3 Linear state space models

3.1 Introduction

The general linear Gaussian state space model can be written in a variety of ways; we shall use the form

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim \mathrm{N}(0, Q_t), \qquad t = 1, \ldots, n, \end{aligned} \tag{3.1}$$

where y_t is a $p \times 1$ vector of observations called the *observation vector* and α_t is an unobserved $m \times 1$ vector called the *state vector*. The idea underlying the model is that the development of the system over time is determined by α_t according to the second equation of (3.1), but because α_t cannot be observed directly we must base the analysis on observations y_t. The first equation of (3.1) is called the *observation equation* and the second is called the *state equation*. The matrices Z_t, T_t, R_t, H_t and Q_t are initially assumed to be known and the error terms ε_t and η_t are assumed to be serially independent and independent of each other at all time points. Matrices Z_t and T_{t-1} can be permitted to depend on $y_1, \ldots, y_{t-1}$. The initial state vector α_1 is assumed to be $\mathrm{N}(a_1, P_1)$ independently of $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_n$, where a_1 and P_1 are first assumed known; we will consider in Chapter 5 how to proceed in the absence of knowledge of a_1 and P_1. In practice, some or all of the matrices Z_t, H_t, T_t, R_t and Q_t will depend on elements of an unknown parameter vector ψ, the estimation of which will be considered in Chapter 7. The same model is used for a classical and a Bayesian analysis. The general linear state space model is the same as (3.1) except that the error densities are written as $\varepsilon_t \sim (0, H_t)$ and $\eta_t \sim (0, Q_t)$, that is, the normality assumption is dropped.

The first equation of (3.1) has the structure of a linear regression model where the coefficient vector α_t varies over time. The second equation represents a first order vector autoregressive model, the Markovian nature of which accounts for many of the elegant properties of the state space model. The local level model (2.3) considered in the last chapter is a simple special case of (3.1). In many applications R_t is the identity. In others, one could define $\eta_t^* = R_t\eta_t$ and $Q_t^* = R_tQ_tR_t'$ and proceed without explicit inclusion of R_t, thus making the model look simpler. However, if R_t is $m \times r$ with $r < m$ and Q_t is nonsingular, there is an obvious advantage in working with nonsingular η_t rather than singular η_t^*. We assume that R_t is a subset of the columns of I_m; in this case R_t is called a

selection matrix since it selects the rows of the state equation which have nonzero disturbance terms; however, much of the theory remains valid if R_t is a general $m \times r$ matrix.

Model (3.1) provides a powerful tool for the analysis of a wide range of problems. In this chapter we shall give substance to the general theory to be presented in Chapter 4 by describing a number of important applications of the model to problems in time series analysis and in spline smoothing analysis.

3.2 Univariate structural time series models

A *structural time series model* is one in which the trend, seasonal and error terms in the basic model (2.1), plus other relevant components, are modelled explicitly. In this section we shall consider structural models for the case where y_t is univariate; we shall extend this to the case where y_t is multivariate in Section 3.3. A detailed discussion of structural time series models, together with further references, has been given by Harvey (1989).

3.2.1 Trend component

The local level model considered in Chapter 2 is a simple form of a structural time series model. By adding a slope term ν_t, which is generated by a random walk, we obtain the model

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{N}\big(0, \sigma_\xi^2\big), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}\big(0, \sigma_\zeta^2\big). \end{aligned} \tag{3.2}$$

This is called the *local linear trend* model. If $\xi_t = \zeta_t = 0$ then $\nu_{t+1} = \nu_t = \nu$, say, and $\mu_{t+1} = \mu_t + \nu$ so the trend is exactly linear and (3.2) reduces to the deterministic linear trend plus noise model. The form (3.2) with $\sigma_\xi^2 > 0$ and $\sigma_\zeta^2 > 0$ allows the trend level and slope to vary over time.

Applied workers sometimes complain that the series of values of μ_t obtained by fitting this model does not look smooth enough to represent their idea of what a trend should look like. This objection can be met by setting $\sigma_\xi^2 = 0$ at the outset and fitting the model under this restriction. Essentially the same effect can be obtained by using in place of the second and third equation of (3.2) the model $\Delta^2 \mu_{t+1} = \zeta_t$, i.e. $\mu_{t+1} = 2\mu_t - \mu_{t-1} + \zeta_t$ where Δ is the first difference operator defined by $\Delta x_t = x_t - x_{t-1}$. This and its extension $\Delta^r \mu_t = \zeta_t$ for $r > 2$ have been advocated for modelling trend in state space models in a series of papers by Young and his collaborators under the name *integrated random walk* models; see, for example, Young, Lane, Ng and Palmer (1991). We see that (3.2) can be written in the form

$$y_t = (1 \quad 0)\begin{pmatrix}\mu_t \\ \nu_t\end{pmatrix} + \varepsilon_t,$$
$$\begin{pmatrix}\mu_{t+1} \\ \nu_{t+1}\end{pmatrix} = \begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}\begin{pmatrix}\mu_t \\ \nu_t\end{pmatrix} + \begin{pmatrix}\xi_t \\ \zeta_t\end{pmatrix},$$

which is a special case of (3.1).

3.2.2 Seasonal component

To model the seasonal term γ_t in (2.1), suppose there are s 'months' per 'year'. Thus for monthly data $s = 12$, for quarterly data $s = 4$ and for daily data, when modelling the weekly pattern, $s = 7$. If the seasonal pattern is constant over time, the seasonal values for months 1 to s can be modelled by the constants $\gamma_1^*, \ldots, \gamma_s^*$ where $\sum_{j=1}^{s} \gamma_j^* = 0$. For the jth 'month' in 'year' i we have $\gamma_t = \gamma_j^*$ where $t = s(i-1)+j$ for $i = 1, 2, \ldots$ and $j = 1, \ldots, s$. It follows that $\sum_{j=0}^{s-1} \gamma_{t+1-j} = 0$ so $\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j}$ with $t = s-1, s, \ldots$. In practice we often wish to allow the seasonal pattern to change over time. A simple way to achieve this is to add an error term ω_t to this relation giving the model

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim \mathrm{N}\big(0, \sigma_\omega^2\big), \tag{3.3}$$

for $t = 1, \ldots, n$ where initialisation at $t = 1, \ldots, s-1$ will be taken care of later by our general treatment of the initialisation question in Chapter 5. An alternative suggested by Harrison and Stevens (1976) is to denote the effect of season j at time t by γ_{jt} and then let γ_{jt} be generated by the quasi-random walk

$$\gamma_{j,t+1} = \gamma_{jt} + \omega_{jt}, \qquad t = (i-1)s + j, \qquad i = 1, 2, \ldots, \qquad j = 1, \ldots, s, \tag{3.4}$$

with an adjustment to ensure that each successive set of s seasonal components sums to zero; see Harvey (1989, §2.3.4) for details of the adjustment.

It is often preferable to express the seasonal in a trigonometric form, one version of which, for a constant seasonal, is

$$\gamma_t = \sum_{j=1}^{[s/2]} (\tilde{\gamma}_j \cos \lambda_j t + \tilde{\gamma}_j^* \sin \lambda_j t), \qquad \lambda_j = \frac{2\pi j}{s}, \qquad j = 1, \ldots, [s/2], \tag{3.5}$$

where $[a]$ is the largest integer $\leq a$ and where the quantities $\tilde{\gamma}_j$ and $\tilde{\gamma}_j^*$ are given constants. For a time-varying seasonal this can be made stochastic by replacing $\tilde{\gamma}_j$ and $\tilde{\gamma}_j^*$ by the random walks

$$\tilde{\gamma}_{j,t+1} = \tilde{\gamma}_{jt} + \tilde{\omega}_{jt}, \quad \tilde{\gamma}_{j,t+1}^* = \tilde{\gamma}_{jt}^* + \tilde{\omega}_{jt}^*, \quad j = 1, \ldots, [s/2], \quad t = 1, \ldots, n, \tag{3.6}$$

where $\tilde{\omega}_{jt}$ and $\tilde{\omega}^*_{jt}$ are independent $\mathrm{N}(0, \sigma^2_\omega)$ variables; for details see Young, Lane, Ng and Palmer (1991). An alternative trigonometric form is the quasi-random walk model

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt}, \tag{3.7}$$

where

$$\begin{aligned} \gamma_{j,t+1} &= \gamma_{jt} \cos \lambda_j + \gamma^*_{jt} \sin \lambda_j + \omega_{jt}, \\ \gamma^*_{j,t+1} &= -\gamma_{jt} \sin \lambda_j + \gamma^*_{jt} \cos \lambda_j + \omega^*_{jt}, \qquad j = 1, \ldots, [s/2], \end{aligned} \tag{3.8}$$

in which the ω_{jt} and ω^*_{jt} terms are independent $\mathrm{N}(0, \sigma^2_\omega)$ variables. We can show that when the stochastic terms in (3.8) are zero, the values of γ_t defined by (3.7) are periodic with period s by taking

$$\begin{aligned} \gamma_{jt} &= \tilde{\gamma}_j \cos \lambda_j t + \tilde{\gamma}^*_j \sin \lambda_j t, \\ \gamma^*_{jt} &= -\tilde{\gamma}_j \sin \lambda_j t + \tilde{\gamma}^*_j \cos \lambda_j t, \end{aligned}$$

which are easily shown to satisfy the deterministic part of (3.8). The required result follows since γ_t defined by (3.5) is periodic with period s. In effect, the deterministic part of (3.8) provides a recursion for (3.5).

The advantage of (3.7) over (3.6) is that the contributions of the errors ω_{jt} and ω^*_{jt} are not amplified in (3.7) by the trigonometric functions $\cos \lambda_j t$ and $\sin \lambda_j t$. We regard (3.3) as the main time domain model and (3.7) as the main frequency domain model for the seasonal component in structural time series analysis. A more detailed discussion of seasonal models is presented in Proietti (2000). In particular, he shows that the seasonal model in trigonometric form with specific variance restrictions for ω_{jt} and ω^*_{jt}, is equivalent to the quasi-random walk seasonal model (3.4).

3.2.3 Basic structural time series model

Each of the four seasonal models of the previous subsection can be combined with either of the trend models to give a structural time series model and all these can be put in the state space form (3.1). For example, for the local linear trend model (3.2) together with model (3.3) we have the observation equation

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{3.9}$$

To represent the model in state space form, we take the state vector as

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_t \quad \gamma_{t-1} \quad \ldots \quad \gamma_{t-s+2})',$$

and take the system matrices as

$$\begin{aligned} Z_t &= \left(Z_{[\mu]}, Z_{[\gamma]}\right), & T_t &= \operatorname{diag}\left(T_{[\mu]}, T_{[\gamma]}\right), \\ R_t &= \operatorname{diag}\left(R_{[\mu]}, R_{[\gamma]}\right), & Q_t &= \operatorname{diag}\left(Q_{[\mu]}, Q_{[\gamma]}\right), \end{aligned} \tag{3.10}$$

where

$$\begin{aligned} Z_{[\mu]} &= (1,0), & Z_{[\gamma]} &= (1,0,\ldots,0), \\ T_{[\mu]} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, & T_{[\gamma]} &= \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \\ R_{[\mu]} &= I_2, & R_{[\gamma]} &= (1,0,\ldots,0)', \\ Q_{[\mu]} &= \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}, & Q_{[\gamma]} &= \sigma_\omega^2. \end{aligned}$$

This model plays a prominent part in the approach of Harvey (1989) to structural time series analysis; he calls it the *basic structural time series model*. The state space form of this basic model with $s = 4$ is therefore

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_t \quad \gamma_{t-1} \quad \gamma_{t-2})',$$

$$Z_t = (1 \quad 0 \quad 1 \quad 0 \quad 0), \qquad T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$R_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 \\ 0 & 0 & \sigma_\omega^2 \end{bmatrix}.$$

Alternative seasonal specifications can also be used within the basic structural model. The Harrison and Stevens (1976) seasonal model referred to below (3.3) has the $(s+2) \times 1$ state vector

$$\alpha_t = (\ \mu_t \quad \nu_t \quad \gamma_t \quad \ldots \quad \gamma_{t-s+1}\ )',$$

where the relevant parts of the system matrices for substitution in (3.10) are given by

$$\begin{aligned} Z_{[\gamma]} &= (1,0,\ldots,0), & T_{[\gamma]} &= \begin{bmatrix} 0 & I_{s-1} \\ 1 & 0 \end{bmatrix}, \\ R_{[\gamma]} &= I_s, & Q_{[\gamma]} &= \sigma_\omega^2 (I_s - \mathbf{1}\mathbf{1}'/s), \end{aligned}$$

in which $\mathbf{1}$ is an $s \times 1$ vector of ones, $\omega_{1t} + \cdots + \omega_{s,t} = 0$ and variance matrix $Q_{[\gamma]}$ has rank $s - 1$.

The seasonal component in trigonometric form (3.8) can be incorporated in the basic structural model with the $(s + 1) \times 1$ state vector

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_{1t} \quad \gamma_{1t}^* \quad \gamma_{2t} \quad \ldots)',$$

and the relevant parts of the system matrices given by

$$\begin{aligned} Z_{[\gamma]} &= (1, 0, 1, 0, 1, \ldots, 1, 0, 1), & T_{[\gamma]} &= \operatorname{diag}(C_1, \quad \ldots, \quad C_{s^*}, \quad -1), \\ R_{[\gamma]} &= I_{s-1}, & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1}. \end{aligned}$$

When we assume that s is even, we have $s^* = s \,/\, 2$ and

$$C_j = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix}, \qquad \lambda_j = \frac{2\pi j}{s}, \qquad j = 1, \ldots, s^*. \tag{3.11}$$

When s is odd, we have $s^* = (s - 1) \,/\, 2$ and

$$\begin{aligned} Z_{[\gamma]} &= (1, 0, 1, 0, 1, \ldots, 1, 0), & T_{[\gamma]} &= \operatorname{diag}(\, C_1, \quad \ldots, \quad C_{s^*} \,), \\ R_{[\gamma]} &= I_{s-1}, & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1}. \end{aligned}$$

where C_j is defined in (3.11) for $j = 1, \ldots, s^*$.

3.2.4 Cycle component

Another important component in some time series is the *cycle* c_t which we can introduce by extending the basic time series model (2.2) to

$$y_t = \mu_t + \gamma_t + c_t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{3.12}$$

In its simplest form c_t is a pure sine wave generated by the relation

$$c_t = \tilde{c} \cos\lambda_c t + \tilde{c}^* \sin\lambda_c t,$$

where λ_c is the frequency of the cycle; the period is $2\pi/\lambda_c$ which is normally substantially greater that the seasonal period s. As with the seasonal, we can allow the cycle to change stochastically over time by means of the relations analogous to (3.8)

$$\begin{aligned} c_{t+1} &= c_t \cos\lambda_c + c_t^* \sin\lambda_c + \tilde{\omega}_t, \\ c_{t+1}^* &= -c_t \sin\lambda_c + c_t^* \cos\lambda_c + \tilde{\omega}_t^*, \end{aligned}$$

where $\tilde{w}_t$ and $\tilde{w}_t^*$ are independent $\mathrm{N}(0, \sigma_{\tilde{w}}^2)$ variables. Cycles of this form fit naturally into the structural time series model framework. The frequency λ_c can be treated as an unknown parameter to be estimated.

The state space representation for the cycle component is similar to a single trigonometric seasonal component but with frequency λ_c. The relevant system matrices for the cycle component are therefore given by

$$\begin{aligned} Z_{[c]} &= (1, 0), & T_{[c]} &= C_c, \\ R_{[c]} &= I_2, & Q_{[c]} &= \sigma_{\tilde{\omega}}^2 I_2, \end{aligned}$$

where the 2×2 matrix C_c is defined as C_j is in (3.11) but with $\lambda_j = \lambda_c$.

In economic time series the cycle component is usually associated with the business cycle. The definition of a business cycle from Burns and Mitchell (1946, p. 3) is typically adopted: 'A cycle consists of expansions occurring at about the same time in many economic activities, followed by similar general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own.' To let our cycle component resemble this definition, we allow its period $2\pi \,/\, \lambda_c$ to range between 1.5 and 12 years and we specify the cycle as a stationary stochastic process. The system matrix C_c for an economic analysis is then given by

$$C_c = \rho_c \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix}, \qquad 1.5 \leq 2\pi \,/\, \lambda_c \leq 12,$$

with damping factor $0 < \rho_c < 1$. We can now define the economic cycle component as

$$c_t = Z_{[c]} \begin{pmatrix} c_t \\ c_t^* \end{pmatrix}, \quad \begin{pmatrix} c_{t+1} \\ c_{t+1}^* \end{pmatrix} = C_c \begin{pmatrix} c_t \\ c_t^* \end{pmatrix} + \begin{pmatrix} \tilde{\omega}_t \\ \tilde{\omega}_t^* \end{pmatrix}, \quad \begin{pmatrix} \tilde{\omega}_t \\ \tilde{\omega}_t^* \end{pmatrix} \sim \mathrm{N}(0, Q_{[c]}), \tag{3.13}$$

for $t = 1, \ldots, n$. When we fit a model with this cycle component to a macroeconomic time series with values for ρ_c and λ_c within their admissable regions, it is justified to interpret the cycle c_t as a business cycle component.

3.2.5 Explanatory variables and intervention effects

Explanatory variables and intervention effects are easily allowed for in the structural model framework. Suppose we have k regressors $x_{1t}, \ldots, x_{kt}$ with regression coefficients $\beta_1, \ldots, \beta_k$ which are constant over time and that we also wish to measure the change in level due to an intervention at time τ. We define an *intervention variable* w_t as follows:

$$\begin{aligned} w_t &= 0, & t &< \tau, \\ &= 1, & t &\geq \tau. \end{aligned}$$

Adding these to the model (3.12) gives

$$y_t = \mu_t + \gamma_t + c_t + \sum_{j=1}^{k} \beta_j x_{jt} + \delta w_t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{3.14}$$

We see that δ measures the change in the level of the series at a known time τ due to an intervention at time τ. The resulting model can readily be put into state space form. For example, if $\gamma_t = c_t = \delta = 0$, $k = 1$ and if μ_t is determined by a local level model, we can take

$$\begin{aligned} \alpha_t &= (\mu_t \quad \beta_{1t})', & Z_t &= (1 \quad x_{1t}), \\ T_t &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & R_t &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad Q_t = \sigma_\xi^2, \end{aligned}$$

in (3.1). Here, although we have attached a suffix t to β_1 it is made to satisfy $\beta_{1,t+1} = \beta_{1t}$ so it is constant. Other examples of intervention variables are the *pulse intervention variable* defined by

$$\begin{aligned} w_t &= 0, \qquad t < \tau, \qquad t > \tau, \\ &= 1, \qquad t = \tau, \end{aligned}$$

and the *slope intervention variable* defined by

$$\begin{aligned} w_t &= 0, & t &< \tau, \\ &= 1 + t - \tau, & t &\geq \tau. \end{aligned}$$

For other forms of intervention variable designed to represent a more gradual change of level or a transient change see Box and Tiao (1975). Coefficients such as δ which do not change over time can be incorporated into the state vector by setting the corresponding state errors equal to zero. Regression coefficients β_{jt} which change over time can be handled straightforwardly in the state space framework by modelling them by random walks of the form

$$\beta_{j,t+1} = \beta_{jt} + \chi_{jt}, \qquad \chi_{jt} \sim \mathrm{N}\big(0, \sigma_\chi^2\big), \qquad j = 1, \ldots, k. \tag{3.15}$$

An example of the use of model (3.14) for intervention analysis is given by Harvey and Durbin (1986) who used it to measure the effect of the British seat belt law on road traffic casualities. Of course, if the cycle term, the regression term or the intervention term are not required, they can be omitted from (3.14). Instead of including regression and intervention coefficients in the state vector, an alternative way of dealing with them is to concentrate them out of the likelihood function and estimate them via regression, as we will show in Subsection 6.2.3.

3.2.6 STAMP

A wide-ranging discussion of structural time series models can be found in Harvey (1989). Supplementary sources for further applications and later work are Harvey and Shephard (1993), Harvey (2006), and Harvey and Koopman (2009). The computer package *STAMP* 8.3 of Koopman, Harvey, Doornik and Shephard (2010) is designed to analyse, model and forecast time series based on univariate and multivariate structural time series models. The package has implemented the Kalman filter and associated algorithms leaving the user free to concentrate on the important part of formulating a model. *STAMP* is a commercial product and more information on it can be obtained from the Internet at

`http://stamp-software.com/`

3.3 Multivariate structural time series models

The methodology of structural time series models lends itself easily to generalisation to multivariate time series. Consider the local level model for a $p \times 1$ vector of observations y_t, that is

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \eta_t, \end{aligned} \tag{3.16}$$

where μ_t, ε_t and η_t are $p \times 1$ vectors and

$$\varepsilon_t \sim \mathrm{N}(0, \Sigma_\varepsilon), \qquad \eta_t \sim \mathrm{N}(0, \Sigma_\eta),$$

with $p \times p$ variance matrices Σ_ε and Σ_η. In this so-called *seemingly unrelated time series equations* model, each series in y_t is modelled as in the univariate case, but the disturbances may be correlated instantaneously across series. In the case of a model with other components such as slope, cycle and seasonal, the disturbances associated with the components become vectors which have $p \times p$ variance matrices. The link across the p different time series is through the correlations of the disturbances driving the components.

3.3.1 Homogeneous models

A seemingly unrelated time series equations model is said to be *homogeneous* when the variance matrices associated with the different disturbances are proportional to each other. For example, the homogeneity restriction for the multivariate local level model is

$$\Sigma_\eta = q\Sigma_\varepsilon,$$

where scalar q is the signal-to-noise ratio. This means that all the series in y_t, and linear combinations thereof, have the same dynamic properties which implies that they have the same autocorrelation function for the stationary form of the model. A homogeneous model is a rather restricted model but it is easy to estimate. For further details we refer to Harvey (1989, Chapter 8).

3.3.2 Common levels

Consider the multivariate local level model without the homogeneity restriction but with the assumption that the rank of Σ_η is $r < p$. The model then contains only r underlying level components. We may refer to these as *common levels*. Recognition of such common factors yields models which may not only have an interesting interpretation, but may also provide more efficient inferences and forecasts. With an appropriate ordering of the series the model may be written as

$$y_t = a + A\mu_t^* + \varepsilon_t,$$
$$\mu_{t+1}^* = \mu_t^* + \eta_t^*,$$

where μ_t^* and η_t^* are $r \times 1$ vectors, a is a $p \times 1$ vector and A is a $p \times r$ matrix. We further assume that

$$a = \begin{pmatrix} 0 \\ a^* \end{pmatrix}, \qquad A = \begin{bmatrix} I_r \\ A^* \end{bmatrix}, \qquad \eta_t^* \sim \mathrm{N}(0, \Sigma_\eta^*),$$

where a^* is a $(p-r) \times 1$ vector and A^* is a $(p-r) \times r$ matrix of nonzero values and where variance matrix Σ_η^* is a $r \times r$ positive definite matrix. The matrix A may be interpreted as a factor loading matrix. When there is more than one common factor ($r > 1$), the factor loadings are not unique. A factor rotation may give components with a more interesting interpretation.

The introduction of common factors can also be extended to other multivariate components such as slope, cycle and seasonal. For example, an illustration of a common business cycle component in a model for a vector of economic time series is given by Valle e Azevedo, Koopman and Rua (2006). Further discussions of multivariate extensions of structural time series models are given by Harvey (1989, Chapter 8) and, Harvey and Koopman (1997) and Koopman, Ooms and Hindrayanto (2009).

3.3.3 Latent risk model

Risk is at the centre of many policy decisions in companies, governments and financial institutions. In risk analysis, measures of exposure to risk, outcomes (or events), and, possibly, losses are analysed. Risk itself cannot be observed. For example, in the case of road safety research, exposure is the number of cars (or the number of kilometres travelled), outcome is the number of accidents and loss is the cost of damages (or the number of fatalities). From these measures we can learn about risk and its associated severity. In a time series context, we typically observe the measures in totals for a country (or a region, or a specific group) and for a specific period (month, quarter or other). The time series for exposure y_{1t}, outcome y_{2t} and loss y_{3t} are typically observed with error and are subject to, possibly, trend, seasonal, cycle and regression effects.

We can carry out a time series analysis with direct interpretable variables via the multiplicative latent risk model

$$y_{1t} = \mu_{1t} \times \xi_{1t}, \qquad y_{2t} = \mu_{1t} \times \mu_{2t} \times \xi_{2t}, \qquad y_{3t} = \mu_{1t} \times \mu_{2t} \times \mu_{3t} \times \xi_{3t},$$

where the unobserved components μ_{1t} is exposure corrected for observation error, μ_{2t} is risk and μ_{3t} is severity while ξ_{jt} are the multiplicative errors with their means equal to one, for $j = 1, 2, 3$. The model implies that expected outcome is exposure times risk while expected loss is outcome times severity. By taking logs we obtain the trivariate latent risk model in additive form

$$\begin{pmatrix} \log y_{1t} \\ \log y_{2t} \\ \log y_{3t} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \theta_t + \varepsilon_t,$$

with the signal vector $\theta_t = (\log \mu_{1t}\,, \log \mu_{2t}\,, \log \mu_{3t})'$ and with the disturbance vector $\varepsilon_t = (\log \xi_{1t}\,, \log \xi_{2t}\,, \log \xi_{3t})'$. The signals for exposure, risk and severity (all three variables in logs) can be modelled simultaneously as linear functions of the state vector. The signal vector for our purpose can be given by

$$\theta_t = S_t \alpha_t, \qquad t = 1, \ldots, n,$$

where S_t is a $3 \times m$ system matrix that relates the signal θ_t with the state vector α_t. The model is placed in state space form (3.1) with

$$Z_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} S_t.$$

Multiple measures for exposure, outcome and/or loss can be incorporated in this framework in a straightforward manner. Typical applications of the latent risk model are for studies on insurance claims, credit card purchases and road safety. Bijleveld, Commandeur, Gould and Koopman (2008) provide further discussions on the latent risk model and show how the general methodology can be effectively used in the assessment of risk.

3.4 ARMA models and ARIMA models

Autoregressive integrated moving average (ARIMA) time series models were employed for this purpose by Box and Jenkins in their pathbreaking (1970) book; see Box, Jenkins and Reinsel (1994) for the current version of this book. As with structural time series models considered in Section 3.2, Box and Jenkins typically regarded a univariate time series y_t as made up of trend, seasonal and irregular components. However, instead of modelling the various components separately, their idea was to eliminate the trend and seasonal by differencing at the outset of the analysis. The resulting differenced series are treated as a stationary time series, that is, a series where characteristic properties such as means, covariances and so on remain invariant under translation through time. Let $\Delta y_t = y_t - y_{t-1}$, $\Delta^2 y_t = \Delta(\Delta y_t)$, $\Delta_s y_t = y_t - y_{t-s}$, $\Delta_s^2 y_t = \Delta_s(\Delta_s y_t)$, and so on, where we

are assuming that we have s 'months' per 'year'. Box and Jenkins suggest that differencing is continued until trend and seasonal effects have been eliminated, giving a new variable $y_t^* = \Delta^d \Delta_s^D y_t$ for $d, D = 0, 1, \ldots$, which we model as a stationary autoregressive moving average ARMA(p, q) model given by

$$y_t^* = \phi_1 y_{t-1}^* + \cdots + \phi_p y_{t-p}^* + \zeta_t + \theta_1 \zeta_{t-1} + \cdots + \theta_q \zeta_{t-q}, \qquad \zeta_t \sim \mathrm{N}\big(0, \sigma_\zeta^2\big), \tag{3.17}$$

with non-negative integers p and q and where ζ_t is a serially independent series of $\mathrm{N}(0, \sigma_\zeta^2)$ disturbances. This can be written in the form

$$y_t^* = \sum_{j=1}^{r} \phi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^{r-1} \theta_j \zeta_{t-j}, \qquad t = 1, \ldots, n, \tag{3.18}$$

where $r = \max(p, q+1)$ and for which some coefficients are zero. Box and Jenkins normally included a constant term in (3.18) but for simplicity we omit this; the modifications needed to include it are straightforward. We use the symbols d, p and q here and elsewhere in their familiar ARIMA context without prejudice to their use in different contexts in other parts of the book.

We now demonstrate how to put these models into state space form, beginning with the case where $d = D = 0$, that is, no differencing is needed, so we can model the series by (3.18) with y_t^* replaced by y_t. Take

$$\begin{aligned} Z_t &= (1 \quad 0 \quad 0 \quad \cdots \quad 0), \\ \alpha_t &= \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_r y_{t-r+1} + \theta_1 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+2} \\ \phi_3 y_{t-1} + \cdots + \phi_r y_{t-r+2} + \theta_2 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \theta_{r-1} \zeta_t \end{pmatrix}, \end{aligned} \tag{3.19}$$

and write the state equation for α_{t+1} as in (3.1) with

$$T_t = T = \begin{bmatrix} \phi_1 & 1 & & 0 \\ \vdots & & \ddots & \\ \phi_{r-1} & 0 & & 1 \\ \phi_r & 0 & \cdots & 0 \end{bmatrix}, \qquad R_t = R = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}, \qquad \eta_t = \zeta_{t+1}. \tag{3.20}$$

This, together with the observation equation $y_t = Z_t \alpha_t$, is equivalent to (3.18) but is now in the state space form (3.1) with $\varepsilon_t = 0$, implying that $H_t = 0$. For example, with $r = 2$ we have the state equation

$$\begin{pmatrix} y_{t+1} \\ \phi_2 y_t + \theta_1 \zeta_{t+1} \end{pmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta_1 \zeta_t \end{pmatrix} + \begin{pmatrix} 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1}.$$

The form given is not the only state space version of an ARMA model but is a convenient one.

We now consider the case of a univariate nonseasonal nonstationary ARIMA model of order p, d and q, with $d > 0$, given by (3.17) with $y_t^* = \Delta^d y_t$. As an example, we first consider the state space form of the ARIMA model with $p = 2$, $d = 1$ and $q = 1$ which is given by

$$y_t = (1 \quad 1 \quad 0)\alpha_t,$$
$$\alpha_{t+1} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \phi_1 & 1 \\ 0 & \phi_2 & 0 \end{bmatrix} \alpha_t + \begin{pmatrix} 0 \\ 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1},$$

with the state vector defined as

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \zeta_t \end{pmatrix},$$

and $y_t^* = \Delta y_t = y_t - y_{t-1}$. This example generalises easily to ARIMA models with $d = 1$ with other values for p and q. The ARIMA model with $p = 2$, $d = 2$ and $q = 1$ in state space form is given by

$$y_t = (1 \quad 1 \quad 1 \quad 0)\alpha_t,$$
$$\alpha_{t+1} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & \phi_1 & 1 \\ 0 & 0 & \phi_2 & 0 \end{bmatrix} \alpha_t + \begin{pmatrix} 0 \\ 0 \\ 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1},$$

with

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ \Delta y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \zeta_t \end{pmatrix},$$

and $y_t^* = \Delta^2 y_t = \Delta(y_t - y_{t-1})$. The relations between y_t, Δy_t and $\Delta^2 y_t$ follow immediately since

$$\Delta y_t = \Delta^2 y_t + \Delta y_{t-1},$$
$$y_t = \Delta y_t + y_{t-1} = \Delta^2 y_t + \Delta y_{t-1} + y_{t-1}.$$

We deal with the unknown nonstationary values y_0 and Δy_0 in the initial state vector α_1 in Subsection 5.6.3 where we describe the initialisation procedure for filtering and smoothing. Instead of estimating y_0 and Δy_0 directly, we treat these

elements of α_1 as diffuse random elements while the other elements, including y_t^*, are stationary which have proper unconditional means and variances. The need to facilitate the initialisation procedure explains why we set up the state space model in this form. The state space forms for ARIMA models with other values for p^*, d and q^* can be represented in similar ways. The advantage of the state space formulation is that the array of techniques that have been developed for state space models are made available for ARMA and ARIMA models. In particular, techniques for exact maximum likelihood estimation and for initialisation are available.

As indicated above, for seasonal series both trend and seasonal are eliminated by the differencing operation $y_t^* = \Delta^d \Delta_s^D y_t$ prior to modelling y_t^* by a stationary ARMA model of the form (3.18). The resulting model for y_t^* can be put into state space form by a straightforward extension of the above treatment. A well-known seasonal ARIMA model is the so-called *airline model* which is given by

$$y_t^* = \Delta\Delta_{12} y_t = \zeta_t - \theta_1 \zeta_{t-1} - \theta_{12}\zeta_{t-12} + \theta_1\theta_{12}\zeta_{t-13}, \tag{3.21}$$

which has a standard ARIMA state space representation.

It is interesting to note that for many state space models an inverse relation holds in the sense that the state space model has an ARIMA representation. For example, if second differences are taken in the local linear trend model (3.2), the terms in μ_t and ν_t disappear and we obtain

$$\Delta^2 y_t = \varepsilon_{t+2} - 2\varepsilon_{t+1} + \varepsilon_t + \xi_{t+1} - \xi_t + \zeta_t.$$

Since the first two autocorrelations of this are nonzero and the rest are zero, we can write it as a moving average series $\zeta_t^* + \theta_1\zeta_{t-1}^* + \theta_2\zeta_{t-2}^*$ where θ_1 and θ_2 are the moving average parameters and the ζ_t^*'s are independent $\mathrm{N}(0, \sigma_{\zeta^*}^2)$ disturbances. In Box and Jenkins' notation this is an ARIMA(0,2,2) model. We obtain the representation

$$\Delta^2 y_t = \zeta_t^* + \theta_1\zeta_{t-1}^* + \theta_2\zeta_{t-2}^*, \tag{3.22}$$

where the local linear trend model imposes a more restricted space on θ_1 and θ_2 than is required for a non-invertible ARIMA(0,2,2) model. A more elaborate discussion on these issues is given by Harvey (1989, §2.5.3).

It is important to recognise that the model (3.22) is less informative than (3.2) since it has lost the information that exists in the form (3.2) about the level μ_t and the slope ν_t. If a seasonal term generated by model (3.3) is added to the local linear trend model, the corresponding ARIMA model has the form

$$\Delta^2\Delta_s y_t = \zeta_t^* + \sum_{j=1}^{s+2} \theta_j \zeta_{t-j}^*,$$

where $\theta_1, \ldots, \theta_{s+2}$ are determined by the four variances σ_ε^2, σ_ξ^2, σ_ζ^2 and σ_ω^2. In this model, information about the seasonal is lost as well as information about the trend. The fact that structural time series models provide explicit information about trend and seasonal, whereas ARIMA models do not, is an important advantage that the structural modelling approach has over ARIMA modelling. We shall make a detailed comparison of the two approaches to time series analysis in Subsection 3.10.1.

3.5 Exponential smoothing

In this section we consider the development of exponential smoothing methods in the 1950s and we examine their relation to simple forms of state space and Box–Jenkins models. These methods have been primarily developed for the purpose of forecasting. The term 'smoothing' is used in a somewhat different context in this section and should not be related to the term 'smoothing' as it is used elsewhere in the book, notably in Chapters 2 and 4.

Let us start with the introduction in the 1950s of the exponentially weighted moving average (EWMA) for one-step ahead forecasting of y_{t+1} given a univariate time series $y_t, y_{t-1}, \ldots$. This has the form

$$\hat{y}_{t+1} = (1-\lambda)\sum_{j=0}^{\infty} \lambda^j y_{t-j}, \qquad 0 < \lambda < 1. \tag{3.23}$$

From (3.23) we deduce immediately the recursion

$$\hat{y}_{t+1} = (1-\lambda)y_t + \lambda \hat{y}_t, \tag{3.24}$$

which is used in place of (3.23) for practical computation. This has a simple structure and requires little storage so it was very convenient for the primitive computers available in the 1950s. As a result, EWMA forecasting became very popular in industry, particularly for sales forecasting of many items simultaneously. We call the operation of calculating forecasts by (3.24) *exponential smoothing*.

Denote the one-step ahead forecast error $y_t - \hat{y}_t$ by u_t and substitute in (3.24) with t replaced by $t-1$; this gives

$$y_t - u_t = (1-\lambda)y_{t-1} + \lambda(y_{t-1} - u_{t-1}),$$

that is,

$$\Delta y_t = u_t - \lambda u_{t-1}. \tag{3.25}$$

Taking u_t to be a series of independent $\mathrm{N}(0, \sigma_u^2)$ variables, we see that we have deduced from the EWMA recursion (3.24) the simple ARIMA model (3.25).

An important contribution was made by Muth (1960) who showed that EWMA forecasts produced by the recursion (3.24) are minimum mean square

error forecasts in the sense that they minimise $\mathrm{E}(\hat{y}_{t+1} - y_{t+1})^2$ for observations $y_t, y_{t-1}, \ldots$ generated by the local level model (2.3), which for convenience we write in the form

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \end{aligned} \tag{3.26}$$

where ε_t and ξ_t are serially independent random variables with zero means and constant variances. Taking first differences of observations y_t generated by (3.26) gives

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t - \varepsilon_{t-1} + \xi_{t-1}.$$

Since ε_t and ξ_t are serially uncorrelated the autocorrelation coefficient of the first lag for Δy_t is nonzero but all higher autocorrelations are zero. This is the autocorrelation function of a moving average model of order one which, with λ suitably defined, we can write in the form

$$\Delta y_t = u_t - \lambda u_{t-1},$$

which is the same as model (3.25).

We observe the interesting point that these two simple forms of state space and ARIMA models produce the same one-step ahead forecasts and that these can be calculated by the EWMA (3.24) which has proven practical value. We can write this in the form

$$\hat{y}_{t+1} = \hat{y}_t + (1 - \lambda)(y_t - \hat{y}_t),$$

which is the Kalman filter for the simple state space model (3.26).

The EWMA was extended by Holt (1957) and Winters (1960) to series containing trend and seasonal. The extension for trend in the additive case is

$$\hat{y}_{t+1} = m_t + b_t,$$

where m_t and b_t are level and slope terms generated by the EWMA type recursions

$$\begin{aligned} m_t &= (1 - \lambda_1) y_t + \lambda_1 (m_{t-1} + b_{t-1}), \\ b_t &= (1 - \lambda_2)(m_t - m_{t-1}) + \lambda_2 b_{t-1}. \end{aligned}$$

In an interesting extension of the results of Muth (1960), Theil and Wage (1964) showed that the forecasts produced by these Holt–Winters recursions are minimum mean square error forecasts for the state space model

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \\ \nu_{t+1} &= \nu_t + \zeta_t, \end{aligned} \tag{3.27}$$

which is the local linear trend model (3.2). Taking second differences of y_t generated by (3.27), we obtain

$$\Delta^2 y_t = \zeta_{t-2} + \xi_{t-1} - \xi_{t-2} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}.$$

This is a stationary series with nonzero autocorrelations at lags 1 and 2 but zero autocorrelations elsewhere. It therefore follows the moving average model

$$\Delta^2 y_t = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2},$$

which is a simple form of ARIMA model.

Adding the seasonal term $\gamma_{t+1} = -\gamma_t - \cdots - \gamma_{t-s+2} + \omega_t$ from (3.3) to the measurement equation of (3.26) gives the model

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \\ \gamma_{t+1} &= -\gamma_t - \cdots - \gamma_{t-s+2} + \omega_t, \end{aligned} \tag{3.28}$$

which is a special case of the structural time series models of Section 3.2. Now take first differences and first seasonal differences of (3.28). We find

$$\Delta\Delta_s y_t = \xi_{t-1} - \xi_{t-s-1} + \omega_{t-1} - 2\omega_{t-2} + \omega_{t-3} + \varepsilon_t - \varepsilon_{t-1} - \varepsilon_{t-s} + \varepsilon_{t-s-1}, \tag{3.29}$$

which is a stationary time series with nonzero autocorrelations at lags 1, 2, $s-1$, s and $s+1$. Consider the airline model (3.21) for general s,

$$\Delta\Delta_s y_t = u_t - \theta_1 u_{t-1} - \theta_s u_{t-s} - \theta_1\theta_s u_{t-s-1},$$

which has been found to fit well many economic time series containing trend and seasonal. It has nonzero autocorrelations at lags 1, $s-1$, s and $s+1$. Now the autocorrelation at lag 2 from model (3.28) arises only from $\mathrm{Var}(\omega_t)$ which in most cases in practice is small. Thus when we add a seasonal component to the models we find again a close correspondence between state space and ARIMA models. A slope component ν_t can be added to (3.28) as in (3.27) without significantly affecting the conclusions.

A pattern is now emerging. Starting with EWMA forecasting, which in appropriate circumstances has been found to work well in practice, we have found that there are two distinct types of models, the state space models and the Box–Jenkins ARIMA models which appear to be very different conceptually but which both give minimum mean square error forecasts from EWMA recursions. The explanation is that when the time series has an underlying structure which is sufficiently simple, then the appropriate state space and ARIMA models are essentially equivalent. It is when we move towards more complex structures that the differences emerge. The above discussion has been based on Durbin (2000b, §3).

3.6 Regression models

The regression model for a univariate series y_t is given by

$$y_t = X_t\beta + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{3.30}$$

for $t = 1, \ldots, n$, where X_t is the $1 \times k$ regressor vector with exogenous variables, β is the $k \times 1$ vector of regression coefficients and H_t is the known variance that possibly varies with t. This model can be represented in the state space form (3.1) with $Z_t = X_t$, $T_t = I_k$ and $R_t = Q_t = 0$, so that $\alpha_t = \alpha_1 = \beta$. The generalised least squares estimator of the regression coefficient vector β is given by

$$\hat{\beta} = \left(\sum_{t=1}^{n} X_t' H_t^{-1} X_t \right)^{-1} \sum_{t=1}^{n} X_t' H_t^{-1} y_t.$$

When the Kalman filter is applied to the state space model that represents the regression model (3.30), it effectively computes $\hat{\beta}$ in a recursive manner. In this case the Kalman filter reduces to the recursive least squares method as developed by Plackett (1950).

3.6.1 Regression with time-varying coefficients

Suppose that in the linear regression model (3.30) we wish the coefficient vector β to vary over time. A suitable model for this is to replace β in (3.30) by α_t and to permit each coefficient α_{it} to vary according to a random walk $\alpha_{i,t+1} = \alpha_{it} + \eta_{it}$. This gives a state equation for the vector α_t in the form $\alpha_{t+1} = \alpha_t + \eta_t$. Since the model is a special case of (3.1) with $Z_t = X_t$, $T_t = R_t = I_k$ and Q_t is the known (diagonal) variance matrix for η_t, it can be handled in a routine fashion by Kalman filter and smoothing techniques.

3.6.2 Regression with ARMA errors

Consider a regression model of the form

$$y_t = X_t\beta + \xi_t, \qquad t = 1, \ldots, n, \tag{3.31}$$

where y_t is a univariate dependent variable, X_t is a $1 \times k$ regressor vector, β is its coefficient vector and ξ_t denotes the error which is assumed to follow an ARMA model of form (3.18); this ARMA model may or may not be stationary and some of the coefficients ϕ_j, θ_j may be zero as long as ϕ_r and θ_{r-1} are not both zero. Let α_t be defined as in (3.19) and let

$$\alpha_t^* = \begin{pmatrix} \beta_t \\ \alpha_t \end{pmatrix},$$

where $\beta_t = \beta$. Writing the state equation implied by (3.20) as $\alpha_{t+1} = T\alpha_t + R\eta_t$, let

$$T^* = \begin{bmatrix} I_k & 0 \\ 0 & T \end{bmatrix}, \qquad R^* = \begin{bmatrix} 0 \\ R \end{bmatrix}, \qquad Z_t^* = (X_t \quad 1 \quad 0 \quad \cdots \quad 0),$$

where T and R are defined in (3.20). Then the model

$$y_t = Z_t^* \alpha_t^*, \qquad \alpha_{t+1}^* = T^* \alpha_t^* + R^* \eta_t,$$

is in state space form (3.1) so Kalman filter and smoothing techniques are applicable; these provide an efficient means of fitting model (3.31). It is evident that the treatment can easily be extended to the case where the regression coefficients are determined by random walks as in Subsection 3.6.1. Moreover, with this approach, unlike some others, it is not necessary for the ARMA model used for the errors to be stationary.

3.7 Dynamic factor models

Principal components and factor analysis are widely used statistical methods in the applied social and behavourial sciences. These methods aim to identify commonalities in the covariance structure of high-dimensional data sets. A factor analysis model can be based on the form given by Lawley and Maxwell (1971), that is

$$y_i = \Lambda f_i + u_i, \qquad f_i \sim \mathrm{N}(0, \Sigma_f), \qquad u_i \sim \mathrm{N}(0, \Sigma_u),$$

where y_i represents a zero-mean data vector containing measured characteristics of subject i, for $i = 1, \ldots, n$, while Λ is the coefficient matrix for the low-dimensional vector f_i of latent variables. The latent vector f_i with its variance matrix Σ_f and the disturbance vector u_j with its variance matrix Σ_u are assumed to be mutually and serially independent for $i, j = 1, \ldots, n$. The factor analysis model implies the decomposition of the variance matrix Σ_y of y_i into

$$\Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_u.$$

Estimation of Λ, Σ_f and Σ_u can be carried out by maximum likelihood procedures; the loading coefficients in Λ and the variance matrices are subject to a set of linear restrictions necessary for identification. A detailed discussion of the maximum likelihood approach to factor analysis is given by Lawley and Maxwell (1971).

In the application of factor analysis in a time series context, the measurements in y_i correspond to a time period t rather than a subject i, we therefore have y_t instead of y_i. The time dependence of the measurements can be accounted for by replacing the serially independence assumption for f_t by a serial dependence

assumption. For example, we can assume that f_t is modelled by a vector autoregressive process. We can also let f_t depend on the state vector α_t in a linear way, that is, $f_t = U_t\alpha_t$ where U_t is typically a known selection matrix.

In applications to economics, y_t may consist of a large set of macroeconomic indicators associated with variables such as income, consumption, investment and unemployment. These variables are all subject to economic activity that can often be related to the business cycle. The dynamic features of the business cycle may be disentangled into a set of factors with possibly different dynamic characteristics. In the context of finance, y_t may consist of a large set of daily prices or returns from individual stocks that make up the indices such as Standard & Poor's 500 and Dow Jones. A set of underlying factors may represent returns from particular porfolio strategies. In the context of marketing, y_t may contain market shares of groups or sub-groups of product brands. The factors may indicate whether marketing strategies in certain periods affect all market shares or only a selection of market shares. In all these cases it is not realistic to assume that the factors are independent over time so we are required to formulate a dynamic process for the factors. The wide variety of applications of factor analysis in a time series context have led to many contributions in the statistics and econometrics literature on the inference of dynamic factor models; see, for example, Geweke (1977), Engle and Watson (1981), Watson and Engle (1983), Litterman and Scheinkman (1991), Quah and Sargent (1993), Stock and Watson (2002), Diebold, Rudebusch and Aruoba (2006) and Doz, Giannone and Reichlin (2012).

The dynamic factor model can be regarded as a state space model in which the state vector consists of latent factors where dynamic properties are formulated in the state equation of the state space model (3.1). The size of the observation vector is typically large while the dimension of the state vector is small so we have $p >> m$. In the example

$$y_t = \Lambda f_t + \varepsilon_t, \qquad f_t = U_t\alpha_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{3.32}$$

the dynamic factor model is a special case of the state space model (3.1) for which the observation equation has $Z_t = \Lambda U_t$. In Chapter 6 we will discuss modifications of the general statistical treatment of the state space model in cases where $p >> m$. These modifications are specifically relevant for the inference of dynamic factor analysis and they ensure feasible methods for a state space analysis.

3.8 State space models in continuous time

In contrast to all the models that we have considered so far, suppose that the observation $y(t)$ is a continuous function of time for t in an interval which we take to be $0 \leq t \leq T$. We shall aim at constructing state space models for $y(t)$ which are the analogues in continuous time for models that we have already

studied in discrete time. Such models are useful not only for studying phenomena which genuinely operate in continuous time, but also for providing a convenient theoretical base for situations where the observations take place at time points $t_1 \leq \cdots \leq t_n$ which are not equally spaced.

3.8.1 Local level model

We begin by considering a continuous version of the local level model (2.3). To construct this, we need a continuous analogue of the Gaussian random walk. This can be obtained from the *Brownian motion process*, defined as the continuous stochastic process $w(t)$ such that $w(0) = 0$, $w(t) \sim \mathrm{N}(0, t)$ for $0 < t < \infty$, where increments $w(t_2) - w(t_1)$, $w(t_4) - w(t_3)$ for $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$ are independent. We sometimes need to consider increments $dw(t)$, where $dw(t) \sim \mathrm{N}(0, dt)$ for dt infinitesimally small. Analogously to the random walk $\alpha_{t+1} = \alpha_t + \eta_t$, $\eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$ for the discrete model, we define $\alpha(t)$ by the continuous time relation $d\alpha(t) = \sigma_\eta dw(t)$ where σ_η is an appropriate positive scale parameter. This suggests that as the continuous analogue of the local level model we adopt the continuous time state space model

$$\begin{aligned} y(t) &= \alpha(t) + \varepsilon(t), \\ \alpha(t) &= \alpha(0) + \sigma_\eta w(t), \qquad 0 \leq t \leq T, \end{aligned} \tag{3.33}$$

where $T > 0$.

The nature of $\varepsilon(t)$ in (3.33) requires careful thought. It must first be recognised that for any analysis that is performed digitally, which is all that we consider in this book, $y(t)$ cannot be admitted into the calculations as a continuous record; we can only deal with it as a series of values observed at a discrete set of time points $0 \leq t_1 < t_2 < \cdots < t_n \leq T$. Second, $\mathrm{Var}[\varepsilon(t)]$ must be bounded significantly away from zero; there is no point in carrying out an analysis when $y(t)$ is indistinguishably close to $\alpha(t)$. Third, in order to obtain a continuous analogue of the local level model we need to assume that $\mathrm{Cov}[\varepsilon(t_i), \varepsilon(t_j)] = 0$ for observational points t_i, t_j $(i \neq j)$. It is obvious that if the observational points are close together it may be advisable to set up an autocorrelated model for $\varepsilon(t)$, for example a low-order autoregressive model; however, the coefficients of this would have to be put into the state vector and the resulting model would not be a continuous local level model. In order to allow $\mathrm{Var}[\varepsilon(t)]$ to vary over time we assume that $\mathrm{Var}[\varepsilon(t)] = \sigma^2(t)$ where $\sigma^2(t)$ is a non-stochastic function of t that may depend on unknown parameters. We conclude that in place of (3.33) a more appropriate form of the model is

$$\begin{aligned} &y(t) = \alpha(t) + \varepsilon(t), && t = t_1, \ldots, t_n, && \varepsilon(t_i) \sim \mathrm{N}[0, \sigma^2(t_i)], \\ &\alpha(t) = \alpha(0) + \sigma_\eta w(t), && 0 \leq t \leq T. \end{aligned} \tag{3.34}$$

We next consider the estimation of unknown parameters by maximum likelihood. Since by definition the likelihood is equal to

$$p[y(t_1)]p[y(t_2)|y(t_1)]\cdots p[y(t_n)|y(t_1),\ldots,y(t_{n-1})],$$

it depends on $\alpha(t)$ only at values $t_1,\ldots,t_n$. Thus for estimation of parameters we can employ the reduced model

$$\begin{aligned} y_i &= \alpha_i + \varepsilon_i, \\ \alpha_{i+1} &= \alpha_i + \eta_i, \qquad i = 1,\ldots,n, \end{aligned} \tag{3.35}$$

where $y_i = y(t_i)$, $\alpha_i = \alpha(t_i)$, $\varepsilon_i = \varepsilon(t_i)$, $\eta_i = \sigma_\eta[w(t_{i+1}) - w(t_i)]$ and where the ε_i's are assumed to be independent. This is a discrete local level model which differs from (2.3) only because the variances of the ε_i's can be unequal; consequently we can calculate the loglikelihood by a slight modification of the method of Subsection 2.10.1 which allows for the variance inequality.

Having estimated the model parameters, suppose that we wish to estimate $\alpha(t)$ at values $t = t_{j_*}$ between t_j and t_{j+1} for $1 \le j < n$. We adjust and extend equations (3.35) to give

$$\begin{aligned} \alpha_{j_*} &= \alpha_j + \eta_j^*, \\ y_{j_*} &= \alpha_{j_*} + \varepsilon_{j_*}, \\ \alpha_{j+1} &= \alpha_{j_*} + \eta_{j_*}^*, \end{aligned} \tag{3.36}$$

where $y_{j_*} = y(t_{j_*})$ is treated as missing, $\eta_j^* = \sigma_\eta[w(t_{j_*}) - w(t_j)]$ and $\eta_{j_*}^* = \sigma_\eta[w(t_{j+1}) - w(t_{j_*})]$. We can now calculate $\mathrm{E}[\alpha_{j_*}|y(t_1),\ldots,y(t_n)]$ and $\mathrm{Var}[\alpha_{j_*}|y(t_1),\ldots,y(t_n)]$ by routine applications of the Kalman filter and smoother for series with missing observations, as described in Section 2.7, with a slight modification to allow for unequal observational error variances.

3.8.2 Local linear trend model

Now let us consider the continuous analogue of the local linear trend model (3.2) for the case where $\sigma_\xi^2 = 0$, so that, in effect, the trend term μ_t is modelled by the relation $\Delta^2\mu_{t+1} = \zeta_t$. For the continuous case, denote the trend by $\mu(t)$ and the slope by $\nu(t)$ by analogy with (3.2). The natural model for the slope is then $d\nu(t) = \sigma_\zeta dw(t)$, where $w(t)$ is standard Brownian motion and $\sigma_\zeta > 0$, which gives

$$\nu(t) = \nu(0) + \sigma_\zeta w(t), \qquad 0 \le t \le T. \tag{3.37}$$

By analogy with (3.2) with $\sigma_\xi^2 = 0$, the model for the trend level is $d\mu(t) = \nu(t)dt$, giving

$$\begin{aligned}\mu(t) &= \mu(0) + \int_0^t \nu(s)\ ds \\ &= \mu(0) + \nu(0)t + \sigma_\zeta \int_0^t w(s)\ ds. \end{aligned} \tag{3.38}$$

As before, suppose that $y(t)$ is observed at times $t_1 \le \cdots \le t_n$. Analogously to (3.34), the observation equation for the continuous model is

$$y(t) = \alpha(t) + \varepsilon(t), \qquad t = t_1, \ldots, t_n, \qquad \varepsilon(t_i) \sim \mathrm{N}[0, \sigma^2(t_i)], \tag{3.39}$$

and the state equation can be written in the form

$$d\begin{bmatrix} \mu(t) \\ \nu(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(t) \\ \nu(t) \end{bmatrix} dt + \sigma_\zeta \begin{bmatrix} 0 \\ dw(t) \end{bmatrix}. \tag{3.40}$$

For maximum likelihood estimation we employ the discrete state space model,

$$\begin{aligned} y_i &= \mu_i + \varepsilon_i, \\ \begin{pmatrix} \mu_{i+1} \\ \nu_{i+1} \end{pmatrix} &= \begin{bmatrix} 1 & \delta_i \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_i \\ \nu_i \end{pmatrix} + \begin{pmatrix} \xi_i \\ \zeta_i \end{pmatrix}, \qquad i = 1, \ldots, n, \end{aligned} \tag{3.41}$$

where $\mu_i = \mu(t_i)$, $\nu_i = \nu(t_i)$, $\varepsilon_i = \varepsilon(t_i)$ and $\delta_i = t_{i+1} - t_i$; also

$$\xi_i = \sigma_\zeta \int_{t_i}^{t_{i+1}} [w(s) - w(t_i)]\, ds,$$

and

$$\zeta_i = \sigma_\zeta [w(t_{i+1}) - w(t_i)],$$

as can be verified from (3.37) and (3.38). From (3.39), $\mathrm{Var}(\varepsilon_i) = \sigma^2(t_i)$. Since $\mathrm{E}[w(s) - w(t_i)] = 0$ for $t_i \le s \le t_{i+1}$, $\mathrm{E}(\xi_i) = \mathrm{E}(\zeta_i) = 0$. To calculate $\mathrm{Var}(\xi_i)$, approximate ξ_i by the sum

$$\frac{\delta_i}{M} \sum_{j=0}^{M-1} (M - j) w_j$$

where $w_j \sim \mathrm{N}(0, \sigma_\zeta^2 \delta_i / M)$ and $\mathrm{E}(w_j w_k) = 0$ $(j \ne k)$. This has variance

$$\sigma_\zeta^2 \frac{\delta_i^3}{M} \sum_{j=0}^{M-1} \left(1 - \frac{j}{M}\right)^2,$$

which converges to

$$\sigma_\zeta^2 \delta_i^3 \int_0^1 x^2 \, dx = \frac{1}{3}\sigma_\zeta^2 \delta_i^3$$

as $M \to \infty$. Also,

$$\begin{aligned}
\mathrm{E}(\xi_i \zeta_i) &= \sigma_\zeta^2 \int_{t_i}^{t_{i+1}} \mathrm{E}[\{w(s) - w(t_i)\}\{w(t_{i+1}) - w(t_i)\}] \, ds \\
&= \sigma_\zeta^2 \int_0^{\delta_i} x \, dx \\
&= \frac{1}{2}\sigma_\zeta^2 \delta_i^2,
\end{aligned}$$

and $\mathrm{E}(\zeta_i^2) = \sigma_\zeta^2 \delta_i$. Thus the variance matrix of the disturbance term in the state equation (3.41) is

$$Q_i = \mathrm{Var}\begin{pmatrix} \xi_i \\ \zeta_i \end{pmatrix} = \sigma_\zeta^2 \delta_i \begin{bmatrix} \frac{1}{3}\delta_i^2 & \frac{1}{2}\delta_i \\ \frac{1}{2}\delta_i & 1 \end{bmatrix}. \tag{3.42}$$

The loglikelihood is then calculated by means of the Kalman filter as in Section 7.2.

As with model (3.34), adjustments (3.36) can also be introduced to model (3.39) and (3.40) in order to estimate the conditional mean and variance matrix of the state vector $[\mu(t), \nu(t)]'$ at values of t other than $t_1, \ldots, t_n$. Chapter 9 of Harvey (1989) may be consulted for extensions to more general models.

3.9 Spline smoothing

3.9.1 Spline smoothing in discrete time

Suppose we have a univariate series $y_1, \ldots, y_n$ of values which are equispaced in time and we wish to approximate the series by a relatively smooth function $\mu(t)$. A standard approach is to choose $\mu(t)$ by minimising

$$\sum_{t=1}^{n} [y_t - \mu(t)]^2 + \lambda \sum_{t=1}^{n} [\Delta^2 \mu(t)]^2 \tag{3.43}$$

with respect to $\mu(t)$ for given $\lambda > 0$. It is important to note that we are considering $\mu(t)$ here to be a discrete function of t at time points $t = 1, \ldots, n$, in contrast to the situation considered in the next section where $\mu(t)$ is a continuous function of time. If λ is small, the values of $\mu(t)$ will be close to the y_t's but $\mu(t)$ may not be smooth enough. If λ is large the $\mu(t)$ series will be smooth but the values of $\mu(t)$ may not be close enough to the y_t's. The function $\mu(t)$ is called a *spline*. Reviews of methods related to this idea are given in Silverman

(1985), Wahba (1990) and Green and Silverman (1994). Note that in this book we usually take t as the time index but it can also refer to other sequentially ordered measures such as temperature, earnings and speed.

Let us now consider this problem from a state space standpoint. Let $\alpha_t = \mu(t)$ for $t = 1, \ldots, n$ and assume that y_t and α_t obey the state space model

$$y_t = \alpha_t + \varepsilon_t, \qquad \Delta^2 \alpha_t = \zeta_t, \qquad t = 1, \ldots, n, \tag{3.44}$$

where $\mathrm{Var}(\varepsilon_t) = \sigma^2$ and $\mathrm{Var}(\zeta_t) = \sigma^2/\lambda$ with $\lambda > 0$. We observe that the second equation of (3.44) is one of the smooth models for trend considered in Section 3.2. For simplicity suppose that α_{-1} and α_0 are fixed and known. The log of the joint density of $\alpha_1, \ldots, \alpha_n, y_1, \ldots, y_n$ is then, apart from irrelevant constants,

$$-\frac{\lambda}{2\sigma^2} \sum_{t=1}^{n} \left(\Delta_t^2 \alpha_t\right)^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - \alpha_t)^2. \tag{3.45}$$

Now suppose that our objective is to smooth the y_t series by estimating α_t by $\hat{\alpha}_t = E(\alpha_t | Y_n)$. We shall employ a technique that we shall use extensively later so we state it in general terms. Suppose $\alpha = (\alpha_1', \ldots, \alpha_n')'$ and $y = (y_1', \ldots, y_n')'$ are jointly normally distributed stacked vectors with density $p(\alpha, y)$ and we wish to calculate $\hat{\alpha} = E(\alpha | y)$. Then $\hat{\alpha}$ is the solution of the equations

$$\frac{\partial \log p(\alpha, y)}{\partial \alpha} = 0.$$

This follows since $\log p(\alpha|y) = \log p(\alpha, y) - \log p(y)$ so $\partial \log p(\alpha|y)/\partial \alpha = \partial \log p(\alpha, y)/\partial \alpha$. Now the solution of the equations $\partial \log p(\alpha|y)/\partial \alpha = 0$ is the mode of the density $p(\alpha|y)$ and since the density is normal the mode is equal to the mean vector $\hat{\alpha}$. The conclusion follows. Since $p(\alpha|y)$ is the conditional distribution of α given y, we call this technique *conditional mode estimation* of $\alpha_1, \ldots, \alpha_n$.

Applying this technique to (3.45), we see that $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ can be obtained by minimising

$$\sum_{t=1}^{n} (y_t - \alpha_t)^2 + \lambda \sum_{t=1}^{n} (\Delta^2 \alpha_t)^2.$$

Comparing this with (3.43), and ignoring for the moment the initialisation question, we see that the spline smoothing problem can be solved by finding $E(\alpha_t | Y_n)$ for model (3.44). This is achieved by a standard extension of the smoothing technique of Section 2.4 that will be given in Section 4.4. It follows that state space techniques can be used for spline smoothing. Treatments along these lines have been given by Kohn, Ansley and Wong (1992). This approach has the advantages that the models can be extended to include extra features such as explanatory variables, calendar variations and intervention effects in the ways indicated earlier in this chapter; moreover, unknown quantities, for example λ in (3.43), can

be estimated by maximum likelihood using methods that we shall describe in Chapter 7.

3.9.2 Spline smoothing in continuous time

Let us now consider the smoothing problem where the observation $y(t)$ is a continuous function of time t for t in an interval which for simplicity we take to be $0 \leq t \leq T$. Suppose that we wish to smooth $y(t)$ by a function $\mu(t)$ given a sample of values $y(t_i)$ for $i = 1, \ldots, n$ where $0 < t_1 < \cdots < t_n < T$. A traditional approach to the problem is to choose $\mu(t)$ to be the twice-differentiable function on $(0, T)$ which minimises

$$\sum_{i=1}^{n} [y(t_i) - \mu(t_i)]^2 + \lambda \int_0^T \left[\frac{\partial^2 \mu(t)}{\partial t^2}\right]^2 dt, \tag{3.46}$$

for given $\lambda > 0$. We observe that (3.46) is the analogue in continuous time of (3.43) in discrete time. This is a well-known problem, a standard treatment to which is presented in Chapter 2 of Green and Silverman (1994). Their approach is to show that the resulting $\mu(t)$ must be a *cubic spline*, which is defined as a cubic polynomial function in t between each pair of time points t_i, t_{i+1} for $i = 0, 1, \ldots, n$ with $t_0 = 0$ and $t_{n+1} = T$, such that $\mu(t)$ and its first two derivatives are continuous at each t_i for $i = 1, \ldots, n$. The properties of the cubic spline are then used to solve the minimisation problem. In contrast, we shall present a solution based on a continuous time state space model of the kind considered in Section 3.8.

We begin by adopting a model for $\mu(t)$ in the form (3.38), which for convenience we reproduce here as

$$\mu(t) = \mu(0) + \nu(0)t + \sigma_\zeta \int_0^t w(s)\, ds, \qquad 0 \leq t \leq T. \tag{3.47}$$

This is a natural model to consider since it is the simplest model in continuous time for a trend with smoothly varying slope. As the observation equation we take

$$y(t_i) = \mu(t_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \qquad i = 1, \ldots, n,$$

where the ε_i's are independent of each other and of $w(t)$ for $0 < t \leq T$. We have taken $\mathrm{Var}(\varepsilon_i)$ to be constant since this is a reasonable assumption for many smoothing problems, and also since it leads to the same solution to the problem of minimising (3.46) as the Green–Silverman approach.

Since $\mu(0)$ and $\nu(0)$ are normally unknown, we represent them by diffuse priors. On these assumptions, Wahba (1978) has shown that on taking

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2},$$

the conditional mean $\hat{\mu}(t)$ of $\mu(t)$ defined by (3.47), given the observations $y(t_1), \ldots, y(t_n)$, is the solution to the problem of minimising (3.46) with respect to $\mu(t)$. We shall not give details of the proof here but will instead refer to discussions of the result by Wecker and Ansley (1983) and Green and Silverman (1994).

The result is important since it enables problems in spline smoothing to be solved by state space methods. We note that Wahba and Wecker and Ansley in the papers cited consider the more general problem in which the second term of (3.46) is replaced by the more general form

$$\lambda \int_0^T \left[\frac{d^m \mu(t)}{dt^m} \right] dt,$$

for $m = 2, 3, \ldots$.

We have reduced the problem of minimising (3.46) to the treatment of a special case of the state space model (3.39) and (3.40) in which $\sigma^2(t_i) = \sigma_\varepsilon^2$ for all i. We can therefore compute $\hat{\mu}(t)$ and $\mathrm{Var}[\mu(t)|y(t_1), \ldots, y(t_n)]$ by routine Kalman filtering and smoothing. We can also compute the loglikelihood and, consequently, estimate λ by maximum likelihood; this can be done efficiently by concentrating out σ_ε^2 by a straightforward extension of the method described in Subsection 2.10.2 and then maximising the concentrated loglikelihood with respect to λ in a one-dimensional search. The implication of these results is that the flexibility and computational power of state space methods can be employed to solve problems in spline smoothing.

3.10 Further comments on state space analysis

In this section we provide discussions and illustrations of state space time series analysis. First we compare the state space and Box–Jenkins approaches to time series analysis. Next we provide examples of how problems in time series analysis can be handled within a state space framework.

3.10.1 State space versus Box–Jenkins approaches

The early development of state space methodology took place in the field of engineering rather than statistics, starting with the pathbreaking paper of Kalman (1960). In this paper Kalman did two crucially important things. He showed that a very wide class of problems could be encapsulated in a simple linear model, essentially the state space model (3.1). Secondly he showed how, due to the Markovian nature of the model, the calculations needed for practical application of the model could be set up in recursive form in a way that was particularly convenient on a computer. A huge amount of work was done in the development of these ideas in the engineering field. In the 1960s to the early 1980s contributions to state space methodology from statisticians and econometricians were isolated

and sporadic. In recent years however there has been a rapid growth of interest in the field in both statistics and econometrics as is indicated by references throughout the book.

The key advantage of the state space approach is that it is based on a structural analysis of the problem. The different components that make up the series, such as trend, seasonal, cycle and calendar variations, together with the effects of explanatory variables and interventions, are modelled separately before being put together in the state space model. It is up to the investigator to identify and model any features in particular situations that require special treatment. In contrast, the Box–Jenkins approach is a kind of 'black box', in which the model adopted depends purely on the data without prior analysis of the structure of the system that generated the data. A second advantage of state space models is that they are flexible. Because of the recursive nature of the models and of the computational techniques used to analyse them, it is straightforward to allow for known changes in the structure of the system over time. Other advantages of a state space analysis are (i) its treatment of missing observations; (ii) explanatory variables can be incorporated into the model without difficulty; (iii) associated regression coefficients can be permitted to vary stochastically over time if this seems to be called for in the application; (iv) trading-day adjustments and other calendar variations can be readily taken care of; (v) no extra theory is required for forecasting since all that is needed is to project the Kalman filter forward into the future.

When employing the Box–Jenkins approach, the elimination of trend and seasonal by differencing may not be a drawback if forecasting is the only object of the analysis. However, in many contexts, particularly in official statistics and some econometric applications, knowledge about such components has intrinsic importance. It is true that estimates of trend and seasonal can be 'recovered' from the differenced series by maximising the residual mean square as in Burman (1980) but this seems an artificial procedure which is not as appealing as modelling the components directly. Furthermore, the requirement that the differenced series should be stationary is a weakness of the theory. In the economic and social fields, real series are never stationary however much differencing is done. The investigator has to face the question, how close to stationarity is close enough? This is a hard question to answer.

In practice it is found that the airline model and similar ARIMA models fit many data sets quite well, but it can be argued that the reason for this is that they are approximately equivalent to plausible state space models. This point is discussed at length by Harvey (1989, pp. 72–73). As we move away from airline-type models, the model identification process in the Box–Jenkins system becomes difficult to apply. The main tool is the sample autocorrelation function which is notoriously imprecise due to its high sampling variability. Practitioners in applied time series analysis are familiar with the fact that many examples can be found where the data appear to be explained equally well by models whose

specifications look very different. The above discussion has been based on Durbin (2000b, §3).

3.10.2 Benchmarking

A common problem in official statistics is the adjustment of monthly or quarterly observations, obtained from surveys and therefore subject to survey errors, to agree with annual totals obtained from censuses and assumed to be free from error. The annual totals are called *benchmarks* and the process is called *benchmarking*. We shall show how the problem can be handled within a state space framework.

Denote the survey observations, which we take to be monthly ($s = 12$), by y_t and the true values they are intended to estimate by y_t^* for $t = 12(i-1)+j$, $i = 1, \ldots, \ell$ and $j = 1, \ldots, 12$, where ℓ is the number of years. Thus the survey error is $y_t - y_t^*$ which we denote by $\sigma_t^s \xi_t^s$ where σ_t^s is the standard deviation of the survey error at time t. The error ξ_t^s is modelled as an AR(1) model with unit variance. In principle, ARMA models of higher order could be used. We assume that the values of σ_t^s are available from survey experts and that the errors are bias free; we will mention the estimation of bias later. The benchmark values are given by $x_i = \sum_{j=1}^{12} y^*_{12(i-1)+j}$ for $i = 1, \ldots, \ell$. We suppose for simplicity of exposition that we have these annual values for all years in the study though in practice the census values will usually lag a year or two behind the survey observations. We take as the model for the observations

$$y_t = \mu_t + \gamma_t + \sum_{j=1}^{k} \delta_{jt} w_{jt} + \varepsilon_t + \sigma_t^s \xi_t^s, \qquad t = 1, \ldots, 12\ell, \tag{3.48}$$

where μ_t is trend, γ_t is seasonal and the term $\sum_{j=1}^{k} \delta_{jt} w_{jt}$ represents systematic effects such as the influence of calendar variations which can have a substantial effect on quantities such as retail sales but which can vary slowly over time.

The series is arranged in the form

$$y_1, \ldots, y_{12}, x_1, y_{13}, \ldots, y_{24}, x_2, y_{25}, \ldots, y_{12\ell}, x_\ell.$$

Let us regard the time point in the series at which the benchmark occurs as $t = (12i)'$; thus the point $t = (12i)'$ occurs in the series between $t = 12i$ and $t = 12i + 1$. It seems reasonable to update the regression coefficients δ_{jt} only once a year, say in January, so we take for these coefficients the model

$$\begin{aligned} \delta_{j,12i+1} &= \delta_{j,12i} + \zeta_{j,12i}, && j = 1, \ldots, k, \qquad i = 1, \ldots, \ell, \\ \delta_{j,t+1} &= \delta_{j,t}, && \text{otherwise.} \end{aligned}$$

Take the integrated random walk model for the trend component and model (3.3) for the seasonal component, that is,

$$\Delta^2 \mu_t = \xi_t, \qquad \gamma_t = -\sum_{j=1}^{11} \gamma_{t-j} + \omega_t;$$

see Section 3.2 for alternative trend and seasonal models. It turns out to be convenient to put the observation errors into the state vector, so we take

$$\alpha_t = \big(\mu_t, \ldots, \mu_{t-11}, \gamma_t, \ldots, \gamma_{t-11}, \delta_{1t}, \ldots, \delta_{kt}, \varepsilon_t, \ldots, \varepsilon_{t-11}, \xi_t^s\big)'.$$

Thus $y_t = Z_t \alpha_t$ where

$$Z_t = \big(1, 0, \ldots, 0, 1, 0, \ldots, 0, w_{1t}, \ldots, w_{kt}, 1, 0, \ldots, 0, \sigma_t^s\big), \qquad t = 1, \ldots, n,$$

and $x_i = Z_t \alpha_t$ where

$$Z_t = \left(1, \ldots, 1, 0, \ldots, 0, \sum_{s=12i-11}^{12i} w_{1s}, \ldots, \sum_{s=12i-11}^{12i} w_{ks}, 1, \ldots, 1, 0\right), \; t = (12i)',$$

for $i = 1, \ldots, \ell$. Using results from Section 3.2 it is easy to write down the state transition from α_t to α_{t+1} for $t = 12i - 11$ to $t = 12i - 1$, taking account of the fact that $\delta_{j,t+1} = \delta_{jt}$. From $t = 12i$ to $t = (12i)'$ the transition is the identity. From $t = (12i)'$ to $t = 12i + 1$, the transition is the same as for $t = 12i - 11$ to $t = 12i - 1$, except that we take account of the relation $\delta_{j,12i+1} = \delta_{j,12i} + \zeta_{j,12i}$ where $\zeta_{j,12i} \neq 0$.

There are many variants of the benchmarking problem. For example, the annual totals may be subject to error, the benchmarks may be values at a particular month, say December, instead of annual totals, the survey observations may be biased and the bias needs to be estimated, more complicated models than the AR(1) model can be used to model y_t^*; finally, the observations may behave multiplicatively whereas the benchmark constraint is additive, thus leading to a nonlinear model. All these variants are dealt with in a comprehensive treatment of the benchmarking problem by Durbin and Quenneville (1997). They also consider a two-step approach to the problem in which a state space model is first fitted to the survey observations and the adjustments to satisfy the benchmark constraints take place in a second stage.

Essentially, this example demonstrates that the state space approach can be used to deal with situations in which the data come from two different sources. Another example of such problems will be given in Subsection 3.10.3 where we model different series which aim at measuring the same phenomenon simultaneously, which are all subject to sampling error and which are observed at different time intervals.

3.10.3 Simultaneous modelling of series from different sources

A different problem in which data come from two different sources has been considered by Harvey and Chung (2000). Here the objective is to estimate the level of UK unemployment and the month-to-month change of unemployment given two different series. Of these, the first is a series y_t of observations obtained from a monthly survey designed to estimate unemployment according to an internationally accepted standard definition (the so-called ILO definition where ILO stands for International Labour Office); this estimate is subject to survey error. The second series consists of monthly counts x_t of the number of individuals claiming unemployment benefit; although these counts are known accurately, they do not themselves provide an estimate of unemployment consistent with the ILO definition. Even though the two series are closely related, the relationship is not even approximately exact and it varies over time. The problem to be considered is how to use the knowledge of x_t to improve the accuracy of the estimate based on y_t alone.

The solution suggested by Harvey and Chung (2000) is to model the bivariate series $(y_t, x_t)'$ by the structural time series model

$$\begin{aligned} \begin{pmatrix} y_t \\ x_t \end{pmatrix} &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0, \Sigma_\varepsilon), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{N}(0, \Sigma_\xi), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}(0, \Sigma_\zeta), \end{aligned} \tag{3.49}$$

for $t = 1, \ldots, n$. Here, μ_t, ε_t, ν_t, ξ_t and ζ_t are 2×1 vectors and Σ_ε, Σ_ξ and Σ_ζ are 2×2 variance matrices. Seasonals can also be incorporated. Many complications are involved in implementing the analysis based on this model, particularly those arising from design features of the survey such as overlapping samples. A point of particular interest is that the claimant count x_t is available one month ahead of the survey value y_t. This extra value of x_t can be easily and efficiently made use of by the missing observations technique discussed in Section 4.10. For a discussion of the details we refer the reader to Harvey and Chung (2000).

This is not the only way in which the information available in x_t can be utilised. For example, in the published discussion of the paper, Durbin (2000a) suggested two further possibilities, the first of which is to model the series $y_t - x_t$ by a structural time series model of one of the forms considered in Section 3.2; unemployment level could then be estimated by $\hat{\mu}_t + x_t$ where $\hat{\mu}_t$ is the forecast of the trend μ_t in the model using information up to time $t-1$, while the month-to-month change could be estimated by $\hat{\nu}_t + x_t - x_{t-1}$ where $\hat{\nu}_t$ is the forecast of the slope ν_t. Alternatively, x_t could be incorporated as an explanatory variable into an appropriate form of model (3.14) with coefficient β_j replaced by β_{jt} which varies over time according to (3.15). In an obvious notation, trend and change would then be estimated by $\hat{\mu}_t + \hat{\beta}_t x_t$ and $\hat{\nu}_t + \hat{\beta}_t x_t - \hat{\beta}_{t-1} x_{t-1}$.

3.11 Exercises

3.11.1

Consider the autoregressive moving average, with constant, plus noise model

$$y_t = y_t^* + \varepsilon_t, \qquad y_t^* = \mu + \phi_1 y_{t-1}^* + \phi_2 y_{t-2}^* + \zeta_t + \theta_1 \zeta_{t-1},$$

for $t = 1, \ldots, n$, where μ is an unknown constant. The disturbances $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\zeta_t \sim \mathrm{N}(0, \sigma_\zeta^2)$ are mutually and serially independent at all times and lags. The autoregressive coefficients ϕ_1 and ϕ_2 are restricted such that y_t^* is a stationary process and with moving average coefficient $0 < \theta_1 < 1$. Represent this model in the state space form (3.1); define the state vector α_t and its initial condition.

3.11.2

The local linear trend model with a smooth slope equation is given by

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_{t+1} = \mu_t + \beta_t + \eta_t, \qquad \Delta^r \beta_t = \zeta_t,$$

for $t = 1, \ldots, n$ and some positive integer r, where the normally distributed disturbances ε_t, η_t and ζ_t are mutually and serially independent at all times and lags. Express the trend model for $r = 3$ in the state space form (3.1); define the state vector α_t and its initial condition.

When a stationary slope component is requested, we consider

$$(1 - \phi L)^r \beta_t = \zeta_t,$$

where $0 < \phi < 1$ is an autoregressive coefficient and L is the lag operator. Express the trend model with such a stationary slope for $r = 2$ in the state space form (3.1); define the state state vector α_t and its initial condition.

3.11.3

The exponential smoothing method of forecasting of Section 3.5 can also be expressed by

$$\hat{y}_{t+1} = \hat{y}_t + \lambda(y_t - \hat{y}_t), \qquad t = 1, \ldots, n.$$

where the new one-step ahead forecast is the old one plus an adjustment for the error that occurred in the last forecast. The steady state solution of the Kalman filter applied to the local level model (2.3) in Section 2.11 can be partly represented by exponentional smoothing. Show this relationship and develop an expression of λ in terms of $q = \sigma_\eta^2 \,/\, \sigma_\varepsilon^2$.

3.11.4

Consider the state space model (3.1) extended with regression effects

$$y_t = X_t \beta + Z_t \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = W_t \beta + T_t \alpha_t + R_t \eta_t,$$

for $t = 1, \ldots, n$, where X_t and W_t are fixed matrices that (partly) consist of exogenous variables and β is a vector of regression coefficient; the matrices have appropriate dimensions. Show that this state space model can be expressed as

$$y_t = X_t^* \beta + Z_t \alpha_t^* + \varepsilon_t, \qquad \alpha_{t+1}^* = T_t \alpha_t^* + R_t \eta_t,$$

for $t = 1, \ldots, n$. Give an expression of X_t^* in terms of X_t, W_t and the other system matrices.