14 Non-Gaussian and nonlinear illustrations

14.1 Introduction

In this chapter we illustrate the methodology of Part II by applying it to different real data sets. In the first example of Section 14.2 we consider the estimation of a multiplicative trend and seasonal model. In Section 14.3 we examine the effects of seat belt legislation on deaths of van drivers due to road accidents in Great Britain modelled by a Poisson distribution. In the third example of Section 14.4 we consider the usefulness of the t-distribution for modelling observation errors in a gas consumption series containing outliers. In Section 14.5 we fit a stochastic volatility model to a series of pound/dollar exchange rates using different methodologies. In the final illustration of Section 14.6 we fit a binary model to the results of the Oxford Cambridge boat race over a long period with many missing observations and we forecast the probability that Oxford will win in 2012.

14.2 Nonlinear decomposition: UK visits abroad

It is common practice in economic time series analyses and seasonal adjustment procedures to take logarithms of the data and to adopt a linear Gaussian time series model for its analysis. The logarithmic transformation converts an exponentially growing trend into a linear trend while it also eliminates or reduces growing seasonal variation and heteroscedasticity in seasonal time series. The logadditive framework appears to work successfully for a model-based decomposition of the time series in trend, seasonal, irregular and other components. It predicates that time series components combine multiplicatively in the implied model for the untransformed series. A full multiplicative model is however not always intended or desired. If heteroscedasticity or changing seasonal variation remains after the logtransformation, applying the logtransformation again is not an attractive solution. Also, if the data is already supplied in units measuring proportional changes, applying the logtransformation can complicate model interpretation.

Koopman and Lee (2009) propose a nonlinear unobserved component time series model that can be used when the time series, possibly after a logtransformation, is not appropriate for an analysis based on the linear model. They generalise the basic structural model of Subsection 3.2.3 by scaling the amplitude

of the seasonal component γ_t via an exponential transformation of the trend component μ_t. The observed time series y_t, either in levels or in logs, is decomposed by the nonlinear model

$$y_t = \mu_t + \exp(c_0 + c_\mu \mu_t)\gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2), \qquad t = 1, \ldots, n, \quad (14.1)$$

where c_0 and c_μ are an unknown fixed coefficients while dynamic specifications of the trend component μ_t are discussed in Subsection 3.2.1 and those of the seasonal component γ_t in Subsection 3.2.2. Coefficient c_0 scales the seasonal effect and therefore we restrict $\sigma_\omega^2 = 1$ in the seasonal models of Subsection 3.2.2. The sign of the coefficient c_μ determines whether the seasonal variation increases or decreases when a positive change in the trend occurs. The model reduces to the linear specification of Subsection 3.2.3 when c_μ is zero. The overall time-varying amplitude of the seasonal component is determined by the combined effect $c_0 + c_\mu \mu_t$.

We consider a data set of monthly visits abroad by UK residents from January 1980 to December 2006. The data is compiled by the Office for National Statistics (ONS), based on the International Passenger Survey. Fig. 14.1 presents the time series in levels and in logs. The time series of visits abroad shows a clear upwards trend, a pronounced seasonal pattern, and a steady increase of the seasonal variation over time. After applying the logtransformation, the increase of

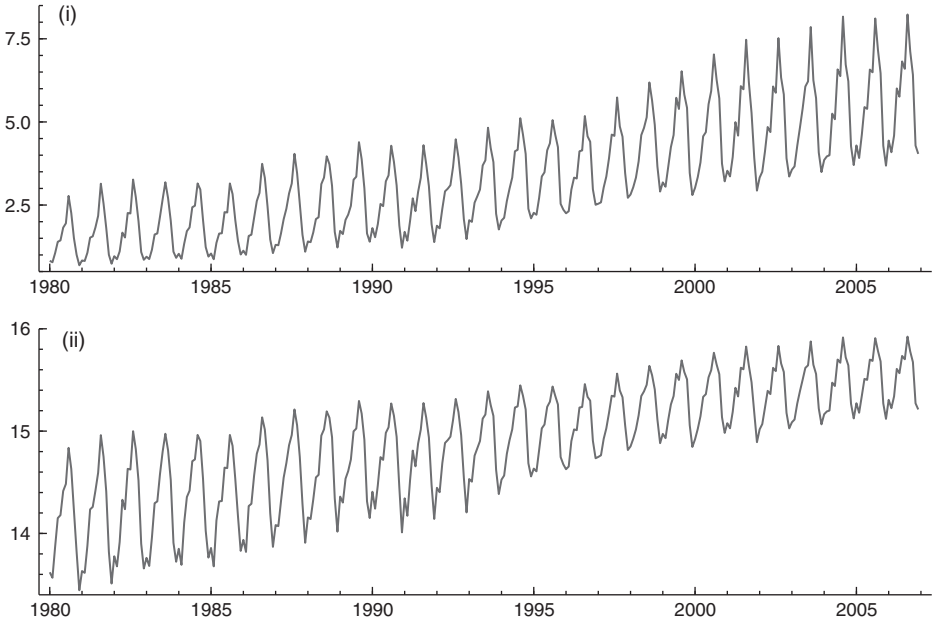

Fig. 14.1 Visits of UK residents abroad (i) in levels (million); (ii) in logarithms.

seasonal variation has been converted into a decrease. This may indicate that the logtransformation is not particularly appropriate for this series.

We therefore consider the model given by (14.1) with a cycle component ψ_t discussed in Subsection 3.2.4 added to capture economic business cycle behaviour from the data, that is $y_t = \mu_t + \psi_t + \exp(c_0 + c_\mu \mu_t)\gamma_t + \varepsilon_t$ with irregular $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. The trend μ_t is specified as the local linear trend (3.2) with $\sigma_\xi^2 = 0$ (smooth trend), the seasonal γ_t has the trigonometric specification (3.5) and the cycle ψ_t is specified as c_t in (3.13). Due to the seasonal effect $\exp(c_0 + c_\mu \mu_t)\gamma_t$, we obtain a nonlinear observation equation $y_t = Z(\alpha_t) + \varepsilon_t$ where the state vector α_t consists of elements associated with the trend, seasonal and cycle components; see Subsections 3.2.3 and 3.2.4. The model is a special case of the nonlinear state space model discussed in Section 9.7.

We apply the extended Kalman filter as developed in Section 10.2 to our model. The Gaussian likelihood function (7.2), with v_t and F_t computed by the extended Kalman filter (10.4) is treated as an approximated likelihood function. The initialisation of the extended Kalman filter is discussed in Koopman and Lee (2009). The numerical optimisation of the approximate loglikelihood function produces the parameter estimates given by

$$\begin{aligned} \hat{\sigma}_\varepsilon = 0.116, \quad \hat{\sigma}_\zeta = 0.00090, \quad \hat{c}_0 = -5.098, \quad \hat{c}_\mu = 0.0984, \\ \hat{\sigma}_\kappa = 0.00088, \quad \hat{\rho} = 0.921, \quad 2\pi/\hat{\lambda}^c = 589. \end{aligned} \tag{14.2}$$

The seasonal component γ_t is scaled by $\exp(c_0 + c_\mu \mu_t)$. In Fig. 14.2, panel (i) presents the scaled seasonal component $\exp(c_0 + c_\mu \mu_t)\gamma_t$ and the unscaled component γ_t as estimated by the extended Kalman smoother discussed in Subsection 10.4.1. The scaled component is changing largely due to the trend component which is plotted in panel (ii) of Fig. 14.2. The unscaled component shows a more stable pattern with almost a constant amplitude over time. A more detailed discussion of this analysis is given by Koopman and Lee (2009).

14.3 Poisson density: van drivers killed in Great Britain

The assessment for the Department of Transport of the effects of seat belt legislation on road traffic accidents in Great Britain, described by Harvey and Durbin (1986) and also discussed in Section 8.2, was based on linear Gaussian methods as described in Part I. One series that was excluded from this study was the monthly numbers of light goods vehicle (van) drivers killed in road accidents from 1969 to 1984. The numbers of deaths of van drivers were too small to justify the use of the linear Gaussian model. A better model for the data is based on the Poisson distribution with mean $\exp(\theta_t)$ and density

$$p(y_t|\theta_t) = \exp\{\theta_t' y_t - \exp(\theta_t) - \log y_t!\}, \qquad t = 1, \ldots, n, \tag{14.3}$$

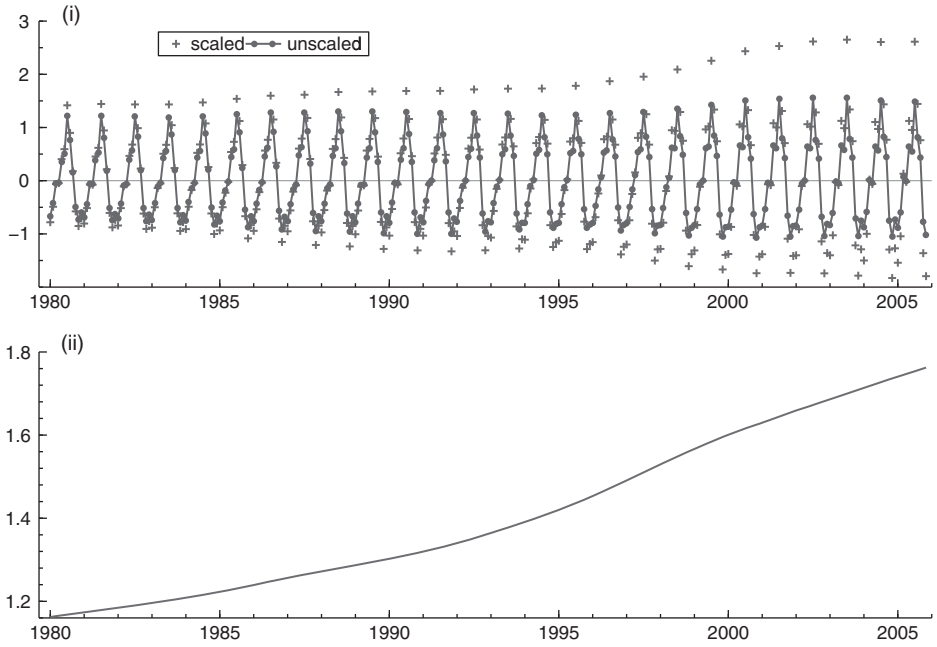

Fig. 14.2 Visits of UK residents abroad: (i) smooth estimates of the scaled and unscaled seasonal components obtained by the extended Kalman filter and smoother; (ii) scaling process $\exp(c_0 + c_\mu \mu_t)$ with μ_t replaced by its smoothed estimate.

as discussed in Subsection 9.3.1. We model θ_t by the relation

$$\theta_t = \mu_t + \gamma_t + \lambda x_t,$$

where the trend μ_t is the random walk

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big), \tag{14.4}$$

λ is the intervention parameter which measures the effects of the seat belt law, x_t is an indicator variable for the post legislation period and the monthly seasonal γ_t is generated by

$$\sum_{j=0}^{11} \gamma_{t+1-j} = \omega_t, \qquad \omega_t \sim \mathrm{N}\big(0, \sigma_\omega^2\big). \tag{14.5}$$

The disturbances η_t and ω_t are mutually independent Gaussian white noise terms with variances $\sigma_\eta^2 = \exp(\psi_\eta)$ and $\sigma_\omega^2 = \exp(\psi_\omega)$, respectively. The parameter

estimates are reported by Durbin and Koopman (1997) as $\hat{\sigma}_\eta = \exp(\hat{\psi}_\eta) = \exp(-3.708) = 0.0245$ and $\hat{\sigma}_\omega = 0$. The fact that $\hat{\sigma}_\omega = 0$ implies that the seasonal is constant over time.

For the Poisson model we have $b_t(\theta_t) = \exp(\theta_t)$. As in Subsection 10.6.4 we have $\dot{b}_t = \ddot{b}_t = \exp(\tilde{\theta}_t)$, so we take

$$A_t = \exp(-\tilde{\theta}_t), \qquad x_t = \tilde{\theta}_t + A_t y_t - 1,$$

where $\tilde{\theta}_t$ is some trial value for θ_t with $t = 1, \ldots, n$. The iterative process for determining the approximating model as described in Section 10.6 converges quickly; usually, between three and five iterations are needed for the Poisson model. The estimated signal $\mu_t + \lambda x_t$ for ψ_η fixed at $\hat{\psi}_\eta$ is computed and its exponentiated values are plotted together with the raw data in panel (i) of Fig. 14.3.

The main objective of the analysis is the estimation of the effect of the seat belt law on the number of deaths. Here, this is measured by λ which is estimated as the value -0.278 with standard error 0.114. The estimate of λ corresponds to a reduction in the number of deaths of 24%. It is clear from the standard

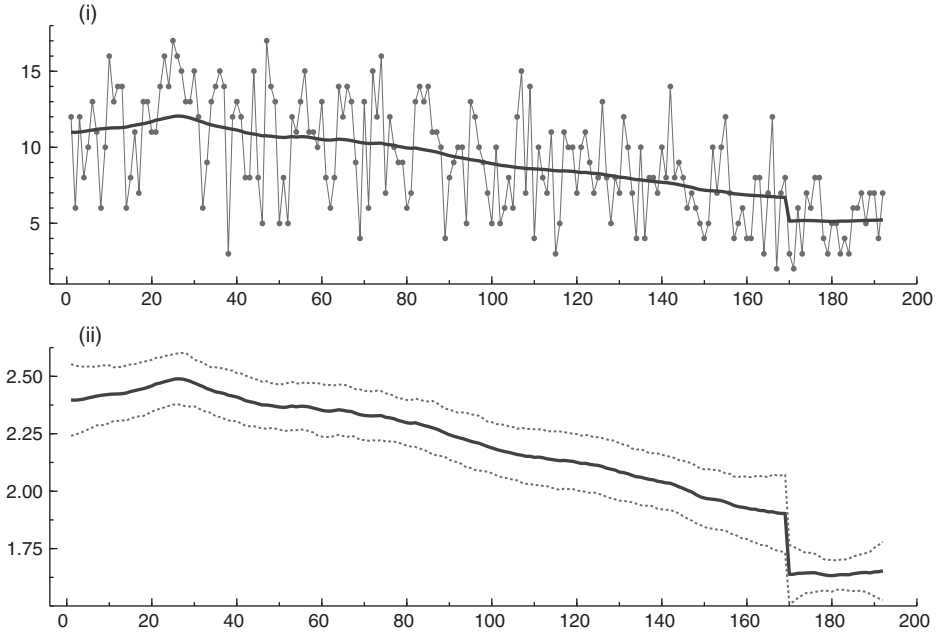

Fig. 14.3 Numbers of van drivers killed: (i) observed time series counts and estimated level including intervention, $\exp(\mu_t + \lambda x_t)$; (ii) estimated level including intervention, $\mu_t + \lambda x_t$, and its confidence interval based on two times standard error.

error that the seat belt law has given some significant reduction to the number of deaths; this is confirmed visually in panel (ii) of Fig. 14.3. What we learn from this exercise so far as the underlying real investigation is concerned is that up to the point where the law was introduced there was a slow regular decline in the number of deaths coupled with a constant multiplicative seasonal pattern, while at that point there was an abrupt drop in the trend of around 25%; afterwards, the trend appeared to flatten out, with the seasonal pattern remaining the same. A more detailed analysis on the basis of this model, and including a Bayesian analysis, is presented by Durbin and Koopman (2000).

14.4 Heavy-tailed density: outlier in gas consumption

In this example we analyse the logged quarterly demand for gas in the UK from 1960 to 1986 which is a series from the standard data set provided by Koopman, Harvey, Doornik and Shephard (2010). We use a structural time series model of the basic form as discussed in Section 3.2:

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \tag{14.6}$$

where μ_t is the local linear trend, γ_t is the seasonal and ε_t is the observation disturbance. The purpose of the investigation underlying the analysis is to study the seasonal pattern in the data with a view to seasonally adjusting the series. It is known that for most of the series the seasonal component changes smoothly over time, but it is also known that there was a disruption in the gas supply in the third and fourth quarters of 1970 which leads to a distortion in the seasonal pattern when a standard analysis based on a Gaussian density for ε_t is employed. The question under investigation is whether the use of a heavy-tailed density for ε_t would improve the estimation of the seasonal in 1970.

To model ε_t we use the t-distribution as in Subsection 9.4.1 with logdensity

$$\log p(\varepsilon_t) = \log a(\nu) + \frac{1}{2}\log\lambda - \frac{\nu+1}{2}\log\left(1+\lambda\varepsilon_t^2\right), \tag{14.7}$$

where

$$a(\nu) = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \qquad \lambda^{-1} = (\nu-2)\,\sigma_\varepsilon^2, \qquad \nu > 2, \qquad t = 1,\ldots,n.$$

The mean of ε_t is zero and the variance is σ_ε^2 for any ν degrees of freedom which need not be an integer. The approximating model can be obtained by the methods described in Section 10.8. When we use the first derivative only, we obtain

$$A_t = \frac{1}{\nu+1}\tilde{\varepsilon}_t^2 + \frac{\nu-2}{\nu+1}\sigma_\varepsilon^2,$$

The iterative scheme is started with $A_t = \sigma_\varepsilon^2$, for $t = 1, \ldots, n$. The number of iterations required for a reasonable level of convergence using the t-distribution is usually higher than for densities from the exponential family; for this example we required around ten iterations. In the classical analysis, the parameters of the model, including the degrees of freedom ν, were estimated by Monte Carlo maximum likelihood as described in Subsection 11.6.2; the estimated value for ν was 12.8.

We now compare the estimated seasonal and irregular components based on the Gaussian model and the model with a t-distribution for ε_t. Fig. 14.4 provide the graphs of the estimated seasonal and irregular for the Gaussian model and the t-model. The most striking feature of these graphs is the greater effectiveness with which the t-model picks and corrects for the outlier relative to the Gaussian model. The t-model estimates are based on 250 simulation samples from the simulation smoother with four antithetics for each sample. We learn from the analysis that the change over time of the seasonal pattern in the data is in fact smooth. We also learn that if model (14.6) is to be used to estimate the seasonal for this or similar cases with outliers in the observations, then a Gaussian model for ε_t is inappropriate and a heavy-tailed model should be used.

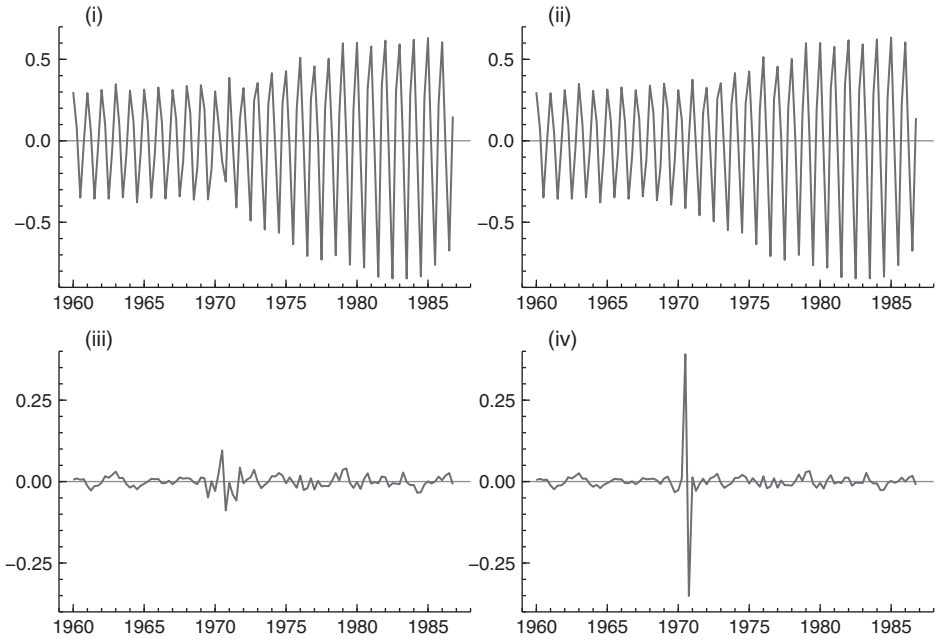

Fig. 14.4 Analyses of gas data: (i) estimated seasonal component from Gaussian model; (ii) estimated seasonal component from t model; (iii) estimated irregular from Gaussian model; (iv) estimated irregular from t model.

14.5 Volatility: pound/dollar daily exchange rates

The stochastic volatility (SV) model is discussed in detail in Section 9.5. In our empirical illustration for the pound/dollar daily exchange rates we consider a basic version of the SV model. The time series of exchange rates is from 1/10/81 to 28/6/85 and have been used by Harvey, Ruiz and Shephard (1994). Denoting the daily exchange rate by x_t, the observations we consider are given by $y_t = \log x_t, x_{t-1}$ for $t = 1, \ldots, n$. A zero-mean stochastic volatility model of the form

$$\begin{aligned} y_t &= \sigma \exp\left(\frac{1}{2}\theta_t\right) u_t, \qquad u_t \sim \mathrm{N}(0,1), \qquad t = 1, \ldots, n, \\ \theta_{t+1} &= \phi\theta_t + \eta_t, \qquad\qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big), \qquad 0 < \phi < 1, \end{aligned} \tag{14.8}$$

was used for analysing these data by Harvey, Ruiz and Shephard (1994). The purpose of the investigations for which this type of analysis is carried out is to study the structure of the volatility of price ratios in the market, which is of considerable interest to financial analysts. The level of θ_t determines the amount of volatility and the value of ϕ measures the autocorrelation present in the volatility process.

14.5.1 Data transformation analysis

We start with providing an approximate solution based on the linear model as suggested in Section 10.5. After the observations y_t are transformed into $\log y_t^2$, we consider the linear model (10.32), that is

$$\log y_t^2 = \kappa + \theta_t + \xi_t, \qquad t = 1, \ldots, n, \tag{14.9}$$

where κ is an unknown constant and ξ_t is a mean-zero disturbance which is not normally distributed. Given the model is linear, we can proceed approximately with the methods developed in Part I. This approach is taken by Harvey, Ruiz and Shephard (1994) who refer to it as a quasi-maximum likelihood (QML) method. Parameter estimation is done via the Kalman filter; smoothed estimates of the volatility component, θ_t, are constructed and forecasts of volatility can be generated. One of the attractions of the QML approach is that it can be carried out straightforwardly using standard software such as *STAMP* of Koopman, Harvey, Doornik and Shephard (2010). This is an advantage compared to the more involved simulation-based methods.

In our illustration of the QML method, we use the same data as analysed by Harvey, Ruiz and Shephard (1994) in which y_t is the first difference of the logged exchange rate between pound sterling and US dollar. To avoid taking logs of zero values, it is common practice to take deviations from its sample mean. The resulting mean-corrected logreturns are then analysed by the QML method.

Parameter estimation is carried out carried out quickly by the Kalman filter. We obtain the following QML estimates:

$$\begin{array}{rclrclrcl} \hat{\sigma}_\xi &=& 2.1521, & \hat{\psi}_1 &=& \log \hat{\sigma}_\xi = 0.7665, & \mathrm{SE}(\hat{\psi}_1) &=& 0.0236, \\ \hat{\sigma}_\eta &=& 0.8035, & \hat{\psi}_2 &=& \log \hat{\sigma}_\eta = -0.2188, & \mathrm{SE}(\hat{\psi}_2) &=& 0.5702, \\ \hat{\phi} &=& 0.9950, & \hat{\psi}_3 &=& \log \frac{\hat{\phi}}{1-\hat{\phi}} = 5.3005, & \mathrm{SE}(\hat{\psi}_3) &=& 1.6245, \end{array}$$

where SE denotes the standard error of the maximum likelihood estimator. We present the results in this form since we estimate the logtransformed parameters, so the standard errors that we calculate apply to them and not to the original parameters of interest.

Once the parameters are estimated, we can compute the smoothed estimate of the signal θ_t using standard Kalman filter and smoother methods. The logreturns for the pound/US dollar exhange series (adjusted for the mean, that is y_t) is depicted in panel (i) of Fig. 14.5. The signal extraction results are presented in panels (ii) and (iii) of Fig. 14.5. In panel (ii) we present the transformed data $\log y_t^2$ together with the smoothed estimate of θ_t from the Kalman filter

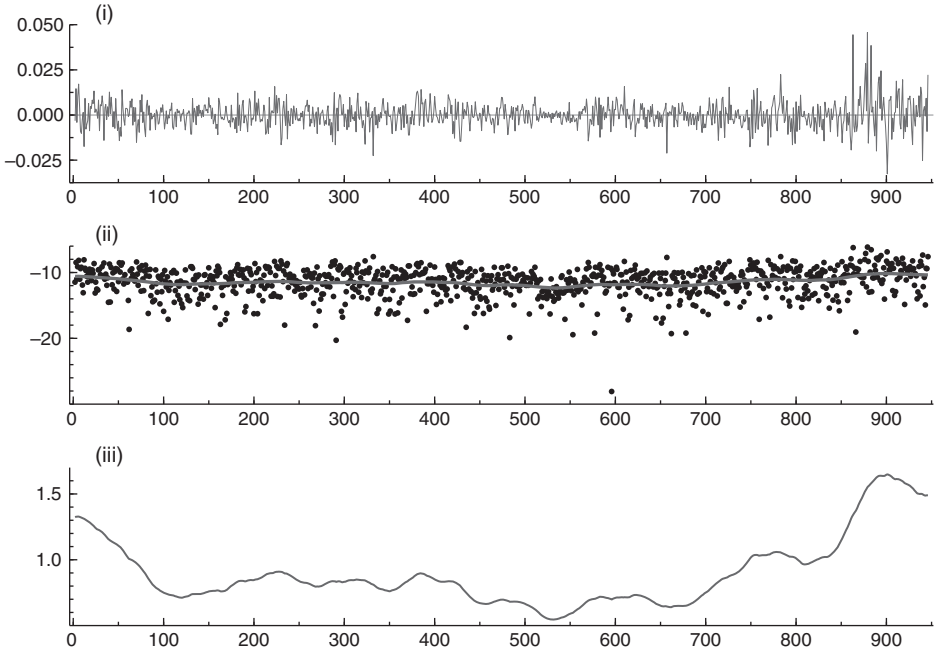

Fig. 14.5 Analyses of pound–dollar exchange rates: (i) daily logreturns of exchange rates, mean-corrected, denoted as y_t; (ii) the $\log y_t^2$ series with the smoothed estimate of $\kappa + \theta_t$;(ii) smoothed estimate of volatility measure $\exp(\theta_t / 2)$.

and smoother using the linear model (10.32). The smoothed estimate of the volatility, which we measure as $\exp(\theta_t/2)$, is displayed in panel (iii).

14.5.2 Estimation via importance sampling

To illustrate the maximum likelihood estimation of parameters in the SV model using importance sampling, we consider the Gaussian logdensity of model (14.8) which is given by

$$\log p(y_t|\theta_t) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\theta_t - \frac{y_t^2}{2\sigma^2}\exp(-\theta_t). \tag{14.10}$$

The linear approximating model based on the estimated mode of θ_t can be obtained by the method of Subsection 10.6.5 with

$$A_t = 2\sigma^2\frac{\exp(\tilde{\theta}_t)}{y_t^2}, \qquad x_t = \tilde{\theta}_t - \frac{1}{2}\tilde{H}_t + 1,$$

for which A_t is always positive. The iterative process can be started with $A_t = 2$ and $x_t = \log(y_t^2/\sigma^2)$, for $t = 1, \ldots, n$, since it follows from (14.8) that $y_t^2/\sigma^2 \approx \exp(\theta_t)$. When y_t is zero or very close to zero, it should be replaced by a small constant value to avoid numerical problems; this device is only needed to obtain the approximating model so we do not depart from our exact treatment. The number of iterations required is usually fewer than ten.

Firstly we focus on the estimation of the parameters. For the method of importance sampling, we take $N = 100$ for computing the loglikelihood function. We carry out the computations as detailed in Section 11.6. During the estimation process, we take the same random numbers for each loglikelihood evaluation so that the loglikelihood is a smooth function of the parameters. We then obtain the following estimates, after convergence of the numerical optimisation:

$$\begin{aligned}
&\hat{\sigma} = 0.6338, && \hat{\psi}_1 = \log\hat{\sigma} = -0.4561, && \mathrm{SE}(\hat{\psi}_1) = 0.1033,\\
&\hat{\sigma}_\eta = 0.1726, && \hat{\psi}_2 = \log\hat{\sigma}_\eta = -1.7569, && \mathrm{SE}(\hat{\psi}_2) = 0.2170,\\
&\hat{\phi} = 0.9731, && \hat{\psi}_3 = \log\frac{\hat{\phi}}{1-\hat{\phi}} = 3.5876, && \mathrm{SE}(\hat{\psi}_3) = 0.5007,
\end{aligned}$$

where SE denotes the standard error of the maximum likelihood estimator which applies to the logtransformed parameters and not to the original parameters of interest.

Secondly our aim is to estimate the underlying volatility θ_t via importance sampling. In panel (i) of Fig. 14.6 we present the data as absolute values of the first differences together with the smoothed estimate of the volatility component θ_t. We observe that the estimates capture the volatility features in the time

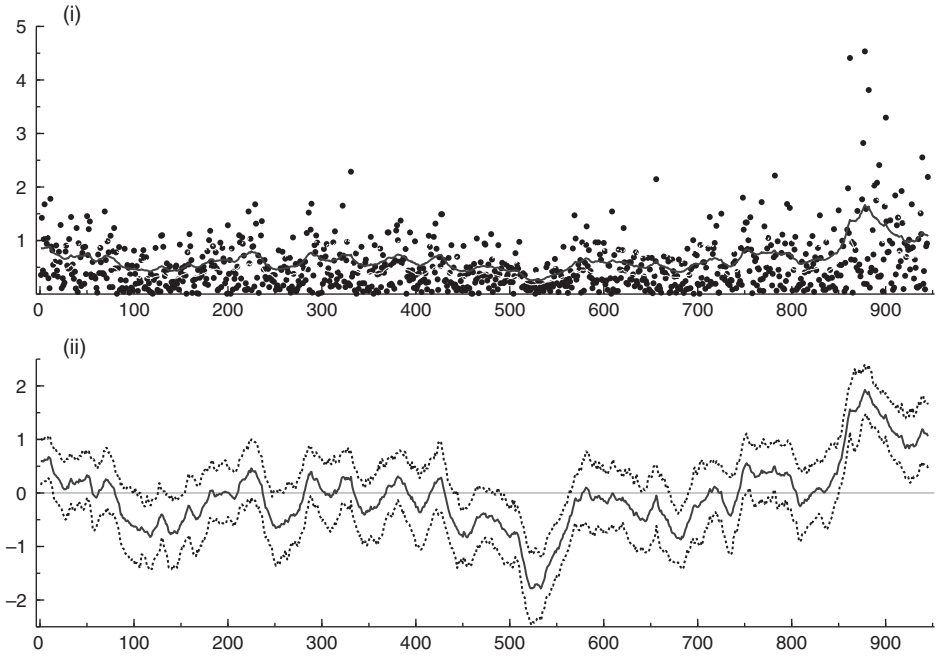

Fig. 14.6 Analyses of pound–dollar exchange rates: (i) difference in absolute values (dots) and smoothed estimate of θ_t; (ii) smoothed estimate of θ_t with 90% confidence interval.

series accurately. Panel (ii) presents the same volatility estimates but here with their 90% confidence interval which are based on standard errors and are also computed by the importance sampling method of Section 11.4.

14.5.3 Particle filtering illustration

To estimate the volatility by filtering for the pound exchange series, we consider both the bootstrap filter and the auxiliary filter and we assess their accuracy in relation to each other. The bootstrap filter is discussed in Subsection 12.4.2 and consists of the following steps for the SV model at a fixed time t and for a given set of particles $\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}$:

(i) Draw N values $\tilde{\theta}_t^{(i)} \sim \mathrm{N}(\phi\theta_{t-1}^{(i)}, \sigma_\eta^2)$.

(ii) Compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\theta_t^{(i)} - \frac{1}{2\sigma^2}\exp(-\theta_t^{(i)})y_t^2\right), \qquad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$.

(iii) Compute

$$\hat{a}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\theta}_t^{(i)}, \qquad \hat{P}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\theta}_t^{(i)\,2} - \hat{a}_{t|t}^2.$$

(iv) Select N new independent particles $\alpha_t^{(i)}$ via stratified sampling; see Subsection 12.3.3.

The implementation of the bootstrap filter is straightforward. To monitor its performance, we compute the efficient sample size (ESS) variable for each t. In panel (i) of Fig. 14.7 we present the ESS variable for each t. In many occasions the number of active particles is sufficiently high. However, at various points in time, the bootstrap filter detoriates and the number of effective particles is below 7500. The filtered estimate $\hat{a}_{t|t}$ of logvolatility θ_t is displayed in panel (ii) together with its 90% confidence interval based on $\hat{P}_{t|t}$. Although the volatility changes over time have the same pattern as the smoothed estimate of logvolatility displayed in panel (ii) of Fig. 14.6, the filtered estimate exhibits a more noisier estimate of logvolatility.

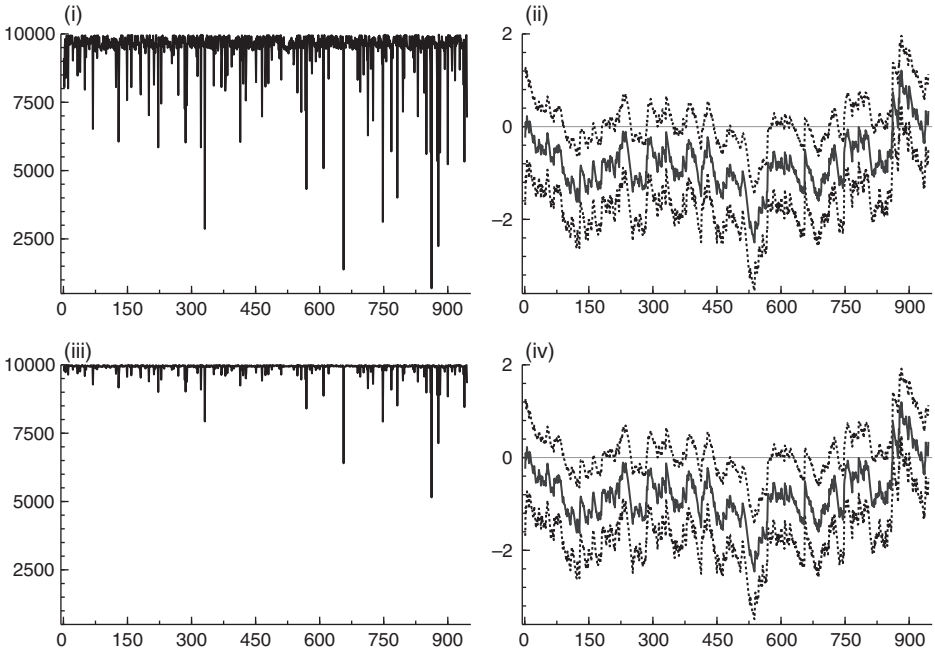

Fig. 14.7 Analyses of pound–dollar exchange rates using bootstrap and auxiliary filters: (i) effective sample size ESS for bootstrap filter, (ii) filtered estimate of θ_t with 90% confidence interval obtained from bootstrap filter, (iii) as (i) for auxiliary filter, (iv) as (ii) obtained from auxiliary filter.

The auxiliary filter is discussed in Section 12.5 and is implemented for the basic SV model. The implementation for $N = 10,000$ is more involved when compared to the bootstrap filter. In panel (iii) of Fig. 14.7 we present the ESS variable for the auxiliary filter; it indicates that the number of effective samples is higher compared to the bootstrap filter. For most t, the number of effective particles is close to $N = 10,000$. The filtered estimate $\hat{a}_{t|t}$ of logvolatility θ_t is displayed in panel (iv) together with its 90% confidence interval based on $\hat{P}_{t|t}$. The filtered estimates from the bootstrap and auxiliary particle filters are virtually the same.

14.6 Binary density: Oxford–Cambridge boat race

In the last illustration we consider the outcomes of the annual boat race between teams representing the universities of Oxford and Cambridge. The race takes place from Putney to Mortlake on the River Thames in the month of March or April. The first took place in 1829 and was won by Oxford and, at the time of writing, the last took place in 2011 and was also won by Oxford; in the year 2010 Cambridge won in an epic battle and denied Oxford the hat-trick. In the first edition of the book, which we finished writing in 2000, we forecasted the probability for a Cambridge win in 2001 as 0.67 and, indeed, Cambridge did win in 2001.

There have been some occasions, especially in the nineteenth century, when the race took place elsewhere and in other months. In the years of both World Wars the race did not take place and there were also some years when the race finished with a dead heat or some other irregularity took place. Thus the time series of yearly outcomes contains missing observations for the years: 1830–1835, 1837, 1838, 1843, 1844, 1847, 1848, 1850, 1851, 1853, 1855, 1877, 1915–1919 and 1940–1945. In Fig. 14.8 the positions of the missing values are displayed as black dots with value 0.5. However, in the analysis we deal with these missing observations as described in Subsection 11.5.5.

The appropriate model for the boat race data is the binary distribution as described in Subsection 9.3.2. We take $y_t = 1$ if Cambridge wins and $y_t = 0$ if Oxford wins. Denoting the probability that Cambridge wins in year t by π_t, as in Subsection 9.3.2 we take $\theta_t = \log[\pi_t/(1 - \pi_t)]$. A winner this year is likely to be a winner next year because of overlapping crew membership, training methods and other factors. We model the transformed probability by the random walk

$$\theta_{t+1} = \theta_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big),$$

where η_t is serially uncorrelated for $t = 1, \ldots, n$.

The method based on mode estimation described in Subsection 10.6.4 provides the approximating model for this case and maximum likelihood estimation for the unknown variance σ_η^2 is carried out as described in Subsection 11.6.

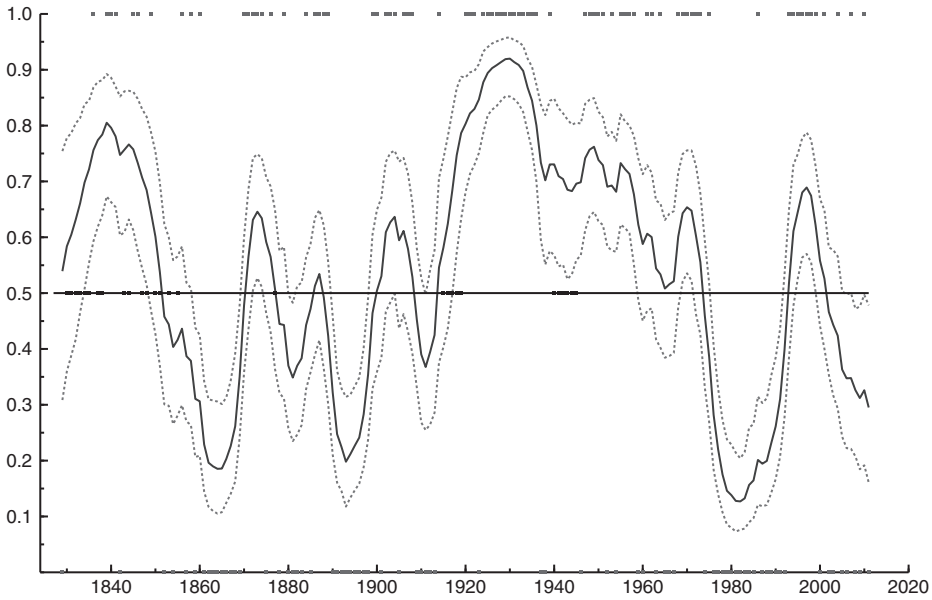

Fig. 14.8 Dot at zero is a win for Oxford, dot at one is a win for Cambridge and dot at 0.5 is a missing value; the solid line is the probability of a Cambridge win and the dotted lines constitute the 50% (asymmetric) confidence interval.

We have estimated the variance as $\hat{\sigma}_\eta^2 = 0.330$. The estimated mean of the probability π_t, indicating a win for Cambridge in year t, is computed using the method described in Subsection 11.5. The resulting time series of π_t is given in Fig. 14.8. The forecasted probability for a Cambridge win in 2012 is 0.30 and therefore we expect the Oxford team to win.