11

Bayesian Time Series and DSGE Models

This chapter covers Bayesian estimation of three popular time series models and returns to the main goal of this book: estimation and inference in DSGE models, this time from a Bayesian perspective. All three types of time series model have a latent variable structure: the data y_t depend on a latent variable x_t and on a vector of parameters α, and the latent variable x_t is a function of another set of parameters θ. In factor models, x_t is a common factor or a common trend; in stochastic volatility models, x_t is a vector of volatilities; and in Markov switching models, x_t is an unobservable finite-state process. While, for the first and the third types of model, classical methods to evaluate the likelihood function are available (see, for example, Sims and Sargent 1977; Hamilton 1989), for the second type, approximations based on either a method of moments or quasi-ML are typically used. Approximations are needed because the density of the observables $f(y \mid \alpha, \theta)$ is a mixture of distributions, that is, $f(y \mid \alpha, \theta) = \int f(y \mid x, \alpha) f(x \mid \theta) \, \mathrm{d}x$. Since the computation of the likelihood function requires a T-dimensional integral, no analytical solution is generally available.

As mentioned in chapter 9, the model for x_t can be interpreted either as a prior or as a description of how the latent variable evolves. This means that all three models have a hierarchical structure which can be handled with the "data-augmentation" technique of Tanner and Wong (1987). Such a technique treats $x = (x_1, \ldots, x_T)$ as a vector of parameters for which we have to compute the conditional posterior — as we have done with the time-varying parameters of a TVC model in chapter 10. Cyclical sampling across the conditional distributions provides, in the limit, posterior draws for the parameters and the unobservable x. The Markov property for x_t is useful to simplify the calculations since we can break the problem of simulating the x vector into the problem of simulating its components in a conditional recursive fashion. For the models we examine, the likelihood is bounded. Therefore, if the priors are proper, the transition kernel induced by the Gibbs sampler (or by the mixed Gibbs–MH sampler) is irreducible, aperiodic, and has an invariant distribution. Hence, sufficient conditions for convergence hold in these setups.

The kernel of (x, α, θ) is the product of the conditional distribution of $(y \mid x, \alpha)$, the conditional distribution of $(x \mid \theta)$, and the prior for (α, θ). Hence, $g(\alpha, \theta \mid y) = \int g(x, \alpha, \theta \mid y) \, \mathrm{d}x$ can be used for inference, while $g(x \mid y)$ provides a solution to the problem of estimating x. The main difference between the setup of this chapter

and a traditional signal extraction problem is that here we produce the distribution of x at each t, not just its conditional mean. It is also important to emphasize that, contrary to classical methods, the tools we describe allow the computation of the exact posterior distribution of x. Therefore, we are able to describe posterior uncertainty surrounding the latent variable and the parameters.

Forecasting $y_{t+\tau}$ and the latent variable $x_{t+\tau}$ is straightforward and can be handled with the tools described in chapter 9. Since many inferential exercises have to do with the problem of obtaining a future measure of the unobserved state (the business cycle in policy circles, the volatility process in business and finance circles, etc.), it is important to have ways to estimate it. Draws for future $x_{t+\tau}$ can be obtained from the marginal posterior of x and the structure of its conditional.

Although this chapter primarily focuses on models with normal errors, more heavy-tailed distributions could also be used, particularly, in finance applications. As in the case of state space models, such an extension presents few complications.

The last section of the chapter studies how to obtain posterior estimates of the structural parameters of DSGE models, how to conduct posterior inference and model comparisons, and reexamines the link between DSGE models and VARs. There is very little new material in this section: we bring together the models discussed in chapter 2 and the ideas contained in chapters 5–7 with the simulation techniques presented in chapter 9 to develop a framework where structural inference can be conducted in false models, taking both parameter and model uncertainty into consideration.

11.1 Factor Models

Factor models are used in many fields of economics and finance. They exploit the insight that there may be a common source of fluctuations in a vector of economic time series. Factor models are therefore alternatives to the (panel) VAR models analyzed in chapter 10. In the latter, detailed cross-variable interdependencies are modeled but no common factor is explicitly considered. Here, most of interdependencies are eschewed and a low-dimensional vector of unobservable variables is assumed to drive the comovements across variables. Clearly, combinations of the two approaches are possible (see, for example, Bernanke et al. 2005; Giannone et al. 2003). The factor structure we consider is

$$\left.\begin{aligned} y_{it} &= \bar{y}_i + \mathbb{Q}_i y_{0t} + e_{it}, \\ A_i^e(\ell) e_{it} &= v_{it}, \\ A^y(\ell) y_{0t} &= v_{0t}, \end{aligned}\right\} \tag{11.1}$$

where $E(v_{it}, v_{i't-\tau}) = 0, \forall i \neq i', i = 1, \ldots, m$, $E(v_{it}, v_{it-\tau}) = \sigma_i^2$ if $\tau = 0$ and zero otherwise, $E(v_{0t}, v_{0t-\tau}) = \sigma_0^2$ if $\tau = 0$ and zero otherwise, and y_{0t} is unobservable. Two features of (11.1) need to be noted. First, the unobservable factor can have arbitrary serial correlation. Second, since the relationship between observables and unobservables is static, e_{it} is allowed to be serially correlated. y_{0t}

could be a scalar or a vector, as long as its dimension is smaller than the dimension of y_t. An interesting case emerges when $e_t = (e_{1t}, \ldots, e_{mt})'$ follows a VAR, i.e., $A^e(\ell)e_t = v_t$, and $A^e(\ell)$ is of order q_e, $\forall i$.

Example 11.1. There are several specifications which fit into this framework. For example, y_{0t} could be a coincident business cycle indicator which moves a vector of macroeconomic time series y_{it}. In this case, e_{it} captures idiosyncratic movements in y_{it}. Alternatively, y_{0t} could be a common stochastic trend while e_{it} is stationary for all i. In this latter case, (11.1) resembles the common trend-UC decomposition studied in chapter 3. Furthermore, many of the models used in finance have a structure similar to (11.1). For example, in a capital asset pricing model (CAPM), y_{0t} is an unobservable market portfolio.

We need restrictions to identify the parameters of (11.1). Since $\mathbb{Q}_i$ and y_{0t} are nonobservable, neither the scale nor the sign of the factor and its loading can be separately identified. For normalization, we choose $\mathbb{Q}_1 > 0$ and assume that σ_0^2 is a fixed constant.

Let $\alpha_{1i} = (\bar{y}_i, \mathbb{Q}_i)$. Let $\alpha = (\alpha_{1i}, \sigma_0^2, \sigma_i^2, A_i^e, A^y,\ i = 1, \ldots, m)$, where $A_i^e = (A_{i,1}^e, \ldots, A_{i,q_i}^e)$ and $A^y = (A_1^y, \ldots, A_{q_0}^y)$, be the vector of parameters of the model. Let $y_i = (y_{i1}, \ldots, y_{it})'$ and $y = (y_1', \ldots, y_m')'$. Given $g(\alpha)$, $g(\alpha \mid y, y_0) \propto f(y \mid \alpha, y_0)g(\alpha)$ and $g(y_0 \mid \alpha, y) \propto f(y \mid \alpha, y_0)f(y_0 \mid \alpha)$. To compute these conditional distributions, we need $f(y \mid \alpha, y_0)$ and $f(y_0 \mid \alpha) = \int f(y, y_0 \mid \alpha)\,\mathrm{d}y$.

Consider first $f(y \mid \alpha, y_0)$. Let $y_i^1 = (y_{i,1}, \ldots, y_{i,q_i})'$ be random and let $y_0^1 = (y_{0,1}, \ldots, y_{0,q_0})'$ be the vector of initial observations on the factors, y_0^1 given, $x_i^1 = [\mathbf{1}, y_0^1]$, where $\mathbf{1} = [1, 1, \ldots, 1]'$, and let $\mathbb{A}_i$ be a $(q_i \times q_i)$ companion matrix representation of $A_i^e(\ell)$. If the errors are normal, $(y_i^1 \mid \bar{y}_i, \mathbb{Q}_i, \sigma_i^2, y_0^1) \sim \mathbb{N}(\bar{y}_i + \mathbb{Q}_i y_0^1, \sigma_i^2 \Sigma_i)$, where Σ_i solves $\Sigma_i = \mathbb{A}_i \Sigma_i \mathbb{A}_i + (1, 0, \ldots, 0)'(1, 0, \ldots, 0)$.

Exercise 11.1. Provide a closed-form solution for Σ_i.

Define $y_i^{1*} = \Sigma_i^{-0.5} y_i^1$ and $x_i^{1*} = \Sigma_i^{-0.5} x_i^1$. To build the rest of the likelihood, let $e_i = [e_{i,q_i+1}, \ldots, e_{i,T}]'$ (this is $(T - q_i) \times 1$ vector); $e_{it} = y_{it} - \bar{y}_i - \mathbb{Q}_i y_{0t}$ and $E = [e_1, \ldots, e_{q_i}]$ (this is a $(T - q_i) \times q_i$ matrix). Similarly, let $y_0 = (y_{01}, \ldots, y_{0t})'$ and $Y_0 = (y_{0,-1}, \ldots, y_{0,-q_0})$. Let y_i^{2*} be a $(T - q_i) \times 1$ vector with the t-row equal to $A_i^e(\ell)y_{it}$ and let x_i^{2*} be a $(T - q_i) \times 2$ matrix with the t-row equal to $(A_i^e(1), A_i^e(\ell)y_{0t})$. Let $x_i^* = [x_i^{1*}, x_i^{2*}]'$ and $y_i^* = [y_i^{1*}, y_i^{2*}]$.

Exercise 11.2. Derive the likelihood of $(y_i^* \mid x_i^*, \alpha)$, when e_t are normally distributed.

To obtain $g(\alpha \mid y, y_0)$, assume that $g(\alpha) = \prod_j g(\alpha_j)$, let σ_0^2 be fixed, and assume that $a_{1i} \sim \mathbb{N}(\bar{\alpha}_{1i}, \bar{\Sigma}_{\alpha_{1i}})$, $A_i^e \sim \mathbb{N}(\bar{A}_i^e, \bar{\Sigma}_{A_i^e})\mathcal{I}_{(-1,1)}$, $A^y \sim \mathbb{N}(\bar{A}^y, \bar{\Sigma}_{A^y})\mathcal{I}_{(-1,1)}$, $\sigma_i^{-2} \sim \mathbb{G}(a_{1i}, a_{2i})$, where $\mathcal{I}_{(-1,1)}$ is an indicator function for stationarity; that is,

the prior for $A_i^e(A^y)$ is normal, truncated outside the range $(-1, 1)$. Then, the conditional posteriors are

$$\left.\begin{aligned}
(\alpha_{1i} \mid y_i, \alpha_{-\alpha_{1i}}) &\sim \mathbb{N}(\tilde{\Sigma}_{\alpha_{1i}}(\bar{\Sigma}_{\alpha_{1i}}^{-1}\bar{\alpha}_{1i} + \sigma_i^{-2}(x_i^*)' y_i^*), \tilde{\Sigma}_{\alpha_{1i}}),\\
(A_i^e \mid y_i, y_0, \alpha_{-A_i^e}) &\sim \mathbb{N}(\tilde{\Sigma}_{A_i^e}(\bar{\Sigma}_{A_i^e}^{-1}\bar{A}_i^e + \sigma_i^{-2}E_i' e_i), \tilde{\Sigma}_{A_i^e})\mathcal{I}_{(-1,1)} \times \mathcal{N}(A_i^e),\\
(A^y \mid y_i, y_0, \alpha_{-A^y}) &\sim \mathbb{N}(\tilde{\Sigma}_{A^y}(\bar{\Sigma}_{A^y}^{-1}\bar{A}^y + \sigma_0^{-2}Y_0' y_0), \tilde{\Sigma}_{A^y})\mathcal{I}_{(-1,1)} \times \mathcal{N}(A^y),\\
(\sigma_i^{-2} \mid y_i, y_0, \alpha_{-\sigma_i}) &\sim \mathbb{G}((a_{1i} + T), a_{2i} + (y_i^* - x_i^*\alpha_{1i,\mathrm{OLS}})^2),
\end{aligned}\right\} \tag{11.2}$$

where $\tilde{\Sigma}_{a_i} = (\bar{\Sigma}_{a_i}^{-1} + \sigma_i^{-2}x_i^{*\prime}x_i^*)^{-1}$, $\tilde{\Sigma}_{A_i^e} = (\bar{\Sigma}_{A_i^e}^{-1} + \sigma_i^{-2}E_i'E_i)^{-1}$, $\tilde{\Sigma}_{A^y} = (\bar{\Sigma}_{A^y}^{-1} + \sigma_0^{-2}Y_0'Y_0)^{-1}$, while

$$\mathcal{N}(A_i^e) = |\Sigma_{A_i^e}|^{-0.5}\exp\{-(1/2\sigma_i^2)(y_i^1 - \bar{y}_i - \mathbb{Q}_i y_0^1)'\Sigma_{A_i^e}^{-1}(y_i^1 - \bar{y}_i - \mathbb{Q}_i y_0^1)\}$$

and

$$\mathcal{N}(A^y) = |\Sigma_{A^y}|^{-0.5}\exp\{-(1/2\sigma_0)(y_0^1 - A^y(\ell)y_{0,-1}^1)'\Sigma_{A^y}^{-1}(y_0^1 - A^y(\ell)y_{0,-1}^1)\}.$$

Sampling $(\bar{y}_i, \mathbb{Q}_i, \sigma_i^2)$ from (11.2) is straightforward. To impose the sign restriction necessary for identification, discard the draws producing $\mathbb{Q}_1 \leqslant 0$. The conditional posterior for $A_i^e(A^y)$ is complicated by the presence of the indicator for stationarity and the conditional distribution of the first $q_i(q_0)$ observations (without these two, drawing these parameters would also be straightforward). Since these distributions are of unknown form, one could use the following variation of the MH algorithm to draw, for example, A_i^e.

Algorithm 11.1.

(1) Draw $(A_i^e)^\dagger$ from $\mathbb{N}(\tilde{\Sigma}_{A_i^e}(\bar{\Sigma}_{A_i^e}^{-1}\bar{A}_i^e + \sigma_i^{-2}E_i'e_i), \tilde{\Sigma}_{A_i^e})$. If $\sum_{j=1}^{q_i}(A_{i,j}^e)^\dagger \geqslant 1$, discard the draw.

(2) Otherwise, draw $\mathfrak{U} \sim \mathbb{U}(0, 1)$. If $\mathfrak{U} < \mathcal{N}((A_i^e)^\dagger)/\mathcal{N}((A_i^e)^{l-1})$, set $(A_i^e)^l = (A_i^e)^\dagger$. Else set $(A_i^e)^l = (A_i^e)^{l-1}$.

(3) Repeat (1) and (2) L times.

The derivation of $g(y_0 \mid \alpha, y)$ is straightforward. Define the $T \times T$ matrix

$$\mathcal{Q}_i^{-1} = \begin{bmatrix} \mathcal{Q}_{i1} \\ \mathcal{Q}_{i2} \end{bmatrix},$$

where $\mathcal{Q}_{i1} = \begin{bmatrix} \Sigma_i^{-0.5} & 0 \end{bmatrix}$ and

$$\mathcal{Q}_{i2} = \begin{bmatrix} -A_{i,q_i}^e & \cdots & -A_{i,1}^e & 1 & 0 & \cdots & 0 \\ 0 & -A_{i,q_i}^e & \cdots & -A_{i,1}^e & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -A_{i,q_i}^e & \cdots & \cdots & 1 \end{bmatrix},$$

Σ_i is a $q_i \times q_i$ matrix, and 0 is a $q_i \times (T - q_i)$ matrix. Similarly, define $\mathcal{Q}_0^{-1}$.

Let $x_i^\dagger = \mathcal{Q}_i^{-1} x_i$ and $y_i^\dagger = \mathcal{Q}_i^{-1}(y_i - \mathbf{1}\bar{y}_i)$. Then the likelihood function is $\prod_{i=1}^m f(y_i^\dagger \mid \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0)$, where $f(y_i^\dagger \mid \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0) = (2\pi\sigma_i^2)^{-0.5T} \times \exp\{-(y_i^\dagger - \mathbb{Q}_i \mathcal{Q}_i^{-1} y_0)'(y_i^\dagger - \mathbb{Q}_i \mathcal{Q}_i^{-1} y_0)/2\sigma_i^2\}$. Since the marginal of the factor is $f(y_0 \mid A^y) = (2\pi\sigma_0^2)^{-0.5T} \exp\{-(\mathcal{Q}_0^{-1} y_0)'(\mathcal{Q}_0^{-1} y_0)/2\sigma_0^2\}$, the joint likelihood is $f(y^\dagger, y_0 \mid \alpha) = \prod_{i=1}^m f(y_i^\dagger \mid \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0) f(y_0 \mid A^y)$. Completing the squares we have

$$g(y_0 \mid y_i^\dagger, \alpha) \sim \mathbb{N}(\tilde{y}_0, \tilde{\Sigma}_{y_0}), \tag{11.3}$$

where $\tilde{y}_0 = \tilde{\Sigma}_{y_0}[\sum_{i=1}^m \mathbb{Q}_i \sigma_i^{-2} (\mathcal{Q}_i^{-1})' \mathcal{Q}_i^{-1}(y_i - \mathbf{1}\bar{y}_i)]$, $\tilde{\Sigma}_{y_0} = [\sum_{i=0}^m \mathbb{Q}_i^2 \sigma_i^{-2} \times (\mathcal{Q}_i^{-1})'(\mathcal{Q}_i^{-1})]^{-1}$ with $\mathbb{Q}_0 = 1$. Note that $\tilde{\Sigma}_{y_0}$ is a $T \times T$ matrix. Given (11.2) and (11.3), the Gibbs sampler can be used to compute the joint conditional posterior of α and of y_0, and their marginals.

To make the Gibbs sampler operative we need to select σ_0^2 and the parameters of the prior distributions. For example, σ_0^2 could be set to the average variance of the innovations in an AR(1) regression for each y_{it}. Since little information is typically available on the loadings and the autoregressive parameters, one could set $\bar{a}_{i1} = \bar{A}_i^e = \bar{A}^y = 0$ and assume a large prior variance. Finally, a relatively diffuse prior for σ_i^{-2} could be chosen, for example, $\mathbb{G}(4, 0.001)$, a distribution without the third and fourth moments.

The calculation of the predictive density of y_{0t} is straightforward and it is left as an exercise for the reader. Note that when the factor is a common business cycle indicator, the construction of this quantity produces the density of a leading indicator.

Exercise 11.3. Describe how to construct the predictive density of $y_{0t+\tau}$, $\tau = 1, 2, \ldots$.

Exercise 11.4. Suppose that $i = 4$ and let $A_i^e(\ell)$ be of first order. In addition, suppose that $\bar{y} = [0.5, 0.8, 0.4, 0.9]'$ and $\mathbb{Q}_1 = [1, 2, 0.4, 0.6, 0.5]'$. Let $A^e = \text{diag}[0.8, 0.7, 0.6, 0.9]$, $A^y = [0.7, -0.3]$, $v_0 \sim$ i.i.d. $\mathbb{N}(0, 5)$, and

$$v \sim \text{i.i.d. } \mathbb{N}\left(0, \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}\right).$$

Let the priors be $(\bar{y}_i, \mathbb{Q}_i) \sim \mathbb{N}(0, 10 * I_2)$, $i = 1, 2, 3, 4$, $A^e \sim \mathbb{N}(0, I_4)\mathcal{I}_{(-1,1)}$, $A^y \sim \mathbb{N}(0, I_2)\mathcal{I}_{(-1,1)}$, and $\sigma_i^{-2} \sim \mathbb{G}(4, 0.001)$, where $\mathcal{I}_{(-1,1)}$ instructs us to drop values such that $\sum_j A_{ij}^e \geqslant 1$ or $\sum_j A_j^y \geqslant 1$. Draw sequences from the posterior of α and construct an estimate of the posterior distribution of y_0.

Exercise 11.5. Let the prior for $(\bar{y}_i, \mathbb{Q}_i, A_i^e, A^y, \sigma_i^{-2})$ be noninformative. Show that the posterior mean of y_0 is the same as the one obtained by running the Kalman filter/smoother on model (11.1).

Example 11.2. We construct a coincident indicator for the euro area business cycle by using quarterly data on real government consumption, real private investment,

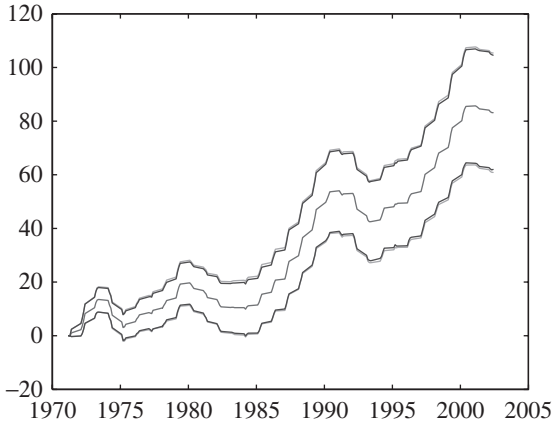

Figure 11.1. Coincident indicator, euro area.

real employment, and real GDP from 1970:1 to 2002:4. We allow an AR(2) structure on the indicator and an AR(1) on the errors of the model. Posterior estimates are obtained by using 10 000 draws from the conditional posterior: 5000 are used as burnout; of the remaining, 1 out of every 5 is used to construct the indicator. The mean value of the indicator together with a 68% confidence band are shown in figure 11.1. The posterior means of the two AR coefficients are 0.711 and 0.025, and the posterior standard errors are 0.177 and 0.134. The coincident indicator we construct shows (classical) recessions, roughly, at the same dates the CEPR selected as recession dates. Furthermore, it displays a considerable slowdown after 2001.

11.1.1 Arbitrage Pricing Theory (APT) Models

Apart from the construction of business cycle or trend indicators, factor models are extensively used in finance (see, for example, Campbell et al. (1997) for references). Here the unobservable factor is a vector of portfolio excess returns, a vector of macroeconomic variables, or a vector of portfolio of real returns, typically restricted to span the mean–variance frontier. APT models are useful since economic theory imposes restrictions on nonlinear combinations of their parameters.

For illustrative purposes, consider a version of an APT model where a vector of m asset returns y_t is related to a vector of k factors y_{0t} according to the linear relationship

$$y_t = \bar{y} + \mathbb{Q}_1 y_{0t} + e_t, \tag{11.4}$$

where $y_0 \sim \mathbb{N}(0, I)$, $e \mid y_0 \sim$ i.i.d. $\mathbb{N}(0, \Sigma_e)$, $\bar{y}$ is a vector of conditional mean returns, $\mathbb{Q}_1$ is an $m \times k$ matrix of loadings, and both $\mathbb{Q}_1$ and y_{0t} are unknown. Traditionally, a model like (11.4) is estimated in two steps: in the first step either the factor loadings or the factors themselves are estimated (with a cross-sectional regression). Then, taking the first-step estimates as if they were the true ones, a second-pass regression (typically, in time series) is used to estimate the other parameters (see, for

example, Roll and Ross 1980). Clearly, this approach suffers from error-in-variables problems and leads to incorrect inference.

A number of authors, starting from Ross (1976), have shown that, as $m \to \infty$, absence of arbitrage opportunities implies that $\bar{y}_i \approx \phi_0 + \sum_{j=1}^{k} \mathbb{Q}_{1ij}\phi_j$, where ϕ_0 is the intercept of the pricing relationship (the so-called zero-beta rate) and ϕ_j is the risk premium on factor $\mathbb{Q}_{1ij}$, $j = 1, 2, \ldots, k$. With the two-step procedure we have described, and treating the estimates of $\mathbb{Q}_{1ij}$ and of $\bar{y}_i$ as given, the restrictions imposed become linear and tests can be easily developed by using restricted and unrestricted estimates of ϕ_j (see Campbell et al. 1997).

One way to test (11.4) is to measure the pricing errors and check their sizes relative to the average returns (with large relative errors indicating an inappropriate specification). This measure is given by $\mathfrak{S} = (1/m)\bar{y}'[I - \mathbb{Q}(\mathbb{Q}'\mathbb{Q})^{-1}\mathbb{Q}']\bar{y}$, where $\mathbb{Q} = (\mathbf{1}, \mathbb{Q}_1)$ and $\mathbf{1}$ is a vector of 1s of dimension m. For fixed m, $\mathfrak{S} \neq 0$, while as $m \to \infty$, $\mathfrak{S} \to 0$. While it is hard to compute the sampling distribution of $\mathfrak{S}$, its exact posterior distribution can be easily obtained with MCMC methods.

For identification we require that $k < \frac{1}{2}m$. Letting $\mathbb{Q}_1^k$ be a lower triangular matrix containing the Choleski transformation of the first k independent rows of $\mathbb{Q}_1$, we also want $\mathbb{Q}_{1ii}^k > 0$, $i = 1, \ldots, k$.

Exercise 11.6. Show that $k < \frac{1}{2}m$ and $\mathbb{Q}_{1ii}^k > 0$, $i = 1, \ldots, k$, are necessary for identification.

Let $\alpha_{i1} = (\bar{y}_i, \mathbb{Q}_i)$. Since the factors capture common components, $\Sigma_e = \text{diag}\{\sigma_i^2\}$. Then $f(\alpha_{i1} \mid y_0, \sigma_i) \propto \exp\{-(\alpha_{i1} - \alpha_{i1,\text{OLS}})'x'x(\alpha_{i1} - \alpha_{i1,\text{OLS}})/2\sigma_i^2\}$, where $x = (\mathbf{1}, y_0)$ is a $T \times (k+1)$ matrix and $\alpha_{i1,\text{OLS}}$ are the OLS estimators of the coefficients in a regression of y_{it} on $(1, y_0)$. We want to compute $g(\alpha \mid y_{0t}, y_t)$ and $g(y_{0t} \mid \alpha, y_t)$, where $\alpha = (\alpha_{1i}, \sigma_i^2,\ i = 1, 2, \ldots)$. We assume independence across i and the following priors: $\mathbb{Q}_{1i} \sim \mathbb{N}(\bar{\mathbb{Q}}_{1i}, \bar{\sigma}^2_{\mathbb{Q}_1})$, $\mathbb{Q}_{1ii} > 0$, $i = 1, \ldots, k$, $\mathbb{Q}_{1i} \sim \mathbb{N}(\bar{\mathbb{Q}}_{1i}, \bar{\omega}^2_{\mathbb{Q}_1})$, $i = k+1, \ldots, m$, $\bar{s}_i^2\sigma_i^{-2} \sim \chi^2(\bar{\nu}_i)$, $\bar{y}_i \sim \mathbb{N}(\bar{y}_{i0}, \bar{\sigma}^2_{\bar{y}_i})$, where $\bar{y}_{i0} = \phi_0 + \sum_j \mathbb{Q}_{1ij}\phi_j$ and ϕ_i are constant. The hyperparameters of all prior distributions are assumed to be known. Note that we impose the theoretical restrictions directly — the prior distribution of $\bar{y}_i$ is conditional on the value of $\mathbb{Q}_1$ — and that by varying $\bar{\sigma}^2_{y_i}$ we can account for different degrees of credence in the ATP restrictions. The conditional posterior distributions for the parameters are easily obtained.

Exercise 11.7. (i) Show that $g(\bar{y}_i \mid y_t, y_{0t}, \mathbb{Q}_1, \sigma_i^2) \sim \mathbb{N}(\tilde{\bar{y}}_i, \tilde{\sigma}^2_{\bar{y}_i})$, where $\tilde{\bar{y}}_i = [\bar{\sigma}^2_{\bar{y}_i}\bar{y}_{i,\text{OLS}} + (\sigma_i^2/T)\bar{y}_{i0}]/[\sigma_i^2/T + \bar{\sigma}^2_{\bar{y}_i}]$, $\tilde{\sigma}^2_{\bar{y}_i} = [(\sigma_i^2\bar{\sigma}^2_{\bar{y}_i})/T]/[\sigma_i^2/T + \bar{\sigma}^2_{\bar{y}_i}]$, $\bar{y}_{i,\text{OLS}} = (1/T)\sum_{t=1}^{T}(y_{it} - \sum_{j=1}^{k}\mathbb{Q}_{1j}y_{0tj})$.

(ii) Show that $g(\mathbb{Q}_{1i} \mid y_t, y_{0t}, \bar{y}_i, \sigma_i^2) \sim \mathbb{N}(\tilde{\mathbb{Q}}_{1i}, \tilde{\Sigma}_{Q_{1i}})$, with $\tilde{\mathbb{Q}}_{1i} = \Sigma_{Q_{1i}} \times (\bar{\mathbb{Q}}_{1i}\bar{\sigma}^{-2}_{\mathbb{Q}_1} + x_i^{\dagger\prime}x_i^{\dagger}\mathbb{Q}_{1i,\text{OLS}}\sigma_i^{-2})$, $\tilde{\Sigma}_{Q_{1i}} = (\bar{\sigma}^{-2}_{\mathbb{Q}_{1i}} + \sigma_i^{-2}x_i^{\dagger\prime}x_i^{\dagger})^{-1}$, $i = 1, \ldots, k$, and $\tilde{\mathbb{Q}}_{1i} = \Sigma_{Q_{1i}}(\bar{\mathbb{Q}}_{1i}\bar{\omega}^{-2}_{\mathbb{Q}_1} + x_i^{\dagger\prime}x_i^{\dagger}\mathbb{Q}_{1i,\text{OLS}}\sigma_i^{-2})$, $\tilde{\Sigma}_{Q_{1i}} = (\bar{\omega}^{-2}_{\mathbb{Q}_1} + \sigma_i^{-2}x_i^{\dagger\prime}x_i^{\dagger})^{-1}$, $i = k+1, \ldots, m$, where $\mathbb{Q}_{1i,\text{OLS}}$ is the OLS estimator of a regression of $(y_{it} - \bar{y}_0)$ on $y_{01}, \ldots, y_{0i-1}$ and $x_i^{\dagger}$ is the matrix x_i without the first row.

(iii) Show that $(\tilde{s}^2\sigma_i^{-2} \mid y_t, y_{0t}, \mathbb{Q}_1, \bar{y}_i) \sim \chi^2(\tilde{\nu})$, where $\tilde{\nu} = \bar{\nu} + T$ and $\tilde{s}_i^2 = \bar{\nu}\bar{s}_i^2 + (T - k - 1)\sum_t(y_{it} - \bar{y}_i - \sum_j \mathbb{Q}_{1j}y_{0tj})^2$.

The joint density of the data and the factor is

$$\begin{bmatrix} y_{0t} \\ y_t \end{bmatrix} \sim \mathbb{N}\left[\begin{pmatrix} 0 \\ \bar{y} \end{pmatrix}, \begin{pmatrix} I & \mathbb{Q}_1' \\ \mathbb{Q}_1 & \mathbb{Q}_1\mathbb{Q}_1' + \Sigma_e \end{pmatrix} \right].$$

Using the properties of conditional normal distributions we have $g(y_{0t} \mid y_t, \alpha) \sim \mathbb{N}(\mathbb{Q}_1'(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1}(y_t - \bar{y}), I - \mathbb{Q}_1'(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1}\mathbb{Q}_1)$, with $(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1} = \Sigma_e^{-1} - \Sigma_e^{-1}\mathbb{Q}_1(I + \mathbb{Q}_1'\Sigma_e^{-1}\mathbb{Q}_1)^{-1}\mathbb{Q}_1'\Sigma_e^{-1}$, where $(I + \mathbb{Q}_1'\Sigma_e^{-1}\mathbb{Q}_1)$ is a $k \times k$ matrix.

Exercise 11.8. Suppose the prior for α is noninformative, that is, $g(\alpha) \propto \prod_j \sigma_{\alpha_j}^{-2}$. Derive the conditional posteriors for $\bar{y}$, $\mathbb{Q}_1$, Σ_e, and y_{0t} in this case.

Exercise 11.9. Using monthly returns data on the stocks listed in Eurostoxx 50 for the last five years, construct five portfolios with the quintiles of the returns. Using informative priors compute the posterior distribution of the pricing error in an APT model using one and two factors (averaging over portfolios). You may want to try two values for σ_0^2, one large and one small. Report a posterior 68% credible set for $\mathfrak{S}$. Do you reject the theory? What can you say about the posterior mean of the proportion of idiosyncratic to total risk?

11.1.2 Conditional Capital Asset Pricing Models

A conditional CAPM combines data-based and model-based approaches to portfolio selection into a specification of the form

$$\left.\begin{aligned} y_{it+1} &= \bar{y}_{it} + \mathbb{Q}_{it}y_{0t+1} + e_{it+1}, \\ \mathbb{Q}_{it} &= x_{1t}\phi_{1i} + v_{1it}, \\ \bar{y}_{it} &= x_{1t}\phi_{2i} + v_{2it}, \\ y_{0t+1} &= x_{2t}\phi_0 + v_{0t+1}, \end{aligned}\right\} \tag{11.5}$$

where $x_t = (x_{1t}, x_{2t})$ is a set of observable variables, $e_{it+1} \sim$ i.i.d. $\mathbb{N}(0, \sigma_e^2)$, $v_{0t+1} \sim$ i.i.d. $\mathbb{N}(0, \sigma_0^2)$, and both v_{1it} and v_{2it} are assumed to be serially correlated, to take into account the possible misspecification of the conditioning variables x_{1t}. Here y_{it+1} is the return on asset i and y_{0t+1} is the return on an unobservable market portfolio. Equations (11.5) fit the factor model structure we have so far considered when $v_{2it} = v_{1it} = 0, \forall t$, x_{2t} are the lags of y_{0t} and $x_{1t} = I$ for all t. Various versions of (11.5) have been considered in the literature.

Example 11.3. Consider the model

$$\left.\begin{aligned} y_{it+1} &= \mathbb{Q}_{it} + e_{it+1}, \\ \mathbb{Q}_{it} &= x_t\phi_i + v_{it}. \end{aligned}\right\} \tag{11.6}$$

Here the return on asset i depends on an unobservable risk premium $\mathbb{Q}_{it}$ and on an idiosyncratic error term, and the risk premium is a function of observable variables.

If we relax the assumption that the cost of risk is constant and allow time variations in the conditional variance of asset i, we have

$$y_{it+1} = x_t \mathbb{Q}_t + e_{it+1}, \quad e_{it} \sim \text{i.i.d. } \mathbb{N}(0, \sigma_{e_i}^2), \tag{11.7}$$

$$\mathbb{Q}_t = \mathbb{Q} + v_t, \qquad v_t \sim \text{i.i.d. } \mathbb{N}(0, \sigma_v^2). \tag{11.8}$$

Here the return on asset i depends on observable variables. The loadings on the observables, assumed to be the same across assets, are allowed to vary over time. Note that by substituting the second expression into the first we have that the model's prediction error is heteroskedastic (the variance is $x_t' x_t \sigma_v^2 + \sigma_{e_i}^2$).

Exercise 11.10. Suppose that $v_{2it} = v_{1it} = 0, \forall t$, and assume that y_{0t+1} is known. Let $\alpha = [\phi_{21}, \ldots, \phi_{2m}, \phi_{11}, \ldots, \phi_{1m}]$. Assume *a priori* that $\alpha \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$. Let the covariance matrix of $e_t = [e_{1t}, \ldots, e_{MT}]$ be Σ_e and assume that, *a priori*, $\Sigma_e^{-1} \sim \mathbb{W}(\bar{\Sigma}, \bar{\nu})$. Show that, conditional on $(y_{it}, y_{0t}, \Sigma_e, x_t)$, the posterior of α is normal with mean $\tilde{\alpha}$ and variance $\tilde{\Sigma}_\alpha$ and that the marginal posterior of Σ_e^{-1} is Wishart with scale matrix $(\bar{\Sigma}^{-1} + \Sigma_{\text{OLS}})^{-1}$ and $\bar{\nu} + T$ degrees of freedom. Show the exact form of $\tilde{\alpha}$, $\tilde{\Sigma}_\alpha$, and Σ_{OLS}.

Exercise 11.11. Assume $v_{2it} = v_{1it} = 0, \forall t$, but allow y_{0t+1} to be unobservable. Postulate a law of motion for y_{0t} of the form $y_{0t+1} = x_{2t}\phi_0 + v_{0t+1}$, where x_{2t} are observables. Describe the steps needed to find the conditional posterior of y_{0t}.

The specification in (11.5) is more complicated than the one in exercises 11.10 and 11.11 because of time variations in the coefficients. To highlight the steps involved in this case, we describe a version of (11.5) where $v_{2it} = 0, \forall t, m = 1, x_t = x_{1t} = x_{2t}$, and we allow for AR(1) errors in the law of motion of $\mathbb{Q}_t$, that is,

$$\left.\begin{aligned} y_{t+1} &= x_t\phi_2 + \mathbb{Q}_t y_{0t+1} + e_{t+1}, \\ \mathbb{Q}_t &= (x_t - \rho x_{t-1})\phi_1 + \rho\mathbb{Q}_{t-1} + v_t, \\ y_{0t} &= x_t\phi_0 + v_{0t}, \end{aligned}\right\} \tag{11.9}$$

where ρ measures the persistence of the shock driving $\mathbb{Q}_t$.

Let $\alpha = [\phi_0, \phi_1, \phi_2, \rho, \sigma_e^2, \sigma_v^2, \sigma_{v_0}^2]$ and let $g(\alpha) = \prod_j g(\alpha_j)$. Assume that $g(\phi_i) \sim \mathbb{N}(\bar{\phi}_i, \bar{\Sigma}_{\phi_i})$, $i = 0, 1, 2$, $g(\rho) \sim \mathbb{N}(0, \bar{\Sigma}_\rho)\mathcal{I}_{(-1,1)}$, $g(\sigma_v^{-2}) \sim \chi(\bar{s}_v^2, \bar{\nu}_v)$, $g(\sigma_e^{-2}, \sigma_{v_0}^{-2}) \propto \sigma_e^{-2}\sigma_{v_0}^{-2}$, and that all hyperparameters are known.

To construct the conditional posterior of $\mathbb{Q}_t$ note that, if ρ is known, $\mathbb{Q}_t$ can be easily simulated as in state space models. Therefore, partition $\alpha = (\alpha_1, \rho)$. Conditional on ρ, the law of motion of $\mathbb{Q}_t$ is $y \equiv \mathbb{Q} - \rho\mathbb{Q}_{-1} = x^+\phi_1 + v$, where $\mathbb{Q} = [\mathbb{Q}_1, \ldots, \mathbb{Q}_t]'$, $x = [x_1, \ldots, x_t]'$, $x^+ = x - \rho x_{-1}$, and $v \sim \text{i.i.d. } \mathbb{N}(0, \sigma_v^2 I_T)$. Setting $\mathbb{Q}_{t=-1} = 0$, we have two sets of equations, one for the first observation and one for the others, $y_0 \equiv \mathbb{Q}_0 = x_0^+\phi_1 + v_0$ and $y_t \equiv \mathbb{Q}_t - \rho\mathbb{Q}_{t-1} = x_t^+\phi_1 + v_t$. When the errors are normal, the likelihood function $f(y \mid x, \phi_1, \rho)$ is proportional to $(\sigma_v^2)^{-0.5T} \exp\{-0.5[(y_0 - x_0^+\phi_1)\sigma_v^{-2}(y_0 - x_0^+\phi_1)' - \sum_{t=1}^T (y_t - x_t^+\phi_1)\sigma_v^{-2} \times (y_t - x_t^+\phi_1)']\}$.

Let $\phi^0_{1,\text{OLS}}$ be the OLS estimator obtained from the first observation and $\phi^1_{1,\text{OLS}}$ the OLS estimator obtained from the other observations. Combining the prior and the likelihood, the posterior kernel of ϕ is proportional to $\exp\{-0.5(\phi^0_1-\phi^0_{1,\text{OLS}})'\times (x^+_0)'\sigma_v^{-2}x^+_0(\phi^1_0-\phi^1_{0,\text{OLS}})-0.5\sum_t(\phi^1_1-\phi^1_{1,\text{OLS}})'(x^+_t)'\sigma_v^{-2}x^+_t(\phi^1_1-\phi^1_{1,\text{OLS}})-0.5(\phi_1-\bar{\phi}_1)'\bar{\Sigma}^{-1}_{\phi_1}(\phi_1-\bar{\phi}_1)\}$. Therefore, the conditional posterior for ϕ_1 is normal. The mean is a weighted average of prior mean and two OLS estimators, i.e., $\tilde{\phi}_1=\tilde{\Sigma}_{\phi_1}(\bar{\Sigma}^{-1}_{\phi_1}\bar{\phi}_1+(x^+_0)'\sigma_v^{-2}y_0+\sum_t(x^+_t)'\sigma_v^{-2}y_t)$ and $\tilde{\Sigma}_{\phi_1}=(\bar{\Sigma}^{-1}_{\phi_1}+(x^+_0)'\sigma_v^{-2}x^+_0+\sum_t(x^+_t)'\sigma_v^{-2}x^+_t)^{-1}$. The conditional posterior for σ_v^2 can be found by using the same logic.

Exercise 11.12. Show that the posterior kernel for σ_v^2 has the form $(\sigma_v^2)^{-0.5(T-1)}\times\exp\{-0.5\sum_t\sigma_v^{-2}(y_t-x^+_t\phi_1)'(y_t-x^+_t\phi_1)\}[(\sigma_v^2/(1-\phi_1^2))^{0.5}]^{-0.5(\bar{\nu}_v+1+2)}\times\exp\{-0.5[\sigma_v^2/(1-\phi_1^2)]^{-1}[(y_0-x^+_0\phi_1)'(y_0-x^+_0\phi_1)+\bar{\nu}_v]\}$. Suggest an algorithm to draw from this (unknown) distribution.

Once the distribution for the components of α_1 is found, we can use the Kalman filter/smoother to construct $\mathbb{Q}_t$ and the posterior of y_{0t}, conditional on ρ. To find the posterior distribution of ρ requires little more work. Conditional on ϕ_1, rewrite the law of motion for $\mathbb{Q}_t$ as $y^\dagger_t\equiv\mathbb{Q}_t-x_t\phi_1=x^\dagger_{t-1}\rho+v_t$, where $x^\dagger_{t-1}=\mathbb{Q}_{t-1}-x_{t-1}\phi_1$. Once again, split the data in two: initial observations $(y^\dagger_1,x^\dagger_0)$ and the rest $(y^\dagger_t,x^\dagger_{t-1})$. The likelihood function is

$$f(y^\dagger\mid x^\dagger,\phi_1,\rho)\propto\sigma_v^{-T}\exp\{-0.5(y^\dagger_1-x^\dagger_0\phi_1)'\sigma_v^{-2}(y^\dagger_1-x^\dagger_0\phi_1)\}+\exp\left\{-0.5\sum_t(y^\dagger_t-x^\dagger_{t-1}\phi_1)'\sigma_v^{-2}(y^\dagger_t-x^\dagger_{t-1}\phi_1)\right\}. \quad (11.10)$$

Let ρ_{OLS} be the OLS estimator of ρ obtained with T data points. Combining the likelihood with the prior produces a kernel of the form $\exp\{-0.5\sum_t(\rho-\rho_{\text{OLS}})'\times(x^\dagger_t)'\sigma_v^{-2}x^\dagger_t(\rho-\rho_{\text{OLS}})+(\rho-\bar{\rho}')\bar{\Sigma}^{-1}_\rho(\rho-\bar{\rho})\}[(\sigma_v^2/(1-\phi_1^2))^{0.5}]^{-0.5(\bar{\nu}_v+1+2)}\times\exp\{-0.5[\sigma_v^2/(1-\phi_1^2)]^{-1}\bar{\nu}_v+(y^\dagger_1)'[\sigma_v^2/(1-\phi_1^2)]^{-1}y^\dagger_1\}$. Hence, the conditional posterior for ρ is normal, truncated outside the range $(-1,1)$, with mean $\tilde{\rho}=\tilde{\Sigma}_\rho(\bar{\Sigma}^{-1}_\rho\bar{\rho}+\sum_t(x^\dagger_t)'\sigma_v^{-2}y^\dagger_t)$, variance $\tilde{\Sigma}_\rho=(\bar{\Sigma}^{-1}_\rho+\sum_t(x^\dagger_t)'\sigma_v^{-2}x^\dagger_t)^{-1}$.

Exercise 11.13. Provide an MH algorithm to draw from the conditional posterior of ρ.

Once $g(\alpha_1\mid\rho,y_{0t},y_t)$, $g(\rho\mid\alpha_1,y_{0t},y_t)$, $g(y_{0t}\mid\alpha_1,\rho,y_t)$ are available, the Gibbs sampler can be used to find the joint posterior of the quantities of interest.

11.2 Stochastic Volatility Models

Stochastic volatility models are alternatives to GARCH or TVC models. In fact, they can account for time-varying volatility and leptokurtosis as GARCH or TVC models but produce excess kurtosis without heteroskedasticity. Since the logarithm of σ_t^2 is assumed to follow an AR process, changes in y_t are driven by shocks in the model for the observables or shocks in the model for the logarithm of σ_t^2. Such

a feature adds flexibility to the specification and produces richer dynamics for the observables as compared with, for example, GARCH-type models, where the same random variable drives both observables and volatilities.

The most basic stochastic volatility specification is

$$\left.\begin{aligned} y_t &= \sigma_t e_t, & e_t &\sim \mathbb{N}(0,1), \\ \ln(\sigma_t^2) &= \rho_0 + \rho_1 \ln(\sigma_{t-1}^2) + \sigma_v v_t, & v_t &\sim \text{i.i.d. } \mathbb{N}(0,1), \end{aligned}\right\} \tag{11.11}$$

where v_t and e_t are independent. In (11.11) we have implicitly assumed that y_t is de-meaned. Hence, this specification could be used to model, for example, asset returns or changes in exchange rates. Also, for simplicity, only one lag of $\ln \sigma_t^2$ is considered.

Let $y = (y_1, \ldots, y_t)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_t^2)$, and let $f(\sigma^2 \mid \rho, \sigma_v)$ be the probability mechanism generating σ^2, where $\rho = (\rho_0, \rho_1)$. The density of the data is $f(y \mid \rho, \sigma_v) = \int f(y \mid \sigma^2) f(\sigma^2 \mid \rho, \sigma_v)\, \mathrm{d}\sigma^2$. As in factor models, we treat σ^2 as an unknown vector of parameters, whose conditional distribution needs to be found.

We postpone the derivation of the conditional distribution of (ρ, σ_v) to a later (more complicated) application and concentrate on the problem of drawing a sample from the conditional posterior of σ_t^2. First, note that, because of the Markov structure, we can break the joint posterior of σ^2 into the product of conditional posteriors of the form $g(\sigma_t^2 \mid \sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t)$, $t = 1, \ldots, T$. Second, these univariate densities have an unusual form: they are the product of a conditional normal for y_t and a lognormal for σ_t^2,

$$\begin{aligned} &g(\sigma_t^2 \mid \sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t) \\ &\qquad \propto f(y_t \mid \sigma_t^2) f(\sigma_t^2 \mid \sigma_{t-1}^2, \rho, \sigma_v) f(\sigma_{t+1}^2 \mid \sigma_t^2, \rho, \sigma_v) \\ &\qquad \propto \frac{1}{\sigma_t} \exp\left\{-\frac{y_t^2}{2\sigma_t^2}\right\} \times \frac{1}{\sigma_t^2} \exp\left\{-\frac{(\ln \sigma_t^2 - E_t(\ln \sigma_t^2))^2}{2 \operatorname{var}(\ln \sigma_t^2)}\right\}, \end{aligned} \tag{11.12}$$

where $E_t(\ln \sigma_t^2) = [\rho_0(1-\rho_1) + \rho_1(\ln \sigma_{t+1}^2 + \ln \sigma_{t-1}^2)]/(1+\rho_1^2)$, $\operatorname{var}(\ln \sigma_t^2) = \sigma_v^2/(1+\rho_1^2)$. Because $g(\sigma_t^2 \mid \sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t)$ is nonstandard, we need either a candidate density to be used as importance sampling or an appropriate transition function to be used in an MH algorithm. There is an array of densities one could use as importance sampling densities. For example, Jacquier et al. (1994) noticed that the first term in (11.12) is the density of an inverse of gamma distributed random variable, that is, $x^{-1} \sim \mathbb{G}(a_1, a_2)$, while the second term can be approximated by an inverse of a gamma distribution (matching first and second moments). The inverse of a gamma is a good "blanketing" density for the lognormal because it dominates the latter on the right tail. Furthermore, the two densities can be combined into one inverse gamma with parameters $\tilde{a}_1 = [1 - 2\exp(\operatorname{var}(\ln \sigma_t^2))]/[1 - \exp(\operatorname{var}(\ln \sigma_t^2))] + 0.5$ and $\tilde{a}_2 = [(\tilde{a}_1 - 1)[\exp(E_t(\ln \sigma_t^2) + 0.5 \operatorname{var}(\ln \sigma_t^2))] + 0.5 y_t^2]$ and draws made from this target density. As an alternative, since the kernel of $\ln(\sigma_t^2)$ is known, we could draw $\ln(\sigma_t^2)$ from $\mathbb{N}(E(\ln \sigma_t^2) - 0.5 \operatorname{var}(\ln \sigma_t^2), \operatorname{var}(\ln \sigma_t^2))$ and accept the draw with probability $\exp\{-y_t^2/2\sigma_t^2\}$ (see Geweke 1994).

Table 11.1. Percentiles of the approximating distributions.

	Percentiles				
	5th	25th	50th	75th	95th
Gamma	0.11	0.70	1.55	3.27	5.05
Normal	0.12	0.73	1.60	3.33	5.13

Example 11.4. We have run a small Monte Carlo experiment to check the quality of these two approximations. Table 11.1 reports the percentiles using 5000 draws from the posterior when $\rho_0 = 0.0$, $\rho_1 = 0.8$, and $\sigma_v = 1.0$. Both approximations appear to produce similar results.

It is worthwhile stressing that (11.11) is a particular nonlinear Gaussian model which can be transformed into a linear but non-Gaussian state space model without loss of information. In fact, letting $x_t = \ln \sigma_t$, $\epsilon_t = \ln e_t^2 + 1.27$, the model (11.11) could be written as

$$\left.\begin{aligned} \ln y_t^2 &= -1.27 + x_t + \epsilon_t, \\ x_{t+1} &= \rho x_t + \sigma_v v_t, \end{aligned}\right\} \tag{11.13}$$

where ϵ_t has zero mean but is nonnormal. A framework like this was encountered in chapter 10 and techniques designed to deal with such models were outlined there. Here it is sufficient to point out that a nonnormal density for ϵ_t can be approximated with a mixture of J normals, that is, $f(\epsilon_t) \approx \sum_j \varrho_j f(\epsilon_t \mid \mathcal{M}_j)$, where each $f(\epsilon_t \mid \mathcal{M}_j) \sim \mathbb{N}(\bar{\epsilon}_j, \sigma_{\epsilon_j}^2)$ and $0 \leqslant \varrho_j \leqslant 1$. Chib (1996) provides details on how this can be done.

Cogley and Sargent (2005) have recently applied the mechanics of stochastic volatility models to a BVAR with time-varying coefficients. Since the setup they use could be employed as an alternative to the linear time-varying conditional structures we studied in chapter 10, we will examine in detail how to obtain conditional posterior estimates for the parameters of such a model.

A VAR model with stochastic volatility has the form

$$\left.\begin{aligned} y_t &= (I_m \otimes \boldsymbol{X}_t)\alpha_t + e_t, \qquad e_t \sim \mathbb{N}(0, \Sigma_t^{\dagger}), \\ \Sigma_t^{\dagger} &= \mathcal{P}^{-1} \Sigma_t (\mathcal{P}^{-1})', \\ \alpha_t &= \mathbb{D}_1 \alpha_{t-1} + v_{1t}, \qquad v_{1t} \sim \mathbb{N}(0, \Sigma_{v_1}), \end{aligned}\right\} \tag{11.14}$$

where $\mathcal{P}$ is a lower triangular matrix with 1s on the main diagonal, $\Sigma_t = \text{diag}\{\sigma_{it}^2\}$,

$$\ln \sigma_{it}^2 = \ln \sigma_{it-1}^2 + \sigma_{v_{2i}} v_{2it}, \tag{11.15}$$

where $\mathbb{D}_1$ is such that α_t is a stationary process. In (11.14) the process for y_t has time-varying coefficients and time-varying variances. To compute conditional posteriors note that it is convenient to block together the α_t and the σ_t^2 and draw a whole sequence for these two vectors of random variables.

We make standard prior assumptions, i.e., $\alpha_0 \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_a)$, $\Sigma_{v_1}^{-1} \sim \mathbb{W}(\bar{\Sigma}_{v_1}, \bar{\nu}_{v_1})$, where $\bar{\Sigma}_{v_1} \propto \bar{\Sigma}_a$, $\bar{\nu}_{v_1} = \dim(\alpha_0) + 1$, $\sigma_{v_{2i}}^{-2} \sim \mathbb{G}(a_1, a_2)$, $\ln \sigma_{i0} \sim \mathbb{N}(\ln \bar{\sigma}_i, \bar{\Sigma}_\sigma)$, and letting ϕ represent the nonzero elements of $\mathcal{P}$, $\phi \sim \mathbb{N}(\bar{\phi}, \bar{\Sigma}_\phi)$.

Given these priors, the calculation of the conditional posterior for $(\alpha_t, \Sigma_{v_1}, \sigma_{v_{2i}})$ is straightforward. The conditional posterior for α_t can be obtained with a run of the Kalman filter as detailed in chapter 10; the conditional posterior for $\Sigma_{v_1}^{-1}$ is $\mathbb{W}((\bar{\Sigma}_{v_1}^{-1} + \sum_t v_{1t} v_{1t}')^{-1}, \bar{\nu}_{v_1} + T)$, and that for $\sigma_{v_{2i}}^{-2}$ is $\mathbb{G}(a_1 + T, a_2 + \sum_t (\ln \sigma_{it}^2 - \ln \sigma_{it-1}^2)^2)$.

Example 11.5. Suppose $y_t = \alpha_t y_{t-1} + e_t, e_t \sim$ i.i.d. $\mathbb{N}(0, \sigma_t^2), \alpha_t = \rho \alpha_{t-1} + v_{1t}$, $v_{1t} \sim$ i.i.d. $\mathbb{N}(0, \sigma_{v_1}^2)$, $\ln \sigma_t^2 = \ln \sigma_{t-1}^2 + \sigma_{v_2} v_{2t}$, $v_{2t} \sim$ i.i.d. $\mathbb{N}(0, 1)$. If $\sigma_{v_2}^{-2} \sim \mathbb{G}(a_{v_2}, b_{v_2})$ and $\sigma_{v_1}^{-2} \sim \mathbb{G}(a_{v_1}, b_{v_1})$, then, given ρ, the conditional posteriors of $(\sigma_{v_1}^{-2}, \sigma_{v_2}^{-2})$ are gamma with parameters $(a_{v_1} + T, b_{v_1} + \sum_t v_{1t}^2)$ and $(a_{v_2} + T, \bar{b}_{v_2} + \sum_t (\ln \sigma_t^2 - \ln \sigma_{t-1}^2)^2)$, respectively.

Exercise 11.14. Derive the conditional posteriors of $(\rho, \sigma_{v_1}^{-2}, \sigma_{v_2}^{-2})$ in example 11.5 when ρ is unknown and has prior $\mathbb{N}(\bar{\rho}, \bar{\sigma}_\rho^2) \mathcal{I}_{(-1,1)}$, where $\mathcal{I}_{(-1,1)}$ is an indicator for stationarity.

To construct the conditional of ϕ, note that, if $\epsilon_t \sim (0, \Sigma_t)$, then $e_t = \mathcal{P}\epsilon_t \sim (0, \mathcal{P}\Sigma_t\mathcal{P}')$. Hence, if e_t is known, and given (y_t, x_t, α_t), the free elements of $\mathcal{P}$ can be estimated as follows. Since $\mathcal{P}$ is lower triangular, the mth equation is

$$\sigma_{mt}^{-1} e_{mt} = \phi_{m1}(-\sigma_{mt}^{-1} e_{1t}) + \cdots + \phi_{m,m-1}(-\sigma_{mt}^{-1} e_{m-1t}) + (\sigma_{mt}^{-1} \epsilon_{mt}). \quad (11.16)$$

Hence, letting $E_{mt} = (-\sigma_{mt}^{-1} e_{1t}, \ldots, -\sigma_{mt-1}^{-1} e_{mt})$, $\varepsilon_{mt} = -\sigma_{mt}^{-1} \epsilon_{mt}$, it is easy to see that the conditional posterior for ϕ_i is normal with mean $\tilde{\phi}_i$ and variance $\tilde{\Sigma}_{\phi_i}$.

Exercise 11.15. Show the form of $\tilde{\phi}_i$ and $\tilde{\Sigma}_{\phi_i}$.

To draw σ_{it}^2 from its conditional distribution, let $\sigma_{(-i)t}^2$ be the sequence of σ_t^2 excluding its ith element and let $e = (e_1, \ldots, e_t)$. Then $g(\sigma_{it}^2 \mid \sigma_{(-i)t}^2, \sigma_{\epsilon_i}, e) = g(\sigma_{it}^2 \mid \sigma_{it-1}^2, \sigma_{it+1}^2, \sigma_{\epsilon_i}, e)$, which is given in (11.12). To draw from this distribution for each i we could choose as candidate distribution $\sigma_{it}^{-2} \exp\{-(\ln \sigma_{it}^2 - E_t(\ln \sigma_{it}^2))^2 / 2 \operatorname{var}(\ln \sigma_t^2)\}$ and accept or reject the draw with probability $(\sigma_{it}^\dagger)^{-1} \times \exp\{-e_{it}^2 / 2(\sigma_{it}^2)^\dagger\} / (\sigma_{it}^{\ell-1})^{-1} \exp\{-e_{it}^2 / 2(\sigma_{it}^2)^{l-1}\}$, where $(\sigma_{it}^2)^{l-1}$ is the last draw and $(\sigma_{it}^2)^\dagger$ is the candidate draw.

Exercise 11.16. Suppose you are interested in predicting future values of y_t. Let $y^{t+\tau} = (y_{t+1}, \ldots, y_{t+\tau})$, $\alpha = (\alpha_1, \ldots, \alpha_t)$, and $y = (y_1, \ldots, y_t)$. Show that, conditional on time t information,

$$\begin{aligned} g(y^{t+\tau} \mid \alpha, \Sigma_t^\dagger, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y) \\ = \iint g(\alpha^{t+\tau} \mid \alpha, \Sigma_t^\dagger, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y) \\ \times g(\Sigma^{\dagger,t+\tau} \mid \alpha^{t+\tau}, \Sigma_t^\dagger, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y) \\ \times f(y^{t+\tau} \mid \alpha^{t+\tau}, \Sigma^{\dagger,t+\tau}, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y) \, \mathrm{d}\alpha^{t+\tau} \, \mathrm{d}\Sigma^{\dagger,t+\tau}. \end{aligned}$$

Describe how to sample (y_{t+1}, y_{t+2}) from this distribution. How would you construct a 68% prediction band?

Stochastic volatility models are typically used to infer values for the unobservable conditional volatilities, both in-sample (smoothing) and out-of-sample (prediction). For example, option pricing formulas require estimates of conditional volatilities and event studies often relate specific occurrences to changes in volatility. Here we concentrate on the smoothing problem, that is, on the computation of $g(\sigma_t^2 \mid y)$, where $y = (y_1, \ldots, y_T)$. An analytic expression for this posterior density is not available but since $g(\sigma_t^2 \mid y) = \int g(\sigma_t^2, \mid \alpha_t, y) g(\alpha_t \mid y)\, \mathrm{d}\alpha_t$ it can be numerically obtained by using the draws of σ_t^2 and α_t. The mean of this distribution can be used as an estimate of the smoothed volatility.

Exercise 11.17. Suppose the volatility model is $\ln \sigma_t^2 = \rho_0 + \rho(\ell) \ln \sigma_{t-1}^2 + \sigma_v v_t$, where $\rho(\ell)$ is unknown of order q. Show how to extend the Gibbs sampler to this case. Assume now that the model is of the form $\ln \sigma_t^2 = \rho_0 + \rho_1 \ln \sigma_{t-1}^2 + \sigma_{v_t} v_t$, where $\sigma_{v_t} = f(x_t)$, x_t are observable variables, and f is linear. Show how to extend the Gibbs sampler to this case.

As with factor models, cycling through the conditionals of $(\Sigma_t^{\dagger}, \alpha_t, \sigma_{v_{2i}}, \Sigma_{v_1}, \phi)$ with the Gibbs sampler produces, in the limit, a sample from the joint posterior.

Uhlig (1994) proposed an alternative specification for a stochastic volatility model which, together with a particular distribution of the innovations of the stochastic volatility term, produces closed-form solutions for the posterior distribution of the parameters and of the unknown vector of volatilities. The approach treats some parameters in the stochastic volatility equation as fixed but has the advantage of producing recursive estimates of the quantities of interest.

Consider an m-variable VAR(q) with stochastic volatility of the form

$$\left.\begin{aligned} Y_t &= AX_t + \mathcal{P}_t^{-1} e_t, \quad && e_t \sim \mathbb{N}(0, I), \\ \Sigma_{t+1} &= \frac{\mathcal{P}_t' v_t \mathcal{P}_t}{\rho}, && v_t \sim \text{Beta}((\nu + k)/2, 1/2), \end{aligned}\right\} \tag{11.17}$$

where X_t contains the lags of the endogenous and the exogenous variables, $\mathcal{P}_t$ is the upper Choleski factor of Σ_{t+1}, ν and ρ are (known) parameters, Beta denotes the m-variate beta distribution, and k is the number of parameters in each equation.

To construct the posterior of the parameters of (11.17) we need a prior for (A, Σ_1). We assume $g_1(A, \Sigma_1) \propto g_0(A) g(A, \Sigma_1 \mid \bar{A}_0, \rho\bar{\Sigma}_A, \bar{\Sigma}_0, \bar{\nu})$, where $g_0(A)$ is a function restricting the prior for A (e.g., to be stationary) and $g(\alpha, \Sigma_1 \mid \bar{A}_0, \rho\bar{\Sigma}_A, \bar{\Sigma}_0, \bar{\nu})$ is of normal-Wishart form, i.e., $g(A \mid \Sigma_1) \sim \mathbb{N}(\bar{A}_0, \rho\bar{\Sigma}_A)$, $g(\Sigma_1^{-1}) \sim \mathbb{W}(\bar{\Sigma}_0, \bar{\nu})$, $\bar{A}_0$, $\bar{\Sigma}_0$, $\bar{\Sigma}_A$, $\bar{\nu}$, ρ known.

Combining the likelihood of (11.17) with these priors and exploiting the fact that the beta distribution conjugates with the gamma distribution, we have that the posterior kernel for (A, Σ_{t+1}) is $\grave{g}_t(A, \Sigma_{t+1}) = \grave{g}_t(A) \grave{g}(A, \Sigma_{t+1} \mid \tilde{A}_t, \rho\tilde{\Sigma}_{At}, \tilde{\Sigma}_t, \nu)$,

where $\grave{g}$ is of normal-Wishart type, $\tilde{\Sigma}_{At} = \rho\tilde{\Sigma}_{At-1} + X_t X_t'$, $\tilde{A}_t = (\rho\tilde{A}_{t-1}\tilde{\Sigma}_{At-1} + Y_t X_t')\tilde{\Sigma}_{At}^{-1}$, $\tilde{\Sigma}_t = \rho\tilde{\Sigma}_{t-1} + (\rho/\nu)e_t(1 - X_t'\tilde{\Sigma}_{At}^{-1}X_t)\tilde{e}_t'$, $\tilde{e}_t = Y_t - \tilde{A}_{t-1}X_t$, and $\grave{g}_t(A) = \grave{g}_{t-1}(A)|(A - \tilde{A}_t)\tilde{\Sigma}_{At}(A - \tilde{A}_t)' + (\nu/\rho)\tilde{\Sigma}_t|^{-0.5}$.

Example 11.6. Consider a univariate AR(1) version of (11.17) of the form

$$y_t = \alpha y_{t-1} + \sigma_t^{-1}e_t, \quad e_t \sim \mathbb{N}(0,1), \tag{11.18}$$

$$\rho\sigma_{t+1}^2 = \sigma_t^2 v_t, \qquad v_t \sim \text{Beta}((\nu+1)/2, 1/2). \tag{11.19}$$

Let $g(\alpha, \sigma_1^2) \propto g_0(\alpha)g(\alpha, \sigma_1^2 \mid \bar{\alpha}_0, \rho\bar{\sigma}_{\alpha_0}^2, \bar{\sigma}_0^2, \bar{\nu})$, where $(\bar{\alpha}_0, \sigma_{\alpha_0}, \bar{\sigma}_0, \bar{\nu})$ are hyperparameters and assume that $g(\alpha, \sigma_1^2 \mid \bar{\alpha}_0, \rho\bar{\sigma}_{\alpha_0}^2, \bar{\sigma}_0^2, \bar{\nu})$ is of normal-inverted gamma type. Recursive posterior estimates of the parameters of $g_t(\alpha)$ are $\tilde{\sigma}_{\alpha,t}^2 = \rho\tilde{\sigma}_{\alpha,t-1}^2 + y_{t-1}^2$, $\tilde{\alpha}_t = (\rho\tilde{\alpha}_{t-1}\sigma_{\alpha,t-1}^2 + y_t y_{t-1})/\sigma_{\alpha,t}^2$, $\tilde{\sigma}_t^2 = \rho\tilde{\sigma}_{t-1}^2 + (\rho/\nu)\tilde{e}_t^2(1 - y_{t-1}^2/\sigma_{\alpha,t}^2)$, $\tilde{e}_t = y_t - \tilde{\alpha}_{t-1}y_{t-1}$, $g_t(\alpha) = g_{t-1}(\alpha)[(\alpha - \tilde{\alpha}_t)^2\sigma_{\alpha,t}^2 + (\nu/\rho)\sigma_t^2]^{-0.5}$. Hence both $\tilde{\sigma}_{\alpha_t}^2$ and $\tilde{\alpha}$ are weighted averages, with ρ measuring the memory of the process. Note that past values of $\tilde{\alpha}$ are weighted by the relative change in $\tilde{\sigma}_{\alpha,t}^2$. When $\sigma_{\alpha,t}^2$ is constant, $\tilde{\alpha}_t = \rho\tilde{\alpha}_{t-1} + y_t y_{t-1}/\rho\sigma_\alpha^2$.

When $\rho = \nu/(\nu+1)$, $\nu/\rho = 1 - \rho$. In this case, $\tilde{\sigma}_t^2$ is a weighted average of $\tilde{\sigma}_{t-1}^2$ and the information contained in the square of the recursive residuals, adjusted for the relative size of y_t^2, to the weighted sum of y_{t-1}^2 up to $t-1$. Note also that $E_{t-1}\sigma_t^2 = \sigma_{t-1}^2(\nu+1)/\rho(\nu+2)$. Hence, when $\rho = (\nu+1)/(\nu+2)$, σ_t^2 is a random walk.

For comparison, it may be useful to map the general prior of (11.17) into a Minnesota-type prior. For example, we could set $\bar{\Sigma}_0 = \text{diag}\{\bar{\sigma}_{0i}\}$ and compute $\bar{\sigma}_{0i}$ from the average square residuals of an AR(1) regression for each i in a training sample. Also, one could set $\bar{\Sigma}_A = \text{blockdiag}[\bar{\Sigma}_{A1}, \bar{\Sigma}_{A2}]$, where the split reflects the distinction between endogenous and exogenous variables. For example, if the second block contains a constant and linear trend, then

$$\bar{\Sigma}_{A2} = \begin{bmatrix} \phi_2 & -\phi_2^2/2 \\ \phi_2^2/2 & -\phi_2^3/3 \end{bmatrix},$$

where ϕ_2 is a hyperparameter, while we could set the diagonal elements of Σ_{A1} equal to $\theta_0^2\theta_1^2/\ell$, where ℓ refers to the lag, and ϕ_1 for the lags of the variables in an equation, and the off-diagonal elements to zero. Unless required by the problem, set $g_0(A) = 1$. Finally, set $\nu \approx 20$ for quarterly data and $\rho = \nu/(\nu+1)$.

Given the generic structure for the posterior of (A_t, Σ_{t+1}) (a time-varying density multiplied by a normal-Wishart density), we need numerical methods to draw posterior sequences. Any of the approaches described in chapter 9 will do it.

Example 11.7. To draw from the posterior we could use the following importance sampling algorithm.

(1) Find the marginal for A_T. Integrating Σ_{T+1} out of $\breve{g}(A_T, \Sigma_{t+1} \mid y)$ we have $\breve{g}(A_T \mid y) = 0.5\sum_t \ln|(A - \tilde{A}_T)\tilde{\Sigma}_{AT}(A - \tilde{A}_T)' + (\nu/\rho)\Sigma_T| - 0.5(k+\nu)|\times (A - \tilde{A}_T)\tilde{\Sigma}_{AT}(A - \tilde{A}_T)' + (\nu/\rho)\Sigma_T|$.

(2) Find the mode of $\breve{g}(A_T \mid y)$ (call it A_T^*) and compute the Hessian at the mode.

(3) Conditional on A_T, $g(\Sigma_{T+1}^{-1} \mid y)$ is $\mathbb{W}([\rho(A-\tilde{A}_T)\tilde{\Sigma}_{AT}(A-\tilde{A}_T)'+\nu\tilde{\Sigma}_T]^{-1}, \nu+k)$.

(4) Draw A_T^l from a multivariate t-distribution centered at A_T^* and with variance equal to the Hessian at the mode and degrees of freedom $\nu \ll T-k(M+1)$. Draw $(\Sigma_{T+1}^{-1})^l$ from the Wishart distribution derived in step (3).

(5) Calculate the importance ratio: $\ln \mathrm{IR}(A_T^l, \Sigma_{T+1}^l) = \text{const.} + \ln(\breve{g}(A_T^l)) - \ln(\breve{g}^{\mathrm{IS}}(A_T^l))$, where $g^{\mathrm{IS}}(A^l)$ is the value of the importance sampling density at A^l.

(6) Use $\bar{h}_L = \sum_{l=1}^L h(A_T^l, \Sigma_{T+1}^l)\mathrm{IR}(A_T^l, \Sigma_{T+1}^l)/\sum_{l=1}^L \mathrm{IR}(A_T^l, \Sigma_{T+1}^l)$ to approximate any function $h(A_T, \Sigma_{T+1})$.

Exercise 11.18. Describe an MH algorithm to draw posterior sequences for (A_T, Σ_{T+1}).

Exercise 11.19 (Cogley). Consider a bivariate model with consumption and income growth of the form $y_t = \bar{y} + A_t(\ell)y_{t-1} + e_t, \alpha_t \equiv \mathrm{vec}(A_t(\ell)) = \alpha_{t-1} + v_{1t}$, $\Sigma_t = \mathrm{diag}\{\sigma_{it}^2\}$, $\ln \sigma_t^2 = \ln \sigma_{t-1}^2 + \sigma_{v_2} v_{2t}$, where $\bar{y}$ is a constant. In a constant-coefficient version of the model the trend growth rate of the two variables is $(I - A(\ell))^{-1}\bar{y}$. Using a Gibbs sampler, describe how to construct a time-varying estimate of the trend growth rate, $(I - A_t(\ell))^{-1}\bar{y}$.

We conclude this section applying Bayesian methods to the estimation of the parameters of a GARCH model.

Example 11.8. Consider the model $y_t = x_t'A + \sigma_t e_t$, $e_t \sim$ i.i.d. $\mathbb{N}(0,1)$, and $\sigma_t^2 = \rho_0 + \rho_1\sigma_{t-1}^2 + \rho_2 e_{t-1}^2$. Assume that $A \sim \mathbb{N}(\bar{A}, \bar{\sigma}_A^2)$, $\rho_0 \sim \mathbb{N}(\bar{\rho}_0, \bar{\sigma}_{\rho_0}^2)$, and that $g(\rho_1, \rho_2)$ is uniform over $[0,1]$ and restricted so that $\rho_1 + \rho_2 \leqslant 1$. The posterior kernel can be easily constructed from these densities. Let $\alpha = (A, \rho_i, i = 0,1,2)$; let the mode of the posterior be α^*, and let $\breve{t}(\cdot)$ be the kernel of a t-distribution with location α^*, scale proportional to the Hessian at the mode, and $\bar{\nu}$ degrees of freedom. Posterior draws for the parameters can be obtained by using, for example, an independence Metropolis algorithm, that is, generate $\alpha^\dagger$ from $\breve{t}(\cdot)$ and accept the draw with probability equal to $\min\{[\breve{g}(\alpha^\dagger \mid y_t)/\breve{t}(\alpha^\dagger)]/[\breve{g}(\alpha^{l-1} \mid y_t)/\breve{t}(\alpha^{l-1})], 1\}$. A t-distribution is appropriate in this case because $\breve{g}(\alpha \mid y_t)/\breve{t}(\alpha)$ is typically bounded from above.

11.3 Markov Switching Models

Markov switching models are extensively used in macroeconomics, in particular, when important relationships are suspected to be functions of an unobservable variable (e.g., the state of a business cycle). Hamilton (1994) provides a classical nonlinear filtering method which can be used to obtain estimates of the parameters and of the unobservable state. Here we consider a Bayesian approach to the problem.

As with factor and stochastic volatility models, the unobservable state is treated as "missing" data and sampled together with other parameters in the Gibbs sampler.

To set up ideas we start from a static model where the slope varies with the state:

$$y_t = x_{1t}A_1 + x_{2t}A_2(\varkappa_t - 1) + e_t, \quad e_t \sim \text{i.i.d. } \mathbb{N}(0, \sigma_e^2). \tag{11.20}$$

Here $\varkappa_t$ is a two-state Markov switching indicator. We take $\varkappa_t = 1$ to be the normal state so that $y_t = x_{1t}A_1 + e_t$. In the extraordinary state, $\varkappa_t = 0$ and $y_t = x_{1t}A_1 - x_{2t}A_2 + e_t$.

We let $p_1 = P(\varkappa_t = 1 \mid \varkappa_{t-1} = 1)$, $p_2 = p(\varkappa_t = 0 \mid \varkappa_{t-1} = 0)$, both of which are unknown; also we let $y^{t-1} = (y_1, \ldots, y_{t-1}, x_{11}, \ldots, x_{1t-1}, x_{21}, \ldots, x_{2t-1})$, $\varkappa^t = (\varkappa_1, \ldots, \varkappa_t)$, $\alpha = (A_1, A_2, \sigma_e^2, \varkappa^t, p_1, p_2)$. We want to obtain the posterior for α. We assume $g(\alpha) = g(A_1, A_2, \sigma_e^2)g(\varkappa^t \mid p_1, p_2)g(p_1, p_2)$. We let $g(p_1, p_2) = p_1^{\bar{d}_{11}}(1-p_1)^{\bar{d}_{12}}p_2^{\bar{d}_{22}}(1-p_2)^{\bar{d}_{21}}$, where $\bar{d}_{ij}$ are the *a priori* proportions of the (i, j) elements in the sample. As usual, we assume $g(A_1, A_2, \sigma_e^{-2}) \propto \mathbb{N}(\bar{A}_1, \bar{\Sigma}_1) \times \mathbb{N}(\bar{A}_2, \bar{\Sigma}_2) \times \mathbb{G}(a_1, a_2)$.

The posterior kernel is $\breve{g}(\alpha \mid y) = \sum_{t=1}^{T} f(y_t \mid \alpha, y^{t-1})g(\alpha)$, where each $f(y_t \mid \alpha, y^{t-1}) \sim \mathbb{N}(Ax_t, \sigma_e^2)$, $x_t = (x_{1t}, x_{2t})$, and $A = (A_1, A_2)$. To sample from this kernel we need starting values for α and $\varkappa_t$ and the following algorithm.

Algorithm 11.2.

(1) Sample (p_1, p_2) from $g(p_1, p_2 \mid y) = p_1^{\bar{d}_{11}+d_{11}}(1-p_1)^{\bar{d}_{12}+d_{12}}p_2^{\bar{d}_{22}+d_{22}} \times (1-p_2)^{\bar{d}_{21}+d_{21}}$, where d_{ij} is the actual number of shifts between state i and state j.

(2) Sample A_i from $\breve{g}(A_i \mid \sigma_e^2, \varkappa^T, y)$. This is the kernel of a normal with mean $\tilde{A} = \tilde{\Sigma}_A(\sum_t x_t y_t/\sigma^2 + \bar{\Sigma}_A^{-1}\bar{A})$ and variance $\tilde{\Sigma}_A = (\sum_t x_t' x_t/\sigma^2 + \bar{\Sigma}^{-1})^{-1}$, where $\bar{A} = (\bar{A}_1, \bar{A}_2)$ and $\bar{\Sigma} = \text{diag}(\bar{\Sigma}_1, \bar{\Sigma}_2)$.

(3) Sample σ_e^{-2} from $\breve{g}(\sigma_e^{-2} \mid \varkappa^T, y, A)$. This is the kernel of a gamma with parameters $a_1 + 0.5(T-1)$ and $a_2 + 0.5\sum_t (y_t - A_1 x_{1t} + A_2 x_{2t}(\varkappa_t - 1))^2$.

(4) Sample $\varkappa^T$ from $\breve{g}(\varkappa^T \mid y,\ A,\ \sigma_e^2,\ p_1,\ p_2)$. As usual we do this in two steps. Given $g(\varkappa_0)$ we run forward into the sample by using $g(\varkappa_t \mid A,\ \sigma_e^2,\ y^t,\ p_1,\ p_2) \propto f(y_t \mid y^{t-1},\ A,\ \sigma_e^2,\ \varkappa_t)\, g(\varkappa_t \mid A,\ \sigma_e^2,\ y^{t-1},\ p_1,\ p_2)$, where $f(y_t \mid y^{t-1},\ A,\ \sigma_e^2,\ \varkappa_t) \sim N(Ax_t, \sigma_e^2)$ and $g(\varkappa_t \mid A,\ \sigma_e^2,\ y^{t-1},\ p_1,\ p_2) = \sum_{\varkappa_{t-1}=0}^{1} g(\varkappa_{t-1} \mid A,\ \sigma_e^2,\ y^{t-1},\ p_1,\ p_2)P(\varkappa_t = i \mid \varkappa_{t-1} = j)$. Then, starting from $\varkappa_T$, we run backward in the sample to smooth estimates, that is, given $g(\varkappa_T \mid y^T, A, \sigma_e^2, p_1, p_2)$, we compute $g(\varkappa_\tau \mid \varkappa_{\tau+1}, y^\tau, A, \sigma_e^2, p_1, p_2) \propto g(\varkappa_\tau \mid A, \sigma_e^2, y^\tau, p_1, p_2)P(\varkappa_\tau = i \mid \varkappa_{\tau+1} = j)^{-1}$, $\tau = T-1, T-2, \ldots$. Note that we have used the Markov properties of $\varkappa_t$ to split the forward and backward problems of drawing T joint values into the problem of drawing T conditional values.

We can immediately see that step (4) of algorithm 11.2 is the same as the one we used to extract the unobservable state in state space models. In fact, the first

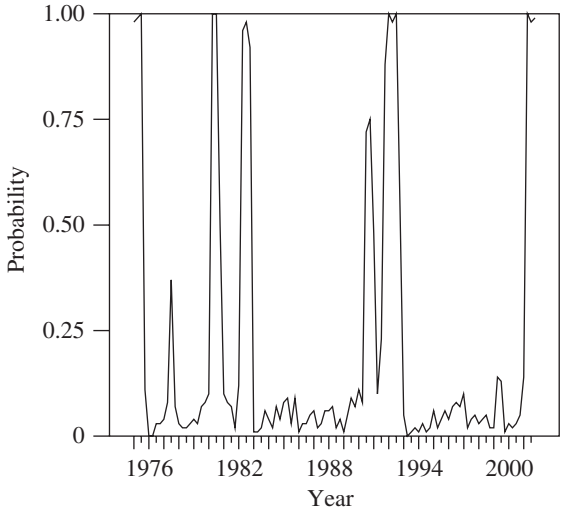

Figure 11.2. Recession probabilities.

part is similar to drawing the AR parameters in a factor model and the second to the estimation of the factor at each stage of the simulation. This is not surprising: a two-state Markov chain model can always be written as a first-order AR process with AR coefficient equal to $p_2 + p_1 - 1$. The difference, as already mentioned, is that the AR process here has binary innovations.

Exercise 11.20. Suppose that $g(p_1, p_2)$ is noninformative. Show the form of the conditional posterior of $(A_1, A_2, \sigma_e^{-2})$. Alter algorithm 11.2 to take into account this change.

Example 11.9. We use equation (11.20) to study fluctuations in EU industrial production. To construct an EU measure we aggregate IP data for Germany, France, and Italy by using GDP weights and let y_t be the yearly changes in industrial production. Data run from 1974:1 to 2001:4. The posterior means are $\tilde{A}_2 = 0.46$ and $\tilde{A}_1 = 0.96$ and the standard deviations are 0.09 for both coefficients. Hence, the annual growth rate in expansions is about two percentage points higher and the difference is statistically significant. Estimates of the probability of being in the extraordinary state (a "recession") are in figure 11.2: the algorithm picks up standard recessions and indicates the presence of a new contractionary phase starting in 2001:1.

11.3.1 A More Complicated Structure

The model we consider here is

$$A^y(\ell)(y_t - \bar{y}(\varkappa_t, x_t)) = \sigma(\varkappa_t)e_t, \tag{11.21}$$

where $A^y(\ell)$ is a polynomial in the lag operator, $\bar{y}(\varkappa_t, x_t)$ is the mean of y_t, which depends on observable regressors x_t and on the unobservable state $\varkappa_t$, $\text{var}(e_t) = 1$,

$\sigma(\varkappa_t)$ also depends on the unobservable state, and $\varkappa_t$ is a two-state Markov chain with transition matrix P. We set $\bar{y}(\varkappa_t, x_t) = x_t' A_0 + A_1 \varkappa_t$, $\sigma^2(\varkappa_t) = \sigma^2 + A_2 \varkappa_t$ and assume $A_2 > 0$, $A_1 > 0$ for identification purposes. Moreover, we restrict the roots of $A^y(\ell)$ to be less than 1.

Let $y^t = (y_1, \ldots, y_t)$, $\varkappa^t = (\varkappa_1, \ldots, \varkappa_t)$; let $\mathbb{A}$ be the companion matrix of $A^y(\ell)$ and $\mathbb{A}_1$ its first m rows. Define $\kappa = A_2/\sigma^2$ and let $\alpha = (A_0, A_1, \mathbb{A}_1, \sigma^2, \kappa, p_{ij})$. The likelihood function is $f(y^t \mid \varkappa^t, \alpha) = f(y^q \mid \varkappa^q, \alpha) \times \prod_{\tau=q+1}^{t} f(y_\tau \mid y^{\tau-1}, \varkappa^{t-1}, \alpha)$, where the first term is the density of the first q observations and the second term is the one-step-ahead conditional density of y_τ.

The density of the first q observations (see derivation in the factor model case) is normal with mean $x^q A_0 + \varkappa^q A_1$ and variance $\sigma^2 \Omega_q$, where $\Omega_q = W_q \Sigma_q W_q$, $\Sigma_q = \mathbb{A}\Sigma_q \mathbb{A}' + (1, 0, 0, \ldots, 0)'(1, 0, 0, \ldots, 0)$, $W_q = \operatorname{diag}\{(1 + \kappa \varkappa_j)^{0.5},\ j = 1, \ldots, q\}$. Using the prediction error decomposition we have that $f(y_\tau \mid y^{\tau-1}, \varkappa^{\tau-1}, \alpha) \propto \exp\{-(y_\tau - y_{\tau|\tau-1})^2 / 2\sigma^2(\varkappa_\tau)\}$, where $y_{\tau|\tau-1} = (1 - A^y(\ell)) y_t + A^y(\ell)(x_\tau' A_0 + A_1 \varkappa_\tau)$. Therefore, y_t is conditionally normal with mean $y_{t|t-1}$ and variance $\sigma^2(\varkappa_t)$. Finally, the joint density of $(y^t, \varkappa^t)$ is $f(y^t \mid \varkappa^t, \alpha) \prod_{\tau=2}^{t} f(\varkappa_\tau \mid \varkappa_{\tau-1}) f(\varkappa_1)$ and the likelihood of the data is $\int f(y^t, \varkappa^t \mid \alpha)\, \mathrm{d}\varkappa^t$. In chapter 3 we produced estimates of $(\alpha, \varkappa^t)$ by using a two-step approach: in the first step α_{ML} is obtained by maximizing the likelihood function; in the second step, inference about $\varkappa^t$ is obtained conditional on α_{ML}. That is,

$$
\begin{aligned}
&f(\varkappa_t, \ldots, \varkappa_{t-\tau+1} \mid y^t, \alpha_{\mathrm{ML}}) \\
&\quad = \sum_{\varkappa_{t-\tau}=0}^{1} f(\varkappa_t, \ldots, \varkappa_{t-\tau} \mid y^{t-1}, \alpha_{\mathrm{ML}}) \\
&\quad \propto f(\varkappa_t \mid \varkappa_{t-1}) f(\varkappa_{t-1}, \ldots, \varkappa_{t-\tau} \mid y^{t-1}, \alpha_{\mathrm{ML}}) f(y_t \mid y^{t-1}, \varkappa^t, \alpha_{\mathrm{ML}}),
\end{aligned}
\tag{11.22}
$$

where the factor of proportionality is given by $f(y_t \mid y^{t-1}, \alpha_{\mathrm{ML}}) = \sum_{\varkappa_t} \cdots \sum_{\varkappa_{t-\tau}} f(y_t, \varkappa_t, \ldots, \varkappa_{t-\tau} \mid y^{t-1}, \alpha_{\mathrm{ML}})$. Since the log likelihood of the sample is $\ln f(y_{q+1}, \ldots, y_t \mid y^q, \alpha) = \sum_\tau \ln f(y_\tau \mid y^{\tau-1}, \alpha)$, once α_{ML} is obtained, transition probabilities can be computed by using $f(\varkappa_t \mid y^t, \alpha_{\mathrm{ML}}) = \int \cdots \int f(\varkappa_t, \ldots, \varkappa_{t-\tau+1} \mid y^t, \alpha_{\mathrm{ML}})\, \mathrm{d}\varkappa_{t-1} \cdots \mathrm{d}\varkappa_{t-\tau+1}$. Note that in this case uncertainty in α_{ML} is not incorporated in the calculations.

To construct the conditional posteriors of the parameters and of the unobservable state, assume that $g(A_0, A_1, \sigma^{-2}) \propto \mathbb{N}(\bar{A}_0, \bar{\Sigma}_{A_0}) \mathbb{N}(\bar{A}_1, \bar{\Sigma}_{A_1}) \mathcal{I}_{(A_1 > 0)} \mathbb{G}(a_1^\sigma, a_2^\sigma)$, where $\mathcal{I}_{(A_1>0)}$ is an indicator function. Further assume that $g((1+\kappa)^{-1}) \sim \mathbb{G}(a_1^\kappa, a_2^\kappa) \mathcal{I}_{(\kappa>0)}$ and $g(\mathbb{A}_1) \sim \mathbb{N}(\bar{\mathbb{A}}_1, \bar{\Sigma}_{\mathbb{A}_1}) \mathcal{I}_{(-1,1)}$, where $\mathcal{I}_{(-1,1)}$ is an indicator for stationarity. Finally, we let $p_{12} = 1 - p_{11} = 1 - p_1$ and $p_{21} = 1 - p_{22} = 1 - p_2$ and $g(p_i) \propto \mathrm{Beta}(\bar{d}_{i1}, \bar{d}_{i2})$, $i = 1, 2$, and assume that all hyperparameters are known.

Exercise 11.21. Let $\alpha_{-\psi}$ be the vector α except for ψ and let $A = (A_0, A_1)$.

(i) Assuming that the first q observations come from the low state, show that the conditional posteriors for the parameters and the unobserved state are

$$\left.\begin{aligned} g(A \mid y^t, \varkappa^t, \alpha_{-A}) &\sim \mathbb{N}(\tilde{A}, \tilde{\Sigma}_A)\mathcal{I}_{A_1>0}, \\ g(\sigma^{-2} \mid y^t, \varkappa^t, \alpha_{-\sigma^2}) &\sim \mathbb{G}(a_1^\sigma + T, a_2^\sigma \\ &\quad + (\Sigma_q^{-0.5} y - \Sigma_q^{-0.5} x A_0 + \Sigma_q^{-0.5} \varkappa A_1)^2), \\ g((1+\kappa)^{-1} \mid y^t, \varkappa^t, \alpha_{-\kappa}) &\sim \mathbb{G}(a_1^\kappa + T_1, a_2^\kappa + \text{rss})\mathcal{I}_{(\kappa>0)}, \\ g(\mathbb{A}_1 \mid y^t, \varkappa^t, \alpha_{-\mathbb{A}_1}) &\sim \mathbb{N}(\tilde{\mathbb{A}}_1, \tilde{\Sigma}_{\mathbb{A}_1})\mathcal{I}_{(-1,1)}|\Omega_q|^{-0.5} \\ &\quad \times \exp\{-(y^q - x^q A)'\Omega_q^{-1}(y^q - x^q A)/2\sigma^2\}, \\ g(p_i \mid y^t, \varkappa^t, \alpha_{-p_i}) &\sim \text{Beta}(\bar{d}_{i1} + d_{i1}, \bar{d}_{i2} + d_{i2}), \quad i = 1, 2, \\ g(\varkappa_t \mid y^t, \alpha_{-\varkappa_t}) &\propto f(\varkappa_t \mid \varkappa_{t-1}) f(\varkappa_{t+1} \mid \varkappa_t) \prod_\tau f(y_\tau \mid y^{\tau-1}, \varkappa^\tau), \end{aligned}\right\} \tag{11.23}$$

where T_1 is the number of elements in T for which $\varkappa_t = 1$, d_{ij} is the number of actual transitions from state i to state j, and $\text{rss} = \sum_{t=1}^{T_1}\{[(1-\kappa x_t^{0.5})(y - x_t' A_0 - \varkappa_t A_1)]/2\}$.

(ii) Show the exact form of $\tilde{\mathbb{A}}_1$, $\tilde{\Sigma}_{\mathbb{A}_1}$, $\tilde{A}$, and $\tilde{\Sigma}_A$.

(iii) Describe how to draw $\mathbb{A}_1$ and A restricted to the correct domain.

Recently, Sims (2001) and Sims and Zha (2004) have used a similar specification to estimate a Markov switching VAR model, where the switch may occur in the lagged dynamics, in the contemporaneous effects, or in both. To illustrate their approach consider the equation

$$A_1(\ell) i_t = \bar{i}(\varkappa_t) + b(\varkappa_t) A_2(\ell)\pi_t + \sigma(\varkappa_t) e_t, \tag{11.24}$$

where $e_t \sim$ i.i.d. $\mathbb{N}(0, 1)$, i_t is the nominal interest rate, π_t is inflation, and $\varkappa_t$ has three states with transition

$$P = \begin{bmatrix} p_1 & 1-p_1 & 0 \\ (1-p_2)/2 & p_2 & (1-p_2)/2 \\ 0 & 1-p_3 & p_3 \end{bmatrix}.$$

The model (11.24) imposes restrictions on the data: the dynamics of interest rates do not depend on the state; the form of the lag distribution on π_t is the same across states, except for a scale factor $b(\varkappa)$; there is no possibility of jumping from state 1 to state 3 (or vice versa) without passing through state 2; finally, the nine elements of P depend only on three parameters.

Let $\alpha = [\text{vec}(A_1(\ell)), \text{vec}(A_2(\ell)), \bar{i}(\varkappa_t), b(\varkappa_t), \sigma(\varkappa_t), p_1, p_2, p_3]$. The marginal likelihood of the data, conditional on the parameters (but integrating out the unobservable state) can be computed numerically and recursively. Let $\mathcal{F}_t$ be the information set at t.

Exercise 11.22. Show that $f(i_t, \varkappa_t \mid \mathcal{F}_{t-1})$ is a mixture of continuous and discrete densities. Show the form of $f(i_t \mid \mathcal{F}_{t-1})$, the marginal of the data, and of $f(\varkappa_t \mid \mathcal{F}_t)$, the updating density.

Once $f(\varkappa_t \mid \mathcal{F}_t)$ is obtained we can compute

$$f(\varkappa_{t+1} \mid \mathcal{F}_t) = \begin{bmatrix} f(\varkappa_t = 1 \mid \mathcal{F}_t) \\ f(\varkappa_t = 2 \mid \mathcal{F}_t) \\ f(\varkappa_t = 3 \mid \mathcal{F}_t) \end{bmatrix}' P$$

and from there we can calculate $f(i_{t+1}, \varkappa_{t+1} \mid i_t, \pi_t, \dots)$, which makes the recursion complete. Given a flat prior on α, the posterior will be proportional to $f(\alpha \mid i_t, \pi_t)$ and posterior estimates of the parameters and of the states can immediately be obtained.

Exercise 11.23. Provide formulas to obtain smoothed estimates of $\varkappa_t$.

More complicated VAR specifications are possible. For example, let $y_t \mathcal{A}_0(\varkappa_t) = x_t' \mathcal{A}_+(\varkappa_t) + e_t$, where x_t includes all lags of y_t and $e_t \sim$ i.i.d. $\mathbb{N}(0, I)$. Assume $\mathcal{A}_+(\varkappa_t) = \mathcal{A}(\varkappa_t) + [I, 0]' \mathcal{A}_0(\varkappa_t)$. Given this specification there are two possibilities: either $\mathcal{A}_0(\varkappa_t) = \bar{A}_0 \Lambda(\varkappa_t)$ and $\mathcal{A}(\varkappa_t) = \bar{\mathcal{A}} \Lambda(\varkappa_t)$ or $\mathcal{A}_0(\varkappa_t)$ free and $\mathcal{A}(\varkappa_t) = \bar{\mathcal{A}}$. In the first specification changes in the contemporaneous and lagged coefficients are proportional; in the second the state affects the contemporaneous relationship but not lagged ones.

Equation (11.24) is an equation of a bivariate VAR. Hence, so long as we are able to keep the posterior of the system in a SUR format (as we have done in chapter 10), the above ideas can be applied to each of the VAR equations.

11.3.2 A General Markov Switching Specification

Finally, we consider a general Markov switching specification which embeds as a special case the two previous ones. So far we have allowed the mean and the variance of y_t to change with the state but we have forced the dynamics to be independent of the state, apart from a scale effect. This is a strong restriction: in fact, it is conceivable that the autocovariance function of the data is different in expansions and in recessions.

The general two-state Markov switching model we consider is

$$y_t = \begin{cases} x_t' A_{01} + Y_t' A_{02} + e_{0t} & \text{if } \varkappa_t = 0, \\ x_t' A_{02} + Y_t' A_{12} + e_{1t} & \text{if } \varkappa_t = 1, \end{cases} \tag{11.25}$$

where x_t is a $1 \times q_2$ vector of exogenous variables for each t, $Y_t' = (y_{t-1}, \dots, y_{t-q_1})$ is a vector of lagged dependent variables and e_{jt}, $j = 0, 1$, are i.i.d. random variables, normally distributed with mean zero and variance σ_j^2. Once again the transition probability for $\varkappa_t$ has diagonal elements p_i. In principle, some of the elements of A_{ji} may be equal to zero for some i, so the model may have different dynamics in different states.

For identification, we choose the first state to be a "recession", so that $A_{02} < A_{12}$ is imposed. We let α_c be the parameters which are common across states, α_i the

parameters which are unique to the state, and $\alpha_{i\mathrm{r}}$ the parameters which are restricted to achieve identification. Then (11.25) can be written as

$$y_t = \begin{cases} X'_{ct}\alpha_{\mathrm{c}} + X'_{0t}\alpha_0 + X'_{rt}\alpha_{0r} + e_{0t} & \text{if } \varkappa_t = 0, \\ X'_{ct}\alpha_{\mathrm{c}} + X'_{1t}\alpha_1 + X'_{rt}\alpha_{1r} + e_{1t} & \text{if } \varkappa_t = 1, \end{cases} \tag{11.26}$$

where $(X'_{ct}, X'_{it}, X'_{rt}) = (x'_t, Y'_t)$ and $(\alpha'_{\mathrm{c}}, \alpha'_i, \alpha'_{i\mathrm{r}}) = (A'_{01}, A'_{02}, A'_{11}, A'_{12})$.

To construct conditional posteriors for the unknowns we assume conjugate priors: $\alpha_{\mathrm{c}} \sim \mathbb{N}(\bar{\alpha}_{\mathrm{c}}, \bar{\Sigma}_{\mathrm{c}})$; $\alpha_i \sim \mathbb{N}(\bar{\alpha}_i, \bar{\Sigma}_i)$; $\alpha_{i\mathrm{r}} \sim \mathbb{N}(\bar{\alpha}_{\mathrm{r}}, \bar{\Sigma}_{\mathrm{r}})\mathcal{I}_{\mathrm{rest}}$; $\bar{s}_i^2\sigma_i^{-2} \sim \chi^2(\bar{\nu}_i)$; $p_i \sim \mathrm{Beta}(d_{i1}, d_{i2})$, $i = 1, 2$, where $\mathcal{I}_{\mathrm{rest}}$ is a function indicating whether the identification restrictions are satisfied. As usual we assume that the hyperparameters $(\bar{\alpha}_{\mathrm{c}}, \bar{\Sigma}_{\mathrm{c}}, \bar{\alpha}_i, \bar{\Sigma}_i, \bar{\alpha}_{\mathrm{r}}, \bar{\Sigma}_{\mathrm{r}}, \bar{\nu}_i, \bar{s}_i^2, \bar{d}_{ij})$ are known or can be estimated from the data. We take the first $\max[q_1, q_0]$ observations as given in constructing the posterior distribution of the parameters and of the latent variable.

Given these priors, it is straightforward to compute conditional posteriors. For example, $g(\alpha_{\mathrm{c}} \mid x_t, y_t)$ has mean $\tilde{\alpha}_{\mathrm{c}} = \tilde{\Sigma}_{\mathrm{c}}(\sum_{t=\min[q_1,q_0]}^{T} X_{ct}y'_{ct}/\sigma_t^2 + \bar{\Sigma}_{\mathrm{c}}^{-1}\bar{\alpha}_{\mathrm{c}})$, variance $\tilde{\Sigma}_{\mathrm{c}} = (\sum_{t=\min[q_1,q_0]}^{T} X_{ct}X'_{ct}/\sigma_t^2 + \bar{\Sigma}_{\mathrm{c}}^{-1})^{-1}$, where $y_{ct} = y_t - X_{it}\alpha_i - X_{rt}\alpha_{i\mathrm{r}}$ and it is normal.

Exercise 11.24. Let T_i be the number of observations in state i.

(i) Show that the conditional posterior of α_i is $\mathbb{N}(\tilde{\alpha}_i, \tilde{\Sigma}_i)$, where $\alpha_i = \tilde{\Sigma}_i \times (\sum_{t=1}^{T_i} X_{it}y'_{it}/\sigma_t^2 + \bar{\Sigma}_i^{-1}\bar{\alpha}_i)$, $\tilde{\Sigma}_i = (\sum_{t=1}^{T_i} X_{it}X'_{it}/\sigma_t^2 + \bar{\Sigma}_i^{-1})^{-1}$, and $y_{it} = y_t - X_{ct}\alpha_{\mathrm{c}} - X_{rt}\alpha_{i\mathrm{r}}$.

(ii) Show that the conditional posterior of α_{r} is $\mathbb{N}(\tilde{\alpha}_{\mathrm{r}}, \tilde{\Sigma}_{\mathrm{r}})$. What are $\tilde{\alpha}_{\mathrm{r}}$ and $\tilde{\Sigma}_{\mathrm{r}}$?

(iii) Show that the conditional posterior of σ_i^{-2} is such that $(\bar{s}_i^2 + \mathrm{rss}_i^2)/\sigma_i^2 \sim \chi^2(\nu_i + T_i - \max[q_1, q_2])$. Write down the expression for rss_i^2.

(iv) Show that the conditional posterior for p_i is $\mathrm{Beta}(\bar{d}_{i1} + d_{i1}, \bar{d}_{i2} + d_{i2})$.

Finally, the conditional posterior for the latent variable $\varkappa_t$ can be computed as usual. Given the Markov properties of the model, we restrict attention to the subsequence $\varkappa_{t,\tau} = (\varkappa_t, \ldots, \varkappa_{t+\tau-1})$. Define $\varkappa_{t(-\tau)}$ as the sequence $\varkappa_t$ with the τth subsequence removed. Then $g(\varkappa_{t,\tau} \mid y, \varkappa_{t(-\tau)}) \propto f(y \mid \varkappa_t, \alpha, \sigma^2) \times g(\varkappa_{t,\tau} \mid \varkappa_{t(-\tau)}, p_i)$, which is a discrete distribution with 2^τ outcomes. Using the Markov property, $g(\varkappa_{t,\tau} \mid \varkappa_{t(-\tau)}, p_i) = g(\varkappa_{t,\tau} \mid \varkappa_{t-1}, \varkappa_{t+\tau}, p_i)$ while $f(y^T \mid \varkappa_t, \alpha) \propto \prod_{j=t}^{t+\tau-1}(1/\sigma_j)\exp\{-e_j^2/2\sigma_j^2\}$. Note that, since the $\varkappa_t$ are correlated, it is a good idea to choose $\tau > 1$.

Exercise 11.25. Write down the components of the conditional posterior for $\varkappa_t$ when $\tau = 1$.

In all Markov switching specifications, it is important to wisely select the initial conditions. One way to do so is to assign all the observations in the training sample to one state, obtain initial estimates for the parameters, and arbitrarily set the parameters of the other state to be equal to the estimates plus or minus a small number (say, 0.1). Alternatively, one can split the points arbitrarily but equally across the two states.

Exercise 11.26. Suppose $\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + e_t$, $e_t \sim$ i.i.d. $\mathbb{N}(0, \sigma_e^2)$ if $\varkappa_t = 0$ and $\Delta y_t = (\alpha_0 + A_0) + (\alpha_1 + A_1)\Delta y_{t-1} + e_t$, $e_t \sim$ i.i.d. $\mathbb{N}(0, (1 + A_2)\sigma_e^2)$ if $\varkappa_t = 1$. Using quarterly GDP growth data for the euro area, construct posterior estimates for A_0, A_1, A_2. Separately test if there is evidence of switching in the intercept, the dynamics, or the variance of Δy_t.

11.4 Bayesian DSGE Models

The use of Bayesian methods to estimate and evaluate Dynamic Stochastic General Equilibrium (DSGE) models does not present new theoretical aspects. We have repeatedly mentioned that DSGE models are false in at least two senses.

- They only provide an approximate representation to the DGP of the actual data. In particular, since the vector of structural parameters is typically of low dimension, strong restrictions are implied both in the short and in the long run.
- The number of driving forces is smaller than the number of endogenous variables so that the covariance matrix of a vector of variables generated by the model is singular.

These features make the estimation and testing of DSGE models with GMM or ML tricky. In fact, with these methods inference is (asymptotically) justified only when the model is the DGP of the data up to a set of unknown parameters, while stochastic singularity prevents numerical routines based on the Hessian from working properly in the search for the maximum of the objective function. In chapter 4 we described a minimalist approach, which only uses qualitative restrictions to identify shocks in the data, and can be employed to examine the match between the theory and the data, when the model is false in the two above senses.

Bayesian methods are also well-suited to dealing with false models. Posterior inference, in fact, does not hinge on the model being the correct DGP and it is feasible even when the covariance matrix of the vector of endogenous variables is singular — we do not need the Hessian to explore the shape of the posterior. Bayesian methods have another advantage over alternatives, which makes them appealing to macroeconomists. Posterior distributions in fact incorporate uncertainty about the parameters and the model specification.

Since log-linearized DSGE models are state space models with nonlinear restrictions on the mapping between reduced-form and structural parameters, posterior estimates of the structural parameters can be obtained, for appropriately designed prior distributions, by using the posterior simulators described in chapter 9. Given the nonlinearity of the mapping, Metropolis, or MH algorithms are generally employed. Numerical methods can also be used to compute marginal likelihoods and Bayes factors; to obtain any posterior function of the structural parameters (for example, impulse responses, variance decompositions, ACFs, turning-point predictions, and

forecasts) and to examine the sensitivity of the results to variations in the prior specification. Once the posterior distribution of the structural parameters is obtained, any interesting inferential exercise becomes trivial.

To estimate the posterior for the structural parameters and for the statistics of interest, and to evaluate the quality of a DSGE model, the following steps are typically used.

Algorithm 11.3.

(1) Construct a log-linear approximation to the DSGE economy and transform it into a state space model. Add measurement errors if the dimension of the vector of endogenous variables used in estimation/evaluation exceeds the dimension of the vector of driving forces of the model.

(2) Specify prior distributions for the structural parameters θ.

(3) Perform prior analysis to study the range of potential outcomes of the model.

(4) Draw sequences from the joint posterior of θ by using Metropolis or MH algorithms. Check convergence.

(5) Compute marginal likelihood numerically by using draws from the prior distribution and the Kalman filter. Compute the marginal likelihood for any alternative or reference model. Calculate Bayes factors or other measures of (relative) forecasting fit.

(6) Construct statistics of economic interest by using the draws in (4) (after an initial set has been discarded). Use loss-based measures to evaluate the discrepancy between the theory and the data.

(7) Examine the sensitivity of the results to the choice of priors.

Step (1) is unnecessary. We will see later on what to do if a nonlinear specification is used. Adding measurement errors helps computationally to reduce the singularity of the covariance matrix of the endogenous variables but it is not needed for the approach to work.

In step (2) prior distributions are generally centered around standard values of the parameters, while standard errors typically reflect subjective prior uncertainty. One could also specify objective prior standard errors, so as to "cover" the range of existing estimates, as we have done in chapter 7. For convenience, the prior distribution for the vector of parameters is assumed to be the product of univariate distributions of each of the parameters. In some applications, it may be convenient to select diffuse priors over a fixed range to avoid imposing too much structure on the data. In general, the form of the prior reflects computational convenience. Conjugate priors are typically preferred. For parameters which must lie in an interval, truncated normal or beta distributions are often chosen.

Step (3) logically precedes posterior analysis and can be used to evaluate whether models have any chance of producing the interesting features we observe in the actual data. This is precisely the analysis we performed in chapter 7, where we

compare statistics of the data with the range of statistics produced by models. While this step is often skipped, it may provide very useful information about the potential outcomes of the models.

Step (4) requires choosing an updating rule and a transition function $\mathfrak{P}(\theta^{\dagger}, \theta^{l-1})$ satisfying the regularity conditions described in chapter 9, estimating joint and marginal distributions by using kernel methods and the draws from the posterior, and checking convergence. In particular, the following steps are needed.

Algorithm 11.4.

(i) Given a $\theta^{\mathbf{0}}$, draw $\theta^{\dagger}$ from $\mathfrak{P}(\theta^{\dagger}, \theta^{\mathbf{0}})$, and compute the prediction error decomposition of the likelihood, i.e., estimate $f(y \mid \theta^{\mathbf{0}})$ and $f(y \mid \theta^{\dagger})$.

(ii) Evaluate the posterior kernel at $\theta^{\dagger}$ and $\theta^{\mathbf{0}}$, i.e., calculate $\breve{g}(\theta^{\dagger}) = f(y \mid \theta^{\dagger}) \times g(\theta^{\dagger})$ and $\breve{g}(\theta^{\mathbf{0}}) = f(y \mid \theta^{\mathbf{0}}) g(\theta^{\mathbf{0}})$.

(iii) Draw $\mathfrak{U} \sim \mathbb{U}(0, 1)$. If $\mathfrak{U} < \min\{[(\breve{g}(\theta^{\dagger})/\breve{g}(\theta^{\mathbf{0}}))][\mathfrak{P}(\theta^{\mathbf{0}}, \theta^{\dagger})/\mathfrak{P}(\theta^{\dagger}, \theta^{\mathbf{0}})], 1\}$, set $\theta^{\mathbf{1}} = \theta^{\dagger}$, otherwise set $\theta^{\mathbf{1}} = \theta^{\mathbf{0}}$.

(iv) Repeat steps (i)–(iii) $\bar{L} + JL$ times. Discard the first $\bar{L}$ draws, keep one draw every L for inference. Alternatively, repeat steps (i)–(iii) J times by using $\bar{L} + 1$ different $\theta^{\mathbf{0}}$, and keep the last draw from each run. Check convergence by using the methods described in chapter 9.

(v) Estimate marginal/joint posteriors with kernel methods. Compute location estimates and credible sets. Compare them with those computed from the prior.

Step (5) requires drawing parameters from the prior, calculating the sequence of prediction errors for each draw, and averaging over draws. To do so, one could use the modified harmonic mean, $\{(1/L) \sum_l [g^{\mathrm{IS}}(\theta^*)/f(y \mid \theta^*) g(\theta^*)]\}^{-1}$, suggested by Gelfand and Dey (1994), where θ^* is a point with high posterior probability and g^{IS} is a density with tail thinner than $f(y \mid \theta) g(\theta)$, or could use the Bayes theorem directly, as suggested by Chib (1995). Similar calculations can be undertaken for any alternative model and Bayes factors can then be numerically computed. When the dimensionality of the parameter space is large, Laplace approximations can reduce the computational burden and give a more accurate picture of the properties of various models. The competitors could be a structural model, which nests the one under consideration (e.g., a model with flexible prices can be obtained by restricting one parameter of a model with sticky prices), a nonnested structural specification (e.g., a model with sticky wages), or a more densely parametrized reduced-form model (e.g., a VAR or a BVAR).

In step (6) loss functions are needed to compare statistics of interest because DSGE models typically have low posterior probability. As we will see later on, posterior odds ratios may not be very informative in such a case.

In step (7), to check the robustness of the results to the choice of prior, one can reweigh the posterior draws by using the techniques described in section 9.5.

11.4.1 Identification

Since log-linearized DSGE models feature a nonlinear mapping between the parameters of the theory and those of the state space representation, and since there is no condition that can be easily employed to check the informational content of the data, any method which is concerned with the estimation of DSGE parameters must deal with potential identification problems. We have already seen aspects of such phenomena in chapters 5 and 6, when dealing with (classical) impulse response matching and maximum likelihood estimation. Since Bayesian inference is based on the likelihood principle, and since the model structure determines, to a large extent, whether parameters are identified or not, all the arguments previously made also apply to a Bayesian context. However, Bayesian methods have two important advantages over classical ones in the presence of identification problems: they can employ information from other data sets to reduce parameter underidentification; they can generate coherent inference even in the presence of identification problems.

Suppose that $\theta = [\theta_1, \theta_2]$, assume that $\Theta = \Theta_1 \times \Theta_2$, and suppose that the likelihood function has no information for θ_2, i.e., $f(y \mid \theta) = f^*(y \mid \theta_1)$. Straightforward application of the Bayes theorem implies that $g(\theta \mid y) = g(\theta_1 \mid y) \times g(\theta_2 \mid \theta_1) \propto f^*(y \mid \theta_1) g(\theta_1, \theta_2)$. Hence, a proper prior for θ can add curvature to a flat likelihood function. This facilitates both the maximization of the posterior, if needed, and its calculations with MCMC methods, and makes the posterior well-behaved. Nevertheless, there is no updating of the prior of $\theta_2 \mid \theta_1$. Hence, a comparison of the prior and the posterior of θ can indicate how informative the data are (priors and posteriors of identified parameters will be different, priors and posteriors of unidentified parameters will not). Furthermore, a sequence of prior distributions with different spreads can be used to assess the extent of identification problems. In fact, the posterior of parameters with dubious identification features will become more and more diffuse, while the posterior of identified parameters will hardly change.

When the space of parameters Θ is not variation free, i.e., $\Theta \neq \Theta_1 \times \Theta_2$, because of stability constraints or restrictions required for the solution to the model to generate nonimaginary time series, the prior of θ_2 could be marginally updated even when the likelihood has no information, since changes in the distribution of θ_2 imply that the domain of θ_1 changes (see, for example, Poirier 1998). In this situation, a comparison of priors and posteriors will not be informative about potential identification problems, unless the parameters constrained by economic requirements are known. This is unlikely to be true in DSGE setups since, for example, the eigenvalues which regulate stability are complicated functions of all the parameters of the model.

Complete lack of identification is typically limited to textbook examples. However, partial or weak identification problems are extremely common. Partial identification occurs when the likelihood displays a ridge in some dimension (see example 6.21), while weak identification implies that the likelihood function is flat in some or all dimensions. Both these phenomena are difficult to detect in practice

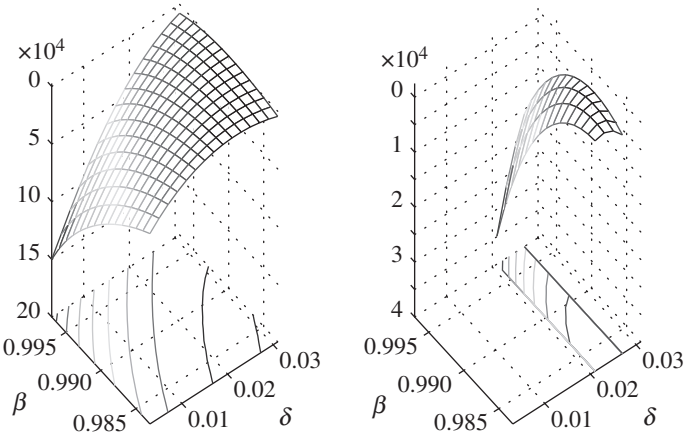

Figure 11.3. Likelihood and posterior, RBC model.

since, in the first case, it is the joint posterior which is indistinguishable from the joint prior (univariate posteriors may move away from univariate priors), while, in the second case, the size of the differences between the priors and the posteriors may depend on the details of MCMC routine employed.

As mentioned, well-behaved priors can induce well-behaved posteriors, even when the data have no information about the parameters. Therefore, it is very important that the priors of potentially nonidentifiable parameters truly contain information external to the data used to estimate the model and effectively reflect the objective uncertainty a researcher faces in specifying it. When these two general principles are not followed, Bayesian inference can mask rather than highlight identification problems. In fact, a sufficiently tight prior may give the illusion that parameter estimation is successful, that the model fits the data well, therefore creating the preconditions for its use for policy purposes. We show how this can occur with the model of example 6.21, which has a likelihood function with both flat sections and ridges.

Example 11.10. Figure 11.3 reproduces the likelihood function presented in the second panel of figure 6.1, which we have seen displays a ridge in β, δ running from $(\delta = 0.005, \beta = 0.975)$ up to $(\delta = 0.03, \beta = 0.99)$, and presents the joint posterior for these two parameters, when a sufficiently tight prior on δ is used. Clearly, while the likelihood has a diagonal ridge, the posterior appears to be much better behaved, since there is very low prior probability that δ lies outside the range $(0.018, 0.025)$.

While there may be reasonable economic arguments for *a priori* limiting the support of δ, they should be clearly spelled out. Furthermore, when bounds are imposed, the prior should be made reasonably uninformative to avoid misleading conclusions. Note that centering estimates at standard calibrated values is not the best strategy to follow since such values are likely to have been obtained with the same data that is employed for estimation, making the prior too data based.

11.4.2 Examples

Next, we present a few examples, highlighting the practical details of the implementation of Bayesian methods for inference in DSGE models.

Example 11.11. The first example is simple. We simulate data from a basic RBC model where the solution is contaminated by measurement errors. Armed with reasonable prior specifications for the structural parameters and a Metropolis algorithm, we examine where the posterior distribution of some crucial parameters lies relative to the "true" parameters we used in the simulations, when samples typical in macroeconomic data are available. We also compare true and estimated moments to give an economic measure of the fit we obtain.

The solution to an RBC model driven by i.i.d. technological disturbances when capital depreciates instantaneously, leisure does not enter the utility function, and the latter is logarithmic in consumption is

$$K_{t+1} = (1-\eta)\beta K_t^{1-\eta}\zeta_t + v_{1t}, \tag{11.27}$$

$$\mathrm{GDP}_t = K_t^{1-\eta}\zeta_t + v_{2t}, \tag{11.28}$$

$$c_t = \eta\beta\,\mathrm{GDP}_t + v_{3t}, \tag{11.29}$$

$$r_t = (1-\eta)\frac{\mathrm{GDP}_t}{K_t} + v_{4t}. \tag{11.30}$$

We have added four measurement errors v_{jt}, $j = 1, 2, 3, 4$, to the equations to reduce the singularity of the system and to mimic the typical situation an investigator is likely to face. Here β is the discount factor, $1-\eta$ the share of capital in production. We simulate 1000 data points by using $k_0 = 100.0$, $(1-\eta) = 0.36$, $\beta = 0.99$, $\ln\zeta_t \sim \mathbb{N}(0, \sigma_\xi^2 = 0.1)$, $v_{1t} \sim \mathbb{N}(0, 0.06)$, $v_{2t} \sim \mathbb{N}(0, 0.02)$, $v_{3t} \sim \mathbb{N}(0, 0.08)$, $v_{4t} \sim \mathbb{U}(0, 0.1)$, and keep only the last 160 data points to reduce the dependence on the initial conditions and match a typical sample size.

We treat σ_ξ^2 as fixed and focus attention on the two economic parameters. We assume that the priors are $(1-\eta) \sim \mathrm{Beta}(4, 9)$ and $\beta \sim \mathrm{Beta}(99, 2)$. Beta distributions are convenient because they are easy to draw from. In fact, if $x \sim \chi^2(2a)$ and $y \sim \chi^2(2b)$, then $z = x/(x+y) \sim \mathrm{Beta}(a, b)$. Since the mean of a $\mathrm{Beta}(a, b)$ is $(a/a+b)$ and the variance is $ab/[(a+b)^2(a+b+1)]$, the prior mean of $1-\eta$ is about 0.31, and the prior mean of β about 0.99. The variances, approximately equal to 0.011 and 0.0002, imply sufficiently loose prior distributions.

We draw 10 000 replications. Given $1-\eta^0 = 0.55$, $\beta^0 = 0.97$, we produce candidates $\theta^\dagger = [(1-\eta)^\dagger, \beta^\dagger]$ by using a reflecting random walk process, i.e., $\theta^\dagger = \bar\theta + (\theta^{l-1} - \bar\theta) + v_\theta^l$, where θ^{l-1} is the previous draw, $\bar\theta$ is the mean of the process and v_θ^l is a vector of errors. The first component of v_θ (corresponding to $1-\eta$) is drawn from a $\mathbb{U}(-0.03, 0.03)$ and the second (corresponding to β) from a $\mathbb{U}(-0.01, 0.01)$ and $\bar\theta = [0.01, 0.001]'$. These ranges produce an acceptance rate of about 75%.

Since we are interested in $(1-\eta)$ and β, we are free to select which equations to use to estimate them. We arbitrarily choose those determining consumption and the real

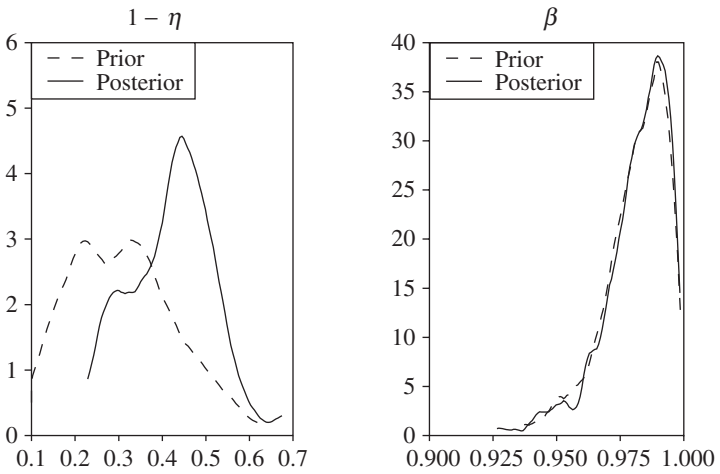

Figure 11.4. Priors and posteriors, basic RBC.

Table 11.2. Variances and covariances.

	True	Posterior 68% range
$\text{var}(c)$	40.16	$[3.65, 5.10 \times 10^{10}]$
$\text{var}(r)$	1.29×10^{-5}	$[2.55 \times 10^{-4}, 136.11]$
$\text{cov}(c, r)$	-0.0092	$[-0.15 \times 10^{-5}, -0.011]$

interest rate. We assume a normal likelihood and since $g(1-\eta, \beta) = g(1-\eta)g(\beta)$, we calculate the prior at the draw for each of the two parameters separately. Since the transition matrix $\mathfrak{P}(\theta^{\dagger}, \theta^{\mathbf{0}})$ is symmetric, the ratio of the kernels at $\theta^{\dagger}$ and θ^{l-1} is all that is needed to accept or reject the candidates.

We discard the first 5000 draws. Out of the last 5000 we keep 1 out of every 5 to reduce the serial correlation present in the draws. We check that the Metropolis algorithm has converged in two ways: splitting the sequences of draws in two and computing a normal test; calculating recursive means for the estimates of each parameter. In both cases, the sequence converged after about 2000 draws.

Figure 11.4 presents the marginal densities of $1-\eta$ and β, estimated with the 1000 saved draws from the prior and the posterior. Two features are worth mentioning. First, the data are more informative about $1-\eta$ than they are about β. Second, both posteriors are unimodal and roughly centered around the true parameter values.

Using the 1000 posterior draws we have calculated three statistics, the variances of consumption and of the real interest rate and the covariance between the two, and compared the posterior 68% credible range with the statistics computed by using the "true" parameters. Table 11.2 shows that the posterior 68% range includes the actual value of the consumption variance but not the one for the real rate or for the

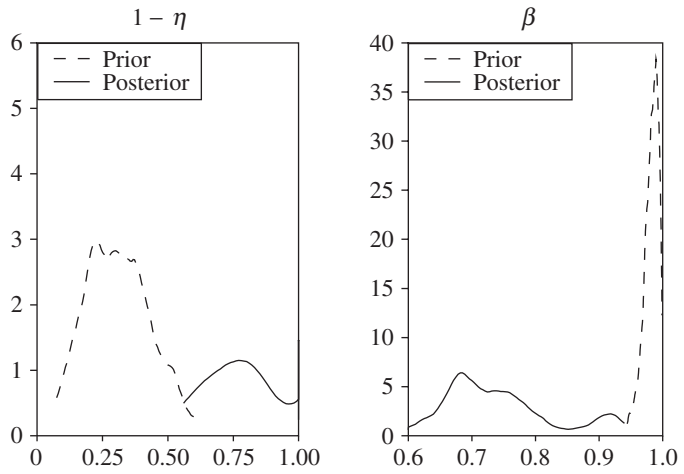

Figure 11.5. Priors and posteriors, RBC with habit persistence.

covariance. Also, there are posterior combinations of parameters which make the two variances very large.

Exercise 11.27. Using the same setup as example 11.11, modify the transition matrix $\mathfrak{P}(\theta^{\dagger}, \theta^{0})$ or the range for σ_v^2 in order to reduce the acceptance rate to about 50%. What would be the consequences of drawing candidates from normals rather than from uniform distributions?

Exercise 11.28. Vary the parameters of the prior for β and $1-\eta$ so as to make them more diffuse. Do the posteriors change? In what way?

Example 11.12. In this example we simulate data from an RBC model with habit in consumption, still assuming that capital depreciates in one period and that leisure does not enter the utility function. We assume $u(c_t, c_{t-1}) = \ln(c_t - \gamma c_{t-1})$, set $\gamma = 0.8$, and add to the solution the same measurement errors used in equations (11.27)–(11.30). We are interested in the shape of the posteriors of β and $1-\eta$ when we mistakenly assume that there is no habit (i.e., we condition on $\gamma = 0$). This experiment is interesting since it can give some indications of the consequences of using a dogmatic (and wrong) prior on some of the parameters of the model.

Perhaps unsurprisingly, the posterior distributions presented in figure 11.5 are very different from those in figure 11.4. What is somewhat unexpected is that the misspecification is so large that the posterior probability for the "true" parameters is roughly zero.

Exercise 11.29. Simulate data from an RBC model with production function $f(K_t, \mathrm{ku}_t, \zeta_t) = (K_t \mathrm{ku}_t)^{1-\eta}\zeta_t$, where ku_t is capital utilization and assume that the depreciation rate depends on the utilization of capital, i.e., $\delta(\mathrm{ku}_t) = \delta_0 + \delta_1 \mathrm{ku}_t^{\delta_2}$,

where $\delta_0 = 0.01$, $\delta_1 = 0.005$, $\delta_2 = 2$. Suppose you mistakenly neglect utilization and estimate a model like the one in equations (11.27)–(11.30). Evaluate the distortions induced by this misspecification.

Example 11.13. The next example considers a standard New Keynesian model with sticky prices and monopolistic competition. Our task here is twofold. First, we want to know how good this model is relative to, say, an unrestricted VAR in capturing the dynamics of the nominal interest rate, the output gap, and inflation. Second, we are interested in knowing the location of the posterior distribution of some important structural parameters. For example, we would like to know how much price stickiness is needed to match actual dynamics, whether policy inertia is an important ingredient to characterize the data, and whether the model has some internal propagation mechanism or if, instead, it relies entirely on the dynamics of the exogenous variables to match the dynamics of the data.

The model economy we use is a simplified version of the structure considered in chapter 2 and comprises a log-linearized (around the steady-state) Euler equation, a New Keynesian Phillips curve, and a Taylor rule. We assume that, in equilibrium, consumption is equal to output and use output in deviation from steady states in the Euler equation directly. Each equation has a shock attached to it: there is an i.i.d. policy shock, ϵ_{3t}, a cost push shock in the Phillips curve, ϵ_{2t}, and an arbitrary demand shock in the Euler equation, ϵ_{4t}. While the latter shock is unnecessary for the estimation, it is clearly needed to match the complexities of the output, inflation, and interest rate processes observed in the real world. The equations are

$$\text{gdpgap}_t = E_t\, \text{gdpgap}_{t+1} - \frac{1}{\varphi}(i_t - E_t\pi_{t+1}) + \epsilon_{4t}, \tag{11.31}$$

$$\pi_t = \beta E_t \pi_{t+1} + \kappa\, \text{gdpgap}_t + \epsilon_{2t}, \tag{11.32}$$

$$i_t = \phi_r i_{t-1} + (1-\phi_r)(\phi_\pi \pi_{t-1} + \phi_{\text{gap}}\, \text{gdpgap}_{t-1}) + \epsilon_{3t}, \tag{11.33}$$

where i_t is the nominal interest rate, π_t is the inflation rate, gdpgap_t is the output gap, $\kappa = (1-\zeta_{\text{p}})(1-\beta\zeta_{\text{p}})(\varphi+\vartheta_N)/\zeta_{\text{p}}$, ζ_{p} is the degree of stickiness in the Calvo setting, β is the discount factor, φ is the risk aversion parameter, ϑ_N is the inverse elasticity of labor supply, ϕ_r is the persistence of the nominal rate, while ϕ_π and ϕ_{gap} measure the responses of interest rates to lagged inflation and lagged output gap movements. We assume that ϵ_{4t} and ϵ_{2t} are AR(1) processes with persistence ρ_4, ρ_2 and variances σ_4^2, σ_2^2, while ϵ_{3t} is i.i.d. $(0, \sigma_3^2)$.

The model has 12 parameters, $\theta = (\beta, \varphi, \vartheta, \zeta_{\text{p}}, \phi_\pi, \phi_{\text{gap}}, \phi_r, \rho_2, \rho_4, \sigma_2^2, \sigma_3^2, \sigma_4^2)$, seven structural, and five auxiliary ones, whose posterior distributions need to be found. Our interest centers in the posterior distributions of $(\zeta_{\text{p}}, \phi_r, \rho_2, \rho_4)$. It is easy to check that ζ_{p} and ϑ_N are not separately identifiable so that inference about ζ_{p} will be meaningful only to the extent that the priors of these two parameters are carefully specified. We use U.S. quarterly detrended data from 1948:1 to 2002:1. We assume that $g(\theta) = \prod_{j=1}^{12} g(\theta_j)$ and use the following priors: $\beta \sim \text{Beta}(98, 3), \varphi \sim \mathbb{N}(1, (0.375)^2), \vartheta_N \sim \mathbb{N}(2, (0.75)^2), \zeta_{\text{p}} \sim \text{Beta}(9, 3), \phi_r \sim$

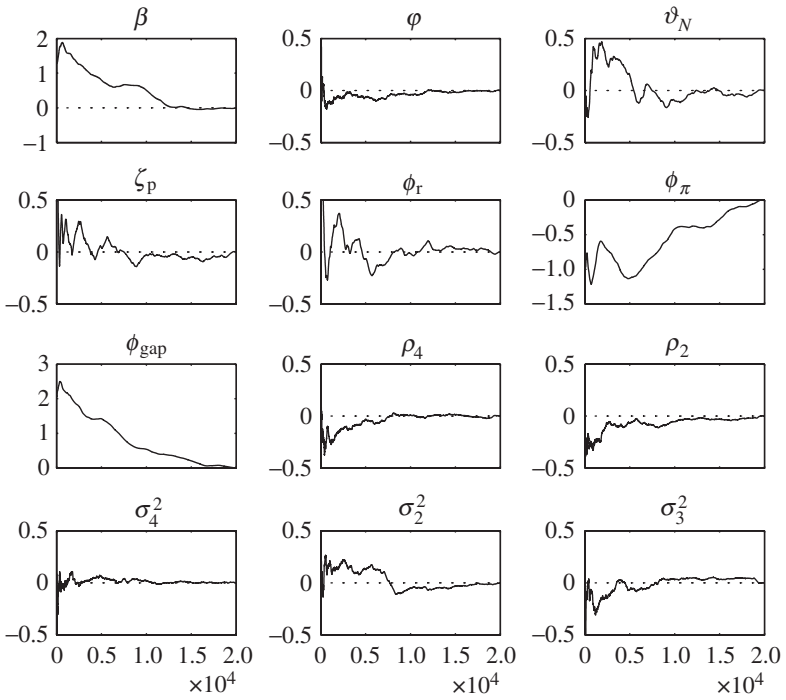

Figure 11.6. CUMSUM statistic.

Beta$(6,2)$, $\phi_\pi \sim \mathbb{N}(1.7,(0.1)^2)$, $\phi_{\mathrm{gap}} \sim \mathbb{N}(0.5,(0.05)^2)$, $\rho_4 \sim$ Beta$(17,3)$, $\rho_2 \sim$ Beta$(17,3)$, $\sigma_i^{-2} \sim \mathbb{G}(4,0.1)$, $i = 2,3,4$.

To generate a candidate vector $\theta^\dagger$, we use a random walk Metropolis algorithm with small uniform errors (the range is tuned up for each parameter so as to achieve a 40% acceptance rate) and check convergence by using a CUMSUM statistic: $(1/J)\sum_j (\theta_j^i - E(\theta_j^i))/\sqrt{\text{var}\,\theta_j^i}$, where $j = 1,2,\ldots,JL+\bar{L}$ and $i = 1,2,\ldots,12$. Figure 11.6, which presents this statistic, indicates that the chain has converged, roughly, after 15 000 draws. Convergence is hard to achieve for ϕ_π and ϕ_{gap}, while it is quickly reached (at times in less than 10 000 iterations) for the other parameters. As shown later the difficulties encountered with ϕ_π and ϕ_{gap} are not necessarily due to subsample instability. Instead, they appear to be related to the near nonidentifiability of these parameters from the data. Figure 11.7 presents prior and posterior distributions (estimated with kernel methods) using 1 out of every 5 of the last 5000 draws. The data appear to be informative in at least two senses. First, posterior distributions often have smaller dispersions than prior ones. Second, in some cases, the whole posterior distribution is shifted relative to the prior. Table 11.3, which presents some statistics of the prior and the posterior, confirms these visual impressions. Note also that, except for isolated cases, posterior distributions are roughly symmetric.

Table 11.3. Prior and posterior statistics.

	Prior		Posterior 1948–2002				
	Mean	Std	Median	Mean	Std	Min	Max
β	0.98	0.01	0.978	0.976	0.007	0.952	0.991
φ	0.99	0.37	0.836	0.841	0.118	0.475	1.214
ϑ_N	2.02	0.75	1.813	2.024	0.865	0.385	4.838
ζ_{p}	0.75	0.12	0.502	0.536	0.247	0.030	0.993
ϕ_r	0.77	0.14	0.704	0.666	0.181	0.123	0.992
ϕ_π	1.69	0.10	1.920	1.945	0.167	1.568	2.361
ϕ_{gap}	0.49	0.05	0.297	0.305	0.047	0.215	0.410
ρ_4	0.86	0.07	0.858	0.857	0.038	0.760	0.942
ρ_2	0.86	0.07	0.842	0.844	0.036	0.753	0.952
σ_4^2	0.017	0.01	0.017	0.017	0.007	0.001	0.035
σ_2^2	0.016	0.01	0.011	0.012	0.008	0.0002	0.036
σ_3^2	0.017	0.01	0.015	0.016	0.007	0.001	0.035

	Posterior 1948–1981		Posterior 1982–2002	
	Mean	Std	Mean	Std
β	0.986	0.008	0.983	0.008
φ	1.484	0.378	1.454	0.551
ϑ_N	2.587	0.849	2.372	0.704
ζ_{p}	0.566	0.200	0.657	0.234
ϕ_r	0.582	0.169	0.695	0.171
ϕ_π	2.134	0.221	1.925	0.336
ϕ_{gap}	0.972	0.119	0.758	0.068
ρ_4	0.835	0.036	0.833	0.036
ρ_2	0.831	0.036	0.832	0.036
σ_4^2	0.017	0.006	0.016	0.007
σ_2^2	0.016	0.006	0.016	0.007
σ_3^2	0.013	0.007	0.014	0.007

As far as the posterior of the four parameters of interest is concerned, note that the shocks are persistent (the posterior mean is 0.85) but there is no pileup of the posterior distribution for the AR parameters around 1. This means that, although the model does not have sufficient internal propagation to replicate the dynamics of the data, no exogenous unit-root-like processes are needed.

The posterior distribution of economic parameters is reasonably centered. The posterior mean of ζ_{p}, the parameter regulating the stickiness in prices, is only 0.5,

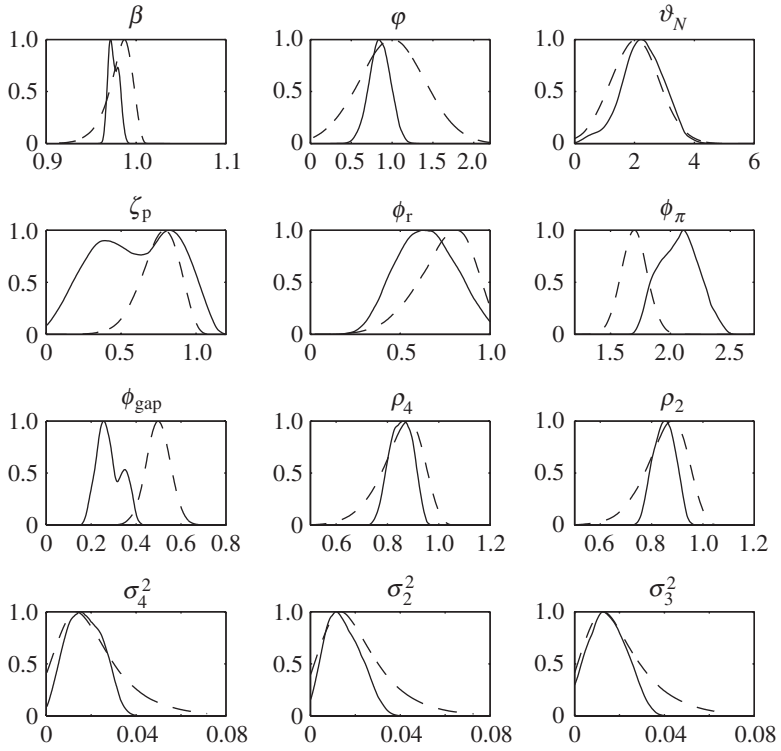

Figure 11.7. Priors (dashed) and posteriors (solid), sticky price model.

implying an average time of about two quarters between price changes — the prior was centered at an average of three quarters. However, since the posterior of ζ_{p} is bimodal, care must be exercised in using the posterior mean as a location measure. ϕ_r, the parameter measuring policy persistence, has a posterior mean of 0.7, implying some degree of policy smoothness, but not an excessive one.

Note that the posterior mean of κ is about 0.5, implying a moderate reaction of inflation to output gap movements. In comparison with the estimates obtained in chapter 5, the mean effect is slightly stronger, even though lower values have nonnegligible posterior probabilities.

The majority of these conclusions remain after splitting the sample in two. For example, ζ_{p} has a posterior mean of 0.566 in the 1948–81 sample and a posterior mean of 0.657 in the 1982–2002 sample. However, since the posterior standard error is around 0.22, differences in the two samples are statistically small. The other parameters also have stable posteriors. In particular, splitting the sample does not change the fact that the coefficients in the policy rule imply a strong reaction of interest rates to inflation.

The location and the shape of the posterior distributions are largely independent of the priors we have selected since priors are broadly noninformative. For example,

reweighing the posterior draws with a prior whose range is 90% of the range of the original prior in all 12 dimensions produces posterior distributions which are qualitatively very similar to those of figure 11.7.

Finally, we examine the forecasting performance of the model by comparing its marginal likelihood to that of a VAR(3) and that of a BVAR(3) with Minnesota prior and standard parameters (tightness equal to 0.1, linear lag decay and weight on other variables equal to 0.5), both with a constant. Bayes factors are small (of the order of 0.19) in both cases, indicating that the model can be improved upon in a forecasting sense. Note that, while both alternatives are more densely parametrized than our DSGE model (30 versus 12 parameters), Bayes factors take model size into account and no adjustment for the number of parameters is needed.

Exercise 11.30. Repeat the estimation of the model of example 11.13 by substituting (11.33) with the rule $i_t = \phi_r i_{t-1} + (1-\phi_r)(\phi_\pi \pi_t + \phi_{\text{gap}} \text{gdpgap}_t) + \epsilon_{3t}$. Compare the results. In particular, describe how the posterior distributions of ϕ_r, ρ_2, and ρ_4 are altered. Evaluate the probability that the data have been generated by a model with indeterminacies (i.e., evaluate what is the posterior probability that $\phi_\pi < 1$). (Hint: set the location of the prior for ϕ_π to 1.0.)

Exercise 11.31. Consider the model of example 11.13, but replace the Phillips curve by the following: $\pi_t = [\omega/(1+\omega\beta)]\pi_{t-1} + [\beta/(1+\omega\beta)] E_t \pi_{t+1} + [\kappa/(1+\omega\beta)] \times \text{gdpgap}_t + \epsilon_{2t}$, where ω is the degree of indexation of prices. Estimate this model and test whether indexation is necessary to match the data. (Hint: be careful about the identification of this parameter.)

Exercise 11.32. Add to the model of example 11.13 the following wage equation: $\Delta w_t = \beta E_t \Delta w_{t+1} + [(1-\zeta_w)(1-\zeta_w\beta)/\zeta_w(1+\varsigma_w \vartheta_N)][\text{mrs}_t - (w_t - p_t)] + \epsilon_{2t}$, where ζ_w is the probability of not changing the wage, ς_w is the elasticity of substitution between types of labor in production, and mrs_t is the marginal rate of substitution. Estimate this model and test whether wage stickiness adds to the fit of the basic sticky price model.

11.4.3 A Few Applied Tips

Although the models we have considered so far are of small scale, it has become standard in central banks and international institutions to estimate large-scale DSGE models with Bayesian methods. Care should be exercised when estimating large-scale models for several reasons.

First, large-scale models, while more articulate and potentially less misspecified, are more prone to identification problems. Furthermore, the variables used in estimation need not carry information about the parameters researchers care about. For example, it is quite common to try to get estimates of import and export price stickiness by using CPI inflation of different countries. Obviously, the informational content of CPI inflation for these parameters may be very small.

Second, as we have seen in chapter 6, the likelihood function of a small-scale DSGE model may have large flat sections or very rocky appearance. The likelihood function of a large-scale DSGE model typically contains both features and, at times, multiple peaks may be present. Calculation of posterior distributions in such a situation is difficult and the prior plays a crucial role in making inference possible. Hence, the choice of prior distributions should be carefully documented, the sensitivity of the results to variations in the spread presented, and the temptation to use reverse engineering (i.e., set a prior so that the posterior is well-behaved and confirms one's "gut" feeling) avoided. Note that multiple peaks in the likelihood may indicate the presence of breaks or multiple regimes and may give important information about features one is interested in examining. Once again, robustness analysis may inform the investigator on the likely presence of these problems.

Third, while it is common to start from a model with a large number of frictions and shocks, Bayesian methods can be used even with models which are misspecified in their dynamics or their probabilistic nature. This means that the type of sequential exercise performed in early calibration exercises (e.g., start from a competitive structure with only technology shocks, add government shocks, introduce noncompetitive markets, etc.) can also be fruitfully employed here. Frictions and shocks which add little to the ability of the model to reproduce interesting features of the data should be discarded. Such an analysis could also help to give some of the black-box shocks estimated in the factor literature an interesting economic content.

Finally, models are hardly built to explain the macroeconomic series that one finds in standard databanks. Therefore, data transformations, such as detrending or outlier elimination, and massaging techniques, such as the selection of appropriately stable sample periods or the elimination of structural breaks, are necessary before the model is taken to the data. When one is interested in the estimation of a model designed to capture only the cyclical properties of the data and dogmatically selects one trend specification, Bayesian and standard classical methods face the same arbitrariness problems and everything we said in chapter 3 applies without change. If more than one alternative trend specification is contemplated, one could put a prior on the various alternatives, compute the posterior probability of each specification, and use the techniques described in the next subsection to undertake inference.

11.4.4 Comparing the Quality of Models to the Data

While Bayesian estimation of structural parameters is simple, it is less straightforward to compare the model outcomes to the actual data and to assess the superiority of a model among alternative candidate specifications. Two methods are available. The first, preferred by macroeconomists, is based on informal analysis of some interesting economic statistics.

Example 11.14. Continuing with example 11.13, we present 68% impulse response bands to interest rate shocks in figure 11.8. While responses are economically reasonable there are three features of the figure which stand out. First, shocks which

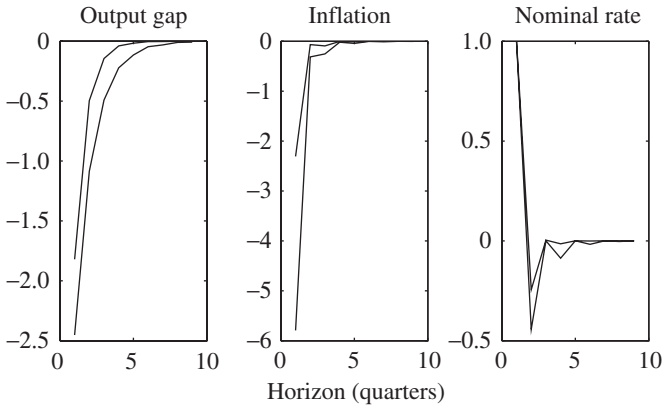

Figure 11.8. Responses to monetary shocks, 1948–2002.

increase interest rates make inflation and the output gap fall with very high probability. Second, responses die out after a few periods. Third, despite the assumed price stickiness, the largest inflation effect is instantaneous.

Figure 11.9 reports response bands obtained by estimating the model over different windows of data, keeping a constant number of observations in each sample. It is remarkable that the sign, the shape, and the magnitude of the posterior 68% credible bands are unchanged as we move from the late 1970s to the early 2000s. Hence, the transmission properties of monetary shocks have hardly changed over the last 30 years.

As an alternative to the presentation of economic statistics of various nested or nonnested models, one could compute measures of forecasting performance of various specifications. As we have seen in chapter 9, the marginal likelihood is the product of one-step-ahead forecast errors. Hence, selecting a model by using Bayes factors, as we did in example 11.13, is equivalent to choosing the specification with smallest one-step (in-sample) MSE. Clearly, out-of-sample forecasting races are also possible, in which case predictive Bayes factors can be computed (see, for example, DeJong et al. 2000). This is easy to do: we leave it to the reader to work out the details.

Exercise 11.33. Show how to construct the predictive density of future $y_{t+\tau}$, $\tau = 1, 2, \ldots$, given the model of example 11.13. (Hint: use the restricted VAR representation of the model.)

Despite their popularity, Bayes factors may not be very informative about the quality of the approximation of the model to the data, in particular, when the models one wishes to compare are grossly misspecified.

Example 11.15. Suppose there are three models, two structural ones ($\mathcal{M}_1$, $\mathcal{M}_2$) and a densely parametrized (e.g., a VAR) reference one ($\mathcal{M}_3$). The Bayes factor between the two structural models is $[f(y, \mathcal{M}_1)/f(y)] \times [f(y)/f(y, \mathcal{M}_2)]$, where

$f(y) = \int f(y, \mathcal{M}_i)\, \mathrm{d}\mathcal{M}_i$. If we use a 0–1 loss function, and assume that the prior probability of each model is 0.5, the posterior risk is minimized by selecting $\mathcal{M}_1$ if the Bayes factor exceeds 1. The presence of a third model does not affect the choice since it only enters in the calculation of $f(y)$, which cancels out of the Bayes factor. If the prior odds do not depend on this third model, the posterior odds ratio will also be independent of it. When $\mathcal{M}_1$ and $\mathcal{M}_2$ are misspecified, they will have low posterior probability relative to $\mathcal{M}_3$, but this has no influence on the inference one makes. Hence, comparing misspecified models with a Bayes factor may be uninteresting: one model may be preferable to another but it may have close to zero posterior probability.

Schorfheide (2000) provided a simple procedure to choose among misspecified models (in his case a cash-in-advance and a working-capital model). The actual data are assumed to be generated by a mixture of the competing structural models and a reference one, which has two characteristics: (i) it is more densely parametrized than the DSGE models; (ii) it can be used to compute a vector of population statistics $h(\theta)$. One such model could be a VAR or a BVAR. Given this setup, loss functions can be used to compare models. In particular, when several alternatives are available, the following algorithm could be used.

Algorithm 11.5.

(1) Compute the posterior distribution for the parameters of each model by using tractable priors and one of the available posterior simulators.

(2) Obtain the marginal likelihood, for each $\mathcal{M}_i$, that is, compute $f(y \mid \mathcal{M}_i) = \int f(y \mid \theta_i, \mathcal{M}_i) g(\theta_i \mid \mathcal{M}_i)\, \mathrm{d}\theta_i$.

(3) Compute posterior probabilities $\tilde{P}_i = \bar{P}_i f(y \mid \mathcal{M}_i) / \sum_i \bar{P}_i f(y \mid \mathcal{M}_i)$, where $\bar{P}_i$ is the prior probability of model i. Note that, if the distribution of y is degenerated under $\mathcal{M}_i$ (e.g., if the number of shocks is smaller than the number of endogenous variables), $\tilde{P}_i = 0$.

(4) Calculate the posterior distribution of any continuous function $h(\theta)$ of the parameters for each model and average by using posterior probabilities, i.e., obtain $g(h(\theta) \mid y, \mathcal{M}_i)$ and $g(h(\theta) \mid y) = \sum_i \tilde{P}_i g(h(\theta) \mid y, \mathcal{M}_i)$. Note that $g(h(\theta) \mid y) = g(h(\theta) \mid y, \mathcal{M}_{i'})$ if all but model i' produce degenerate distributions.

(5) Set up a loss function $\mathfrak{L}(h_T, h_i(\theta))$ measuring the discrepancy between model i's predictions of $h(\theta)$ and data h_T. Since the optimal predictor in model $\mathcal{M}_i$ is $\hat{h}_i(\theta) = \operatorname{argmin}_{h_i(\theta)} \int \mathfrak{L}(h_T, h_i(\theta)) g(h_i(\theta) \mid y, \mathcal{M}_i)\, \mathrm{d}h_T$, one can compare models by using the risk of $\hat{h}_i(\theta)$ under the overall posterior distribution $g(h(\theta) \mid y)$, i.e., $\min \mathfrak{R}(\hat{h}_i(\theta) \mid y) = \min \int \mathfrak{L}(h_T, \hat{h}_i(\theta)) g(h(\theta) \mid y)\, \mathrm{d}h_T$.

Since $\mathfrak{R}(\hat{h}_i(\theta) \mid y)$ measures how well model $\mathcal{M}_i$ predicts h_T, a model is preferable to another if it has a lower risk. Note also that, while model comparison is relative, $g(h(\theta) \mid y)$ takes into account information from all models. Taking

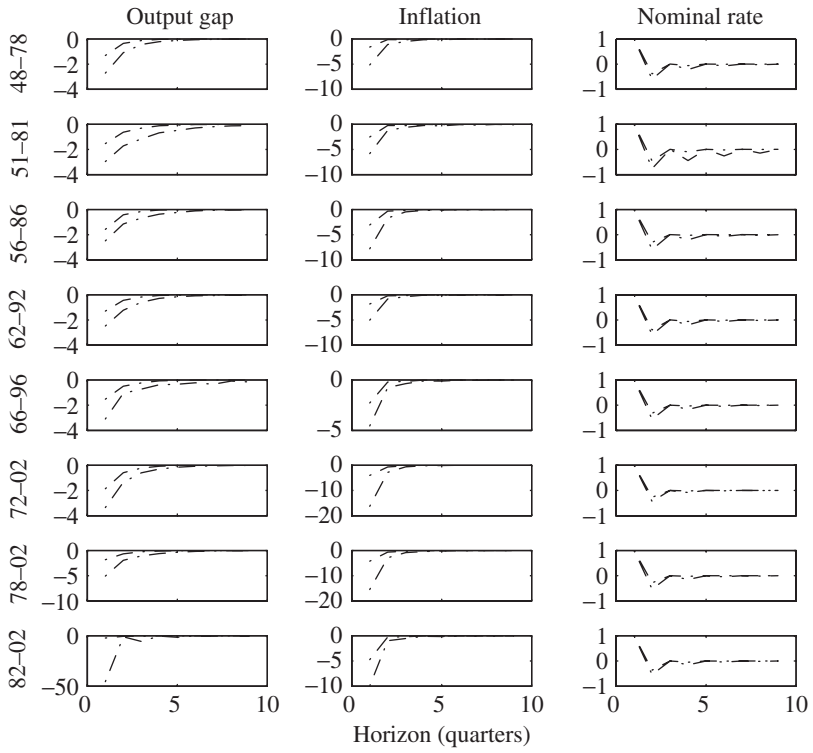

Figure 11.9. Responses to a monetary shock, various samples.

step (5) further, one should note that, for each i, θ can be selected so as to minimize $\mathfrak{R}(\hat{h}_i(\theta) \mid y)$. Such an estimate provides a lower bound to the posterior risk obtained by the "best" candidate model in the dimensions represented by h_T.

To make algorithm 11.5 operative a loss function must be selected. We have presented a few options in chapter 9. For DSGE models, the most useful are as follows.

(a) Quadratic loss: $\mathfrak{L}_2(h_T, h(\theta)) = [h_T - h(\theta)]' W [h_T - h(\theta)]$, where W is an arbitrary positive definite weighting matrix.

(b) Penalized loss: $\mathfrak{L}_p(h_T, h(\theta)) = \mathfrak{I}_{[g(h(\theta)|y) < g(h_T|y)]}$, where $\mathfrak{I}_{[x_1 < x_2]} = 1$ if $x_1 < x_2$.

(c) χ^2 loss: $\mathfrak{L}_{\chi^2}(h(\theta), h_T) = \mathfrak{I}_{[\mathfrak{Q}_{\chi^2}(h(\theta)|y) > \mathfrak{Q}_{\chi^2}(h_T|y)]}$, where $\mathfrak{Q}_{\chi^2}(h(\theta) \mid y) = [h(\theta) - E(h(\theta) \mid y)]' \Sigma^{-1}_{h(\theta)} [h(\theta) - E(h(\theta) \mid y)]$, $\Sigma_{h(\theta)}$ is the covariance of $h(\theta)$, and $\mathfrak{I}_{[x_1 > x_2]} = 1$ if $x_1 > x_2$.

(d) 0–1 loss: $\mathfrak{L}(h_T, h(\theta), \epsilon) = 1 - \mathfrak{I}_{\epsilon(h(\theta))}(h_T)$, where $\epsilon(h(\theta))$ is an ϵ-neighborhood of $h(\theta)$.

Three features of these loss functions should be mentioned. First, with penalized and χ^2 loss functions, two DSGE models are compared on the basis of the height of

the posterior distribution at $h_i(\theta)$. Second, with a quadratic loss function, comparison is based on the weighted distance between $h_i(\theta)$ and the posterior mean. Third, as already mentioned, a 0–1 loss implies that $\mathcal{M}_1$ is preferred if the posterior odds exceed 1.

Exercise 11.34. (i) Show that $\mathfrak{R}_2 = [h_T - E(h(\theta)) \mid y]' W [h_T - E(h(\theta)) \mid y] + \varrho_0$, where ϱ_0 does not depend on $Eh(\theta)$. How would you choose W optimally?

(ii) Show that, if $g(\theta \mid y)$ is normal, $\mathfrak{L}_2 = \mathfrak{L}_{\chi^2}$ and the optimal predictor is $E(h(\theta) \mid y, \mathcal{M}_i)$.

(iii) Show that the optimal predictor for the $\mathfrak{L}_p$ loss is the mode of $g(h(\theta) \mid y, \mathcal{M}_i)$.

Two interesting special cases obtain when the $\mathfrak{L}_2$ loss is used.

Exercise 11.35 (Schorfheide). Suppose there are three models. Suppose that $\tilde{P}_1 \xrightarrow{\text{P}} 1$, $E(h_i(\theta) \mid y_T, \mathcal{M}_i) \xrightarrow{\text{P}} \bar{h}_i(\theta)$, and $\bar{h}_1(\theta) - \bar{h}_2(\theta) = \delta_\theta$, where $|\delta_\theta| > 0$. Show that, as $T \to \infty$, $\mathfrak{R}(\hat{h}_1(\theta)) \xrightarrow{\text{P}} 0$ and $\mathfrak{R}(\hat{h}_2(\theta)) \xrightarrow{\text{P}} \delta_\theta' W \delta_\theta$. Suppose now that, as $T \to \infty$, $\tilde{P}_{3,T} \to 1$ and $E(h_i(\theta) \mid y, \mathcal{M}_i) \xrightarrow{\text{P}} \bar{h}_i(\theta)$. Show that $E(h(\theta) \mid y) - E(h(\theta) \mid y, \mathcal{M}_3) \xrightarrow{\text{P}} 0$.

Exercise 11.35 reaches a couple of interesting conclusions. First, if for any positive definite W model $\mathcal{M}_1$ is better than $\mathcal{M}_2$ with probability 1, model selection using $\mathfrak{L}_2$ is consistent and gives the same result as a posterior odds ratio in large samples. To restate this concept differently, under these conditions, $\mathfrak{L}_2$-model comparison is based on the relative one-step-ahead predictive ability. Second, if the two models are so misspecified that their posterior probability goes to zero as $T \to \infty$, the ranking of these models only depends on the discrepancy between $E(h(\theta) \mid y, \mathcal{M}_3) \approx E(h(\theta) \mid y)$ and $\hat{h}_i(\theta)$, $i = 1, 2$. If $\mathcal{M}_3$ is any empirical model, then using an $\mathfrak{L}_2$ loss is equivalent to comparing sample and population moments obtained from different models. This means that, when one makes decisions based on the $\mathfrak{L}_2$ loss function and the models are highly misspecified, an informal comparison between the predictions of the model and the data, as is done in the simplest calibration exercises, is optimal from a Bayesian point of view. Intuitively, this surprising outcome obtains because the posterior variance of $h(\theta)$ does not affect the ranking of models — this conclusion does not hold with the $\mathfrak{L}_p$ or the $\mathfrak{L}_{\chi^2}$ loss functions.

Example 11.16. Continuing with example 11.13, we calculate the risk associated with the model when $h(\theta)$ represents the persistence of inflation and persistence is measured by the height of the spectrum at zero frequency. This number is large (227.09), reflecting the inability of the model to generate persistence in inflation. In comparison, for example, the risk generated by a univariate AR(1) is 38.09.

11.4.5 DSGEs and VARs, Once Again

As mentioned in chapter 10, it is possible to use a DSGE model to construct a prior for reduced-form VAR coefficients. Such an approach is advantageous since it jointly

allows posterior estimation of both reduced-form and structural parameters. We have already derived the posterior for VAR parameters in section 10.2.5. Here we describe how to obtain posterior distributions for the structural ones. Let $f(y \mid \alpha, \Sigma_e)$ be the likelihood function of the data, conditional on the VAR parameters, let $g(\alpha, \Sigma_e \mid \theta)$ be the prior for the VAR parameters, conditional on the DSGE model parameters, and $g(\theta)$ the prior distribution for the DSGE parameters. Here $g(\alpha, \Sigma_e \mid \theta)$ is the prior for the reduced-form parameters induced by the prior on the structural parameters and the details of the model. The joint posterior of VAR and structural parameters is $g(\alpha, \Sigma_e, \theta \mid y) = g(\alpha, \Sigma_e, \mid \theta, y)g(\theta \mid y)$.

We have seen that $g(\alpha, \Sigma_e, \mid \theta, y)$ has a normal-inverted Wishart form so that it can be easily computed analytically or by simulation. The computation of $g(\theta \mid y)$ is more complicated since its form is unknown. The kernel of this distribution is $\breve{g}(\theta \mid y) = f(y \mid \theta)g(\theta)$, where

$$\begin{aligned} f(y \mid \theta) &= \int f(y \mid \alpha, \Sigma_e) g(\alpha, \Sigma_e, \theta)\, \mathrm{d}\alpha\, \mathrm{d}\Sigma_e \\ &= \frac{f(y \mid \alpha, \Sigma_e) g(\alpha, \Sigma_e \mid \theta)}{g(\alpha, \Sigma_e \mid y, \theta)}. \end{aligned} \tag{11.34}$$

Since the posteriors of (α, Σ_e) depend on θ only through y, $g(\alpha, \Sigma_e \mid y, \theta) = g(\alpha, \Sigma_e \mid y)$ and we can use the fact that both the numerator and the denominator of (11.34) have normal-inverted Wishart format to obtain

$$\begin{aligned} f(y \mid \theta) = {} & \frac{|(X^{\mathrm{s}})'(\theta)X^{\mathrm{s}}(\theta) + X'X|^{-0.5m} |(T_{\mathrm{s}} + T)\tilde{\Sigma}_e(\theta)|^{-0.5(T_{\mathrm{s}}+T-k)}}{|(X^{\mathrm{s}})'(\theta)X^{\mathrm{s}}(\theta)|^{-0.5m} |T_{\mathrm{s}}\bar{\Sigma}_e^{s}(\theta)|^{-0.5(T_{\mathrm{s}}-k)}} \\ & \times \frac{(2\pi)^{-0.5mT} 2^{0.5m(T_{\mathrm{s}}+T-k)} \prod_{i=1}^{m} \Gamma((T_{\mathrm{s}} + T - k + 1 - i)/2)}{2^{0.5m(T_1-k)} \prod_{i=1}^{m} \Gamma((T_{\mathrm{s}} - k + 1 - i)/2)}, \end{aligned} \tag{11.35}$$

where $\tilde{\Sigma}_e(\theta) = (1/(1+\kappa)T)\{(y^{\mathrm{s}})'y^{\mathrm{s}} + y'y - [(y^{\mathrm{s}})'X^{\mathrm{s}} + y'X][(X^{\mathrm{s}})'X^{\mathrm{s}} + X'X]^{-1} \times [(X^{\mathrm{s}})'y^{\mathrm{s}} + X'y]\}$ and $\bar{\Sigma}_e^{\mathrm{s}} = (1/T_{\mathrm{s}})\{(y^{\mathrm{s}})'y^{\mathrm{s}} - (y^{\mathrm{s}})'x^{\mathrm{s}}[(x^{\mathrm{s}})'x^{\mathrm{s}}]^{-1}(x^{\mathrm{s}})'y^{\mathrm{s}}\}$, T_{s} is the number of observations from the DSGE model added to the actual data, Γ is the gamma function, $X = (I \otimes \boldsymbol{X})$ includes all the lags of y, the superscript "s" indicates simulated data, and k is the number of coefficients in each VAR equation.

Exercise 11.36. Suggest an algorithm to draw sequences from $g(\theta \mid y)$.

11.4.6 Nonlinear Specifications

So far we have focused attention on DSGE models that are (log-)linearized around some pivotal point. As seen in chapter 2, there are applications for which (log-) linearizations are unappealing; for example, when economic experiments involve changes of regime or large perturbations of the relationships. In these cases one may want to work directly with the nonlinear version of the model and some steps of the algorithms of this chapter need to be modified to take this into account. Consider

the model

$$y_{2t+1} = h_1(y_{2t}, \epsilon_{1t}, \theta), \tag{11.36}$$

$$y_{1t} = h_2(y_{2t}, \epsilon_{2t}, \theta), \tag{11.37}$$

where ϵ_{2t} are measurement errors, ϵ_{1t} are structural shocks, θ is a vector of structural parameters, y_{2t} is the vector of states, and y_{1t} is the vector of controls. Let $y_t = (y_{1t}, y_{2t})$, $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})$, $y^{t-1} = (y_0, \ldots, y_{t-1})$, and $\epsilon^t = (\epsilon_1, \ldots, \epsilon_t)$. Integrating the initial conditions and the shocks out, the likelihood of the model can be written as (see Fernandez-Villaverde and Rubio-Ramirez 2003a,b)

$$\mathcal{L}(y^T, \theta) = \int \left[\prod_{t=1}^{T} \int f(y_t \mid \epsilon^t, y^{t-1}, y_{20}, \theta) f(\epsilon^t \mid y^{t-1}, y_{20}, \theta) \, \mathrm{d}\epsilon^t \right] f(y_{20}, \theta) \, \mathrm{d}y_{20}, \tag{11.38}$$

where y_{20} is the initial state. Clearly, (11.38) is intractable. However, if we have L draws for y_{20} from $f(y_{20}, \theta)$ and L draws for $\epsilon^{t|t-1}$ from $f(\epsilon^t \mid y^{t-1}, y_{20}, \theta)$, $t = 1, \ldots, T$, we can approximate (11.38) with

$$\mathcal{L}(y^T, \theta) = \frac{1}{L} \left[\prod_{t=1}^{T} \frac{1}{L} \sum_{l} f(y_t \mid \epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta) \right]. \tag{11.39}$$

Drawing from $f(y_{20}, \theta)$ is simple, but drawing from $f(\epsilon^t \mid y^{t-1}, y_{20}, \theta)$ is, in general, complicated. Fernandez-Villaverde and Rubio-Ramirez suggest using $f(\epsilon^{t-1} \mid y^{t-1}, y_{20}, \theta)$ as importance sampling for $f(\epsilon^t \mid y^{t-1}, y_{20}, \theta)$. We summarize their approach in the next algorithm.

Algorithm 11.6.

(1) Draw y_{20}^l from $f(y_{20}, \theta)$. Draw $\epsilon^{t|t-1,l}$ L times from $f(\epsilon^t \mid y^{t-1}, y_{20}^l, \theta) = f(\epsilon^{t-1} \mid y^{t-1}, y_{20}^l, \theta) f(\epsilon_t \mid \theta)$.

(2) Set $\mathrm{IR}_t^l = f(y_t \mid \epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta) / \sum_{l=1}^{L} f(y_t \mid \epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)$ and assign it as a weight to each draw $\epsilon^{t|t-1,l}$.

(3) Resample from $\{\epsilon^{t|t-1,l}\}_{l=1}^{L}$ with probabilities equal to IR_t^l. Call this draw $\epsilon^{t,l}$.

(4) Repeat steps (1)–(3) for every $t = 1, 2, \ldots, T$.

Step (3) is crucial to making the algorithm work. If omitted, only one particle will asymptotically remain and the integral in (11.38) will diverge as $T \to \infty$. The resampling step prevents this from happening. Note that such a step is similar to the one employed in genetic algorithms: you resample from candidates which have high probability and create new branches at each step.

Clearly, algorithm 11.6 is computationally demanding: in fact, at each iteration, the model needs to be solved to find an expression for $f(y^t \mid \epsilon^t, y^{t-1}, y_{20}, \theta)$. At

this point only the most basic RBC model has been estimated by nonlinear likelihood methods and some gains have been reported by Fernandez-Villaverde and Rubio-Ramirez (2004). When Bayesian analysis is performed, algorithm 11.6 must be inserted between steps (3) and (4) of algorithm 11.3. This makes such an approach very demanding on currently available computers.

11.4.7 Which Approach to Use?

There is surprisingly little work comparing estimation/evaluation approaches in models which are misspecified, tightly parametrized, and feature fewer driving forces than endogenous variables. Ruge-Murcia (2002) is one recent example. Despite the lack of formal evidence, there are a few general ideas which may be useful to the applied investigator.

First, there are economic and statistical advantages in jointly estimating a system of structural equations. From an economic point of view, this is appealing since parameter estimates are obtained by employing all the model's restrictions. On the other hand, statistical efficiency is enhanced when all available information is used. Joint estimation may be problematic when a researcher is not necessarily willing to subscribe to all the details of a model. After all, tight parameter estimates which are economically unreasonable are hard to justify and interpret.

Misspecification, a theme we have repeatedly touched upon in several chapters of this book, creates problems for full-information estimation techniques in at least two ways. When the number of shocks is smaller than the number of endogenous variables, parameter estimates can be obtained only from a restricted number of series — essentially transforming full-information methods into limited-information ones. Furthermore, since not all variables have the same informational content about the parameters of interest, one is forced to experiment, with little guidance from economic or statistical theory. Second, if the model cannot be considered the DGP of the data (because of the assumptions made or because of the purely qualitative nature of the behavioral relationships it describes), both full-information estimation and testing are problematic. Maximum likelihood, in fact, attempts to minimize the largest discrepancy between the model's equations and the data. That is to say, it will choose parameter estimates that are best in the dimensions where misspecification is the largest. Therefore, it is likely to produce estimates which are either unreasonable from an economic point of view or on the boundary of the parameter space.

There are a few solutions to these problems. Adding measurement errors may eliminate the singularity of the system but it cannot remedy dynamic misspecification problems. Adding serially correlated measurement errors, on the other hand, may solve both problems, but such an approach lacks economic foundations. Roughly speaking, it amounts to giving up the idea that the model is a good representation of the data, both in an economic and in a statistical sense. The methods we have described in the last three chapters can elegantly deal with these problems. The prior plays the role of a penalty function and if appropriately specified, it may make a full-information approach look for a local, but economically interesting, maximum

of the problem. In addition, it may reduce both biases and skewness in ML estimates. However, it is still to be proved that computer-intensive MCMC methods have good size and power properties in the types of model we have studied in this book. The simple examples we have presented suggest that a lot more work needs to be done.

The alternative is to use less information and therefore be theoretically less demanding about the quality of the approximation of the model to the data. Still, the singularity of the system imposes restrictions on the vector of moments (functions) used to estimate the structural parameters — the functions must be linearly independent, otherwise the asymptotic covariance matrix of the estimates will not be well-defined. Nevertheless, there are situations when the model is extremely singular (for example, there is one source of shocks and ten endogenous variables) and limited-information procedures like GMM, SMM, or indirect inference may paradoxically use more information than ML. We have also mentioned that limited-information approaches may fall into logical inconsistencies whenever they claim to approximate only parts of the DGP. To avoid these inconsistencies, what an investigator wants to explain and what she does not should naturally have a block recursive structure, which is hardly a feature of currently available DSGE models.

Despite the remarkable progress in the specification of DSGE models, one may still prefer to take the point of view that models are still too stylized to credibly represent the data and choose an estimation approach where only the qualitative implications (as opposed to the quantitative ones) are entertained. Such an approach sidesteps both the singularity and the misspecification issues, since qualitative implications can be embedded, as seen in chapter 4, as identification devices for structural VAR models. Combining DSGE and VAR models either informally or more formally, as in Del Negro and Schorfheide (2004), seems to be the most promising way to compare stylized models and the data.

In terms of computations, a VAR-based approach has clear advantages. Bayesian and ML estimation are time-consuming especially when the objective function is not well-behaved (a typical case with DSGE models), while SMM and indirect inference may require substantial computer capabilities and may be subject to important identification problems. GMM is a close competitor, but its severe small-sample problems may well wipe out the gains from simplicity. This makes GMM (and simulation estimators) unsuitable for macroeconomic problems where samples are typically short and breaks or regime changes make the time series of data heterogeneous.

It is also important to stress that different small-sample distributions for the structural parameters do not necessarily translate into statistically and economically large differences in the interesting functions a researcher wants to compute. For example, Ruge-Murcia (2002) documents that ML, GMM, SMM, and indirect inference have somewhat different small-sample biases and markedly different efficiency properties. Yet, small-sample impulse response bands computed with estimates obtained with the four approaches are similar in size and shape.