2 Local level model

2.1 Introduction

The purpose of this chapter is to introduce the basic techniques of state space analysis, such as filtering, smoothing, initialisation and forecasting, in terms of a simple example of a state space model, the local level model. This is intended to help beginners grasp the underlying ideas more quickly than they would if we were to begin the book with a systematic treatment of the general case. We shall present results from both the classical and Bayesian perspectives, assuming normality, and also from the standpoint of minimum variance linear unbiased estimation when the normality assumption is dropped.

A *time series* is a set of observations $y_1, \ldots, y_n$ ordered in time. The basic model for representing a time series is the additive model

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{2.1}$$

Here, μ_t is a slowly varying component called the *trend*, γ_t is a periodic component of fixed period called the *seasonal* and ε_t is an irregular component called the *error* or *disturbance*. In general, the observation y_t and the other variables in (2.1) can be vectors but in this chapter we assume they are scalars. In many applications, particularly in economics, the components combine multiplicatively, giving

$$y_t = \mu_t \gamma_t \varepsilon_t. \tag{2.2}$$

By taking logs however and working with logged values model (2.2) reduces to model (2.1), so we can use model (2.1) for this case also.

To develop suitable models for μ_t and γ_t we need the concept of a *random walk*. This is a scalar series α_t determined by the relation $\alpha_{t+1} = \alpha_t + \eta_t$ where the η_t's are independent and identically distributed random variables with zero means and variances σ_η^2.

Consider a simple form of model (2.1) in which $\mu_t = \alpha_t$ where α_t is a random walk, no seasonal is present and all random variables are normally distributed. We assume that ε_t has constant variance σ_ε^2. This gives the model

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \alpha_{t+1} &= \alpha_t + \eta_t, \qquad & \eta_t &\sim \mathrm{N}\big(0, \sigma_\eta^2\big), \end{aligned} \tag{2.3}$$

Time Series Analysis by State Space Methods, Second Edition, J Durbin, S J Koopman © J Durbin, S J Koopman 2012. Published in 2012 by Oxford University Press

for $t = 1, \ldots, n$ where the ε_t's and η_t's are all mutually independent and are independent of α_1. This model is called the *local level model.* Although it has a simple form, this model is not an artificial special case and indeed it provides the basis for the analysis of important real problems in practical time series analysis; for example, the local level model provides the basis for our analysis of the Nile data that we start in Subsection 2.2.5. It exhibits the characteristic structure of state space models in which there is a series of unobserved values $\alpha_1, \ldots, \alpha_n$, called the *states*, which represents the development over time of the system under study, together with a set of *observations* $y_1, \ldots, y_n$ which are related to the α_t's by the state space model (2.3). The object of the methodology that we shall develop is to infer relevant properties of the α_t's from a knowledge of the observations $y_1, \ldots, y_n$. The model (2.3) is suitable for both classical and Bayesian analysis. Where the ε_t's and the η_t's are not normally distributed we obtain equivalent results from the standpoint of minimum variance linear unbiased estimation.

We assume initially that $\alpha_1 \sim \mathrm{N}(a_1,P_1)$ where a_1 and P_1 are known and that σ^2_ε and σ^2_η are known. Since random walks are non-stationary the model is non-stationary. By non-stationary here we mean that distributions of random variables y_t and α_t depend on time t.

For applications of model (2.3) to real series, we need to compute quantities such as the mean of α_t given $y_1, \ldots, y_{t-1}$ or the mean of α_t given $y_1, \ldots, y_n$, together with their variances; we also need to fit the model to data by calculating maximum likelihood estimates of the parameters σ^2_ε and σ^2_η. In principle, this could be done by using standard results from multivariate normal theory as described in books such as Anderson (2003). In this approach the observations y_t generated by the local level model are represented as the $n \times 1$ vector Y_n such that

$$Y_n \sim \mathrm{N}(\mathtt{1} a_1, \Omega), \quad \text{with} \quad Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathtt{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \Omega = \mathtt{1}\mathtt{1}'P_1 + \Sigma, \quad (2.4)$$

where the (i,j)th element of the $n \times n$ matrix Σ is given by

$$\Sigma_{ij} = \begin{cases} (i-1)\sigma^2_\eta, & i < j \\ \sigma^2_\varepsilon + (i-1)\sigma^2_\eta, & i = j, \qquad i,j = 1, \ldots, n, \\ (j-1)\sigma^2_\eta, & i > j \end{cases} \quad (2.5)$$

which follows since the local level model implies that

$$y_t = \alpha_1 + \sum_{j=1}^{t-1} \eta_j + \varepsilon_t, \qquad t = 1, \ldots, n. \quad (2.6)$$

Starting from this knowledge of the distribution of Y_n, estimation of conditional means, variances and covariances is in principle a routine matter using standard results in multivariate analysis based on the properties of the multivariate normal distribution. However, because of the serial correlation between the observations y_t, the routine computations rapidly become cumbersome as n increases. This naive approach to estimation can be improved upon considerably by using the filtering and smoothing techniques described in the next three sections. In effect, these techniques provide efficient computing algorithms for obtaining the same results as those derived by multivariate analysis theory. The remaining sections of this chapter deal with other important issues such as fitting the local level model and forecasting future observations.

2.2 Filtering

2.2.1 The Kalman filter

The object of *filtering* is to update our knowledge of the system each time a new observation y_t is brought in. We shall first develop the theory of filtering for the local level model (2.3) where the ε_t's and η_t's are assumed normal from the standpoint of classical analysis. Since in this case all distributions are normal, conditional joint distributions of one set of observations given another set are also normal. Let Y_{t-1} be the vector of observations $(y_1, \dots, y_{t-1})'$ for $t = 2, 3, \dots$ and assume that the conditional distribution of α_t given Y_{t-1} is $\mathrm{N}(a_t,P_t)$ where a_t and P_t are known. Assume also that the conditional distribution of α_t given Y_t is $\mathrm{N}(a_{t|t}, P_{t|t})$. The distribution of α_{t+1} given Y_t is $\mathrm{N}(a_{t+1}, P_{t+1})$. Our object is to calculate $a_{t|t}$, $P_{t|t}$, a_{t+1} and P_{t+1} when y_t is brought in. We refer to $a_{t|t}$ as the *filtered estimator* of the state α_t and a_{t+1} as the *one-step ahead predictor* of α_{t+1}. Their respective associated variances are $P_{t|t}$ and P_{t+1}.

An important part is played by the one-step ahead prediction error v_t of y_t. Then $v_t = y_t - a_t$ for $t = 1, \dots, n$, and

$$\begin{aligned}
\mathrm{E}(v_t|Y_{t-1}) &= \mathrm{E}(\alpha_t + \varepsilon_t - a_t|Y_{t-1}) = a_t - a_t = 0,\\
\mathrm{Var}(v_t|Y_{t-1}) &= \mathrm{Var}(\alpha_t + \varepsilon_t - a_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2,\\
\mathrm{E}(v_t|\alpha_t, Y_{t-1}) &= \mathrm{E}(\alpha_t + \varepsilon_t - a_t|\alpha_t, Y_{t-1}) = \alpha_t - a_t,\\
\mathrm{Var}(v_t|\alpha_t, Y_{t-1}) &= \mathrm{Var}(\alpha_t + \varepsilon_t - a_t|\alpha_t, Y_{t-1}) = \sigma_\varepsilon^2,
\end{aligned} \tag{2.7}$$

for $t = 2, \dots, n$. When Y_t is fixed, Y_{t-1} and y_t are fixed so Y_{t-1} and v_t are fixed and vice versa. Consequently, $p(\alpha_t|Y_t) = p(\alpha_t|Y_{t-1}, v_t)$. We have

$$\begin{aligned}
p(\alpha_t|Y_{t-1}, v_t) &= p(\alpha_t, v_t|Y_{t-1})/p(v_t|Y_{t-1})\\
&= p(\alpha_t|Y_{t-1})p(v_t|\alpha_t, Y_{t-1})/p(v_t|Y_{t-1})\\
&= \text{constant} \times \exp(-\frac{1}{2}Q),
\end{aligned} \tag{2.8}$$

where

$$\begin{aligned}Q &= (\alpha_t - a_t)^2/P_t + (v_t - \alpha_t + a_t)^2/\sigma_\varepsilon^2 - v_t^2/(P_t + \sigma_\varepsilon^2)\\ &= \left(\frac{1}{P_t} + \frac{1}{\sigma_\varepsilon^2}\right)(\alpha_t - a_t)^2 - 2(\alpha_t - a_t)\frac{v_t}{\sigma_\varepsilon^2} + \left(\frac{1}{\sigma_\varepsilon^2} - \frac{1}{P_t + \sigma_\varepsilon^2}\right)v_t^2\\ &= \frac{P_t + \sigma_\varepsilon^2}{P_t\,\sigma_\varepsilon^2}\left(\alpha_t - a_t - \frac{P_t\,v_t}{P_t + \sigma_\varepsilon^2}\right)^2.\end{aligned} \tag{2.9}$$

Thus

$$p(\alpha_t|Y_t) = \mathrm{N}\left(a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t\ ,\ \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}\right). \tag{2.10}$$

But $a_{t|t}$ and $P_{t|t}$ have been defined such that $p(\alpha_t|Y_t) = \mathrm{N}(a_{t|t}, P_{t|t})$. It follows that

$$a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t, \tag{2.11}$$

$$P_{t|t} = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}. \tag{2.12}$$

Since $a_{t+1} = \mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(\alpha_t + \eta_t|Y_t)$ and $P_{t+1} = \mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}(\alpha_t + \eta_t|Y_t)$ from (2.3), we have

$$\begin{aligned}a_{t+1} &= \mathrm{E}(\alpha_t|Y_t) &= a_{t|t},\\ P_{t+1} &= \mathrm{Var}(\alpha_t|Y_t) + \sigma_\eta^2 &= P_{t|t} + \sigma_\eta^2,\end{aligned}$$

giving

$$a_{t+1} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t, \tag{2.13}$$

$$P_{t+1} = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2, \tag{2.14}$$

for $t = 2, \ldots, n$. For $t = 1$ we delete the symbol Y_{t-1} in the above derivation and we find that all results from (2.7) to (2.13) hold for $t = 1$ as well as for $t = 2, \ldots, n$.

In order to make these results consistent with the treatment of filtering for the general linear state space model in Subsection 4.3.1, we introduce the notation

$$F_t = \mathrm{Var}(v_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2, \qquad K_t = P_t/F_t,$$

where F_t is referred to as the variance of the prediction error v_t and K_t is known as the *Kalman gain*. Using (2.11) to (2.14) we can then write the full set of relations for updating from time t to time $t+1$ in the form

$$\begin{aligned} v_t &= y_t - a_t, & F_t &= P_t + \sigma_\varepsilon^2, \\ a_{t|t} &= a_t + K_t v_t, & P_{t|t} &= P_t(1 - K_t), \\ a_{t+1} &= a_t + K_t v_t, & P_{t+1} &= P_t(1 - K_t) + \sigma_\eta^2, \end{aligned} \tag{2.15}$$

for $t = 1, \dots, n$, where $K_t = P_t \,/\, F_t$.

We have assumed that a_1 and P_1 are known; however, more general initial specifications for a_1 and P_1 will be dealt with in Section 2.9. Relations (2.15) constitute the celebrated *Kalman filter* for the local level model. It should be noted that P_t depends only on σ_ε^2 and σ_η^2 and does not depend on Y_{t-1}. We include the case $t = n$ in (2.15) for convenience even though a_{n+1} and P_{n+1} are not normally needed for anything except forecasting. A set of relations such as (2.15) which enables us to calculate quantities for $t + 1$ given those for t is called a *recursion*.

2.2.2 Regression lemma

The above derivation of the Kalman filter can be regarded as an application of a regression lemma for the bivariate normal distribution. Suppose that x and y are jointly normally distributed variables with

$$\mathrm{E}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \mathrm{Var}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix},$$

with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}. The joint distribution is

$$p(x, y) = p(y)\, p(x|y),$$

by the definition of the conditional density $p(x|y)$. But it can also be verified by direct multiplication. We have

$$p(x, y) = \frac{A}{2\pi} \exp\left\{ -\frac{1}{2}\sigma_y^{-2}(y - \mu_y)^2 - \frac{1}{2}\sigma_x^{-2}\left[x - \mu_x - \sigma_{xy}\sigma_y^{-2}(y - \mu_y)\right]^2 \right\},$$

where $A = \sigma_x^2 - \sigma_y^{-2}\sigma_{xy}$. It follows that the conditional distribution of x given y is normal and independent of y with mean and variance given by

$$\mathrm{E}(x|y) = \mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \qquad \mathrm{Var}(x|y) = \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}.$$

To apply this lemma to the Kalman filter, let $v_t = y_t - a_t$ and keep Y_{t-1} fixed. Take $x = \alpha_t$ so that $\mu_x = a_t$ and $y = v_t$. It follows that $\mu_y = \mathrm{E}(v_t) = 0$. Then, $\sigma_x^2 = \mathrm{Var}(\alpha_t) = P_t$, $\sigma_y^2 = \mathrm{Var}(v_t) = \mathrm{Var}(\alpha_t - a_t + \varepsilon_t) = P_t + \sigma_\varepsilon^2$ and $\sigma_{xy} = P_t$. We obtain the conditional distribution for α_t given v_t by

$$\mathrm{E}(\alpha_t|v_t) = a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \qquad \mathrm{Var}(\alpha_t|v_t) = P_{t|t} = \frac{P_t}{P_t + \sigma_\varepsilon^2}.$$

In a similar way we can obtain the equations for a_{t+1} and P_{t+1} by application of this regression lemma.

2.2.3 Bayesian treatment

To analyse the local level model from a Bayesian standpoint, we assume that the data are generated by model (2.3). In this approach α_t and y_t are regarded as a parameter and a constant, respectively. Before the observation y_t is taken, the prior distribution of α_t is $p(\alpha_t|Y_{t-1})$. The likelihood of α_t is $p(y_t|\alpha_t, Y_{t-1})$. The posterior distribution of α_t given y_t is given by the Bayes theorem which is proportional to the product of these. In particular we have

$$p(\alpha_t|Y_{t-1}, y_t) = p(\alpha_t|Y_{t-1})\, p(y_t|\alpha_t, Y_{t-1}) \,/\, p(y_t|Y_{t-1}).$$

Since $y_t = \alpha_t + \varepsilon_t$ we have $\mathrm{E}(y_t|Y_{t-1}) = a_t$ and $\mathrm{Var}(y_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2$, so

$$p(\alpha_t|Y_{t-1}, y_t) = \text{constant} \times \exp(-\frac{1}{2}Q),$$

where

$$\begin{aligned} Q &= (\alpha_t - a_t)^2/P_t \;+\; (\alpha_t - a_t)^2/\sigma_\varepsilon^2 \;-\; (y_t - a_t)^2/(P_t + \sigma_\varepsilon^2) \\ &= \frac{P_t + \sigma_\varepsilon^2}{P_t\, \sigma_\varepsilon^2}\left(\alpha_t - a_t - \frac{P_t}{P_t + \sigma_\epsilon^2}(y_t - a_t)\right)^2. \end{aligned} \qquad (2.16)$$

This is a normal density which we denote by $\mathrm{N}(a_{t|t}, P_{t|t})$. Thus the posterior mean and variance are

$$\begin{aligned} a_{t|t} &= a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \\ P_{t|t} &= \frac{P_t\, \sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}, \end{aligned} \qquad (2.17)$$

which are the same as (2.11) and (2.12) on putting $v_t = y_t - a_t$. The case $t = 1$ has the same form. Similarly, the posterior density of α_{t+1} given y_t is $p(\alpha_{t+1}|Y_{t-1}, y_t) = p(\alpha_t + \eta_t|Y_{t-1}, y_t)$, which is normal with mean $a_{t|t}$ and variance $P_{t|t} + \sigma_\eta^2$. Denoting this by $\mathrm{N}(a_{t+1}, P_{t+1})$, we have

$$\begin{aligned} a_{t+1} &= a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \\ P_{t+1} &= P_{t|t} + \sigma_\eta^2 = \frac{P_t\, \sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2, \end{aligned} \qquad (2.18)$$

which are, of course, the same as (2.13) and (2.14). It follows that the Kalman filter from a Bayesian point of view has the same form (2.15) as the Kalman filter from the standpoint of classical inference. This is an important result; as will be seen in Chapter 4 and later chapters, many inference results for the state α_t are the same whether approached from a classical or a Bayesian standpoint.

2.2.4 Minimum variance linear unbiased treatment

In some situations, some workers object to the assumption of normality in model (2.3) on the grounds that the observed time series they are concerned with do not appear to behave in a way that corresponds with the normal distribution. In these circumstances an alternative approach is to treat the filtering problem as a problem in the estimation of α_t and α_{t+1} given Y_t and to confine attention to estimates that are linear unbiased functions of y_t; we then choose those estimates that have minimum variance. We call these estimates *minimum variance linear unbiased estimates* (MVLUE).

Taking first the case of α_t, we seek an estimate $\bar{\alpha}_t$ which has the linear form $\bar{\alpha}_t = \beta + \gamma y_t$ where β and γ are constants given Y_{t-1} and which is unbiased in the sense that the estimation error $\bar{\alpha}_t - \alpha_t$ has zero mean in the conditional joint distribution of α_t and y_t given Y_{t-1}. We therefore have

$$\begin{aligned} \mathrm{E}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) &= \mathrm{E}(\beta + \gamma y_t - \alpha_t | Y_{t-1}) \\ &= \beta + \gamma a_t - a_t = 0, \end{aligned} \tag{2.19}$$

so $\beta = a_t(1-\gamma)$ which gives $\bar{\alpha}_t = a_t + \gamma(y_t - a_t)$. Thus $\bar{\alpha}_t - \alpha_t = \gamma(\alpha_t - a_t + \varepsilon_t) - (\alpha_t - a_t)$. Now $\mathrm{Cov}(\alpha_t - a_t + \varepsilon_t, \alpha_t - a_t) = P_t$ so we have

$$\begin{aligned} \mathrm{Var}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) &= \gamma^2(P_t + \sigma_\varepsilon^2) - 2\gamma P_t + P_t \\ &= \left(P_t + \sigma_\varepsilon^2\right)\left(\gamma - \frac{P_t}{P_t + \sigma_\varepsilon^2}\right)^2 + P_t - \frac{P_t^2}{P_t + \sigma_\varepsilon^2}. \end{aligned} \tag{2.20}$$

This is minimised when $\gamma = P_t/(P_t + \sigma_\varepsilon^2)$ which gives

$$\bar{\alpha}_t = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \tag{2.21}$$

$$\mathrm{Var}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}. \tag{2.22}$$

Similarly, if we estimate α_{t+1} given Y_{t-1} by the linear function $\bar{\alpha}_{t+1}^* = \beta^* + \gamma^* y_t$ and require this to have the unbiasedness property $\mathrm{E}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1}) = 0$, we find that $\beta^* = a_t(1-\gamma^*)$ so $\bar{\alpha}_{t+1}^* = a_t + \gamma^*(y_t - a_t)$. By the same argument as for $\bar{\alpha}_t$ we find that $\mathrm{Var}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1})$ is minimised when $\gamma^* = P_t/(P_t + \sigma_\varepsilon^2)$ giving

$$\bar{\alpha}_{t+1}^* = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \tag{2.23}$$

$$\mathrm{Var}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1}) = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2. \tag{2.24}$$

We have therefore shown that the estimates of $\bar{\alpha}_t$ and $\bar{\alpha}_{t+1}$ given by the MVLUE approach and their variances are exactly the same as the values $a_{t|t}$, a_{t+1}, $P_{t|t}$ and P_{t+1} in (2.11) to (2.14) that are obtained by assuming normality, both from a classical and from a Bayesian standpoint. It follows that the values given by the Kalman filter recursion (2.15) are MVLUE. We shall show in Subsection 4.3.1 that the same is true for the general linear Gaussian state space model (4.12).

2.2.5 Illustration

In this subsection we shall illustrate the output of the Kalman filter using observations from the river Nile. The data set consists of a series of readings of the annual flow volume at Aswan from 1871 to 1970. The series has been analysed by Cobb (1978) and Balke (1993). We analyse the data using the local level model (2.3) with $a_1 = 0$, $P_1 = 10^7$, $\sigma_\varepsilon^2 = 15,099$ and $\sigma_\eta^2 = 1,469.1$. The values for a_1 and P_1 were chosen arbitrarily for illustrative purposes. The values for σ_ε^2 and σ_η^2 are the maximum likelihood estimates which we obtain in Subsection 2.10.3. The values of a_t together with the raw data, P_t, v_t and F_t, for $t = 2, \dots, n$, given by the Kalman filter, are presented graphically in Fig. 2.1.

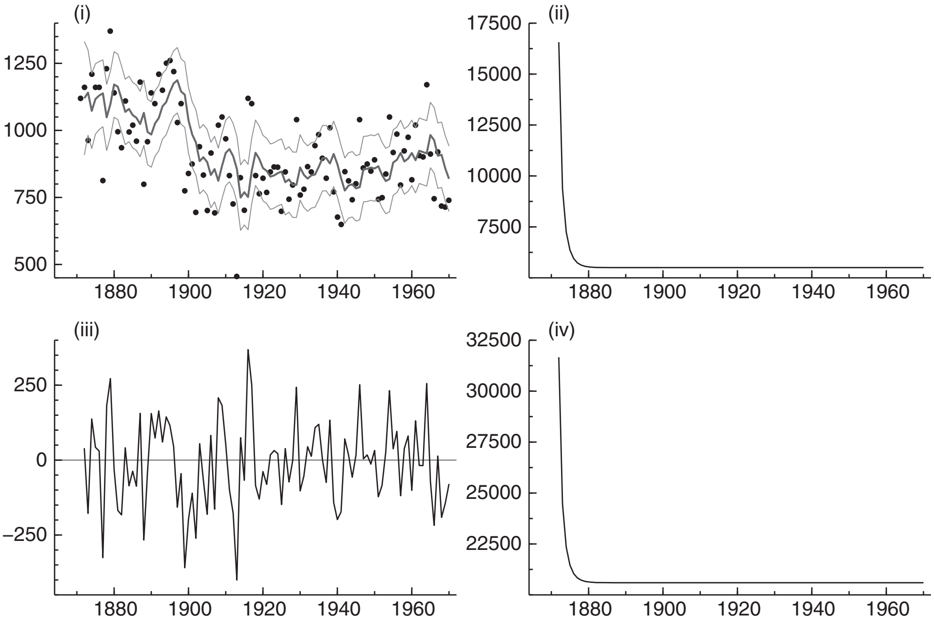

Fig. 2.1 Nile data and output of Kalman filter: (i) data (dots), filtered state a_t (solid line) and its 90% confidence intervals (light solid lines); (ii) filtered state variance P_t; (iii) prediction errors v_t; (iv) prediction variance F_t.

The most obvious feature of the four graphs is that P_t and F_t converge rapidly to constant values which confirms that the local level model has a steady state solution; for discussion of the concept of a steady state see Section 2.11. However, it was found that the fitted local level model converged numerically to a steady state in around 25 updates of P_t although the graph of P_t seems to suggest that the steady state was obtained after around 10 updates.

2.3 Forecast errors

The Kalman filter residual $v_t = y_t - a_t$ and its variance F_t are the one-step ahead forecast error and the one-step ahead forecast error variance of y_t given Y_{t-1} as defined in Section 2.2. The forecast errors $v_1, \dots, v_n$ are sometimes called *innovations* because they represent the new part of y_t that cannot be predicted from the past for $t = 1, \dots, n$. We shall make use of v_t and F_t for a variety of results in the next sections. It is therefore important to study them in detail.

2.3.1 Cholesky decomposition

First we show that $v_1, \dots, v_n$ are mutually independent. The joint density of $y_1, \dots, y_n$ is

$$p(y_1, \dots, y_n) = p(y_1) \prod_{t=2}^{n} p(y_t | Y_{t-1}). \tag{2.25}$$

We then transform from $y_1, \dots, y_n$ to $v_1, \dots, v_n$. Since each v_t equals y_t minus a linear function of $y_1, \dots, y_{t-1}$ for $t = 2, \dots, n$, the Jacobian is one. From (2.25) and making the substitution we have

$$p(v_1, \dots, v_n) = \prod_{t=1}^{n} p(v_t), \tag{2.26}$$

since $p(v_1) = p(y_1)$ and $p(v_t) = p(y_t | Y_{t-1})$ for $t = 2, \dots, n$. Consequently, the v_t's are independently distributed.

We next show that the forecast errors v_t are effectively obtained from a Cholesky decomposition of the observation vector Y_n. The Kalman filter recursions compute the forecast error v_t as a linear function of the initial mean a_1 and the observations $y_1, \dots, y_t$ since

$$\begin{aligned}
v_1 &= y_1 - a_1, \\
v_2 &= y_2 - a_1 - K_1(y_1 - a_1), \\
v_3 &= y_3 - a_1 - K_2(y_2 - a_1) - K_1(1 - K_2)(y_1 - a_1), \quad \text{and so on.}
\end{aligned}$$

It should be noted that K_t does not depend on the initial mean a_1 and the observations $y_1, \dots, y_n$; it depends only on the initial state variance P_1 and the disturbance variances σ^2_ε and σ^2_η. Using the definitions in (2.4), we have

$$v = C(Y_n - \mathbf{1}a_1), \quad \text{with} \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

where matrix C is the lower triangular matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ c_{21} & 1 & 0 & & 0 \\ c_{31} & c_{32} & 1 & & 0 \\ & & & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & 1 \end{bmatrix},$$

$$\begin{aligned} c_{i,i-1} &= -K_{i-1}, \\ c_{ij} &= -(1 - K_{i-1})(1 - K_{i-2}) \cdots (1 - K_{j+1})K_j, \end{aligned} \tag{2.27}$$

for $i = 2, \ldots, n$ and $j = 1, \ldots, i - 2$. The distribution of v is therefore

$$v \sim \mathrm{N}(0, C\Omega C'), \tag{2.28}$$

where $\Omega = \mathrm{Var}(Y_n)$ as given by (2.4). On the other hand we know from (2.7), (2.15) and (2.26) that $\mathrm{E}(v_t) = 0$, $\mathrm{Var}(v_t) = F_t$ and $\mathrm{Cov}(v_t, v_j) = 0$, for $t, j = 1, \ldots, n$ and $t \neq j$; therefore,

$$v \sim \mathrm{N}(0, F), \quad \text{with} \quad F = \begin{bmatrix} F_1 & 0 & 0 & & 0 \\ 0 & F_2 & 0 & & 0 \\ 0 & 0 & F_3 & & 0 \\ & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & F_n \end{bmatrix}.$$

It follows that $C\Omega C' = F$. The transformation of a symmetric positive definite matrix (say Ω) into a diagonal matrix (say F) using a lower triangular matrix (say C) by means of the relation $C\Omega C' = F$ is known as the *Cholesky decomposition* of the symmetric matrix. The Kalman filter can therefore be regarded as essentially a Cholesky decomposition of the variance matrix implied by the local level model (2.3). This result is important for understanding the role of the Kalman filter and it will be used further in Subsections 2.5.4 and 2.10.1. Note also that $F^{-1} = (C')^{-1}\Omega^{-1}C^{-1}$ so we have $\Omega^{-1} = C'F^{-1}C$.

2.3.2 Error recursions

Define the *state estimation error* as

$$x_t = \alpha_t - a_t, \quad \text{with} \quad \mathrm{Var}(x_t) = P_t. \tag{2.29}$$

We now show that the state estimation errors x_t and forecast errors v_t are linear functions of the initial state error x_1 and the disturbances ε_t and η_t analogously to the way that α_t and y_t are linear functions of the initial state and the disturbances for $t = 1, \ldots, n$. It follows directly from the Kalman filter relations (2.15) that

$$\begin{aligned} v_t &= y_t - a_t \\ &= \alpha_t + \varepsilon_t - a_t \\ &= x_t + \varepsilon_t, \end{aligned}$$

and

$$\begin{aligned} x_{t+1} &= \alpha_{t+1} - a_{t+1} \\ &= \alpha_t + \eta_t - a_t - K_t v_t \\ &= x_t + \eta_t - K_t(x_t + \varepsilon_t) \\ &= L_t x_t + \eta_t - K_t \varepsilon_t, \end{aligned}$$

where

$$L_t = 1 - K_t = \sigma_\varepsilon^2 / F_t. \tag{2.30}$$

Thus analogously to the local level model relations

$$y_t = \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = \alpha_t + \eta_t,$$

we have the error relations

$$v_t = x_t + \varepsilon_t, \qquad x_{t+1} = L_t x_t + \eta_t - K_t \varepsilon_t, \qquad t = 1, \ldots, n, \tag{2.31}$$

with $x_1 = \alpha_1 - a_1$. These relations will be used in the next section. We note that P_t, F_t, K_t and L_t do not depend on the initial state mean a_1 or the observations $y_1, \ldots, y_n$ but only on the initial state variance P_1 and the disturbance variances σ_ε^2 and σ_η^2. We note also that the recursion for P_{t+1} in (2.15) can alternatively be derived by

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(x_{t+1}) = \mathrm{Cov}(x_{t+1}, \alpha_{t+1}) = \mathrm{Cov}(x_{t+1}, \alpha_t + \eta_t) \\ &= L_t \mathrm{Cov}(x_t, \alpha_t + \eta_t) + \mathrm{Cov}(\eta_t, \alpha_t + \eta_t) - K_t \mathrm{Cov}(\varepsilon_t, \alpha_t + \eta_t) \\ &= L_t P_t + \sigma_\eta^2 \;=\; P_t(1 - K_t) + \sigma_\eta^2. \end{aligned}$$

2.4 State smoothing

2.4.1 Smoothed state

We now consider the estimation of $\alpha_1, \ldots, \alpha_n$ in model (2.3) given the entire sample Y_n. Since all distributions are normal, the conditional density of α_t given Y_n

is $\mathrm{N}(\hat{\alpha}_t, V_t)$ where $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ and $V_t = \mathrm{Var}(\alpha_t|Y_n)$. We call $\hat{\alpha}_t$ the *smoothed state*, V_t the *smoothed state variance* and the operation of calculating $\hat{\alpha}_1, \dots, \hat{\alpha}_n$ *state smoothing*. Similar arguments to those in Subsections 2.2.3 and 2.2.4 can be used to justify the same formulae for a Bayesian analysis and a MVLUE approach.

The forecast errors $v_1, \dots, v_n$ are mutually independent and $v_t, \dots, v_n$ are independent of $y_1, \dots, y_{t-1}$ with zero means. Moreover, when $y_1, \dots, y_n$ are fixed, Y_{t-1} and $v_t, \dots, v_n$ are fixed and vice versa. By an extension of the lemma of Subsection 2.2.2 to the multivariate case we have the regression relation for the conditional distribution of α_t and $v_t, \dots, v_n$ given Y_{t-1},

$$\hat{\alpha}_t = a_t + \sum_{j=t}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} v_j. \tag{2.32}$$

Now $\mathrm{Cov}(\alpha_t, v_j) = \mathrm{Cov}(x_t, v_j)$ for $j = t, \dots, n$, and

$$\begin{aligned}\mathrm{Cov}(x_t, v_t) &= \mathrm{E}[x_t(x_t + \varepsilon_t)] = \mathrm{Var}(x_t) = P_t,\\ \mathrm{Cov}(x_t, v_{t+1}) &= \mathrm{E}[x_t(x_{t+1} + \varepsilon_{t+1})] = \mathrm{E}[x_t(L_t x_t + \eta_t - K_t \varepsilon_t)] = P_t L_t,\end{aligned}$$

where x_t is defined in (2.29) and L_t in (2.30). Similarly,

$$\begin{aligned}\mathrm{Cov}(x_t, v_{t+2}) &= P_t L_t L_{t+1},\\ &\ \vdots\\ \mathrm{Cov}(x_t, v_n) &= P_t L_t L_{t+1} \dots L_{n-1}.\end{aligned} \tag{2.33}$$

Substituting in (2.32) gives

$$\begin{aligned}\hat{\alpha}_t &= a_t + P_t \frac{v_t}{F_t} + P_t L_t \frac{v_{t+1}}{F_{t+1}} + P_t L_t L_{t+1} \frac{v_{t+2}}{F_{t+2}} + \dots + P_t L_t L_{t+1} \dots L_{n-1} \frac{v_n}{F_n}\\ &= a_t + P_t r_{t-1},\end{aligned}$$

where

$$\begin{aligned}r_{t-1} = \frac{v_t}{F_t} &+ L_t \frac{v_{t+1}}{F_{t+1}} + L_t L_{t+1} \frac{v_{t+2}}{F_{t+2}} + L_t L_{t+1} L_{t+2} \frac{v_{t+3}}{F_{t+3}} + \dots +\\ &+ L_t L_{t+1} \dots L_{n-1} \frac{v_n}{F_n}\end{aligned} \tag{2.34}$$

is a weighted sum of innovations after $t-1$. The value of this at time t is

$$\begin{aligned}r_t = \frac{v_{t+1}}{F_{t+1}} &+ L_{t+1} \frac{v_{t+2}}{F_{t+2}} + L_{t+1} L_{t+2} \frac{v_{t+3}}{F_{t+3}} + \cdots\\ &+ L_{t+1} L_{t+2} \dots L_{n-1} \frac{v_n}{F_n}.\end{aligned} \tag{2.35}$$

Obviously, $r_n = 0$ since no observations are available after time n. By substituting from (2.35) into (2.34), it follows that the values of r_{t-1} can be evaluated using the backwards recursion

$$r_{t-1} = \frac{v_t}{F_t} + L_t r_t, \tag{2.36}$$

with $r_n = 0$, for $t = n, n-1, \ldots, 1$. The smoothed state can therefore be calculated by the backwards recursion

$$r_{t-1} = F_t^{-1} v_t + L_t r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad t = n, \ldots, 1, \tag{2.37}$$

with $r_n = 0$. The relations in (2.37) are collectively called the *state smoothing recursion.*

2.4.2 Smoothed state variance

The error variance of the smoothed state, $V_t = \text{Var}(\alpha_t | Y_n)$, is derived in a similar way. By the multivariate extension of the regression lemma of Subsection 2.2.2 applied to the conditional distribution of α_t and $v_t, \ldots, v_n$ given Y_{t-1} we have

$$\begin{aligned} V_t = \text{Var}(\alpha_t | Y_n) &= \text{Var}(\alpha_t | Y_{t-1}, v_t, \ldots, v_n) \\ &= P_t - \sum_{j=t}^{n} [\text{Cov}(\alpha_t, v_j)]^2 F_j^{-1}, \end{aligned} \tag{2.38}$$

where the expressions for $\text{Cov}(\alpha_t, v_j)$ $\text{Cov}(x_t, v_j)$ are given by (2.33). Substituting these into (2.38) leads to

$$\begin{aligned} V_t &= P_t - P_t^2 \frac{1}{F_t} - P_t^2 L_t^2 \frac{1}{F_{t+1}} - P_t^2 L_t^2 L_{t+1}^2 \frac{1}{F_{t+2}} - \cdots - P_t^2 L_t^2 L_{t+1}^2 \cdots L_{n-1}^2 \frac{1}{F_n} \\ &= P_t - P_t^2 N_{t-1}, \end{aligned} \tag{2.39}$$

where

$$\begin{aligned} N_{t-1} = \frac{1}{F_t} &+ L_t^2 \frac{1}{F_{t+1}} + L_t^2 L_{t+1}^2 \frac{1}{F_{t+2}} + L_t^2 L_{t+1}^2 L_{t+2}^2 \frac{1}{F_{t+3}} + \cdots \\ &+ L_t^2 L_{t+1}^2 \cdots L_{n-1}^2 \frac{1}{F_n}, \end{aligned} \tag{2.40}$$

is a weighted sum of the inverse variances of innovations after time $t-1$. Its value at time t is

$$N_t = \frac{1}{F_{t+1}} + L_{t+1}^2 \frac{1}{F_{t+2}} + L_{t+1}^2 L_{t+2}^2 \frac{1}{F_{t+3}} + \cdots + L_{t+1}^2 L_{t+2}^2 \cdots L_{n-1}^2 \frac{1}{F_n}, \tag{2.41}$$

and, obviously, $N_n = 0$ since no variances are available after time n. Substituting from (2.41) into (2.40) it follows that the value for N_{t-1} can be calculated using the backwards recursion

$$N_{t-1} = \frac{1}{F_t} + L_t^2 N_t, \tag{2.42}$$

with $N_n = 0$, for $t = n, n-1, \ldots, 1$. We observe from (2.35) and (2.41) that $N_t = \mathrm{Var}(r_t)$ since the forecast errors v_t are independent.

By combining these results, the error variance of the smoothed state can be calculated by the backwards recursion

$$N_{t-1} = F_t^{-1} + L_t^2 N_t, \qquad V_t = P_t - P_t^2 N_{t-1}, \qquad t = n, \ldots, 1, \tag{2.43}$$

with $N_n = 0$. The relations in (2.43) are collectively called the *state variance smoothing recursion*. From the standard error $\sqrt{V_t}$ of $\hat{\alpha}_t$ we can construct confidence intervals for α_t for $t = 1, \ldots, n$. It is also possible to derive the smoothed covariances between the states, that is, $\mathrm{Cov}(\alpha_t, \alpha_s | Y_n)$, $t \neq s$, using similar arguments. We shall not give them here but will derive them for the general case in Section 4.7.

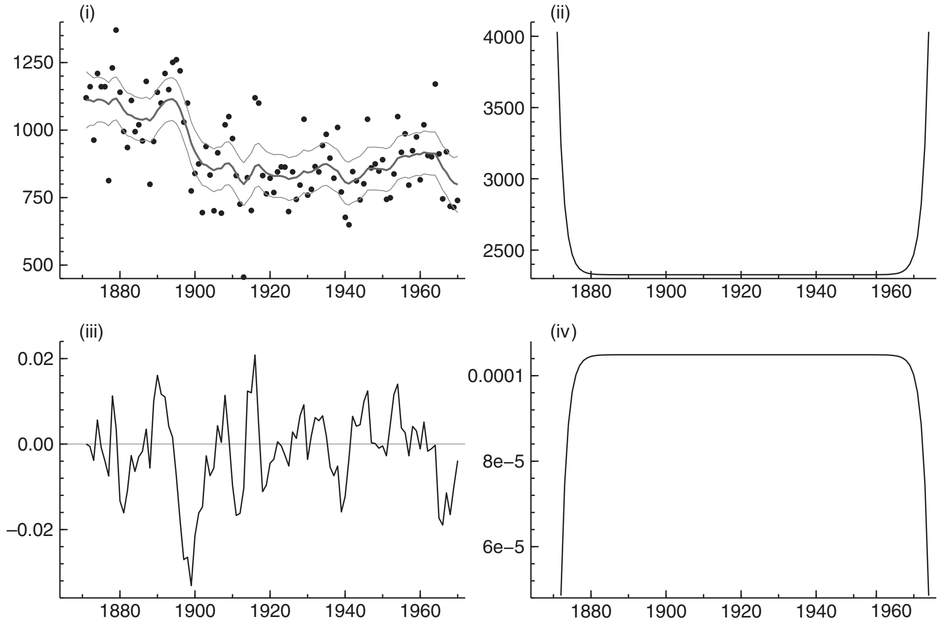

Fig. 2.2 Nile data and output of state smoothing recursion: (i) data (dots), smoothed state $\hat{\alpha}_t$ and its 90% confidence intervals; (ii) smoothed state variance V_t; (iii) smoothing cumulant r_t; (iv) smoothing variance cumulant N_t.

2.4.3 Illustration

We now show the results of state smoothing for the Nile data of Subsection 2.2.5 using the same local level model. The Kalman filter is applied first and the output v_t, F_t, a_t and P_t is stored for $t = 1, \ldots, n$. Figure 2.2 presents the output of the backwards smoothing recursions (2.37) and (2.43); that is $\hat{\alpha}_t$, V_t, r_t and N_t. The plot of $\hat{\alpha}_t$ includes the 90% confidence bands for α_t. The graph of $\mathrm{Var}(\alpha_t|Y_n)$ shows that the conditional variance of α_t is larger at the beginning and end of the sample, as it obviously should be on intuitive grounds. Comparing the graphs of a_t and $\hat{\alpha}_t$ in Fig. 2.1 and 2.2, we see that the graph of $\hat{\alpha}_t$ is much smoother than that of a_t, except at time points close to the end of the series, as it should be.

2.5 Disturbance smoothing

In this section we consider the calculation of the smoothed observation disturbance $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n) = y_t - \hat{\alpha}_t$ and the smoothed state disturbance $\hat{\eta}_t = \mathrm{E}(\eta_t|Y_n) = \hat{\alpha}_{t+1} - \hat{\alpha}_t$ together with their error variances. Of course, these could be calculated directly from a knowledge of $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ and covariances $\mathrm{Cov}(\alpha_t, \alpha_j|Y_n)$ for $j \leq t$. However, it turns out to be computationally advantageous to compute them from r_t and N_t without first calculating $\hat{\alpha}_t$, particularly for the general model discussed in Chapter 4. The merits of smoothed disturbances are discussed in Section 4.5. For example, the estimates $\hat{\varepsilon}_t$ and $\hat{\eta}_t$ are useful for detecting outliers and structural breaks, respectively; see Subsection 2.12.2. For the sake of brevity we shall restrict the treatment of this section to classical inference based on the assumption of normality as in model (2.3).

In order to economise on the amount of algebra in this chapter we shall present the required recursions for the local level model without proof, referring the reader to Section 4.5 for derivations of the analogous recursions for the general model.

2.5.1 Smoothed observation disturbances

From (4.58) in Section 4.5.1, the smoothed observation disturbance $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n)$ is calculated by

$$\hat{\varepsilon}_t = \sigma_\varepsilon^2 u_t, \qquad t = n, \ldots, 1, \tag{2.44}$$

where

$$u_t = F_t^{-1} v_t - K_t r_t, \tag{2.45}$$

and where the recursion for r_t is given by (2.36). The scalar u_t is referred to as the *smoothing error*. Similarly, from (4.65) in Section 4.5.2, the smoothed variance $\mathrm{Var}(\varepsilon_t|Y_n)$ is obtained by

$$\mathrm{Var}(\varepsilon_t|Y_n) = \sigma_\varepsilon^2 - \sigma_\varepsilon^4 D_t, \qquad t = n, \ldots, 1, \tag{2.46}$$

where

$$D_t = F_t^{-1} + K_t^2 N_t, \tag{2.47}$$

and where the recursion for N_t is given by (2.42). Since from (2.35) v_t is independent of r_t, and $\mathrm{Var}(r_t) = N_t$, we have

$$\mathrm{Var}(u_t) = \mathrm{Var}\big(F_t^{-1} v_t - K_t r_t\big) = F_t^{-2}\,\mathrm{Var}(v_t) + K_t^2\,\mathrm{Var}(r_t) = D_t.$$

Consequently, from (2.44) we obtain $\mathrm{Var}(\hat{\varepsilon}_t) = \sigma_\varepsilon^4 D_t$.

Note that the methods for calculating $\hat{\alpha}_t$ and $\hat{\varepsilon}_t$ are consistent since $K_t = P_t F_t^{-1}$, $L_t = 1 - K_t = \sigma_\varepsilon^2 F_t^{-1}$ and

$$\begin{aligned}
\hat{\varepsilon}_t &= y_t - \hat{\alpha}_t \\
&= y_t - a_t - P_t r_{t-1} \\
&= v_t - P_t\big(F_t^{-1} v_t + L_t r_t\big) \\
&= F_t^{-1} v_t (F_t - P_t) - \sigma_\varepsilon^2 P_t F_t^{-1} r_t \\
&= \sigma_\varepsilon^2\big(F_t^{-1} v_t - K_t r_t\big), \qquad t = n, \ldots, 1.
\end{aligned}$$

Similar equivalences can be shown for V_t and $\mathrm{Var}(\varepsilon_t|Y_n)$.

2.5.2 Smoothed state disturbances

From (4.63) in Subsection 4.5.1, the smoothed mean of the disturbance $\hat{\eta}_t = \mathrm{E}(\eta_t|Y_n)$ is calculated by

$$\hat{\eta}_t = \sigma_\eta^2 r_t, \qquad t = n, \ldots, 1, \tag{2.48}$$

where the recursion for r_t is given by (2.36). Similarly, from (4.68) in Subsection 4.5.2, the smoothed variance $\mathrm{Var}(\eta_t|Y_n)$ is computed by

$$\mathrm{Var}(\eta_t|Y_n) = \sigma_\eta^2 - \sigma_\eta^4 N_t, \qquad t = n, \ldots, 1, \tag{2.49}$$

where the recursion for N_t is given by (2.42). Since $\mathrm{Var}(r_t) = N_t$, we have $\mathrm{Var}(\hat{\eta}_t) = \sigma_\eta^4 N_t$. These results are interesting because they give an interpretation to the values r_t and N_t; they are the scaled smoothed estimator of $\eta_t = \alpha_{t+1} - \alpha_t$ and its unconditional variance, respectively.

The method of calculating $\hat{\eta}_t$ is consistent with the definition $\eta_t = \alpha_{t+1} - \alpha_t$ since

$$\begin{aligned}
\hat{\eta}_t &= \hat{\alpha}_{t+1} - \hat{\alpha}_t \\
&= a_{t+1} + P_{t+1} r_t - a_t - P_t r_{t-1} \\
&= a_t + K_t v_t - a_t + P_t L_t r_t + \sigma_\eta^2 r_t - P_t\big(F_t^{-1} v_t + L_t r_t\big) \\
&= \sigma_\eta^2 r_t.
\end{aligned}$$

Similar consistencies can be shown for N_t and $\mathrm{Var}(\eta_t|Y_n)$.

2.5.3 Illustration

The smoothed disturbances and their related variances for the analysis of the Nile data and the local level model introduced in Subsection 2.2.5 are calculated by the above recursions and presented in Fig. 2.3. We note from the graphs of $\text{Var}(\varepsilon_t|Y_n)$ and $\text{Var}(\eta_t|Y_n)$ the extent that these conditional variances are larger at the beginning and end of the sample. Obviously, the plot of r_t in Fig. 2.2 and the plot of $\hat{\eta}_t$ in Fig. 2.3 are the same apart from a different scale.

2.5.4 Cholesky decomposition and smoothing

We now consider the calculation of $\hat{\varepsilon}_t = \text{E}(\varepsilon_t|Y_n)$ by direct regression of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ on the observation vector Y_n defined in (2.4) to obtain $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$, that is,

$$\begin{aligned}\hat{\varepsilon} &= \text{E}(\varepsilon) + \text{Cov}(\varepsilon, Y_n)\,\text{Var}(Y_n)^{-1}[Y_n - \text{E}(Y_n)]\\ &= \text{Cov}(\varepsilon, Y_n)\Omega^{-1}(Y_n - 1a_1),\end{aligned}$$

where, here and later, when necessary, we treat Y_n as the observation vector $(y_1, \dots, y_n)'$. It is obvious from (2.6) that $\text{Cov}(\varepsilon, Y_n) = \sigma_\varepsilon^2 I_n$; also, from the

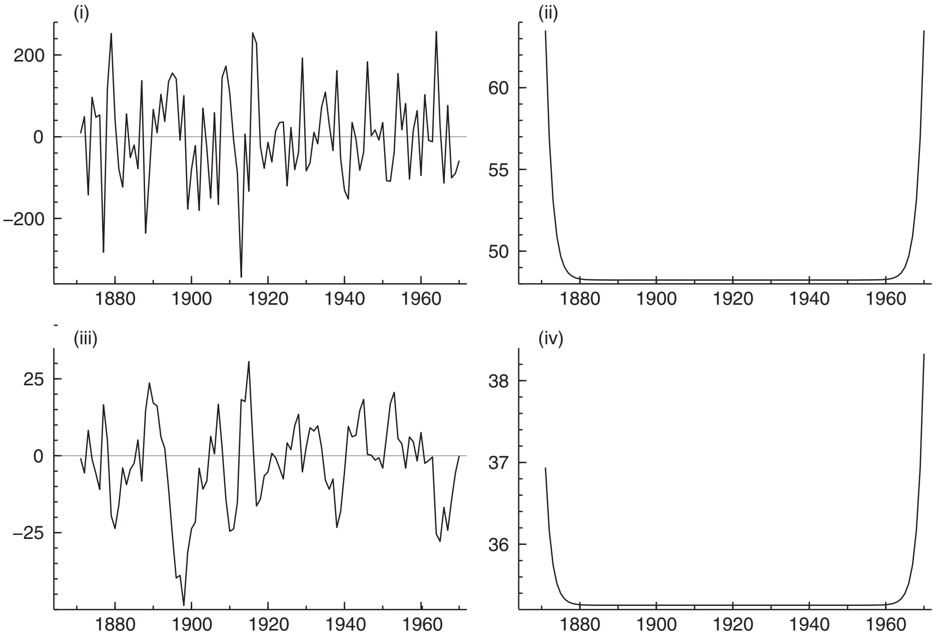

Fig. 2.3 Output of disturbance smoothing recursion: (i) observation error $\hat{\varepsilon}_t$; (ii) observation error variance $\text{Var}(\varepsilon_t|Y_n)$; (iii) state error $\hat{\eta}_t$; (iv) state error variance $\text{Var}(\eta_t|Y_n)$.

Cholesky decomposition considered in Subsection 2.3.1 we have $\Omega^{-1} = C'F^{-1}C$ and $C(Y_n - 1a_1) = v$. We therefore have

$$\hat{\varepsilon} = \sigma_\varepsilon^2 C' F^{-1} v,$$

which, by consulting the definitions of the lower triangular elements of C in (2.27), also leads to the disturbance equations (2.44) and (2.45). Thus

$$\hat{\varepsilon} = \sigma_\varepsilon^2 u, \qquad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

where

$$u = C'F^{-1}v \quad \text{with} \quad v = C(Y_n - 1a_1).$$

It follows that

$$u = C'F^{-1}C(Y_n - 1a_1) = \Omega^{-1}(Y_n - 1a_1), \tag{2.50}$$

where $\Omega = \mathrm{Var}(Y_n)$ and $F = C\Omega C'$, as is consistent with standard regression theory.

2.6 Simulation

It is simple to draw samples generated by the local level model (2.3). We first draw the random normal deviates

$$\varepsilon_t^+ \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \qquad \eta_t^+ \sim \mathrm{N}\big(0, \sigma_\eta^2\big), \qquad t = 1, \ldots, n. \tag{2.51}$$

Then we generate observations using the local level recursion as follows

$$y_t^+ = \alpha_t^+ + \varepsilon_t^+, \qquad \alpha_{t+1}^+ = \alpha_t^+ + \eta_t^+, \qquad t = 1, \ldots, n, \tag{2.52}$$

for some starting value α_1^+.

For the implementation of classical and Bayesian simulation methods and for the treatment of nonlinear and non-Gaussian models, which will be discussed in Part II of this book, we may require samples generated by the local level model conditional on the observed time series $y_1, \ldots, y_n$. Such samples can be obtained by use of the simulation smoother developed for the general linear Gaussian state space model in Section 4.9. For the local level model, a simulated sample for the disturbances ε_t, $t = 1, \ldots, n$, given the observations $y_1, \ldots, y_n$ can be obtained using the method of mean corrections as discussed in Subsection 4.9.1. It requires the drawing of the samples ε_t^+ and η_t^+ as in (2.51) and using them to draw y_t^+ as in (2.52). Then, a conditional draw for ε_t given Y_n is obtained by

$$\tilde{\varepsilon}_t = \varepsilon_t^+ - \hat{\varepsilon}_t^+ + \hat{\varepsilon}_t, \tag{2.53}$$

for $t = 1, \ldots, n$, where $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t | Y_n)$ and $\hat{\varepsilon}_t^+ = \mathrm{E}(\varepsilon_t | Y_n^+)$ with $Y_n^+ = (y_1^+, \ldots, y_n^+)'$ and where both are computed via the disturbance smoothing equations (2.44) and (2.45). This set of computations is sufficient to obtain a conditional draw of ε_t given Y_n, for $t = 1, \ldots, n$. Given a sample $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$, we obtain simulated samples for α_t and η_t via the relations

$$\tilde{\alpha}_t = y_t - \tilde{\varepsilon}_t, \qquad \tilde{\eta}_t = \tilde{\alpha}_{t+1} - \tilde{\alpha}_t,$$

for $t = 1, \ldots, n$.

2.6.1 Illustration

To illustrate the difference between simulating a sample from the local level model unconditionally and simulating a sample conditional on the observations, we consider the Nile data and the local level model of Subsection 2.2.5. In Fig. 2.4 (i) we present the smoothed state $\hat{\alpha}_t$ and a sample generated by the local level model unconditionally. The two series have seemingly nothing in common. In the next panel, again the smoothed state is presented but now together with a sample

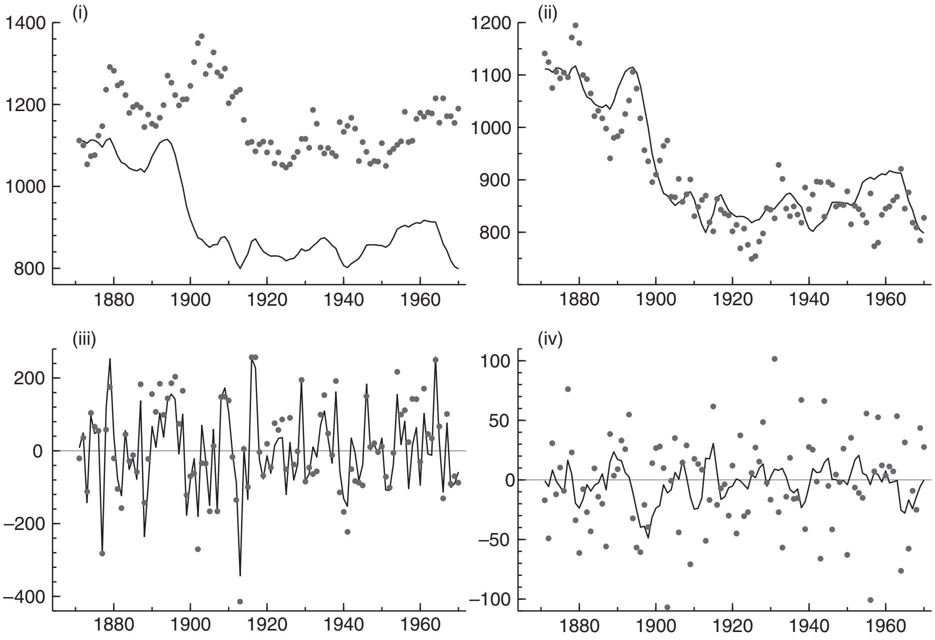

Fig. 2.4 Simulation: (i) smoothed state $\hat{\alpha}_t$ (solid line) and sample α_t^+ (dots); (ii) smoothed state $\hat{\alpha}_t$ (solid line) and sample $\tilde{\alpha}_t$ (dots); (iii) smoothed observation error $\hat{\varepsilon}_t$ (solid line) and sample $\tilde{\varepsilon}_t$ (dots); (iv) smoothed state error $\hat{\eta}_t$ (solid line) and sample $\tilde{\eta}_t$ (dots).

generated conditional on the observations. Here we see that the generated sample is much closer to $\hat{\alpha}_t$. The remaining two panels present the smoothed disturbances together with a sample from the corresponding disturbances conditional on the observations.

2.7 Missing observations

A considerable advantage of the state space approach is the ease with which missing observations can be dealt with. Suppose we have a local level model where observations y_j, with $j = \tau, \dots, \tau^* - 1$, are missing for $1 < \tau < \tau^* \leq n$. For the filtering stage, the most obvious way to deal with the situation is to define a new series y_t^* where $y_t^* = y_t$ for $t = 1, \dots, \tau - 1$ and $y_t^* = y_{t+\tau^*-\tau}$ for $t = \tau, \dots, n^*$ with $n^* = n - (\tau^* - \tau)$. The model for y_t^* with time scale $t = 1, \dots, n^*$ is then the same as (2.3) with $y_t = y_t^*$ except that $\alpha_\tau = \alpha_{\tau-1} + \eta_{\tau-1}$ where $\eta_{\tau-1} \sim \mathrm{N}[0, (\tau^* - \tau)\sigma_\eta^2]$. Filtering for this model can be treated by the methods developed in Chapter 4 for the general state space model. The treatment is readily extended if more than one group of observations is missing.

It is, however, easier and more transparent to proceed as follows, using the original time domain. For filtering at times $t = \tau, \dots, \tau^* - 1$, we have

$$\begin{aligned}
\mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{\tau-1}) &= \mathrm{E}\left(\alpha_\tau + \sum_{j=\tau}^{t-1}\eta_j \middle| Y_{\tau-1}\right) = a_\tau, \\
\mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(\alpha_{t+1}|Y_{\tau-1}) &= \mathrm{E}\left(\alpha_\tau + \sum_{j=\tau}^{t}\eta_j \middle| Y_{\tau-1}\right) = a_\tau, \\
\mathrm{Var}(\alpha_t|Y_t) = \mathrm{Var}(\alpha_t|Y_{\tau-1}) &= \mathrm{Var}\left(\alpha_\tau + \sum_{j=\tau}^{t-1}\eta_j \middle| Y_{\tau-1}\right) = P_\tau + (t-\tau)\sigma_\eta^2, \\
\mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}(\alpha_t|Y_{\tau-1}) &= \mathrm{Var}\left(\alpha_\tau + \sum_{j=\tau}^{t}\eta_j \middle| Y_{\tau-1}\right) = P_\tau + (t-\tau+1)\sigma_\eta^2.
\end{aligned}$$

We can compute them recursively by

$$\begin{aligned}
a_{t|t} &= a_t, & P_{t|t} &= P_t, \\
a_{t+1} &= a_t, & P_{t+1} &= P_t + \sigma_\eta^2,
\end{aligned} \qquad t = \tau, \dots, \tau^* - 1, \qquad (2.54)$$

the remaining values a_t and P_t being given as before by (2.15) for $t = 1, \dots, \tau - 1$ and $t = \tau^*, \dots, n$. The consequence is that we can use the original filter (2.15) for all t by taking $K_t = 0$ at the missing time points. The same procedure is used when more than one group of observations is missing. It follows that allowing for missing observations when using the Kalman filter is extremely simple.

The forecast error recursions from which we derive the smoothing recursions are given by (2.31). These error-updating equations at the missing time points become

$$v_t = x_t + \varepsilon_t, \qquad x_{t+1} = x_t + \eta_t, \qquad t = \tau, \ldots, \tau^* - 1,$$

since $K_t = 0$ and therefore $L_t = 1$. The covariances between the state at the missing time points and the innovations after the missing period are given by

$$\begin{aligned} \mathrm{Cov}(\alpha_t, v_{\tau^*}) &= P_t, \\ \mathrm{Cov}(\alpha_t, v_j) &= P_t L_{\tau^*} L_{\tau^*+1} \ldots L_{j-1}, \quad j = \tau^* + 1, \ldots, n, \quad t = \tau, \ldots, \tau^* - 1. \end{aligned}$$

By deleting the terms associated with the missing time points, the state smoothing equation (2.32) for the missing time points becomes

$$\hat{\alpha}_t = a_t + \sum_{j=\tau^*}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} v_j, \qquad t = \tau, \ldots, \tau^* - 1.$$

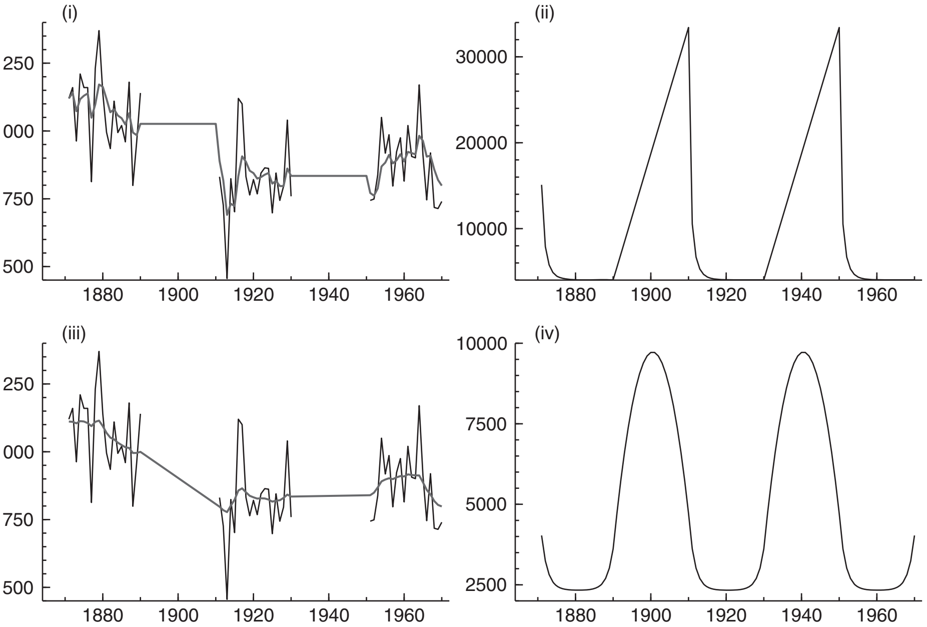

Fig. 2.5 Filtering and smoothing output when observations are missing: (i) data and filtered state a_t (extrapolation); (ii) filtered state variance P_t; (iii) data and smoothed state $\hat{\alpha}_t$ (interpolation); (iv) smoothed state variance V_t.

Substituting the covariance terms into this and taking into account the definition (2.34) leads directly to

$$r_{t-1} = r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad t = \tau, \dots, \tau^* - 1. \tag{2.55}$$

The consequence is that we can use the original state smoother (2.37) for all t by taking $K_t = 0$, and hence $L_t = 1$, at the missing time points. This device applies to any missing observation within the sample period. In the same way the equations for the variance of the state error and the smoothed disturbances can be obtained by putting $K_t = 0$ at missing time points.

2.7.1 Illustration

Here we consider the Nile data and the same local level model as before; however, we treat the observations at time points $21, \dots, 40$ and $61, \dots, 80$ as missing. The Kalman filter is applied first and the output v_t, F_t, a_t and P_t is stored for $t = 1, \dots, n$. Then, the state smoothing recursions are applied. The first two graphs in Fig. 2.5 are the Kalman filter values of a_t and P_t, respectively. The last two graphs are the smoothing output $\hat{\alpha}_t$ and V_t, respectively.

Note that the application of the Kalman filter to missing observations can be regarded as extrapolation of the series to the missing time points, while smoothing at these points is effectively interpolation.

2.8 Forecasting

Let $\bar{y}_{n+j}$ be the minimum mean square error forecast of y_{n+j} given the time series $y_1, \dots, y_n$ for $j = 1, 2, \dots, J$ with J as some pre-defined positive integer. By minimum mean square error forecast here we mean the function $\bar{y}_{n+j}$ of $y_1, \dots, y_n$ which minimises $\mathrm{E}[(y_{n+j} - \bar{y}_{n+j})^2 | Y_n]$. Then $\bar{y}_{n+j} = \mathrm{E}(y_{n+j}|Y_n)$. This follows immediately from the well-known result that if x is a random variable with mean μ the value of λ that minimises $\mathrm{E}(x-\lambda)^2$ is $\lambda = \mu$; see Exercise 4.14.3. The variance of the forecast error is denoted by $\bar{F}_{n+j} = \mathrm{Var}(y_{n+j}|Y_n)$. The theory of forecasting for the local level model turns out to be surprisingly simple; we merely regard forecasting as filtering the observations $y_1, \dots, y_n, y_{n+1}, \dots, y_{n+J}$ using the recursion (2.15) and treating the last J observations $y_{n+1}, \dots, y_{n+J}$ as missing, that is, taking $K_t = 0$ in (2.15).

Letting $\bar{a}_{n+j} = \mathrm{E}(\alpha_{n+j}|Y_n)$ and $\bar{P}_{n+j} = \mathrm{Var}(\alpha_{n+j}|Y_n)$, it follows immediately from equation (2.54) with $\tau = n + 1$ and $\tau^* = n + J$ in §2.7 that

$$\bar{a}_{n+j+1} = \bar{a}_{n+j}, \qquad \bar{P}_{n+j+1} = \bar{P}_{n+j} + \sigma_\eta^2, \qquad j = 1, \dots, J - 1,$$

with $\bar{a}_{n+1} = a_{n+1}$ and $\bar{P}_{n+1} = P_{n+1}$ obtained from the Kalman filter (2.15). Furthermore, we have

$$\bar{y}_{n+j} = \mathrm{E}(y_{n+j}|Y_n) = \mathrm{E}(\alpha_{n+j}|Y_n) + \mathrm{E}(\varepsilon_{n+j}|Y_n) = \bar{a}_{n+j},$$
$$\bar{F}_{n+j} = \mathrm{Var}(y_{n+j}|Y_n) = \mathrm{Var}(\alpha_{n+j}|Y_n) + \mathrm{Var}(\varepsilon_{n+j}|Y_n) = \bar{P}_{n+j} + \sigma_\varepsilon^2,$$

for $j = 1, \ldots, J$. The consequence is that the Kalman filter can be applied for $t = 1, \ldots, n + J$ where we treat the observations at times $n + 1, \ldots, n + J$ as missing. Thus we conclude that forecasts and their error variances are delivered by applying the Kalman filter in a routine way with $K_t = 0$ for $t = n+1, \ldots, n+J$. The same property holds for the general linear Gaussian state space model as we shall show in Section 4.11. For a Bayesian treatment a similar argument can be used to show that the posterior mean and variance of the forecast of y_{n+j} is obtained by treating $y_{n+1}, \ldots, y_{n+j}$ as missing values, for $j = 1, \ldots, J$.

2.8.1 Illustration

The Nile data set is now extended by 30 missing observations allowing the computation of forecasts for the observations $y_{101}, \ldots, y_{130}$. Only the Kalman filter

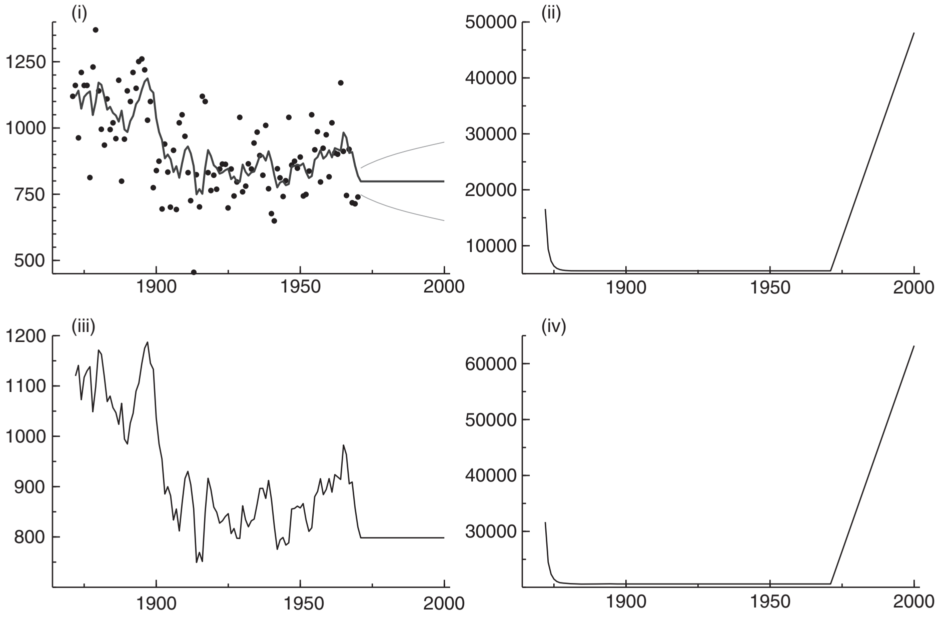

Fig. 2.6 Nile data and output of forecasting: (i) data (dots), state forecast a_t and 50% confidence intervals; (ii) state variance P_t; (iii) observation forecast $\mathrm{E}(y_t|Y_{t-1})$; (iv) observation forecast variance F_t.

is required. The graphs in Fig. 2.6 contain $\hat{y}_{n+j|n} = a_{n+j|n}$, $P_{n+j|n}$, $a_{n+j|n}$ and $F_{n+j|n}$, respectively, for $j = 1, \dots, J$ with $J = 30$. The confidence interval for $\mathrm{E}(y_{n+j}|Y_n)$ is $\hat{y}_{n+j|n} \pm k\sqrt{F}_{n+j|n}$ where k is determined by the required probability of inclusion; in Fig. 2.6 this probability is 50%.

2.9 Initialisation

We assumed in our treatment of the linear Gaussian model in previous sections that the distribution of the initial state α_1 is $\mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. We now consider how to start up the filter (2.15) when nothing is known about the distribution of α_1, which is the usual situation in practice. In this situation it is reasonable to represent α_1 as having a *diffuse prior* density, that is, fix a_1 at an arbitrary value and let $P_1 \to \infty$. From (2.15) we have

$$v_1 = y_1 - a_1, \qquad F_1 = P_1 + \sigma_\varepsilon^2,$$

and, by substituting into the equations for a_2 and P_2 in (2.15), it follows that

$$a_2 = a_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}(y_1 - a_1), \tag{2.56}$$

$$\begin{aligned} P_2 &= P_1\left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2}\right) + \sigma_\eta^2 \\ &= \frac{P_1}{P_1 + \sigma_\varepsilon^2}\sigma_\varepsilon^2 + \sigma_\eta^2. \end{aligned} \tag{2.57}$$

Letting $P_1 \to \infty$, we obtain $a_2 = y_1$, $P_2 = \sigma_\varepsilon^2 + \sigma_\eta^2$; we can then proceed normally with the Kalman filter (2.15) for $t = 2, \dots, n$. This process is called *diffuse initialisation* of the Kalman filter and the resulting filter is called *the diffuse Kalman filter*. We note the interesting fact that the same values of a_t and P_t for $t = 2, \dots, n$ can be obtained by treating y_1 as fixed and taking $\alpha_1 \sim \mathrm{N}(y_1, \sigma_\varepsilon^2)$. Specifically, we have $a_{1|1} = y_1$ and $P_{1|1} = \sigma_\varepsilon^2$. It follows from (2.18) for $t = 1$ that $a_2 = y_1$ and $P_2 = \sigma_\varepsilon^2 + \sigma_\eta^2$. This is intuitively reasonable in the absence of information about the marginal distribution of α_1 since $(y_1 - \alpha_1) \sim \mathrm{N}(0, \sigma_\varepsilon^2)$.

We also need to take account of the diffuse distribution of the initial state α_1 in the smoothing recursions. It is shown above that the filtering equations for $t = 2, \dots, n$ are not affected by letting $P_1 \to \infty$. Therefore, the state and disturbance smoothing equations are also not affected for $t = n, \dots, 2$ since these only depend on the Kalman filter output. From (2.37), the smoothed mean of the state α_1 is given by

$$\begin{aligned} \hat{\alpha}_1 &= a_1 + P_1\left[\frac{1}{P_1 + \sigma_\varepsilon^2}v_1 + \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2}\right)r_1\right] \\ &= a_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}v_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}\sigma_\varepsilon^2 r_1. \end{aligned}$$

Letting $P_1 \to \infty$, we obtain $\hat{\alpha}_1 = a_1 + v_1 + \sigma_\varepsilon^2 r_1$ and by substituting for v_1 we have

$$\hat{\alpha}_1 = y_1 + \sigma_\varepsilon^2 r_1.$$

The smoothed conditional variance of the state α_1 given Y_n is, from (2.43)

$$\begin{aligned} V_1 &= P_1 - P_1^2 \left[\frac{1}{P_1 + \sigma_\varepsilon^2} + \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 N_1 \right] \\ &= P_1 \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} \right) - \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 \sigma_\varepsilon^4 N_1 \\ &= \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right) \sigma_\varepsilon^2 - \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 \sigma_\varepsilon^4 N_1. \end{aligned}$$

Letting $P_1 \to \infty$, we obtain $V_1 = \sigma_\varepsilon^2 - \sigma_\varepsilon^4 N_1$.

The smoothed means of the disturbances for $t = 1$ are given by

$$\hat{\varepsilon}_1 = \sigma_\varepsilon^2 u_1, \quad \text{with} \quad u_1 = \frac{1}{P_1 + \sigma_\varepsilon^2} v_1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} r_1,$$

and $\hat{\eta}_1 = \sigma_\eta^2 r_1$. Letting $P_1 \to \infty$, we obtain $\hat{\varepsilon}_1 = -\sigma_\varepsilon^2 r_1$. Note that r_1 depends on the Kalman filter output for $t = 2, \ldots, n$. The smoothed variances of the disturbances for $t = 1$ depend on D_1 and N_1 of which only D_1 is affected by $P_1 \to \infty$; using (2.47),

$$D_1 = \frac{1}{P_1 + \sigma_\varepsilon^2} + \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 N_1.$$

Letting $P_1 \to \infty$, we obtain $D_1 = N_1$ and therefore $\mathrm{Var}(\hat{\varepsilon}_1) = \sigma_\varepsilon^4 N_1$. The variance of the smoothed estimate of η_1 remains unaltered as $\mathrm{Var}(\hat{\eta}_1) = \sigma_\eta^4 N_1$.

The initial smoothed state $\hat{\alpha}_1$ under diffuse conditions can also be obtained by assuming that y_1 is fixed and $\alpha_1 = y_1 - \varepsilon_1$ where $\varepsilon_1 \sim N(0, \sigma_\varepsilon^2)$. For example, for the smoothed mean of the state at $t = 1$, we have now only $n - 1$ varying y_t's so that

$$\hat{\alpha}_1 = a_1 + \sum_{j=2}^{n} \frac{\mathrm{Cov}(\alpha_1, v_j)}{F_j} v_j$$

with $a_1 = y_1$. It follows from (2.56) that $a_2 = a_1 = y_1$. Further, $v_2 = y_2 - a_2 = \alpha_2 + \varepsilon_2 - y_1 = \alpha_1 + \eta_1 + \varepsilon_2 - y_1 = -\varepsilon_1 + \eta_1 + \varepsilon_2$. Consequently, $\mathrm{Cov}(\alpha_1, v_2) = \mathrm{Cov}(-\varepsilon_1, -\varepsilon_1 + \eta_1 + \varepsilon_2) = \sigma_\varepsilon^2$. We therefore have from (2.32),

$$\begin{aligned} \hat{\alpha}_1 &= a_1 + \frac{\sigma_\varepsilon^2}{F_2} v_2 + \frac{(1 - K_2)\sigma_\varepsilon^2}{F_3} v_3 + \frac{(1 - K_2)(1 - K_3)\sigma_\varepsilon^2}{F_4} v_4 + \cdots \\ &= y_1 + \sigma_\varepsilon^2 r_1, \end{aligned}$$

as before with r_1 as defined in (2.34) for $t = 1$. The equations for the remaining $\hat{\alpha}_t$'s are the same as previously. The same results may be obtained by Bayesian arguments.

Use of a diffuse prior for initialisation is the approach preferred by most time series analysts in the situation where nothing is known about the initial value α_1. However, some workers find the diffuse approach uncongenial because they regard the assumption of an infinite variance as unnatural since all observed time series have finite values. From this point of view an alternative approach is to assume that α_1 is an unknown constant to be estimated from the data by maximum likelihood. The simplest form of this idea is to estimate α_1 by maximum likelihood from the first observation y_1. Denote this maximum likelihood estimate by $\hat{\alpha}_1$ and its variance by $\mathrm{Var}(\hat{\alpha}_1)$. We then initialise the Kalman filter by taking $a_{1|1} = \hat{\alpha}_1$ and $P_{1|1} = \mathrm{Var}(\hat{\alpha}_1)$. Since when α_1 is fixed $y_1 \sim \mathrm{N}(\alpha_1, \sigma_\varepsilon^2)$, we have $\hat{\alpha}_1 = y_1$ and $\mathrm{Var}(\hat{\alpha}_1) = \sigma_\varepsilon^2$. We therefore initialise the filter by taking $a_{1|1} = y_1$ and $P_{1|1} = \sigma_\varepsilon^2$. But these are the same values as we obtain by assuming that α_1 is diffuse. It follows that we obtain the same initialisation of the Kalman filter by representing α_1 as a random variable with infinite variance as by assuming that it is fixed and unknown and estimating it from y_1. We shall show that a similar result holds for the general linear Gaussian state space model in Subsection 5.7.3.

2.10 Parameter estimation

We now consider the fitting of the local level model to data from the standpoint of classical inference. In effect, this amounts to deriving formulae on the assumption that the additional parameters are known and then replacing these by their maximum likelihood estimates. Bayesian treatments will be considered for the general linear Gaussian model in Chapter 13. Parameters in state space models are often called *hyperparameters*, possibly to distinguish them from elements of state vectors which can plausibly be thought of as random parameters; however, in this book we shall just call them *additional parameters*, since with the usual meaning of the word parameter this is what they are. We will discuss methods for calculating the loglikelihood function and the maximisation of it with respect to the additional parameters, σ_ε^2 and σ_η^2.

2.10.1 Loglikelihood evaluation

Since

$$p(y_1, \ldots, y_t) = p(Y_{t-1})p(y_t|Y_{t-1}),$$

for $t = 2, \ldots, n$, the joint density of $y_1, \ldots, y_n$ can be expressed as

$$p(Y_n) = \prod_{t=1}^{n} p(y_t|Y_{t-1}),$$

where $p(y_1|Y_0) = p(y_1)$. Now $p(y_t|Y_{t-1}) = \mathrm{N}(a_t, F_t)$ and $v_t = y_t - a_t$ so on taking logs and assuming that a_1 and P_1 are known the loglikelihood is given by

$$\log L = \log p(Y_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right). \tag{2.58}$$

The exact loglikelihood can therefore be constructed easily from the Kalman filter (2.15).

Alternatively, let us derive the loglikelihood for the local level model from the representation (2.4). This gives

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}(Y_n - a_1\mathtt{1})'\Omega^{-1}(Y_n - a_1\mathtt{1}), \tag{2.59}$$

which follows from the multivariate normal distribution $Y_n \sim \mathrm{N}(a_1\mathtt{1}, \Omega)$. Using results from §2.3.1, $\Omega = CFC'$, $|C| = 1$, $\Omega^{-1} = C'F^{-1}C$ and $v = C(Y_n - a_1\mathtt{1})$; it follows that

$$\log|\Omega| = \log|CFC'| = \log|C||F||C| = \log|F|,$$

and

$$(Y_n - a_1\mathtt{1})'\Omega^{-1}(Y_n - a_1\mathtt{1}) = v'F^{-1}v.$$

Substitution and using the results $\log|F| = \sum_{t=1}^{n}\log F_t$ and $v'F^{-1}v = \sum_{t=1}^{n}F_t^{-1}v_t^2$ lead directly to (2.58).

The loglikelihood in the diffuse case is derived as follows. All terms in (2.58) remain finite as $P_1 \to \infty$ with Y_n fixed except the term for $t = 1$. It thus seems reasonable to remove the influence of P_1 as $P_1 \to \infty$ by defining the *diffuse loglikelihood* as

$$\begin{aligned}
\log L_d &= \lim_{P_1\to\infty}\left(\log L + \frac{1}{2}\log P_1\right)\\
&= -\frac{1}{2}\lim_{P_1\to\infty}\left(\log\frac{F_1}{P_1} + \frac{v_1^2}{F_1}\right) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right)\\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right), \qquad (2.60)
\end{aligned}$$

since $F_1/P_1 \to 1$ and $v_1^2/F_1 \to 0$ as $P_1 \to \infty$. Note that v_t and F_t remain finite as $P_1 \to \infty$ for $t = 2, \ldots, n$.

Since P_1 does not depend on σ_ε^2 and σ_η^2, the values of σ_ε^2 and σ_η^2 that maximise $\log L$ are identical to the values that maximise $\log L + \frac{1}{2}\log P_1$. As $P_1 \to \infty$, these latter values converge to the values that maximise $\log L_d$ because first and second derivatives with respect to σ_ε^2 and σ_η^2 converge, and second derivatives are finite and strictly negative. It follows that the maximum likelihood estimators of σ_ε^2 and σ_η^2 obtained by maximising (2.58) converge to the values obtained by maximising (2.60) as $P_1 \to \infty$.

We estimate the unknown parameters σ_ε^2 and σ_η^2 by maximising expression (2.58) or (2.60) numerically according to whether a_1 and P_1 are known or unknown. In practice it is more convenient to maximise numerically with respect to the quantities $\psi_\varepsilon = \log \sigma_\varepsilon^2$ and $\psi_\eta = \log \sigma_\eta^2$. An efficient algorithm for numerical maximisation is implemented in the *STAMP* 8.3 package of Koopman, Harvey, Doornik and Shephard (2010). This optimisation procedure is based on the quasi-Newton scheme BFGS for which details are given in Subsection 7.3.2.

2.10.2 Concentration of loglikelihood

It can be advantageous to re-parameterise the model prior to maximisation in order to reduce the dimensionality of the numerical search for the estimation of the parameters. For example, for the local level model we can put $q = \sigma_\eta^2/\sigma_\varepsilon^2$ to obtain the model

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \alpha_{t+1} &= \alpha_t + \eta_t, \qquad & \eta_t &\sim \mathrm{N}\big(0, q\sigma_\varepsilon^2\big), \end{aligned}$$

and estimate the pair σ_ε^2, q in preference to $\sigma_\varepsilon^2, \sigma_\eta^2$. Put $P_t^* = P_t/\sigma_\varepsilon^2$ and $F_t^* = F_t/\sigma_\varepsilon^2$; from (2.15) and Section 2.9, we have

$$\begin{aligned} v_t &= y_t - a_t, \qquad & F_t^* &= P_t^* + 1, \\ a_{t+1} &= a_t + K_t v_t, \qquad & P_{t+1}^* &= P_t^*(1 - K_t) + q, \end{aligned}$$

where $K_t = P_t/F_t = P_t^*/F_t^*$ for $t = 2, \ldots, n$ and these relations are initialised with $a_2 = y_1$ and $P_2^* = 1 + q$. Note that F_t^* depends on q but not on σ_ε^2. The loglikelihood (2.60) then becomes

$$\log L_d = -\frac{n}{2}\log(2\pi) - \frac{n-1}{2}\log \sigma_\varepsilon^2 - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t^* + \frac{v_t^2}{\sigma_\varepsilon^2 F_t^*}\right). \tag{2.61}$$

By maximising (2.61) with respect to σ_ε^2, for given $F_2^*, \ldots, F_n^*$, we obtain

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-1}\sum_{t=2}^{n}\frac{v_t^2}{F_t^*}. \tag{2.62}$$

The value of $\log L_d$ obtained by substituting $\hat{\sigma}_\varepsilon^2$ for σ_ε^2 in (2.61) is called the *concentrated diffuse loglikelihood* and is denoted by $\log L_{dc}$, giving

$$\log L_{dc} = -\frac{n}{2}\log(2\pi) - \frac{n-1}{2} - \frac{n-1}{2}\log \hat{\sigma}_\varepsilon^2 - \frac{1}{2}\sum_{t=2}^{n}\log F_t^*. \tag{2.63}$$

This is maximised with respect to q by a one-dimensional numerical search.

Table 2.1 Estimation of parameters of local level model by maximum likelihood.

Iteration	q	ψ	Score	Loglikelihood
0	1	0	-3.32	-495.68
1	0.0360	-3.32	0.93	-492.53
2	0.0745	-2.60	0.25	-492.10
3	0.0974	-2.32	-0.001	-492.07
4	0.0973	-2.33	0.0	-492.07

2.10.3 Illustration

The estimates of the variances σ_ε^2 and $\sigma_\eta^2 = q\sigma_\varepsilon^2$ for the Nile data are obtained by maximising the concentrated diffuse loglikelihood (2.63) with respect to ψ where $q = \exp(\psi)$. In Table 2.1 the iterations of the BFGS procedure are reported starting with $\psi = 0$. The relative percentage change of the loglikelihood goes down very rapidly and convergence is achieved after 4 iterations. The final estimate for ψ is -2.33 and hence the estimate of q is $\hat{q} = 0.097$. The estimate of σ_ε^2 given by (2.62) is 15099 which implies that the estimate of σ_η^2 is $\hat{\sigma}_\eta^2 = \hat{q}\hat{\sigma}_\varepsilon^2 = 0.097 \times 15099 = 1469.1$.

2.11 Steady state

We now consider whether the Kalman filter (2.15) converges to a *steady state* as $n \to \infty$. This will be the case if P_t converges to a positive value, $\bar{P}$ say. Obviously, we would then have $F_t \to \bar{P} + \sigma_\varepsilon^2$ and $K_t \to \bar{P}/(\bar{P} + \sigma_\varepsilon^2)$. To check whether there is a steady state, put $P_{t+1} = P_t = \bar{P}$ in (2.15) and verify whether the resulting equation in $\bar{P}$ has a positive solution. The equation is

$$\bar{P} = \bar{P}\left(1 - \frac{\bar{P}}{\bar{P} + \sigma_\varepsilon^2}\right) + \sigma_\eta^2,$$

which reduces to the quadratic

$$x^2 - xq - q = 0, \tag{2.64}$$

where $x = \bar{P}/\sigma_\varepsilon^2$ and $q = \sigma_\eta^2/\sigma_\varepsilon^2$, with the solution

$$x = \left(q + \sqrt{q^2 + 4q}\right)/2.$$

This is positive when $q > 0$ which holds for nontrivial models. The other solution to (2.64) is inapplicable since it is negative for $q > 0$. Thus all non-trivial local level models have a steady state solution.

The practical advantage of knowing that a model has a steady state solution is that, after convergence of P_t to $\bar{P}$ has been verified as close enough, we can stop computing F_t and K_t and the filter (2.15) reduces to the single relation

$$a_{t+1} = a_t + \bar{K} v_t,$$

with $\bar{K} = \bar{P}/(\bar{P} + \sigma_\varepsilon^2)$ and $v_t = y_t - a_t$. While this has little consequence for the simple local level model we are concerned with here, it is a useful property for the more complicated models we shall consider in Chapter 4, where P_t can be a large matrix.

2.12 Diagnostic checking

2.12.1 Diagnostic tests for forecast errors

The assumptions underlying the local level model are that the disturbances ε_t and η_t are normally distributed and serially independent with constant variances. On these assumptions the standardised one-step ahead forecast errors

$$e_t = \frac{v_t}{\sqrt{F_t}}, \qquad t = 1, \ldots, n, \tag{2.65}$$

(or for $t = 2, \ldots, n$ in the diffuse case) are also normally distributed and serially independent with unit variance. We can check that these properties hold by means of the following large-sample diagnostic tests:

- Normality

  The first four moments of the standardised forecast errors are given by

$$m_1 = \frac{1}{n} \sum_{t=1}^{n} e_t,$$

$$m_q = \frac{1}{n} \sum_{t=1}^{n} (e_t - m_1)^q, \qquad q = 2, 3, 4,$$

  with obvious modifications in the diffuse case. Skewness and kurtosis are denoted by S and K, respectively, and are defined as

$$S = \frac{m_3}{\sqrt{m_2^3}}, \qquad K = \frac{m_4}{m_2^2},$$

  and it can be shown that when the model assumptions are valid they are asymptotically normally distributed as

$$S \sim \mathrm{N}\left(0, \frac{6}{n}\right), \qquad K \sim \mathrm{N}\left(3, \frac{24}{n}\right);$$

see Bowman and Shenton (1975). Standard statistical tests can be used to check whether the observed values of S and K are consistent with their asymptotic densities. They can also be combined as

$$N = n\left\{\frac{S^2}{6} + \frac{(K-3)^2}{24}\right\},$$

which asymptotically has a χ^2 distribution with 2 degrees of freedom on the null hypothesis that the normality assumption is valid. The *QQ plot* is a graphical display of ordered residuals against their theoretical quantiles. The 45 degree line is taken as a reference line (the closer the residual plot to this line, the better the match).

- Heteroscedasticity
  A simple test for heteroscedasticity is obtained by comparing the sum of squares of two exclusive subsets of the sample. For example, the statistic

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=1}^{h} e_t^2},$$

  is $F_{h,h}$-distributed for some preset positive integer h, under the null hypothesis of homoscedasticity. Here, e_t is defined in (2.65) and the sum of h squared forecast errors in the denominator starts at $t = 2$ in the diffuse case.
- Serial correlation
  When the local level model holds, the standardised forecast errors are serially uncorrelated as we have shown in Subsection 2.3.1. Therefore, the correlogram of the forecast errors should reveal serial correlation insignificant. A standard portmanteau test statistic for serial correlation is based on the Box–Ljung statistic suggested by Ljung and Box (1978). This is given by

$$Q(k) = n(n+2)\sum_{j=1}^{k}\frac{c_j^2}{n-j},$$

  for some preset positive integer k where c_j is the jth correlogram value

$$c_j = \frac{1}{nm_2}\sum_{t=j+1}^{n}(e_t - m_1)(e_{t-j} - m_1).$$

More details on diagnostic checking will be given in Section 7.5.

2.12.2 Detection of outliers and structural breaks

The standardised smoothed residuals are given by

$$\begin{aligned} u_t^* &= \hat{\varepsilon}_t/\sqrt{\mathrm{Var}(\hat{\varepsilon}_t)} = D_t^{-\frac{1}{2}}u_t, \\ r_t^* &= \hat{\eta}_t/\sqrt{\mathrm{Var}(\hat{\eta}_t)} = N_t^{-\frac{1}{2}}r_t, \qquad t = 1,\ldots,n; \end{aligned}$$

see Section 2.5 for details on computing the quantities u_t, D_t, r_t and N_t. Harvey and Koopman (1992) refer to these standardised residuals as *auxiliary residuals* and they investigate their properties in detail. For example, they show that the auxiliary residuals are autocorrelated and they discuss their autocorrelation function. The auxiliary residuals can be useful in detecting outliers and structural breaks in time series because $\hat{\varepsilon}_t$ and $\hat{\eta}_t$ are estimators of ε_t and η_t. An outlier in a series that we postulate as generated by the local level model is indicated by a large (positive or negative) value for $\hat{\varepsilon}_t$, or u_t^*, and a break in the level α_{t+1} is indicated by a large (positive or negative) value for $\hat{\eta}_t$, or r_t^*. A discussion of the use of auxiliary residuals for the general model will be given in Section 7.5.

2.12.3 Illustration

We consider the fitted local level model for the Nile data as obtained in Subsection 2.10.3. A plot of e_t is given in Fig. 2.7 together with the histogram, the QQ plot and the correlogram. These plots are satisfactory and they suggest that the assumptions underlying the local level model are valid for the Nile data. This is largely confirmed by the following diagnostic test statistics

$$S = -0.03, \quad K = 0.09, \quad N = 0.05, \quad H(33) = 0.61, \quad Q(9) = 8.84.$$

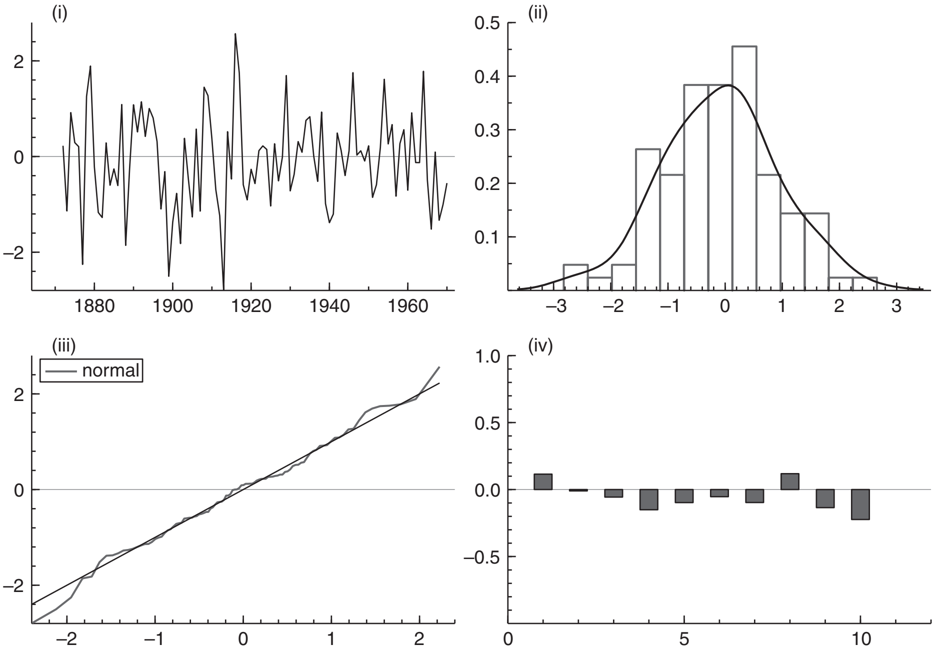

Fig. 2.7 Diagnostic plots for standardised prediction errors: (i) standardised residual; (ii) histogram plus estimated density; (iii) ordered residuals; (iv) correlogram.

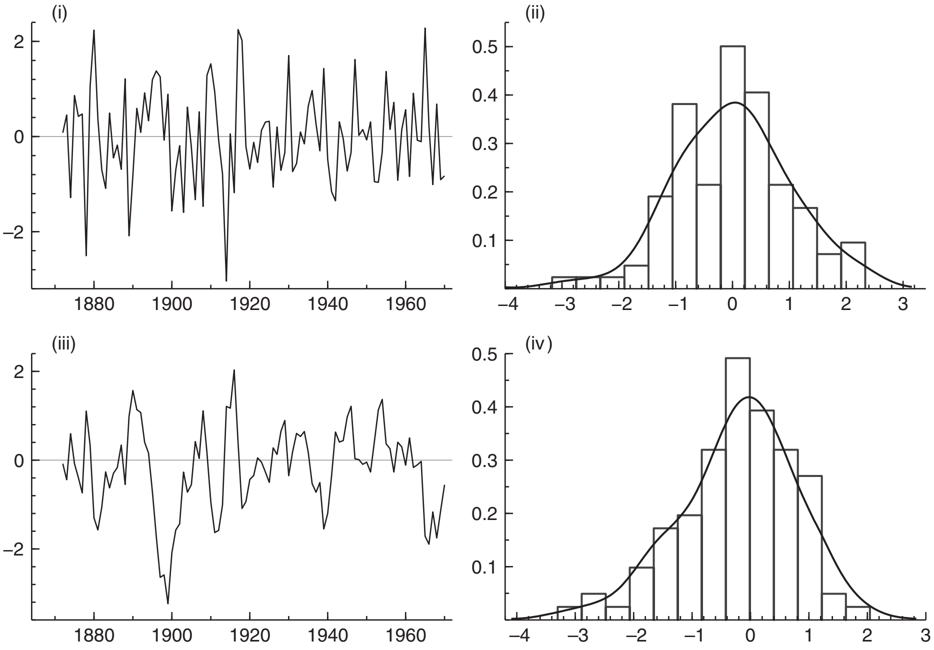

Fig. 2.8 Diagnostic plots for auxiliary residuals: (i) observation residual u_t^*; (ii) histogram and estimated density for u_t^*; (iii) state residual r_t^*; (iv) histogram and estimated density for r_t^*.

The low value for the heteroscedasticity statistic H indicates a degree of heteroscedasticity in the residuals. This is apparent in the plots of u_t^* and r_t^* together with their histograms in Fig. 2.8. These diagnostic plots indicate outliers in 1913 and 1918 and a level break in 1899. The plot of the Nile data confirms these findings.

2.13 Exercises

2.13.1

Consider the local level model (2.3).

(a) Give a model representation for $x_t = y_t - y_{t-1}$, for $t = 2, \ldots, n$.

(b) Show that the model for x_t in (a) can have the same statistical properties as the model given by $x_t = \xi_t + \theta\,\xi_{t-1}$ where $\xi_t \sim \mathrm{N}(0, \sigma_\xi^2)$ are independent disturbances with variance $\sigma_\xi^2 > 0$ and for some value θ.

(c) For what value of θ, in terms of σ_ε^2 and σ_η^2, are the model representations for x_t in (a) and (b) equivalent? Comment.

2.13.2

(a) Using the derivations as in Subsection 2.4.2, develop backwards recursions for the evaluation of $\text{Cov}(\alpha_{t+1}, \alpha_t | Y_n)$ for $t = n, \ldots, 1$.

(b) Using the derivations as in Subsection 2.5.1 and 2.5.2, develop backwards recursions for the evaluation of $\text{Cov}(\varepsilon_t, \eta_t | Y_n)$ for $t = n, \ldots, 1$.

2.13.3

Consider the loglikelihood expression (2.59) and show that the maximum likelihood estimator of a_1 is given by

$$\hat{a}_1 = \frac{1}{n} \sum_{t=1}^{n} u_t^o,$$

where u_t^o is defined as in (2.45) but obtained from the Kalman filter and smoothing recursions with initialisation $a_1 = 0$. Note that we treat the initial state variance P_1 here as a known and finite value.