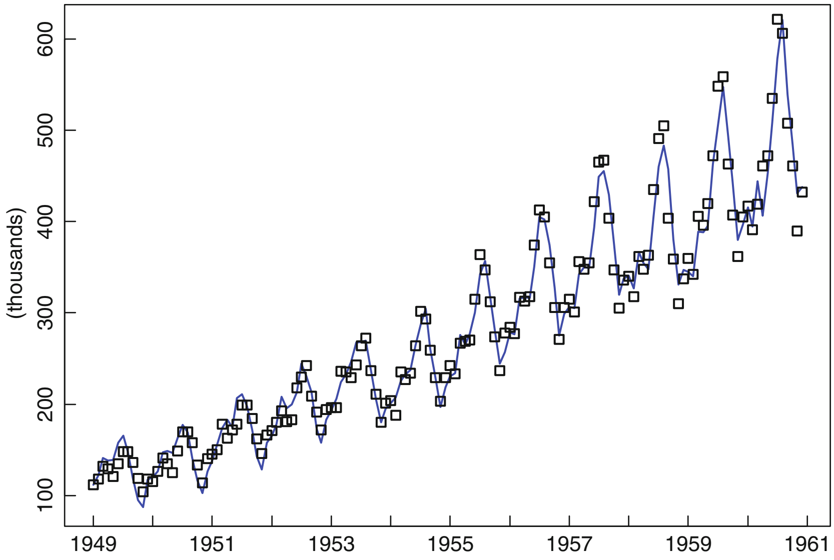

Figure 9-5 The one-step predictors $\hat{Y}_t$ for the airline passenger data (*solid line*) and the actual data (*square boxes*)

must be nearly constant and equal to $\hat{X}_{12} = 2.171$. The first three components of the predictors $\hat{\mathbf{X}}_t$ are plotted in Figure 9-4. Notice that the first component varies like a random walk around a straight line, while the second component is nearly constant as a result of $\hat{\sigma}_2^2 \approx 0$. The third component, corresponding to the seasonal component, exhibits a clear seasonal cycle that repeats roughly the same pattern throughout the 12 years of data. The one-step predictors $\hat{X}_{t1} + \hat{X}_{t3}$ of Y_t are plotted in Figure 9-5 (solid line) together with the actual data (square boxes). For this model the predictors follow the movement of the data quite well.

□

9.6 State-Space Models with Missing Observations

State-space representations and the associated Kalman recursions are ideally suited to the analysis of data with missing values, as was pointed out by Jones (1980) in the context of maximum likelihood estimation for ARMA processes. In this section we shall deal with two missing-value problems for state-space models. The first is the evaluation of the (Gaussian) likelihood based on $\{\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\}$, where $i_1, i_2, \ldots, i_r$ are positive integers such that $1 \le i_1 < i_2 < \cdots < i_r \le n$. (This allows for observation of the process $\{\mathbf{Y}_t\}$ at irregular intervals, or equivalently for the possibility that $(n-r)$ observations are missing from the sequence $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$.) The solution of this problem will, in particular, enable us to carry out maximum likelihood estimation for ARMA and ARIMA processes with missing values. The second problem to be considered is the minimum mean squared error estimation of the missing values themselves.

9.6.1 The Gaussian Likelihood of $\{\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\}$, $1 \le i_1 < i_2 < \cdots < i_r \le n$

Consider the state-space model defined by equations (9.1.1) and (9.1.2) and suppose that the model is completely parameterized by the components of the vector $\boldsymbol{\theta}$. If there are no missing observations, i.e., if $r = n$ and $i_j = j, j = 1, \ldots, n$, then the likelihood of the observations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ is easily found as in Section 9.5 to be

$$L(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = (2\pi)^{-nw/2} \left(\prod_{j=1}^{n} \det \Delta_j \right)^{-1/2} \exp\left[-\frac{1}{2} \sum_{j=1}^{n} \mathbf{I}_j' \Delta_j^{-1} \mathbf{I}_j \right],$$

where $\mathbf{I}_j = \mathbf{Y}_j - P_{j-1}\mathbf{Y}_j$ and Δ_j, $j \ge 1$, are the one-step predictors and error covariance matrices found from (9.4.7) and (9.4.9) with $\mathbf{Y}_0 = \mathbf{1}$.

To deal with the more general case of possibly irregularly spaced observations $\{\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\}$, we introduce a new series $\{\mathbf{Y}_t^*\}$, related to the process $\{\mathbf{X}_t\}$ by the modified observation equation

$$\mathbf{Y}_t^* = G_t^* \mathbf{X}_t + \mathbf{W}_t^*, \quad t = 1, 2, \ldots, \tag{9.6.1}$$

where

$$G_t^* = \begin{cases} G_t & \text{if } t \in \{i_1, \ldots, i_r\}, \\ 0 & \text{otherwise,} \end{cases} \qquad \mathbf{W}_t^* = \begin{cases} \mathbf{W}_t & \text{if } t \in \{i_1, \ldots, i_r\}, \\ \mathbf{N}_t & \text{otherwise,} \end{cases} \tag{9.6.2}$$

and $\{\mathbf{N}_t\}$ is iid with

$$\mathbf{N}_t \sim \mathrm{N}(\mathbf{0}, I_{w \times w}), \quad \mathbf{N}_s \perp \mathbf{X}_1, \quad \mathbf{N}_s \perp \begin{bmatrix} \mathbf{V}_t \\ \mathbf{W}_t \end{bmatrix}, \quad s, t = 0, \pm 1, \ldots. \tag{9.6.3}$$

Equations (9.6.1) and (9.1.2) constitute a state-space representation for the new series $\{\mathbf{Y}_t^*\}$, which coincides with $\{\mathbf{Y}_t\}$ at each $t \in \{i_1, i_2, \ldots, i_r\}$, and at other times takes random values that are independent of $\{\mathbf{Y}_t\}$ with a distribution independent of $\boldsymbol{\theta}$.

Let $L_1\left(\boldsymbol{\theta}; \mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_r}\right)$ be the Gaussian likelihood based on the observed values $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_r}$ of $\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}$ under the model defined by (9.1.1) and (9.1.2). Corresponding to these observed values, we define a new sequence, $\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*$, by

$$\mathbf{y}_t^* = \begin{cases} \mathbf{y}_t & \text{if } t \in \{i_1, \ldots, i_r\}, \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{9.6.4}$$

Then it is clear from the preceding paragraph that

$$L_1\left(\boldsymbol{\theta}; \mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_r}\right) = (2\pi)^{(n-r)w/2} L_2\left(\boldsymbol{\theta}; \mathbf{y}_1^*, \ldots, \mathbf{y}_n^*\right), \tag{9.6.5}$$

where L_2 denotes the Gaussian likelihood under the model defined by (9.6.1) and (9.1.2).

In view of (9.6.5) we can now compute the required likelihood L_1 of the realized values $\{\mathbf{y}_t, t = i_1, \ldots, i_r\}$ as follows:

i. Define the sequence $\{\mathbf{y}_t^*, t = 1, \ldots, n\}$ as in (9.6.4).

ii. Find the one-step predictors $\hat{\mathbf{Y}}_t^*$ of $\mathbf{Y}_t^*$, and their error covariance matrices Δ_t^*, using Kalman prediction and equations (9.4.7) and (9.4.9) applied to the state-space representation (9.6.1) and (9.1.2) of $\{\mathbf{Y}_t^*\}$. Denote the realized values of the predictors, based on the observation sequence $\left\{\mathbf{y}_t^*\right\}$, by $\left\{\hat{\mathbf{y}}_t^*\right\}$.

iii. The required Gaussian likelihood of the irregularly spaced observations $\{\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_r}\}$ is then, by (9.6.5),

$$L_1(\boldsymbol{\theta}; \mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_r}) = (2\pi)^{-rw/2} \left(\prod_{j=1}^{n} \det \Delta_j^* \right)^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \mathbf{i}_j^{*\prime} \Delta_j^{*-1} \mathbf{i}_j^* \right\},$$

where $\mathbf{i}_j^*$ denotes the observed innovation $\mathbf{y}_j^* - \hat{\mathbf{y}}_j^*, j = 1, \ldots, n$.

Example 9.6.1. An AR(1) Series with One Missing Observation

Let $\{Y_t\}$ be the causal AR(1) process defined by

$$Y_t - \phi Y_{t-1} = Z_t, \qquad \{Z_t\} \sim \mathrm{WN}\left(0, \sigma^2\right).$$

To find the Gaussian likelihood of the observations y_1, y_3, y_4, and y_5 of Y_1, Y_3, Y_4, and Y_5 we follow the steps outlined above.

i. Set $y_i^* = y_i,\ i = 1, 3, 4, 5$ and $y_2^* = 0$.

ii. We start with the state-space model for $\{Y_t\}$ from Example 9.1.1, i.e., $Y_t = X_t,\ X_{t+1} = \phi X_t + Z_{t+1}$. The corresponding model for $\{Y_t^*\}$ is then, from (9.6.1),

$$Y_t^* = G_t^* X_t + W_t^*,\ t = 1, 2, \ldots,$$

where

$$X_{t+1} = F_t X_t + V_t,\ t = 1, 2, \ldots,$$

$$F_t = \phi, \quad G_t^* = \begin{cases} 1 & \text{if } t \neq 2, \\ 0 & \text{if } t = 2, \end{cases} \quad V_t = Z_{t+1}, \quad W_t^* = \begin{cases} 0 & \text{if } t \neq 2, \\ N_t & \text{if } t = 2, \end{cases}$$

$$Q_t = \sigma^2, \qquad R_t^* = \begin{cases} 0 & \text{if } t \neq 2, \\ 1 & \text{if } t = 2, \end{cases} \qquad S_t^* = 0,$$

and $X_1 = \sum_{j=0}^{\infty} \phi^j Z_{1-j}$. Starting from the initial conditions

$$\hat{X}_1 = 0, \qquad \Omega_1 = \sigma^2 / \left(1 - \phi^2\right),$$

and applying the recursions (9.4.1) and (9.4.2), we find (Problem 9.19) that

$$\Theta_t \Delta_t^{-1} = \begin{cases} \phi & \text{if } t = 1, 3, 4, 5, \\ 0 & \text{if } t = 2, \end{cases} \qquad \Omega_t = \begin{cases} \sigma^2 / \left(1 - \phi^2\right) & \text{if } t = 1, \\ \sigma^2 \left(1 + \phi^2\right) & \text{if } t = 3, \\ \sigma^2 & \text{if } t = 2, 4, 5, \end{cases}$$

and

$$\hat{X}_1 = 0, \quad \hat{X}_2 = \phi Y_1, \quad \hat{X}_3 = \phi^2 Y_1, \quad \hat{X}_4 = \phi Y_3, \quad \hat{X}_5 = \phi Y_4.$$

From (9.4.7) and (9.4.9) with $h = 1$, we find that

$$\hat{Y}_1^* = 0, \quad \hat{Y}_2^* = 0, \quad \hat{Y}_3^* = \phi^2 Y_1, \quad \hat{Y}_4^* = \phi Y_3, \quad \hat{Y}_5^* = \phi Y_4,$$

with corresponding mean squared errors

$$\Delta_1^* = \sigma^2 / \left(1 - \phi^2\right), \quad \Delta_2^* = 1, \quad \Delta_3^* = \sigma^2 \left(1 + \phi^2\right), \quad \Delta_4^* = \sigma^2, \quad \Delta_5^* = \sigma^2.$$

iii. From the preceding calculations we can now write the likelihood of the original data as

$$L_1(\phi, \sigma^2;\, y_1, y_3, y_4, y_5) = \sigma^{-4}(2\pi)^{-2}\left[\left(1-\phi^2\right) / \left(1+\phi^2\right)\right]^{1/2}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}\left[y_1^2\left(1-\phi^2\right) + \frac{(y_3-\phi^2 y_1)^2}{1+\phi^2} + (y_4-\phi y_3)^2 + (y_5-\phi y_4)^2\right]\right\}.$$

□

Remark 1. If we are given observations $y_{1-d}, y_{2-d}, \ldots, y_0, y_{i_1}, y_{i_2}, \ldots, y_{i_r}$ of an ARIMA(p, d, q) process at times $1 - d, 2 - d, \ldots, 0, i_1, \ldots, i_r$, where $1 \le i_1 < i_2 < \cdots < i_r \le n$, a similar argument can be used to find the Gaussian likelihood of $y_{i_1}, \ldots, y_{i_r}$ *conditional on* $Y_{1-d} = y_{1-d}, Y_{2-d} = y_{2-d}, \ldots, Y_0 = y_0$. Missing values among the first d observations $y_{1-d}, y_{2-d}, \ldots, y_0$ can be handled by treating them as unknown parameters for likelihood maximization. For more on ARIMA series with missing values see Brockwell and Davis (1991) and Ansley and Kohn (1985). □

9.6.2 Estimation of Missing Values for State-Space Models

Given that we observe only $\mathbf{Y}_{i_1}, \mathbf{Y}_{i_2}, \ldots, \mathbf{Y}_{i_r}$, $1 \le i_1 < i_2 < \cdots < i_r \le n$, where $\{\mathbf{Y}_t\}$ has the state-space representation (9.1.1) and (9.1.2), we now consider the problem of finding the minimum mean squared error estimators $P\left(\mathbf{Y}_t | \mathbf{Y}_0, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\right)$ of $\mathbf{Y}_t$, $1 \le t \le n$, where $\mathbf{Y}_0 = \mathbf{1}$. To handle this problem we again use the modified process $\{\mathbf{Y}_t^*\}$ defined by (9.6.1) and (9.1.2) with $\mathbf{Y}_0^* = \mathbf{1}$. Since $\mathbf{Y}_s^* = \mathbf{Y}_s$ for $s \in \{i_1, \ldots, i_r\}$ and $\mathbf{Y}_s^* \perp \mathbf{X}_t, \mathbf{Y}_0$ for $1 \le t \le n$ and $s \notin \{0, i_1, \ldots, i_r\}$, we immediately obtain the minimum mean squared error state estimators

$$P\left(\mathbf{X}_t | \mathbf{Y}_0, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\right) = P\left(\mathbf{X}_t | \mathbf{Y}_0^*, \mathbf{Y}_1^*, \ldots, \mathbf{Y}_n^*\right), \quad 1 \le t \le n. \tag{9.6.6}$$

The right-hand side can be evaluated by application of the Kalman fixed-point smoothing algorithm to the state-space model (9.6.1) and (9.1.2). For computational purposes the observed values of $\mathbf{Y}_t^*$, $t \notin \{0, i_1, \ldots, i_r\}$, are quite immaterial. They may, for example, all be set equal to zero, giving the sequence of *observations* of $\mathbf{Y}_t^*$ defined in (9.6.4).

To evaluate $P\left(\mathbf{Y}_t | \mathbf{Y}_0, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\right)$, $1 \le t \le n$, we use (9.6.6) and the relation

$$\mathbf{Y}_t = G_t \mathbf{X}_t + \mathbf{W}_t. \tag{9.6.7}$$

Since $E\left(\mathbf{V}_t \mathbf{W}_t'\right) = S_t = 0, \quad t = 1, \ldots, n$, we find from (9.6.7) that

$$P\left(\mathbf{Y}_t | \mathbf{Y}_0, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_r}\right) = G_t P\left(\mathbf{X}_t | \mathbf{Y}_0^*, \mathbf{Y}_1^*, \ldots, \mathbf{Y}_n^*\right). \tag{9.6.8}$$

Example 9.6.2. An AR(1) Series with One Missing Observation

Consider the problem of estimating the missing value Y_2 in Example 9.6.1 in terms of $Y_0 = 1, Y_1, Y_3, Y_4$, and Y_5. We start from the state-space model $X_{t+1} = \phi X_t + Z_{t+1}$, $Y_t = X_t$, for $\{Y_t\}$. The corresponding model for $\{Y_t^*\}$ is the one used in Example 9.6.1. Applying the Kalman smoothing equations to the latter model, we find that

$$\begin{aligned}
&P_1X_2 = \phi Y_1, \quad P_2X_2 = \phi Y_1, \quad P_3X_2 = \frac{\phi(Y_1+Y_3)}{(1+\phi^2)},\\
&P_4X_2 = P_3X_2, \quad P_5X_2 = P_3X_2,\\
&\Omega_{2,2} = \sigma^2, \quad \Omega_{2,3} = \phi\sigma^2, \quad \Omega_{2,t} = 0, \quad t \ge 4,
\end{aligned}$$

and

$$\Omega_{2|1} = \sigma^2, \quad \Omega_{2|2} = \sigma^2, \quad \Omega_{2|t} = \frac{\sigma^2}{(1+\phi^2)}, \quad t \ge 3,$$

where $P_t(\cdot)$ here denotes $P\left(\cdot|Y_0^*, \ldots, Y_t^*\right)$ and $\Omega_{t,n}$, $\Omega_{t|n}$ are defined correspondingly. We deduce from (9.6.8) that the minimum mean squared error estimator of the missing value Y_2 is

$$P_5Y_2 = P_5X_2 = \frac{\phi(Y_1+Y_3)}{\left(1+\phi^2\right)},$$

with mean squared error

$$\Omega_{2|5} = \frac{\sigma^2}{\left(1+\phi^2\right)}.$$ □

Remark 2. Suppose we have observations $Y_{1-d}, Y_{2-d}, \ldots, Y_0, Y_{i_1}, \ldots, Y_{i_r}$ $(1 \le i_1 < i_2 \cdots < i_r \le n)$ of an ARIMA(p,d,q) process. Determination of the best linear estimates of the missing values Y_t, $t \notin \{i_1, \ldots, i_r\}$, in terms of Y_t, $t \in \{i_1, \ldots, i_r\}$, and the components of $\mathbf{Y}_0 := (Y_{1-d}, Y_{2-d}, \ldots, Y_0)'$ can be carried out as in Example 9.6.2 using the state-space representation of the ARIMA series $\{Y_t\}$ from Example 9.3.3 and the Kalman recursions for the corresponding state-space model for $\{Y_t^*\}$ defined by (9.6.1) and (9.1.2). See Brockwell and Davis (1991) for further details. □

We close this section with a brief discussion of a direct approach to estimating missing observations. This approach is often more efficient than the methods just described, especially if the number of missing observations is small and we have a simple (e.g., autoregressive) model. Consider the general problem of computing $E(\mathbf{X}|\mathbf{Y})$ when the random vector $(\mathbf{X}', \mathbf{Y}')'$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. (In the missing observation problem, think of $\mathbf{X}$ as the vector of the missing observations and $\mathbf{Y}$ as the vector of observed values.) Then the joint probability density function of $\mathbf{X}$ and $\mathbf{Y}$ can be written as

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}), \tag{9.6.9}$$

where $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ is a multivariate normal density with mean $E(\mathbf{X}|\mathbf{Y})$ and covariance matrix $\Sigma_{\mathbf{X}|\mathbf{Y}}$ (see Proposition A.3.1). In particular,

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^q \det \Sigma_{\mathbf{X}|\mathbf{Y}}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - E(\mathbf{X}|\mathbf{y}))'\Sigma_{\mathbf{X}|\mathbf{Y}}^{-1}(\mathbf{x} - E(\mathbf{X}|\mathbf{y}))\right\}, \tag{9.6.10}$$

where $q = \dim(\mathbf{X})$. It is clear from (9.6.10) that $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ (and also $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$) is maximum when $\mathbf{x} = E(\mathbf{X}|\mathbf{y})$. Thus, the best estimator of $\mathbf{X}$ in terms of $\mathbf{Y}$ can be found by maximizing the joint density of $\mathbf{X}$ and $\mathbf{Y}$ with respect to $\mathbf{x}$. For autoregressive processes it is relatively straightforward to carry out this optimization, as shown in the following example.

Example 9.6.3. Estimating Missing Observations in an AR Process

Suppose $\{Y_t\}$ is the AR(p) process defined by

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}\left(0, \sigma^2\right),$$

and $\mathbf{Y} = (Y_{i_1}, \ldots, Y_{i_r})'$, with $1 \le i_1 < \cdots < i_r \le n$, are the observed values. If there are no missing observations in the first p observations, then the best estimates of the missing values are found by minimizing

$$\sum_{t=p+1}^{n} (Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p})^2 \tag{9.6.11}$$

with respect to the missing values (see Problem 9.20). For the AR(1) model in Example 9.6.2, minimization of (9.6.11) is equivalent to minimizing

$$(Y_2 - \phi Y_1)^2 + (Y_3 - \phi Y_2)^2$$

with respect to Y_2. Setting the derivative of this expression with respect to Y_2 equal to 0 and solving for Y_2 we obtain $E(Y_2|Y_1, Y_3, Y_4, Y_5) = \phi(Y_1 + Y_3)/\left(1 + \phi^2\right)$.

□

9.7 The EM Algorithm

The expectation-maximization (EM) algorithm is an iterative procedure for computing the maximum likelihood estimator when only a subset of the complete data set is available. Dempster et al. (1977) demonstrated the wide applicability of the EM algorithm and are largely responsible for popularizing this method in statistics. Details regarding the convergence and performance of the EM algorithm can be found in Wu (1983).

In the usual formulation of the EM algorithm, the "complete" data vector $\mathbf{W}$ is made up of "observed" data $\mathbf{Y}$ (sometimes called incomplete data) and "unobserved" data $\mathbf{X}$. In many applications, $\mathbf{X}$ consists of values of a "latent" or unobserved process occurring in the specification of the model. For example, in the state-space model of Section 9.1, $\mathbf{Y}$ could consist of the observed vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ and $\mathbf{X}$ of the unobserved state vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The EM algorithm provides an iterative procedure for computing the maximum likelihood estimator based only on the observed data $\mathbf{Y}$. Each iteration of the EM algorithm consists of two steps. If $\theta^{(i)}$ denotes the estimated value of the parameter θ after i iterations, then the two steps in the $(i+1)$th iteration are

E-step. Calculate $Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}\left[\ell(\theta; \mathbf{X}, \mathbf{Y})|\mathbf{Y}\right]$

and

M-step. Maximize $Q(\theta|\theta^{(i)})$ with respect to θ.

Then $\theta^{(i+1)}$ is set equal to the maximizer of Q in the M-step. In the E-step, $\ell(\theta; \mathbf{x}, \mathbf{y}) = \ln f(\mathbf{x}, \mathbf{y}; \theta)$, and $E_{\theta^{(i)}}(\cdot|\mathbf{Y})$ denotes the conditional expectation relative to the conditional density $f\left(\mathbf{x}|\mathbf{y}; \theta^{(i)}\right) = f\left(\mathbf{x}, \mathbf{y}; \theta^{(i)}\right)/f\left(\mathbf{y}; \theta^{(i)}\right)$.

It can be shown that $\ell\left(\theta^{(i)}; \mathbf{Y}\right)$ is nondecreasing in i, and a simple heuristic argument shows that if $\theta^{(i)}$ has a limit $\hat{\theta}$ then $\hat{\theta}$ must be a solution of the likelihood equations $\ell'\left(\hat{\theta}; \mathbf{Y}\right) = 0$. To see this, observe that $\ln f(\mathbf{x}, \mathbf{y}; \theta) = \ln f(\mathbf{x}|\mathbf{y}; \theta) + \ell(\theta; \mathbf{y})$, from which we obtain

$$Q\left(\theta|\theta^{(i)}\right) = \int (\ln f(\mathbf{x}|\mathbf{Y}; \theta)) f\left(\mathbf{x}|\mathbf{Y}; \theta^{(i)}\right) d\mathbf{x} + \ell(\theta; \mathbf{Y})$$

and

$$Q'(\theta|\theta^{(i)}) = \int \left[\frac{\partial}{\partial\theta} f(\mathbf{x}|\mathbf{Y};\theta)\right] / f(\mathbf{x}|\mathbf{Y};\theta) f\left(\mathbf{x}|\mathbf{Y};\theta^{(i)}\right) d\mathbf{x} + \ell'(\theta;\mathbf{Y}).$$

Now replacing θ with $\theta^{(i+1)}$, noticing that $Q'(\theta^{(i+1)}|\theta^{(i)}) = 0$, and letting $i \to \infty$, we find that

$$0 = \int \frac{\partial}{\partial\theta}\left[f(\mathbf{x}|\mathbf{Y};\theta)\right]_{\theta=\hat\theta} d\mathbf{x} + \ell'\left(\hat\theta;\mathbf{Y}\right) = \ell'\left(\hat\theta;\mathbf{Y}\right).$$

The last equality follows from the fact that

$$0 = \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta}\left[\int (f(\mathbf{x}|\mathbf{Y};\theta)\, d\mathbf{x}\right]_{\theta=\hat\theta} = \int \left[\frac{\partial}{\partial\theta} f(\mathbf{x}|\mathbf{Y};\theta)\right]_{\theta=\hat\theta} d\mathbf{x}.$$

The computational advantage of the EM algorithm over direct maximization of the likelihood is most pronounced when the calculation and maximization of the exact likelihood is difficult as compared with the maximization of Q in the M-step. (There are some applications in which the maximization of Q can easily be carried out explicitly.)

9.7.1 Missing Data

The EM algorithm is particularly useful for estimation problems in which there are missing observations. Suppose the complete data set consists of $Y_1, \ldots, Y_n$ of which r are observed and $n-r$ are missing. Denote the observed and missing data by $\mathbf{Y} = (Y_{i_1}, \ldots, Y_{i_r})'$ and $\mathbf{X} = (Y_{j_1}, \ldots, Y_{j_{n-r}})'$, respectively. Assuming that $\mathbf{W} = (\mathbf{X}', \mathbf{Y}')'$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ, which depends on the parameter $\boldsymbol\theta$, the log-likelihood of the complete data is given by

$$\ell(\boldsymbol\theta;\mathbf{W}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\Sigma) - \frac{1}{2}\mathbf{W}'\Sigma\mathbf{W}.$$

The E-step requires that we compute the expectation of $\ell(\boldsymbol\theta;\mathbf{W})$ with respect to the conditional distribution of $\mathbf{W}$ given $\mathbf{Y}$ with $\boldsymbol\theta = \boldsymbol\theta^{(i)}$. Writing $\Sigma(\boldsymbol\theta)$ as the block matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

which is conformable with $\mathbf{X}$ and $\mathbf{Y}$, the conditional distribution of $\mathbf{W}$ given $\mathbf{Y}$ is multivariate normal with mean $\left[\begin{smallmatrix}\hat{\mathbf{X}}\\ \mathbf{Y}\end{smallmatrix}\right]$ and covariance matrix $\left[\begin{smallmatrix}\Sigma_{11|2}(\boldsymbol\theta) & 0\\ 0 & 0\end{smallmatrix}\right]$, where $\hat{\mathbf{X}} = E_{\boldsymbol\theta}(\mathbf{X}|\mathbf{Y}) = \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}$ and $\Sigma_{11|2}(\boldsymbol\theta) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (see Proposition A.3.1). Using Problem A.8, we have

$$E_{\boldsymbol\theta^{(i)}}\left[(\mathbf{X}',\mathbf{Y}')\Sigma^{-1}(\boldsymbol\theta)(\mathbf{X}',\mathbf{Y}')'|\mathbf{Y}\right] = \text{trace}\left(\Sigma_{11|2}(\boldsymbol\theta^{(i)})\Sigma_{11|2}^{-1}(\boldsymbol\theta)\right) + \hat{\mathbf{W}}'\Sigma^{-1}(\boldsymbol\theta)\hat{\mathbf{W}},$$

where $\hat{\mathbf{W}} = \left(\hat{\mathbf{X}}', \mathbf{Y}'\right)'$. It follows that

$$Q\left(\boldsymbol\theta|\boldsymbol\theta^{(i)}\right) = \ell\left(\boldsymbol\theta, \hat{\mathbf{W}}\right) - \frac{1}{2}\text{trace}\left(\Sigma_{11|2}\left(\boldsymbol\theta^{(i)}\right)\Sigma_{11|2}^{-1}(\boldsymbol\theta)\right).$$

The first term on the right is the log-likelihood based on the complete data, but with $\mathbf{X}$ replaced by its "best estimate" $\hat{\mathbf{X}}$ calculated from the previous iteration. If the increments $\boldsymbol\theta^{(i+1)} - \boldsymbol\theta^{(i)}$ are small, then the second term on the right is nearly constant ($\approx n-r$) and can be ignored. For ease of computation in this application we shall use the modified version

$$\tilde{Q}\left(\boldsymbol\theta|\boldsymbol\theta^{(i)}\right) = \ell\left(\boldsymbol\theta;\hat{\mathbf{W}}\right).$$

With this adjustment, the steps in the EM algorithm are as follows:

E-step. Calculate $E_{\boldsymbol{\theta}^{(i)}}(\mathbf{X}|\mathbf{Y})$ (e.g., with the Kalman fixed-point smoother) and form $\ell\big(\boldsymbol{\theta}; \hat{\mathbf{W}}\big)$.

M-step. Find the maximum likelihood estimator for the "complete" data problem, i.e., maximize $\ell\big(\boldsymbol{\theta} : \hat{\mathbf{W}}\big)$. For ARMA processes, ITSM can be used directly, with the missing values replaced with their best estimates computed in the E-step.

Example 9.7.1. The Lake Data

It was found in Example 5.2.5 that the AR(2) model

$$W_t - 1.0415W_{t-1} + 0.2494W_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, .4790)$$

was a good fit to the mean-corrected lake data $\{W_t\}$. To illustrate the use of the EM algorithm for missing data, consider fitting an AR(2) model to the mean-corrected data assuming that there are 10 missing values at times $t = 17, 24, 31, 38, 45, 52, 59, 66, 73$, and 80. We start the algorithm at iteration 0 with $\hat{\phi}_1^{(0)} = \hat{\phi}_2^{(0)} = 0$. Since this initial model represents white noise, the first E-step gives, in the notation used above, $\hat{W}_{17} = \cdots = \hat{W}_{80} = 0$. Replacing the "missing" values of the mean-corrected lake data with 0 and fitting a mean-zero AR(2) model to the resulting complete data set using the maximum likelihood option in ITSM, we find that $\hat{\phi}_1^{(1)} = 0.7252$, $\hat{\phi}_2^{(1)} = 0.0236$. (Examination of the plots of the ACF and PACF of this new data set suggests an AR(1) as a better model. This is also borne out by the small estimated value of ϕ_2.) The updated missing values at times $t = 17, 24, \ldots, 80$ are found (see Section 9.6 and Problem 9.21) by minimizing

$$\sum_{j=0}^{2}\left(W_{t+j} - \hat{\phi}_1^{(1)}W_{t+j-1} - \hat{\phi}_2^{(1)}W_{t+j-2}\right)^2$$

with respect to W_t. The solution is given by

$$\hat{W}_t = \frac{\hat{\phi}_2^{(1)}(W_{t-2} + W_{t+2}) + \left(\hat{\phi}_1^{(1)} - \hat{\phi}_1^{(1)}\hat{\phi}_2^{(1)}\right)(W_{t-1} + W_{t+1})}{1 + \left(\hat{\phi}_1^{(1)}\right)^2 + \left(\hat{\phi}_2^{(1)}\right)^2}.$$

The M-step of iteration 1 is then carried out by fitting an AR(2) model using ITSM applied to the updated data set. As seen in the summary of the results reported in Table 9.1, the EM algorithm converges in four iterations with the final parameter estimates reasonably close to the fitted model based on the complete data set. (In Table 9.1, estimates of the missing values are recorded only for the first three.) Also notice how $-2\ell\left(\boldsymbol{\theta}^{(i)}, \mathbf{W}\right)$ decreases at every iteration. The standard errors of the parameter estimates produced from the last iteration of ITSM are based on a "complete" data set and, as such, underestimate the true sampling errors. Formulae for adjusting the standard errors to reflect the true sampling error based on the observed data can be found in Dempster et al. (1977).

□

9.8 Generalized State-Space Models

As in Section 9.1, we consider a sequence of state variables $\{X_t,\ t \geq 1\}$ and a sequence of observations $\{Y_t,\ t \geq 1\}$. For simplicity, we consider only one-dimensional state and observation variables, since extensions to higher dimensions can be carried out with

Table 9.1 Estimates of the missing observations at times $t = 17, 24, 31$ and the AR estimates using the EM algorithm in Example 9.7.1

Iteration i	$\hat{W}_{17}$	$\hat{W}_{24}$	$\hat{W}_{31}$	$\hat{\phi}_1^{(i)}$	$\hat{\phi}_2^{(i)}$	$-2\ell\left(\boldsymbol{\theta}^{(i)}, \mathbf{W}\right)$
0				0	0	322.60
1	0	0	0	0.7252	0.0236	244.76
2	0.534	0.205	0.746	1.0729	−0.2838	203.57
3	0.458	0.393	0.821	1.0999	−0.3128	202.25
4	0.454	0.405	0.826	1.0999	−0.3128	202.25

little change. Throughout this section it will be convenient to write $\mathbf{Y}^{(t)}$ and $\mathbf{X}^{(t)}$ for the t dimensional column vectors $\mathbf{Y}^{(t)} = (Y_1, Y_2, \ldots, Y_t)'$ and $\mathbf{X}^{(t)} = (X_1, X_2, \ldots, X_t)'$.

There are two important types of state-space models, "parameter driven" and "observation driven," both of which are frequently used in time series analysis. The observation equation is the same for both, but the state vectors of a parameter-driven model evolve independently of the past history of the observation process, while the state vectors of an observation-driven model depend on past observations.

9.8.1 Parameter-Driven Models

In place of the observation and state equations (9.1.1) and (9.1.2), we now make the assumptions that Y_t given $\left(X_t, \mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$ is independent of $\left(\mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$ with conditional probability density

$$p(y_t|x_t) := p\left(y_t|x_t, \mathbf{x}^{(t-1)}, \mathbf{y}^{(t-1)}\right), \quad t = 1, 2, \ldots, \tag{9.8.1}$$

and that X_{t+1} given $\left(X_t, \mathbf{X}^{(t-1)}, \mathbf{Y}^{(t)}\right)$ is independent of $\left(\mathbf{X}^{(t-1)}, \mathbf{Y}^{(t)}\right)$ with conditional density function

$$p(x_{t+1}|x_t) := p\left(x_{t+1}|x_t, \mathbf{x}^{(t-1)}, \mathbf{y}^{(t)}\right) \quad t = 1, 2, \ldots. \tag{9.8.2}$$

We shall also assume that the initial state X_1 has probability density p_1. The joint density of the observation and state variables can be computed directly from (9.8.1)–(9.8.2) as

$$\begin{aligned}
p(y_1, \ldots, y_n, x_1, \ldots, x_n) &= p\left(y_n|x_n, \mathbf{x}^{(n-1)}, \mathbf{y}^{(n-1)}\right) p\left(x_n, \mathbf{x}^{(n-1)}, \mathbf{y}^{(n-1)}\right) \\
&= p(y_n|x_n) p\left(x_n|\mathbf{x}^{(n-1)}, \mathbf{y}^{(n-1)}\right) p\left(\mathbf{y}^{(n-1)}, \mathbf{x}^{(n-1)}\right) \\
&= p(y_n|x_n) p(x_n|x_{n-1}) p\left(\mathbf{y}^{(n-1)}, \mathbf{x}^{(n-1)}\right) \\
&= \cdots \\
&= \left(\prod_{j=1}^{n} p(y_j|x_j)\right)\left(\prod_{j=2}^{n} p(x_j|x_{j-1})\right) p_1(x_1),
\end{aligned}$$

and since (9.8.2) implies that $\{X_t\}$ is Markov (see Problem 9.22),

$$p(y_1, \ldots, y_n|x_1, \ldots, x_n) = \left(\prod_{j=1}^{n} p(y_j|x_j)\right). \tag{9.8.3}$$

We conclude that $Y_1, \ldots, Y_n$ are conditionally independent given the state variables $X_1, \ldots, X_n$, so that the dependence structure of $\{Y_t\}$ is inherited from that of the state process $\{X_t\}$. The sequence of state variables $\{X_t\}$ is often referred to as the **hidden** or **latent** generating process associated with the observed process.

In order to solve the **filtering** and **prediction** problems in this setting, we shall determine the conditional densities $p\left(x_t|\mathbf{y}^{(t)}\right)$ of X_t given $\mathbf{Y}^{(t)}$, and $p\left(x_t|\mathbf{y}^{(t-1)}\right)$ of X_t given $\mathbf{Y}^{(t-1)}$, respectively. The minimum mean squared error estimates of X_t based on $\mathbf{Y}^{(t)}$ and $\mathbf{Y}^{(t-1)}$ can then be computed as the conditional expectations, $E\left(X_t|\mathbf{Y}^{(t)}\right)$ and $E\left(X_t|\mathbf{Y}^{(t-1)}\right)$.

An application of Bayes's theorem, using the assumption that the distribution of Y_t given $\left(X_t, \mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$ does not depend on $\left(\mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$, yields

$$p\left(x_t|\mathbf{y}^{(t)}\right) = p(y_t|x_t)p\left(x_t|\mathbf{y}^{(t-1)}\right)/p\left(y_t|\mathbf{y}^{(t-1)}\right) \tag{9.8.4}$$

and

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = \int p\left(x_t|\mathbf{y}^{(t)}\right)p(x_{t+1}|x_t)\,d\mu(x_t). \tag{9.8.5}$$

(The integral relative to $d\mu(x_t)$ in (9.8.4) is interpreted as the integral relative to dx_t in the continuous case and as the sum over all values of x_t in the discrete case.) The initial condition needed to solve these recursions is

$$p\left(x_1|\mathbf{y}^{(0)}\right) := p_1(x_1). \tag{9.8.6}$$

The factor $p\left(y_t|\mathbf{y}^{(t-1)}\right)$ appearing in the denominator of (9.8.4) is just a scale factor, determined by the condition $\int p\left(x_t|\mathbf{y}^{(t)}\right)\,d\mu(x_t) = 1$. In the generalized state-space setup, prediction of a future state variable is less important than forecasting a future value of the observations. The relevant forecast density can be computed from (9.8.5) as

$$p\left(y_{t+1}|\mathbf{y}^{(t)}\right) = \int p(y_{t+1}|x_{t+1})p\left(x_{t+1}|\mathbf{y}^{(t)}\right)\,d\mu(x_{t+1}). \tag{9.8.7}$$

Equations (9.8.1)–(9.8.2) can be regarded as a Bayesian model specification. A classical Bayesian model has two key assumptions. The first is that the data $Y_1, \ldots, Y_t$, given an unobservable parameter ($\mathbf{X}^{(t)}$ in our case), are independent with specified conditional distribution. This corresponds to (9.8.3). The second specifies a **prior distribution** for the parameter value. This corresponds to (9.8.2). The **posterior distribution** is then the conditional distribution of the parameter given the data. In the present setting the posterior distribution of the component X_t of $\mathbf{X}^{(t)}$ is determined by the solution (9.8.4) of the filtering problem.

Example 9.8.1. Consider the simplified version of the linear state-space model of Section 9.1,

$$Y_t = GX_t + W_t, \quad \{W_t\} \sim \text{iid N}(0, R), \tag{9.8.8}$$

$$X_{t+1} = FX_t + V_t, \quad \{V_t\} \sim \text{iid N}(0, Q), \tag{9.8.9}$$

where the noise sequences $\{W_t\}$ and $\{V_t\}$ are independent of each other. For this model the probability densities in (9.8.1)–(9.8.2) become

$$p_1(x_1) = n(x_1; EX_1, \text{Var}(X_1)), \tag{9.8.10}$$

$$p(y_t|x_t) = n(y_t; Gx_t, R), \tag{9.8.11}$$

$$p(x_{t+1}|x_t) = n(x_{t+1}; Fx_t, Q), \tag{9.8.12}$$

where $n\left(x; \mu, \sigma^2\right)$ is the normal density with mean μ and variance σ^2 defined in Example (a) of Section A.1.

To solve the filtering and prediction problems in this new framework, we first observe that the filtering and prediction densities in (9.8.4) and (9.8.5) are both normal. We shall write them, using the notation of Section 9.4, as

$$p\left(x_t|\mathbf{Y}^{(t)}\right) = n(x_t; X_{t|t}, \Omega_{t|t}) \tag{9.8.13}$$

and

$$p\left(x_{t+1}|\mathbf{Y}^{(t)}\right) = n\left(x_{t+1}; \hat{X}_{t+1}, \Omega_{t+1}\right). \tag{9.8.14}$$

From (9.8.5), (9.8.12), (9.8.13), and (9.8.14), we find that

$$\begin{aligned}
\hat{X}_{t+1} &= \int_{-\infty}^{\infty} x_{t+1} p(x_{t+1}|\mathbf{Y}^{(t)}) dx_{t+1} \\
&= \int_{-\infty}^{\infty} x_{t+1} \int_{-\infty}^{\infty} p(x_t|\mathbf{Y}^{(t)}) p(x_{t+1}|x_t)\, dx_t\, dx_{t+1} \\
&= \int_{-\infty}^{\infty} p(x_t|\mathbf{Y}^{(t)}) \left[\int_{-\infty}^{\infty} x_{t+1} p(x_{t+1}|x_t)\, dx_{t+1}\right] dx_t \\
&= \int_{-\infty}^{\infty} F x_t p(x_t|\mathbf{Y}^{(t)})\, dx_t \\
&= F X_{t|t}
\end{aligned}$$

and (see Problem 9.23)

$$\Omega_{t+1} = F^2 \Omega_{t|t} + Q.$$

Substituting the corresponding densities (9.8.11) and (9.8.14) into (9.8.4), we find by equating the coefficient of x_t^2 on both sides of (9.8.4) that

$$\Omega_{t|t}^{-1} = G^2 R^{-1} + \Omega_t^{-1} = G^2 R^{-1} + (F^2 \Omega_{t-1|t-1} + Q)^{-1}$$

and

$$X_{t|t} = \hat{X}_t + \Omega_{t|t} G R^{-1} \left(Y_t - G\hat{X}_t\right).$$

Also, from (9.8.4) with $p\left(x_1|\mathbf{y}^{(0)}\right) = n(x_1; EX_1, \Omega_1)$ we obtain the initial conditions

$$X_{1|1} = EX_1 + \Omega_{1|1} G R^{-1} (Y_1 - GEX_1)$$

and

$$\Omega_{1|1}^{-1} = G^2 R^{-1} + \Omega_1^{-1}.$$

The Kalman prediction and filtering recursions of Section 9.4 give the same results for $\hat{X}_t$ and $X_{t|t}$, since for Gaussian systems best linear mean square estimation is equivalent to best mean square estimation.

□

Example 9.8.2. A non-Gaussian Example

In general, the solution of the recursions (9.8.4) and (9.8.5) presents substantial computational problems. Numerical methods for dealing with non-Gaussian models are discussed by Sorenson and Alspach (1971) and Kitagawa (1987). Here we shall illustrate the recursions (9.8.4) and (9.8.5) in a very simple special case. Consider the state equation

$$X_t = aX_{t-1}, \tag{9.8.15}$$

with observation density

$$p(y_t|x_t) = \frac{(\pi x_t)^{y_t} e^{-\pi x_t}}{y_t!}, \quad y_t = 0, 1, \ldots, \tag{9.8.16}$$

where π is a constant between 0 and 1. The relationship in (9.8.15) implies that the transition density [in the discrete sense—see the comment after (9.8.5)] for the state variables is

$$p(x_{t+1}|x_t) = \begin{cases} 1, & \text{if } x_{t+1} = ax_t, \\ 0, & \text{otherwise.} \end{cases}$$

We shall assume that X_1 has the gamma density function

$$p_1(x_1) = g(x_1; \alpha, \lambda) = \frac{\lambda^\alpha x_1^{\alpha-1} e^{-\lambda x_1}}{\Gamma(\alpha)}, \quad x_1 > 0.$$

(This is a simplified model for the evolution of the number X_t of individuals at time t infected with a rare disease, in which X_t is treated as a continuous rather than an integer-valued random variable. The observation Y_t represents the number of infected individuals observed in a random sample consisting of a small fraction π of the population at time t.) Because the transition distribution of $\{X_t\}$ is not continuous, we use the integrated version of (9.8.5) to compute the prediction density. Thus,

$$\begin{aligned} P\left(X_t \le x|\mathbf{y}^{(t-1)}\right) &= \int_0^\infty P(X_t \le x|x_{t-1}) p\left(x_{t-1}|\mathbf{y}^{(t-1)}\right) dx_{t-1} \\ &= \int_0^{x/a} p\left(x_{t-1}|\mathbf{y}^{(t-1)}\right) dx_{t-1}. \end{aligned}$$

Differentiation with respect to x gives

$$p\left(x_t|\mathbf{y}^{(t-1)}\right) = a^{-1} p_{X_{t-1}|\mathbf{Y}^{(t-1)}}\left(a^{-1}x_t|\mathbf{y}^{(t-1)}\right). \tag{9.8.17}$$

Now applying (9.8.4), we find that

$$\begin{aligned} p(x_1|y_1) &= p(y_1|x_1)p_1(x_1)/p(y_1) \\ &= \left(\frac{(\pi x_1)^{y_1} e^{-\pi x_1}}{y_1!}\right)\left(\frac{\lambda^\alpha x_1^{\alpha-1} e^{-\lambda x_1}}{\Gamma(\alpha)}\right)\left(\frac{1}{p(y_1)}\right) \\ &= c(y_1) x_1^{\alpha+y_1-1} e^{-(\pi+\lambda)x_1}, \quad x_1 > 0, \end{aligned}$$

where $c(y_1)$ is an integration factor ensuring that $p(\cdot|y_1)$ integrates to 1. Since $p(\cdot|y_1)$ has the form of a gamma density, we deduce (see Example (d) of Section A.1) that

$$p(x_1|y_1) = g(x_1; \alpha_1, \lambda_1), \tag{9.8.18}$$

where $\alpha_1 = \alpha + y_1$ and $\lambda_1 = \lambda + \pi$. The prediction density, calculated from (9.8.5) and (9.8.18), is

$$\begin{aligned} p\left(x_2|\mathbf{y}^{(1)}\right) &= a^{-1} p_{X_1|\mathbf{Y}^{(1)}}\left(a^{-1}x_2|\mathbf{y}^{(1)}\right) \\ &= a^{-1} g\left(a^{-1}x_2; \alpha_1, \lambda_1\right) \\ &= g(x_2; \alpha_1, \lambda_1/a). \end{aligned}$$

Iterating the recursions (9.8.4) and (9.8.5) and using (9.8.17), we find that for $t \ge 1$,

$$p\left(x_t|\mathbf{y}^{(t)}\right) = g(x_t; \alpha_t, \lambda_t) \tag{9.8.19}$$

and

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = a^{-1}g\left(a^{-1}x_{t+1};\alpha_t,\lambda_t\right)$$
$$= g(x_{t+1};\alpha_t,\lambda_t/a), \tag{9.8.20}$$

where $\alpha_t = \alpha_{t-1} + y_t = \alpha + y_1 + \cdots + y_t$ and $\lambda_t = \lambda_{t-1}/a + \pi = \lambda a^{1-t} + \pi\left(1 - a^{-t}\right)/(1 - a^{-1})$. In particular, the minimum mean squared error estimate of x_t based on $\mathbf{y}^{(t)}$ is the conditional expectation α_t/λ_t with conditional variance α_t/λ_t^2. From (9.8.7) the probability density of Y_{t+1} given $\mathbf{Y}^{(t)}$ is

$$\begin{aligned} p(y_{t+1}|\mathbf{y}^{(t)}) &= \int_0^\infty \left(\frac{(\pi x_{t+1})^{y_{t+1}}e^{-\pi x_{t+1}}}{y_{t+1}!}\right) g(x_{t+1};\alpha_t,\lambda_t/a)\,dx_{t+1} \\ &= \frac{\Gamma(\alpha_t + y_{t+1})}{\Gamma(\alpha_t)\Gamma(y_{t+1}+1)}\left(1 - \frac{\pi}{\lambda_{t+1}}\right)^{\alpha_t}\left(\frac{\pi}{\lambda_{t+1}}\right)^{y_{t+1}} \\ &= nb(y_{t+1};\alpha_t, 1 - \pi/\lambda_{t+1}), \qquad y_{t+1} = 0, 1, \ldots, \end{aligned}$$

where $nb(y;\alpha,p)$ is the negative binomial density defined in example (i) of Section A.1. Conditional on $\mathbf{Y}^{(t)}$, the best one-step predictor of Y_{t+1} is therefore the mean, $\alpha_t\pi/(\lambda_{t+1} - \pi)$, of this negative binomial distribution. The conditional mean squared error of the predictor is $\mathrm{Var}\left(Y_{t+1}|\mathbf{Y}^{(t)}\right) = \alpha_t\pi\lambda_{t+1}/(\lambda_{t+1} - \pi)^2$ (see Problem 9.25).

□

Example 9.8.3. A Model for Time Series of Counts

We often encounter time series in which the observations represent count data. One such example is the monthly number of newly recorded cases of poliomyelitis in the U.S. for the years 1970–1983 plotted in Figure 9-6. Unless the actual counts are large and can be approximated by continuous variables, Gaussian and linear time series models are generally inappropriate for analyzing such data. The parameter-driven specification provides a flexible class of models for modeling count data. We now discuss a specific model based on a Poisson observation density. This model is similar to the one presented by Zeger (1988) for analyzing the polio data. The observation density is assumed to be Poisson with mean $\exp\{x_t\}$, i.e.,

$$p(y_t|x_t) = \frac{e^{x_t y_t}e^{-e^{x_t}}}{y_t!}, \qquad y_t = 0, 1, \ldots, \tag{9.8.21}$$

while the state variables are assumed to follow a regression model with Gaussian AR(1) noise. If $\mathbf{u}_t = (u_{t1}, \ldots, u_{tk})'$ are the regression variables, then

$$X_t = \boldsymbol{\beta}'\mathbf{u}_t + W_t, \tag{9.8.22}$$

where $\boldsymbol{\beta}$ is a k-dimensional regression parameter and

$$W_t = \phi W_{t-1} + Z_t, \qquad \{Z_t\} \sim \mathrm{IID\ N}\left(0, \sigma^2\right).$$

The transition density function for the state variables is then

$$p(x_{t+1}|x_t) = n(x_{t+1};\boldsymbol{\beta}'\mathbf{u}_{t+1} + \phi\left(x_t - \boldsymbol{\beta}'\mathbf{u}_t\right), \sigma^2). \tag{9.8.23}$$

The case $\sigma^2 = 0$ corresponds to a log-linear model with Poisson noise.

Estimation of the parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \phi, \sigma^2\right)'$ in the model by direct numerical maximization of the likelihood function is difficult, since the likelihood cannot be written down in closed form. (From (9.8.3) the likelihood is the n-fold integral,

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{\sum_{t=1}^{n}(x_t y_t - e^{x_t})\right\} L\left(\boldsymbol{\theta};\mathbf{x}^{(n)}\right)\ (dx_1\cdots dx_n)\Big/\prod_{i=1}^{n}(y_i!),$$

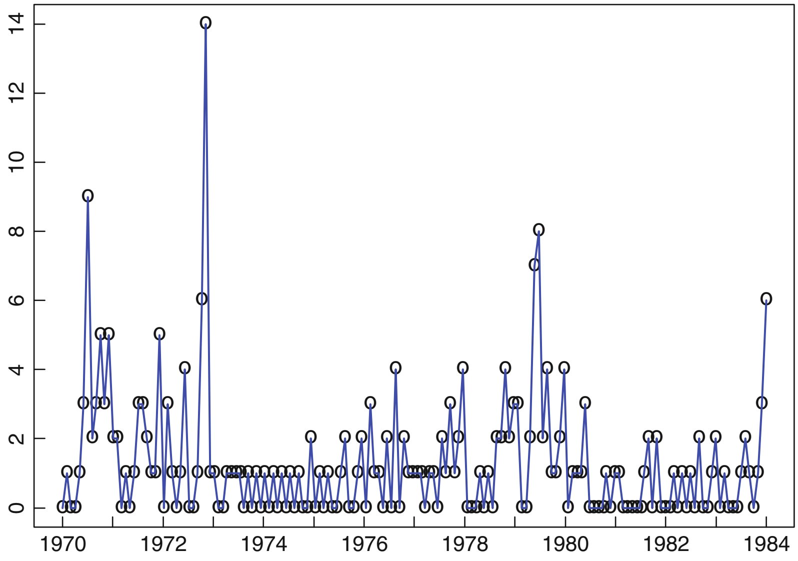

Figure 9-6 Monthly number of U.S. cases of polio, January 1970–December 1983

where $L(\boldsymbol{\theta};\mathbf{x})$ is the likelihood based on $X_1,\ldots,X_n$.) To overcome this difficulty, Chan and Ledolter (1995) proposed an algorithm, called Monte Carlo EM (MCEM), whose iterates $\theta^{(i)}$ converge to the maximum likelihood estimate. To apply this algorithm, first note that the conditional distribution of $\mathbf{Y}^{(n)}$ given $\mathbf{X}^{(n)}$ does not depend on $\boldsymbol{\theta}$, so that the likelihood based on the complete data $\left(\mathbf{X}^{(n)\prime},\mathbf{Y}^{(n)\prime}\right)'$ is given by

$$L\left(\boldsymbol{\theta};\mathbf{X}^{(n)},\mathbf{Y}^{(n)}\right)=f\left(\mathbf{Y}^{(n)}|\mathbf{X}^{(n)}\right)L\left(\boldsymbol{\theta};\mathbf{X}^{(n)}\right).$$

The E-step of the algorithm (see Section 9.7) requires calculation of

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})&=E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\mathbf{X}^{(n)},\mathbf{Y}^{(n)})|\mathbf{Y}^{(n)}\right)\\&=E_{\boldsymbol{\theta}^{(i)}}\left(\ln f(\mathbf{Y}^{(n)}|\mathbf{X}^{(n)})|\mathbf{Y}^{(n)}\right)+E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\mathbf{X}^{(n)})|\mathbf{Y}^{(n)}\right).\end{aligned}$$

We delete the first term from the definition of Q, since it is independent of $\boldsymbol{\theta}$ and hence plays no role in the M-step of the EM algorithm. The new Q is redefined as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})=E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\mathbf{X}^{(n)})|\mathbf{Y}^{(n)}\right). \tag{9.8.24}$$

Even with this simplification, direct calculation of Q is still intractable. Suppose for the moment that it is possible to generate replicates of $\mathbf{X}^{(n)}$ from the conditional distribution of $\mathbf{X}^{(n)}$ given $\mathbf{Y}^{(n)}$ when $\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}$. If we denote m independent replicates of $\mathbf{X}^{(n)}$ by $\mathbf{X}_1^{(n)},\ldots,\mathbf{X}_m^{(n)}$, then a Monte Carlo approximation to Q in (9.8.24) is given by

$$Q_m\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}\right)=\frac{1}{m}\sum_{j=1}^{m}\ln L\left(\boldsymbol{\theta};\mathbf{X}_j^{(n)}\right).$$

The M-step is easy to carry out using Q_m in place of Q (especially if we condition on $X_1=0$ in all the simulated replicates), since L is just the Gaussian likelihood of the regression model with AR(1) noise treated in Section 6.6. The difficult steps in the algorithm are the generation of replicates of $\mathbf{X}^{(n)}$ given $\mathbf{Y}^{(n)}$ and the choice of m. Chan and Ledolter (1995) discuss the use of the Gibb's sampler for generating the desired replicates and give some guidelines on the choice of m.

In their analyses of the polio data, Zeger (1988) and Chan and Ledolter (1995) included as regression components an intercept, a slope, and harmonics at periods of 6 and 12 months. Specifically, they took

$$\mathbf{u}_t=(1,t/1000,\cos(2\pi t/12),\sin(2\pi t/12),\cos(2\pi t/6),\sin(2\pi t/6))'.$$

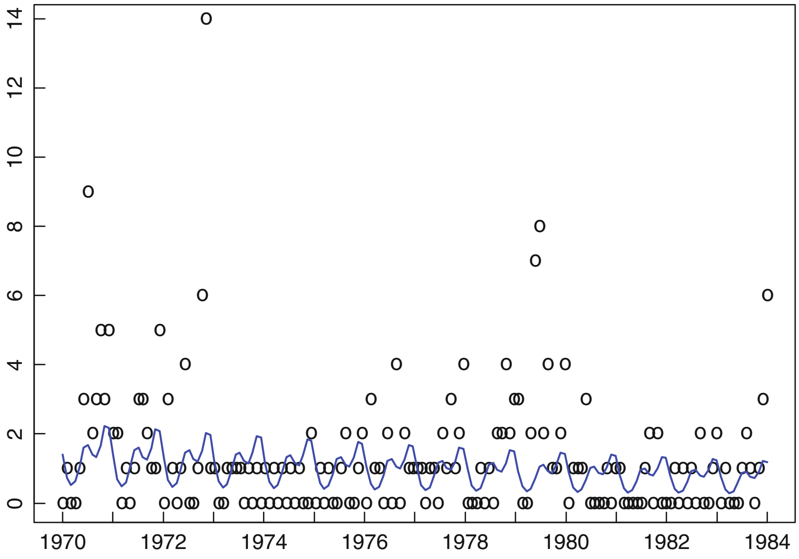

Figure 9-7 Trend estimate for the monthly number of U.S. cases of polio, January 1970–December 1983

The implementation of Chan and Ledolter's MCEM method by Kuk and Cheng (1994) gave estimates $\hat{\boldsymbol{\beta}} = (0.247, -3.871, 0.162, -0.482, 0.414, -0.011)'$, $\hat{\phi} = 0.648$, and $\hat{\sigma}^2 = 0.281$. The estimated trend function $\hat{\boldsymbol{\beta}}'\mathbf{u}_t$ is displayed in Figure 9-7. The negative coefficient of $t/1000$ indicates a slight downward trend in the monthly number of polio cases.

□

9.8.2 Observation-Driven Models

Again we assume that Y_t, conditional on $\left(X_t, \mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$, is independent of $\left(\mathbf{X}^{(t-1)}, \mathbf{Y}^{(t-1)}\right)$. These models are specified by the conditional densities

$$p(y_t|x_t) = p\left(y_t|\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}\right), \quad t = 1, 2, \ldots, \tag{9.8.25}$$

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = p_{X_{t+1}|\mathbf{Y}^{(t)}}\left(x_{t+1}|\mathbf{y}^{(t)}\right), \quad t = 0, 1, \ldots, \tag{9.8.26}$$

where $p\left(x_1|\mathbf{y}^{(0)}\right) := p_1(x_1)$ for some prespecified initial density $p_1(x_1)$. The advantage of the observation-driven state equation (9.8.26) is that the posterior distribution of X_t given $\mathbf{Y}^{(t)}$ can be computed directly from (9.8.4) without the use of the updating formula (9.8.5). This then allows for easy computation of the forecast function in (9.8.7) and hence of the joint density function of $(Y_1, \ldots, Y_n)'$,

$$p(y_1, \ldots, y_n) = \prod_{t=1}^{n} p\left(y_t|\mathbf{y}^{(t-1)}\right). \tag{9.8.27}$$

On the other hand, the mechanism by which the state X_{t-1} makes the transition to X_t is not explicitly defined. In fact, without further assumptions there may be state sequences $\{X_t\}$ and $\{X_t^*\}$ with different distributions for which both (9.8.25) and (9.8.26) hold (see Example 9.8.6). Both sequences, however, lead to the same joint distribution, given by (9.8.27), for $Y_1, \ldots, Y_n$. The ambiguity in the specification of the distribution of the state variables can be removed by assuming that X_{t+1} given $\left(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}\right)$ is independent of $\mathbf{X}^{(t)}$, with conditional distribution (9.8.26), i.e.,

$$p\left(x_{t+1}|\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right) = p_{X_{t+1}|\mathbf{Y}^{(t)}}\left(x_{t+1}|\mathbf{y}^{(t)}\right). \tag{9.8.28}$$

With this modification, the joint density of $\mathbf{Y}^{(n)}$ and $\mathbf{X}^{(n)}$ is given by (cf. (9.8.3))

$$\begin{aligned} p\left(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}\right) &= p(y_n|x_n)p\left(x_n|\mathbf{y}^{(n-1)}\right)p\left(\mathbf{y}^{(n-1)}, \mathbf{x}^{(n-1)}\right) \\ &= \cdots \\ &= \prod_{t=1}^{n}\left(p(y_t|x_t)p\left(x_t|\mathbf{y}^{(t-1)}\right)\right). \end{aligned}$$

Example 9.8.4. An AR(1) Process

An AR(1) process with iid noise can be expressed as an observation driven model. Suppose $\{Y_t\}$ is the AR(1) process

$$Y_t = \phi Y_{t-1} + Z_t,$$

where $\{Z_t\}$ is an iid sequence of random variables with mean 0 and some probability density function $f(x)$. Then with $X_t := Y_{t-1}$ we have

$$p(y_t|x_t) = f(y_t - \phi x_t)$$

and

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = \begin{cases} 1, & \text{if } x_{t+1} = y_t, \\ 0, & \text{otherwise.} \end{cases}$$

□

Example 9.8.5. Suppose the observation-equation density is given by

$$p(y_t|x_t) = \frac{x_t^{y_t}e^{-x_t}}{y_t!}, \quad y_t = 0, 1, \ldots, \tag{9.8.29}$$

and the state equation (9.8.26) is

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = g(x_t; \alpha_t, \lambda_t), \tag{9.8.30}$$

where $\alpha_t = \alpha + y_1 + \cdots + y_t$ and $\lambda_t = \lambda + t$. It is possible to give a parameter-driven specification that gives rise to the same state equation (9.8.30). Let $\{X_t^*\}$ be the parameter-driven state variables, where $X_t^* = X_{t-1}^*$ and X_1^* has a gamma distribution with parameters α and λ. (This corresponds to the model in Example 9.8.2 with $\pi = a = 1$.) Then from (9.8.19) we see that $p\left(x_t^*|\mathbf{y}^{(t)}\right) = g(x_t^*; \alpha_t, \lambda_t)$, which coincides with the state equation (9.8.30). If $\{X_t\}$ are the state variables whose joint distribution is specified through (9.8.28), then $\{X_t\}$ and $\{X_t^*\}$ cannot have the same joint distributions. To see this, note that

$$p\left(x_{t+1}^*|x_t^*\right) = \begin{cases} 1, & \text{if } x_{t+1}^* = x_t^*, \\ 0, & \text{otherwise,} \end{cases}$$

while

$$p\left(x_{t+1}|\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right) = p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = g(x_t; \alpha_t, \lambda_t).$$

If the two sequences had the same joint distribution, then the latter density could take only the values 0 and 1, which contradicts the continuity (as a function of x_t) of this density.

□

9.8.3 Exponential Family Models

The exponential family of distributions provides a large and flexible class of distributions for use in the observation equation. The density in the observation equation is said to belong to an **exponential family** (in natural parameterization) if

$$p(y_t|x_t) = \exp\{y_t x_t - b(x_t) + c(y_t)\}, \tag{9.8.31}$$

where $b(\cdot)$ is a twice continuously differentiable function and $c(y_t)$ does not depend on x_t. This family includes the normal, exponential, gamma, Poisson, binomial, and many other distributions frequently encountered in statistics. Detailed properties of the exponential family can be found in Barndorff-Nielsen (1978), and an excellent treatment of its use in the analysis of linear models is given by McCullagh and Nelder (1989). We shall need only the following important facts:

$$e^{b(x_t)} = \int \exp\{y_t x_t + c(y_t)\}\, \nu(dy_t), \tag{9.8.32}$$

$$b'(x_t) = E(Y_t|x_t), \tag{9.8.33}$$

$$b''(x_t) = \mathrm{Var}(Y_t|x_t) := \int y_t^2 p(y_t|x_t)\, \nu(dy_t) - \left[b'(x_t)\right]^2, \tag{9.8.34}$$

where integration with respect to $\nu(dy_t)$ means integration with respect to dy_t in the continuous case and summation over all values of y_t in the discrete case.

Proof. The first relation is simply the statement that $p(y_t|x_t)$ integrates to 1. The second relation is established by differentiating both sides of (9.8.32) with respect to x_t and then multiplying through by $e^{-b(x_t)}$ (for justification of the differentiation under the integral sign see Barndorff-Nielsen 1978). The last relation is obtained by differentiating (9.8.32) twice with respect to x_t and simplifying. ■

Example 9.8.6. The Poisson Case

If the observation Y_t, given $X_t = x_t$, has a Poisson distribution of the form (9.8.21), then

$$p(y_t|x_t) = \exp\left\{y_t x_t - e^{x_t} - \ln y_t!\right\}, \quad y_t = 0, 1, \ldots, \tag{9.8.35}$$

which has the form (9.8.31) with $b(x_t) = e^{x_t}$ and $c(y_t) = -\ln y_t!$. From (9.8.33) we easily find that $E(Y_t|x_t) = b'(x_t) = e^{x_t}$. This parameterization is slightly different from the one used in Examples 9.8.2 and 9.8.5, where the conditional mean of Y_t given x_t was πx_t and not e^{x_t}. For this observation equation, define the family of densities

$$f(x; \alpha, \lambda) = \exp\{\alpha x - \lambda b(x) + A(\alpha, \lambda)\}, \quad -\infty < x < \infty, \tag{9.8.36}$$

where $\alpha > 0$ and $\lambda > 0$ are parameters and $A(\alpha, \lambda) = -\ln \Gamma(\alpha) + \alpha \ln \lambda$. Now consider state densities of the form

$$p(x_{t+1}|\mathbf{y}^{(t)}) = f(x_{t+1}; \alpha_{t+1|t}, \lambda_{t+1|t}), \tag{9.8.37}$$

where $\alpha_{t+1|t}$ and $\lambda_{t+1|t}$ are, for the moment, unspecified functions of $\mathbf{y}^{(t)}$. (The subscript $t+1|t$ on the parameters is a shorthand way to indicate dependence on the conditional distribution of X_{t+1} given $\mathbf{Y}^{(t)}$.) With this specification of the state densities, the parameters $\alpha_{t+1|t}$ are related to the best one-step predictor of Y_t through the formula

$$\alpha_{t+1|t}/\lambda_{t+1|t} = \hat{Y}_{t+1} := E\left(Y_{t+1}|\mathbf{y}^{(t)}\right). \tag{9.8.38}$$

Proof. We have from (9.8.7) and (9.8.33) that

$$\begin{aligned} E(Y_{t+1}|\mathbf{y}^{(t)}) &= \sum_{y_{t+1}=0}^{\infty} \int_{-\infty}^{\infty} y_{t+1} p(y_{t+1}|x_{t+1}) p\left(x_{t+1}|\mathbf{y}^{(t)}\right) \, dx_{t+1} \\ &= \int_{-\infty}^{\infty} b'(x_{t+1}) p\left(x_{t+1}|\mathbf{y}^{(t)}\right) \, dx_{t+1}. \end{aligned}$$

Addition and subtraction of $\alpha_{t+1|t}/\lambda_{t+1|t}$ then gives

$$\begin{aligned} E(Y_{t+1}|\mathbf{y}^{(t)}) &= \int_{-\infty}^{\infty} \left(b'(x_{t+1}) - \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}} \right) p\left(x_{t+1}|\mathbf{y}^{(t)}\right) \, dx_{t+1} + \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}} \\ &= \int_{-\infty}^{\infty} -\lambda_{t+1|t}^{-1} p'\left(x_{t+1}|\mathbf{y}^{(t)}\right) \, dx_{t+1} + \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}} \\ &= \left[-\lambda_{t+1|t}^{-1} p\left(x_{t+1}|\mathbf{y}^{(t)}\right) \right]_{x_{t+1}=-\infty}^{x_{t+1}=\infty} + \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}} \\ &= \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}}. \end{aligned}$$

■

Letting $A_{t|t-1} = A(\alpha_{t|t-1}, \lambda_{t|t-1})$, we can write the posterior density of X_t given $\mathbf{Y}^{(t)}$ as

$$\begin{aligned} p\left(x_t|\mathbf{y}^{(t)}\right) &= \exp\{y_t x_t - b(x_t) + c(y_t)\} \exp\{\alpha_{t|t-1} x_t - \lambda_{t|t-1} b(x_t) \\ &\quad + A_{t|t-1}\}/p\left(y_t|\mathbf{y}^{(t-1)}\right) \\ &= \exp\{\lambda_{t|t}\left(\alpha_{t|t} x_t - b(x_t)\right) - A_{t|t}\}, \\ &= f(x_t; \alpha_t, \lambda_t), \end{aligned}$$

where we find, by equating coefficients of x_t and $b(x_t)$, that the coefficients λ_t and α_t are determined by

$$\lambda_t = 1 + \lambda_{t|t-1}, \qquad (9.8.39)$$

$$\alpha_t = y_t + \alpha_{t|t-1}. \qquad (9.8.40)$$

The family of prior densities in (9.8.37) is called a **conjugate family of priors** for the observation equation (9.8.35), since the resulting posterior densities are again members of the same family.

As mentioned earlier, the parameters $\alpha_{t|t-1}$ and $\lambda_{t|t-1}$ can be quite arbitrary: Any nonnegative functions of $\mathbf{y}^{(t-1)}$ will lead to a consistent specification of the state densities. One convenient choice is to link these parameters with the corresponding parameters of the posterior distribution at time $t-1$ through the relations

$$\lambda_{t+1|t} = \delta\lambda_t \left(= \delta(1 + \lambda_{t|t-1})\right), \qquad (9.8.41)$$

$$\alpha_{t+1|t} = \delta\alpha_t \left(= \delta(y_t + \alpha_{t|t-1})\right), \qquad (9.8.42)$$

where $0 < \delta < 1$ (see Remark 4 below). Iterating the relation (9.8.41), we see that

$$\begin{aligned} \lambda_{t+1|t} = \delta(1 + \lambda_{t|t-1}) &= \delta + \delta\lambda_{t|t-1} \\ &= \delta + \delta(\delta + \delta\lambda_{t-2|t-2}) \\ &= \cdots \end{aligned}$$

$$= \delta + \delta^2 + \cdots + \delta^t + \delta^t \lambda_{1|0} \tag{9.8.43}$$
$$\to \delta/(1-\delta)$$

as $t \to \infty$. Similarly,

$$\begin{aligned}\alpha_{t+1|t} &= \delta y_t + \delta\alpha_{t|t-1}\\ &= \cdots\\ &= \delta y_t + \delta^2 y_{t-1} + \cdots + \delta^t y_1 + \delta^t \alpha_{1|0}.\end{aligned} \tag{9.8.44}$$

For large t, we have the approximations

$$\lambda_{t+1|t} = \delta/(1-\delta) \tag{9.8.45}$$

and

$$\alpha_{t+1|t} = \delta \sum_{j=0}^{t-1} \delta^j y_{t-j}, \tag{9.8.46}$$

which are exact if $\lambda_{1|0} = \delta/(1-\delta)$ and $\alpha_{1|0} = 0$. From (9.8.38) the one-step predictors are linear and given by

$$\hat{Y}_{t+1} = \frac{\alpha_{t+1|t}}{\lambda_{t+1|t}} = \frac{\sum_{j=0}^{t-1} \delta^j y_{t-j} + \delta^{t-1}\alpha_{1|0}}{\sum_{j=0}^{t-1} \delta^j + \delta^{t-1}\lambda_{1|0}}. \tag{9.8.47}$$

Replacing the denominator with its limiting value, or starting with $\lambda_{1|0} = \delta/(1-\delta)$, we find that $\hat{Y}_{t+1}$ is the solution of the recursions

$$\hat{Y}_{t+1} = (1-\delta)y_t + \delta\hat{Y}_t, \quad t = 1, 2, \ldots, \tag{9.8.48}$$

with initial condition $\hat{Y}_1 = (1-\delta)\delta^{-1}\alpha_{1|0}$. In other words, under the restrictions of (9.8.41) and (9.8.42), the best one-step predictors can be found by exponential smoothing.

□

Remark 1. The preceding analysis for the Poisson-distributed observation equation holds, almost verbatim, for the general family of exponential densities (9.8.31). (One only needs to take care in specifying the correct range for x and the allowable parameter space for α and λ in (9.8.37).) The relations (9.8.43)–(9.8.44), as well as the exponential smoothing formula (9.8.48), continue to hold even in the more general setting, provided that the parameters $\alpha_{t|t-1}$ and $\lambda_{t|t-1}$ satisfy the relations (9.8.41)–(9.8.42). □

Remark 2. Equations (9.8.41)–(9.8.42) are equivalent to the assumption that the prior density of X_t given $\mathbf{y}^{(t-1)}$ is proportional to the δ-power of the posterior distribution of X_{t-1} given $\mathbf{Y}^{(t-1)}$, or more succinctly that

$$\begin{aligned}f(x_t; \alpha_{t|t-1}, \lambda_{t|t-1}) &= f(x_t; \delta\alpha_{t-1|t-1}, \delta\lambda_{t-1|t-1})\\ &\propto f^\delta(x_t; \alpha_{t-1|t-1}, \lambda_{t-1|t-1}).\end{aligned}$$

This power relationship is sometimes referred to as the **power steady model** (Grunwald et al. 1993; Smith 1979). □

Remark 3. The transformed state variables $W_t = e^{X_t}$ have a gamma state density given by

$$p\left(w_{t+1}|\mathbf{y}^{(t)}\right) = g(w_{t+1}; \alpha_{t+1|t}, \lambda_{t+1|t})$$

(see Problem 9.26). The mean and variance of this conditional density are

$$E\left(W_{t+1}|\mathbf{y}^{(t)}\right) = \alpha_{t+1|t} \quad \text{and} \quad \text{Var}\left(W_{t+1}|\mathbf{y}^{(t)}\right) = \alpha_{t+1|t}/\lambda^2_{t+1|t}. \qquad \square$$

Remark 4. If we regard the random walk plus noise model of Example 9.2.1 as the prototypical state-space model, then from the calculations in Example 9.8.1 with $G = F = 1$, we have

$$E\left(X_{t+1}|\mathbf{Y}^{(t)}\right) = E\left(X_t|\mathbf{Y}^{(t)}\right)$$

and

$$\text{Var}\left(X_{t+1}|\mathbf{Y}^{(t)}\right) = \text{Var}\left(X_t|\mathbf{Y}^{(t)}\right) + Q > \text{Var}\left(X_t|\mathbf{Y}^{(t)}\right).$$

The first of these equations implies that the best estimate of the next state is the same as the best estimate of the current state, while the second implies that the variance increases. Under the conditions (9.8.41), and (9.8.42), the same is also true for the state variables in the above model (see Problem 9.26). This was, in part, the rationale behind these conditions given in Harvey and Fernandes (1989). □

Remark 5. While the calculations work out neatly for the power steady model, Grunwald et al. (1994) have shown that such processes have degenerate sample paths for large t. In the Poisson example above, they argue that the observations Y_t converge to 0 as $t \rightarrow \infty$ (see Figure 9-12). Although such models may still be useful in practice for modeling series of moderate length, the efficacy of using such models for describing long-term behavior is doubtful. □

Example 9.8.7. Goals Scored by England Against Scotland

The time series of the number of goals scored by England against Scotland in soccer matches played at Hampden Park in Glasgow is graphed in Figure 9-8. The matches have been played nearly every second year, with interruptions during the war years. We will treat the data $y_1, \ldots, y_{52}$ as coming from an equally spaced time series model $\{Y_t\}$. Since the number of goals scored is small (see the frequency histogram in Figure 9-9), a model based on the Poisson distribution might be deemed appropriate. The observed relative frequencies and those based on a Poisson distribution with mean equal to $\bar{y}_{52} = 1.269$ are contained in Table 9.2. The standard chi-squared goodness of fit test, comparing the observed frequencies with expected frequencies based on a Poisson model, has a p-value of 0.02. The lack of fit with a Poisson distribution is hardly unexpected, since the sample variance (1.652) is much larger than the sample mean, while the mean and variance of the Poisson distribution are equal. In this case the data are said to be *overdispersed* in the sense that there is more variability in the data than one would expect from a sample of independent Poisson-distributed variables. Overdispersion can sometimes be explained by serial dependence in the data.

Dependence in count data can often be revealed by estimating the probabilities of transition from one state to another. Table 9.3 contains estimates of these probabilities, computed as the average number of one-step transitions from state y_t to state y_{t+1}. If the data were independent, then in each column the entries should be nearly the same. This is certainly not the case in Table 9.3. For example, England is very unlikely to be shut out or score 3 or more goals in the next match after scoring at least three goals in the previous encounter.

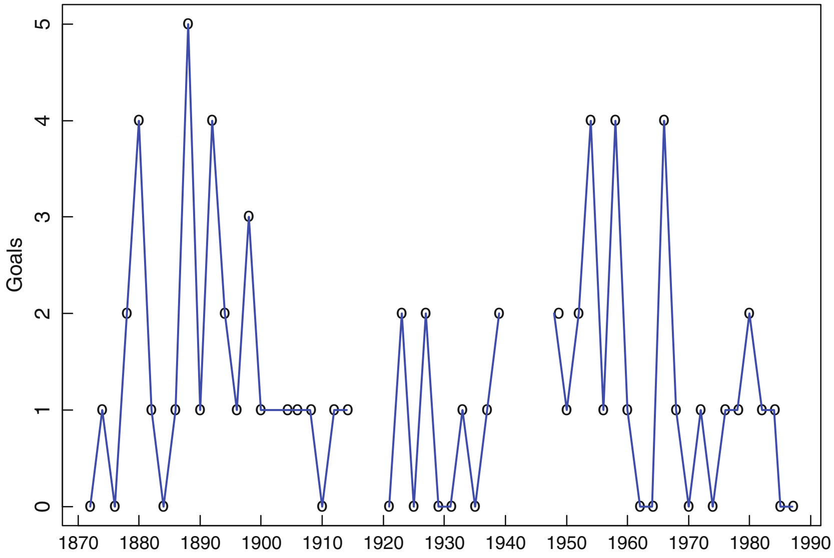

Figure 9-8
Goals scored by England against Scotland at Hampden Park, Glasgow, 1872–1987

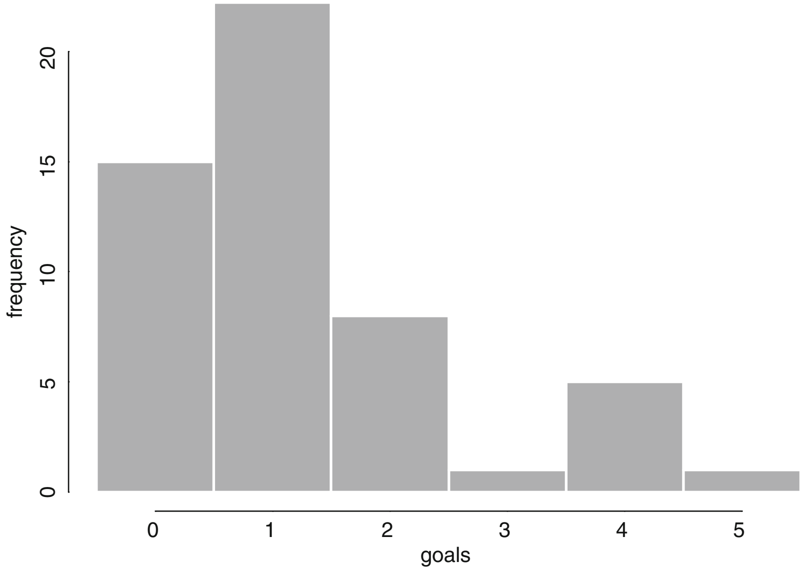

Figure 9-9
Histogram of the data in Figure 9-8

Table 9.2 Relative frequency and fitted Poisson distribution of goals scored by England against Scotland

	Number of goals					
	0	1	2	3	4	5
Relative frequency	0.288	0.423	0.154	0.019	0.096	0.019
Poisson distribution	0.281	0.356	0.226	0.096	0.030	0.008

Harvey and Fernandes (1989) model the dependence in this data using an observation-driven model of the type described in Example 9.8.6. Their model assumes a Poisson observation equation and a log-gamma state equation:

$$p(y_t|x_t) = \frac{\exp\{y_t x_t - e^{x_t}\}}{y_t!}, \qquad y_t = 0, 1, \dots,$$

$$p\left(x_t|\mathbf{y}^{(t-1)}\right) = f(x_t; \alpha_{t|t-1}, \lambda_{t|t-1}), \qquad -\infty < x < \infty,$$

Table 9.3 Transition probabilities for the number of goals scored by England against Scotland

	y_{t+1}				
	$p(y_{t+1} \mid y_t)$	0	1	2	≥ 3
	0	0.214	0.500	0.214	0.072
y_t	1	0.409	0.272	0.136	0.182
	2	0.250	0.375	0.125	0.250
	≥ 3	0	0.857	0.143	0

Table 9.4 Prediction density of Y_{53} given $\mathbf{Y}^{(52)}$ for data in Figure 9-7

	Number of goals					
	0	1	2	3	4	5
$p(y_{53} \mid \mathbf{y}^{(52)})$	0.472	0.326	0.138	0.046	0.013	0.004

for $t = 1, 2, \ldots,$ where f is given by (9.8.36) and $\alpha_{1|0} = 0, \lambda_{1|0} = 0$. The power steady conditions (9.8.41)–(9.8.42) are assumed to hold for $\alpha_{t|t-1}$ and $\lambda_{t|t-1}$. The only unknown parameter in the model is δ. The log-likelihood function for δ based on the conditional distribution of $y_1, \ldots, y_{52}$ given y_1 is given by [see (9.8.27)]

$$\ell\left(\delta, \mathbf{y}^{(n)}\right) = \sum_{t=1}^{n-1} \ln p\left(y_{t+1} | \mathbf{y}^{(t)}\right), \tag{9.8.49}$$

where $p\left(y_{t+1} | \mathbf{y}^{(t)}\right)$ is the negative binomial density [see Problem 9.25(c)]

$$p\left(y_{t+1} | \mathbf{y}^{(t)}\right) = nb\left(y_{t+1}; \alpha_{t+1|t}, (1 + \lambda_{t+1|t})^{-1}\right),$$

with $\alpha_{t+1|t}$ and $\lambda_{t+1|t}$ as defined in (9.8.44) and (9.8.43). (For the goal data, $y_1 = 0$, which implies $\alpha_{2|1} = 0$ and hence that $p\left(y_2 | y^{(1)}\right)$ is a degenerate density with unit mass at $y_2 = 0$. Harvey and Fernandes avoid this complication by conditioning the likelihood on $y^{(\tau)}$, where τ is the time of the first nonzero data value.)

Maximizing this likelihood with respect to δ, we obtain $\hat{\delta} = 0.844$. (Starting equations (9.8.43)–(9.8.44) with $\alpha_{1|0} = 0$ and $\lambda_{1|0} = \delta/(1 - \delta)$, we obtain $\hat{\delta} = 0.732$.) With 0.844 as our estimate of δ, the prediction density of the next observation Y_{53} given $\mathbf{y}^{(52)}$ is $nb(y_{53}; \alpha_{53|52}, (1+\lambda_{53|52})^{-1}$. The first five values of this distribution are given in Table 9.4. Under this model, the probability that England will be held scoreless in the next match is 0.471. The one-step predictors, $\hat{Y}_1 = 0, \hat{Y}_2, \ldots, \hat{Y}_{52}$ are graphed in Figure 9-10. (This graph can be obtained by using the ITSM option `Smooth>Exponential` with $\alpha = 0.154$.)

Figures 9-11 and 9-12 contain two realizations from the fitted model for the goal data. The general appearance of the first realization is somewhat compatible with the goal data, while the second realization illustrates the convergence of the sample path to 0 in accordance with the result of Grunwald et al. (1994).

□

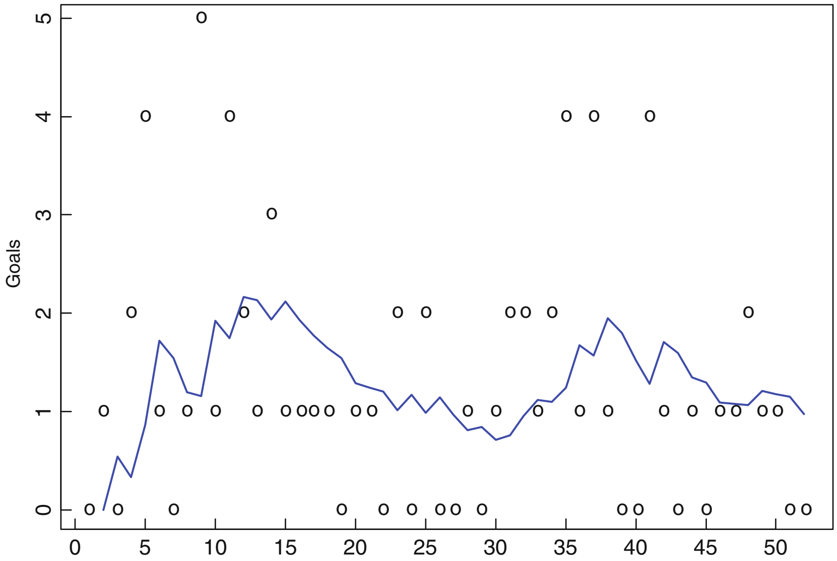

Figure 9-10
One-step predictors of the goal data

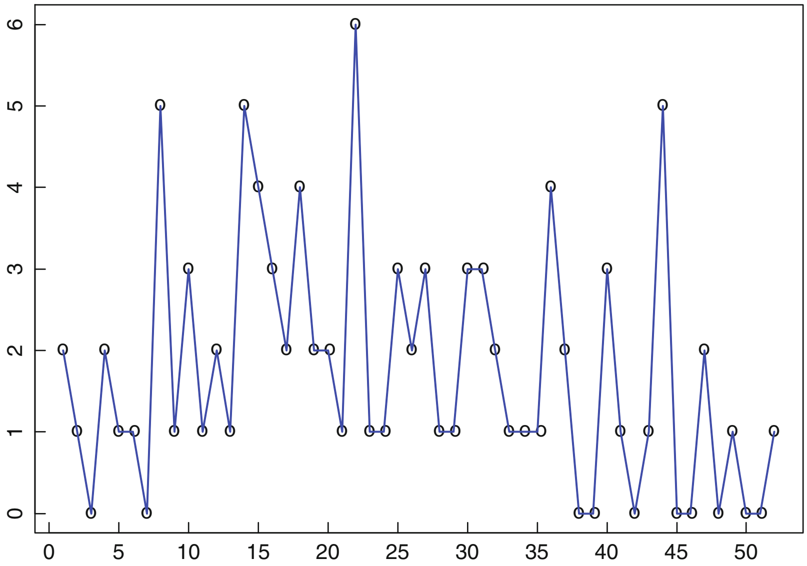

Figure 9-11
A simulated time series from the fitted model to the goal data

Example 9.8.8. The Exponential Case

Suppose Y_t given X_t has an exponential density with mean $-1/X_t$ ($X_t < 0$). The observation density is given by

$$p(y_t|x_t) = \exp\{y_t x_t + \ln(-x_t)\}, \quad y_t > 0,$$

which has the form (9.8.31) with $b(x) = -\ln(-x)$ and $c(y) = 0$. The state densities corresponding to the family of conjugate priors (see (9.8.37)) are given by

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = \exp\{\alpha_{t+1|t}\,x_{t+1} - \lambda_{t+1|t}\,b(x_{t+1}) + A_{t+1|t}\}, \quad -\infty < x < 0.$$

(Here $p(x_{t+1}|\mathbf{y}^{(t)})$ is a probability density when $\alpha_{t+1|t} > 0$ and $\lambda_{t+1|t} > -1$.) The one-step prediction density is

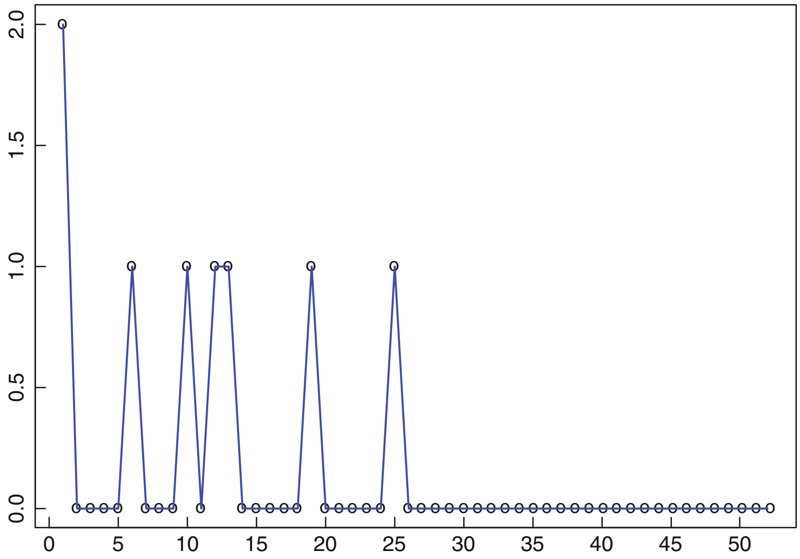

Figure 9-12
A second simulated time series from the fitted model to the goal data

$$p\left(y_{t+1}|\mathbf{y}^{(t)}\right) = \int_{-\infty}^{0} e^{x_{t+1}y_{t+1}+\ln(-x_{t+1})+\alpha_{t+1|t}x-\lambda_{t+1|t}b(x)+A_{t+1|t}}\,dx_{t+1}$$

$$= (\lambda_{t+1|t}+1)\alpha_{t+1|t}^{\lambda_{t+1|t}+1}(y_{t+1}+\alpha_{t+1|t})^{-\lambda_{t+1|t}-2}, \qquad y_{t+1} > 0$$

(see Problem 9.28). While $E(Y_{t+1}|\mathbf{y}^{(t)}) = \alpha_{t+1|t}/\lambda_{t+1|t}$, the conditional variance is finite if and only if $\lambda_{t+1|t} > 1$. Under assumptions (9.8.41)–(9.8.42), and starting with $\lambda_{1|0} = \delta/(1-\delta)$, the exponential smoothing formula (9.8.48) remains valid.

□

Problems

9.1 Show that if all the eigenvalues of F are less than 1 in absolute value (or equivalently that $F^k \to 0$ as $k \to \infty$), the unique stationary solution of equation (9.1.11) is given by the infinite series

$$\mathbf{X}_t = \sum_{j=0}^{\infty} F^j V_{t-j-1}$$

and that the corresponding observation vectors are

$$\mathbf{Y}_t = \mathbf{W}_t + \sum_{j=0}^{\infty} GF^j \mathbf{V}_{t-j-1}.$$

Deduce that $\{(\mathbf{X}_t', \mathbf{Y}_t')'\}$ is a multivariate stationary process. (Hint: Use a vector analogue of the argument in Example 2.2.1.)

9.2 In Example 9.2.1, show that $\theta = -1$ if and only if $\sigma_v^2 = 0$, which in turn is equivalent to the signal M_t being constant.

9.3 Let F be the coefficient of $\mathbf{X}_t$ in the state equation (9.3.4) for the causal AR(p) process

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

Establish the stability of (9.3.4) by showing that

$$\det(zI - F) = z^p \phi\left(z^{-1}\right),$$

and hence that the eigenvalues of F are the reciprocals of the zeros of the autoregressive polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$.

9.4 By following the argument in Example 9.3.3, find a state-space model for $\{Y_t\}$ when $\{\nabla\nabla_{12} Y_t\}$ is an ARMA(p, q) process.

9.5 For the local linear trend model defined by equations (9.2.6)–(9.2.7), show that $\nabla^2 Y_t = (1 - B)^2 Y_t$ is a 2-correlated sequence and hence, by Proposition 2.1.1, is an MA(2) process. Show that this MA(2) process is noninvertible if $\sigma_u^2 = 0$.

9.6 a. For the seasonal model of Example 9.2.2, show that $\nabla_d Y_t = Y_t - Y_{t-d}$ is an MA(1) process.

b. Show that $\nabla\nabla_d Y_t$ is an MA$(d + 1)$ process where $\{Y_t\}$ follows the seasonal model with a local linear trend as described in Example 9.2.3.

9.7 Let $\{Y_t\}$ be the MA(1) process

$$Y_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

Show that $\{Y_t\}$ has the state-space representation

$$Y_t = [1 \quad 0]\mathbf{X}_t,$$

where $\{\mathbf{X}_t\}$ is the unique stationary solution of

$$\mathbf{X}_{t+1} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{X}_t + \begin{bmatrix} 1 \\ \theta \end{bmatrix} Z_{t+1}.$$

In particular, show that the state vector $\mathbf{X}_t$ can written as

$$\mathbf{X}_t = \begin{bmatrix} 1 & \theta \\ \theta & 0 \end{bmatrix} \begin{bmatrix} Z_t \\ Z_{t-1} \end{bmatrix}.$$

9.8 Verify equations (9.3.16)–(9.3.18) for an ARIMA(1,1,1) process.

9.9 Consider the two state-space models

$$\begin{cases} \mathbf{X}_{t+1,1} = F_1 \mathbf{X}_{t1} + \mathbf{V}_{t1}, \\ \mathbf{Y}_{t1} \quad = G_1 \mathbf{X}_{t1} + \mathbf{W}_{t1}, \end{cases}$$

and

$$\begin{cases} \mathbf{X}_{t+1,2} = F_2 \mathbf{X}_{t2} + \mathbf{V}_{t2}, \\ \mathbf{Y}_{t2} \quad = G_2 \mathbf{X}_{t2} + \mathbf{W}_{t2}, \end{cases}$$

where $\{(\mathbf{V}'_{t1}, \mathbf{W}'_{t1}, \mathbf{V}'_{t2}, \mathbf{W}'_{t2})'\}$ is white noise. Derive a state-space representation for $\{(\mathbf{Y}'_{t1}, \mathbf{Y}'_{t2})'\}$.

9.10 Use Remark 1 of Section 9.4 to establish the linearity properties of the operator P_t stated in Remark 3.

9.11 a. Show that if the matrix equation $XS{=}B$ can be solved for X, then $X{=}BS^{-1}$ is a solution for *any* generalized inverse S^{-1} of S.

b. Use the result of (a) to derive the expression for $P(\mathbf{X}|\mathbf{Y})$ in Remark 4 of Section 9.4.

9.12 In the notation of the Kalman prediction equations, show that every vector of the form

$$\mathbf{Y} = A_1\mathbf{X}_1 + \cdots + A_t\mathbf{X}_t$$

can be expressed as

$$\mathbf{Y} = B_1\mathbf{X}_1 + \cdots + B_{t-1}\mathbf{X}_{t-1} + C_t\mathbf{I}_t,$$

where $B_1, \dots, B_{t-1}$ and C_t are matrices that depend on the matrices $A_1, \dots, A_t$. Show also that the converse is true. Use these results and the fact that $E(\mathbf{X}_s\mathbf{I}_t) = 0$ for all $s < t$ to establish (9.4.3).

9.13 In Example 9.4.1, verify that the steady-state solution of the Kalman recursions (9.1.2) is given by $\Omega_t = \left(\sigma_v^2 + \sqrt{\sigma_v^4 + 4\sigma_v^2\sigma_w^2}\right)/2$.

9.14 Show from the difference equations for Ω_t in Example 9.4.1 that $(\Omega_{t+1} - \Omega)(\Omega_t\Omega) \geq 0$ for all $\Omega_t \geq 0$, where Ω is the steady-state solution for Ω_t given in Problem 9.13.

9.15 Show directly that for the MA(1) model (9.2.3), the parameter θ is equal to $-\left(2\sigma_w^2 + \sigma_v^2 - \sqrt{\sigma_v^4 + 4\sigma_v^2\sigma_w^2}\right)/\left(2\sigma_w^2\right)$, which in turn is equal to $-\sigma_w^2/(\Omega + \sigma_w^2)$, where Ω is the steady-state solution for Ω_t given in Problem 9.13.

9.16 Use the ARMA(0,1,1) representation of the series $\{Y_t\}$ in Example 9.4.1 to show that the predictors defined by

$$\hat{Y}_{n+1} = aY_n + (1-a)\hat{Y}_n, \quad n = 1, 2, \dots,$$

where $a = \Omega/(\Omega + \sigma_w^2)$, satisfy

$$Y_{n+1} - \hat{Y}_{n+1} = Z_{n+1} + (1-a)^n\left(Y_0 - Z_0 - \hat{Y}_1\right).$$

Deduce that if $0 < a < 1$, the mean squared error of $\hat{Y}_{n+1}$ converges to $\Omega + \sigma_w^2$ for any initial predictor $\hat{Y}_1$ with finite mean squared error.

9.17 a. Using equations (9.4.1) and (9.4.16), show that $\hat{\mathbf{X}}_{t+1} = F_t\mathbf{X}_{t|t}$.

b. From (a) and (9.4.16) show that $\mathbf{X}_{t|t}$ satisfies the recursions

$$\mathbf{X}_{t|t} = F_{t-1}\mathbf{X}_{t-1|t-1} + \Omega_tG_t'\Delta_t^{-1}(\mathbf{Y}_t - G_tF_{t-1}\mathbf{X}_{t-1|t-1})$$

for $t = 2, 3, \dots$, with $\mathbf{X}_{1|1} = \hat{\mathbf{X}}_1 + \Omega_1G_1'\Delta_1^{-1}\left(\mathbf{Y}_1 - G_1\hat{\mathbf{X}}_1\right)$.

9.18 In Section 9.5, show that for fixed Q^*, $-2\ln L\left(\boldsymbol{\mu}, Q^*, \sigma_w^2\right)$ is minimized when $\boldsymbol{\mu}$ and σ_w^2 are given by (9.5.10) and (9.5.11), respectively.

9.19 Verify the calculation of $\Theta_t\Delta_t^{-1}$ and Ω_t in Example 9.6.1.

9.20 Verify that the best estimates of missing values in an AR(p) process are found by minimizing (9.6.11) with respect to the missing values.

9.21 Suppose that $\{Y_t\}$ is the AR(2) process

$$Y_t = \phi_1Y_{t-1} + \phi_2Y_{t-2} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}\left(0, \sigma^2\right),$$

and that we observe $Y_1, Y_2, Y_4, Y_5, Y_6, Y_7$. Show that the best estimator of Y_3 is

$$(\phi_2(Y_1 + Y_5) + (\phi_1 - \phi_1\phi_2)(Y_2 + Y_4))/\left(1 + \phi_1^2 + \phi_2^2\right).$$

9.22 Let X_t be the state at time t of a parameter-driven model (see (9.8.2)). Show that $\{X_t\}$ is a Markov chain and that (9.8.3) holds.

9.23 For the generalized state-space model of Example 9.8.1, show that $\Omega_{t+1} = F^2\Omega_{t|t} + Q$.

9.24 If Y and X are random variables, show that

$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X)).$$

9.25 Suppose that Y and X are two random variables such that the distribution of Y given X is Poisson with mean πX, $0 < \pi \le 1$, and X has the gamma density $g(x; \alpha, \lambda)$.

a. Show that the posterior distribution of X given Y also has a gamma density and determine its parameters.

b. Compute $E(X|Y)$ and $\mathrm{Var}(X|Y)$.

c. Show that Y has a negative binomial density and determine its parameters.

d. Use (c) to compute $E(Y)$ and $\mathrm{Var}(Y)$.

e. Verify in Example 9.8.2 that $E\left(Y_{t+1}|\mathbf{Y}^{(t)}\right) = \alpha_t\pi/(\lambda_{t+1} - \pi)$ and $\mathrm{Var}\left(Y_{t+1}|\mathbf{Y}^{(t)}\right) = \alpha_t\pi\lambda_{t+1}/(\lambda_{t+1} - \pi)^2$.

9.26 For the model of Example 9.8.6, show that

a. $E\left(X_{t+1}|\mathbf{Y}^{(t)}\right) = E\left(X_t|\mathbf{Y}^{(t)}\right)$, $\mathrm{Var}\left(X_{t+1}|\mathbf{Y}^{(t)}\right) > \mathrm{Var}\left(X_t|\mathbf{Y}^{(t)}\right)$, and

b. the transformed sequence $W_t = e^{X_t}$ has a gamma state density.

9.27 Let $\{V_t\}$ be a sequence of independent exponential random variables with $EV_t = t^{-1}$ and suppose that $\{X_t, t \ge 1\}$ and $\{Y_t, t \ge 1\}$ are the state and observation random variables, respectively, of the parameter-driven state-space system

$$X_1 = V_1,$$

$$X_t = X_{t-1} + V_t, \quad t = 2, 3, \ldots,$$

where the distribution of the observation Y_t, conditional on the random variables $Y_1, Y_2, \ldots, Y_{t-1}, X_t$, is Poisson with mean X_t.

a. Determine the observation and state transition density functions $p(y_t|x_t)$ and $p(x_{t+1}|x_t)$ in the parameter-driven model for $\{Y_t\}$.

b. Show, using (9.8.4)–(9.8.6), that

$$p(x_1|y_1) = g(x_1; y_1 + 1, 2)$$

and

$$p(x_2|y_1) = g(x_2; y_1 + 2, 2),$$

where $g(x; \alpha, \lambda)$ is the gamma density function (see Example (d) of Section A.1).

c. Show that

$$p\left(x_t|\mathbf{y}^{(t)}\right) = g(x_t; \alpha_t + t, t + 1)$$

and

$$p\left(x_{t+1}|\mathbf{y}^{(t)}\right) = g(x_{t+1}; \alpha_t + t + 1, t + 1),$$

where $\alpha_t = y_1 + \cdots + y_t$.

d. Conclude from (c) that the minimum mean squared error estimates of X_t and X_{t+1} based on $Y_1, \ldots, Y_t$ are

$$X_{t|t} = \frac{t + Y_1 + \cdots + Y_t}{t+1}$$

and

$$\hat{X}_{t+1} = \frac{t + 1 + Y_1 + \cdots + Y_t}{t+1},$$

respectively.

9.28 Let Y and X be two random variables such that Y given X is exponential with mean $1/X$, and X has the gamma density function with

$$g(x; \lambda + 1, \alpha) = \frac{\alpha^{\lambda+1} x^{\lambda} \exp\{-\alpha x\}}{\Gamma(\lambda + 1)}, \quad x > 0,$$

where $\lambda > -1$ and $\alpha > 0$.

a. Determine the posterior distribution of X given Y.

b. Show that Y has a Pareto distribution

$$p(y) = (\lambda + 1)\alpha^{\lambda+1}(y + \alpha)^{-\lambda-2}, \quad y > 0.$$

c. Find the mean and variance of Y. Under what conditions on α and λ does the latter exist?

d. Verify the calculation of $p\left(y_{t+1}|\mathbf{y}^{(t)}\right)$ and $E\left(Y_{t+1}|\mathbf{y}^{(t)}\right)$ for the model in Example 9.8.8.

9.29 Consider an observation-driven model in which Y_t given X_t is binomial with parameters n and X_t, i.e.,

$$p(y_t|x_t) = \binom{n}{y_t} x_t^{y_t}(1 - x_t)^{n-y_t}, \quad y_t = 0, 1, \ldots, n.$$

a. Show that the observation equation with state variable transformed by the logit transformation $W_t = \ln(X_t/(1 - X_t))$ follows an exponential family

$$p(y_t|w_t) = \exp\{y_t w_t - b(w_t) + c(y_t)\}.$$

Determine the functions $b(\cdot)$ and $c(\cdot)$.

b. Suppose that the state X_t has the beta density

$$p(x_{t+1}|\mathbf{y}^{(t)}) = f(x_{t+1}; \alpha_{t+1|t}, \lambda_{t+1|t}),$$

where

$$f(x; \alpha, \lambda) = [B(\alpha, \lambda)]^{-1} x^{\alpha-1}(1 - x)^{\lambda-1}, \quad 0 < x < 1,$$

$B(\alpha, \lambda) := \Gamma(\alpha)\Gamma(\lambda)/\Gamma(\alpha + \lambda)$ is the beta function, and $\alpha, \lambda > 0$. Show that the posterior distribution of X_t given Y_t is also beta and express its parameters in terms of y_t and $\alpha_{t|t-1}, \lambda_{t|t-1}$.

c. Under the assumptions made in (b), show that $E\left(X_t|\mathbf{Y}^{(t)}\right) = E\left(X_{t+1}|\mathbf{Y}^{(t)}\right)$ and $\mathrm{Var}\left(X_t|\mathbf{Y}^{(t)}\right) < \mathrm{Var}\left(X_{t+1}|\mathbf{Y}^{(t)}\right)$.

d. Assuming that the parameters in (b) satisfy (9.8.41)–(9.8.42), show that the one-step prediction density $p\left(y_{t+1}|\mathbf{y}^{(t)}\right)$ is beta-binomial,

$$p(y_{t+1}|\mathbf{y}^{(t)}) = \frac{B(\alpha_{t+1|t} + y_{t+1}, \lambda_{t+1|t} + n - y_{t+1})}{(n + 1)B(y_{t+1} + 1, n - y_{t+1} + 1)B(\alpha_{t+1|t}, \lambda_{t+1|t})},$$

and verify that $\hat{Y}_{t+1}$ is given by (9.8.47).