8
Introduction to Time Series: State Space Models

8.1 INTRODUCTION

Time series data is, as its name suggests, ordered by time. It commonly arises in the fields of macroeconomics (e.g. one might have data on the unemployment rate observed every quarter since 1960) and finance (e.g. one might have data on the price of a particular stock every day for a year). There is a huge literature on econometric methods for time series, and it is impossible to do it justice in just one chapter. In this chapter we offer an introduction to one class of models referred to as *state space models* which are commonly-used with time series data. We make this choice for three reasons. First, as we shall see, state space models are hierarchical in nature. As was stressed in the previous chapter, Bayesian methods with hierarchical priors are particularly attractive. Secondly, Bayesian analysis of the main alternative approach[1] to time series econometrics has already been covered in detail in a recent textbook: Bauwens, Lubrano and Richard (1999). To avoid overlap, the present book offers a different way of looking at time series issues. Thirdly, state space models are not so much a different class of models than is used in Bauwens, Lubrano and Richard (1999), but rather offer a different way of writing the same models.[2] Hence, by using state space models, we can address all the same issues as Bauwens, Lubrano and Richard (1999), but stay in a hierarchical framework which is both familiar and computationally convenient.

We have already introduced many time series concepts in Chapter 6, Section 6.5, which discussed the linear regression model with autocorrelated errors. You may wish to review this material to remind yourself of basic concepts and notation. For instance, with time series we use t and T instead of i and N, so that y_t for $t = 1, \ldots, T$ indicates observations on the dependent variable from

[1] For readers with some knowledge of time series methods, note that this alternative approach includes autoregressive moving average (ARMA) models and extensions to dynamic regression models which allow for the discussion of issues like unit roots and cointegration.

[2] For instance, there is a state space representation for any ARMA model.

period 1 through T. Before discussing state space models, it is worth briefly mentioning that the techniques discussed in Chapter 6 (Section 6.5) can take you quite far in practice. For instance, the linear regression model with autocorrelated errors is a time series model which may be appropriate in many cases. This model has the errors, ε_t, following an AR(p) process. A common univariate time series model (i.e. a model for investigating the behavior of the single time series y) has y_t following an AR(p) process:

$$(1 - \rho_1 L - \cdots - \rho_p L^p) y_t = u_t \tag{8.1}$$

Computational methods for Bayesian analysis of this model are a straightforward simplification of those presented previously. In fact, (8.1) is really just a linear regression model where the explanatory variables are lags of the dependent variable

$$y_t = \rho_1 y_{t-1} + \cdots + \rho_p y_{t-p} + u_t \tag{8.2}$$

Thus, all the basic regression techniques discussed in previous chapters are relevant. Equation (8.2) can even be extended to include other explanatory variables (and their lags) while still remaining within the regression framework:

$$y_t = \rho_1 y_{t-1} + \cdots + \rho_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_q x_{t-q} + u_t \tag{8.3}$$

However, several complications arise in this regression-based approach. Loosely speaking, a good deal of the time series literature relates to placing restrictions on (or otherwise transforming) the coefficients of (8.3). There are also some important issues relating to prior elicitation which do not arise in cross-sectional contexts.[3]

Even if we stay within the class of state space models, we cannot possibly offer more than a superficial coverage of a few key issues in a single chapter. Accordingly, we will begin with the simplest univariate state space model called the *local level* model. Most of the basic issues involving prior elicitation and computation can be discussed in the context of this model. We then proceed to a more general state space model. For readers interested in more detail West and Harrison (1997) is a popular Bayesian textbook reference in this field.[4] Kim and Nelson (1999) is another Bayesian book which introduces and extends state space models.

In this chapter, we also use state space models to introduce empirical Bayes methods. These methods are increasingly popular with hierarchical models of all sorts. They provide a data-based method for eliciting prior hyperparameters. For the researcher who does not wish to subjectively elicit informative priors and

[3]In addition to Bauwens, Lubrano and Richard (1999), the reader interested in more detail is referred to the papers in themed issues of *Econometric Theory* (volume 10, August/October, 1994) and the *Journal of Applied Econometrics* (volume 6, October/December, 1991).

[4]A few other recent journal articles on Bayesian analysis of state space models are Carlin, Polson and Stoffer (1992), Carter and Kohn (1994), de Jong and Shephard (1995), Fruhwirth-Schnatter (1995), Koop and van Dijk (2000) and Shively and Kohn (1997). Durbin and Koopman (2001) is a good textbook source which has some Bayesian content.

does not wish to use a noninformative prior (e.g. since Bayes factors are hard to interpret with improper priors), empirical Bayesian methods offer an attractive alternative.[5]

8.2 THE LOCAL LEVEL MODEL

The local level model is given by

$$y_t = \alpha_t + \varepsilon_t \tag{8.4}$$

where ε_t is i.i.d. $N(0, h^{-1})$. The unique aspect of this model is the term α_t which is not observed and is assumed to follow a *random walk*

$$\alpha_{t+1} = \alpha_t + u_t \tag{8.5}$$

where u_t is i.i.d. $N(0, \eta h^{-1})$ and ε_t and u_s are independent of one another for all s and t. In (8.4) t runs from 1 through T while in (8.5) it runs from 1 through $T-1$. Equation (8.5) does not explicitly provide an expression for α_1, which is referred to as an *initial condition*. Equation (8.4) is referred to as the *observation* (or measurement) *equation*, while (8.5) is referred to as the *state equation*.

Before discussing Bayesian inference in the local level model, it is worthwhile to spend some time motivating this model. In Chapter 6, Section 6.5, we discussed the AR(1) model, and noted that if the coefficient on the lagged dependent variable, ρ, equalled one then the time series was nonstationary. Here it can be seen that (8.5) implies that α_t is nonstationary. In particular, it implies that α_t has a *stochastic trend*. The term *stochastic trend* arises from the fact that models such as (8.5) imply that a series can wander widely (i.e. trend) over time, but that an element of randomness enters the trend behavior. That is, in contrast to a *deterministic trend* such as

$$\alpha_t = \alpha + \beta t$$

where the variable is an exact function of time, a stochastic trend involves a random error, u_t. The fact that (8.5) implies that α_t exhibits trend behavior can be seen by noting that (8.5) can be written as

$$\alpha_t = \alpha_1 + \sum_{j=1}^{t-1} u_j \tag{8.6}$$

and, thus (ignoring the initial condition) $var(\alpha_t) = (t-1)\eta h^{-1}$. In addition, α_t and α_{t-1} tend to be close to one another (i.e. $E(\alpha_t|\alpha_{t-1}) = 0$). In words, the stochastic trend term has variance which is increasing with time (and thus can wander over an increasing wide range), but α_t changes only gradually over

[5]Carlin and Louis (2000) provides an excellent introduction to empirical Bayesian methods, although it is a statistics as opposed to econometrics textbook.

time. This is consistent with the intuitive concept of a trend as something which increases (or decreases) gradually over time.

To return to the local level model, we can see that (8.4) decomposes the observed series, y_t, into a trend component, α_t, and an error or irregular component, u_t.[6] In general, state space models can be interpreted as decomposing an observed time series into various parts. In the local level model, there are two components, a trend and an error. In more complicated state space models, the observed series can be decomposed into more components (e.g. trend, error and seasonal components).

It is worth mentioning that the local level model has been used for measuring the relative sizes of the trend and irregular components. This motivates the way that we have written the variances of the two errors (i.e. error variances are written as h^{-1} and ηh^{-1}). In this manner, η is directly interpreted as the size of the random walk relative to the error variance in the measurement equation. That is, it can be seen that if $\eta \rightarrow 0$, then the error drops out of (8.5) and $\alpha_t = \alpha_1$ for all t and (8.4) becomes $y_t = \alpha_1 + \varepsilon_t$. In this case, y_t exhibits random fluctuations around a constant level, α_1, and is not trending at all. However, as η becomes larger (i.e. the variance of u_t becomes larger), then the stochastic trend term plays a bigger role. Examining η is, thus, a nice way of measuring the importance of trend behavior in an economic time series. For the reader with previous knowledge of time series econometrics, note that the test of whether $\eta = 0$ is one way of testing for a *unit root*. We will not discuss unit root testing in any detail here. Suffice it to note that, unit root testing has played an important role in modern empirical macroeconomics, and that state space models allow for this to be done in an intuitive and straightforward manner.

Another way of interpreting (8.4) and (8.5) is by noting that α_t is the mean (or level) of y_t. Since this mean is varying over time, the terminology *local level model* is used. Interpreting α_t in this way, as a parameter, is natural in a Bayesian setup. That is, (8.4) can be interpreted as a very simple example of a linear regression model involving only an intercept. The innovative thing is that the intercept varies over time. Thus, the local level model is a simple example of a *time varying parameter model*. More sophisticated state space models can allow for time varying regression coefficients or time varying error variances. If $\alpha = (\alpha_1, \ldots, \alpha_T)'$ is interpreted as a vector of parameters then, as Bayesians, we must elicit a prior for it. But (8.5) provides us with such a prior. That is, (8.5) can be interpreted as defining a hierarchical prior for α. Note that, with such an interpretation, the local level model is very similar to the individual effects panel data model of Chapter 7 (Section 7.3) with $T = 1$. Of course, the individual effects model has an intercept which varies across individuals, while the local level model has an intercept which varies across time, but the basic structure of

[6]For the macroeconomist, some imperfect intuition for this would be that the trend term captures the long run trend growth of the economy (e.g. due to growth of the labor force, building up of capital stock and gradual technical improvements), whereas the irregular component reflects the random short term shocks hitting the economy (e.g. business cycle effects).

the two models is the same. Thus, the basic tools developed in Chapter 7 using an independent Normal-Gamma prior can be used here with some modifications. For this reason, in this section we do something new. We use a natural conjugate prior and introduce a new type of prior elicitation procedure.

That is, Bayesian methods using an independent Normal-Gamma prior are very similar to those described in Chapter 7, so we do not repeat them here. In particular, a Gibbs sampler with data augmentation can be developed as in Chapter 7. In Section 8.3 we develop such an algorithm in the context of a more general state space model. This can be used for the local level model and the reader interested in using the independent Normal-Gamma prior is referred to Section 8.3. In the present section, we will use a natural conjugate framework to introduce empirical Bayesian methods.

8.2.1 The Likelihood Function and Prior

If we define $y = (y_1, \ldots, y_T)'$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$, then we can write the local level model in matrix notation as

$$y = I_T \alpha + \varepsilon \tag{8.7}$$

If we make the standard error assumptions, that ε has a multivariate Normal distribution with mean 0_T and covariance matrix $h^{-1} I_T$, then this model is simply a Normal linear regression model where the matrix of explanatory variables is the identity matrix (i.e. $X = I_T$) and α is the T-vector of regression coefficients. Thus, the likelihood function has the standard form for the Normal linear regression model (e.g. see Chapter 3, (3.3)).

Of course, as in any Bayesian exercise, we can use any prior we wish. However, the state equation given in (8.5) suggests a hierarchical prior. We use one involving natural conjugate form. To draw out the similarities with results in Chapter 3 for the Normal linear regression model with natural conjugate prior, it is convenient to write this model in a slightly different way. To do this we begin by defining the $(T-1) \times T$ *first difference matrix*:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & 0 & -1 & 1 \end{bmatrix} \tag{8.8}$$

To draw out the connections with the state space model, note that

$$D\alpha = \begin{pmatrix} \alpha_2 - \alpha_1 \\ \cdot \\ \cdot \\ \alpha_T - \alpha_{T-1} \end{pmatrix}$$

and thus the state equation given in (8.5) can be written as:

$$D\alpha = u$$

where $u = (u_1, \ldots, u_{T-1})'$. The assumption that u is Normal can thus be interpreted as saying that the state equation is defining a Normal hierarchical prior for $D\alpha$.

To specify a complete prior for all the parameters in the model, we also need to specify a prior for h and α_1. To do this, we first write (8.7) as

$$y = W\theta + \varepsilon \tag{8.9}$$

where

$$\theta = \begin{pmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \cdot \\ \cdot \\ \alpha_T - \alpha_{T-1} \end{pmatrix}$$

and

$$W = \begin{pmatrix} 1 & 0'_{T-1} \\ \iota_{T-1} & C \end{pmatrix}$$

where ι_{T-1} is an $(T-1)$-vector of ones. Direct matrix multiplication can be used to verify that (8.9) is exactly equivalent to (8.7). Direct matrix inversion can be used to show that C is a $(T-1) \times (T-1)$ lower triangular matrix with all non-zero elements equalling one (it is the inverse of D with its first column removed). That is, C has all elements on or below the diagonal equalling 1, and all elements above the diagonal equalling 0.

We begin by eliciting a natural conjugate prior for θ and h:

$$\theta, h \sim NG(\underline{\theta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}) \tag{8.10}$$

The reader is referred to Chapter 3 for a reminder of notation and properties of this Normal-Gamma prior.

We consider a particular structure for $\underline{\theta}$ and $\underline{V}$ which embodies the prior information contained in the state equation:

$$\underline{\theta} = \begin{pmatrix} \underline{\theta}_1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix} \tag{8.11}$$

$$\underline{V} = \begin{pmatrix} \underline{V}_{11} & 0'_{T-1} \\ 0_{T-1} & \eta I_{T-1} \end{pmatrix} \tag{8.12}$$

Note that this prior implies $\alpha_{t+1} - \alpha_t$ is $N(0, \eta h^{-1})$, which is exactly what we assumed at the beginning of this section. The fact that this prior depends upon the parameter η makes it hierarchical. In addition, we have provided a prior for the initial condition, α_1, as being $N(\underline{\theta}_1, h^{-1}\underline{V}_{11})$.

At this point it is worth summarizing what we have done. We have written the local level model as a familiar Normal linear regression model with natural conjugate prior. The fact that this is a time series problem involving a state space model manifests itself solely through the prior we choose. In a Bayesian paradigm, the interpretation of the state equation as being a prior is natural and attractive. However, it is worth mentioning that the non-Bayesian econometrician would interpret our hierarchical prior as part of a likelihood function. As stressed in the previous chapter, in many models there is a degree of arbitrariness as to what part of a model is labelled the 'likelihood function' and what part is labelled the 'prior'.

8.2.2 The Posterior

Using standard results for the Normal linear regression model with natural conjugate prior (see Chapter 3), it follows that the posterior for θ and h, denoted by $p(\theta, h|y)$ is NG$(\overline{\theta}, \overline{V}, \overline{s}^{-2}, \overline{\nu})$ where

$$\overline{\theta} = \overline{V}(\underline{V}^{-1}\underline{\theta} + W'y) \tag{8.13}$$

$$\overline{V} = (\underline{V}^{-1} + W'W)^{-1} \tag{8.14}$$

$$\overline{\nu} = \underline{\nu} + T \tag{8.15}$$

and

$$\overline{\nu s}^2 = \underline{\nu s}^2 + (y - W\overline{\theta})'(y - W\overline{\theta}) + (\overline{\theta} - \underline{\theta})'\underline{V}^{-1}(\overline{\theta} - \underline{\theta}) \tag{8.16}$$

The properties of the Normal-Gamma distribution imply that it is easy to transform back from the parameterization in (8.9) to the original parameterization given in (8.7). That is, $p(\theta|h, y)$ is Normal and we know linear combinations of Normal are Normal (see Appendix B, Theorem B.10). Thus, if the posterior for (θ, h) is NG$(\overline{\theta}, \overline{V}, \overline{s}^{-2}, \overline{\nu})$ then the posterior for (α, h) is NG$(\overline{\alpha}, \overline{V}_\alpha, \overline{s}^{-2}, \overline{\nu})$ where

$$\overline{\alpha} = W\overline{\theta} \tag{8.17}$$

and

$$\overline{V}_\alpha = W\overline{V}W' \tag{8.18}$$

Since we have used a natural conjugate prior, analytical posterior results are available and there is no need for a posterior simulator. It is also interesting to note that the local level model is a regression model where the number of regression coefficients is equal to the number of observations. In a regression analysis, it is usually the case that the number of regression coefficients is much less than the number of observations (i.e. in the notation of previous chapters $k \ll N$). However, the local level model shows that prior information can, in many cases, be used to provide valid posterior inferences even in models with a huge number of parameters. Expressed in another way, the question arises as to

why we don't just obtain a degenerate posterior distribution at the point $\alpha = y$. After all, setting $\alpha_t = y_t$ for all t would yield a perfectly fitting model in the sense that $\varepsilon_t = 0$ for all t. It can be verified that the likelihood is infinite at this point. However, the Bayesian posterior is not located at this point of infinite likelihood because of prior information. The state equation says that α_{t+1} and α_t are close to one another, which pulls the posterior away from the point of perfect fit. In the state space literature, this is referred to as smoothing the state vector.

Since the model considered here is simply a Normal linear regression model with natural conjugate prior, model comparison and prediction can be done using methods outlined in Chapter 3.

8.2.3 Empirical Bayesian Methods

In previous chapters, we have either elicited priors subjectively or used noninformative priors. In the present context, this would mean choosing values for $\underline{\theta}$, $\underline{V}$, $\underline{s}^{-2}$, $\underline{\nu}$ or setting them to their noninformative values (see Chapter 3, Section 3.5) of $\underline{\nu} = 0$ and $\underline{V}^{-1} = 0_{T\times T}$.[7] However, both of these approaches had potential drawbacks. Subjective elicitation of priors may be difficult to do, or it may be subject to criticism by other researchers with different priors. Noninformative priors often make it difficult to do Bayesian model comparison since the resulting marginal likelihood may be undefined. Accordingly, some Bayesians use so-called empirical Bayes methods which surmount these two problems. The local level model is a convenient place to introduce empirical Bayes methods because some interesting issues arise in its application. However, empirical Bayes methods can be used with any model and are particularly popular with hierarchical prior models such as those of Chapter 7 and the present chapter. It should be noted, however, that empirical Bayesian methods have been criticized for implicitly double-counting the data. That is, the data is first used to select prior hyperparameter values. Once these values are selected, the data are used a second time in a standard Bayesian analysis.

Empirical Bayesian methods involve estimating prior hyperparameters from the data, rather than subjectively choosing values for them or setting them to noninformative values. The marginal likelihood is the preferred tool for this. In particular, for any choice of prior hyperparameters a marginal likelihood can be calculated. The values of the prior hyperparameters which yield the largest marginal likelihood are those used in an empirical Bayes analysis. However, searching over all possible prior hyperparameters can be a very difficult thing to do. Accordingly, empirical Bayes methods are often used on one or two key prior hyperparameters. Here we show how this might be done for the local level model.

The prior for the local level model specified in (8.10), (8.11) and (8.12) depends upon four hyperparameters η, $\underline{\theta}_1$, $\underline{V}_{11}$, $\underline{s}^{-2}$ and $\underline{\nu}$. Of these, η is almost invariably the most important and seems a candidate for the empirical Bayes approach. After

[7] Remember that, with these noninformative choices, the values of $\underline{\theta}$ and $\underline{s}^{-2}$ are irrelevant.

all, it can be interpreted as relating to the size of the random walk component in the state space model and it may be hard to elicit subjectively a value for it. Furthermore, setting it to an apparently 'noninformative' limiting value, $\eta \rightarrow \infty$, makes little sense since this implies the stochastic trend term completely dominates the irregular component. This is not 'noninformative', but rather quite informative. Accordingly, we focus on η. We will begin by assuming the researcher is able to subjectively elicit values for $\underline{\theta}_1$, $\underline{V}_{11}$, $\underline{s}^{-2}$ and $\underline{\nu}$.

The results of Chapter 3 (see (3.34)) imply that the marginal likelihood for the present model takes the form

$$p(y|\eta) = c\left(\frac{|\overline{V}|}{|\underline{V}|}\right)^{\frac{1}{2}} (\overline{\nu s}^2)^{-\frac{\overline{\nu}}{2}} \tag{8.19}$$

where

$$c = \frac{\Gamma\left(\frac{\overline{\nu}}{2}\right)(\underline{\nu s}^2)^{\frac{\underline{\nu}}{2}}}{\Gamma\left(\frac{\underline{\nu}}{2}\right)\pi^{\frac{T}{2}}} \tag{8.20}$$

The notation in (8.19) makes clear that we are treating the marginal likelihood as a function of η (i.e. in previous chapters we used notation $p(y)$ or $p(y|M_j)$ to denote the marginal likelihood, but here we make explicit the dependence on η). The standard way of carrying out an empirical Bayes analysis would be to choose, $\widehat{\eta}$, the value of η which maximizes $p(y|\eta)$ in (8.19). $\widehat{\eta}$ would then be plugged in (8.12), and posterior analysis could then be done in the standard way using (8.13)–(8.18). In the present model, $\widehat{\eta}$ could be found by using grid search methods. That is, the researcher could simply try every value for η in some appropriate grid and choose $\widehat{\eta}$ as being the value which maximizes $p(y|\eta)$.

A more formal way of carrying out empirical Bayesian estimation would involve explicitly treating η as a parameter and using the laws of conditional probability to carry out Bayesian inference. If η is treated as an unknown parameter, then Bayes theorem implies $p(\eta|y) \propto p(y|\eta)p(\eta)$ where $p(\eta)$ is a prior and we can write

$$p(\eta|y) \propto c\left(\frac{|\overline{V}|}{|\underline{V}|}\right)^{\frac{1}{2}} (\overline{\nu s}^2)^{-\frac{\overline{\nu}}{2}} p(\eta) \tag{8.21}$$

This posterior can be used to make inferences about η. If interest centers on the other parameters in the model, then we can use the fact that

$$p(\theta, h, \eta|y) = p(\theta, h|y, \eta)p(\eta|y)$$

Since $p(\theta, h|y, \eta)$ is Normal-Gamma (i.e. conditional on a specific value for η the posterior results in (8.13)–(8.18) hold) and $p(\eta|y)$ is one-dimensional, Monte Carlo integration can be used to carry out posterior inference in this model. That is, drawing from $p(\eta|y) \propto p(y|\eta)p(\eta)$ and, conditional upon this draw, drawing from $p(\theta, h|y, \eta)$ yields a draw from the joint posterior. As an aside, just how one draws from $p(\eta|y)$ depends on the exact form of $p(\eta)$. However, a simple

way of drawing from any univariate distribution involves approximating it by a discrete alternative. That is, evaluating $p(\eta|y)$ at B different points on a grid, $\eta_1, \ldots, \eta_B$, will yield $p(\eta_1|y), \ldots, p(\eta_B|y)$. Draws of η taken from the resulting discrete distribution (i.e. the distribution defined by $p(\eta = \eta_i) = p(\eta_i|y)$ for $i = 1, \ldots, B$), will be approximately equal to draws from $p(\eta|y)$. As B increases, the quality of the approximation will get better. In the empirical illustration below, we use this crude but effective strategy for carrying out Bayesian inference in the local level model.

The empirical Bayes methods for the local level model as described so far requires the researcher to choose $\underline{\theta}_1, \underline{V}_{11}, \underline{s}^{-2}$ and $\underline{\nu}$ (and $p(\eta)$ for the second approach outlined in the preceding paragraph). It is common to make noninformative choices for such prior hyperparameters and, for most models with hierarchical priors (e.g. the panel data models of Chapter 7), such a strategy works well. However, with the local level model, such a strategy does not work. It is worthwhile to discuss in detail why this is so, as it illustrates a problem which can occur in Bayesian inference in models with large numbers of parameters.

Consider first what happens when we set $\underline{\nu}$ and $\underline{V}_{11}^{-1}$ to their limiting values $\underline{\nu} = \underline{V}_{11}^{-1} = 0$. With these choices, the values of $\underline{s}^2$ and $\underline{\theta}_1$ are irrelevant. For these noninformative choices, it can be directly verified that $p(\theta, \sigma^{-2}|y, \eta)$ is a well-defined posterior. However, with regards to the marginal likelihood, two problems arise. First, the integrating constant in (8.20) is indeterminate. This is the standard problem we have discussed previously (e.g. see Chapter 2, Section 2.5). Insofar as interest centers on η, or the marginal likelihood is used for comparing the present model to another with the same noninformative prior for the error variance, this first problem is not a serious one. The constant c either does not enter or cancels out of any derivation (e.g. a Bayes factor) and can be ignored. Secondly, the term $\overline{\nu s}^2$ goes to zero as $\eta \to \infty$. To see this, note that with all the hyperparameters set to noninformative values $\overline{\theta} = (W'W)^{-1}W'y$ and $y - W\overline{\theta} = 0_T$. We will not provide a formal proof, but it is the case that this degeneracy is enough to imply that the marginal likelihood in (8.10) becomes infinite as $\eta \to \infty$. Hence, an empirical Bayes analysis will set $\widehat{\eta} \to \infty$ for any data set. It can be shown that this implies $E(\alpha|y) = y$ and no smoothing of the state vector occurs. Thus, empirical Bayes methods fail in the local level model when we set $\underline{\nu}$ and $\underline{V}_{11}^{-1}$ to noninformative values. This problem (which does not arise in most models) occurs because the number of explanatory variables in the linear regression model given in (8.7) is equal to the number of observations and, thus, it is possible for the regression line to fit perfectly. The general point to note here is that, in models with a large number of parameters, the researcher must be very careful when working with improper noninformative priors.

In the local level model, we have seen that we cannot use empirical Bayes methods with $\underline{\nu} = \underline{V}_{11}^{-1} = 0$. However, it can be verified that if we set either $\underline{\nu} > 0$ or $\underline{V}_{11}^{-1} > 0$ (and make an appropriate choice for $\underline{s}^2$ or $\underline{\theta}_1$), then we can use empirical Bayes methods. Intuitively, either of these will stop $\overline{\nu s}^2$ in (8.16)

from going to zero as $\eta \rightarrow \infty$. It is worth stressing that empirical Bayes methods work in the present model if *either* $\underline{\nu} > 0$ or $\underline{V}_{11}^{-1} > 0$, it is not necessary to have an informative prior for both h and θ_1.

In the alternative approach which involves treating η as a parameter (see (8.21)), a similar pathology occurs if we set $\underline{\nu} = \underline{V}_{11}^{-1} = 0$ and use an improper prior for η. For instance, if we set $\underline{\nu} = \underline{V}_{11}^{-1} = 0$ and choose $p(\eta)$ to be an improper Uniform distribution over the interval $(0, \infty)$ then it turns out that $p(\eta|y)$ is not a valid probability density function (i.e. it is improper). However, if we either set $\underline{\nu} > 0$ or $\underline{V}_{11}^{-1} > 0$ or choose $p(\eta)$ to be a proper p.d.f. then $p(\eta|y)$ is a valid posterior density. Thus, if we treat η as an unknown parameter, Bayesian inference can be carried out if prior informative about η or h or θ_1 is available.

8.2.4 Empirical Illustration: The Local Level Model

To illustrate empirical Bayesian inference in the local level model, we artificially generated data from the model given in (8.4) and (8.5) with $\eta = 1$, $h = 1$ and $\theta_1 \equiv \alpha_1 = 1$. For a prior we use $\theta, h \sim \text{NG}(\underline{\theta}, \underline{V}, \underline{s}^{-2}, \underline{\nu})$ with $\underline{\theta}$ and $\underline{V}$ as described in (8.11) and (8.12). We begin by considering four priors. The first of these is weakly informative for all parameters and sets $\underline{\nu} = 0.01$, $\underline{s}^{-2} = 1$, $\underline{\theta}_1 = 1$ and $\underline{V}_{11} = 100$. Note that this prior is centered over the values used to generate the data (i.e. $\underline{s}^{-2} = 1$ and $\underline{\theta}_1 = 1$), but expresses extreme uncertainty about these values. That is, the prior for h contains as much information as 0.01 of an observation and the prior variance for the initial condition is 100. The second prior is the same as the first, except that it is completely noninformative for h (i.e. $\underline{\nu} = 0$). The third prior is the same as the first, except that it is completely noninformative for θ_1 (i.e. $\underline{V}_{11}^{-1} = 0$). The fourth prior is completely noninformative for both parameters (i.e. $\underline{\nu} = \underline{V}_{11}^{-1} = 0$). Of course, the preceding discussion implies that empirical Bayesian methods should fail for this last prior.

Figure 8.1 plots the marginal likelihoods for a grid of values of η between 0 and 10. The plots corresponding to the four priors are very similar to one another. For the first three priors, we find empirical Bayes estimates of η being $\widehat{\eta} = 0.828$, $\widehat{\eta} = 0.828$ and $\widehat{\eta} = 0.823$, respectively. In fact, even the completely noninformative case (which has $\widehat{\eta} \rightarrow \infty$) would yield $\widehat{\eta} = 0.829$ if we limit consideration to the interval $(0, 10)$. The pathology noted with the use of a completely noninformative prior only occurs for extremely large values of η. Equation (8.21) can be used to derive $p(\eta|y)$ and, since we have not specified $p(\eta)$, our empirical illustration implicitly holds for the case where $p(\eta)$ is an improper Uniform prior over the interval $(0, \infty)$. Interpreted in this manner, our empirical illustration shows that if we use a completely noninformative prior for all parameters, $p(\eta|y)$ is a skewed (improper) distribution. It has a mode at the point $\eta = 0.829$, but then gradually increases to infinity as $\eta \rightarrow \infty$. Using results solely based on Figure 8.1 is equivalent to using a Uniform prior over

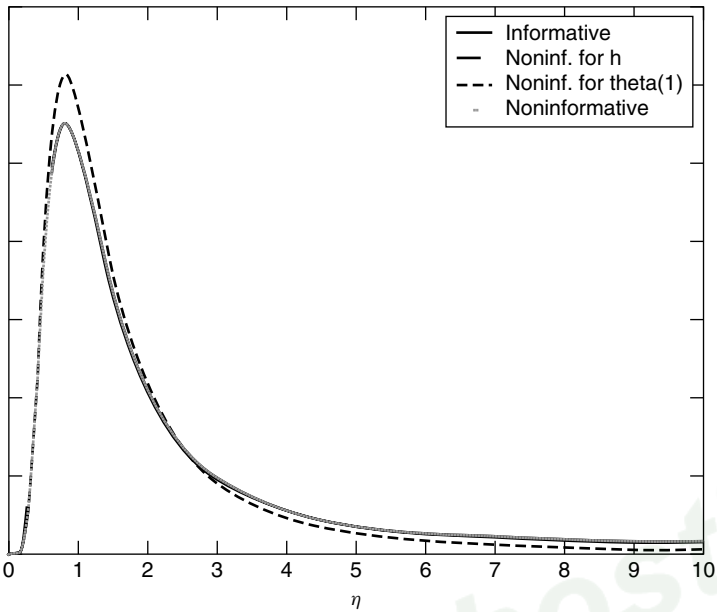

Figure 8.1 Marginal Likelihoods for Four Different Priors

the interval (0, 10) for η. Using such a prior for η is enough to ensure sensible empirical Bayes results.

In summary, users of empirical Bayes methods are often interested in focusing on one parameter and using noninformative priors over the rest. In the local level model with natural conjugate prior, this amounts to setting $\underline{\nu} = \underline{V}_{11}^{-1} = 0$ and using empirical Bayesian methods to estimate η. In the previous subsection we have shown that, in theory, this is not possible to do since we will always obtain $\widehat{\eta} \rightarrow \infty$. However, in practice, the empirical illustration shows that this pathology is probably not an important problem. That is, only a minuscule amount of prior information about either h or the initial condition or η is required to ensure empirical Bayesian methods will work.

So far we have focused exclusively on η, however it is often the case that interest centers on the state equation and, in particular, estimating the stochastic trend in the model. To investigate how well empirical Bayes methods work in this regard, we focus on the second prior of the previous section which uses a minuscule amount of prior information about h (i.e. $\underline{\nu} = 0.01$, $\underline{s}^{-2} = 1$), but is noninformative in all other respects. The other priors yield virtually identical results. We simulate four artificial data sets all of which have $h = 1$ and $\theta_1 = 1$ but have $\eta = 0$, 0.1, 1 and 100, respectively.

Figures 8.2a–8.2d plot the four data sets along with $E(\alpha|y)$ obtained using (8.17) for the value of η chosen using empirical Bayes methods. $E(\alpha|y)$ is referred to as the 'Fitted Trend' in the figures. Remember that α can be interpreted

as the stochastic trend in the time series, and is often of interest in a time series analysis. Before discussing the stochastic trend it is worthwhile to discuss the data itself. A wide variety of values for η have been chosen to show its role in determining the properties of the data. In Figure 8.2a we see how time series with no stochastic trend ($\eta = 0$) exhibit random fluctuations about a mean. However, as η increases, the trend behavior becomes more and more apparent. As η becomes very large (see Figure 8.2d), the stochastic trend becomes predominant and the series wanders smoothly over a wide range of values.

The estimates of η selected by empirical Bayes are similar to those used to generate the artificial data and the resulting fitted trends are quite sensible. In Figure 8.2a, where there is no trend, the fitted stochastic trend is almost nonexistent (i.e. it is close to simply being a horizontal line). In Figure 8.2d, where the trend predominates, the fitted stochastic trend matches the data very closely (indeed it is hard to see the difference between the two lines in Figure 8.2d). Figures 8.2b and 8.2c present intermediate cases.

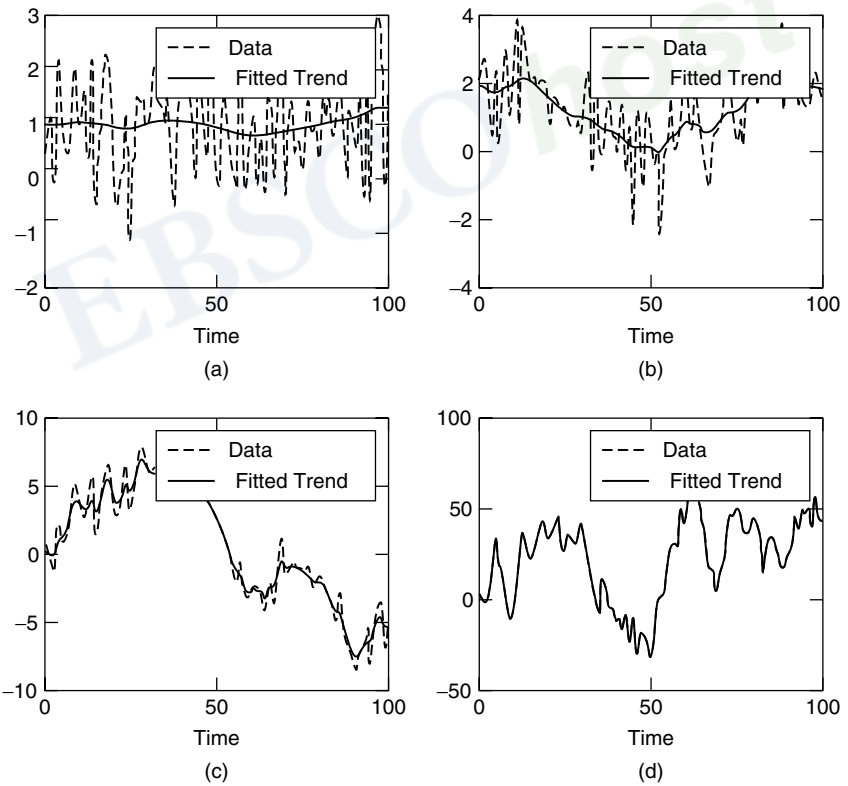

Figure 8.2 (a) Data Set with $\eta = 0$; (b) Data Set with $\eta = 0.1$; (c) Data Set with $\eta = 1$; (d) Data Set with $\eta = 100$

8.3 A GENERAL STATE SPACE MODEL

In this section, we discuss a more general state space model which we will simply refer to it as the *state space model* and write as

$$y_t = X_t\beta + Z_t\alpha_t + \varepsilon_t \tag{8.22}$$

and

$$\alpha_{t+1} = T_t\alpha_t + u_t \tag{8.23}$$

This model uses slightly different notation from the local level model, in that we allow α_t to be a $p \times 1$ vector containing p state equations. We assume ε_t to be i.i.d. $N(0, h^{-1})$, but u_t is now a $p \times 1$ vector which is i.i.d. $N(0, H^{-1})$ and ε_t and u_s are independent of one another for all s and t. X_t and Z_t are $1 \times k$ and $1 \times p$ vectors, respectively, containing explanatory variables and/or known constants. T_t is a $p \times p$ matrix of known constants. The case where T_t contains unknown parameters can be handled in a straightforward fashion, as noted below.

This state space model is not the most general possible (see the next section for a discussion of extensions), but it does encompass a wide variety of models. To understand the types of behavior the state space model allows for, it is useful to discuss several special cases. First, the local level model is a special case of (8.22) and (8.23) if $p = 1$, $k = 0$, $T_t = 1$ and $Z_t = 1$ and, thus, this model can be used decompose a time series into a stochastic trend and irregular component. Secondly, (8.22) can reduce to a Normal linear regression model of the sort considered in Chapters 3 and 4 if $Z_t = 0$. Thirdly, it can reduce to a Normal linear regression model with time varying parameters if Z_t contains some or all of the explanatory variables. Fourthly, there are many so-called *structural time series models* which can be put in the form of (8.22) and (8.23). The reader is referred to Durbin and Koopman (2001, Chapter 3) for a discussion of such models, including issues such as seasonality, and how commonly-used Autoregressive Integrated Moving Average (or ARIMA) models can be put in state space form. Here we will show how one common structural time series model referred to as the *local linear trend model* can be put in state space form. This model is similar to the local level model, but allows the trend to evolve over time. Thus,

$$y_t = \mu_t + \varepsilon_t$$

$$\mu_{t+1} = \mu_t + \upsilon_t + \xi_t$$

and

$$\upsilon_{t+1} = \upsilon_t + \zeta_t$$

where ξ_t is i.i.d. $N(0, \sigma_\xi^2)$, ζ_t is i.i.d. $N(0, \sigma_\zeta^2)$ and all the errors are independent of one another. It can be seen that this local linear trend model can be put in the

form of the state space model by setting

$$\alpha_t = \begin{pmatrix} \mu_t \\ v_t \end{pmatrix}$$

$$u_t = \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix}$$

$$T_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$Z_t = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$H = \begin{pmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{pmatrix}$$

and $\beta = 0$. In short, a wide variety of useful regression and time series models can be written as state space models.

8.3.1 Bayesian Computation in the State Space Model

We have stressed throughout this book that an advantage of Bayesian inference is that it is often modular in nature. Methods for posterior computation in many complicated models can be developed by simply combining results from simpler models. The state space model is a good example of how this can be done. Hence, rather than go through the steps of writing out the likelihood, prior and posterior, we jump straight to the issue of Bayesian computation, and show how we can draw on results from earlier chapters to carry out Bayesian inference in this model. As we shall see, a complication for posterior simulation arises since the posterior conditional for α analogous to Chapter 7 (7.17) will not be independent across time (i.e. (8.23) implies that α_t and α_{t-1} will not be independent of one another). Thus, we cannot easily draw from the α_ts one at a time and a direct implementation of the Gibbs sampler would involve drawing from a T-dimensional Normal distribution. In general, this can be a bit slow, but De Jong and Shephard (1995) describe an efficient method for Gibbs sampling in this class of models.

An examination of (8.22) reveals that, if α_t for $t = 1, \ldots, T$ were known (as opposed to being unobserved), then the state space model would reduce to a Normal linear regression model:

$$y_t^* = X_t\beta + \varepsilon_t$$

where $y_t^* = y_t - Z_t\alpha_t$. Thus, all the results of previous chapters for the Normal linear regression model could be used, except the dependent variable would be y_t^* instead of y. This suggests that a Gibbs sampler with data augmentation can be set up for the state space model. That is, depending on the prior chosen, $p(\beta, h|y, \alpha_1, \ldots, \alpha_T)$ will have one of the simple forms given in Chapters 3 or 4. Similarly, if α_t for $t = 1, \ldots, T$ were known then the state equations given in

(8.23) are a simple variant of the Seemingly Unrelated Regression (SUR) linear regression model discussed in Chapter 6 (Section 6.6) and $p(H|y, \alpha_1, \ldots, \alpha_T)$ has a familiar form.[8] Thus, if we can derive a method for taking random draws from $p(\alpha_1, \ldots, \alpha_T|y, \beta, h, H)$ then we have completely specified a Gibbs sampler with data augmentation which allows for Bayesian inference in the state space model. In the following material, we develop such a Gibbs sampler for a particular prior choice, but we stress that other priors can be used with minor modifications.

Here we will use an independent Normal-Gamma prior for β and h, a Wishart prior for H and the prior implied by the state equation for $\alpha_1, \ldots, \alpha_T$. In particular, we assume a prior of the form

$$p(\beta, h, H, \alpha_1, \ldots, \alpha_T) = p(\beta)p(h)p(H)p(\alpha_1, \ldots, \alpha_T|H)$$

where

$$p(\beta) = f_N(\beta|\underline{\beta}, \underline{V}) \tag{8.24}$$

$$p(h) = f_G(h|\underline{s}^{-2}, \underline{\nu}) \tag{8.25}$$

and

$$p(H) = f_W(H|\underline{\nu}_H, \underline{H}) \tag{8.26}$$

For the elements of the state vector we treat (8.23) as a hierarchical prior. If we treat the time index for (8.23) as beginning at 0 (i.e. $t = 0, 1, \ldots, T$) and assume $\alpha_0 = 0$, then the state equation even provides a prior for the initial condition. Formally, this amounts to writing this part of the prior as

$$p(\alpha_1, \ldots, \alpha_T|H) = p(\alpha_1|H)p(\alpha_2|\alpha_1, H) \ldots p(\alpha_T|\alpha_{T-1}, H)$$

where, for $t = 1, \ldots, T-1$

$$p(\alpha_{t+1}|\alpha_t, H) = f_N(\alpha_{t+1}|T_t\alpha_t, H) \tag{8.27}$$

and

$$p(\alpha_1|H) = f_N(\alpha_1|0, H) \tag{8.28}$$

Note that H is playing a similar role to η in the local level model. However, H is a $p \times p$ matrix, so it would be difficult to use empirical Bayes methods with this high dimensional model. Furthermore, we are no longer using a natural conjugate prior so that the analytical results of Section 8.2 no longer hold.

The reasoning above suggests that our end goal is a Gibbs sampler with data augmentation which sequentially draws from $p(\beta|y, \alpha_1, \ldots, \alpha_T)$, $p(h|y, \alpha_1, \ldots, \alpha_T)$, $p(H|y, \alpha_1, \ldots, \alpha_T)$ and $p(\alpha_1, \ldots, \alpha_T|y, \beta, h, H)$. The first three of these posterior conditional distributions can be dealt with by using

[8]The case where T_t contains unknown parameters would involve drawing from $p(H, T_1, \ldots, T_T|y, \alpha_1, \ldots, \alpha_T)$ which can usually be done fairly easily. In the common time-invariant case where $T_1 = \cdots = T_T$, $p(H, T_1, \ldots, T_T|y, \alpha_1, \ldots, \alpha_T)$ will have precisely the form of a SUR model.

results from previous chapters. In particular, from Chapter 4 (Section 4.2.2) we find

$$\beta|y, h, \alpha_1, \ldots, \alpha_T \sim N(\overline{\beta}, \overline{V}) \tag{8.29}$$

and

$$h|y, \beta, \alpha_1, \ldots, \alpha_T \sim G(\overline{s}^{-2}, \overline{\nu}) \tag{8.30}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + h\sum_{t=1}^{T} X_t' X_t\right)^{-1} \tag{8.31}$$

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + h\sum_{t=1}^{T} X_t'(y_t - Z_t\alpha_t)\right) \tag{8.32}$$

$$\overline{\nu} = T + \underline{\nu} \tag{8.33}$$

and

$$\overline{s}^2 = \frac{\sum_{t=1}^{T}(y_t - X_t\beta - Z_t\alpha_t)^2 + \underline{\nu}\underline{s}^2}{\overline{\nu}} \tag{8.34}$$

Using results for the SUR model (with no explanatory variables) from Chapter 6 (Section 6.6.3) we obtain

$$H|y, \alpha_1, \ldots, \alpha_T \sim W(\overline{\nu}_H, \overline{H}) \tag{8.35}$$

where

$$\overline{\nu}_H = T + \underline{\nu}_H \tag{8.36}$$

and

$$\overline{H} = \left[\underline{H}^{-1} + \sum_{t=0}^{T-1}(\alpha_{t+1} - T_t\alpha_t)(\alpha_{t+1} - T_t\alpha_t)'\right]^{-1} \tag{8.37}$$

To complete our Gibbs sampler, we need to derive $p(\alpha_1, \ldots, \alpha_T|y, \beta, h, H)$ and a means of drawing from it. Although it is not hard to write out this multivariate Normal distribution, it can be hard to draw from it in practice since it is T-dimensional, and its elements can be highly correlated with one another. Accordingly, there have been many statistical papers which seek to find efficient ways of drawing from this distribution (Carter and Kohn (1994) and DeJong and Shephard (1995) are two prominent contributions to this literature). Here we present the method described in DeJong and Shephard (1995), which has been found to work very well in many applications. The reader interested in proofs

and derivations can look at this paper. DeJong and Shephard (1995) works with a slightly more general version of our state space model, written as

$$y_t = X_t\beta + Z_t\alpha_t + G_t v_t \tag{8.38}$$

and

$$\alpha_{t+1} = T_t\alpha_t + J_t v_t \tag{8.39}$$

for $t = 1, \ldots, T$ in (8.38) and $t = 0, \ldots, T$ in (8.39) and $\alpha_0 = 0$. v_t is i.i.d. $N(0, h^{-1}I_{p+1})$. Other variables and parameters are as defined for our state space model. It can be see that our state space model is equivalent to the one given in (8.38) and (8.39) if we set

$$v_t = \begin{pmatrix} \varepsilon_t \\ u_t \end{pmatrix}$$

G_t to be a $(p+1)$ row vector given by

$$G_t = (\ 1 \quad 0 \quad . \quad . \quad 0\ )$$

and J_t to be a $p \times (p+1)$ matrix given by

$$J_t[\ 0_p \quad A\ ]$$

where A is a $p \times p$ matrix implicitly defined by

$$H^{-1} = \frac{1}{h}AA'$$

Since the Gibbs sampler involves drawing from $p(\alpha_1, \ldots, \alpha_T | y, \beta, h, H)$, everything in (8.38) and (8.39) except for α_t and v_t can be treated as known. The contribution of DeJong and Shephard (1995)[9] was to develop an efficient algorithm for drawing from $\eta_t = F_t v_t$ for various choices of F_t. Draws from η_t can then be transformed into draws from α_t. We set out their algorithm for arbitrary F_t, but note that the usual choice is to set $F_t = J_t$, as this yields draws from the state equation errors which can be directly transformed into the required draws from α_t.

DeJong and Shephard (1995) refer to their algorithm as the *simulation smoother*. The simulation smoother begins by setting $a_1 = 0$, $P_1 = J_0 J_0'$ and calculating for $t = 1, \ldots, T$ the quantities:[10]

$$e_t = y_t - X_t\beta - Z_t a_t \tag{8.40}$$

$$D_t = Z_t P_t Z_t' + G_t G_t' \tag{8.41}$$

$$K_t = (T_t P_t Z_t' + J_t G_t') D_t' \tag{8.42}$$

$$a_{t+1} = T_t a_t + K_t e_t \tag{8.43}$$

[9]There are other advantages of the algorithm proposed by DeJong and Shephard (1995) involving computer storage requirements and avoiding certain degeneracies which will not be discussed here.

[10]For readers with some knowledge of the state space literature, these calculations are referred to as running the *Kalman filter*.

and

$$P_{t+1} = T_t P_t (T_t - K_t Z_t)' + J_t (J_t - K_t G_t) \tag{8.44}$$

and storing the quantities e_t, D_t and K_t. Then a new set of quantities are calculated in reverse time order (i.e. $t = T, T-1, \ldots, 1$). These begin by setting $r_T = 0$ and $U_T = 0$, and then calculating

$$C_t = F_t (I - G_t' D_t^{-1} G_t - [J_t - K_t G_t]' U_t [J_t - K_t G_t]) F_t' \tag{8.45}$$

$$\xi_t \sim N(0, h^{-1} C_t) \tag{8.46}$$

$$V_t = F_t (G_t' D_t^{-1} Z_t + [J_t - K_t G_t]' U_t [T_t - K_t Z_t]') \tag{8.47}$$

$$r_{t-1} = Z_t' D_t^{-1} e_t + (T_t - K_t Z_t)' r_t - V_t' C_t^{-1} \xi_t \tag{8.48}$$

$$U_{t-1} = Z_t' D_t^{-1} Z_t + (T_t - K_t Z_t)' U_t (T_t - K_t Z_t) + V_t' C_t^{-1} V_t \tag{8.49}$$

and

$$\eta_t = F_t (G_t' D_t^{-1} e_t + [J_t - K_t G_t]' r_t) + \xi_t \tag{8.50}$$

where $G_0 = 0$. This algorithm will yield $\eta = (\eta_0, \ldots, \eta_T)'$, and it can be proved that this is a random draw from $p(\eta | y, \beta, h, H)$. Depending on the form for F_t, this can be transformed into the required random draw of α_t to $t = 1, \ldots, T$. For the common choice of $F_t = J_t$, this algorithm provides draws from the errors in the state equation (i.e. $\eta_t = J_t v_t$) which can be transformed into draws from α_t using (8.39) and the fact that $\alpha_0 = 0$.

These formulae may look complicated. However, the algorithm is simply a series of a calculations involving matrices that are of low dimension plus random draws from the Normal distribution to get ξ_t. This greatly speeds up computation since manipulating high dimensional (e.g. $T \times T$) matrices is very slow indeed. Furthermore, for most applications the matrices F_t, G_t, J_t and T_t will have simple forms, and thus the previous equations will simplify. Thus, with a bit of care, programming up this component of the Gibbs sampler is a straightforward task.

In summary, a Gibbs sampler with data augmentation which sequentially draws from $p(\beta | y, \alpha_1, \ldots, \alpha_T)$, $p(h | y, \alpha_1, \ldots, \alpha_T)$, $p(H | y, \alpha_1, \ldots, \alpha_T)$ and $p(\alpha_1, \ldots, \alpha_T | y, \beta, h, H)$ has been derived using results from previous chapters along with an algorithm developed in DeJong and Shephard (1995). Given output from such a posterior simulator, posterior inference can be carried out as in previous chapters (see Chapter 4, Sections 4.2.3 and 4.2.4). Predictive inference in this model can be carried out using the strategy outlined in Chapter 4, Section 4.2.6. Posterior predictive p-values or HPDIs can be calculated to shed light on the fit and appropriateness of the model. The marginal likelihood for the state space model can be calculated using the method of Chib (see Chapter 7, Section 7.5). The implementation of the Chib method is similar to that described for the individual effects model of Chapter 7 with $\alpha_1, \ldots, \alpha_T$ being treated as latent data.

8.3.2 Empirical Illustration: The State Space Model

To illustrate Bayesian methods in the state space model, we use one of the data sets and some of the models analyzed in Koop and Potter (2001). The data set has been used by economic historians interested in epochs such as the industrial revolution and the Great Depression (e.g. see Greasley and Oxley, 1994). It consists of the annual percentage change in UK industrial production from 1701 to 1992. There are many questions of interest which can be investigated with this data set. In this empirical illustration, we will focus on the question of whether the basic structure of the time series model driving the growth in industrial production is changing over time. To this end we consider an AR(p) model with time varying coefficients:

$$y_t = \alpha_{0t} + \alpha_{1t} y_{t-1} + \cdots + \alpha_{pt} y_{t-p} + \varepsilon_t \tag{8.51}$$

where for $i = 0, \ldots, p$

$$\alpha_{it+1} = \alpha_{i,t} + u_{it} \tag{8.52}$$

We assume ε_t to be i.i.d. $N(0, h^{-1})$ and u_{it} to i.i.d. $N(0, \lambda_i h^{-1})$ where ε_t, u_{is} and u_{jr} are independent of one another for all s, t, r, i and j. In words, this is an autoregressive model, but the autoregressive coefficients (and the intercept) may be gradually evolving over time. It can be seen that this model is a special case of the state space model in (8.22) and (8.23) if we exclude X_t, and define

$$\alpha_t = \begin{pmatrix} \alpha_{0t} \\ \alpha_{1t} \\ \cdot \\ \cdot \\ \alpha_{pt} \end{pmatrix}$$

$$u_t = \begin{pmatrix} u_{0t} \\ u_{1t} \\ \cdot \\ \cdot \\ u_{pt} \end{pmatrix}$$

$$Z_t = \begin{pmatrix} 1 & y_{t-1} & \cdot & \cdot & y_{t-p} \end{pmatrix}$$

and set $T_t = I_{p+1}$ and

$$H^{-1} = h^{-1} \begin{bmatrix} \lambda_0 & 0 & 0 & . & 0 \\ 0 & \lambda_1 & . & . & . \\ . & 0 & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & \lambda_p \end{bmatrix}$$

We choose $p = 1$ to illustrate results for this model, although our previous work with this data set indicates larger values of p should be used in a serious piece of empirical work. To simplify issues relating to initial conditions, for our dependent

variable we use data beginning in 1705. This means that the y_{t-4} term in (8.51) will always be observed.

We use an informative prior for the parameters h and λ_i for $i = 0, \ldots, p$.[11] For h we use the Gamma prior from (8.25) with $\underline{\nu} = 1$ and $\underline{s}^{-2} = 1$. Since the data is measured as a percentage change, the prior for h is centered over a value which implies over 95% of the errors are less than 2%. However it is relatively noninformative, since the prior contains the same information as one data point (i.e. $\underline{\nu} = 1$). Note that if H were not a diagonal matrix we would probably want to use a Wishart prior for it, here we have assumed the state equations to have errors which are uncorrelated with one another and, hence, we only need elicit $p+1$ univariate priors for the λ_is. Thus, the Wishart prior for H given in (8.36) simplifies here to

$$p(\lambda_i^{-1}) = f_G(\lambda_i^{-1} | \underline{\lambda}_i^{-1}, \underline{\nu}_i)$$

for $i = 0, \ldots, p$. We choose the relatively noninformative values of $\underline{\nu}_i = 1$ for all i, but center the prior for λ_i over 1 by setting $\underline{\lambda}_i = 1$. Since $AR(p)$ coefficients tend to be quite small (e.g. in the stationary AR(1) case the coefficient is less than one in absolute value), this prior allows for fairly substantive changes in the coefficients over time. With this prior, the conditional posterior for H given in (8.35) simplifies to

$$p(\lambda_i^{-1} | y, \alpha_1, \ldots, \alpha_T) = f_G(\lambda_i^{-1} | \overline{\lambda}_i^{-1}, \overline{\nu}_i)$$

for $i = 0, \ldots, p$, where

$$\overline{\nu}_i = T + \underline{\nu}_i$$

and

$$\overline{\lambda}_i = \frac{h \sum_{t=0}^{T-1} (\alpha_{i,t+1} - \alpha_{it})(\alpha_{i,t+1} - \alpha_t)' + \underline{\nu}_i \underline{\lambda}_i}{\overline{\nu}_i}$$

Table 8.1 contains posterior results for the state space model using the data and prior discussed above. The Gibbs sampler was run for 21 000 replications, with 1000 burn-in replications discarded and 20 000 replications retained. The last column of Table 8.1 presents Geweke's convergence diagnostic which indicates that convergence of the posterior simulator has been achieved. Posterior means and standard deviations for λ_0 and λ_1 indicate that a substantial amount of parameter variation occurs both in the intercept and the AR(1) coefficient. Thus, in addition to there being a stochastic trend in the growth in industrial

[11] Note that we are not using the natural conjugate prior and, hence, the results relating to noninformative prior which we derived for the local level model do not apply here. The results of Fernandez, Osiewalski and Steel (1997) are relevant, and imply that a proper prior is required for these parameters if we are to obtain a proper posterior.

Table 8.1 Posterior Results for State Space Model

	Mean	Stand. Dev.	Geweke's CD
h	0.17	0.04	0.99
λ_0	0.93	0.16	0.28
λ_1	0.61	0.11	0.64

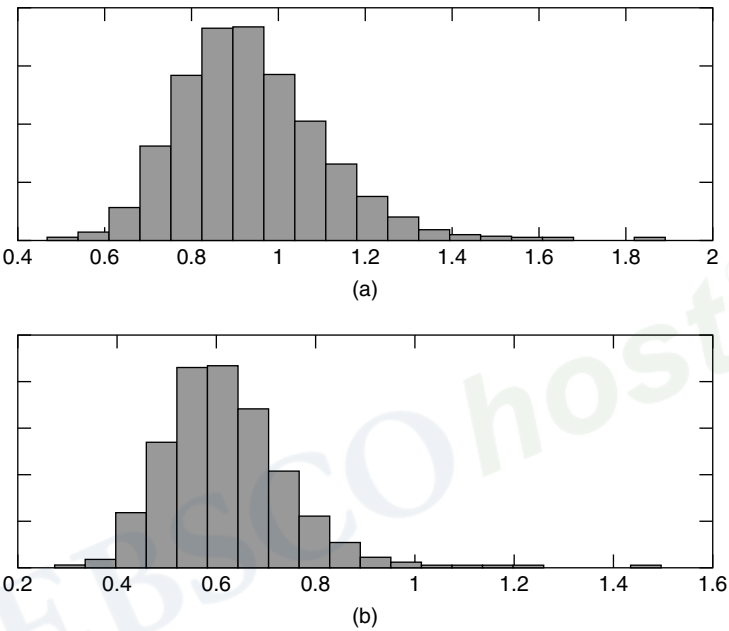

Figure 8.3 (a) Posterior Density for λ_0; (b) Posterior Density for λ_1

production, the AR process itself is changing over time. These findings are supported by an examination of Figures 8.3a and 8.3b which plot the entire posterior densities of each of these parameters.[12]

8.4 EXTENSIONS

The state space model introduced in (8.22) and (8.23) covers a wide range of interesting time series models (e.g. the local linear trend model, time varying parameter models, models with seasonality, etc.). However, there are numerous

[12]We stress that this is only an illustration of Bayesian methods in the state space models and should not necessarily be taken to imply a particular model for industrial production. A serious piece of empirical work involving this time series would involve a consideration of other models (e.g. a model with a structural break).

extensions of this model that can be made in a straightforward manner. Some of these extensions can be handled using methods outlined in this book. For instance, we have discussed the *Normal* state space model, but extending this to the *Student-t* state space model can be done by adding one block to the Gibbs sampler. That is, the Gibbs sampler of Chapter 6, Section 6.4 for handling the linear regression model with Student-t errors can be combined with the Gibbs sampler developed in this chapter.

Other important extensions cannot be directly handled using the methods outlined in this book, but a small amount of additional reading or thought would suffice to create methods for Bayesian inference in these models. Examples of this sort include various nonlinear, regime shifting or structural break models (see Kim and Nelson (1999), Bauwens, Lubrano and Richard (1999) or Chapter 12, Section 12.4.1 for a discussion of some of these models). In finance, a particularly important model, known as the *stochastic volatility model*, is a modest extension of that discussed in this chapter. The stochastic volatility model has time varying error variances of the sort that seem to occur in stock returns and many other financial time series (see Jacquier, Polson and Rossi (1994) for the first Bayesian work on this model). It can be written as

$$y_t = \varepsilon_t$$

where ε_t is i.i.d. $N(0, \sigma_t^2)$

$$\log(\sigma_t^2) = \log(\sigma^2) + \beta \log(\sigma_{t-1}^2) + u_t$$

where u_t is i.i.d. $N(0, 1)$. If we define $\alpha_t \equiv \log(\sigma_t^2)$, then it can be seen that this is a state space model where the state equation relates to the conditional variance of the error as opposed to the conditional mean. A Gibbs sampler which can be used to carry out Bayesian inference in this model is very similar to the one we have developed in this chapter. In fact, the algorithm of DeJong and Shephard (1995) can be directly used to draw from the posterior of the states, α_t, conditional on the other parameters in the model. Hence, all that is required is an algorithm to draw from the other parameters conditional on the states. But this is relatively straightforward (see DeJong and Shephard, 1995, for details).

Perhaps the most important extension of the state space model discussed in this chapter is to allow y_t to be a vector instead of a scalar. After all, economists are usually interested in multivariate relationships between time series. This extension is very easy since (8.22) can be re-interpreted with y_t being a q-vector containing q different time series and very little will change in the development of the posterior simulator. In fact, the DeJong and Shephard (1995) model and algorithm in (8.38)–(8.50) has been deliberately written so as to hold in the multivariate case. Thus, these equations can be used to draw from $p(\alpha_1, \ldots, \alpha_T | y, \beta, h, H)$ in a posterior simulator. $p(H | y, \alpha_1, \ldots, \alpha_T)$ will be similarly unaffected by the move from a univariate to a multivariate state space model. A Gibbs sampler for the multivariate model can be completed by drawing on methods for the SUR model (see Chapter 6, Section 6.6.3) to derive $p(\beta | y, \alpha_1, \ldots, \alpha_T)$ and

$p(h|y, \alpha_1, \ldots, \alpha_T)$. Thus, Bayesian inference in the multivariate state space model involves only minor alterations of the methods of Section 8.3.

As an aside, we should mention that multivariate state space models can be used to investigate the presence of *cointegration* in time series. Cointegration is an important concept in empirical macroeconomics and relates to the number of stochastic trends which are present in a set of time series. To motivate this concept further, consider the following multivariate state space model:

$$y_t = Z_t\alpha_t + \varepsilon_t \tag{8.53}$$

and

$$\alpha_{t+1} = \alpha_t + u_t \tag{8.54}$$

where y_t is a q-vector of q time series and α_t is a p-vector of state equations and H is a diagonal matrix. If $p = q$ and $Z_t = I_q$, then this becomes a multivariate local level model. In this case, each of the q time series contains a stochastic trend. That is, we can write

$$y_{it} = \alpha_{it} + \varepsilon_{it}$$

and

$$\alpha_{i,t+1} = \alpha_{it} + u_{it}$$

for $i = 1, \ldots, q$ and each individual time series follows a local level model. There are q independent stochastic trends driving the q time series.

Consider what happens, however, if $p < q$ and Z_t is a $q \times p$ matrix of constants. In this case, there are p stochastic trends driving the q time series. Since there are fewer stochastic trends than time series, some of the trends must be common to more than one time series. For this reason, if $p < q$ this model is referred to as a *common trends model*. Other ways of expressing this common trend behavior is to say the q time series are trending together or *co-trending* or *cointegrated*.

The macroeconometric literature on cointegration is very voluminous; suffice it to note here that cointegration is thought to hold in many economic time series. That is, many economic time series seem to exhibit stochastic trend behavior. However, economic theory also suggests many time series variables should be related through equilibrium concepts. In practice, these two considerations suggest cointegration should occur. Suppose, for instance, that y_{1t} and y_{2t} are two time series which should be equal to one another in equilibrium. In reality, we expect perturbations and random shocks to imply that equilibrium is rarely, if ever, perfectly achieved. A bivariate local level model with a single stochastic trend would fit with this theoretical expectation. That is, both time series would exhibit stochastic trend behavior. Furthermore, (8.53) and (8.54) with $q = 2$, $p = 1$ and $Z_t = 1$ can be written as

$$y_{1t} = y_{2t} + (\varepsilon_{1t} - \varepsilon_{2t})$$

Hence, $y_{1t} = y_{2t}$ apart from a random equilibrium error $(\varepsilon_{1t} - \varepsilon_{2t})$. Thus, it can be argued that cointegration is how macroeconomic equilibrium concepts should manifest themselves empirically. Economic theories used to justify cointegration include purchasing power parity, the permanent income hypothesis and various theories of money demand and asset pricing.

Cointegration is, thus, a potentially important thing to look for in a wide variety of applications in macroeconomics and finance. In multivariate state space models the number of common trends can be directly investigated by comparing models with different numbers of state equations. For instance, a researcher could calculate the marginal likelihood for the model described in (8.53) and (8.54) for various values of p. If substantial posterior probability is assigned to models with $p < q$, then the researcher could conclude that evidence exists for cointegration.

8.5 SUMMARY

In this chapter, we have introduced a fairly general state space model along with an interesting special case referred to as the local level model. State space models are commonly used when working with time series data and are suitable for modeling a wide variety of behaviors (e.g. trending, cycling or seasonal). State space models are especially attractive for the Bayesian since they can be interpreted as flexible models with hierarchical priors. Thus, the interpretation and computational methods are similar to those for other models such as the individual effects or random coefficients panel data models.

For the local level model, we used a natural conjugate prior and showed how this allowed for analytical results. We introduced a new method for prior elicitation referred to as empirical Bayes. This method, which is especially popular in models involving hierarchical priors, chooses as values for prior hyperparameters those which maximize the marginal likelihood. Such an approach allows the researcher to avoid subjective elicitation of prior hyperparameters or using a noninformative prior. An empirical illustration involving artificial data showed how the empirical Bayes approach could be implemented in practice.

For the more general state space model, we used an independent Normal-Gamma prior and showed how this, along with a hierarchical prior defined by the state equation, meant that a posterior simulator was required. Such a posterior simulator was developed by combining results for the Normal linear regression model, and the SUR model along with a method developed in DeJong and Shephard (1995) for drawing from the states (i.e. α_t for $t = 1, \ldots, T$). The state space model thus provides a good illustration of the modular nature of Bayesian computation where model extensions often simply involve adding a new block to a Gibbs sampler. An application of interest to economic historians, involving a long time series of industrial production and an AR(p) model with time varying coefficients, was used to illustrate Bayesian inference in the state space model.

The chapter ended with a discussion of several possible extensions to the state space models considered in this chapter. Of particular importance were the stochastic volatility model and multivariate state space models. The former of these extensions is commonly used with financial time series while the latter can be used in macroeconomic applications to investigate cointegration and related issues. We stressed how Bayesian analysis of these extensions can be implemented through minor modifications of the posterior simulator described in Section 8.3.

Time series econometrics is such a huge field that a single chapter such as the present one necessarily skips over many important issues. Chapter 12 (Section 12.4.1) provides a brief discussion of some additional time series topics.

8.6 EXERCISES

The exercises in this chapter are closer to being small projects than standard textbook questions. Remember that some data sets and MATLAB programs are available on the website associated with this book.

1. (a) Use the derivations in Section 8.2 and Chapter 6, Section 6.4 to obtain a posterior simulator for the local level model with independent Student-t errors. You may use whatever prior you wish for the model parameters (although the natural conjugate one of Section 6.4 will be the easiest).
   (b) Write a program which uses your result from part (a) and test the program on either real data (e.g. the industrial production data from the empirical illustration in Section 8.3.2) or artificial data generated according to various data generating processes.
   (c) Add to your program code for calculating the marginal likelihood for the local level model with independent Student-t errors. For your data set(s) calculate the Bayes factor comparing the local level model with Normal errors to the local level model with Student-t errors.
2. *Unit root testing with the local level model.* To do this exercise, use either real data (e.g. the industrial production data from the empirical illustration in Section 8.3.2) or artificial data generated according to various data generating processes of your choice. Use the local level model with natural conjugate Normal-Gamma prior described in Section 8.2 with the second variant on the empirical Bayesian methodology described in Section 8.2.3. That is, treat η as an unknown parameter, choose a prior for η of your choice (e.g. a Gamma or Uniform prior) and obtain $p(\eta|y)$. Remember that a unit root is present in the model $M_1 : \eta > 0$, but is not present in the model $M_2 : \eta = 0$. You want to calculate the Bayes factor comparing M_1 to M_2.
   (a) Derive the formula for the marginal likelihood of M_2.
   (b) Using your result for part (a), write a program for calculating the required Bayes factor and test the program using your data set(s).

(c) Consider an approximate strategy where you calculate the Bayes factor comparing M_1 to $M_2^*: \eta = a$ where a is a very small number. Using the Savage–Dickey density ratio, derive a formula for calculating the Bayes factor comparing M_1 to M_2^*.

(d) For your data set(s) compare the approximate Bayes factor of part (c) to that obtained in part (b) for various values of a (e.g. $a = 0.01$, 0.0001, 0.0000001, etc.). How well does the approximate strategy work?

3. Use the methods of Section 8.3 for the general state space model to answer this question. Use either real data (e.g. the industrial production data from the empirical illustration in Section 8.3.2) or artificial data generated according to various data generating processes of your choice.

(a) Write a program which carries out posterior simulation in the local level model described in Section 8.3.

(b) Write a program which carries out posterior simulation in the local linear trend model described in Section 8.3.

(c) Test the programs you have written in parts (a) and (b) using your data set(s).

(d) Modify your programs to calculate the marginal likelihood for each model and, hence, calculate the Bayes factor comparing the local level model to local linear trend model and test your program using your data set(s). Remember that meaningful marginal likelihoods can only be calculated with informative priors and, hence, choose informative priors of your choice.

(e) Carry out a prior sensitivity analysis to investigate which aspects of prior elicitation seem to be most important for model comparison involving the local level and local linear trend models.

This Page Intentionally Left Blank