7 Maximum likelihood estimation of parameters

7.1 Introduction

So far we have developed methods for estimating parameters which can be placed in the state vector of model (4.12). In virtually all applications in practical work the models depend on additional parameters which have to be estimated from the data; for example, in the local level model (2.3) the variances σ_ε^2 and σ_η^2 are unknown and need to be estimated. In classical analyses, these additional parameters are assumed to be fixed but unknown whereas in Bayesian analyses they are assumed to be random variables. Because of the differences in assumptions the treatment of the two cases is not the same. In this chapter we deal with classical analyses in which the additional parameters are fixed and are estimated by maximum likelihood. The Bayesian treatment for these parameters is discussed as part of a general Bayesian discussion of state space models in Chapter 13 of Part II.

For the linear Gaussian model we shall show that the likelihood can be calculated by a routine application of the Kalman filter, even when the initial state vector is fully or partially diffuse. We also give the details of the computation of the likelihood when the univariate treatment of multivariate observations is adopted as suggested in Section 6.4. We go on to consider how the loglikelihood can be maximised by means of iterative numerical procedures. An important part in this process is played by the score vector and we show how this is calculated, both for the case where the initial state vector has a known distribution and for the diffuse case. A useful device for maximisation of the loglikelihood in some cases, particularly in the early stages of maximisation, is the EM algorithm; we give details of this for the linear Gaussian model. We go on to consider biases in estimates due to errors in parameter estimation. The chapter ends with a discussion of some questions of goodness-of-fit and diagnostic checks.

7.2 Likelihood evaluation

7.2.1 Loglikelihood when initial conditions are known

We first assume that the initial state vector has density $\mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. The likelihood is

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

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

where $Y_t = (y_1', \dots, y_t')'$. In practice we generally work with the loglikelihood

$$\log L(Y_n) = \sum_{t=1}^{n} \log p(y_t|Y_{t-1}), \tag{7.1}$$

where $p(y_1|Y_0) = p(y_1)$. For model (3.1), $\mathrm{E}(y_t|Y_{t-1}) = Z_t a_t$. Putting $v_t = y_t - Z_t a_t$, $F_t = \mathrm{Var}(y_t|Y_{t-1})$ and substituting $\mathrm{N}(Z_t a_t, F_t)$ for $p(y_t|Y_{t-1})$ in (7.1), we obtain

$$\log L(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} \left(\log|F_t| + v_t' F_t^{-1} v_t \right). \tag{7.2}$$

The quantities v_t and F_t are calculated routinely by the Kalman filter (4.24) so $\log L(Y_n)$ is easily computed from the Kalman filter output. We assume that F_t is nonsingular for $t = 1, \dots, n$. If this condition is not satisfied initially it is usually possible to redefine the model so that it is satisfied. The representation (7.2) of the loglikelihood was first given by Schweppe (1965). Harvey (1989, §3.4) refers to it as the *prediction error decomposition*.

7.2.2 Diffuse loglikelihood

We now consider the case where some elements of α_1 are diffuse. As in Section 5.1, we assume that $\alpha_1 = a + A\delta + R_0\eta_0$ where a is a known constant vector, $\delta \sim \mathrm{N}(0, \kappa I_q)$, $\eta_0 \sim \mathrm{N}(0, Q_0)$ and $A'R_0 = 0$, giving $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where $P_1 = \kappa P_\infty + P_*$ and $\kappa \to \infty$. From (5.6) and (5.7),

$$F_t = \kappa F_{\infty,t} + F_{*,t} + O(\kappa^{-1}) \quad \text{with} \quad F_{\infty,t} = Z_t P_{\infty,t} Z_t', \tag{7.3}$$

where by definition of d, $P_{\infty,t} \neq 0$ for $t = 1, \dots, d$. The number of diffuse elements in α_1 is q which is the dimensionality of vector δ. Thus the loglikelihood (7.2) will contain a term $-\frac{1}{2} q \log 2\pi\kappa$ so $\log L(Y_n)$ will not converge as $\kappa \to \infty$. Following de Jong (1991), we therefore define the *diffuse loglikelihood* as

$$\log L_d(Y_n) = \lim_{\kappa \to \infty} \left[\log L(Y_n) + \frac{q}{2} \log \kappa \right]$$

and we work with $\log L_d(Y_n)$ in place of $\log L(Y_n)$ for estimation of unknown parameters in the diffuse case. Similar definitions for the diffuse loglikelihood function have been adopted by Harvey and Phillips (1979) and Ansley and Kohn (1986). As in Section 5.2, and for the same reasons, we assume that $F_{\infty,t}$ is positive definite or is a zero matrix. We also assume that q is a multiple of p. This covers the important special case of univariate series and is generally satisfied in practice for multivariate series; if not, the series can be dealt with as if it were univariate as in Section 6.4.

Suppose first that $F_{\infty,t}$ is positive definite and therefore has rank p. From (7.3) we have for $t = 1, \dots, d$,

$$F_t^{-1} = \kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2}).$$

It follows that

$$\begin{aligned} -\log|F_t| &= \log\left|F_t^{-1}\right| = \log\left|\kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2})\right| \\ &= -p \log \kappa + \log\left|F_{\infty,t}^{-1} + O(\kappa^{-1})\right|, \end{aligned}$$

and

$$\lim_{\kappa\to\infty} (-\log|F_t| + p \log \kappa) = \log\left|F_{\infty,t}^{-1}\right| = -\log|F_{\infty,t}|.$$

Moreover,

$$\begin{aligned} \lim_{\kappa\to\infty} v_t' F_t^{-1} v_t &= \lim_{\kappa\to\infty} \left[v_t^{(0)} + \kappa^{-1} v_t^{(1)} + O(\kappa^{-2})\right]' \left[\kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2})\right] \\ &\quad \times \left[v_t^{(0)} + \kappa^{-1} v_t^{(1)} + O(\kappa^{-2})\right] \\ &= 0 \end{aligned}$$

for $t = 1, \dots, d$, where $v_t^{(0)}$ and $v_t^{(1)}$ are defined in Subsection 5.2.1.

When $F_{\infty,t} = 0$, it follows from Subsection 5.2.1 that $F_t = F_{*,t} + O(\kappa^{-1})$ and $F_t^{-1} = F_{*,t}^{-1} + O(\kappa^{-1})$. Consequently,

$$\lim_{\kappa\to\infty} (-\log|F_t|) = -\log|F_{*,t}| \quad \text{and} \quad \lim_{\kappa\to\infty} v_t' F_t^{-1} v_t = v_t^{(0)\prime} F_{*,t}^{-1} v_t^{(0)}.$$

Putting these results together, we obtain the diffuse loglikelihood as

$$\log L_d(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{d} w_t - \frac{1}{2} \sum_{t=d+1}^{n} \left(\log|F_t| + v_t' F_t^{-1} v_t\right), \quad (7.4)$$

where

$$w_t = \begin{cases} \log|F_{\infty,t}|, & \text{if } F_{\infty,t} \text{ is positive definite,} \\ \log|F_{*,t}| + v_t^{(0)\prime} F_{*,t}^{-1} v_t^{(0)}, & \text{if } F_{\infty,t} = 0, \end{cases}$$

for $t = 1, \dots, d$. The expression (7.4) for the diffuse loglikelihood is given by Koopman (1997).

7.2.3 Diffuse loglikelihood via augmented Kalman filter

In the notation of Subsection 5.7.3, the joint density of δ and Y_n for given κ is

$$\begin{aligned} p(\delta, Y_n) &= p(\delta)p(Y_n|\delta) \\ &= p(\delta)\sum_{t=1}^{n} p(v_{\delta,t}) \\ &= (2\pi)^{-(np+q)/2}\kappa^{-q/2}\prod_{t=1}^{n}|F_{\delta,t}|^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}\left(\frac{\delta'\delta}{\kappa} + S_{a,n} + 2b_n'\delta + \delta' S_{A,n}\delta\right)\right], \end{aligned} \tag{7.5}$$

where $v_{\delta,t}$ is defined in (5.35), b_n and $S_{A,n}$ are defined in (5.40) and $S_{a,n} = \sum_{t=1}^{n} v_{a,t}' F_{\delta,t}^{-1} v_{a,t}$. From (5.41) we have $\bar{\delta}_n = \mathrm{E}(\delta|Y_n) = -(S_{A,n} + \kappa^{-1}I_q)^{-1}b_n$. The exponent of (7.5) can now be rewritten as

$$-\frac{1}{2}[S_{a,n} + (\delta - \bar{\delta}_n)'(S_{A,n} + \kappa^{-1}I_q)(\delta - \bar{\delta}_n) - \bar{\delta}_n'(S_{A,n} + \kappa^{-1}I_q)\bar{\delta}_n],$$

as is easily verified. Integrating out δ from $p(\delta, Y_n)$ we obtain the marginal density of Y_n. After taking logs, the loglikelihood appears as

$$\begin{aligned} \log L(Y_n) = &-\frac{np}{2}\log 2\pi - \frac{q}{2}\log\kappa - \frac{1}{2}\log|S_{A,n} + \kappa^{-1}I_q| \\ &- \frac{1}{2}\sum_{t=1}^{n}\log|F_{\delta,t}| - \frac{1}{2}[S_{a,n} - \bar{\delta}_n'(S_{A,n} + \kappa^{-1}I_q)\bar{\delta}_n]. \end{aligned} \tag{7.6}$$

Adding $\frac{q}{2}\log\kappa$ and letting $\kappa \to \infty$ we obtain the diffuse loglikelihood

$$\begin{aligned} \log L_d(Y_n) = &-\frac{np}{2}\log 2\pi - \frac{1}{2}\log|S_{A,n}| - \frac{1}{2}\sum_{t=1}^{n}\log|F_{\delta,t}| \\ &- \frac{1}{2}\left(S_{a,n} - b_n' S_{A,n}^{-1} b_n\right), \end{aligned} \tag{7.7}$$

which is due to de Jong (1991). In spite of its very different structure (7.7) necessarily has the same numerical value as (7.4).

It is shown in Subsection 5.7.3 that the augmented Kalman filter can be collapsed at time point $t = d$. We could therefore form a partial likelihood based on Y_d for fixed κ, integrate out δ and let $\kappa \to \infty$ as in (7.7). Subsequently we could add the contribution from innovations $v_{d+1}, \ldots, v_n$ obtained from the collapsed Kalman filter. However, we will not give detailed formulae here.

These results were originally derived by de Jong (1988b, 1999). The calculations required to compute (7.7) are more complicated than those required to compute (7.4). This is another reason why we ourselves prefer the initialisation technique of Section 5.2 to the augmentation device of Section 5.7. A further reason for prefering our computation of (7.4) is given in Subsection 7.3.5.

7.2.4 Likelihood when elements of initial state vector are fixed but unknown

Now let us consider the case where δ is treated as fixed. The density of Y_n given δ is, as in the previous section,

$$p(Y_n|\delta) = (2\pi)^{-np/2}\prod_{t=1}^{n}|F_{\delta,t}|^{-1/2}\exp\left[-\frac{1}{2}(S_{a,n}+2b_n'\delta_n+\delta_n'S_{A,n}\delta_n)\right]. \qquad (7.8)$$

The usual way to remove the influence of an unknown parameter vector such as δ from the likelihood is to estimate it by its maximum likelihood estimate, $\hat{\delta}_n$ in this case, and to employ the concentrated loglikelihood $\log L_c(Y_n)$ obtained by substituting $\hat{\delta}_n = -S_{A,n}^{-1}b_n$ for δ in $p(Y_n|\delta)$. This gives

$$\log L_c(Y_n) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}|F_{\delta,t}| - \frac{1}{2}(S_{a,n} - b_n'S_{A,n}^{-1}b_n). \qquad (7.9)$$

Comparing (7.7) and (7.9) we see that the only difference between them is the presence in (7.7) of the term $-\frac{1}{2}\log|S_{A,n}|$. The relation between (7.7) and (7.9) was demonstrated by de Jong (1988b) using a different argument. A modification of the augmented Kalman filter and the corresponding diffuse loglikelihood function is proposed by Francke, Koopman and de Vos (2010). Their modification ensures that the loglikelihood function has the same value regardless of the way the regression effects are treated in the model.

Harvey and Shephard (1990) argue that parameter estimation should preferably be based on the loglikelihood function (7.4) for which the initial vector δ is treated as diffuse not fixed. They have shown for the local level model of Chapter 2 that maximising (7.9) with respect to the signal to noise ratio q leads to a much higher probability of estimating q to be zero compared to maximising (7.4). This is undesirable from a forecasting point of view since this results in no discounting of past observations.

7.2.5 Likelihood when a univariate treatment of multivariate series is employed

When the univariate treatment of the Kalman filter for multivariate time series is considered as in Section 6.4, we transform the vector time series $y_1, \ldots, y_n$ into the univariate time series

$$y_{1,1}, \ldots, y_{1,p_1}, y_{2,1}, \ldots, y_{n,p_n}.$$

The univariate Kalman filter (6.12) and (6.13) can then be applied to the model equations (6.10) and (6.11) for this univariate series. It produces the prediction error $v_{t,i}$ and its scalar variance $F_{t,i}$ for $t = 1, \ldots, n$ and $i = 1, \ldots, p_t$, where the prediction error is the error in predicting $y_{t,i}$ as a function of the 'past' observations $y_{1,1}, \ldots, y_{1,p_1}, y_{2,1}, \ldots, y_{t,i-1}$ for $i = 2, \ldots p_t$ and $y_{1,1}, \ldots, y_{1,p_1}, y_{2,1}, \ldots, y_{t-1,p_{t-1}}$ for $i = 1$. Since the model (6.10) and (6.11) for the univariate series is fully consistent with the original model for the vector time series $y_1, \ldots, y_n$ and the Kalman filter is adapted correctly, the loglikelihood function when initial conditions are known is given by

$$\log L(Y_n) = -\frac{1}{2}\sum_{t=1}^{n}\left[p_t^* \log 2\pi + \sum_{i=1}^{p_t} \iota_{t,i}(\log F_{t,i} + v_{t,i}^2 \,/\, F_{t,i})\right].$$

where $\iota_{t,i}$ equals zero if $F_{t,i} = 0$ and unity otherwise, and $p_t^* = \sum_{i=1}^{p_t} \iota_{t,i}$. The occurence of $F_{t,i} = 0$ is due to the singularity of F_t as discussed in Section 6.4.

The variables associated with the exact initial Kalman filter of Section 5.2 can also redefined when it is applied to the transformation of a multivariate series into a univariate series. When the univariate treatment is used, the relevant variables for the diffuse loglikelihood function considered in Subsection 7.2.2 are given by the scalars $v_{t,i}^{(0)}$, $F_{\infty,t,i}$ and $F_{*,t,i}$ for $t = 1, \ldots, d$, and $v_{t,i}$ and $F_{t,i}$ for $t = d+1, \ldots, n$, with $i = 1, \ldots, p_t$. The diffuse loglikelihood function (7.4) is then computed by

$$\begin{aligned}\log L_d(Y_n) = &-\frac{1}{2}\sum_{t=1}^{n}\sum_{i=1}^{p_t} \iota_{t,i} \log 2\pi - \frac{1}{2}\sum_{t=1}^{d}\sum_{i=1}^{p_t} w_{t,i} \\ &-\frac{1}{2}\sum_{t=d+1}^{n}\sum_{i=1}^{p_t} \iota_{t,i}(\log F_{t,i} + v_{t,i}^2 \,/\, F_{t,i}),\end{aligned}$$

where $\iota_{t,i}$ equal zero if $F_{*,t,i} = 0$ and unity otherwise, and where

$$w_{t,i} = \begin{cases} \log F_{\infty,t,i}, & \text{if} \quad F_{\infty,t,i} > 0, \\ \iota_{t,i}(\log F_{*,t,i} + v_t^{(0)\,2} \,/\, F_{*,t,i}), & \text{if} \quad F_{\infty,t,i} = 0, \end{cases}$$

for $t = 1, \ldots, d$ and $i = 1, \ldots, p_t$.

7.2.6 Likelihood when the model contains regression effects

When regression effects are present in the state space model, similar adjustments as for the diffuse loglikelihood function are required. The diffuse loglikelihood can be regarded as a loglikelihood function of a linear transformation of the original time series. When such a loglikelihood function is used for parameter estimation, the transformation must not depend on the parameter vector itself. When it does, the likelihood function is not a smooth function with respect to the parameter vector. Francke, Koopman and de Vos (2010) show that a proper transformation can be obtained via a simple modification of the augmented Kalman filter. They provide the details with a comparison of the different likelihood functions and some illustrations.

7.2.7 Likelihood when large observation vector is collapsed

When a large $p \times 1$ observation vector y_t needs to be treated in the analysis, we proposed in Section 6.5 to collapse y_t into a transformed observation vector with its dimension equal to the dimension of the state vector m. It is shown in Section 6.5 that the collapse strategy becomes computationally beneficial when $p >> m$. The Kalman filter can be applied to a small observation vector without additional costs. We show next that the likelihood function for the original model can be computed using the Kalman filter for a small observation vector together with some additional computing. Consider y_t^* and y_t^+ as defined in Subsection 6.5.2, or their counterparts $\bar{y}_t^*$ and $\bar{y}_t^+$ as defined in Subsection 6.5.3, for which the models

$$y_t^* = \alpha_t + \varepsilon_t^*, \qquad y_t^+ = \varepsilon_t^+,$$

are adopted where $\varepsilon_t^* \sim \mathrm{N}(0, H_t^*)$ and $\varepsilon_t^+ \sim \mathrm{N}(0, H_t^+)$ are serially and mutually independent. It follows that the original loglikelihood function is subject to the relation

$$\log L(Y_n) = \log L(Y_n^*) + \log L(Y_n^+) + \sum_{t=1}^{n} \log |A_t|,$$

where $Y_n^* = (y_1^{*\prime}, \ldots, y_n^{*\prime})'$, $Y_n^+ = (y_1^{+\prime}, \ldots, y_n^{+\prime})'$ and $A_t = (A_t^{*\prime}, A_t^{+\prime})'$ such that $A_t y_t = (y_t^{*\prime}, y_t^{+\prime})'$ for $t = 1, \ldots, n$. The term $|A_t|$ can be expressed more conveniently via the relations

$$|A_t|^2 = |A_t A_t'| = |H_t^{-1}||A_t H_t A_t'| = |H_t|^{-1}|H_t^*||H_t^+|.$$

Since the scaling of matrix A_t^+ is not relevant for its construction in Subsections 6.5.2 and 6.5.3, we choose A_t^+ such that $|H_t^+| = 1$. We then have

$$\log |A_t| = \frac{1}{2} \log \frac{|H_t^*|}{|H_t|}, \qquad t = 1, \ldots, n,$$

which can be computed efficiently since matrix H_t^* has a small dimension while H_t is the variance matrix of the original observation disturbance ε_t and is typically a diagonal matrix or has a convenient structure. The loglikelihood function $\log L(Y_n^*)$ is computed by the Kalman filter applied to the model $y_t^* = \alpha_t + \varepsilon_t^*$ and $\log L(Y_n^+)$ is given by

$$\log L(Y_n^+) = -\frac{(p-m)n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} y_t^{+\prime} (H_t^+)^{-1} y_t^+,$$

since we have $|H_t^+| = 1$. The term $y_t^{+\prime}(H_t^+)^{-1} y_t^+$ insists that matrix A_t^+ actually needs to be computed. However, it is shown in Jungbacker and Koopman (2008, Lemma 2) that this term can be computed as

$$y_t^{+\prime}(H_t^+)^{-1} y_t^+ = e_t' H_t^{-1} e_t,$$

where $e_t = y_t - Z_t y_t^*$ in or, more generally, $e_t = y_t - Z_t \bar{y}_t^*$ for any nonsingular matrix C_t as defined in Subsection 6.5.3. The computation of $e_t' H_t^{-1} e_t$ is simple and does not require the construction of matrix A_t^+ or $\bar{A}_t^+$.

7.3 Parameter estimation

7.3.1 Introduction

So far in this book we have assumed that the system matrices Z_t, H_t, T_t, R_t and Q_t in model (3.1) are known for $t = 1, \ldots, n$. We now consider the more usual situation in which at least some of the elements of these matrices depend on a vector ψ of unknown parameters. We shall estimate ψ by maximum likelihood. To make explicit the dependence of the loglikelihood on ψ we write $\log L(Y_n|\psi)$, $\log L_d(Y_n|\psi)$ and $\log L_c(Y_n|\psi)$. In the diffuse case we shall take it for granted that for models of interest, estimates of ψ obtained by maximising $\log L(Y_n|\psi)$ for fixed κ converge to the estimates obtained by maximising the diffuse loglikelihood $\log L_d(Y_n|\psi)$ as $\kappa \to \infty$.

7.3.2 Numerical maximisation algorithms

A wide range of numerical search algorithms are available for maximising the loglikelihood with respect to unknown parameters. Many of these are based on Newton's method which solves the equation

$$\partial_1(\psi) = \frac{\partial \log L(Y_n|\psi)}{\partial \psi} = 0, \tag{7.10}$$

using the first-order Taylor series

$$\partial_1(\psi) \simeq \tilde{\partial}_1(\psi) + \tilde{\partial}_2(\psi)(\psi - \tilde{\psi}), \tag{7.11}$$

for some trial value $\tilde{\psi}$, where

$$\tilde{\partial}_1(\psi) = \partial_1(\psi)|_{\psi=\tilde{\psi}}, \qquad \tilde{\partial}_2(\psi) = \partial_2(\psi)|_{\psi=\tilde{\psi}},$$

with

$$\partial_2(\psi) = \frac{\partial^2 \log L(Y_n|\psi)}{\partial \psi \partial \psi'}. \tag{7.12}$$

By equating (7.11) to zero we obtain a revised value $\bar{\psi}$ from the expression

$$\bar{\psi} = \tilde{\psi} - \tilde{\partial}_2(\psi)^{-1} \tilde{\partial}_1(\psi).$$

This process is repeated until it converges or until a switch is made to another optimisation method. If the Hessian matrix $\partial_2(\psi)$ is negative definite for all ψ the loglikelihood is said to be concave and a unique maximum exists for the likelihood. The *gradient* $\partial_1(\psi)$ determines the direction of the step taken to the optimum and the Hessian modifies the size of the step. It is possible to overstep the maximum in the direction determined by the vector

$$\tilde{\pi}(\psi) = -\tilde{\partial}_2(\psi)^{-1}\tilde{\partial}_1(\psi),$$

and therefore it is common practice to include a line search along the gradient vector within the optimisation process. We obtain the algorithm

$$\bar{\psi} = \tilde{\psi} + s\tilde{\pi}(\psi),$$

where various methods are available to find the optimum value for s which is usually found to be between 0 and 1. In practice it is often computationally demanding or impossible to compute $\partial_1(\psi)$ and $\partial_2(\psi)$ analytically. Numerical evaluation of $\partial_1(\psi)$ is usually feasible. A variety of computational devices are available to approximate $\partial_2(\psi)$ in order to avoid computing it analytically or numerically. For example, the *STAMP* package of Koopman, Harvey, Doornik and Shephard (2010) and the *Ox* matrix programming system of Doornik (2010) both use the so-called BFGS (Broyden–Fletcher–Goldfarb–Shannon) method which approximates the Hessian matrix using a device in which at each new value for ψ a new approximate inverse Hessian matrix is obtained via the recursion

$$\bar{\partial}_2(\psi)^{-1} = \tilde{\partial}_2(\psi)^{-1} + \left(s + \frac{g'g^*}{\tilde{\pi}(\psi)'g}\right)\frac{\tilde{\pi}(\psi)\tilde{\pi}(\psi)'}{\tilde{\pi}(\psi)'g} - \frac{\tilde{\pi}(\psi)g^{*\prime} + g^*\tilde{\pi}(\psi)'}{\tilde{\pi}(\psi)'g},$$

where g is the difference between the gradient $\tilde{\partial}_1(\psi)$ and the gradient for a trial value of ψ prior to $\tilde{\psi}$ and

$$g^* = \tilde{\partial}_2(\psi)^{-1}g.$$

The BFGS method ensures that the approximate Hessian matrix remains negative definite. The details and derivations of the Newton's method of optimisation and the BFGS method in particular can be found, for example, in Fletcher (1987).

Model parameters are sometimes constrained. For example, the parameters in the local level model (2.3) must satisfy the constraints $\sigma_\varepsilon^2 \geq 0$ and $\sigma_\eta^2 \geq 0$ with $\sigma_\varepsilon^2 + \sigma_\eta^2 > 0$. However, the introduction of constraints such as these within the numerical procedure is inconvenient and it is preferable that the maximisation is performed with respect to quantities which are unconstrained. For this example we therefore make the transformations $\psi_\varepsilon = \frac{1}{2}\log\sigma_\varepsilon^2$ and $\psi_\eta = \frac{1}{2}\log\sigma_\eta^2$ where $-\infty < \psi_\varepsilon, \psi_\eta < \infty$, thus converting the problem to one in unconstrained maximisation. The parameter vector is $\psi = [\psi_\varepsilon, \psi_\eta]'$. Similarly, if we have a parameter χ which is restricted to the range $[-a, a]$ where a is positive we can make a transformation to ψ_χ for which

$$\chi = \frac{a\psi_\chi}{\sqrt{1+\psi_\chi^2}}, \qquad -\infty < \psi_\chi < \infty.$$

7.3.3 The score vector

We now consider details of the calculation of the gradient or *score vector*

$$\partial_1(\psi) = \frac{\partial \log L(Y_n|\psi)}{\partial \psi}.$$

As indicated in the last section, this vector is important in numerical maximisation since it specifies the direction in the parameter space along which a search should be made.

We begin with the case where the initial vector α_1 has the distribution $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. Let $p(\alpha, Y_n|\psi)$ be the joint density of α and Y_n, let $p(\alpha|Y_n, \psi)$ be the conditional density of α given Y_n and let $p(Y_n|\psi)$ be the marginal density of Y_n for given ψ. We now evaluate the score vector $\partial \log L(Y_n|\psi)/\partial\psi = \partial \log p(Y_n|\psi)/\partial\psi$ at the trial value $\tilde{\psi}$. We have

$$\log p(Y_n|\psi) = \log p(\alpha, Y_n|\psi) - \log p(\alpha|Y_n, \psi).$$

Let $\tilde{\mathrm{E}}$ denote expectation with respect to density $p(\alpha|Y_n, \tilde{\psi})$. Since $p(Y_n|\psi)$ does not depend on α, taking $\tilde{\mathrm{E}}$ of both sides gives

$$\log p(Y_n|\psi) = \tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] - \tilde{\mathrm{E}}[\log p(\alpha|Y_n, \psi)].$$

To obtain the score vector at $\tilde{\psi}$, we differentiate both sides with respect to ψ and put $\psi = \tilde{\psi}$. Assuming that differentiating under integral signs is legitimate,

$$\begin{aligned}\tilde{\mathrm{E}}\left[\left.\frac{\partial \log p(\alpha|Y_n, \psi)}{\partial\psi}\right|_{\psi=\tilde{\psi}}\right] &= \int \frac{1}{p(\alpha|Y_n, \tilde{\psi})} \left.\frac{\partial p(\alpha|Y_n, \psi)}{\partial\psi}\right|_{\psi=\tilde{\psi}} p(\alpha|Y_n, \tilde{\psi})\, d\alpha \\ &= \frac{\partial}{\partial\psi} \int p(\alpha|Y_n, \psi)\, d\alpha \bigg|_{\psi=\tilde{\psi}} = 0.\end{aligned}$$

Thus

$$\left.\frac{\partial \log p(Y_n|\psi)}{\partial\psi}\right|_{\psi=\tilde{\psi}} = \tilde{\mathrm{E}}\left[\frac{\partial \log p(\alpha, Y_n|\psi)}{\partial\psi}\right]\bigg|_{\psi=\tilde{\psi}}.$$

With substitutions $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ and $\varepsilon_t = y_t - Z_t\alpha_t$ and putting $\alpha_1 - a_1 = \eta_0$ and $P_1 = Q_0$, we obtain

$$\begin{aligned}\log p(\alpha, Y_n|\psi) = {} & \text{constant} \\ & - \frac{1}{2}\sum_{t=1}^{n} \big(\log|H_t| + \log|Q_{t-1}| + \varepsilon_t' H_t^{-1}\varepsilon_t + \eta_{t-1}' Q_{t-1}^{-1}\eta_{t-1} \big).\end{aligned} \quad (7.13)$$

On taking the expectation $\tilde{\mathrm{E}}$ and differentiating with respect to ψ, this gives the score vector at $\psi = \tilde{\psi}$,

$$\left.\frac{\partial \log L(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = -\frac{1}{2}\frac{\partial}{\partial \psi}\sum_{t=1}^{n}\big[\big(\log|H_t| + \log|Q_{t-1}| \\ + \mathrm{tr}\big[\{\hat{\varepsilon}_t\hat{\varepsilon}_t' + \mathrm{Var}(\varepsilon_t|Y_n)\}H_t^{-1}\big] \\ + \mathrm{tr}\big[\{\hat{\eta}_{t-1}\hat{\eta}_{t-1}' + \mathrm{Var}(\eta_{t-1}|Y_n)\}Q_{t-1}^{-1}\big]\big|\psi\big)\big|_{\psi=\tilde{\psi}}\big], \quad (7.14)$$

where $\hat{\varepsilon}_t$, $\hat{\eta}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\eta_{t-1}|Y_n)$ are obtained for $\psi = \tilde{\psi}$ as in Section 4.5.

Only the terms in H_t and Q_t in (7.14) require differentiation with respect to ψ. Since in practice H_t and Q_t are often simple functions of ψ, this means that the score vector is often easy to calculate, which can be a considerable advantage in numerical maximisation of the loglikelihood. A similar technique can be developed for the system matrices Z_t and T_t but this requires more computations which involve the state smoothing recursions. Koopman and Shephard (1992), to whom the result (7.14) is due, therefore conclude that the score values for ψ associated with system matrices Z_t and T_t can be evaluated better numerically than analytically.

We now consider the diffuse case. In Section 5.1 we specified the initial state vector α_1 as

$$\alpha_1 = a + A\delta + R_0\eta_0, \qquad \delta \sim \mathrm{N}(0, \kappa I_q), \qquad \eta_0 \sim \mathrm{N}(0, Q_0),$$

where Q_0 is nonsingular. Equation (7.13) is still valid except that $\alpha_1 - a_1 = \eta_0$ is now replaced by $\alpha_1 - a = A\delta + R_0\eta_0$ and $P_1 = \kappa P_\infty + P_*$ where $P_* = R_0Q_0R_0'$. Thus for finite κ the term

$$-\frac{1}{2}\frac{\partial}{\partial \psi}(q\log\kappa + \kappa^{-1}\mathrm{tr}\{\hat{\delta}\hat{\delta}' + \mathrm{Var}(\delta|Y_n)\})$$

must be included in (7.14). Defining

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \lim_{\kappa\to\infty}\frac{\partial}{\partial \psi}\left[\log L(Y_n|\psi) + \frac{q}{2}\log\kappa\right],$$

analogously to the definition of $\log L_d(Y_n)$ in Subsection 7.2.2, and letting $\kappa \to \infty$, we have that

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \left.\frac{\partial \log L(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}}, \quad (7.15)$$

which is given in (7.14). In the event that α_1 consists only of diffuse elements, so the vector η_0 is null, the terms in Q_0 disappear from (7.14).

As an example consider the local level model (2.3) with η replaced by ξ, for which

$$\psi = \begin{pmatrix} \psi_\varepsilon \\ \psi_\xi \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\log\sigma_\varepsilon^2 \\ \frac{1}{2}\log\sigma_\xi^2 \end{pmatrix},$$

with a diffuse initialisation for α_1. Then ψ is the unknown parameter vector of the kind mentioned in Section 7.1. We have on substituting $y_t - \alpha_t = \varepsilon_t$ and $\alpha_{t+1} - \alpha_t = \xi_t$,

$$\begin{aligned}\log p(\alpha, Y_n|\psi) = &-\frac{2n-1}{2}\log 2\pi - \frac{n}{2}\log\sigma_\varepsilon^2 - \frac{n-1}{2}\log\sigma_\xi^2 \\ &- \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^n \varepsilon_t^2 - \frac{1}{2\sigma_\xi^2}\sum_{t=2}^n \xi_{t-1}^2,\end{aligned}$$

and

$$\begin{aligned}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] = &-\frac{2n-1}{2}\log 2\pi - \frac{n}{2}\log\sigma_\varepsilon^2 - \frac{n-1}{2}\log\sigma_\xi^2 - \frac{1}{2\sigma_\varepsilon^2} \\ &\times \sum_{t=1}^n \big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\} - \frac{1}{2\sigma_\xi^2}\sum_{t=2}^n \big\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\big\},\end{aligned}$$

where the conditional means and variances for ε_t and ξ_t are obtained from the Kalman filter and disturbance smoother with σ_ε^2 and σ_ξ^2 implied by $\psi = \tilde{\psi}$. To obtain the score vector, we differentiate both sides with respect to ψ where we note that, with $\psi_\varepsilon = \frac{1}{2}\log\sigma_\varepsilon^2$,

$$\begin{aligned}\frac{\partial}{\partial\sigma_\varepsilon^2}\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\}\right] &= \frac{1}{\sigma_\varepsilon^2} - \frac{1}{\sigma_\varepsilon^4}\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\}, \\ \frac{\partial\sigma_\varepsilon^2}{\partial\psi_\varepsilon} &= 2\sigma_\varepsilon^2.\end{aligned}$$

The terms $\hat{\varepsilon}_t$ and $\mathrm{Var}(\varepsilon_t|Y_n)$ do not vary with ψ since they have been calculated on the assumption that $\psi = \tilde{\psi}$. We obtain

$$\begin{aligned}\frac{\partial\log L_d(Y_n|\psi)}{\partial\psi_\varepsilon} &= -\frac{1}{2}\frac{\partial}{\partial\psi_\varepsilon}\sum_{t=1}^n\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\}\right] \\ &= -n + \frac{1}{\sigma_\varepsilon^2}\sum_{t=1}^n\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\}.\end{aligned}$$

In a similar way we have

$$\begin{aligned}\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi_\xi} &= -\frac{1}{2}\frac{\partial}{\partial \psi_\xi}\sum_{t=2}^{n}\left[\log \sigma_\xi^2 + \frac{1}{\sigma_\xi^2}\left\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\right\}\right] \\ &= 1 - n + \frac{1}{\sigma_\xi^2}\sum_{t=2}^{n}\left\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\right\}.\end{aligned}$$

The score vector for ψ of the local level model evaluated at $\psi = \tilde{\psi}$ is therefore

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \begin{bmatrix} \tilde{\sigma}_\varepsilon^2 \sum_{t=1}^{n}\left(u_t^2 - D_t\right) \\ \tilde{\sigma}_\xi^2 \sum_{t=2}^{n}\left(r_{t-1}^2 - N_{t-1}\right) \end{bmatrix},$$

with $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$ from $\tilde{\psi}$. This result follows since from Subsections 2.5.1 and 2.5.2 $\hat{\varepsilon}_t = \tilde{\sigma}_\varepsilon^2 u_t$, $\mathrm{Var}(\varepsilon_t|Y_n) = \tilde{\sigma}_\varepsilon^2 - \tilde{\sigma}_\varepsilon^4 D_t$, $\hat{\xi}_t = \tilde{\sigma}_\xi^2 r_t$ and $\mathrm{Var}(\xi_t|Y_n) = \tilde{\sigma}_\xi^2 - \tilde{\sigma}_\xi^4 N_t$.

It is very satisfactory that after so much algebra we obtain such a simple expression for the score vector which can be computed efficiently using the disturbance smoothing equations of Section 4.5. We can compute the score vector for the diffuse case efficiently because it is shown in Section 5.4 that no extra computing is required for disturbance smoothing when dealing with a diffuse initial state vector. Finally, score vector elements associated with variances or variance matrices in more complicated models such as multivariate structural time series models continue to have similar relatively simple expressions. Koopman and Shephard (1992) give for these models the score vector for parameters in H_t, R_t and Q_t as the expression

$$\begin{aligned}\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} &= \frac{1}{2}\sum_{t=1}^{n}\mathrm{tr}\left\{(u_t u_t' - D_t)\frac{\partial H_t}{\partial \psi}\right\} \\ &\quad \left. + \frac{1}{2}\sum_{t=2}^{n}\mathrm{tr}\left\{(r_{t-1}r_{t-1}' - N_{t-1})\frac{\partial R_t Q_t R_t'}{\partial \psi}\right\}\right|_{\psi=\tilde{\psi}}, \qquad (7.16)\end{aligned}$$

where u_t, D_t, r_t and N_t are evaluated by the Kalman filter and smoother as discussed in Sections 4.5 and 5.4.

7.3.4 The EM algorithm

The EM algorithm is a well-known tool for iterative maximum likelihood estimation which for many state space models has a particularly neat form. The earlier EM methods for the state space model were developed by Shumway and Stoffer (1982) and Watson and Engle (1983). The EM algorithm can be used either entirely instead of, or in place of the early stages of, direct numerical maximisation of the loglikelihood. It consists of an E-step (expectation) and

M-step (maximisation) for which the former involves the evaluation of the conditional expectation $\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)]$ and the latter maximises this expectation with respect to the elements of ψ. The details of estimating unknown elements in H_t and Q_t are given by Koopman (1993) and they are close to those required for the evaluation of the score function. Taking first the case of a_1 and P_1 known and starting with (7.13), we evaluate $\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)]$ and as in (7.14) we obtain

$$\begin{aligned}\frac{\partial}{\partial\psi}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] = -\frac{1}{2}\frac{\partial}{\partial\psi}\sum_{t=1}^{n}\big[&\log|H_t| + \log|Q_{t-1}| \\ &+ \mathrm{tr}\big[\{\hat{\varepsilon}_t\hat{\varepsilon}_t' + \mathrm{Var}(\varepsilon_t|Y_n)\}H_t^{-1}\big] \\ &+ \mathrm{tr}\big[\{\hat{\eta}_{t-1}\hat{\eta}_{t-1}' + \mathrm{Var}(\eta_{t-1}|Y_n)\}Q_{t-1}^{-1}\big]\big|\psi\big], \quad (7.17)\end{aligned}$$

where $\hat{\varepsilon}_t$, $\hat{\eta}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\eta_{t-1}|Y_n)$ are computed assuming $\psi = \tilde{\psi}$, while H_t and Q_{t-1} retain their original dependence on ψ. The equations obtained by setting (7.17) equal to zero are then solved for the elements of ψ to obtain a revised estimate of ψ. This is taken as the new trial value of ψ and the process is repeated either until adequate convergence is achieved or a switch is made to numerical maximisation of $\log L(Y_n|\psi)$. The latter option is often used since although the EM algorithm usually converges fairly rapidly in the early stages, its rate of convergence near the maximum is frequently substantially slower than numerical maximisation; see Watson and Engle (1983) and Harvey and Peters (1984) for discussion of this point. As for the score vector in the previous section, when α_1 is diffuse we merely redefine η_0 and Q_0 in such a way that they are consistent with the initial state vector model $\alpha_1 = a + A\delta + R_0\eta_0$ where $\delta \sim \mathrm{N}(0, \kappa I_q)$ and $\eta_0 \sim \mathrm{N}(0, Q_0)$ and we ignore the part associated with δ. When α_1 consists only of diffuse elements the term in Q_0^{-1} disappears from (7.17).

To illustrate, we apply the EM algorithm to the local level model as in the previous section but now we take

$$\psi = \begin{pmatrix}\sigma_\varepsilon^2 \\ \sigma_\xi^2\end{pmatrix},$$

as the unknown parameter vector. The E-step involves the Kalman filter and disturbance smoother to obtain $\hat{\varepsilon}_t$, $\hat{\xi}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\xi_{t-1}|Y_n)$ of (7.17) given $\psi = \tilde{\psi}$. The M-step solves for σ_ε^2 and σ_ξ^2 by equating (7.17) to zero. For example, in a similar way as in the previous section we have

$$\begin{aligned}-2\frac{\partial}{\partial\sigma_\varepsilon^2}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] &= \frac{\partial}{\partial\sigma_\varepsilon^2}\sum_{t=1}^{n}\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\}\right] \\ &= \frac{n}{\sigma_\varepsilon^2} - \frac{1}{\sigma_\varepsilon^4}\sum_{t=1}^{n}\big\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\big\} \\ &= 0,\end{aligned}$$

and similarly for the term in σ_ξ^2. New trial values for σ_ε^2 and σ_ξ^2 are therefore obtained from

$$\bar{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{t=1}^{n}\left\{\hat{\varepsilon}_t^2 - \mathrm{Var}(\varepsilon_t|Y_n)\right\} = \tilde{\sigma}_\varepsilon^2 + \frac{1}{n}\tilde{\sigma}_\varepsilon^4\sum_{t=1}^{n}\left(u_t^2 - D_t\right),$$

$$\bar{\sigma}_\xi^2 = \frac{1}{n-1}\sum_{t=2}^{n}\left\{\hat{\xi}_{t-1}^2 - \mathrm{Var}(\xi_{t-1}|Y_n)\right\} = \tilde{\sigma}_\xi^2 + \frac{1}{n-1}\tilde{\sigma}_\xi^4\sum_{t=2}^{n}\left(r_{t-1}^2 - N_{t-1}\right),$$

since $\hat{\varepsilon}_t = \tilde{\sigma}_\varepsilon^2 u_t$, $\mathrm{Var}(\varepsilon_t|Y_n) = \tilde{\sigma}_\varepsilon^2 - \tilde{\sigma}_\varepsilon^4 D_t$, $\hat{\xi}_t = \tilde{\sigma}_\xi^2 r_t$ and $\mathrm{Var}(\xi_t|Y_n) = \tilde{\sigma}_\xi^2 - \tilde{\sigma}_\xi^4 N_t$. The disturbance smoothing values u_t, D_t, r_t and N_t are based on $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$. The new values $\bar{\sigma}_\varepsilon^2$ and $\bar{\sigma}_\xi^2$ replace $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$ and the procedure is repeated until either convergence has been attained or until a switch is made to numerical optimisation. Similar elegant results are obtained for more general time series models where unknown parameters occur only in the H_t and Q_t matrices.

7.3.5 Estimation when dealing with diffuse initial conditions

It was shown in previous sections that only minor adjustments are required for parameter estimation when dealing with a diffuse initial state vector. The diffuse loglikelihood requires either the exact initial Kalman filter or the augmented Kalman filter. In both cases the diffuse loglikelihood is calculated in much the same way as for the nondiffuse case. No real new complications arise when computing the score vector or when estimating parameters via the EM algorithm. There is a compelling argument however for using the exact initial Kalman filter of Section 5.2 rather than the augmented Kalman filter of Section 5.7 for the estimation of parameters. For most practical models, the matrix $P_{\infty,t}$ and its associated matrices $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$ do not depend on parameter vector ψ. This may be surprising but, for example, by studying the illustration given in Subsection 5.6.1 for the local linear trend model we see that the matrices $P_{\infty,t}$, $K_t^{(0)} = T_t M_{\infty,t} F_{\infty,t}^{-1}$ and $L_t^{(0)} = T_t - K_t^{(0)} Z_t$ do not depend on σ_ε^2, σ_ξ^2 or on σ_ζ^2. On the other hand, we see that all the matrices reported in Subsection 5.7.4, which deals with the augmentation approach to the same example, depend on $q_\xi = \sigma_\xi^2/\sigma_\varepsilon^2$ and $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$. Therefore, every time that the parameter vector ψ changes during the estimation process we need to recalculate the augmented part of the augmented Kalman filter while we do not have to recalculate the matrices related to $P_{\infty,t}$ for the exact initial Kalman filter.

First we consider the case where only the system matrices H_t, R_t and Q_t depend on the parameter vector ψ. The matrices $F_{\infty,t} = Z_t P_{\infty,t} Z_t'$ and $M_{\infty,t} = P_{\infty,t} Z_t'$ do not depend on ψ since the update equation for $P_{\infty,t}$ is given by

$$P_{\infty,t+1} = T_t P_{\infty,t}\left(T_t - K_t^{(0)} Z_t\right)',$$

where $K_t^{(0)} = T_t M_{\infty,t} F_{\infty,t}^{-1}$ and $P_{\infty,1} = AA'$, for $t = 1, \ldots, d$. Thus for all quantities related to $P_{\infty,t}$ the parameter vector ψ does not play a role. The same holds for computing a_{t+1} for $t = 1, \dots, d$ since

$$a_{t+1} = T_t a_t + K_t^{(0)} v_t,$$

where $v_t = y_t - Z_t a_t$ and $a_1 = a$. Here again no quantity depends on ψ. The update equation

$$P_{*,t+1} = T_t P_{*,t} \big(T_t - K_t^{(0)} Z_t \big)' - K_t^{(0)} F_{\infty,t} K_t^{(1)\prime} + R_t Q_t R_t',$$

where $K_t^{(1)} = T_t M_{*,t} F_{\infty,t}^{-1} - K_t^{(0)} F_{*,t} F_{\infty,t}^{-1}$ depends on ψ. Thus we compute vector v_t and matrices $K_t^{(0)}$ and $F_{\infty,t}$ for $t = 1, \dots, d$ once at the start of parameter estimation and we store them. When the Kalman filter is called again for likelihood evaluation we do not need to recompute these quantities and we only need to update the matrix $P_{*,t}$ for $t = 1, \dots, d$. This implies considerable computational savings during parameter estimation using the EM algorithm or maximising the diffuse loglikelihood using a variant of Newton's method.

For the case where ψ also affects the system matrices Z_t and T_t we achieve the same computational savings for all nonstationary models we have considered in this book. The matrices Z_t and T_t may depend on ψ but the parts of Z_t and T_t which affect the computation of $P_{\infty,t}$, $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$ for $t = 1, \dots, d$ do not depend on ψ. It should be noted that the rows and columns of $P_{\infty,t}$ associated with elements of α_1 which are not elements of δ are zero for $t = 1, \dots, d$. Thus the columns of Z_t and the rows and columns of T_t related to stationary elements of the state vector do not influence the matrices $P_{\infty,t}$, $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$. In the nonstationary time series models of Chapter 3 such as the ARIMA and structural time series models, all elements of ψ which affect Z_t and T_t only relate to the stationary part of the model, for $t = 1, \dots, d$. The parts of Z_t and T_t associated with δ only have values equal to zero and unity. For example, the ARIMA(2,1,1) model of Section 3.4 shows that $\psi = (\phi_1\,, \phi_2\,, \theta_1\,, \sigma^2)'$ does not influence the elements of Z_t and T_t associated with the first element of the state vector.

7.3.6 Large sample distribution of estimates

It can be shown that under reasonable assumptions about the stability of the model over time, the distribution of $\hat{\psi}$ for large n is approximately

$$\hat{\psi} \sim \mathrm{N}(\psi, \Omega), \tag{7.18}$$

where

$$\Omega = \left[-\frac{\partial^2 \log L}{\partial \psi \partial \psi'} \right]^{-1}. \tag{7.19}$$

This distribution has the same form as the large sample distribution of maximum likelihood estimators from samples of independent and identically distributed observations. The result (7.18) is discussed by Hamilton (1994) in Section 5.8 for general time series models and in Section 13.4 for the special case of linear Gaussian state space models. In his discussion, Hamilton gives a number of references to theoretical work on the subject.

7.3.7 Effect of errors in parameter estimation

Up to this point we have followed standard classical statistical methodology by first deriving estimates of quantities of interest on the assumption that the parameter vector ψ is known and then replacing ψ in the resulting formulae by its maximum likelihood estimate $\hat{\psi}$. We now consider the estimation of the biases in the estimates that might arise from following this procedure. Since an analytical solution in the general case seems intractable, we employ simulation. We deal with cases where $\mathrm{Var}(\hat{\psi}) = O(n^{-1})$ so the biases are also of order n^{-1}.

The technique that we propose is simple. Pretend that $\hat{\psi}$ is the true value of ψ. From (7.18) and (7.19) we know that the approximate large sample distribution of the maximum likelihood estimate of ψ given that the true ψ is $\hat{\psi}$ is $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, where $\hat{\Omega}$ is Ω given by (7.19) evaluated at $\psi = \hat{\psi}$. Draw a simulation sample of N independent values $\psi^{(i)}$ from $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, $i = 1, \dots, N$. Denote by e a scalar, vector or matrix quantity that we wish to estimate from the sample Y_n and let

$$\hat{e} = \mathrm{E}(e|Y_n)|_{\psi=\hat{\psi}}$$

be the estimate of e obtained by the methods of Chapter 4. For simplicity we focus on smoothed values, though an essentially identical technique holds for filtered estimates. Let

$$e^{(i)} = \mathrm{E}(e|Y_n)|_{\psi=\psi^{(i)}}$$

be the estimate of e obtained by taking $\psi = \psi^{(i)}$, for $i = 1, \dots, N$. Then estimate the bias by

$$\hat{B}_e = \frac{1}{N}\sum_{i=1}^{N} e^{(i)} - \hat{e}. \tag{7.20}$$

The accuracy of $\hat{B}_e$ can be improved significantly by the use of antithetic variables, which are discussed in detail in Subsection 11.4.3 in connection with the use of importance sampling in the treatment of non-Gaussian models. For example, we can balance the sample of $\psi^{(i)}$'s for location by taking only $N/2$ draws from $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, where N is even, and defining $\psi^{(N-i+1)} = 2\hat{\psi} - \psi^{(i)}$ for $i = 1, \dots, N/2$. Since $\psi^{(N-i+1)} - \hat{\psi} = -(\psi^{(i)} - \hat{\psi})$ and the distribution of $\psi^{(i)}$ is symmetric about $\hat{\psi}$, the distribution of $\psi^{(N-i+1)}$ is the same as that of $\psi^{(i)}$. In this way we not only reduce the numbers of draws required from the $\mathrm{N}(\hat{\psi}, \hat{\Omega})$ distribution by half, but we introduce negative correlation between the $\psi^{(i)}$'s which will reduce sample variation and we have arranged the simulation sample so that the sample mean $(\psi^{(1)} + \cdots + \psi^{(N)})/N$ is equal to the population mean $\hat{\psi}$.

We can balance the sample for scale by a technique described in Subsection 11.4.3 using the fact that

$$\big(\psi^{(i)} - \hat{\psi}\big)'\hat{\Omega}^{-1}\big(\psi^{(i)} - \hat{\psi}\big) \sim \chi^2_w,$$

where w is the dimensionality of ψ; however, our expectation is that in most cases balancing for location only would be sufficient. The mean square error matrix due to simulation can be estimated in a manner similar to that described in Subsection 11.6.5.

Of course, we are not proposing that bias should be estimated as a standard part of routine time series analysis. We have included a description of this technique in order to assist workers in investigating the degree of bias in particular types of problems; in most practical cases we would expect the bias to be small enough to be neglected.

Simulation for correcting for bias due to errors in parameter estimates has previously been suggested by Hamilton (1994). His methods differ from ours in two respects. First he uses simulation to estimate the entire function under study, which in his case is a mean square error matrix, rather than just the bias, as in our treatment. Second, he has omitted a term of the same order as the bias, namely n^{-1}, as demonstrated for the local level model that we considered in Chapter 2 by Quenneville and Singh (1997). This latter paper corrects Hamilton's method and provides interesting analytical and simulation results but it only gives details for the local level model. Different methods based on parametric and nonparametric bootstrap samples have been proposed by Stoffer and Wall (1991, 2004) and Pfeffermann and Tiller (2005).

7.4 Goodness of fit

Given the estimated parameter vector $\hat{\psi}$, we may want to measure the fit of the model under consideration for the given time series. Goodness of fit measures for time series models are usually associated with forecast errors. A basic measure of fit is the forecast variance F_t which could be compared with the forecast variance of a naive model. For example, when we analyse a time series with time-varying trend and seasonal, we could compare the forecast variance of this model with the forecast variance of the time series after adjusting it with fixed trend and seasonal.

When dealing with competing models, we may want to compare the loglikelihood value of a particular fitted model, as denoted by $\log L(Y_n|\hat{\psi})$ or $\log L_d(Y_n|\hat{\psi})$, with the corresponding loglikelihood values of competing models. Generally speaking, the larger the number of parameters that a model contains the larger its loglikelihood. In order to have a fair comparison between models with different numbers of parameters, information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are used. For a univariate series they are given by

$$\text{AIC} = n^{-1}[-2\log L(Y_n|\hat{\psi}) + 2w], \qquad \text{BIC} = n^{-1}[-2\log L(Y_n|\hat{\psi}) + w\log n],$$

and with diffuse initialisation they are given by

$$\text{AIC} = n^{-1}[-2\log L_d(Y_n|\hat{\psi}) + 2(q+w)],$$
$$\text{BIC} = n^{-1}[-2\log L_d(Y_n|\hat{\psi}) + (q+w)\log n],$$

where w is the dimension of ψ. Models with more parameters or more nonstationary elements obtain a larger penalty. More details can be found in Harvey (1989). In general, a model with a smaller value of AIC or BIC is preferred.

7.5 Diagnostic checking

The diagnostic statistics and graphics discussed in Section 2.12 for the local level model (2.3) can be used in the same way for all univariate state space models. The basic diagnostics of Subsection 2.12.1 for normality, heteroscedasticity and serial correlation are applied to the one-step ahead forecast errors defined in (4.13) after standardisation by dividing by the standard deviation $F_t^{1/2}$. In the case of multivariate models, we can consider the standardised individual elements of the vector

$$v_t \sim \text{N}(0, F_t), \qquad t = d+1, \ldots, n,$$

but the individual elements are correlated since matrix F_t is not diagonal. The innovations can be transformed such that they are uncorrelated:

$$v_t^s = B_t v_t, \qquad F_t^{-1} = B_t' B_t.$$

It is then appropriate to apply the basic diagnostics to the individual elements of v_t^s. Another possibility is to apply multivariate generalisations of the diagnostic tests to the full vector v_t^s. A more detailed discussion on diagnostic checking can be found in Harvey (1989) and throughout the *STAMP* manual of Koopman, Harvey, Doornik and Shephard (2010).

Auxiliary residuals for the general state space model are constructed by

$$\hat{\varepsilon}_t^s = B_t^\varepsilon \hat{\varepsilon}_t, \qquad [\text{Var}(\hat{\varepsilon}_t)]^{-1} = B_t^{\varepsilon\prime} B_t^\varepsilon,$$
$$\hat{\eta}_t^s = B_t^\eta \hat{\eta}_t, \qquad [\text{Var}(\hat{\eta}_t)]^{-1} = B_t^{\eta\prime} B_t^\eta,$$

for $t = 1, \ldots, n$. The auxiliary residual $\hat{\varepsilon}_t^s$ can be used to identify outliers in the y_t series. Large absolute values in $\hat{\varepsilon}_t^s$ indicate that the behaviour of the observed value cannot be appropriately represented by the model under consideration. The usefulness of $\hat{\eta}_t^s$ depends on the interpretation of the state elements in α_t implied by the design of the system matrices T_t, R_t and Q_t. The way these auxiliary residuals can be exploited depends on their interpretation. For the local level model considered in Subsections 7.3.3 and 7.3.4 it is clear that the state is the time-varying level and ξ_t is the change of the level for time $t+1$.

It follows that structural breaks in the series y_t can be identified by detecting large absolute values in the series for $\hat{\xi}_t^s$. In the same way, for the univariate local linear trend model (3.2), the second element of $\hat{\xi}_t^s$ can be exploited to detect slope changes in the series y_t. Harvey and Koopman (1992) have formalised these ideas further for the structural time series models of Section 3.2 and they constructed some diagnostic normality tests for the auxiliary residuals.

It is argued by de Jong and Penzer (1998) that such auxiliary residuals can be computed for any element of the state vector and that they can be considered as t-tests for the hypotheses

$$H_0 : (\alpha_{t+1} - T_t \alpha_t - R_t \eta_t)_i = 0,$$

the appropriate large-sample statistic for which is computed by

$$r_{it}^s = r_{it} / \sqrt{N_{ii,t}},$$

for $i = 1, \ldots, m$, where $(\cdot)_i$ is the ith element of the vector within brackets, r_{it} is the ith element of the vector r_t and $N_{ij,t}$ is the (i, j)th element of the matrix N_t; the recursions for evaluating r_t and N_t are given in Subsection 4.5.3. The same applies to the measurement equation for which t-test statistics for the hypotheses

$$H_0 : (y_t - Z_t \alpha_t - \varepsilon_t)_i = 0,$$

are computed by

$$e_{it}^s = e_{it} / \sqrt{D_{ii,t}},$$

for $i = 1, \ldots, p$, where the equations for computing e_t and D_t are given in Subsection 4.5.3. These diagnostics can be regarded as model specification tests. Large values in r_{it}^s and e_{it}^s, for some values of i and t, may reflect departures from the overall model and they may indicate specific adjustments to the model.