11 Importance sampling for smoothing

11.1 Introduction

In this chapter we develop the methodology of importance sampling based on simulation for the analysis of observations from the non-Gaussian and nonlinear models that we have specified in Chapter 9. Unlike the treatment of linear models in Chapters 2 and 4, we deal with smoothing first and leave filtering for Chapter 12. The main reason is that filtering is dealt with by the method of particle filtering which is based on importance sampling methods such as those discussed in this chapter. We show that importance sampling methods can be adopted for the estimation of functions of the state vector and the estimation of the error variance matrices of the resulting estimates. We shall also develop estimates of conditional densities, distribution functions and quantiles of interest, given the observations. Of key importance is the method of estimating unknown parameters by maximum likelihood. The methods are based on standard ideas in simulation methodology and, in particular, importance sampling. In this chapter we will develop the basic ideas of importance sampling that we employ in our methodology for nonlinear non-Gaussian state space models. Details of applications to particular models will be given in later sections.

Importance sampling was introduced as early as Kahn and Marshall (1953) and Marshall (1956) and was described in books by Hammersley and Handscomb (1964, Section 5.4) and Ripley (1987, Chapter 5). It was first used in econometrics by Kloek and Van Dijk (1978) in their work on computing posterior densities.

The general model of interest in this chapter is given by

$$y_t \sim p(y_t|\alpha_t), \qquad \alpha_{t+1} = T_t(\alpha_t) + R_t\eta_t, \qquad \eta_t \sim p(\eta_t), \tag{11.1}$$

where both $p(y_t|\alpha_t)$ and $p(\eta_t)$, for $t = 1, \ldots, n$, can be non-Gaussian densities. We also give attention to the special case of a signal model with a linear Gaussian state equation, that is,

$$y_t \sim p(y_t|\theta_t), \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q_t), \tag{11.2}$$

for $t = 1, \ldots, n$, where $\theta_t = Z_t\alpha_t$ and R_t is a selection matrix with $R_t'R_t = I_r$ and r is the number of disturbances in η_t; see Table 4.1. Denote the stacked vectors

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

$(\alpha_1', \dots, \alpha_{n+1}')'$, $(\theta_1', \dots, \theta_n')'$ and $(y_1', \dots, y_n')'$ by α, θ and Y_n, respectively. In order to keep the exposition simple we shall assume in this section and the next section that the initial density $p(\alpha_1)$ is nondegenerate and known. The case where some of the elements of α_1 are diffuse will be considered in Subsection 11.4.4.

We mainly focus on the estimation of the conditional mean

$$\bar{x} = \mathrm{E}[x(\alpha)|Y_n] = \int x(\alpha)p(\alpha|Y_n)\,d\alpha, \tag{11.3}$$

of an arbitrary function $x(\alpha)$ of α given the observation vector Y_n. This formulation includes estimates of quantities of interest such as the mean $\mathrm{E}(\alpha_t|Y_n)$ of the state vector α_t given Y_n and its conditional variance matrix $\mathrm{Var}(\alpha_t|Y_n)$; it also includes estimates of the conditional density and distribution function of $x(\alpha)$ given Y_n when $x(\alpha)$ is scalar. The conditional density $p(\alpha|Y_n)$ depends on an unknown parameter vector ψ, but in order to keep the notation simple we shall not indicate this dependence explicitly in this chapter; the estimation of ψ is considered in Section 11.6.

In theory, we could draw a random sample of values from the distribution with density $p(\alpha|Y_n)$ and estimate $\bar{x}$ by the sample mean of the corresponding values of $x(\alpha)$. In practice, however, since explicit expressions are not available for $p(\alpha|Y_n)$ for the models of Chapter 9, this idea is not feasible. Instead, we seek a density as close to $p(\alpha|Y_n)$ as possible for which random draws are available, and we sample from this, making an appropriate adjustment to the integral in (11.3). This technique is called *importance sampling* and the density is referred to as the *importance density.* The techniques we shall describe will be based on Gaussian importance densities since these are available for the problems we shall consider and they work well in practice. We shall use the generic notation $g(\cdot)$, $g(\cdot,\cdot)$ and $g(\cdot|\cdot)$ for marginal, joint and conditional densities, respectively.

11.2 Basic ideas of importance sampling

Consider model (11.1) and let $g(\alpha|Y_n)$ be an importance density which is chosen to resemble $p(\alpha|Y_n)$ as closely as is reasonably possible while being easy to sample from; we have from (11.3),

$$\bar{x} = \int x(\alpha)\frac{p(\alpha|Y_n)}{g(\alpha|Y_n)}g(\alpha|Y_n)\,d\alpha = \mathrm{E}_g\left[x(\alpha)\frac{p(\alpha|Y_n)}{g(\alpha|Y_n)}\right], \tag{11.4}$$

where E_g denotes expectation with respect to the importance density $g(\alpha|Y_n)$. For the models of Chapter 9, $p(\alpha|Y_n)$ and $g(\alpha|Y_n)$ are complicated algebraically, whereas the corresponding joint densities $p(\alpha, Y_n)$ and $g(\alpha, Y_n)$ are straightforward. We therefore put $p(\alpha|Y_n) = p(\alpha, Y_n)/p(Y_n)$ and $g(\alpha|Y_n) = g(\alpha, Y_n)/g(Y_n)$ in (11.4), giving

$$\bar{x} = \frac{g(Y_n)}{p(Y_n)}\mathrm{E}_g\left[x(\alpha)\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}\right]. \tag{11.5}$$

Putting $x(\alpha) = 1$ in (11.5) we have

$$1 = \frac{g(Y_n)}{p(Y_n)} \mathrm{E}_g \left[\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} \right], \tag{11.6}$$

and effectively obtain an expression for the observation density

$$p(Y_n) = g(Y_n) \, \mathrm{E}_g \left[\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} \right]. \tag{11.7}$$

Taking the ratio of (11.5) and (11.6) gives

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha) w(\alpha, Y_n)]}{\mathrm{E}_g[w(\alpha, Y_n)]}, \quad \text{where} \quad w(\alpha, Y_n) = \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}. \tag{11.8}$$

In model (11.1)

$$p(\alpha, Y_n) = p(\alpha_1) \prod_{t=1}^{n} p(\eta_t) p(y_t | \alpha_t), \tag{11.9}$$

where $\eta_t = R_t' \left[\alpha_{t+1} - T_t(\alpha_t) \right]$ for $t = 1, \dots, n$, since $R_t' R_t = I^r$.

Expression (11.8) provides the basis for importance sampling. We could in principle obtain a Monte Carlo estimate $\hat{x}$ of $\bar{x}$ in the following way. Choose a series of independent draws $\alpha^{(1)}, \dots, \alpha^{(N)}$ from the distribution with density $g(\alpha | Y_n)$ and take

$$\hat{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}, \quad \text{where} \quad x_i = x\big(\alpha^{(i)}\big) \quad \text{and} \quad w_i = w\big(\alpha^{(i)}, Y_n\big). \tag{11.10}$$

Since the draws are independent, the law of large numbers applies and under assumptions which are usually satisfied in cases of practical interest, $\hat{x}$ converges to $\bar{x}$ probabilistically as $N \to \infty$.

An important special case is where the observations are non-Gaussian but the state equation is linear Gaussian, that is, where we consider model (11.2). We then have $p(\alpha) = g(\alpha)$ so

$$\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} = \frac{p(\alpha) p(Y_n | \alpha)}{g(\alpha) g(Y_n | \alpha)} = \frac{p(Y_n | \alpha)}{g(Y_n | \alpha)} = \frac{p(Y_n | \theta)}{g(Y_n | \theta)},$$

where θ is the stacked vector of the signals $\theta_t = Z_t \alpha_t$ for $t = 1, \dots, n$. Thus (11.8) becomes the simpler formula

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha) w^*(\theta, Y_n)]}{\mathrm{E}_g[w^*(\theta, Y_n)]}, \quad \text{where} \quad w^*(\theta, Y_n) = \frac{p(Y_n | \theta)}{g(Y_n | \theta)}; \tag{11.11}$$

its estimate $\hat{x}$ is given by an obvious analogue of (11.10). The advantage of (11.11) relative to (11.8) is that the dimensionality of θ_t is often much smaller than that of α_t. In the important case in which y_t is univariate, θ_t is a scalar. Furthermore, once a sample is available for θ, we can deduce a sample from α using the argument in Subsection 4.13.6.

11.3 Choice of an importance density

For the computation of the estimate (11.10), we need to sample N time series for $\alpha_1, \ldots, \alpha_n$ from the importance density $g(\alpha|Y_n)$. We need to choose this density carefully. To obtain a feasible procedure, sampling from $g(\alpha|Y_n)$ should be relatively easy and computationally fast. The importance density $g(\alpha|Y_n)$ should also be sufficiently close to $p(\alpha|Y_n)$ such that the Monte Carlo variance of $\hat{x}$ is small. Hence the choice of the importance density is essential for the quality of the estimate (11.10). For example, consider the choice of $g(\alpha|Y_n)$ being equal to $p(\alpha)$. The simulation from $p(\alpha)$ is simple by all means since we can rely directly on the model specification for α_t in (11.1) or (11.2). All selected α's from this density will however have no relation with the observed vector Y_n. Almost all draws will have no support from $p(\alpha, Y_n)$ and we will obtain a poor estimate of $x(\alpha)$ with high Monte Carlo variance. A more succesful choice is to consider models for the observation density $g(Y_n|\alpha)$ and the state density $g(\alpha)$ that are both linear Gaussian; it implies that the importance density $g(\alpha|Y_n) = g(Y_n|\alpha)\, g(\alpha) \,/\, g(Y_n)$ is Gaussian also, where $g(Y_n)$ is the likelihood function that is not relevant for α. An importance density should be chosen within the class of linear Gaussian state space models that closely resembles or approximates the model density $p(\alpha, Y_n)$. From the discussion in Section 4.9, we have learned that simulating from $g(\alpha|Y_n)$ is feasible by means of a simulation smoothing algorithm.

In Sections 10.6 and 10.7 it is shown how the mode of the smoothed density $p(\alpha|Y_n)$ can be obtained. The mode is obtained by repeated linearisations of the smoothed density around a trial estimate of its mode for α. The linearisation is based on an approximating linear model and allows the use of the Kalman filter and smoother. After convergence to the mode, the approximating model can effectively be treated as the importance density $g(\alpha|Y_n)$. Given the linear model, we can apply a simulation smoother to generate importance samples from the density $g(\alpha|Y_n)$.

Alternative choices of an importance density for nonlinear non-Gaussian state space models are proposed by Danielsson and Richard (1993), Liesenfeld and Richard (2003) and Richard and Zhang (2007); they refer to their method as *efficient importance sampling*. Their importance densities are based on the minimisation of the variance of the importance weights in logs, that is the variance of $\log w(\alpha, Y_n)$ where $w(\alpha, Y_n)$ is defined in (11.8). The construction of such an importance density requires simulation and is computationally demanding. Lee and Koopman (2004) compare the performances of the simulation-based methods when different importance densities are adopted. Koopman, Lucas and Scharth (2011) show that efficient importance sampling can also be implemented by using numerical integration, simulation smoothing and control variables. They show that their *numerically accelarated importance sampling* method leads to significant computational and numerical efficiency gains when compared to other importance sampling methods. We notice that a control variable is a traditional device for improving the efficiency of estimates obtained by simulation; see also the discussion in Durbin and Koopman (2000). A related device is the antithetic variable which we discuss in Subsection 11.4.3.

11.4 Implementation details of importance sampling

11.4.1 Introduction

In this section we describe the practical implementation details of importance sampling for our general class of models. The first step is to select an appropriate importance density $g(\alpha|Y_n)$ from which it is practical to generate samples of α or, when appropriate, θ. The next step is to express the relevant formulae in terms of variables which are as simple as possible; we do this in Subsection 11.4.2. In 11.4.3 we describe antithetic variables, which increase the efficiency of the simulation by introducing a balanced structure into the simulation sample. Questions of initialisation of the approximating linear Gaussian model are considered in Subsection 11.4.4. We can take a moderate value for N when computing $\hat{x}$ in practice; typical values are $N = 100$ and $N = 250$.

11.4.2 Practical implementation of importance sampling

Up to this point we have based our exposition of the ideas underlying the use of importance sampling on α and Y_n since these are the basic vectors of interest in the state space model. However, for practical computations it is important to express formulae in terms of variables that are as simple as possible. In particular, in place of the α_t's it is usually more convenient to work with the state disturbance terms $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$. We therefore consider how to reformulate the previous results in terms of η rather than α.

By repeated substitution from the relation $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$, for $t = 1, \ldots, n$, we express $x(\alpha)$ as a function of α_1 and η; for notational convenience and because we intend to deal with the initialisation in Subsection 11.4.4, we suppress the dependence on α_1 and write $x(\alpha)$ as a function of η in the form $x^*(\eta)$. We next note that we could have written (11.3) in the form

$$\bar{x} = \mathrm{E}[x^*(\eta)|Y_n] = \int x^*(\eta)p(\eta|Y_n)\, d\eta. \tag{11.12}$$

Analogously to (11.8) we have

$$\bar{x} = \frac{\mathrm{E}_g[x^*(\eta)w^*(\eta, Y_n)]}{\mathrm{E}_g[w^*(\eta, Y_n)]}, \qquad w^*(\eta, Y_n) = \frac{p(\eta, Y_n)}{g(\eta, Y_n)}. \tag{11.13}$$

In this formula, E_g denotes expectation with respect to importance density $g(\eta|Y_n)$, which is the conditional density of η given Y_n in the approximating model, and

$$p(\eta, Y_n) = \prod_{t=1}^{n} p(\eta_t)p(y_t|\theta_t),$$

where $\theta_t = Z_t \alpha_t$. In the special case where $y_t = \theta_t + \varepsilon_t$, $p(y_t|\theta_t) = p(\varepsilon_t)$. In a similar way, for the same special case,

$$g(\eta, Y_n) = \prod_{t=1}^{n} g(\eta_t) g(\varepsilon_t).$$

For cases where the state equation is not linear and Gaussian, formula (11.13) provides the basis for the simulation estimates. When the state is linear and Gaussian, $p(\eta_t) = g(\eta_t)$ so in place of $w^*(\eta, Y_n)$ in (11.13) we take

$$w^*(\theta, Y_n) = \prod_{t=1}^{n} \frac{p(y_t|\theta_t)}{g(\varepsilon_t)}. \tag{11.14}$$

For the case $p(\eta_t) = g(\eta_t)$ and $y_t = \theta_t + \varepsilon_t$, we replace $w^*(\eta, Y_n)$ by

$$w^*(\varepsilon) = \prod_{t=1}^{n} \frac{p(\varepsilon_t)}{g(\varepsilon_t)}. \tag{11.15}$$

11.4.3 Antithetic variables

The simulations are based on random draws of η from the importance density $g(\eta|Y_n)$ using the simulation smoother as described in Section 4.9; this computes efficiently a draw of η as a linear function of rn independent standard normal deviates where r is the dimension of vector η_t and n is the number of observations. Efficiency is increased by the use of antithetic variables. An *antithetic variable* in this context is a function of a random draw of η which is equiprobable with η and which, when included together with η in the estimate of $\bar{x}$ increases the efficiency of the estimation. We assume that the importance density is Gaussian. We shall employ two types of antithetic variables. The first is the standard one given by $\check{\eta} = 2\hat{\eta} - \eta$ where $\hat{\eta} = \mathrm{E}_g(\eta|Y_n)$ is obtained from the disturbance smoother as described in Section 4.5. Since $\check{\eta} - \hat{\eta} = -(\eta - \hat{\eta})$ and η is normal, the two vectors η and $\check{\eta}$ are equi-probable. Thus we obtain two simulation samples from each draw of the simulation smoother; moreover, values of conditional means calculated from the two samples are negatively correlated, giving further efficiency gains. When this antithetic is used we say that the simulation sample is *balanced for location.*

The second antithetic variable was developed by Durbin and Koopman (1997). Let u be the vector of rn $N(0,1)$ variables that is used in the simulation smoother to generate η and let $c = u'u$; then $c \sim \chi^2_{rn}$. For a given value of c let $q = \Pr(\chi^2_{rn} < c) = F(c)$ and let $\acute{c} = F^{-1}(1-q)$. Then as c varies, c and $\acute{c}$ have the same distribution. Now take, $\acute{\eta} = \hat{\eta} + \sqrt{\acute{c}/c}(\eta - \hat{\eta})$. Then $\acute{\eta}$ has the same distribution as η. This follows because c and $(\eta - \hat{\eta})/\sqrt{c}$ are independently distributed. Finally, take $\grave{\eta} = \hat{\eta} + \sqrt{\acute{c}/c}(\check{\eta} - \hat{\eta})$. When this antithetic is used we say that the simulation sample is *balanced for scale.* By using both antithetics we obtain a set of four equiprobable values of η for each run of the simulation smoother giving a simulation sample which is balanced for location and scale.

The number of antithetics can be increased without difficulty. For example, take c and q as above. Then q is uniformly distributed on $(0, 1)$ and we write $q \sim U(0, 1)$. Let $q_1 = q + 0.5$ modulo 1; then $q_1 \sim U(0, 1)$ and we have a balanced set of four $U(0, 1)$ variables, q, q_1, $1 - q$ and $1 - q_1$. Take $\acute{c} = F^{-1}(1 - q)$ as before and similarly $c_1 = F^{-1}(q_1)$ and $\acute{c}_1 = F^{-1}(1 - q_1)$. Then each of c_1 and $\acute{c}_1$ can be combined with η and $\breve{\eta}$ as was $\acute{c}$ previously and we emerge with a balanced set of eight equiprobable values of η for each simulation. In principle this process could be extended indefinitely by taking $q_1 = q$ and $q_{j+1} = q_j + 2^{-k}$ modulo 1, for $j = 1, \ldots, 2^{k-1}$ and $k = 2, 3, \ldots$; however, two or four values of q are probably enough in practice. By using the standard normal distribution function applied to elements of u, the same idea could be used to obtain a new balanced value η_1 from η so by taking $\breve{\eta}_1 = 2\hat{\eta} - \eta_1$ we would have four values of η to combine with the four values of c. In the following we will assume that we have generated N draws of η using the simulation smoother and the antithetic variables; this means that N is a multiple of the number of different values of η obtained from a single draw of the simulation smoother. For example, when 250 simulation samples are drawn by the smoother and the one or two basic antithetics are employed, one for location and the other for scale, $N = 1000$. In practice, we have found that satisfactory results are obtained by only using the two basic antithetics.

In theory, importance sampling could give an inaccurate result on a particular occasion if in the basic formulae (11.13) very high values of $w^*(\eta, Y_n)$ are associated with very small values of the importance density $g(\eta|Y_n)$ in such a way that together they make a significant contribution to $\bar{x}$, and if also, on this particular occasion, these values happen to be over- or under-represented; for further discussion of this point see Gelman, Carlin, Stern and Rubin (1995, p. 307). In practice, we have not experienced difficulties from this source in any of the examples we have considered.

11.4.4 Diffuse initialisation

We now consider the situation where the model is non-Gaussian and some elements of the initial state vector are diffuse, the remaining elements having a known joint density; for example, they could come from stationary series. Assume that α_1 is given by (5.2) with $\eta_0 \sim p_0(\eta_0)$ where $p_0(\cdot)$ is a known density. It is legitimate to assume that δ is normally distributed as in (5.3) since we intend to let $\kappa \to \infty$. The joint density of α and Y_n is

$$p(\alpha, Y_n) = p(\eta_0) g(\delta) \prod_{t=1}^{n} p(\eta_t) p(y_t|\theta_t), \tag{11.16}$$

with $\eta_0 = R_0'(\alpha_1 - a), \delta = A'(\alpha_1 - a)$ and $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ for $t = 1, \ldots, n$, since $p(y_t|\alpha_t) = p(y_t|\theta_t)$.

As in Sections 10.6 and 10.7, we can find the mode of $p(\alpha|Y_n)$ by differentiating $\log p(\alpha, Y_n)$ with respect to $\alpha_1, \dots, \alpha_{n+1}$. For given κ the contribution from $\partial \log g(\delta)/\partial\alpha_1$ is $-A\delta/\kappa$ which $\to 0$ as $\kappa \to \infty$. Thus in the limit the mode equation is the same as (10.54) except that $\partial \log p(\alpha_1)/\partial\alpha_1$ is replaced by $\partial \log p(\eta_0)/\partial\alpha_1$. In the case that α_1 is entirely diffuse, the term $p(\eta_0)$ does not enter into (11.16), so the procedure given in Subsection 10.6.3 for finding the mode applies without change.

When $p(\eta_0)$ exists but is non-Gaussian, it is preferable to incorporate a normal approximation to it, $g(\eta_0)$ say, in the approximating Gaussian density $g(\alpha, Y_n)$, rather than include a linearised form of its derivative $\partial \log p(\eta_0)/\partial\eta_0$ within the linearisation of $\partial \log p(\alpha, Y_n)/\partial\alpha$. The reason is that we are then able to initialise the Kalman filter for the linear Gaussian approximating model by means of the standard initialisation routines developed in Chapter 5. For $g(\eta_0)$ we could take either the normal distribution with mean vector and variance matrix equal to those of $p(\eta_0)$ or with mean vector equal to the mode of $p(\eta_0)$ and variance matrix equal to $[-\partial^2 p(\eta_0)/\partial\eta_0\partial\eta_0']^{-1}$. For substitution in the basic formula (11.8) we take

$$w(\alpha, Y_n) = \frac{p(\eta_0)p(\alpha_2, \dots, \alpha_{n+1}, Y_n|\eta_0)}{g(\eta_0)g(\alpha_2, \dots, \alpha_{n+1}, Y_n|\eta_0)}, \tag{11.17}$$

since the denisties $p(\delta)$ and $g(\delta)$ are the same and therefore cancel out; thus $w(\alpha, Y_n)$ remains unchanged as $\kappa \to \infty$. The corresponding equation for (11.13) becomes simply

$$w^*(\eta, Y_n) = \frac{p(\eta_0)p(\eta_1, \dots, \eta_n, Y_n)}{g(\eta_0)g(\eta_1, \dots, \eta_n, Y_n)}. \tag{11.18}$$

While the expressions (11.17) and (11.18) are technically manageable, the practical worker may well believe in a particular situation that knowledge of $p(\eta_0)$ contributes such a small amount of information to the investigation that it can be simply ignored. In that event the factor $p(\eta_0)/g(\eta_0)$ disappears from (11.17), which amounts to treating the whole vector α_1 as diffuse, and this simplifies the analysis significantly. Expression (11.17) then reduces to

$$w(\alpha, Y_n) = \prod_{t=1}^{n} \frac{p(\alpha_t|\alpha_{t-1})p(y_t|\alpha_t)}{g(\alpha_t|\alpha_{t-1})g(y_t|\alpha_t)}.$$

Expression (11.18) reduces to

$$w^*(\eta, Y_n) = \prod_{t=1}^{n} \frac{p(\eta_t)p(y_t|\theta_t)}{g(\eta_t)g(y_t|\theta_t)},$$

with $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ for $t = 1, \dots, n$.

For nonlinear models, the initialisation of the Kalman filter is similar and the details are handled in the same way.

11.5 Estimating functions of the state vector

We will discuss the estimation of general functions of the state vector based on importance sampling for analysing data from non-Gaussian and nonlinear models. We start by showing how the method enables us to estimate mean and variance functions of the state vector using simulation and antithetic variables. We also derive estimates of the additional variances of estimates due to simulation. We use these results to obtain estimates of conditional densities and distribution functions of scalar functions of the state. Then we continue by investigating how the methods can be used for forecasting and estimating missing observations in a data set.

11.5.1 Estimating mean functions

We will consider details of the estimation of conditional means $\bar{x}$ of functions $x^*(\eta)$ of the stacked state error vector and the estimation of error variances of our estimates. Let

$$w^*(\eta) = \frac{p(\eta, Y_n)}{g(\eta, Y_n)},$$

taking the dependence of $w^*(\eta)$ on Y_n as implicit since Y_n is constant from now on. Then (11.13) gives

$$\bar{x} = \frac{\mathrm{E}_g\left[x^*(\eta) w^*(\eta)\right]}{\mathrm{E}_g\left[w^*(\eta)\right]}, \tag{11.19}$$

which is estimated by

$$\hat{x} = \frac{\sum_{i=1}^N x_i w_i}{\sum_{i=1}^N w_i}, \tag{11.20}$$

where

$$x_i = x^*\big(\eta^{(i)}\big), \qquad w_i = w^*\big(\eta^{(i)}\big) = \frac{p\big(\eta^{(i)}, Y_n\big)}{g\big(\eta^{(i)}, Y_n\big)},$$

and $\eta^{(i)}$ is the ith draw of η from the importance density $g(\eta|Y_n)$ for $i = 1, \ldots, N$.

11.5.2 Estimating variance functions

For the case where $x^*(\eta)$ is a vector we could at this point present formulae for estimating the matrix $\mathrm{Var}[x^*(\eta)|Y_n]$ and also the variance matrix due to simulation of $\hat{x} - \bar{x}$. However, from a practical point of view the covariance terms are of little interest so it seems sensible to focus on variance terms by taking $x^*(\eta)$ as a scalar for estimation of variances; extension to include covariance terms is straightforward. We estimate $\mathrm{Var}[x^*(\eta)|Y_n]$ by

$$\widehat{\mathrm{Var}}[x^*(\eta)|Y_n] = \frac{\sum_{i=1}^N x_i^2 w_i}{\sum_{i=1}^N w_i} - \hat{x}^2. \tag{11.21}$$

The estimation error due to the simulation is

$$\hat{x} - \bar{x} = \frac{\sum_{i=1}^{N} w_i(x_i - \bar{x})}{\sum_{i=1}^{N} w_i}.$$

To estimate the variance of this, consider the introduction of the antithetic variables as described in Subsection 11.4.3 and for simplicity restrict the exposition to the case of the two basic antithetics for location and scale; the extension to a larger number of antithetics is straightforward. Denote the sum of the four values of $w_i(x_i - \bar{x})$ that come from the jth run of the simulation smoother by v_j and the sum of the corresponding values of $w_i(x_i - \hat{x})$ by $\hat{v}_j$. For N large enough, since the draws from the simulation smoother are independent, the variance due to simulation is, to a good approximation,

$$\mathrm{Var}_s(\hat{x}) = \frac{1}{4N} \frac{\mathrm{Var}(v_j)}{[\mathrm{E}_g\{w^*(\eta)\}^2]}, \tag{11.22}$$

which we estimate by

$$\widehat{\mathrm{Var}}_s(\hat{x}) = \frac{\sum_{j=1}^{N/4} \hat{v}_j^2}{\left(\sum_{i=1}^{N} w_i\right)^2}. \tag{11.23}$$

The ability to estimate simulation variances so easily is an attractive feature of our methods.

11.5.3 Estimating conditional densities

When $x^*(\eta)$ is a scalar function the above technique can be used to estimate the conditional distribution function and the conditional density function of x given Y_n. Let $G[x|Y_n] = \Pr[x^*(\eta) \leq x|Y_n]$ and let $I_x(\eta)$ be an indicator which is unity if $x^*(\eta) \leq x$ and is zero if $x^*(\eta) > x$. Then $G(x|Y_n) = \mathrm{E}_g(I_x(\eta)|Y_n)$. Since $I_x(\eta)$ is a function of η we can treat it in the same way as $x^*(\eta)$. Let S_x be the sum of the values of w_i for which $x_i \leq x$, for $i = 1, \ldots, N$. Then estimate $G(x|Y_n)$ by

$$\hat{G}(x|Y_n) = \frac{S_x}{\sum_{i=1}^{N} w_i}. \tag{11.24}$$

This can be used to estimate quantiles. We order the values of x_i and we order the corresponding values w_i accordingly. The ordered sequences for x_i and w_i are denoted by $x_{[i]}$ and $w_{[i]}$, respectively. The $100k\%$ quantile is given by $x_{[m]}$ which is chosen such that

$$\frac{\sum_{i=1}^{m} w_{[i]}}{\sum_{i=1}^{N} w_{[i]}} \approx k.$$

We may interpolate between the two closest values for m in this approximation to estimate the $100k\%$ quantile. The approximation error becomes smaller as N increases.

11.5.4 Estimating conditional distribution functions

Similarly, if δ is the interval $(x - \frac{1}{2}d, x + \frac{1}{2}d)$ where d is suitably small and positive, let S^{δ} be the sum of the values of w_i for which $x^*(\eta) \in \delta$. Then the estimate of the conditional density $p(x|Y_n)$ of x given Y_n is

$$\hat{p}(x|Y_n) = d^{-1}\frac{S^{\delta}}{\sum_{i=1}^{N} w_i}. \tag{11.25}$$

This estimate can be used to construct a histogram.

We now show how to generate a sample of M independent values from the estimated conditional distribution of $x^*(\eta)$ using importance resampling; for further details of the method see Gelfand and Smith (1999) and Gelman, Carlin, Stern and Rubin (1995). Take $x^{[k]} = x_j$ with probability $w_j / \sum_{i=1}^{N} w_i$ for $j = 1, \ldots, N$. Then

$$\Pr\big(x^{[k]} \leq x\big) = \frac{\sum_{x_j \leq x} w_j}{\sum_{i=1}^{N} w_i} = \hat{G}(x|Y_n).$$

Thus $x^{[k]}$ is a random draw from the distribution function given by (11.24). Doing this M times with replacement gives a sample of $M \leq N$ independent draws. The sampling can also be done without replacement but the values are not then independent.

11.5.5 Forecasting and estimating with missing observations

The treatment of missing observations and forecasting by the methods of this chapter is straightforward. For missing observations, our objective is to estimate $\bar{x} = \int x^*(\eta)p(\eta|Y_n)\, d\eta$ where the stacked vector Y_n contains only those observational elements actually observed. We achieve this by omitting from the linear Gaussian approximating model the observational components that correspond to the missing elements in the original model. Only the Kalman filter and smoother algorithms are needed in the determination of the approximating model and we described in Section 4.10 how the filter is modified when observational vectors or elements are missing. For the simulation, the simulation smoother of Section 4.9 must be similarly modified to allow for the missing elements.

For forecasting, our objective is to estimate $\bar{y}_{n+j} = \mathrm{E}(y_{n+j}|Y_n)$, $j = 1, \ldots, J$, where we assume that $y_{n+1}, \ldots y_{n+J}$ and $\alpha_{n+2}, \ldots, \alpha_{n+J}$ have been generated by model (9.3) and (9.4), noting that α_{n+1} has already been generated by (9.4) with $t = n$. It follows from (9.3) that

$$\bar{y}_{n+j} = \mathrm{E}[\mathrm{E}(y_{n+j}|\theta_{n+j})|Y_n], \tag{11.26}$$

for $j = 1, \ldots, J$, where $\theta_{n+j} = Z_{n+j}\alpha_{n+j}$, with $Z_{n+1}, \ldots, Z_{n+J}$ assumed known. We estimate this as in Section 11.5 with $x^*(\eta) = \mathrm{E}(y_{n+j}|\theta_{n+j})$, extending the simulation smoother for $t = n + 1, \ldots, n + J$.

For exponential families,

$$\mathrm{E}(y_{n+j}|\theta_{n+j}) = \dot{b}_{n+j}(\theta_{n+j}),$$

as in Section 9.3 for $t \leq n$, so we take $x^*(\eta) = \dot{b}_{n+j}(\theta_{n+j})$, for $j = 1, \ldots, J$. For the model $y_t = \theta_t + \varepsilon_t$ in (9.7) we take $x^*(\eta) = \theta_t$.

11.6 Estimating loglikelihood and parameters

In this section we consider the estimation of the parameter vector ψ by maximum likelihood. Since analytical methods are not feasible we employ techniques based on simulation using importance sampling. We shall find that the techniques we develop are closely related to those we employed earlier in this chapter for estimation of the mean of $x(\alpha)$ given Y_n. Monte Carlo estimation of ψ by maximum likelihood using importance sampling was considered briefly by Shephard and Pitt (1997) and in more detail by Durbin and Koopman (1997) for the special case where $p(y_t|\theta_t)$ is non-Gaussian but α_t is generated by a linear Gaussian model. In this section we will begin by considering first the general case where both $p(y_t|\theta_t)$ and the state error density $p(\eta_t)$ in (9.4) are non-Gaussian and will specialise later to the simpler case where $p(\eta_t)$ is Gaussian. We will also consider the case where the state space models are nonlinear. Our approach will be to estimate the loglikelihood by simulation and then to estimate ψ by maximising the resulting value numerically.

11.6.1 Estimation of likelihood

The likelihood $L(\psi)$ is defined by $L(\psi) = p(Y_n|\psi)$, where for convenience we suppress the dependence of $L(\psi)$ on Y_n, so we have

$$L(\psi) = \int p(\alpha, Y_n)\, d\alpha.$$

Dividing and multiplying by the importance density $g(\alpha|Y_n)$ as in Section 11.2 gives

$$\begin{aligned} L(\psi) &= \int \frac{p(\alpha, Y_n)}{g(\alpha|Y_n)} g(\alpha|Y_n)\, d\alpha \\ &= g(Y_n) \int \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} g(\alpha|Y_n)\, d\alpha \\ &= L_g(\psi)\, \mathrm{E}_g\left[w(\alpha, Y_n)\right], \end{aligned} \tag{11.27}$$

where $L_g(\psi) = g(Y_n)$ is the likelihood of the approximating linear Gaussian model that we employ to obtain the importance density $g(\alpha|Y_n)$, E_g denotes expectation with respect to density $g(\alpha|Y_n)$, and $w(\alpha, Y_n) = p(\alpha, Y_n)/g(\alpha, Y_n)$

as in (11.8). Indeed we observe that (11.27) is essentially equivalent to (11.6). We note the elegant feature of (11.27) that the non-Gaussian likelihood $L(\psi)$ has been obtained as an adjustment to the linear Gaussian likelihood $L_g(\psi)$, which is easily calculated by the Kalman filter; moreover, the adjustment factor $\mathrm{E}_g[w(\alpha, Y_n)]$ is readily estimable by simulation. Obviously, the closer the importance joint density $g(\alpha, Y_n)$ is to the non-Gaussian density $p(\alpha, Y_n)$, the smaller will be the simulation sample required.

For practical computations we follow the practice discussed in Subsection 11.4.2 and Section 11.5 of working with the signal $\theta_t = Z_t\alpha_t$ in the observation equation and the state disturbance η_t in the state equation, rather than with α_t directly, since these lead to simpler computational procedures. In place of (11.27) we therefore use the form

$$L(\psi) = L_g(\psi)\,\mathrm{E}_g[w^*(\eta, Y_n)], \tag{11.28}$$

where $L(\psi)$ and $L_g(\psi)$ are the same as in (11.27) but E_g and $w^*(\eta, Y_n)$ have the interpretations discussed in Subsection 11.4.2. We then suppress the dependence on Y_n and write $w^*(\eta)$ in place of $w^*(\eta, Y_n)$ as in Section 11.5. We employ antithetic variables as in Subsection 11.4.3, and analogously to (11.20) our estimate of $L(\psi)$ is

$$\hat{L}(\psi) = L_g(\psi)\bar{w}, \tag{11.29}$$

where $\bar{w} = (1/N)\sum_{i=1}^{N} w_i$, with $w_i = w^*(\eta^{(i)})$ where $\eta^{(1)}, \dots \eta^{(N)}$ is the simulation sample generated by the importance density $g(\eta|Y_n)$.

11.6.2 Maximisation of loglikelihood

We estimate ψ by the value $\hat{\psi}$ of ψ that maximises $\hat{L}(\psi)$. In practice, it is numerically more stable to maximise

$$\log \hat{L}(\psi) = \log L_g(\psi) + \log \bar{w}, \tag{11.30}$$

rather than to maximise $\hat{L}(\psi)$ directly because the likelihood value can become very large. Moreover, the value of ψ that maximises $\log \hat{L}(\psi)$ is the same as the value that maximises $\hat{L}(\psi)$.

To calculate $\hat{\psi}$, $\log \hat{L}(\psi)$ is maximised by any convenient iterative numerical optimisation technique, as discussed, for example in Subsection 7.3.2. To ensure stability of the iterative process, it is important to use the same random numbers from the simulation smoother for each value of ψ. To start the iteration, an initial value of ψ can be obtained by maximising the approximate loglikelihood

$$\log L(\psi) \approx \log L_g(\psi) + \log w(\hat{\eta}),$$

where $\hat{\eta}$ is the mode of $g(\eta|Y_n)$ that is determined during the process of approximating $p(\eta|Y_n)$ by $g(\eta|Y_n)$; alternatively, the more accurate non-simulated approximation given in expression (21) of Durbin and Koopman (1997) may be used.

11.6.3 Variance matrix of maximum likelihood estimate

Assuming that appropriate regularity conditions are satisfied, the estimate of the large-sample variance matrix of $\hat{\psi}$ is given by the standard formula

$$\hat{\Omega} = \left[-\frac{\partial^2 \log L(\psi)}{\partial \psi \partial \psi'} \right]^{-1} \Bigg|_{\psi = \hat{\psi}}, \tag{11.31}$$

where the derivatives of $\log L(\psi)$ are calculated numerically from values of ψ in the neighbourhood of $\hat{\psi}$.

11.6.4 Effect of errors in parameter estimation

In the above treatment we have performed classical analyses in the traditional way by first assuming that the parameter vector is known and then substituting the maximum likelihood estimate $\hat{\psi}$ for ψ. The errors $\hat{\psi} - \psi$ give rise to biases in the estimates of functions of the state and disturbance vectors, but since the biases are of order n^{-1} they are usually small enough to be neglected. It may, however, be important to investigate the amount of bias in particular cases. In Subsection 7.3.7 we described techniques for estimating the bias for the case where the state space model is linear and Gaussian. Exactly the same procedure can be used for estimating the bias due to errors $\hat{\psi} - \psi$ for the non-Gaussian and nonlinear models considered in this chapter.

11.6.5 Mean square error matrix due to simulation

We have denoted the estimate of ψ that is obtained from the simulation by $\hat{\psi}$; let us denote by $\tilde{\psi}$ the 'true' maximum likelihood estimate of ψ that would be obtained by maximising the exact $\log L(\psi)$ without simulation, if this could be done. The error due to simulation is $\hat{\psi} - \tilde{\psi}$, so the mean square error matrix is

$$\text{MSE}(\hat{\psi}) = \text{E}_g[(\hat{\psi} - \tilde{\psi})(\hat{\psi} - \tilde{\psi})'].$$

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

$$\frac{\partial \log \hat{L}(\psi)}{\partial \psi} = 0,$$

which on expansion about $\tilde{\psi}$ gives approximately,

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} + \frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial \psi \partial \psi'} (\hat{\psi} - \tilde{\psi}) = 0,$$

where

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{\partial \log \hat{L}(\psi)}{\partial \psi} \Bigg|_{\psi = \tilde{\psi}}, \qquad \frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial \psi \partial \psi'} = \frac{\partial^2 \log \hat{L}(\psi)}{\partial \psi \partial \psi'} \Bigg|_{\psi = \tilde{\psi}},$$

giving

$$\hat{\psi} - \tilde{\psi} = \left[-\frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial \psi \partial \psi'} \right]^{-1} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi}.$$

Thus to a first approximation we have

$$\mathrm{MSE}(\hat{\psi}) = \hat{\Omega}\, \mathrm{E}_g \left[\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi'} \right] \hat{\Omega}, \tag{11.32}$$

where $\hat{\Omega}$ is given by (11.31).

From (11.30) we have

$$\log \hat{L}(\tilde{\psi}) = \log L_g(\tilde{\psi}) + \log \bar{w},$$

so

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} + \frac{1}{\bar{w}} \frac{\partial \bar{w}}{\partial \psi}.$$

Similarly, for the true loglikelihood $\log L(\tilde{\psi})$ we have

$$\frac{\partial \log L(\tilde{\psi})}{\partial \psi} = \frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} + \frac{\partial \log \mu_w}{\partial \psi},$$

where $\mu_w = \mathrm{E}_g(\bar{w})$. Since $\tilde{\psi}$ is the 'true' maximum likelihood estimator of ψ,

$$\frac{\partial \log L(\tilde{\psi})}{\partial \psi} = 0.$$

Thus

$$\frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} = -\frac{\partial \log \mu_w}{\partial \psi} = -\frac{1}{\mu_w} \frac{\partial \mu_w}{\partial \psi},$$

so we have

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{1}{\bar{w}} \frac{\partial \bar{w}}{\partial \psi} - \frac{1}{\mu_w} \frac{\partial \mu_w}{\partial \psi}.$$

It follows that, to a first approximation,

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{1}{\bar{w}} \frac{\partial}{\partial \psi}(\bar{w} - \mu_w),$$

and hence

$$\mathrm{E}_g \left[\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi'} \right] = \frac{1}{\bar{w}^2} \mathrm{Var} \left(\frac{\partial \bar{w}}{\partial \psi} \right).$$

Taking the case of two antithetics, denote the sum of the four values of w obtained from each draw of the simulation smoother by w_j^* for $j = 1, \ldots, N/4$. Then $\bar{w} = N^{-1} \sum_{j=1}^{N/4} w_j^*$, so

$$\text{Var}\left(\frac{\partial \bar{w}}{\partial \psi}\right) = \frac{4}{N} \text{Var}\left(\frac{\partial w_j^*}{\partial \psi}\right).$$

Let $q^{(j)} = \partial w_j^* / \partial \psi$, which we calculate numerically at $\psi = \hat{\psi}$, and let $\bar{q} = (4/N) \sum_{j=1}^{N/4} q^{(j)}$. Then estimate (11.32) by

$$\widehat{\text{MSE}}(\hat{\psi}) = \hat{\Omega} \left[\left(\frac{4}{N\bar{w}}\right)^2 \sum_{j=1}^{N/4} \left(q^{(j)} - \bar{q}\right)\left(q^{(j)} - \bar{q}\right)' \right] \hat{\Omega}. \qquad (11.33)$$

The square roots of the diagonal elements of (11.33) may be compared with the square roots of the diagonal elements of $\hat{\Omega}$ in (11.31) to obtain relative standard errors due to simulation.

11.7 Importance sampling weights and diagnostics

In this section we briefly draw attention to ways to validate the effectiveness of the importance sampling method. The importance weight function $w(\alpha, Y_n)$ as defined in (11.8) is instrumental for this purpose. Geweke (1989) argued that importance sampling should only be used in settings where the variance of the importance weights is known to exist. Failure of this condition can lead to slow and unstable convergence of the estimator as the central limit theorem governing convergence fails to hold. Robert and Casella (2010, §4.3) provide examples of importance samplers that fail this condition and show that ignoring the problem can result in strongly biased estimators. While the variance conditions can be checked analytically in low dimensional problems, proving that they are met in high dimensional cases such as time series can be challenging.

Monahan (1993, 2001) and Koopman, Shephard and Creal (2009) have developed diagnostic procedures to check for the existence of the variance of the importance weights. It is based on the application of extreme value theory. Limit results from extreme value theory imply that we can learn about the variance of the importance weights by studying the behaviour of their distribution in the right hand tail. Test statistics are then formulated to test whether the tail of the distribution allows for a properly defined variance. If the characteristics of the tails are not sufficient, we reject the hypothesis that the variance of the importance weights exists. A set of graphical diagnostics can be deducted from the hypothesis and a complete insight can be obtained.