13 Bayesian estimation of parameters

13.1 Introduction

The parameters of a state space model can be divided into two classes, the state parameters and the additional parameters. For example, in the local level model (2.3) the state parameters are functions of $\alpha_1, \ldots, \alpha_n$ and the additional parameters are σ_ε^2 and σ_η^2. There are two methods of estimation of these parameters, classical analysis and Bayesian analysis; these arise from two different theories of probability. Classical analysis comes from a theory in which parameters are fixed and observations are random variables; Bayesian analysis comes from a theory in which state parameters are random variables and the observations are fixed. In state space analysis, state estimates are the same whether classical analysis or Bayesian analysis is employed by Lemmas 1 to 4 of Chapter 4. However, different treatments are needed for the estimation of additional parameters. classical treatments are based on fixed parameters and we have shown in previous chapters how to use the method of maximum likelihood for their estimation. For Bayesian treatments it turns out that we need to use simulation-based methods for the estimation of additional parameters, even for linear models. Thus it seemed logical to delay the Bayesian treatment of linear models until after simulation methods for nonlinear and non-Gaussian models had been dealt with. It was then natural to run on with the Bayesian treatment of nonlinear and non-Gaussian models. This explains why we did not attempt to deal with the Bayesian estimation of additional parameters in Part I of the book. This chapter deals with the Bayesian analysis of additional parameters as well as state parameters.

A Bayesian treatment of state space models is provided by the monograph of West and Harrison (1989, 1997). Other previous work on Bayesian analysis of non-Gaussian state space models has been based mainly on Markov chain Monte Carlo (MCMC) methods; we note in particular here the contributions by Carlin, Polson and Stoffer (1992), Frühwirth-Schnatter (1994, 2004), Shephard (1994b), Carter and Kohn (1994, 1996, 1997), Shephard and Pitt (1997), Cargnoni, Muller and West (1997) and Gamerman (1998). General accounts of Bayesian methodology and computation are given by Bernardo and Smith (1994), Gelman, Carlin, Stern and Rubin (1995) and Frühwirth-Schnatter (2006).

We first develop the analysis of the linear Gaussian state space model by constructing importance samples of additional parameters. We then show how to

combine these with Kalman filter and smoother outputs to obtain the estimates of state parameters required. We first do this for a proper prior in Subsection 13.2.1 and then extend to non-informative priors in Subsection 13.2.2. The discussion is extended to nonlinear and non-Gaussian models in Subsection 13.3. In Section 13.4 a brief description is given of the alternative simulation technique, Markov chain Monte Carlo (MCMC) methods. Although we prefer to use importance sampling methods for the problems considered in this book, MCMC methods are also used in time series applications.

13.2 Posterior analysis for linear Gaussian model

13.2.1 Posterior analysis based on importance sampling

Let ψ denote the vector of additional parameters. We consider a Bayesian analysis for the situation where the parameter vector ψ is not fixed and known; instead, we treat ψ as a random vector with a known prior density $p(\psi)$, which to begin with we take as a proper prior, leaving the non-informative case until later. For discussions of choice of prior in a general Bayesian analysis see Gelman, Carlin, Stern and Rubin (1995) and Bernardo and Smith (1994). The problems we shall consider amount essentially to the estimation of the posterior mean of a function $x(\alpha)$ of the stacked state vector α,

$$\bar{x} = \mathrm{E}[x(\alpha)|Y_n]. \tag{13.1}$$

Let $\bar{x}(\psi) = \mathrm{E}[x(\alpha)|\psi, Y_n]$ be the conditional expectation of $x(\alpha)$ given ψ and Y_n. We shall restrict consideration in this section to those functions $x(\alpha)$ for which $\bar{x}(\psi)$ can be readily calculated by the Kalman filter and smoother. This restricted class of functions still, however, includes many important cases such as the posterior mean and variance matrix of α_t and forecasts $\mathrm{E}(y_{n+j}|Y_n)$ for $j = 1, 2, \ldots$. The treatment of initialisation in Chapter 5 permits elements of the initial state vector α_1 to have either proper or diffuse prior densities.

We begin by attempting an analysis which is free of importance sampling. We have

$$\bar{x} = \int \bar{x}(\psi) p(\psi|Y_n)\, d\psi. \tag{13.2}$$

By Bayes theorem $p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi)$ where K is the normalising constant defined by

$$K^{-1} = \int p(\psi)p(Y_n|\psi)\, d\psi. \tag{13.3}$$

We therefore have

$$\bar{x} = \frac{\int \bar{x}(\psi)p(\psi)p(Y_n|\psi)\, d\psi}{\int p(\psi)p(Y_n|\psi)\, d\psi}. \tag{13.4}$$

Now $p(Y_n|\psi)$ is the likelihood, which for the linear Gaussian model is calculated by the Kalman filter as shown in Section 7.2. In principle, simulation could be applied directly to formula (13.4) by drawing a random sample $\psi^{(1)}, \ldots, \psi^{(N)}$

from the distribution with density $p(\psi)$ and then estimating the numerator and denominatior of (13.4) by the sample means of $\bar{x}(\psi)p(Y_n|\psi)$ and $p(Y_n|\psi)$ respectively. However, this estimator is inefficient in cases of practical interest.

We provide a treatment based on the simulation method of importance sampling that we have explored in Chapter 11. Suppose that simulation from density $p(\psi|Y_n)$ is impractical. Let $g(\psi|Y_n)$ be an importance density which is as close as possible to $p(\psi|Y_n)$ while at the same time permitting simulation. From (13.2) we have

$$\begin{aligned}\bar{x} &= \int \bar{x}(\psi)\frac{p(\psi|Y_n)}{g(\psi|Y_n)}g(\psi|Y_n)\,d\psi \\ &= \mathrm{E}_g\left[\bar{x}(\psi)\frac{p(\psi|Y_n)}{g(\psi|Y_n)}\right] \\ &= K\mathrm{E}_g[\bar{x}(\psi)z^g(\psi,Y_n)] \qquad (13.5)\end{aligned}$$

by Bayes theorem where E_g denotes expectation with respect to density $g(\psi|Y_n)$,

$$z^g(\psi,Y_n) = \frac{p(\psi)p(Y_n|\psi)}{g(\psi|Y_n)}, \qquad (13.6)$$

and K is a normalising constant. By replacing $\bar{x}(\psi)$ by 1 in (13.5) we obtain

$$K^{-1} = \mathrm{E}_g[z^g(\psi,Y_n)],$$

so the posterior mean of $x(\alpha)$ can be expressed as

$$\bar{x} = \frac{\mathrm{E}_g[\bar{x}(\psi)z^g(\psi,Y_n)]}{\mathrm{E}_g[z^g(\psi,Y_n)]}. \qquad (13.7)$$

This expression is evaluated by simulation. We choose random samples of N draws of ψ, denoted by $\psi^{(i)}$, from the importance density $g(\psi|Y_n)$ and estimate $\bar{x}$ by

$$\hat{x} = \frac{\sum_{i=1}^N \bar{x}\big(\psi^{(i)}\big)z_i}{\sum_{i=1}^N z_i}, \qquad (13.8)$$

where

$$z_i = \frac{p\big(\psi^{(i)}\big)p\big(Y_n|\psi^{(i)}\big)}{g\big(\psi^{(i)}|Y_n\big)}. \qquad (13.9)$$

As an importance density for $p(\psi|Y_n)$ we take its large sample normal approximation

$$g(\psi|Y_n) = \mathrm{N}(\hat{\mu},\hat{\Omega}),$$

where $\hat{\psi}$ is the solution to the equation

$$\frac{\partial \log p(\psi|Y_n)}{\partial \psi} = \frac{\partial \log p(\psi)}{\partial \psi} + \frac{\partial \log p(Y_n|\psi)}{\partial \psi} = 0, \tag{13.10}$$

and

$$\hat{\Omega}^{-1} = -\frac{\partial^2 \log p(\psi)}{\partial \psi \partial \psi'} - \frac{\partial^2 \log p(Y_n|\psi)}{\partial \psi \partial \psi'}\bigg|_{\psi=\hat{\psi}}. \tag{13.11}$$

For a discussion of this large sample approximation to $p(\psi|Y_n)$ see Gelman, Carlin, Stern and Rubin (1995, Chapter 4) and Bernardo and Smith (1994, §5.3). Since $p(Y_n|\psi)$ can easily be computed by the Kalman filter for $\psi = \psi^{(i)}$, $p(\psi)$ is given and $g(\psi|Y_n)$ is Gaussian, the value of z_i is easy to compute. The draws for $\psi^{(i)}$ are independent and therefore $\hat{x}$ converges probabilistically to $\bar{x}$ as $N \to \infty$ under very general conditions.

The value $\hat{\psi}$ is computed iteratively by an obvious extension of the technique of maximum likelihood estimation as discussed in Chapter 7 while the second derivatives can be calculated numerically. Once $\hat{\psi}$ and $\hat{\Omega}$ are computed, it is straightforward to generate samples from $g(\psi|Y_n)$ by use of a standard normal random number generator. Where needed, efficiency can be improved by the use of antithetic variables, which we discuss in Subsection 11.4.3. For example, for each draw $\psi^{(i)}$ we could take another value $\tilde{\psi}^{(i)} = 2\hat{\psi} - \psi^{(i)}$, which is equiprobable with $\psi^{(i)}$. The use of $\psi^{(i)}$ and $\tilde{\psi}^{(i)}$ together introduces balance in the sample.

The posterior mean of the parameter vector ψ is $\bar{\psi} = \mathrm{E}(\psi|Y_n)$. An estimate $\tilde{\psi}$ of $\bar{\psi}$ is obtained by putting $\bar{x}(\psi^{(i)}) = \psi^{(i)}$ in (13.8) and taking $\tilde{\psi} = \hat{x}$. Similarly, an estimate $\tilde{\mathrm{V}}(\psi|Y_n)$ of the posterior variance matrix $\mathrm{Var}(\psi|Y_n)$ is obtained by putting $\bar{x}(\psi^{(i)}) = \psi^{(i)}\psi^{(i)\prime}$ in (13.8), taking $\tilde{S} = \hat{x}$ and then taking $\tilde{\mathrm{V}}(\psi|Y_n) = \tilde{S} - \tilde{\psi}\tilde{\psi}'$.

To estimate the posterior distribution function of an element ψ_1 of ψ, which is not necessarily the first element of ψ, we introduce the indicator function $I_1(\psi_1^{(i)})$ which equals one if $\psi_1^{(i)} \leq \psi_1$ and zero otherwise, where $\psi_1^{(i)}$ is the value of ψ_1 in the ith simulated value of ψ and ψ_1 is fixed. Then $\mathrm{F}(\psi_1|Y_n) = \Pr(\psi_1^{(i)} \leq \psi_1) = \mathrm{E}[I_1(\psi_1^{(i)})|Y_n]$ is the posterior distribution function of ψ_1. Putting $\bar{x}(\psi^{(i)}) = I_1(\psi^{(i)})$ in (13.8), we estimate $\mathrm{F}(\psi_1|Y_n)$ by $\tilde{\mathrm{F}}(\psi_1|Y_n) = \hat{x}$. This is equivalent to taking $\tilde{\mathrm{F}}(\psi_1|Y_n)$ as the sum of values of z_i for which $\psi_1^{(i)} \leq \psi_1$ divided by the sum of all values of z_i. Similarly, if δ is the interval $(\psi_1 - \frac{1}{2}d, \psi_1 + \frac{1}{2}d)$ where d is small and positive then we can estimate the posterior density of ψ_1 by $\tilde{p}(\psi_1|Y_n) = d^{-1}S^{\delta}/\sum_{i=1}^{N} z_i$ where S^{δ} is the sum of the values of z_i for which $\psi_1^{(i)} \in \delta$.

13.2.2 Non-informative priors

For cases where a proper prior is not available we may wish to use a noninformative prior in which we assume that the prior density is proportional to a specified

function $p(\psi)$ in a domain of ψ of interest even though the integral $\int p(\psi)d\psi$ does not exist. For a discussion of noninformative priors see, for example, Chapters 3 and 4 of Gelman, Carlin, Stern and Rubin (1995). Where it exists, the posterior density is $p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi)$ as in the proper prior case, so all the previous formulae apply without change. This is why we use the same symbol $p(\psi)$ for both cases even though in the noninformative case $p(\psi)$ is not a density. An important special case is the diffuse prior for which $p(\psi) = 1$ for all ψ.

13.3 Posterior analysis for a nonlinear non-Gaussian model

In this section we develop Bayesian techniques for estimating posterior means and posterior variance matrices of functions of the state vector for a nonlinear non-Gaussian model. We also show how to estimate posterior distribution and density functions of scalar functions of the state vector. It turns out that the basic ideas of importance sampling and antithetic variables developed for classical analysis in Chapter 11 can be applied with little essential change to the Bayesian case. Different considerations apply to questions regarding the posterior distribution of the parameter vector and we deal with these in Subsection 13.3.3. The treatment is based on the methods developed by Durbin and Koopman (2000).

13.3.1 Posterior analysis of functions of the state vector

We first obtain some basic formulae analogous to those derived in Section 11.2 for the classical case. Suppose that we wish to calculate the posterior mean $\bar{x} = \mathrm{E}[x(\alpha)|Y_n]$ of a function $x(\alpha)$ of the stacked state vector α given the stacked observation vector Y_n. As we shall show, this is a general formulation which enables us not only to estimate posterior means of quantities of interest such as the trend or seasonal, but also posterior variance matrices and posterior distribution functions and densities of scalar functions of the state. We shall estimate $\bar{x}$ by simulation techniques based on importance sampling and antithetic variables analogous to those developed in Chapter 11 for the classical case.

We have

$$\begin{aligned}\bar{x} &= \int x(\alpha)p(\psi,\alpha|Y_n)\,d\psi d\alpha \\ &= \int x(\alpha)p(\psi|Y_n)p(\alpha|\psi,Y_n)\,d\psi d\alpha. \end{aligned} \tag{13.12}$$

As an importance density for $p(\psi|Y_n)$ we take its large sample normal approximation

$$g(\psi|Y_n) = \mathrm{N}(\hat{\psi}, \hat{V}),$$

where $\hat{\psi}$ is the solution of the equation

$$\frac{\partial \log p(\psi|Y_n)}{\partial \psi} = \frac{\partial \log p(\psi)}{\partial \psi} + \frac{\partial \log p(Y_n|\psi)}{\partial \psi} = 0, \tag{13.13}$$

and

$$\hat{V}^{-1} = -\frac{\partial^2 \log p(\psi)}{\partial\psi\partial\psi'} - \frac{\partial^2 \log p(Y_n|\psi)}{\partial\psi\partial\psi'}\bigg|_{\psi=\hat{\psi}}. \tag{13.14}$$

For a discussion of this large sample approximation to $p(\psi|Y_n)$ see Gelman, Carlin, Stern and Rubin (1995, Chapter 4) and Bernardo and Smith (1994, §5.3).

Let $g(\alpha|\psi, Y_n)$ be a Gaussian importance density for α given ψ and Y_n which is obtained from an approximating linear Gaussian model in the way described in Chapter 11. From (13.12),

$$\begin{aligned}\bar{x} &= \int x(\alpha)\frac{p(\psi|Y_n)p(\alpha|\psi,Y_n)}{g(\psi|Y_n)g(\alpha|\psi,Y_n)}g(\psi|Y_n)g(\alpha|\psi,Y_n)\,d\psi d\alpha \\ &= \int x(\alpha)\frac{p(\psi|Y_n)g(Y_n|\psi)p(\alpha,Y_n|\psi)}{g(\psi|Y_n)p(Y_n|\psi)g(\alpha,Y_n|\psi)}g(\psi,\alpha|Y_n)\,d\psi d\alpha.\end{aligned}$$

By Bayes theorem,

$$p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi),$$

in which K is a normalising constant, so we have

$$\begin{aligned}\bar{x} &= K\int x(\alpha)\frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\alpha,Y_n|\psi)}{g(\alpha,Y_n|\psi)}g(\psi,\alpha|Y_n)\,d\psi d\alpha \\ &= K\mathrm{E}_g\left[x(\alpha)z(\psi,\alpha,Y_n)\right], \end{aligned}\tag{13.15}$$

where E_g denotes expectation with respect to the importance joint density

$$g(\psi,\alpha|Y_n) = g(\psi|Y_n)g(\alpha|\psi,Y_n),$$

and where

$$z(\psi,\alpha,Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\alpha,Y_n|\psi)}{g(\alpha,Y_n|\psi)}. \tag{13.16}$$

In this formula, $g(Y_n|\psi)$ is the likelihood for the approximating Gaussian model, which is easily calculated by the Kalman filter.

Taking $x(\alpha) = 1$ in (13.15) gives

$$K^{-1} = \mathrm{E}_g[z(\psi,\alpha,Y_n)],$$

so we have finally

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha)z(\psi,\alpha,Y_n)]}{\mathrm{E}_g[z(\psi,\alpha,Y_n)]}. \tag{13.17}$$

We note that (13.17) differs from the corresponding formula (11.8) in the classical inference case only in the replacement of $w(\alpha, Y_n)$ by $z(\psi, \alpha, Y_n)$ and the inclusion of ψ in the importance density $g(\psi, \alpha|Y_n)$.

In the important special case in which the state equation error η_t is $\mathrm{N}(0, Q_t)$, then α is Gaussian so we can write its density as $g(\alpha)$ and use this as the state density for the approximating model. This gives $p(\alpha, Y_n|\psi) = g(\alpha)p(Y_n|\theta, \psi)$ and $g(\alpha, Y_n|\psi) = g(\alpha)g(Y_n|\theta, \psi)$, where θ is the stacked vector of signals $\theta_t = Z_t\alpha_t$, so (13.16) simplifies to

$$z(\psi, \alpha, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(Y_n|\theta, \psi)}{g(Y_n|\theta, \psi)}. \tag{13.18}$$

For cases where a proper prior is not available, we may wish to use a non-informative prior in which we assume that the prior density is proportional to a specified function $p(\psi)$ in a domain of ψ of interest even though the integral $\int p(\psi)d\psi$ does not exist. The posterior density, where it exists, is

$$p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi),$$

which is the same as in the proper prior case, so all the previous formulae apply without change. This is why we can use the same symbol $p(\psi)$ in both cases even when $p(\psi)$ is not a proper density. An important special case is the diffuse prior for which $p(\psi) = 1$ for all ψ. For a general discussion of noninformative priors, see, for example, Gelman, Carlin, Stern and Rubin (1995, Chapters 2 and 3).

13.3.2 Computational aspects of Bayesian analysis

For practical computations based on these ideas we express the formulae in terms of variables that are as simple as possible as in Section 11.4, Subsections 11.5.3 and 11.5.5 for the classical analysis. This means that to the maximum feasible extent we employ formulae based on the disturbance terms $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ and $\varepsilon_t = y_t - \theta_t$ for $t = 1, \ldots, n$. By repeated substitution for α_t we first obtain $x(\alpha)$ as a function $x^*(\eta)$ of η. We then note that in place of (13.12) we obtain the posterior mean of $x^*(\eta)$,

$$\bar{x} = \int x^*(\eta)p(\psi|Y_n)p(\eta|\psi, Y_n)\, d\psi d\eta. \tag{13.19}$$

By reductions analogous to those above we obtain in place (13.17)

$$\bar{x} = \frac{\mathrm{E}_g[x^*(\eta)z^*(\psi, \eta, Y_n)]}{\mathrm{E}_g[z^*(\psi, \eta, Y_n)]}, \tag{13.20}$$

where

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}, \tag{13.21}$$

and E_g denotes expectation with respect to the importance density $g(\psi, \eta|Y_n)$.

Let $\psi^{(i)}$ be a random draw from the importance density for ψ, $g(\psi|Y_n) = \mathrm{N}(\hat{\psi}, \hat{V})$, where $\hat{\psi}$ satisfies (13.13) and $\hat{V}$ is given by (13.14), and let $\eta^{(i)}$ be a random draw from density $g(\eta|\psi^{(i)}, Y_n)$ for $i = 1, \ldots, N$. To obtain this we need an approximation to the mode $\hat{\eta}^{(i)}$ of density $g(\eta|\psi^{(i)}, Y_n)$ but this is rapidly obtained in a few iterations from the mode of $g(\eta|\hat{\psi}, Y_n)$. Let

$$x_i = x^*\big(\eta^{(i)}\big), \qquad z_i = z^*\big(\psi^{(i)}, \eta^{(i)}, Y_n\big), \tag{13.22}$$

and consider as an estimate of $\bar{x}$ the ratio

$$\hat{x} = \frac{\sum_{i=1}^N x_i z_i}{\sum_{i=1}^N z_i}. \tag{13.23}$$

The efficiency of this estimate can obviously be improved by the use of antithetic variables. For $\eta^{(i)}$ we can use the location and scale antithetics described in Subsection 11.4.3. Antithethics may not be needed for $\psi^{(i)}$ since $\hat{V} = O(n^{-1})$ but it is straightforward to allow for them if their use is worthwhile; for example, it would be an easy matter to employ the location antithetic $\tilde{\psi}^{(i)} = 2\hat{\psi} - \psi^{(i)}$.

There is flexibility in the way the pairs $\psi^{(i)}$, $\eta^{(i)}$ are chosen, depending on the number of antithetics employed and the way the values of ψ and η are combined. For example, one could begin by making a random selection ψ^s of ψ from $\mathrm{N}(\hat{\psi}, \hat{V})$. Next we compute the antithetic value $\tilde{\psi}^s = 2\hat{\psi} - \psi^s$. For each of the values ψ^s and $\tilde{\psi}^s$ one could draw separate values of η from $g(\eta|\psi, Y_n)$, and then employ the two antithetics for each η that are described in Subsection 11.4.3. Thus in the sample there are four values of η combined with each value of ψ so N is a multiple of four and the number of draws of η from the simulation smoother is $N/4$. For estimation of variances due to simulation we need however to note that, since ψ^s and $\tilde{\psi}^s$ are related, there are only $N/8$ independent draws from the joint importance density $g(\psi, \eta|Y_n)$. For the purpose of estimating posterior variances of scalar quantities, assume that $x^*(\eta)$ is a scalar. Then, as in (11.21), the estimate of its posterior variance is

$$\widehat{\mathrm{Var}}[x^*(\eta)|Y_n] = \frac{\sum_{i=1}^N x_i^2 z_i}{\sum_{i=1}^N z_i} - \hat{x}^2. \tag{13.24}$$

Let us now consider the estimation of variance of the estimate $\hat{x}$ of the posterior mean of scalar $x^*(\eta)$ due to simulation. As indicated above the details depend on the way values of ψ and η are combined. For the example we considered, with a single antithetic for ψ and two antithetics for η, combined in the way described, let $\hat{v}_j^\dagger$ be the sum of the eight associated values of $z_i(x_i - \hat{x})$. Then as in (11.23), the estimate of the variance of $\hat{x}$ due to errors of simulation is

$$\widehat{\mathrm{Var}}_s(\hat{x}) = \frac{\sum_{j=1}^{N/8} \hat{v}_j^{\dagger 2}}{\Big(\sum_{i=1}^N z_i\Big)^2}. \tag{13.25}$$

For the estimation of posterior distribution functions and densities of scalar $x^*(\eta)$, let $I_x(\eta)$ be an indicator which is unity if $x^*(\eta) \leq x$ and is zero if $x^*(\eta) > x$. Then the posterior distribution function is estimated by (11.24) provided that w_i is replaced by z_i. With the same proviso, the posterior density of $x^*(\eta)$ is estimated by (11.25). Samples of independent values from the estimated posterior distribution can be obtained by a method analogous to that described by a method at the end of Subsection 11.5.3.

13.3.3 Posterior analysis of parameter vector

In this section we consider the estimation of posterior means, variances, distribution functions and densities of functions of the parameter vector ψ. Denote by $\nu(\psi)$ the function of ψ whose posterior properties we wish to investigate. Using Bayes theorem, the posterior mean of $\nu(\psi)$ is

$$\begin{aligned}\bar{\nu} &= \mathrm{E}[\nu(\psi)|Y_n] \\ &= \int \nu(\psi) p(\psi|Y_n)\, d\psi \\ &= K \int \nu(\psi) p(\psi) p(Y_n|\psi)\, d\psi \\ &= K \int \nu(\psi) p(\psi) p(\eta, Y_n|\psi)\, d\psi d\eta, \end{aligned} \tag{13.26}$$

where K is a normalising constant. Introducing importance densities $g(\psi|Y_n)$ and $g(\eta|\psi, Y_n)$ as in Subsection 13.3.2, we have

$$\begin{aligned}\bar{\nu} &= K \int \nu(\psi) \frac{p(\psi) g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)} g(\psi, \eta|Y_n)\, d\psi d\eta \\ &= K \mathrm{E}_g[\nu(\psi) z^*(\psi, \eta, Y_n)], \end{aligned} \tag{13.27}$$

where E_g denotes expectation with respect to the joint importance density $g(\psi, \eta|Y_n)$ and

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi) g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}.$$

Putting $\nu(\psi) = 1$ in (13.27) we obtain as in (13.20),

$$\bar{\nu} = \frac{\mathrm{E}_g[\nu(\psi) z^*(\psi, \eta, Y_n)]}{\mathrm{E}_g[z^*(\psi, \eta, Y_n)]}. \tag{13.28}$$

In the simulation, take $\psi^{(i)}$ and $\eta^{(i)}$ as in Subsection 13.3.2 and let $\nu_i = \nu(\psi^{(i)})$. Then the estimates $\hat{\nu}$ of $\bar{\nu}$ and $\widehat{\mathrm{Var}}[\nu(\psi)|Y_n]$ of $\mathrm{Var}[\nu(\psi)|Y_n]$ are given by (13.23) and (13.24) by replacing x_i by ν_i. Similarly, the variance of $\hat{\nu}$ due to simulation can, for the antithetics considered in Subsection 11.5.3, be calculated

by defining $v_j^\dagger$ as the sum of the eight associated values of $z_i(\nu_i - \bar{\nu})$ and using (13.25) to obtain the estimate $\widehat{\mathrm{Var}}_s(\hat{\nu})$. Estimates of the posterior distribution and density functions are obtained by the indicator function techniques described at the end of Subsection 13.3.2. While $\hat{\nu}$ can be a vector, for the remaining estimates $\nu(\psi)$ has to be a scalar quantity.

The estimate of the posterior density $p[\nu(\psi)|Y_n]$ obtained in this way is essentially a histogram estimate, which is accurate at values of $\nu(\psi)$ near the midpoint of the intervals containing them. An alternative estimate of the posterior density of a particular element of ψ, which is accurate at any value of the element, was proposed by Durbin and Koopman (2000). Without loss of generality take this element to be the first element of ψ and denote it by ψ_1. Denote the remaining elements by ψ_2. Let $g(\psi_2|\psi_1, Y_n)$ be the approximate conditional density of ψ_2 given ψ_1 and Y_n, which is easily obtained by applying standard regression theory to $g(\psi|Y_n)$, where $g(\psi|Y_n) = \mathrm{N}(\hat{\mu}, \hat{V})$. We take $g(\psi_2|\psi_1, Y_n)$ as an importance density in place of $g(\psi|Y_n)$. Then

$$\begin{aligned} p(\psi_1|Y_n) &= \int p(\psi|Y_n)\, d\psi_2 \\ &= K \int p(\psi) p(Y_n|\psi)\, d\psi_2 \\ &= K \int p(\psi) p(\eta, Y_n|\psi)\, d\psi_2 d\eta \\ &= K \mathrm{E}_g[\tilde{z}(\psi, \eta, Y_n)], \end{aligned} \tag{13.29}$$

where E_g denotes expectation with respect to importance density $g(\psi_2|\psi_1, Y_n)$ and

$$\tilde{z}(\psi, \eta, Y_n) = \frac{p(\psi) g(Y_n|\psi)}{g(\psi_2|\psi_1, Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}. \tag{13.30}$$

Let $\tilde{\psi}_2^{(i)}$ be a draw from $g(\psi_2|\psi_1, Y_n)$, let $\tilde{\psi}^{(i)} = (\psi_1, \tilde{\psi}_2^{(i)\prime})'$ and let $\tilde{\eta}^{(i)}$ be a draw from $g(\eta|\tilde{\psi}^{(i)}, Y_n)$. Then take

$$\tilde{z}_i = \frac{p\big(\tilde{\psi}^{(i)}\big) g\big(Y_n|\tilde{\psi}^{(i)}\big)}{g\big(\tilde{\psi}_2^{(i)}|\psi_1, Y_n\big)} \frac{p\big(\tilde{\eta}^{(i)}, Y_n|\tilde{\psi}^{(i)}\big)}{g\big(\tilde{\eta}^{(i)}, Y_n|\tilde{\psi}^{(i)}\big)}. \tag{13.31}$$

Now as in (13.28),

$$K^{-1} = \mathrm{E}_g[z^*(\psi, \eta, Y_n)],$$

where E_g denotes expectation with respect to importance density $g(\psi|Y_n)$ $g(\eta|\psi, Y_n)$ and

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi) g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}.$$

Let ψ_i^* be a draw from $g(\psi|Y_n)$ and let η_i^* be a draw from $g(\eta|\psi_i^*, Y_n)$. Then take

$$z_i^* = \frac{p(\psi_i^*)g(Y_n|\psi_i^*)}{g(\psi_i^*|Y_n)} \frac{p(\eta_i^*, Y_n|\psi_i^*)}{g(\eta_i^*, Y_n|\psi_i^*)}, \tag{13.32}$$

and estimate $p(\psi_i|Y_n)$ by the simple form

$$\hat{p}(\psi_i|Y_n) = \sum_{i=1}^{N} \tilde{z}_i / \sum_{i=1}^{N} z_i^*. \tag{13.33}$$

The simulations for the numerator and denominator of (13.33) are different since for the numerator only ψ_2 is drawn, whereas for the denominator the whole vector ψ is drawn. The variability of the ratio can be reduced however by employing the same set of N(0, 1) deviates employed for choosing η from $p(\eta|\tilde{\psi}^{(i)}, Y_n)$ in the simulation smoother as for choosing η from $p(\eta|\psi_i^*, Y_n)$. The variability can be reduced further by first selecting ψ_{1i}^* from $g(\psi_1|Y_n)$ and then using the same set of N(0, 1) deviates to select ψ_{2i}^* from $g(\psi_2|\psi_{1i}^*, Y_n)$ as were used to select $\tilde{\psi}_2^{(i)}$ from $g(\psi_2|\psi_1, Y_n)$ when computing $\tilde{z}_i$; in this case $g(\psi^*|Y_n)$ in (13.32) is replaced by $g(\psi_1^*)g(\psi_2^*|\psi_{1i}^*, Y_n)$.

To improve efficiency, antithetics may be used for draws of ψ and η in the way suggested in Subsection 13.3.2.

13.4 Markov chain Monte Carlo methods

An alternative approach to Bayesian analysis based on simulation is provided by the *Markov chain Monte Carlo* (MCMC) method which has received a substantial amount of interest in the statistical and econometric literature on time series. We briefly outline here the basic ideas of MCMC as applied to state space models. Frühwirth-Schnatter (1994) was the first to give a full Bayesian treatments of the linear Gaussian model using MCMC techniques. The proposed algorithms for simulation sample selection were later refined by Carter and Kohn (1994) and de Jong and Shephard (1995). This work resulted in the simulation smoother of Durbin and Koopman (2002) which we discussed in Section 4.9. We showed there how to generate random draws from the conditional densities $p(\varepsilon|Y_n, \psi)$, $p(\eta|Y_n, \psi)$ and $p(\alpha|Y_n, \psi)$ for a given parameter vector ψ. Now we briefly discuss how this technique can be incorporated into a Bayesian MCMC analysis in which we treat the parameter vector as stochastic.

The basic idea is as follows. We evaluate the posterior mean of $x(\alpha)$ or of the parameter vector ψ via simulation by choosing samples from an augmented joint density $p(\psi, \alpha|Y_n)$. In the MCMC procedure, the sampling from this joint density is implemented as a Markov chain. After initialisation for ψ, say $\psi = \psi^{(0)}$ we repeatedly cycle through the two simulation steps:

(1) sample $\alpha^{(i)}$ from $p(\alpha|Y_n, \psi^{(i-1)})$;
(2) sample $\psi^{(i)}$ from $p(\psi|Y_n, \alpha^{(i)})$;

for $i = 1, 2 \ldots$. After a number of 'burning-in' iterations we are allowed to treat the samples from step (2) as being generated from the density $p(\psi|Y_n)$. The attraction of this MCMC scheme is that sampling from conditional densities is easier than sampling from the marginal density $p(\psi|Y_n)$. The circumstances under which subsequent samples from the marginal densities $p(\alpha|Y_n, \psi^{(i-1)})$ and $p(\psi|Y_n, \alpha^{(i)})$ converge to samples from the joint density $p(\psi, \alpha|Y_n)$ are considered in books on MCMC, for example, Gamerman and Lopes (2006). It is not straightforward to develop appropriate diagnostics which indicate whether convergence within the MCMC process has taken place, as is discussed, for example, in Gelman (1995).

There exist various implementations of the basic MCMC algorithm for the state space model. For example, Carlin, Polson and Stoffer (1992) propose samplling individual state vectors from $p(\alpha_t|Y_n, \alpha^t, \psi)$ where α^t is equal to α excluding α_t. It turns out that this approach to sampling is inefficient. It is argued by Frühwirth-Schnatter (1994) that it is more efficient to sample all the state vectors directly from the density $p(\alpha|Y_n, \psi)$. She provides the technical details of implementation. de Jong and Shephard (1995) have developed this approach further by concentrating on the disturbance vectors ε_t and η_t instead of the state vector α_t. The details regarding the resulting simulation smoother were given in Section 4.9.

Implementing the two steps of the MCMC is not as straightforward as suggested so far. Sampling from the density $p(\alpha|Y_n, \psi)$ for a given ψ is done by using the simulation smoother of Section 4.9. Sampling from $p(\psi|Y_n, \alpha)$ depends partly on the model for ψ and is usually only possible up to proportionality. To sample under such circumstances, *accept-reject* algorithms have been developed; for example, the Metropolis algorithm is often used for this purpose. Details and an excellent general review of these matters are given by Gilks, Richardson and Spiegelhalter (1996). Applications to state space models have been developed by Carter and Kohn (1994), Shephard (1994b), Gamerman (1998) and Frühwirth-Schnatter (2006).

In the case of structural time series models of Section 3.2 for which the parameter vector consists only of variances of disturbances associated with the components, the distribution of the parameter vector can be modelled such that sampling from $p(\psi|Y_n, \alpha)$ in step (2) is relatively straightforward. For example, a model for a variance can be based on the inverse gamma distribution with logdensity

$$\log p(\sigma^2|c, s) = -\log \Gamma\left(\frac{c}{2}\right) - \frac{c}{2}\log\frac{s}{2} - \frac{c+2}{2}\log\sigma^2 - \frac{s}{2\sigma^2}, \qquad \text{for } \sigma^2 > 0,$$

and $p(\sigma^2|c, s) = 0$ for $\sigma^2 \leq 0$; see, for example, Poirier (1995). We denote this density by $\sigma^2 \sim \text{IG}(c/2, s/2)$ where c determines the shape and s determines the scale of the distribution. It has the convenient property that if we take this as

the prior density of σ^2 and we take a sample $u_1, \ldots, u_n$ of independent $\mathrm{N}(0, \sigma^2)$ variables, the posterior density of σ^2 is

$$p(\sigma^2|u_1, \ldots, u_n) = \mathrm{IG}\left[(c+n)/2, \left(s + \sum_{i=1}^{n} u_i^2\right) \Big/ 2\right];$$

for further details see, for example, Poirier (1995). For the implementation of step (2) a sample value of σ^2 is chosen from this density. We can take u_t as an element of ε_t or η_t obtained by the simulation smoother in step (1). Further details of this approach are given by Frühwirth-Schnatter (1994, 2006) and Carter and Kohn (1994).