# Chapter 10: Bayesian VARs

We have seen in chapter 4 that VAR models can be used to characterize any vector of time series under a minimal set of conditions. We have also seen that since VARs are reduced form models, identification restrictions, motivated by economic theory, are needed to conduct meaningful policy analysis. Reduced form VARs are also typically unsuitable for forecasting out-of-sample. To reasonably approximate the Wold representation it is in fact necessary to have a VAR with long lags. A generous parametrization means that unrestricted VARs are not operational alternatives to either standard macroeconometric models, where insignificant coefficients are purged out of the specification, or to parsimonious time series models since, with a limited number of degrees of freedom, estimates of VAR coefficients are imprecise and forecasts have large standard errors.

It is useful to think of the construction of an empirical model as the process of combining historical and a-priori information, both of statistical and of economic nature. Alternative modeling techniques provide different a-priori information or different relative weights to sample and prior information. Unrestricted VARs employ a-priori information very sparsely - in choosing the variables of the VAR; in selecting the lag length of the model; in imposing identification restrictions. Because of this choice, overfitting may obtain when the data set is short, sample information is weak or the number of parameters is large. In-sample overfitting typically translates into poor forecasting performance, both in unconditional and conditional sense. Bayesian methods can solve these problems: they can make in-sample fitting less dramatic and improve out-of-sample performance. While Bayesian VAR (BVAR) were originally devised to improve macroeconomic forecasts, they have evolved dramatically and they are used now for a variety of purposes.

This chapter describes Bayesian methods for a variety of VAR models. First, we present the decomposition of the likelihood function of a VAR and the construction of the posterior distribution for a number of prior specifications. We also show the link between posterior mean estimates and classical estimates obtained when the coefficients of the VAR model are subject to stochastic linear constraints. The third section describes Bayesian structural VARs and block recursive structures which arise e.g., in models with some exogenous variables or in two country models with (overidentifying) linear restrictions on the contemporaneous impact of the shocks. The fourth section, discusses time varying BVAR models. These models have a state space representation and this helps in constructing both empirical Bayes and fully hierarchical posterior estimates of the VAR coefficients and of

the covariance matrix. We show that these structures generate a variety of distributional patterns and can be used to model series with thick tails, with smoothly evolving pattern, or displaying coefficients switching over a finite number of states.

The fifth section deals with multiple BVAR models: these structures are becoming popular in empirical practice, for example, when comparing the effects of monetary policy shocks in different countries or the growth behavior in different regions, and present interesting complications relative to single unit BVAR models. We show how to obtain posterior estimates of the coefficients of the model for each unit and how to obtain estimates of the mean effect across units, which often is the center of interest for applied investigators. We also describe a procedure to endogenously group units with similar characteristics. This is useful when one wants to distinguish the impact of certain shocks on e.g. small or large firms, or when policy advice requires some particular endogenous classifications (e.g. income per-capita, education level, indebtness, etc.). The last part of the section studies Bayesian Panel VAR models with cross unit interdependencies. These models are suited to study e.g., the transmission of shocks across countries or the effects of increased interdependencies in various world economies. Because of the large number of parameters, it is impossible to estimate them with classical methods and suitable (prior) restrictions need to be imposed for estimation. With such a respecification, these models are easily estimable with Monte Carlo Markov Chain methods.

Since the chapter deals with models of increasing complexity, increasingly complex methods will be used to compute posteriors. The techniques described in chapter 9 are handy here: conjugate priors allow the derivation of analytic forms for the conditional posteriors; Markov Chain Monte Carlo methods are used to draw sequences from the posterior distributions.

## 10.1 The Likelihood function of an m variable VAR(q)

Throughout this chapter we assume that the VAR has the form $y_t = A(L)y_{t-1} + C\bar{y}_t + e_t,\ e_t \sim (0, \Sigma_e)$, where $y_t$ includes $m$ variables, each of which has q lags, while the constant and other deterministic variables (trends, seasonal dummies) are collected into the $m_c \times 1$ vector $\bar{y}_t$. Hence, the number of regressors in each equation is $k = mq + m_c$ and there are $mk$ coefficients in the VAR.

Following the steps described in chapter 4, we can rewrite the VAR in two alternative formats, both of which will be used in this chapter:

$$\mathsf{Y} = \mathsf{X}\mathsf{A} + \mathsf{E} \tag{10.1}$$

$$y = (I_m \otimes \mathsf{X})\alpha + e \quad e \sim (0, \Sigma_e \otimes I_T) \tag{10.2}$$

where $\mathsf{Y}$ and $\mathsf{E}$ are $T \times m$ matrices and $\mathsf{X}$ is a $T \times k$ matrix, $\mathsf{X}_t = [y'_{t-1}, \ldots, y'_{t-q}, \bar{y}'_t]$; $y$ and $e$ are $mT \times 1$ vectors, $I_m$ is the identify matrix of dimension $m$, and $\alpha = vec(\mathsf{A})$ is a $mk \times 1$ vector. Using (10.2) the likelihood function is

$$\mathcal{L}(\alpha, \Sigma_e) \propto |\Sigma_e \otimes I_T|^{-0.5} \exp\{-0.5(y - (I_m \otimes \mathsf{X})\alpha)'(\Sigma_e^{-1} \otimes I_T)(y - (I_m \otimes \mathsf{X})\alpha)\} \tag{10.3}$$

To derive a useful decomposition of (10.3) note that

$$\begin{aligned}(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-1}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ (\Sigma_e^{-0.5}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-0.5}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ [(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha)]'[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha)]\end{aligned}$$

Also $(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha = (\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols}+(\Sigma_e^{-0.5}\otimes \mathsf{X})(\alpha_{ols}-\alpha)$ where $\alpha_{ols}=(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})^{-1}(\Sigma_e^{-1}\otimes \mathsf{X})'y$. Therefore:

$$\begin{aligned}(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-1}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ ((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols}) &+ \quad (10.4)\\ (\alpha_{ols}-\alpha)'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha_{ols}-\alpha) & \quad (10.5)\end{aligned}$$

The term in (10.4) is independent of $\alpha$ and looks like a sum of squared errors. The one in (10.5) looks like the scaled square error of $\alpha_{ols}$. Putting the pieces back together we have:

$$\begin{aligned}\mathcal{L}(\alpha,\Sigma_e) &\propto |\Sigma_e\otimes I_T|^{-0.5}\exp\{-0.5(\alpha-\alpha_{ols})'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha-\alpha_{ols}) \\ &- 0.5[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})]\} \\ &= |\Sigma_e|^{-0.5k}\exp\{-0.5(\alpha-\alpha_{ols})'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha-\alpha_{ols})\} \\ &\times |\Sigma_e|^{-0.5(T-k)}\exp\{-0.5tr[(\Sigma_e^{-0.5}\otimes I_T)y \\ &- (\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})]\} \\ &\propto \mathbb{N}(\alpha|\alpha_{ols},\Sigma_e,\mathsf{X},y)\times\mathbb{W}(\Sigma_e^{-1}|y,\mathsf{X},\alpha_{ols},T-k-m-1) \quad (10.6)\end{aligned}$$

where $tr$ is the trace of a matrix. The likelihood function of a VAR(q) can therefore be decomposed into the product of a Normal density for $\alpha$, conditional on the OLS estimate $\alpha_{ols}$ and on $\Sigma_e$, and a Wishart density for $\Sigma_e^{-1}$, conditional on $\alpha_{ols}$, with scale matrix $[(y-(I_m\otimes \mathsf{X})\alpha_{ols})'(y-(I_m\otimes \mathsf{X})\alpha_{ols})]^{-1}$, and $(T-k-m-1)$ degrees of freedom (see the Appendix for the form of various distributions).

Hence, under appropriate conjugate prior restrictions, we can analytically derive the conditional posterior distribution for the VAR coefficients and the covariance matrix of the reduced form shocks. As we have seen in chapter 9, a Normal-Wishart prior conjugates the two blocks of the likelihood. Therefore, under these assumptions, the conditional posterior for $\alpha$ will be Normal and the conditional posterior of $\Sigma_e^{-1}$ will be Wishart. Other prior assumptions on $\alpha$ and $\Sigma_e$ also allow analytical computation of conditional posteriors. We examine them in the next section.

## 10.2 Priors for VARs

In this section we consider four alternative types of prior specification:

1. A Normal prior for $\alpha$ with $\Sigma_e$ fixed.

2. A non-informative prior for both $\alpha$ and $\Sigma_e$.

3. A Normal prior for $\alpha$, and a non-informative prior for $\Sigma_e$.

4. A Conditionally conjugate prior, i.e. a Normal for $\alpha$, and a Wishart for $\Sigma_e^{-1}$.

We examine in details the derivation of the posterior distribution for the VAR coefficients for case 1. Let the prior be $\alpha = \bar{\alpha} + v_a, \quad v_a \sim N(0, \bar{\Sigma}_a)$, with $\bar{\Sigma}_a$ fixed. Then

$$\begin{aligned} g(\alpha) &\propto |\bar{\Sigma}_a|^{-0.5} exp[-0.5(\alpha-\bar{\alpha})'\bar{\Sigma}_a^{-1}(\alpha-\bar{\alpha})] \\ &\propto |\bar{\Sigma}_a|^{-0.5} exp[-0.5(\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha}))'\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha})] \end{aligned} \tag{10.7}$$

Let $\mathcal{Y} = [\bar{\Sigma}_a^{-0.5}\bar{\alpha}, (\Sigma_e^{-0.5} \otimes I_T)y]'$; $\mathcal{X} = [\bar{\Sigma}_a^{-0.5}, (\Sigma_e^{-0.5} \otimes \mathsf{X})]'$. Then:

$$\begin{aligned} g(\alpha|y) &\propto |\bar{\Sigma}_a|^{-0.5} \exp\{-0.5(\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha}))'\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha})\} \times |\Sigma_e \otimes I_T|^{-0.5} \\ &\times \exp\{(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes \mathsf{X})\alpha)'(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes \mathsf{X})\alpha)\} \\ &\propto \exp\{-0.5(\mathcal{Y} - \mathcal{X}\alpha)'(\mathcal{Y} - \mathcal{X}\alpha)\} \\ &\propto \exp\{-0.5(\alpha-\tilde{\alpha})'\mathcal{X}'\mathcal{X}(\alpha-\tilde{\alpha}) + (\mathcal{Y} - \mathcal{X}\tilde{\alpha})'(\mathcal{Y} - \mathcal{X}\tilde{\alpha})\} \end{aligned} \tag{10.8}$$

where

$$\tilde{\alpha} = (\mathcal{X}'\mathcal{X})^{-1}(\mathcal{X}'\mathcal{Y}) = [\bar{\Sigma}_a^{-1} + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}[\bar{\Sigma}_a^{-1}\bar{\alpha} + (\Sigma_e^{-1} \otimes \mathsf{X})'y] \tag{10.9}$$

Since $\Sigma_e$ and $\bar{\Sigma}_a$ are fixed, the second term in (10.8) is a constant independent of $\alpha$ and

$$g(\alpha|y) \propto exp[-0.5(\alpha-\tilde{\alpha})'\mathcal{X}'\mathcal{X}(\alpha-\tilde{\alpha})] \propto exp[-0.5(\alpha-\tilde{\alpha})'\tilde{\Sigma}_a^{-1}(\alpha-\tilde{\alpha})] \tag{10.10}$$

Hence, the posterior density of $\alpha$ is Normal with mean $\tilde{\alpha}$ and variance $\tilde{\Sigma}_a = [\bar{\Sigma}_a^{-1} + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}$. For (10.10) to be operational we need $\bar{\Sigma}_a$ and $\Sigma_e$. Typically, $\bar{\Sigma}_a$ is arbitrarily chosen (e.g. to have a loose prior) and one uses e.g., $\Sigma_{e,ols} = \frac{1}{T-1}\sum_{t=1}^{T} e'_{t,ols}e_{t,ols}$, $e_{t,ols} = y_t - (I_m \otimes \mathsf{X})\alpha_{ols}$, in the formulas.

### 10.2.1 Least square under uncertain restrictions

The posterior mean for $\alpha$ displayed in (10.9) has the same format as a classical estimator obtained with Theil's mixed type approach when coefficients are stochastically restricted. To illustrate this point consider a univariate AR(q) with no constant:

$$\begin{aligned} \mathsf{Y} &= \mathsf{X}A + \mathsf{E} \qquad \mathsf{E} \sim (0, \Sigma_e) \\ A &= \bar{A} + v_a \qquad v_a \sim (0, \bar{\Sigma}_a) \end{aligned} \tag{10.11}$$

where $A = [A_1, \ldots A_q]'$, $\mathsf{X}_t = [y_{t-1}, \ldots y_{t-q}]$. Set $\mathcal{Y}_t = [\mathsf{Y}_t, \bar{A}']'$, $\mathcal{X}_t = [\mathsf{X}_t, I]'$, $E_t = [\mathsf{E}_t, v'_a]'$. Then $\mathcal{Y}_t = \mathcal{X}_t A + E_t$, where $E_t \sim (0, \Sigma_E)$, and $\Sigma_E$ is assumed known. The (generalized) least square estimator is $A_{GLS} = (\mathcal{X}'\Sigma_E^{-1}\mathcal{X})^{-1}(\mathcal{X}'\Sigma_E^{-1}\mathcal{Y})$, which is identical to $\tilde{A}$, the mean of the posterior of $A$ obtained with fixed $\Sigma_e$, fixed $\bar{\Sigma}_a$ and a Normal prior for $A$. There

is a simple but useful interpretation of this result. Prior restrictions on VAR coefficients can be treated as dummy observations which are added to the system of VAR equations. The posterior estimator will efficiently combine sample and prior information using their precisions as weights. Additional restrictions can be tagged on to the system in exactly the same fashion and posterior estimates can be obtained by combining the vector of prior restrictions with the data. We will exploit this feature later on, when we design restrictions intended to capture the existence of trends, seasonal fluctuations, etc.

Exercise 10.1 *(Hoerl and Kennard) Suppose that* $\bar{A} = 0$ *in (10.11). Show that the posterior mean of* $A$ *is* $\tilde{A} = (\bar{\Sigma}_a^{-1} + \mathsf{X}'\Sigma_e^{-1}\mathsf{X})^{-1}(\mathsf{X}'\Sigma_e^{-1}\mathsf{Y})$. *Show that if* $\Sigma_e = \sigma_e^2 \times I_T$, $\bar{\Sigma}_a = \sigma_v^2 \times I_q$, $\tilde{A} = (I_q + \frac{\sigma_e^2}{\sigma_v^2}(\mathsf{X}'\mathsf{X})^{-1})^{-1}A_{ols}$, *where* $A_{ols}$ *is the OLS estimator of* $A$.

There are two important features of exercise 10.1. First, since the restriction $\bar{A} = 0$ imposes the belief that all the coefficients are small, it is appropriate if $y_t$ is the growth rate of financial variables like exchange rates or stock prices. Second, the last part of the exercise indicates that the posterior estimator increases the smallest eigenvalues of the data matrix by the factor $\frac{\sigma_e^2}{\sigma_v^2}$. Hence, it is useful when the $(\mathsf{X}'\mathsf{X})$ matrix is ill-conditioned (e.g. when near multi-collinearity is present).

Exercise 10.2 *Treating* $\tilde{\alpha}$ *in (10.9) as a classical estimator, show what conditions insure its consistency and its asymptotic normality.*

There is an alternative representation of the prior for case 1. Set $R\alpha = r + v_a$, $v_a \sim \mathbb{N}(0, I)$, where $R$ is a square matrix. Then $g(\alpha)$ is $\mathbb{N}(R^{-1}r, R^{-1}R^{-1'})$ and $\tilde{\alpha} = [R'R + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}[R'r + (\Sigma_e^{-1} \otimes \mathsf{X})'y]$. This last expression has two advantages over (10.9). First, it does not require the inversion of the $mk \times mk$ matrix $\bar{\Sigma}_a$, which could be complicated in large scale VARs. Second, zero restrictions on some coefficients are easy to impose - in (10.9) this must be done setting some diagonal elements of $\bar{\Sigma}_a$ to infinity.

Exercise 10.3 *Using* $R\alpha = r + v_a$, $v_a \sim \mathbb{N}(0, I)$ *as a prior, show that* $\sqrt{T}(\tilde{\alpha} - \alpha_{ols}) \xrightarrow{P} 0$ *as* $T \to \infty$.

The intuition for the result of exercise 10.3 is clear: since as $T$ grows, the importance of the data increases relative to the prior, $\tilde{\alpha}$ coincides with the unrestricted OLS estimator.

### 10.2.2 The Minnesota prior

The so-called Minnesota (Litterman) prior is a special case of Case 1 prior when $\bar{\alpha}$ and $\Sigma_\alpha$ are functions of a small number of hyperparameters. In particular (see, for example, RATS (2000)) this prior assumes that $\bar{\alpha} = 0$ except for $\bar{\alpha}_{i1} = 1$, $i = 1, \ldots, m$; that $\Sigma_a$ is diagonal and that the $\sigma_{ij,\ell}$ element corresponding to lag $\ell$ of variable $j$ in equation $i$ has the form:

$$\sigma_{ij,\ell} = \frac{\phi_0}{h(\ell)} \qquad \text{if} \quad i = j, \forall \ell$$

$$
\begin{aligned}
&= \phi_0 \times \frac{\phi_1}{h(\ell)} \times (\frac{\sigma_j}{\sigma_i})^2 \ \text{ otherwise when } i \neq j, j \text{ endogenous}, \forall \ell \\
&= \phi_0 \times \phi_2 \ \text{ for j exogenous} \qquad (10.12)
\end{aligned}
$$

Here $\phi_i,\ i = 0, 1, 2$ are hyperparameters, $(\frac{\sigma_j}{\sigma_i})^2$ is a scaling factor and $h(\ell)$ a deterministic function of $\ell$. The prior (10.12) captures features of interest to the investigator: $\phi_0$ represents the tightness on the variance of the first lag; $\phi_1$ the relative tightness of other variables; $\phi_2$ the relative tightness of the exogenous variables and $h(\ell)$ the relative tightness of the variance of lags other than the first one. Typically, one assumes an harmonic decay $h(\ell) = \ell^{\phi_3}$(a special case of which is $h(\ell) = \ell$, a linear decay) or a geometric decay $h(\ell) = \phi_3^{-\ell+1}, \phi_3 > 0$. Since $\sigma_i, i = 1, \ldots, m$ are unknown, consistent estimates of the standard errors of the variables $i, j$ are used in (10.12).

To understand the logic of this prior note that the $m$ time series are a-priori represented as random walks. This specification is selected because univariate random walk models are typically good in forecasting macroeconomic time series. Note also that the random walk hypothesis is imposed a-priori: a posteriori, each time series may follow a more complicated process if there is sufficient information in the data to require it.

The variance-covariance matrix is a-priori selected to be diagonal. Hence, there is no relationship among the coefficients of various VAR equations. Moreover, the most recent lags of a variable are expected to contain more information about the variable's current value than do earlier lags. Hence, the variance of lag $\ell_2$ is smaller than the variance of lag $\ell_1$ if $\ell_2 > \ell_1$ for every endogenous variable of the model. Furthermore, since lags of other variables typically have less information than lags of own variables, $\phi_1 \leq 1$. Note that, if $\phi_1 = 0$, the VAR is a-priori collapsed into a vector of univariate models. Finally, $\phi_2$ regulates the relative importance of the information contained in the exogenous variables and $\phi_0$ controls the relative importance of sample and prior information. From (10.9) if $\phi_0$ is large, prior information becomes diffuse so the posterior distribution mirrors sample information. If $\phi_0$ is small, prior information dominates.

A graphical representation of this prior is in figure 10.1: all coefficients have zero prior mean (except the first own lag) and prior distributions become more concentrated for coefficients on longer lags. Moreover, the prior distributions of the lags of the variables not appearing on the left hand side of the equation are more concentrated than those of the own lags.

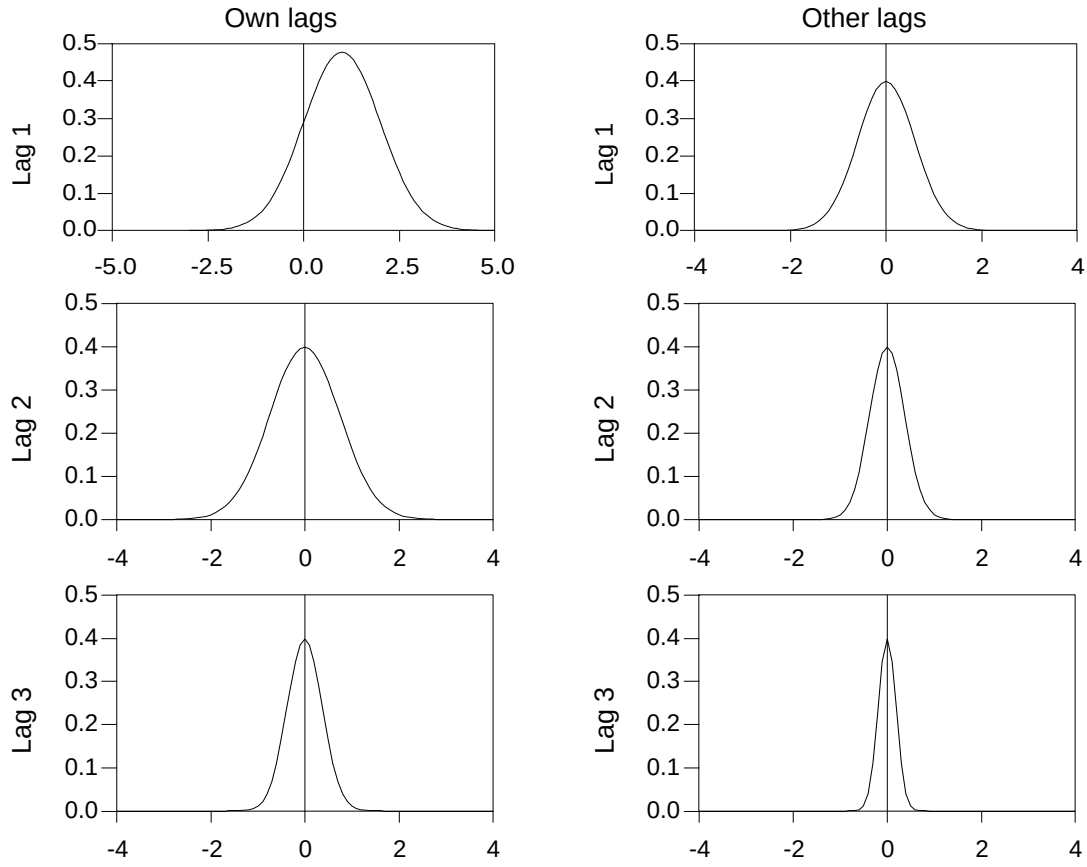


Figure 10.1: Minnesota prior.

Example 10.1 *To see what the Minnesota prior implies consider a VAR(2) with* $q = 2$, $\bar{y} = 0$, $h(\ell) = \ell$. *In this case* $\bar{\alpha} = [1, 0, 0, 0, 0, 1, 0, 0]$, *and*

$$\bar{\Sigma}_a = \begin{bmatrix} \phi_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_0\phi_1(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\phi_0}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_0\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2} \end{bmatrix}$$

There are considerable advantages in specifying $\bar{\Sigma}_a$ to be diagonal. Since the same variables appear on the right hand side of each equations, a diagonal $\bar{\Sigma}_a$ implies a diagonal $\tilde{\Sigma}_a$ so that $\tilde{\alpha}$ is the same as the vector of $\tilde{\alpha}_i$ computed equation by equation. This property is lost with other prior specifications, regardless of the assumption made on $\bar{\Sigma}_a$.

Exercise 10.4 *Using the logic of seemingly unrelated regressions show that when $g(\alpha)$ is of Minnesota type, estimating the VAR jointly gives the same posterior estimator for the coefficients of equation $i$ as estimating each VAR equation separately.*

The dimension of $\alpha$ for moderate VARs is typically large: for example, if there are 5 endogenous variables, 5 lags and a constant, $k = 26$ and a $mk = 130$. With standard macro data (say, forty years of quarterly data (T=160)), maximum likelihood estimates are unlikely to have reasonable properties. The Minnesota type makes this large number of coefficients depend on a smaller vector of hyperparameters. If these are the objects estimated from the data, a better precision is expected because of the sheer dimensionality reduction (the noise to signal ratio is smaller; the number of data points per parameter increased), and out-of-sample forecasts can be improved. Note that even when the prior is false, in the sense that it does not reflect well sample information, this approach may reduce the MSE of the estimates. A number of authors have shown that VARs with a Minnesota prior produce superior forecasts to those of, say, univariate ARIMA models or traditional multivariate simultaneous equations (see e.g. Robertson and Tallman (1999) for a recent assessment). Therefore, it is not surprising that BVARs are routinely used for short-term macroeconomic forecasting in Central Banks and international institutions.

It is useful to contrast the Minnesota approach and other methods used to deal with the "curse of dimensionality". In classical approaches, "unimportant" lags are purged from the specification using t-test or similar procedures (see e.g. Favero (2001)). This approach therefore imposes strong a-priori restrictions on what variables and which lags should be in the VAR. However, dogmatic restrictions are unpalatable because they are hard to justify on both economic and statistical grounds. The Minnesota prior introduces restrictions in a flexible way: it imposes probability distributions on the coefficients of the VAR which reduce the dimensionality of the problem and, at the same time, give a reasonable account of the uncertainty faced by an investigator.

The choice of $\phi = (\phi_0,\ \phi_1,\ \phi_2,\ \phi_3)$ is important since if the prior is too loose, overfitting is hard to avoid; while if it is too tight, the data is not allowed to speak. There are three approaches one can use. In the first two, one obtains estimates of $\phi$ and plug-in these estimates into the expression for $\bar{\alpha}$ and $\bar{\Sigma}_a$. Then the posterior distribution of $\alpha$ can be obtained from (10.9) in an Empirical Bayes fashion, conditional on the $\phi$ estimates. In the third approach, one treats $\phi$ as random, assumes a prior distribution and computes fully hierarchical posterior estimates of $\alpha$. To do this we need MCMC methods. For now we focus on the first two methods.

One way to choose $\phi$ is to use simple rules of thumb or experience. The RATS manual (2000), for example, suggests as default values $\phi_0 = 0.2,\ \phi_1 = 0.5, \phi_2 = 10^5$, an harmonic

specification for $h(\ell)$ with $\phi_3 = 1$ or 2, implying a relatively loose prior on the VAR coefficients and an uninformative prior for the exogenous variables. These values work reasonably well in forecasting a number of macroeconomic and financial variables and should be used as a benchmark or as starting points for further investigations.

The alternative is to estimate $\phi$ using the information contained in the data. In particular, the predictive density $f(\phi|y) = \int \mathcal{L}(\alpha|y,\phi)g(\alpha|\phi)d\alpha$, constructed on a training sample $(-\tau,\ldots,0)$, could be used. The next example shows how to do this in a simple model.

**Example 10.2** *Suppose $y_t = Ax_t + e_t$, where $A$ is a random scalar,$e_t \sim \mathbb{N}(0,\sigma_e^2)$; $\sigma_e^2$ known and let $A = \bar{A} + v_a$; $v_a \sim \mathbb{N}(0,\bar{\sigma}_a^2)$, $\bar{A}$ is fixed and $\bar{\sigma}_a^2 = h(\phi)^2$ where $\phi$ is a vector of hyperparameters. Then $y_t = \bar{A}x_t + \epsilon_t$ where $\epsilon_t = e_t + v_a x_t$ and the posterior kernel is:*

$$\breve{g}(\alpha,\theta|y) = \frac{1}{\sqrt{2\pi}\sigma_e h(\phi)}\exp\{-0.5\frac{(y-Ax)^2}{\sigma_e^2} - 0.5\frac{(A-\bar{A})^2}{h(\phi)^2}\} \tag{10.13}$$

*where $y = [y_1,\ldots y_t]', x = [x_1,\ldots x_t]'$.Integrating (10.13) with respect to $A$ we obtain*

$$f(\phi|y) = \frac{1}{\sqrt{2\pi h(\phi)^2 tr|x'x| + \sigma_e^2}}\exp\{-0.5\frac{(y-\bar{A}x)^2}{\sigma_e^2 + h(\phi)^2 tr|x'x|}\} \tag{10.14}$$

*which can be constructed and maximized, e.g., using the prediction error decomposition generated by the Kalman filter.*

While in example 10.2 $A$ is a scalar, the same logic applies when $\alpha$ is a vector.

**Exercise 10.5** *Let $y_t = A(\ell)y_{t-1} + e_t$, $e_t \sim \mathbb{N}(0,\Sigma_e)$, $\Sigma_e$ known, let $\alpha = vec(A_1,\ldots,A_q)'$ $= \bar{\alpha} + v_a$, $\bar{\alpha}$ known and $\bar{\Sigma}_a = h(\phi)^2$. Show $f(\phi|y)$ and its prediction error decomposition.*

**Exercise 10.6** *Suppose that $\bar{A} = h_1(\phi)$ and $\bar{\Sigma}_a = h_2(\phi)$ in example 10.2. Derive the first order conditions for the optimal $\phi$. Describe how to numerically find ML-II estimates of $\phi$.*

We summarize the features of the posterior distribution of $\alpha$ and $\Sigma_e$ obtained with the other three prior specifications in the next exercises (see Kadiyala and Karlsson (1997)).

**Exercise 10.7** *Suppose that $g(\alpha,\Sigma_e^{-1}) \propto |\Sigma_e^{-1}|^{0.5(m+1)}$. Show that the joint posterior has a Normal-Wishart shape with $(\alpha|\Sigma_e,y) \sim \mathbb{N}(\alpha_{ols},(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})^{-1})$ ; $(\Sigma_e^{-1}|y) \sim \mathbb{W}([(y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols})]^{-1}, T-k)$ and that $(\alpha|y)$ has a t-distribution with parameters $((y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols}),\alpha_{ols},T-k)$, where $\alpha_{ols}$ is the OLS estimator of $\alpha$. Conclude that, a-posteriori, the elements of $\alpha$ are dependent (Hint: Stare at the variance of $\alpha$).*

**Exercise 10.8** *Suppose that the joint prior for $(\alpha,\Sigma_e^{-1})$ is Normal-diffuse, i.e. $g(\alpha) \sim \mathbb{N}(\bar{\alpha},\bar{\Sigma}_a)$ where both $\bar{\alpha}$ and $\bar{\Sigma}_a$ are known and $g(\Sigma_e) \propto |\Sigma_e^{-1}|^{0.5(m+1)}$. Show that $g(\alpha|y) \propto \exp\{0.5(\alpha-\bar{\alpha})'\bar{\Sigma}_a^{-1}(\alpha-\bar{\alpha})\} \times |(y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols}) + (\alpha-\alpha_{ols})'(\mathsf{X}'\mathsf{X})(\alpha-$*

$\alpha_{ols})|^{-0.5T}$. *Conclude that* $g(\alpha|y)$ *is the product of the normal prior and the same t-distribution found in exercise 10.7. Argue that there is posterior dependence among equations, even when* $\bar{\Sigma}_a$ *is diagonal.*

Exercise 10.9 *Let* $g(\alpha|\Sigma_e) \sim \mathbb{N}(\bar{\alpha}, \Sigma_e \otimes \bar{\Omega})$ *and* $g(\Sigma_e^{-1}) \sim \mathbb{W}(\bar{\Sigma}^{-1}, \bar{\nu})$. *Show that* $g(\alpha|\Sigma_e, y) \sim \mathbb{N}(\tilde{\alpha}, \Sigma_e \otimes \tilde{\Omega})$, $g(\Sigma_e^{-1}|y) \sim \mathbb{W}(\tilde{\Sigma}^{-1}, T + \bar{\nu})$. *Give the form of* $\tilde{\alpha}, \tilde{\Omega}, \tilde{\Sigma}^{-1}$. *Show that* $(\alpha|y)$ *has a t-distribution with parameters* $(\tilde{\Omega}^{-1}, \tilde{\Sigma}, \tilde{\alpha}, T + \bar{\nu})$. *Assume that* $\bar{\Omega} = \text{diag}\{\bar{\omega}_{ii}\}$ *where* $\bar{\omega}_{ii}$ *is parametrized as in the Minnesota prior (except that* $\phi_1 = 1$*); suppose that* $\bar{\nu} = m + 2$ *and that* $\bar{\sigma}_{ii} = \text{diag}(\bar{\Sigma}) = (\bar{\nu} - m - 1)s_i^2$, *where* $s_i^2$ *is the estimated variance of* $e_i$. *Show that there is posterior dependence among the equations.*

### 10.2.3 Adding other prior restrictions

We can add a number of other statistical restrictions to the standard Minnesota prior without altering the form of the posterior moments. For example, an investigator may be interested in studying the dynamics at seasonal frequencies and therefore want to use the seasonal information to set up prior restrictions. The simplest way to deal with seasonality is to include a set of dummies in the VAR and treat their coefficients in the same way as the coefficients on the constant.

Example 10.3 *In quarterly data, a prior for a bivariate VAR(2) with four seasonal dummies has mean equal to* $\bar{\alpha} = [1, 0, 0, 0, 0, 0, 0, 0, 0|0, 1, 0, 0, 0, 0, 0, 0, 0]$ *and the block of* $\Sigma_a$ *corresponding to the seasonal dummies has diagonal elements,* $\sigma_{dd} = \phi_0\phi_s$. *Here* $\phi_s$ *represents the tightness of the seasonal information (and a large* $\phi_s$ *implies little prior information).*

Seasonality, however, is hardly deterministic (in that case, it would be easy to eliminate it if we did not want it) and seasonal dummies only roughly account for seasonal variations. As an alternative, note that seasonal series display a peak (or a wide mass) in the spectrum at some or all seasonal frequencies. When a series has a peak at frequency $\omega_0$ it must be the case that in the model $y_t = D(\ell)e_t$, $|D(\omega_0)|^2$ is large. A large $|D(\omega_0)|^2$ implies that $|A(\omega_0)|^2$ should be small, where $A(\ell) = D(\ell)^{-1}$, which in turns implies $\sum_{j=1}^{\infty} A_j cos(\omega_0 j) \approx -1$.

Example 10.4 *In quarterly data,* $\omega_0 = \frac{\pi}{2}, \pi$ *(cycles corresponding to 4 and 2 quarters) and a peak at, say,* $\frac{\pi}{2}$ *implies that* $-A_2 + A_4 - A_6 + A_8 + \ldots$ *must be close to* $-1$.

The same idea applies to multivariate models. Omitting constants, the MA representation is $y_t = D(\ell)e_t$ and the spectral density of $y_t$ is $\mathcal{S}_y(\omega) = |D(\omega)|^2 \frac{\Sigma_e}{2\pi}$. Since $D(\omega) = \sum_j D_j(cos(\omega j) + isin(\omega j))$, a peak in $\mathcal{S}_y$ at $\omega_0$ implies that $\sum_j D_j cos(\omega_0 j)$ is large and $\sum_{j=1}^{\infty} A_j cos(\omega_0 j) \approx -1$.

We can cast these restrictions in the form $R\alpha = r + v_a$, where $r = [-1, \ldots, -1]'$, $R$ is a $m_1 \times mk$ matrix and $m_1$ is the number of seasonal frequencies. In quarterly data, if the first variable of the VAR displays seasonality at both $\frac{\pi}{2}, \pi$ then:

$$R = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & -1 & \ldots & 0 \\ -1 & 1 & -1 & 1 & -1 & 1 & \ldots & 0 \end{bmatrix}$$

These restrictions can be added to those of the original (Minnesota) prior and combined with the data using the logic of Theil's mixed type estimation, once $\Sigma_{v_a}$ is selected. The same approach can also be used to account for the presence of peaks in other parts of the spectrum, as it is shown in the next exercise.

Exercise 10.10 *(Canova)*
*(i) Show that a peak in the spectral density at frequency zero in variable i implies* $\sum_{j=1}^{\infty} A_{ji} \approx -1$. *Cast this constraint in the form of an uncertain linear restriction.*
*(ii) Show that a large mass in the band* $(\frac{2\pi}{j} \pm \varepsilon)$, *some* $j$, $\varepsilon$ *small, in variable i implies* $\sum_{j=1}^{\infty} A_{ji} cos(j\omega_0) \approx -1$, *for all* $\omega_0$ *in the band. Cast these constraints in the form of uncertain linear restrictions.*
*(iii) Show that a high coherence at* $\omega_0 = \frac{\pi}{2}$ *in series i and* $i'$ *of a VAR implies that* $\sum_{j=1}^{\infty}(-1)^j A_{i'i'(2j)} + \sum_{j=1}^{\infty}(-1)^j A_{ii(2j)} \approx -2$. *Cast this constraint in the form of an uncertain linear restriction.*

Other types of probabilistic constraints can be imposed in a similar way. As long as $r$, $R$ and $\text{var}(v_a)$ are fixed, combining prior and sample information presents no conceptual difficulty: the dimensionality of $R$ and of $r$ changes, but the form of the posterior moments of $\alpha$ is unchanged.

### 10.2.4 Some Applied tips

There are few practical issues a researcher faces in setting-up a Minnesota prior for a VAR. First, in simple applications it is typical to use default values for the hyperparameters $\phi$. While this is a good starting point, it is not clear that this choice is appropriate in all forecasting situations or when structural inference is required. In these cases, sensitivity analysis may give information about interesting local derivatives, e.g. how much the MSE of the forecasts change when $\phi$ varies within a small range of the default value. If differences are large, should hyperparameters be chosen to get the best out-of-sample performance? Since hyperparameters describe features of the prior they should be chosen using the predictive density. Using ex-post MSE statistics poses few operational problems. Which forecasting horizon should be chosen to select the hyperparameters? If different horizons require different parameters, how should one proceed? The use of the predictive density provides a natural answer to these questions. Since predictive densities can be decomposed into the product of one-step ahead prediction errors, hyperparameters chosen optimizing the predictive density minimize the one-step ahead prediction error in the training sample.

Second, in certain applications the defaults values of the Minnesota prior are clearly inappropriate: for example, a mean of one on the first lag for growth rates is unlikely to be useful. In others, one may want to have additional parameters controlling, e.g., the relative importance of certain variables in one equation or across equations. For example, one would expect lags of other variables to be less important when the left hand side of an equation there is a financial variable, but very important when there is a macroeconomic variable.

Alterations of the Minnesota prior in this direction do not change the form of the posterior so long as $\bar{\Sigma}_a$ is diagonal and $\Sigma_e$ fixed.

Although the emphasis of this section has been on type 1 priors, all the arguments made remain valid when a general Normal-Wishart prior are used. Conditional on $\Sigma_e$ the posterior for $\alpha$ is still normal. However, equation-by-equation computations are no longer efficient since the posterior covariance matrix obtained using the whole system is different from the covariance matrix obtained using each equation separately. For VARs with 5 or 6 variables and 4 or 5 lags, system wide calculations are not computationally demanding, given existing computer technology. For larger scale models such as the one of Leeper, Sims and Zha (1996), intelligent choices for the prior may dramatically simplify the computations.

How do one selects the variables to be included in a BVAR? Using the same logic described in chapter 9, specifications with different variables can be treated as different models. Therefore, a posterior odds ratio or the Leamer's version of it can be used to select the specification that best fit the data in a training sample. Consequently, one chooses the specification with the smallest one-step ahead prediction error will be preferred. Such calculations can be performed both in nested and non-nested models.

Example 10.5 *(Forecasting inflation) We use a BVAR with a Minnesota prior to forecast inflation rates in Italy. The features of inflation rates have changed dramatically in the 90's all over the world and in Italy in particular. In fact, while the autocovariance function displays remarkable persistence in the 80's (AR(1) coefficient equals 0.85), it decays pretty quickly in the 90's (AR(1) coefficient equals 0.48). In this situation, using 1980's data to choose a model or its hyperparameters may severely impair its ability to forecast in the 90's. As a benchmark for comparison we use a univariate ARIMA model, chosen using standard Box-Jenkins methods, and a three variable unrestricted VAR, including the annualized three month inflation, the unemployment rate and the annualized three month rent inflation, each with four lags. These variables were chosen among a set of ten candidates using Leamer's posterior odds ratio approach. We present results for two alternative specifications: a BVAR with hyperparameters sets using rules of thumb and one with hyperparameters chosen to maximize the predictive density using data from 1980:1 to 1995:4. The prior variance is characterized by a general tightness parameter, a decay parameter and a parameter for lags of other variables. In the first case they are set to 0.2, 1, 0.5, respectively. In the second, they are optimally estimated (point estimates 0.14, 2.06, 1.03). The prior variance on the constant is diffuse. In table 10.1 we report one year ahead Theil-U statistics (the ratio of the MSE of the model to the MSE of a random walk) for the four specifications. Posterior standard error for the two BVAR are in parenthesis.*

| *Sample* | *ARIMA* | *VAR* | *BVAR1* | *BVAR2* |
|---|---|---|---|---|
| *1996:1-2000:4* | *1.04* | *1.47* | *1.09 (0.03)* | *0.97 (0.02)* |
| *1990:1-1995:4* | *0.99* | *1.24* | *1.04 (0.04)* | *0.94 (0.03)* |

Table 10.1: One year ahead Theil-U statistics.

*Three features deserve comments. First, forecasting Italian inflation one year ahead is difficult: all models have a hard time to beat a random walk and three of them do worse. Second, an unrestricted VAR performs poorly. Third, a BVAR with default choices is better than a unrestricted VAR but not better than an ARIMA model. Finally, a BVAR with optimally chosen parameters, outperforms both random walk and ARIMA models at the one year horizon but the gains are small. The results are robust: repeating the exercise using data from 1980:1 to 1989:4 to choose the variables, the hyperparameters and estimate the models and data from 1991:1 to 1995:4 to forecast produces qualitatively similar Theil-Us.*

### 10.2.5 Priors derived from DSGE models

The priors we have considered so far are either statistically motivated or based on rules-of-thumb useful for forecasting macroeconomic time series. In both cases, economic theory plays no role, except perhaps in establishing the range of values for the prior distributions. To be able to use BVARs for purposes other than forecasting, one may want to consider priors based on economic theory. In addition, one may be interested in knowing if theory based priors are as good as statistically based priors in forecasting, unconditionally, out-of-sample.

Here we consider priors which are derived from DSGE models. The nature of the model and a prior for the structural parameters imply a prior for the reduced form VAR coefficients. One can dogmatically take these restrictions or simply consider their qualitative content in constructing posterior distributions. In this setup prior information measures the confidence a researcher has that the DSGE structure has generated the observed data.

An alternative representation for the log-linearized solution of a DSGE model is:

$$y_{2t+1} = \mathcal{A}_{22}(\theta)y_{2t} + \mathcal{A}_{23}(\theta)y_{3t+1} \tag{10.15}$$

$$y_{1t} = \mathcal{A}_{12}(\theta)y_{2t} \tag{10.16}$$

where $y_{2t}$ is a $m_2 \times 1$ vector including the states and the driving forces; $y_{1t}$ is $m_1 \times 1$ vecotr including all the endogenous variables and $y_{3t+1}$ are the shocks. Here $\mathcal{A}_{jj'}(\theta)$ are time invariant functions of $\theta$, the vector of structural (preferences, technologies, policy) parameters of the model. It is easy to transform (10.15)-(10.16) into a (restricted) VAR(1) for $y_t = [y_{1t}, y_{2t}]'$ of the form

$$\begin{bmatrix} 0 & 0 \\ 0 & I_{m_2} \end{bmatrix} y_{t+1} = \begin{bmatrix} -I_{m_1} & \mathcal{A}_{12}(\theta) \\ 0 & \mathcal{A}_{22}(\theta) \end{bmatrix} y_t + \begin{bmatrix} 0 \\ \mathcal{A}_{23}(\theta) \end{bmatrix} y_{3t+1} \tag{10.17}$$

or $\mathcal{A}_0 y_{t+1} = \mathcal{A}_1(\theta)y_t + \epsilon_{t+1}(\theta)$ where $\epsilon_{t+1}(\theta) = \begin{bmatrix} 0 \\ \mathcal{A}_{23}(\theta) \end{bmatrix} y_{3t+1}$. Hence, given a prior for $\theta$, the model implies a prior for $\mathcal{A}_{12}(\theta), \mathcal{A}_{22}(\theta), \mathcal{A}_{23}(\theta)$. In turn these priors imply restrictions for the reduced form parameters $A(\ell) = \mathcal{A}_0^{-1}\mathcal{A}_1(\ell)$ and $\Sigma_e = \mathcal{A}_0^{-1}\Sigma_\epsilon \mathcal{A}_0^{-1}$. Expressions for the priors for $\mathcal{A}_{12}(\theta), \mathcal{A}_{22}(\theta), \mathcal{A}_{23}(\theta)$ can be obtained using $\delta$-approximations, i.e. if $\theta \sim \mathbb{N}(\bar{\theta}, \bar{\Sigma}_\theta)$, $vec(\mathcal{A}_{12}(\theta)) \sim \mathbb{N}(vec(\mathcal{A}_{12}(\bar{\theta})), \frac{\partial vec(\mathcal{A}_{12}(\theta))}{\partial \theta} \bar{\Sigma}_\theta \frac{\partial vec(\mathcal{A}_{12}(\theta))}{\partial \theta}')$, etc.

**Example 10.6** *Consider a VAR(q): $y_{t+1} = A(\ell)y_t + e_t$. From (10.17) the prior for $A_1$ is Normal with mean $\mathcal{A}_0^G\mathcal{A}_1(\bar{\theta})$, where $\mathcal{A}_0^G$ is the generalized inverse of $\mathcal{A}_0$ and variance equal to $\Sigma_a = (A_0^G \otimes I_{m_1+m_2})\Sigma_{a_1}(A_0^G \otimes I_{m_1+m_2})'$); where $\Sigma_{a_1}$ is the variance of $vec(\mathcal{A}_1(\theta))$. A DSGE prior for $A_2, A_3, \ldots$ has a dogmatic form: mean zero and zero variance.*

Since the states of a DSGE model typically include unobservable variables (e.g. the Lagrangian multiplier or the driving forces of the model) or variables measured with error (e.g. the capital stock), it may be more convenient to set up prior restrictions for a VAR composed only of the endogenous variables, as the next example shows.

**Example 10.7** *(Ingram and Whiteman). A RBC model with utility function $u(c_t, c_{t-1}, N_t, N_{t-1}) = \ln(c_t) + \ln(1 - N_t)$ implies a law of motion for the states of the form*

$$\begin{bmatrix} K_{t+1} \\ \ln\zeta_{t+1} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{kk}(\theta) & \mathcal{A}_{k\zeta}(\theta) \\ 0 & \rho_\zeta \end{bmatrix} \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix} + \begin{bmatrix} 0 \\ \epsilon_{1t+1} \end{bmatrix} \equiv \mathcal{A}_{22}(\theta) \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix} + \epsilon_{t+1} \quad (10.18)$$

*where $K_t$ is the capital stock and $\zeta_t$ is a technological disturbance. The equilibrium mapping between the endogenous variables and the states is $[c_t, N_t, gdp_t, inv_t]' = \mathcal{A}_{12}(\theta) \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix}$ where $c_t$ is consumption, $N_t$ hours, $gdp_t$ output and $inv_t$ investments. Here $\mathcal{A}_{12}(\theta)$ and $\mathcal{A}_{22}(\theta)$ are function of $\eta$, the share of labor in production, $\beta$ the discount factor, $\delta$ the depreciation rate, $\rho_\zeta$ the AR parameter of the technology shock. Let $y_{1t} = [c_t, N_t, gdp_t, inv_t]'$ and $y_{2t} = [k_t, \ln\zeta_t]'$, $\theta = (\eta, \beta, \delta, \rho_\zeta)$. Then $y_{1t} = A(\theta)y_{1t-1} + e_{1t}$, where $A(\theta) = \mathcal{A}_{12}(\theta)\mathcal{A}_{22}(\theta)(\mathcal{A}_{12}(\theta)'\mathcal{A}_{12}(\theta))^{-1}\mathcal{A}_{12}(\theta)$, $e_{1t} = \mathcal{A}_{12}(\theta)\epsilon_t$ and $(\mathcal{A}_{12}(\theta)'\mathcal{A}_{12}(\theta))^{-1}\mathcal{A}_{12}(\theta)$ is the generalized inverse of $\mathcal{A}_{12}(\theta)$. If $g(\theta)$ is $\theta \sim \mathbb{N}(\begin{bmatrix} 0.58 \\ 0.988 \\ 0.025 \\ 0.95 \end{bmatrix}, \begin{bmatrix} 0.0006 & & & \\ & 0.0005 & & \\ & & 0.0004 & \\ & & & 0.00015 \end{bmatrix})$, the prior mean of $A(\theta)$ is $A(\bar{\theta}) = \begin{bmatrix} 0.19 & 0.33 & 0.13 & -0.02 \\ 0.45 & 0.67 & 0.29 & -0.10 \\ 0.49 & 1.32 & 0.40 & 0.17 \\ 1.35 & 4.00 & 1.18 & 0.64 \end{bmatrix}$ which implies, e.g., substantial feedback from consumption, output and hours to investment (see the last row). The prior variance for $A(\theta)$ is $\Sigma_A = \frac{\partial A(\theta)}{\partial \theta'}\bar{\Sigma}_\theta\frac{\partial A(\theta)}{\partial \theta'}$, where $\frac{\partial A(\theta)}{\partial \theta'}$ is a $16 \times 4$ vector. Hence, a RBC prior for $y_{1t}$ implies a normal prior on the first lag with mean $A(\bar{\theta})$ and variance proportional to $\Sigma_A$. To relax the dogmatic prior restriction on higher lags, we could assume a Normal prior with zero mean and variance $\propto \frac{\Sigma_A}{h(\ell)}$ where $h(\ell)$ is a decaying function of $\ell$.*

**Exercise 10.11** *(RBC cointegrating prior). In example 10.7 suppose that $(\ln\zeta_t)$ has a unit root. Then all endogenous variables must have a unit root and the stochastic trend is a common one.*
*(i) Argue that $(I - \mathcal{A}_{kk}(\theta), -\mathcal{A}_{k\zeta}(\theta))$ must be a cointegrating vector for $k_t$.*
*(ii) Argue that $(I_4, -\mathcal{A}_{12}(\theta))$ must be a cointegrating vector for $y_{1t}$*
*(iii) Given a Normal prior on $\theta$, derive a cointegrating prior for the $\mathcal{A}'s$.*

**Exercise 10.12** *Suppose consumers maximize* $u(c_t, c_{t-1}, N_t) = \ln c_t - \epsilon_{2t} \ln N_t$ *subject to the constraint* $c_t + B_{t+1} \leq y_t + (1 + r_t^B)B_t - T_t$ *where* $y_t = N_t \epsilon_{1t}$, $\epsilon_{1t}$ *is a technology shock with mean* $\bar{\epsilon}_1$ *and variance* $\sigma^2_{\epsilon_1}$ *and* $\epsilon_{2t}$ *is a labor supply shock with mean of* $\bar{\epsilon}_2$ *and variance* $\sigma^2_{\epsilon_2}$. *Here* $T_t$ *are lump sum taxes,* $B_t$ *are real bonds and the government finances a random stream of expenditure using lump sum taxes and real bonds according to the budget constraint* $G_t - T_t = B_{t+1} - (1 + r_t^B)B_t$. *In this model there are three shocks: two supply type shocks* $(\epsilon_{1t}, \epsilon_{2t})$ *and one demand type shock* $(G_t)$.
*i) Find a log-linearized solution for* $N_t$, $y_t, c_t$ *and labor productivity* $(np_t)$.
*ii) Use the results in i) to construct a prior for a bivariate VAR in hours and output. Derive the posterior distribution for the VAR parameters and the covariance matrix of the shocks. Be precise about the assumptions and the choices you make (Careful, there are three shocks and two variables). Would it make a difference for the answer if you would have used a trivariate model with consumption or labor productivity?*
*iii) Describe how to construct impulse responses to* $G_t$ *shocks using posterior estimates.*
*iv) Suppose that, for identification purposes, an investigator makes the assumption that demand shocks have zero contemporaneous effect on hours. Is this assumption reasonable in the logic of the model? Under what conditions the estimated demand shocks you recover from posterior analysis correctly represent* $G_t$ *shocks?*

Del Negro and Schorfheide (2003) have suggested an alternative way to append priors derived from DSGE models onto a VAR. The advantage of their approach is that the posterior distributions for both VAR and DSGE parameters can be simultaneously obtained. The basic specification they use differs from the one so far described in an important way. Up to now a DSGE model has provided only the "form" of the prior restrictions (zero mean on lags greater than one, etc.). Here the prior is more tightly based on the data produced by the DSGE model.

The logic of the approach is simple. Since the prior can be thought as an additional observation tagged on to the VAR, one way to add DSGE information is to augment the VAR for the actual data with a prior based on data simulated from the model. The proportion of actual and simulated data points then reflects the relative importance that a researcher gives to the two types of information.

Let the data be represented by a VAR with parameters $(\alpha, \Sigma_e)$. Assume that $g(\alpha, \Sigma_e)$ is of the form $\alpha \sim \mathbb{N}(\bar{\alpha}(\theta), \bar{\Sigma}(\theta))$; $\Sigma_e^{-1} \sim \mathbb{W}(T_s \bar{\Sigma}_e(\theta), T_s - k)$ where

$$\begin{aligned} \bar{\alpha}(\theta) &= ((X^s)'X^s)^{-1}((X^s)'y^s) \\ \bar{\Sigma}(\theta) &= \Sigma_e(\theta) \otimes ((X^s)'X^s)^{-1} \\ \bar{\Sigma}_e(\theta) &= (y^s - X^s\bar{\alpha}(\theta))(y^s - (X^s)\bar{\alpha}(\theta))' \end{aligned} \tag{10.19}$$

Here $y^s$ is data simulated from the DSGE model, $X^s = (I_m \otimes \mathsf{X}^s)$ is a matrix of lags in the VAR representation of simulated data and $\theta$ the structural parameters. In (10.19), the moments of $g(\alpha, \Sigma_e)$ depend on $\theta$ through the simulated data $(y^s, \mathsf{X}^s)$. If $T_s$ measures the length of simulated data, $\kappa = \frac{T_s}{T}$ controls the relative importance of the information

contained in actual and simulated data. Clearly, if $\kappa \to 0$, the actual data dominates and if $\kappa \to \infty$, the simulated data dominates.

The model has a hierarchical structure $f(\alpha, \Sigma_e|y)g(\alpha|\theta)g(\Sigma_e|\theta)g(\theta)$. Conditional on $\theta$, the posterior for $\alpha, \Sigma_e$ are easily derived. In fact, since the likelihood and the prior are conjugate $(\alpha|\theta, y, \Sigma_e) \sim \mathbb{N}(\tilde{\alpha}(\theta), \tilde{\Sigma}(\theta))$; $(\Sigma_e^{-1}|\theta, y) \sim \mathbb{W}((\kappa+T)\tilde{\Sigma}_e(\theta), T+\kappa-k)$ where

$$\begin{aligned}
\tilde{\alpha}(\theta) &= (\kappa\frac{(X^s)'X^s}{T^s} + \frac{X'X}{T})^{-1}(\kappa\frac{(X^s)'y^s}{T^s} + \frac{X'y}{T}) \\
\tilde{\Sigma}(\theta) &= \Sigma_e(\theta) \otimes ((X^s)'X^s + X'X)^{-1} \qquad (10.20) \\
\tilde{\Sigma}_e(\theta) &= \frac{1}{(1+\kappa)T}[(y^s)'y^s + y'y) - ((y^s)'X^s + y'X)((X^s)'X^s + X'X)^{-1}((X^s)'y^s + X'y)]
\end{aligned}$$

where $X = (I \otimes \mathsf{X})$. The posterior for $\theta$ can be computed using the hierarchical structure of the model. In fact, $g(\theta|y) \propto f(\alpha, \Sigma_e, y|\theta)g(\theta)$ where $f(\alpha, \Sigma_e, y|\theta) \propto |\Sigma_e|^{-0.5(T-m-1)} \exp\{-0.5 tr[\Sigma_e^{-1}(y-X\alpha)'(y-X\alpha)\} \times |\bar{\Sigma}_e(\theta)|^{-0.5(T_s-m-1)} \exp\{-0.5tr[\Sigma_e^{-1}(y^s - X^s\bar{\alpha}(\theta))'(y^s - X^s\bar{\alpha}(\theta))\}$. We will discuss how to draw from this posterior in chapter 11.

Exercise 10.13 *Use the fact that $g(\alpha, \Sigma_e, \theta|y) = g(\alpha, \Sigma_e|y, \theta)g(\theta|y)$, to suggest an algorithm to draw sequences for $(\alpha, \Sigma_e)$. How do you compute impulse responses in the VAR?*

Exercise 10.14 *Suppose $g(\Sigma_e)$ is non-informative. Show the form of $(\tilde{\alpha}, \tilde{\Sigma}_e)$ in this case.*

All posterior moments in (10.20) are conditional on a value of $\kappa$. Since this parameter regulates the relative importance of sample and prior information it is important to appropriately select it. As in standard BVAR, there are two ways to proceed. First, we can use a rule of thumb, e.g. set $\kappa = 1$, meaning that $T$ simulated data are added to the actual ones. Second, we can choose it to maximize the predictive density of the model.

Exercise 10.15 *Show the form of $f(y|\kappa)$. Describe how to find its maximum numerically.*

Exercise 10.16 *Consider the working capital model described in exercise 1.14 of chapter 2 driven by shocks to technology, government expenditure and the monetary policy rule. Choose appropriate priors for the parameters (for example, Normal, Gamma or Beta for parameters that lie in an interval). Simulate data for output, inflation and the nominal interest rate. Combine this data with actual data for output, inflation and the nominal interest rate. Explore the predictive density of inflation numerically for different values of $\kappa$. Is there a relationship between the $\kappa$ which maximizes the predictive density and the one which minimizes the MSE of the forecasts? How would you compare such a model against a sticky price, sticky wage model?*

### 10.2.6 Probability distributions for forecasts: Fan Charts

BVAR models can be used to construct probability distributions for future events and therefore are well suited to produce e.g. fan charts or probabilities of turning points. To see how this can be done, set $\bar{y} = 0$ and rewrite the VAR model in a companion form

$$\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t \tag{10.21}$$

where $\mathbb{Y}_t$ and $\mathbb{E}_t$ are $mq \times 1$ vectors, $\mathbb{A}$ is a $mq \times mq$ matrix.

Repeatedly substituting we have $\mathbb{Y}_t = \mathbb{A}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{A}^j \mathbb{E}_{t-j}$ or $y_t = \mathbb{S}\mathbb{A}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{A}^j e_{t-j}$ where $\mathbb{S}$ is such that $\mathbb{S}\mathbb{Y}_t = y_t$, $\mathbb{S}\mathbb{E}_t = e_t$ and $\mathbb{S}'\mathbb{S}E_t = E_t$. A "point" forecast for $y_{t+\tau}$ is obtained plugging-in some location measures of the posterior of $\mathbb{A}$ into $y_t(\tau) = \mathbb{S}\mathbb{A}^\tau \mathbb{Y}_t$. Call this point forecast $\hat{y}_t(\tau)$. The forecast error is $y_{t+\tau} - \hat{y}_t(\tau) = \sum_{j=0}^{\tau-1} \mathbb{A}^j e_{t+\tau-j} + [y_t(\tau) - \hat{y}_t(\tau)]$ and the variance of the forecast error can be computed once posterior estimates of $\mathbb{A}$ are available. This is easy when $\tau = 1$. For $\tau \geq 2$ only approximate expressions for the MSE are available (see e.g. Lutkepohl (1991), p. 88).

Exercise 10.17 *Show the MSE of the forecasts when $\tau = 1$.*

When a distribution of forecasts is actually needed we can exploit the fact that we can draw from $g(\alpha|y)$. We describe how "fan charts" can be obtained for case 1. prior with the obvious extension if also $\Sigma_e$ is a random variable. Let $\tilde{\mathcal{P}}\tilde{\mathcal{P}}'$ be any orthogonal factorization of $\Sigma_e$. Then, at a given $t$:

Algorithm 10.1

1) *Draw $v_a^l$ from a $\mathbb{N}(0,1)$ and set $\alpha^l = \tilde{\alpha} + \tilde{\mathcal{P}}^{-1} v_a^l$, $l = 1, \ldots L$.*

2) *Construct point forecasts $y_t^l(\tau)$, $\tau = 1, 2, , \ldots$ conditioning on $\alpha^l$.*

3) *Construct distributions at each $\tau$ using kernel methods and extract percentiles.*

Exercise 10.18 *Consider case 4. prior (i.e. a Normal prior for $\alpha$ and a Wishart prior for $\Sigma_e^{-1}$). Modify algorithm 10.1 to fit this situation.*

Algorithm 10.1 can also be used recursively, using estimates of $\tilde{\alpha}$ which are updated through the sample. The only difference is that $\tilde{\alpha}$ and $\tilde{\mathcal{P}}$ now depend on $t$.

Example 10.8 *In certain situations one wants to compute "average" forecasts at step $\tau$, i.e. may want to compute the predictive density $f(y_{t+\tau} \mid y_t) = \int f(y_{t+\tau} \mid y_t, \alpha) g(\alpha \mid y_t) d\alpha$ where $f(y_{t+\tau} \mid y_t, \alpha)$ is the conditional density of the future observation vector, given $\alpha$ and the model, and $g(\alpha \mid y_t)$ is the posterior of $\alpha$ at $t$. Given draws from algorithm 10.1 and the model then $\hat{y}_t(\tau) = L^{-1} \sum_{l=1}^{L} y_t^l(\tau)$ and its numerical variance is $L^{-1} \sum_{l=1}^{L} \sum_{j=-J(L)}^{J(L)} \mathcal{K}(j) ACF_\tau^l(j)$, where $\mathcal{K}(j)$ is a kernel and $ACF_\tau(j)$ the autocovariance of $\hat{y}_t(\tau)$ at lag $j$.*

Turning point probabilities can also be computed from the numerically constructed predictive density of future observations. For example, given $y_t^l(\tau),\ l = 1, \dots L$ we only need to check if e.g. a two quarters rule is satisfied for each draw $\alpha^l$. The fraction of draws for which the condition is satisfied is an estimate of the probability of the event at $t + \tau$.

Example 10.9 *Continuing with example 10.5, figure 10.2 presents BVAR based 68 and 95 percent bands for inflation forecasts one year ahead where we recursively update posterior estimates. The forecasting sample is 1996:1-1998:2. The bands are relatively tight reflecting very precise estimates. This precision can also be seen from the distribution of the forecasts one year ahead, constructed with data up to 1995:4. We calculate the distribution of the number of downturns that the annualized inflation rate is expected to experience over the sample 1996:1-2000:4. Downturns are identified with a two quarters rule. In the actual data there are four downturns. The median number of forecasted downturns is three. Moreover, in 90 per cent of the cases the model underpredicts the actual number of downturns and it never produces more than four downturns.*

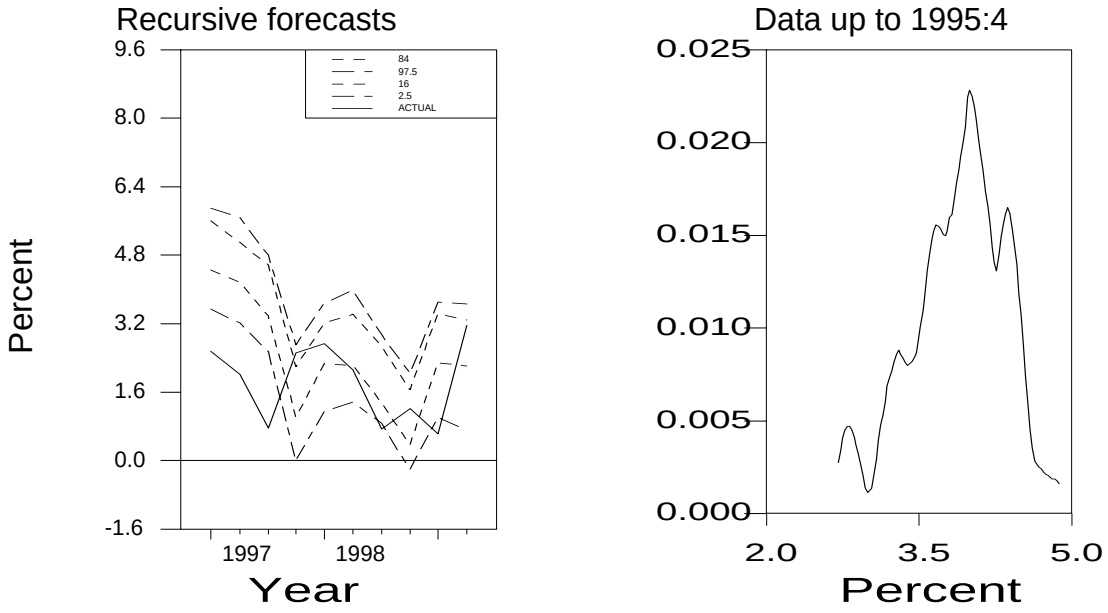


Figure 10.2: Forecasts of Italian inflation.

## 10.3 Structural BVARs

The priors we have specified in section 10.2 are designed for reduced form VAR models. What kind of priors are reasonable for structural VARs?

There are two approaches in the literature. A naive one, employed by Canova (1991), Gordon and Leeper (1994), is to use a Normal-Wishart structure for reduced form parameters $(\alpha, \Sigma_e)$. Then draws for the structural parameters are made conditional on the identification restrictions. Hence, if $\Sigma_e = \mathcal{A}_0^{-1}\mathcal{A}_0^{-1'}$, then $A_j = \mathcal{A}_0^{-1}\mathcal{A}_j$, where $A_j$ are VAR coefficients. This approach is appropriate if $\mathcal{A}_0$ is just identified since there is a unique mapping between draws of $\Sigma_e$ and draws of $\mathcal{A}_0$. When $\mathcal{A}_0$ is overidentified this method neglects the (over-identifying) restrictions. In this case, it is better to work with the structural model, and the prior suggested by Sims and Zha (1998). Consider the following structural model, where $\mathcal{A}_0$ is non singular and $\bar{y}$ only includes deterministic variables:

$$\mathcal{A}_0 y_t - \mathcal{A}(\ell) y_{t-1} + \mathcal{C}\bar{y}_t = \epsilon_t \quad \epsilon_t \sim (0, I) \tag{10.22}$$

where $\mathcal{A}(\ell) = \mathcal{A}_1\ell + \ldots \mathcal{A}_q\ell^q$. Staking the $t$ observations we have:

$$\mathsf{Y}\mathcal{A}_0 - \mathsf{X}\mathcal{A}_- = \varepsilon \tag{10.23}$$

where $\mathsf{Y}$ is a $T \times m$, $\mathsf{X}$ is a $T \times k$ matrix of lagged and exogenous variables, $k = mq + m_c$; $\varepsilon$ is a $T \times m$ matrix. Let $\mathsf{Z} = [\mathsf{Y}, -\mathsf{X}]$; $\mathcal{A} = [\mathcal{A}_0, \mathcal{A}_-]'$. The likelihood function is:

$$L(\mathcal{A}|y) \propto |\mathcal{A}_0|^T \exp\{-0.5tr(\mathsf{Z}\mathcal{A})'(\mathsf{Z}\mathcal{A})\} = |\mathcal{A}_0|^T \exp\{-0.5b'(I_{mk} \otimes \mathsf{Z}'\mathsf{Z})b\} \tag{10.24}$$

where $b = vec(\mathcal{A})$ is a $m(k+m) \times 1$ vector; $b_0 = vec(\mathcal{A}_0)$ is a $m^2 \times 1$ vector; $b_- = vec(\mathcal{A}_-)$ is a $mk \times 1$ vector and $I_{mk}$ is a $(mk \times mk)$ matrix.

Suppose $g(b) = g(b_0)g(b_-|b_0)$ where $g(b_0)$ may have singularities (due to zero identification restrictions) and let $g(b_-|b_0) \sim \mathbb{N}(\bar{h}(b_0), \bar{\Sigma}(b_0))$. The posterior is :

$$g(b|y) \propto g(b_0)|\mathcal{A}_0|^T|\Sigma(b_0)|^{-0.5} \exp\{-0.5[b'(I_{mk}\otimes\mathsf{Z}'\mathsf{Z})b\} \exp\{(b_- - \bar{h}(b_0))'\bar{\Sigma}(b_0)^{-1}(b_- - \bar{h}(b_0))\} \tag{10.25}$$

Since $b'(I_{mk}\otimes\mathsf{Z}'\mathsf{Z})b = b_0'(I_{mk}\otimes\mathsf{Y}'\mathsf{Y})b_0 + b_-'(I_{nk}\otimes\mathsf{X}'\mathsf{X})b_- - 2b_-'(I_{mk}\otimes\mathsf{X}'\mathsf{Y})b_0$, conditional on $b_0$, the quantity in the exponent in (10.25) is quadratic in $b_-$ so that $g(b_-|b_0, y) \sim \mathbb{N}(\tilde{h}(b_0), \tilde{\Sigma}(b_0))$ where $\tilde{h}(b_0) = \tilde{\Sigma}(b_0)((I_{mk} \otimes \mathsf{X}'\mathsf{Y})\hat{h}(b_0) + \bar{\Sigma}(b_0)^{-1}\bar{h}(b_0))$ and $\tilde{\Sigma}(b_0) = ((I \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(a_0)^{-1})^{-1}$. Furthermore

$$\begin{aligned} g(b_0|y) \propto\ & g(b_0)|\mathcal{A}_0|^T|(I_{mk} \otimes \mathsf{X}'\mathsf{X})\bar{\Sigma}(b_0) + I|^{-0.5} \\ & exp\{-0.5[b_0'(I_{mk} \otimes \mathsf{Y}'\mathsf{Y})b_0 + h(b_0)'\bar{\Sigma}(b_0)^{-1}h(b_0) - \tilde{h}(b_0)\tilde{\Sigma}(b_0)\tilde{h}(b_0)]\} \end{aligned} \tag{10.26}$$

Since $dim(b_-) = mk$, the calculation of $g(b_-|b_0, y)$ may be time consuming. Equation by equation computations are possible if the structural model is in SUR format, i.e. if we can run $m$ separate least square regressions with $k$ parameters each. To do this we need to choose $\bar{\Sigma}(b_0)$ appropriately. For example, if $\bar{\Sigma}(b_0) = \bar{\Sigma}_1 \otimes \bar{\Sigma}_2$ and $\bar{\Sigma}_1 \propto I$, then even if $\bar{\Sigma}_{2i} \neq \bar{\Sigma}_{2j}$, independence across equations is guaranteed since $(I_{mk} \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(b_0)^{-1} \propto (I_{mk} \otimes \mathsf{X}'\mathsf{X}) + \text{diag}\{\bar{\Sigma}_{21}, \ldots, \bar{\Sigma}_{2m}\} = \text{diag}\{\bar{\Sigma}_{21} + \mathsf{X}'\mathsf{X}, \ldots, \bar{\Sigma}_{2m} + \mathsf{X}'\mathsf{X}\}$.

Note that if we had started from a reduced form VAR (as we have done in exercise 10.9) the structure of $\tilde{\Sigma}(b_0)$ would have been $\tilde{\Sigma}(b_0) = [(\Sigma_e \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(b_0)^{-1}]^{-1}$, where $\Sigma_e$ is

the covariance matrix of the disturbances. This means that to maintain the computations simple $\bar{\Sigma}(b_0)$ must allow correlation across equations (contrary, for example, to what the Minnesota prior assumes).

It is interesting to map structural priors into Minnesota priors. Let $\mathcal{A}_0$ be given and let the VAR be $y_t = A(\ell)y_{t-1} + C\bar{y}_t + e_t$. Let $\alpha = vec[A_1, \dots A_q, C]$. Since $A(\ell) = [\mathcal{A}_-\mathcal{A}_0^{-1}]$; $E(\alpha) = [I_m, 0, \dots 0]$ and $var(\alpha) = \bar{\Sigma}_\alpha$ where $\bar{\Sigma}_\alpha$ was defined in (10.12) imply

$$E(\mathcal{A}_-|\mathcal{A}_0) = [\mathcal{A}_0, 0, \dots, 0] \tag{10.27}$$

$$var(\mathcal{A}_-|\mathcal{A}_0) = \text{diag}(b_{-(ijl)}) = \frac{\phi_0\phi_1}{h(\ell)\sigma_j^2} \quad i, j = 1, \dots m,\ \ell = 1, \dots, q \tag{10.28}$$

$$= \phi_0\phi_2 \qquad \text{otherwise} \tag{10.29}$$

where $i$ stands for equation, $j$ for variable, $\ell$ for lag, $\phi_0$ ($\phi_1$) controls the tightness of the prior variance of $\mathcal{A}_0$, ($\mathcal{A}_+$) and $\phi_2$ the tightness of the prior variance of $\mathcal{C}$.

Three features of (10.27)-(10.29) are worth mentioning: (i) there is no distinction between own and other coefficients since, in simultaneous equation models, no normalization with one right hand side variable is available; (ii) the scale factors differ from those of reduced form BVARs since $var(\epsilon_t) = I$; (iii) since $\alpha = vec[\mathcal{A}_+\mathcal{A}_0^{-1}]$ beliefs about $\alpha$ may be correlated across equations (if beliefs about $\mathcal{A}_0$ are).

As in a reduced form BVARs, stochastic linear restrictions can be added to the specification and combined with the data using the logic of Theil's mixed estimation.

Exercise 10.19 *(Controlling for trends: sum of coefficients restrictions) Suppose the average value of lagged $y_i$'s (say, $\bar{\bar{y}}_i$) is a good predictor of $y_{it}$ for equation $i$. Write this information as $\mathsf{Y}^\dagger\mathcal{A}_0 - \mathsf{X}^\dagger\mathcal{A}_- = V$ where $y^\dagger = \{y_{ij}^\dagger\} = \phi_3\bar{\bar{y}}_i$ if $i = j$ and zero otherwise, $i, j = 1, \dots m$; $x^\dagger = \{x_{i\tau}^\dagger\} = \phi_3\bar{\bar{y}}_i$ if $i = j$, for $\tau < k$ and zero otherwise, $i = 1, \dots m,\ \tau = 1, \dots k$. Construct the posterior for $b_-$ under this restriction.*

Adding the sum of coefficient restrictions introduces correlation among the coefficients of a variable in an equation. When $\phi_3 \to \infty$, the restriction implies a model in first difference, i.e. the model has $m$ unit roots and no cointegration.

Exercise 10.20 *(Controlling for seasonality: seasonal sum of coefficients restrictions). Suppose the average value of $y_{t-j}$ is good predictor of $y_t$ for each equation. Setup this restriction as a dummy observation and construct the posterior for $b_-$.*

Exercise 10.21 *(Controlling for cointegration: initial dummy restriction) Suppose we set up an initial dummy observation of the form $\mathsf{Y}^\ddagger\mathcal{A}_0 - \mathsf{X}^\ddagger\mathcal{A}_- = V$ where $y^\ddagger = \{y_j^\ddagger\} = \phi_4\bar{\bar{y}}_j$ if $j = 1, \dots M$, $x^\ddagger = \{x_\tau^\ddagger\} = \phi_4\bar{\bar{y}}_j$ if $\tau \leq k-1$ and $X^\ddagger = \phi_4$ if $\tau = k$. Construct the posterior for $b_-$ under this additional restriction.*

The prior of exercise 10.21 forces all the variables to be stationary. In fact, if $\phi_4 \to \infty$, the dummy observation becomes $[I - \mathcal{A}_0^{-1}\mathcal{A}(1)]\bar{y}_0 + \mathcal{A}_0^{-1}\mathcal{C} = 0$. If $\mathcal{C} = 0$, there is a one unit root, while if $\mathcal{C} \neq 0$ there are no unit roots.

To calculate (10.26) we need $g(b_0)$. Since for identification purposes, some elements of $b_0$ may be forced to be zero, we make a distinction between hard restrictions (those imposing identification, possibly of blocks of equations) and soft restrictions (those involving a prior on non-zero coefficients). Since little is typically known about $b_0$, a non-informative prior should be preferred i.e. $g(b_0^0) \propto 1$ where $b_0^0$ are the non-zero elements of $b_0$. In some occasions, a Normal prior may also be appropriate.

**Example 10.10** *Suppose we have $m(m-1)/2$ restrictions so that $\mathcal{A}_0$ is just identified. Assume, for example, that $\mathcal{A}_0$ is lower triangular and let $b_0^0$ be the nonzero elements of $\mathcal{A}_0$. Suppose $g(b_0^0) = \prod_i g(b_{0i}^0)$, where each $g(b_{0i}^0)$ is $\mathbb{N}(0, \sigma^2(b_0^0))$ so that the coefficients of, say, GDP and unemployment in the first equation may be related to each other but are unrelated with the coefficients of GDP and unemployment in other equations. Set, for example, $\sigma^2(b_{0ij}^0) = (\frac{\phi_5}{\sigma_i})^2$ i.e. all the elements of equation i have the same variance. Since the system is just identified one can also use a Wishart prior for $\Sigma_e^{-1}$, with $\bar{\nu}$ degrees of freedom and scale matrix $\bar{\Sigma}$ to derive a prior for $b_0^0$. Since a lower triangular $\mathcal{A}_0$ is just the Choleski factor of $\Sigma_e^{-1}$, if $\bar{\nu} = m+1$, $\bar{\Sigma} = \text{diag}\ (\frac{\phi_5}{\sigma_i})^2$, then a prior for $b_0^0$ is proportional to $\mathbb{N}(0, \sigma^2(b_0^0))$, where the factor of proportionality is the Jacobian of the transformation, i.e. $|\frac{\partial \Sigma_e^{-1}}{\partial \mathcal{A}_0}| = 2^m \prod_{j=1}^m b_{jj}^j$. Since the likelihood contains a term $|\mathcal{A}_0|^T = \prod_{j=1}^T b_{jj}^T$, ignoring the Jacobian is irrelevant if $T >> m$.*

The posterior $g(b_0|y)$ can not be computed analytically. To simulate a sequence we can use one of the algorithms we described in chapter 9. For example, one could:

**Algorithm 10.2**

1) *Calculated posterior mode $b_0^*$ of $g(b_0|y)$ and the Hessian at $b_0^*$.*

2) *Draw $b_0$ from a normal centered at $b_0^*$ with covariance equal to the Hessian at $b_0^*$ or a t-distribution with the same mean and covariance and $\nu = m+1$ degrees of freedom.*

3) *Use importance sampling to weight the draws, checking the magnitude of $IR^l = \frac{\tilde{g}(b_0^l)}{g^{IS}(b_0^l)}$, where $g^{IS}(b_0)$ is an importance density, and $l = 1, \ldots, L$.*

As alternative one could use a Metropolis-Hastings (MH) algorithm with a Normal or a t-distribution as the target, or the restricted Gibbs sampler of Waggoner and Zha (2003).

**Exercise 10.22** *Describe how to use a MH algorithm to draw a sequence from $g(b_0|y)$.*

It is immediate to extend the framework to the case where non-contemporaneous restrictions are used to identify the VAR.

**Exercise 10.23** *Suppose $\mathcal{A}_0$ is just identified using long run restrictions. How would you modify the prior for $\mathcal{A}_0$ to account for this?*

Exercise 10.24 *Suppose $\mathcal{A}_0$ is overidentified. How should the prior for $\mathcal{A}_0$ be changed?*

Exercise 10.25 *Suppose $\mathcal{A}_0$ is identified using sign restrictions. Let $\Sigma_e = \tilde{\mathcal{P}}(\omega)\tilde{\mathcal{P}}'(\omega)$, where $\omega$ is an angle. How would you modify the prior for $\mathcal{A}_0$ to take this into account? How would you modify the algorithm to draw from the posterior distribution of $\mathcal{A}_0$? (Hint: treat $\omega$ as a random variable and select an appropriate prior distribution)*

There are a number of extensions one can consider. Here we analyze two:

1. Structural VAR models with exogenous stochastic variables: e.g. oil prices in a structural VAR for domestic variables.

2. Structural VAR models with block exogenous variables and overidentifying restrictions in some block, e.g. a two-country structural model where one is block exogenous.

We assume that $y_t$ is demeaned so that $\bar{y}_t$ is omitted from the model. For the case of structural models with exogenous variables, let

$$\mathcal{A}_{i0}y_t - \mathcal{A}_i(\ell)y_{t-1} = \epsilon_{it} \quad \epsilon_{it} \sim \mathbb{N}(0, I) \tag{10.30}$$

where $i = 1, \ldots, n$ refers to the number of blocks; $m = \sum_{i=1}^{n} m_i$ with $m_i$ equations in each block; $\epsilon_{it}$ is $m_i \times 1$ for each $i$, $\mathcal{A}_i(\ell) = (\mathcal{A}_{i1}(\ell), \ldots, \mathcal{A}_{in}(\ell))$ and each $\mathcal{A}_{ij}(\ell)$ is a $m_i \times m_j$ matrix for each $\ell$. (10.30) is just the block representation of (10.22). Rewrite (10.30) as

$$y_{it} = A_i(\ell)y_{it-1} + e_{it} \tag{10.31}$$

where $A_i(\ell) = (0_{i-}, I_i, 0_{i+}) - \mathcal{A}_{i0}^{-1}\mathcal{A}_i(\ell)$; $0_{i-}$ is a matrix of zeros of dimension $m_i \times m_{i_-}$, $0_{i+}$ is a matrix of zeros of dimension $m_i \times m_{i_+}$, where $m_{i_-} = 0$ for $i = 1$ and $m_{i_-} = \sum_{j}^{i-1} m_j$ for $i = 2, \ldots, n$; $m_{i_+} = 0$ for $i = n$ and $m_{i_+} = \sum_{j=i+1}^{n} m_j$ for $i = 1, \ldots, n-1$ and where $E(e_t e_t') = \text{diag}\{\Sigma_{ii}\} = \text{diag}\{\mathcal{A}_{i0}^{-1}\mathcal{A}_{i0}^{-1\prime}\}$. Stacking the $T$ observations to have

$$\mathsf{Y}_i = \mathsf{X}_i\mathsf{A}_i + \mathsf{E}_i \tag{10.32}$$

where $\mathsf{Y}_i$ and $\mathsf{E}_i$ are $T \times m_i$ matrices, $\mathsf{X}_i$ is a $T \times k_i$ matrix and $k_i$ is the number of coefficients in each block. The likelihood function is

$$\begin{aligned} f(A_i, \Sigma_{ii}|y_T, \ldots, y_1, y_0 \ldots) &\propto \prod_{i=1}^{n} |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_i)'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_i)\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \\ &\propto \prod_{i=1}^{n} |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0} \\ &+ (\mathsf{A}_i - \mathsf{A}_{i,ols})'\mathsf{X}_i'\mathsf{X}_i(\mathsf{A}_i - \mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \end{aligned} \tag{10.33}$$

where $\mathsf{A}_{i,ols} = (\mathsf{X}_i'\mathsf{X}_i)^{-1}(\mathsf{X}_i'\mathsf{Y}_i)$ and $tr$ indicates the trace of the matrix. Suppose $g(\mathcal{A}_{i0}, \mathcal{A}_i) \propto |\mathcal{A}_{i0}|^{k_i}$. Then the posterior for $\mathcal{A}_{i0}$ and $\alpha_i = vec(\mathcal{A}_i)$ has the same form as the likelihood and

$$\begin{aligned} g(\mathcal{A}_{i0}|y) &\propto |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \quad (10.34) \\ g(\alpha_i|\mathcal{A}_{i0}, y) &\sim \mathbb{N}(\alpha_{i,ols}, (\mathcal{A}_{i0}'\mathcal{A}_{i0})^{-1} \otimes (\mathsf{X}_i'\mathsf{X}_i)^{-1}) \quad (10.35) \end{aligned}$$

where $\alpha_{i,ols} = vec(\mathsf{A}_{i,ols})$. As before, if $\mathcal{A}_{i0}$ is the Choleski factor of $\Sigma_{ii}^{-1}$ and $g(\Sigma_{ii}^{-1}) \propto |\Sigma_{ii}^{-1}|^{0.5k_i}$, then the posterior for $\Sigma_{ii}^{-1}$ has Wishart form with parameters $([(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})]^{-1}, T - m_i - 1)$. Hence, one could draw from the posterior of $\Sigma_{ii}^{-1}$ and use the Choleski restrictions to draw $\mathcal{A}_{i0}$. When $\mathcal{A}_{i0}$ is overidentified, we need to draw $\mathcal{A}_{i0}$ from the marginal posterior (10.35), which is of unknown form. To do so one could use, e.g., a version of the importance sampling algorithm 10.2.

Exercise 10.26 *Extend algorithm 10.2 to the case where the VAR has different lags in different blocks.*

Exercise 10.27 *Suppose $g(\mathcal{A}_i) \sim \mathbb{N}(\bar{\mathcal{A}}_i, \bar{\Sigma}_{\mathcal{A}})$. Show the form of $g(\alpha_i|\mathcal{A}_{i0}, y)$ in this case.*

For the case of block exogenous variables with overidentifying restrictions, suppose there are linear restrictions on $\mathcal{A}_{ij0}, \quad j > i$. This case is different from the previous case since overidentifying restrictions were placed on $\mathcal{A}_{ii0}$. Define $\mathcal{A}_i^*(\ell) = \mathcal{A}_{i0} - \mathcal{A}_i(\ell), \ i = 1, \ldots n$ and rewrite the system as $\mathcal{A}_{i0}y_t = \mathcal{A}_i^*(\ell)y_t + \epsilon_{it}$. Stacking the observations we have

$$\mathsf{Y}\mathcal{A}_{i0}' = \mathsf{X}_i\mathbb{A}_i^* + \epsilon_i \quad (10.36)$$

where $\mathsf{X}_i$ is a $T \times k_i^*$ matrix including all right hand side variables, $k_i^* = k_i - m_{i+1} - \ldots - m_n$; $\mathbb{A}_i^*$ is a $k_i^* \times m_i$ companion matrix of $\mathcal{A}_i^*(\ell)$; $\epsilon_i$ a $T \times m_i$ matrix; $Y = [Y_1, \ldots, Y_n]$ is a $T \times m$ matrix ; $\mathcal{A}_{i0} = \{\mathcal{A}_{i10}, \ldots \mathcal{A}_{in0}; \ \mathcal{A}_{ij0} = 0, j < i\}$ is a $m \times m_i$ matrix. Let $\mathsf{A}_{i,ols}^* = (\mathsf{X}_i'\mathsf{X}_i)^{-1}\mathsf{X}_i'\mathsf{Y}$ and let the prior for $(\mathcal{A}_i(0), \mathsf{A}_i^*)$ be non-informative. Letting $\alpha_i^* = vec(\mathsf{A}_i^*)$, the posteriors are:

$$\begin{aligned} g(\mathcal{A}_{i0}|y) &\propto |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols}^*)'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols}^*)\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \\ g(\alpha_i^*|\mathcal{A}_{i0}, y) &\sim \mathbb{N}(\alpha_{i,ols}^*, (I_i \otimes (\mathsf{X}_i'\mathsf{X}_i)^{-1})) \quad (10.37) \end{aligned}$$

Exercise 10.28 *Describe how to draw posterior sequences for $(\alpha_i^*, \mathcal{A}_{i0})$ from (10.37).*

We conclude with an example illustrating the techniques described in this section.

Example 10.11 *We take monthly US data from 1959:1 to 2003:1 for the log of GDP, the log of CPI, log of M2, the Federal funds rate and log of commodity prices. We are interested in the dynamic responses of the first four variables to an identified monetary policy shock and in knowing how much of the variance of output and inflation is explained by monetary policy shocks. We use contemporaneous restrictions and overidentify the system by assuming*

*that the monetary authority only looks at money when manipulating the Federal funds rate. Hence, the system has a Choleski form (in the order in which the variables are listed) except for the (3,1) entry which is set to zero. We assume* $b_0^0 \sim \mathbb{N}(0, I)$ *and use as importance sampling a Normal centered at the mode and with dispersion equal to the Hessian at the mode. We monitor the draws using the importance ratio and find that in only 11 out of 1000 draws the weight given to the draw is large.*

*The median response and the 68% band for each variable are in figure 10.3. Both output and money persistently decline in response to an interest rate increase. The response of prices is initially zero but turns positive and significant after a few quarters - a reminiscence of what is typically called the "price puzzle". Monetary shocks explain 4-18 per cent of the variance of output at the 20 quarters horizon and only 10-17 per cent of the variance of prices. One may wonder what moves prices then: it turns out that output shocks explain 45-60 per cent of the variability of prices in the sample.*

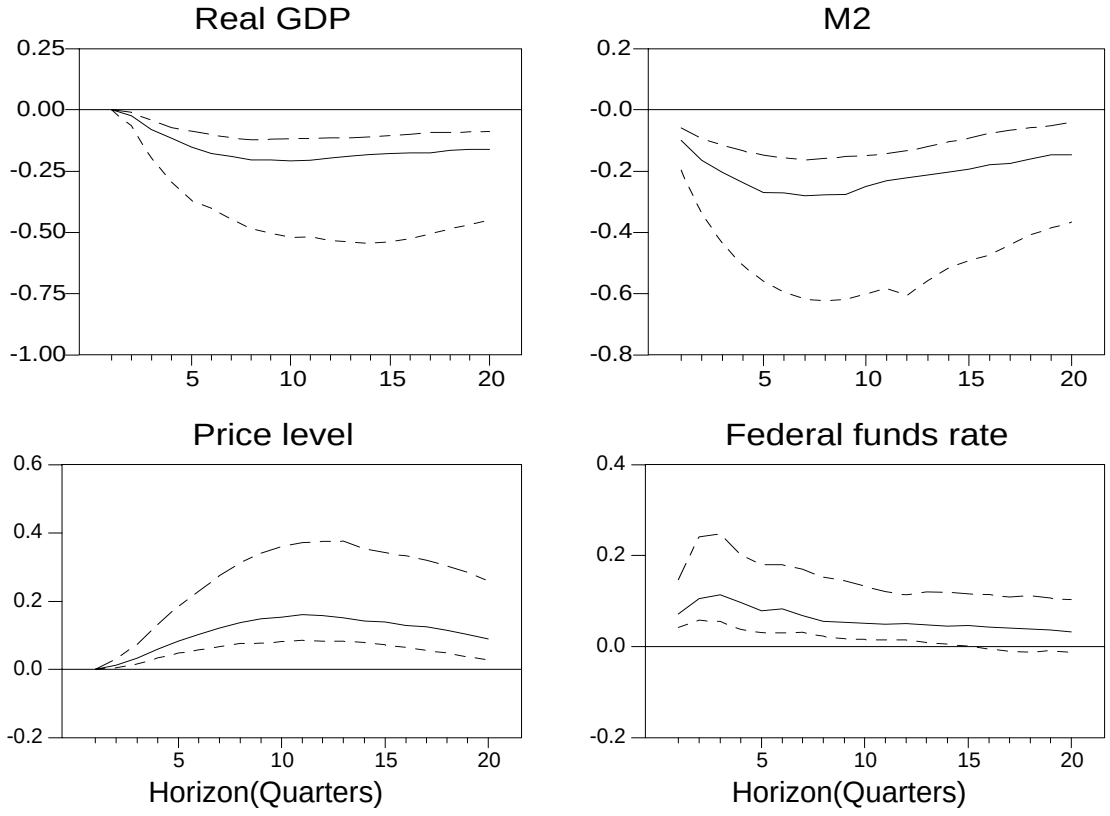


Figure 10.3: Median and 68% band for the responses to a US monetary policy shock.

## 10.4 Time Varying Coefficients BVARs

Economic time series tend to show evolving features. One could think of these changes as abrupt and model the switch as a structural break (either in the intercept, in the slope coefficients or in both). Alternatively, one may suspect that changes are related to some unobservable state, for example, the business cycle, in which case the coefficients or the covariance matrix or both could be made a function of a finite order Markov Chain (as we will do in Chapter 11). Since structural changes are rare but the coefficients tend evolve continuously one may finally prefer a model with smoothly changing coefficients. Time varying coefficient models have a long history in applied work going back, at least, to Cooley and Prescott (1973), and classical estimation methods, ranging from generalized least square (Swaamy (1970)) to Kalman filtering, are available. Here we treat the law of motion of the coefficients as the first layer of an hierarchical prior and specify, in a second layer, the distributions for the parameters of this law of motion.

The model we consider is of the form

$$y_t = A_t(\ell)y_{t-1} + C_t\bar{y}_t + e_t \qquad e_t \sim \mathbb{N}(0, \Sigma_e) \tag{10.38}$$

$$\alpha_t = \mathbb{D}_1\alpha_{t-1} + \mathbb{D}_0\bar{\alpha} + v_t \qquad v_t \sim \mathbb{N}(0, \Sigma_t) \tag{10.39}$$

where $\alpha_t = vec[A_t(\ell), C_t]$ and $\mathbb{D}_0, \mathbb{D}_1$ are $mk \times mk$ matrices. (10.39) allows for stationary and non-stationary behavior in $\alpha_t$. For example, the law of motion of the coefficients displays reversion towards the mean $\bar{\alpha}$ if the roots of $\mathbb{D}_1$ are all less than one in absolute value. In principle, $\Sigma_t$ depends on time, therefore imparting conditional heteroschedastic movements to both the coefficients and the variables of a VAR.

The specification in (10.38)-(10.39) is flexible and can generate a variety of non-linearities in the conditional moment structure. In fact, substituting (10.39) into (10.38) we have

$$y_t = (I_m \otimes \mathsf{X}_t)(\mathbb{D}_1\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}) + (I_m \otimes \mathsf{X}_t)v_t + e_t = X_t\alpha_t^{\dagger} + e_t^{\dagger} \tag{10.40}$$

where $(I_m \otimes \mathsf{X}_t)$ is the matrix of regressors. Depending on the nature of the $X_t$ and the relationship between $X_t$ and $v_t$, (10.40) encompasses several specifications used in the literature. We consider three such cases in the next example.

Example 10.12 *Suppose $m = 1$, that $X_t$ and $v_t$ are conditionally independent and that var($v_t$) $= \Sigma_v$. Then, $y_t$ is conditionally heteroschedastic with mean $X_t\alpha_t^{\dagger}$ and variance $\Sigma_e + X_t'\Sigma_v X_t$. In addition, if $X_t$ includes lagged dependent variables and a constant and $(v_t|X_t) \sim \mathbb{N}(0, \Sigma_v)$, then (10.40) generates a conditionally normal ARMA-ARCH structure. Finally, if $X_t$ includes latent variables or variables which are not perfectly predictable at t, then $y_t$ is non-Gaussian and heteroschedastic (as in Clark's (1973) mixture model).*

Exercise 10.29 *i) Suppose $m = 1$, $X_t = (X_{1t}, X_{2t})$ and assume $X_{1t}$ is correlated with $v_t$. Show that (10.40) produces a version of the bilinear model of Granger and Anderson (1978). ii) Suppose $v_t = v_{1t} + v_{2t}$, where $v_{1t}$ is independent of $X_t$ and $v_{2t}$ and has covariance matrix $\Sigma_1$, and $v_{2t}$ is perfectly correlated with $X_t$. Show that (10.38)- (10.39) can generate a model with features similar to an ARCH-M model (see Engle, Lilien and Robbins (1987)).*

(10.38)-(10.39) also include, as a special case, Hamilton's (1989) two-state shift model.

**Exercise 10.30** *Suppose* $\Delta y_t = a_0 + a_1 \varkappa_t + \Delta y_t^c$ *where* $\varkappa_t = (1-p_2) + (p_1+p_2-1)\varkappa_{t-1} + e_t^x$, $e_t^x$ *is a binomial random variable and* $\Delta y_t^c = A(\ell)\Delta y_{t-1}^c + e_t^c$. *Cast such a model into a TVC framework (Hint: Find its state space format and match coefficients with (10.38)-(10.39)).*

The model can also generate non-normalities in $y_t$. Typically, such a feature is produced when $X_t$ is a latent variable. However, even when $X_t$ includes only observable variables, $e_t$ and $v_t$ are independently distributed and $v_t$ and $X_t$ conditionally independent, (10.38)-(10.39) can generate non-normalities. To see this set $m = 1$ and define $\hat{e}_{t+\tau} = (\mathbb{D}_1^{\tau+1}\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}\sum_{j=0}^{\tau}\mathbb{D}_1{}^j)'(X_{t+\tau} - E_{t-1}X_{t+\tau}) + (\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau} - E_{t-1}(\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau}) + v_{t+\tau}'X_{t+\tau} + e_{t+\tau}$.

**Exercise 10.31** *Show that, for fixed* $t$ *and all* $\tau$, $E_{t-1}y_{t+\tau} = (\mathbb{D}_1^{\tau+1}\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}\sum_{j=0}^{\tau}\mathbb{D}_1{}^j)'$ $E_{t-1}X_{t+\tau} + E_{t-1}(\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau}$; $var_{t-1}y_{t+\tau} = E_{t-1}(\hat{e}_{t+\tau})^2$; $sk_{t-1}(y_{t+\tau}) = \frac{E_{t-1}(\hat{e}_{t+\tau})^3}{(var_{t-1}y_{t+\tau})^{\frac{3}{2}}}$; $kt_{t-1}(y_{t+\tau}) = \frac{E_{t-1}(\hat{e}_{t+\tau})^4}{(var_{t-1}y_{t+\tau})^2}$ *where* $sk_{t-1}$ *and* $kt_{t-1}$ *are the conditional skewness and kurtosis coefficients. Show that for* $\tau = 0$, $sk_{t-1}(y_t) = 0$, $kt_{t-1}(y_t) = 3$, *i.e.* $y_t$ *is conditionally normal.*

For $\tau = 1$ the conditional mean of $y_{t+1}$ is nonlinear and equal to $E_{t-1}(\alpha_{t+1}'X_{t+1}) = (\mathbb{D}_1^2\alpha_{t-1} + \mathbb{D}_0(I+\mathbb{D}_1)\bar{\alpha})'E_{t-1}X_{t+1} + E_{t-1}v_t'\mathbb{D}_1X_{t+1}$ where $E_{t-1}X_{t+1} = [E_{t-1}y_t, y_{t-1}, \ldots y_{t-\ell+1}]$, while its conditional variance is $E_{t-1}((\mathbb{D}_1{}^2\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}(1+\mathbb{D}_1))'(X_{t+1} - E_{t-1}X_{t+1}) + (v_t'\mathbb{D}_1{}'X_{t+1} - E_{t-1}v_t'\mathbb{D}_1{}'X_{t+1}) + v_{t+1}'X_{t+1} + e_{t+1})^2$. Note that $(X_{t+1} - E_{t-1}X_{t+1})' = [e_t^{\dagger}, 0, \ldots, 0]$ and that $((v_t'\mathbb{D}_1{}'X_{t+1}) - E_{t-1}(v_t'\mathbb{D}_1{}'X_{t+1}))$ involves, among other things, terms of the form $v_t'\mathbb{D}_1{}'e_t$. Hence, even when $v_t$ and $e_t$ are normal and independent, $y_{t+1}$ is conditionally non-normal because the prediction errors involve the product of normal random variables. The above argument holds for any $\tau \geq 1$.

### 10.4.1 Minnesota style prior

If (10.38) is the model for the data and (10.39) the first layer for the prior, we need to specify $\bar{\alpha}$, the evolution of $\Sigma_t$ and the form of $\mathbb{D}_1$ and $\mathbb{D}_0$. For example, we could use:

$$\mathbb{D}_1 = \phi_0\, I, \qquad \mathbb{D}_0 = I - \mathbb{D}_1 \tag{10.41}$$

$$\bar{\alpha}_{ij\ell} = 1 \qquad \text{if } i = j,\ \ell = 1 \tag{10.42}$$

$$\bar{\alpha}_{ij\ell} = 0 \qquad \text{otherwise} \tag{10.43}$$

$$\Sigma_t = \sigma_t\Sigma_0 \tag{10.44}$$

$$\Sigma_{0ij\ell} = \phi_1\frac{h_1(i,j)}{h_2(\ell)}(\frac{\sigma_j}{\sigma_i})^2 \qquad h_1(i,i) = 1 \tag{10.45}$$

$$\Sigma_{0ij\ell} = \phi_1\phi_4 \qquad \text{if exogenous} \tag{10.46}$$

where $\sigma_t = \phi_3^t + \phi_2\frac{1-\phi_3^{t-1}}{1-\phi_3}$. As in the basic Minnesota prior we assume that $\Sigma_e$ is fixed, but there is no conceptual difficulty in assuming, e.g., a Wishart prior for $\Sigma_e^{-1}$.

With (10.41) the law of motion of the coefficients has a first order autoregressive structure with decay toward the mean. $\phi_0$ controls the speed of the decay: for $\phi_0 = 0$ the coefficients are random around $\bar{\alpha}$ and for $\phi_0 = 1$ they are random walks. Higher order processes can be obtained by substituting the identity matrix in (10.41) with an appropriate matrix. The prior mean and the prior variance for the time zero coefficients are identical to those of the basic Minnesota prior except that we allow a general pattern of weights for different variables in different equations via the function $h_1(i,j)$. The variance of the innovation in the coefficients evolves linearly. The nature of time variations can be clearly understood using: $\Sigma_t = V_0\Sigma_0 + V_1\Sigma_{t-1}$, which has the same structure as the law of motion of the coefficients, and which reduces to the expression in (10.44) if $V_0 = \phi_2 \times I, V_1 = \phi_3 \times I$. For $\phi_3 = 0$ the coefficients are time varying but no heteroschedasticity is allowed, while for $\phi_2 = 0$ the variance of the coefficients is geometrically related to $\Sigma_0$. Finally, if $\phi_2 = \phi_3 = 0$, time variations and heteroschedasticity are absent.

Empirical Bayes methods can be employed to estimate the hyperparameters $\phi$ on a training sample of data going from $(-\tau, 0)$. As usual, the predictive density can be constructed and evaluated numerically using the Kalman filter.

Exercise 10.32 *Write down the predictive density for the TVC-VAR model. Specify exactly how to use the Kalman filter to numerically maximize the predictive density.*

Posterior inference can be conducted conditional on the estimates of $\phi$, i.e., we use $g(\alpha|y, \hat{\phi}_{ML-II}) \propto f(y|\alpha)g(\alpha|\hat{\phi}_{ML-II})$ in place of $g(\alpha|y)$. Note that while the full posterior averages over all possible values of $\phi$, the empirical-Bayes posterior uses ML-II estimates. Clearly, if $f(y|\phi)$ is flat in the hyperparameter space, differences will be minor.

Example 10.13 *Continuing with example 10.5, we add time variations to the coefficients of the BVAR and forecast inflation using the same style of Minnesota prior outlined above, but set $\phi_3 = 0$. We use a simplex algorithm to maximize the predictive density with respect to $\phi$'s. The optimal values are $\phi_0 = 0.98$, $\phi_1 = 0.11, \phi_2 = 0.1e-8, \phi_4 = 1000$, while $h_1(i,j) = 0.4\ \forall i,j,\ h_2(\ell) = \ell^{0.4}$. The Theil-U statistics one year ahead are 0.93 for the sample 1996:1-2000:4 and 0.89 for the sample 1991:1-1995:4 (the posterior standard error is 0.03 in both cases). Therefore, time variations in the coefficients appear to be important in forecasting Italian inflation. However, time variations in the variance hardly matter. In fact, setting $\phi_2 = 0$, the Theil-U are 0.95 and 0.90, respectively.*

Exercise 10.33 *(Ciccarelli and Rebucci) Suppose $y_{1t} = A_{11}(\ell)y_{1t-1} + y_{2t}A_{12}$ and $y_{2t} = A_{22}(\ell)y_{1t-1} + v_t$ and suppose a researcher estimates $y_{1t} = A(\ell)y_{1t-1} + e_t$.*
*i) Show that $A_{ols}(\ell)$ is biased unless $A_{22}(\ell) = 0$.*
*ii) Consider the approximating model $y_{1t} = A(\ell)y_{1t-1} + A^c(\ell)y_{1t-1} + e_t$ where $A^c(\ell) = A_{22}(\ell)A_{12}$ and $e_t = v_tA_{12}$. Clearly, the estimated model sets $A^c(\ell) = 0$, otherwise perfect collinearity would result. Suppose $\alpha = vec(A^c(\ell), A(\ell)) \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$ where $\bar{\alpha} = (0, \bar{\alpha}_2)$ and $\bar{\Sigma}_\alpha = diag[\bar{\Sigma}_{\alpha_1}, \bar{\Sigma}_{\alpha_2}]$. Show that $g(\alpha|y) \sim \mathbb{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$. Show the form of $\tilde{\alpha}, \tilde{\Sigma}_\alpha$. In particular, show that, in the formula for the posterior mean, the OLS estimator receives less weight*

*than in standard problems. Show that the posterior for $A^c(\ell)$ is centered away from zero to correct for the skewness produced by omitting a set of regressors. How would your answer change if coefficients are functions of time?*

### 10.4.2 Hierarchical prior

A BVAR with time varying coefficients is a state space model where the coefficients (variances) play the role of the unobservable states. Full hierarchical estimation of such models do not present difficulties once it is understood that time-varying and time invariant features can be jointly estimated. The Gibbs sampling is particularly useful for this purpose.

Here we consider a simple version of the model (10.38)-(10.39) and leave the discussion of a more complicated setup to a later section. The specification we employ has the form:

$$\begin{aligned} y_t &= X_t\alpha_t + e_t \qquad & e_t &\sim \mathbb{N}(0,\Sigma_e) \\ \alpha_t &= \mathbb{D}_1\alpha_{t-1} + v_t \qquad & v_t &\sim \mathbb{N}(0,\Sigma_a) \end{aligned} \tag{10.47}$$

where $X_t = (I_m \otimes \mathsf{X}_t)$. We assume that $\mathbb{D}_1$ is known and discuss in an exercise how to estimate it, in the case it is not. Posterior draws from the distribution of the unknown parameters $(\Sigma_e, \Sigma_a)$ and of the unobserved state $\{\alpha_t\}_{t=1}^T$ can be obtained with the Gibbs sampler. Let $\alpha^t = (\alpha_0, \ldots, \alpha_t)$, $y^t = (y_0, \ldots, y_t)$. To use the Gibbs sampler we need three conditional posteriors: $(\Sigma_a|y^t, \alpha^t, \Sigma_e)$, $(\Sigma_e|y^t, \alpha^t, \Sigma_a)$ and $(\alpha^t|y^t, \Sigma_e, \Sigma_a)$.

Suppose that $g(\Sigma_e^{-1}, \Sigma_a^{-1}) = g(\Sigma_e^{-1})g(\Sigma_a^{-1})$ and that each is Wishart with $\bar\nu_0$ and $\bar\nu_1$ degrees of freedom and scale matrices $\bar\Sigma_e, \bar\Sigma_a$, respectively. Then, since $e_t, v_t$ are normal

$$(\Sigma_e^{-1}|y^t, \alpha^t, \Sigma_a^{-1}) \sim \mathbb{W}(\bar\nu_0 + T, (\bar\Sigma_e^{-1} + \sum_t (y_t - X_t\alpha_t)(y_t - X_t\alpha_t)')^{-1})$$

$$(\Sigma_a^{-1}|y^t, \alpha^t, \Sigma_e^{-1}) \sim \mathbb{W}(\bar\nu_1 + T, (\bar\Sigma_a^{-1} + \sum_t (\alpha_t - \mathbb{D}_1\alpha_{t-1})(\alpha_t - \mathbb{D}_1\alpha_{t-1})')^{-1})$$

To obtain the conditional posterior of $\alpha^t$ notice that $g(\alpha^t|y^t, \Sigma_e, \Sigma_a) = g(\alpha_t|y^t, \Sigma_e, \Sigma_a)$ $g(\alpha_{t-1}|y^t, \alpha_t, \Sigma_e, \Sigma_a) \cdots g(\alpha_0|y^t, \alpha_1, \Sigma, V)$. Therefore, a sequence $\alpha^t$ can be obtained drawing each element from the corresponding conditional posterior while $\alpha_t$ is drawn from the marginal $g(\alpha_t|y^t, \Sigma_e, \Sigma_a)$. Let $\alpha_\tau^t = (\alpha_\tau, \ldots, \alpha_t)$ and $y_\tau^t = (y_\tau, \ldots, y_t)$. Then

$$\begin{aligned} g(\alpha_\tau|y^t, \alpha_{\tau+1}^t, \Sigma_e, \Sigma_a) &\propto g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a) g(\alpha_{\tau+1}|y^\tau, \alpha_\tau, \Sigma_e, \Sigma_a) \\ &\times f(y_{\tau+1}^t, \alpha_{\tau+1}^t|y_\tau, \alpha_\tau, \alpha_{\tau+1}, \Sigma_e, \Sigma_a) \\ &= g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a) g(\alpha_{\tau+1}|\alpha_\tau, \Sigma_e, \Sigma_a) \end{aligned} \tag{10.48}$$

The first two terms involve posterior distributions obtained with data up to $\tau$ and the last term the distribution of the data and the coefficients from $\tau+1$ until $t$. The last line follows from the fact that $\alpha_\tau$ is independent of $y_{\tau+1}^t, \alpha_{\tau+1}^t$, conditional on $(y^\tau, \Sigma_e, \Sigma_a)$. It is immediate to recognize that the two densities in (10.48) can be computed from the smoothing and the predictive equations of the Kalman filter (see chapter 6). Let $\alpha_{t|t} \equiv E(\alpha_t|y^t, \Sigma_e, \Sigma_a) = \alpha_{t|t-1} + \mathsf{K}_t(y_t - X_t\alpha_{t|t-1})$; $\Sigma_{t|t} \equiv var(\alpha_t|y^t, \Sigma_a, \Sigma_e) = (I - \mathsf{K}_t X_t)\Sigma_{t|t-1}$ where $\alpha_{t|t-1} =$

$\mathbb{D}_1\alpha_{t-1|t-1}$, $\mathsf{K}_t = \Sigma_{t|t-1}X_t'(X_t\Sigma_{t|t-1}X_t' + \Sigma_e)^{-1}$, and $\Sigma_{t|t-1} \equiv var(\alpha_t|y^{t-1}, \Sigma_e, \Sigma_a) = \mathbb{D}_1\Sigma_{t-1|t-1}\mathbb{D}_1' + \Sigma_a$. Using the linearity of the model and the Gaussian structure of (10.47), $g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a)$ is normal with mean $\alpha_{\tau|\tau}$ and variance $\Sigma_{\tau|\tau}$, while $g(\alpha_{\tau+1}|y^\tau, \alpha_\tau, \Sigma_e, \Sigma_a)$ is normal with mean $\mathbb{D}_1\alpha_\tau$ and variance $\Sigma_a$. Therefore, given a prior for $\alpha_0$, all conditional densities are Gaussian and to keep track of these distributions we only need to update conditional means and variances. Hence, to draw samples from $g(\alpha^t|y^t, \Sigma, \Sigma_a)$ we use the following:

### Algorithm 10.3

1) *Run the Kalman filter, save* $\alpha_{t|t}$, $\Sigma_t = \Sigma_{t|t} - \mathbb{M}_t\Sigma_{t+1|t}\mathbb{M}_t'$, *and* $\mathbb{M}_t = \Sigma_{t|t}\Sigma_{t+1|t}^{-1}$.

2) *Draw* $\alpha_t^l \sim \mathbb{N}(\alpha_{t|t}, \Sigma_{t|t})$, $\alpha_{t-j}^l \sim \mathbb{N}(\alpha_{t-j|t-j} + \mathbb{M}_{t-j}(\alpha_{t-j+1}^l - \alpha_{t-j|t-j}), \Sigma_{t-j})$, $j \geq 1$.

3) *Repeat* $l = 1, \ldots L$ *times*

It is straightforward to allow for an unknown $\mathbb{D}_1$ and a time-varying $\Sigma_a$.

**Exercise 10.34** *Assume that* $\mathbb{D}_1$ *is unknown and assume a normal prior on its nonzero elements i.e.* $\mathbb{D}_1^0 \sim \mathbb{N}(\bar{\mathbb{D}}_1, \bar{\sigma}_{D_1}^2)$. *Show that* $g(\mathbb{D}_1^0|\alpha^t, y^t, \Sigma_e, \Sigma_a) \sim \mathbb{N}((\alpha_{t-1}'\Sigma_a^{-1}\alpha_{t-1} + \sigma_{D_1}^{-2})^{-1}$ $(\alpha_{t-1}'\Sigma_a^{-1}\alpha_t + \sigma_{D_1}^{-2}\bar{\mathbb{D}}_1); (\alpha_{t-1}'\Sigma_a^{-1}\alpha_{t-1} + \sigma_{D_1}^{-2})^{-1})$.

**Exercise 10.35** *Let* $\Sigma_{at} = \sigma_t\Sigma_a$. *How would you construct the conditional posterior distribution for* $\Sigma_{at}$? *(Hint: treat* $\sigma_t$ *as a parameter and assume a conjugate prior).*

The next extension is useful to compute the likelihood of DSGE models which are not linearized around the steady state.

**Exercise 10.36** *(Non-linear state space models) Consider the state space model:*

$$\begin{aligned} y_t &= f_{1t}(\alpha_t) + e_t \qquad & e_t \sim \mathbb{N}(0, \Sigma_e) \\ \alpha_t &= f_{2t}(\alpha_{t-1}) + v_t \qquad & v_t \sim \mathbb{N}(0, \Sigma_a) \end{aligned} \tag{10.49}$$

*where* $f_{1t}$ *and* $f_{2t}$ *are given but perhaps depend on unknown parameters. Show that* $(\alpha_t|\alpha_{j\neq t}, \Sigma_e, \Sigma_a, y^t) \propto \mathsf{h}_1(\alpha_t)\mathsf{h}_2(\alpha_t)\mathbb{N}(f_{2t}(\alpha_{t-1}), \Sigma_a)$ *where* $\mathsf{h}_1(\alpha_t) = \exp\{-0.5(\alpha_{t+1} - f_{2t}(\alpha_t))'$ $\Sigma_a^{-1}(\alpha_{t+1} - f_{2t}(\alpha_t))\}$; $\mathsf{h}_2(\alpha_t) = \exp\{-0.5(y_t - f_{1t}(\alpha_t))'\Sigma_e^{-1}(y_t - f_{1t}(\alpha_t))\}$. *Describe how to use an acceptance sampling algorithm to draw from this posterior distribution.*

Finally, we consider the case of non-normal errors. While for macroeconomic data the assumption of normality is, by and large, appropriate, for robustness purposes it may be useful to allow for non-normalities. As noted, the conditional moments of (10.47) are nonlinear for $\tau \geq 1$. To generate non-normalities, when $\tau = 0$, it is sufficient to add a nuisance parameter $\phi_5$ to the variance of the error term, i.e., $(\alpha_t|\alpha_{t-1}, \phi_5, \Sigma_a) \sim \mathbb{N}(\mathbb{D}_1\alpha_{t-1}, \phi_5\Sigma_a)$ where $g(\phi_5)$ is chosen to mimic a distribution of interest. For example, suppose that $\phi_5$ is exponentially distributed with mean equal to 2. Since $g(\alpha_t|\alpha_{t-1}, \Sigma_a, \phi_5)$ is normal with mean $\mathbb{D}_1\alpha_{t-1}$ and

variance $\phi_5\Sigma_a$; $g(\phi_5|y^t,\alpha^t,\Sigma_a) \propto \sqrt{\frac{1}{\phi_5}}\exp\{-0.5[\phi_5+(\alpha_t-\mathbb{D}_1\alpha_{t-1})'\phi_5^{-1}\Sigma_a^{-1}(\alpha_t-\mathbb{D}_1\alpha_{t-1})]\}$ which is the kernel of the generalized inverse Gaussian distribution. A similar approach can be used to model non-normalities in the measurement equation.

Exercise 10.37 *Suppose* $(y_t|\alpha_t,x_t,\phi_6,\Sigma_e) \sim \mathbb{N}(x_t\alpha_t,\phi_6\Sigma_e)$ *and that* $g(\phi_6)$ *is* $exp(2)$. *Show the form of the conditional posterior for* $\phi_6$. *Describe how to draw sequences for* $\phi_6$.

Exercise 10.38 *Let* $y_t = x_t\alpha_t,\ t=1,\dots T$ *where conditional on* $x_t$ $\alpha_t' = (\alpha_{1t},\dots\alpha_{kt})$*is iid with mean* $\bar{\alpha}$ *and variance* $\bar{\Sigma}_\alpha$, $|\bar{\Sigma}_\alpha| \neq 0$. *Assume that* $\bar{\alpha}$ *and* $\bar{\Sigma}_\alpha$ *are known and let* $\alpha = (\alpha_1,\dots,\alpha_t)$.
*i) Show that the minimum MSE estimator of* $\alpha$ *is* $\tilde{\alpha} = (I_T\otimes\bar{\Sigma}_\alpha)x'\Omega^{-1}y + (I_{Tk} - (I_T\otimes\bar{\Sigma}_\alpha)x'\Omega^{-1}x)(\mathbf{1}\otimes\bar{\alpha})$ *where* $\Omega = x(I_T\otimes\Sigma_\alpha)x'$,$x = diag(x_1',\dots x_t')$ *and* $\mathbf{1} = [1,\dots,1]'$.
*ii) Show that if* $\bar{\alpha} = \alpha_0 + v_a$, $v_a \sim (0,\Sigma_{\bar{a}})$ *and* $\Sigma_{\bar{a}}$ *is known, the best minimum MSE estimator of* $\bar{\alpha}$ *equals* $(x'\Omega^{-1}x + \Sigma_{\bar{a}}^{-1})^{-1}(x'\Omega^{-1}y + \Sigma_{\bar{a}}^{-1}\alpha_0)$. *Show that as* $\Sigma_{\bar{a}} \to \infty$ *the optimal MSE estimator is the GLS estimator.*

Exercise 10.39 *(Cooley and Prescott) Let* $y_t = x_t\alpha_t$ *where* $x_t$ *is a* $1\times k$ *vector;* $\alpha_t = \alpha_t^P + e_t$; $\alpha_t^P = \alpha_{t-1}^P + v_t$ *where* $e_t \sim (0,(1-\varrho)\sigma^2\Sigma_e)$, $v_t \sim (0,\varrho\sigma^2\Sigma_v)$ *and assume* $\Sigma_e$, $\Sigma_v$ *known. Here* $\varrho$ *represents the speed of adjustment of* $\alpha_t$ *to structural changes (for* $\varrho \to 1$ *permanent changes are large relative to transitory ones). Let* $y = [y_1,\dots,y_T]'$, $x = [x_1,\dots,x_T]'$ *and* $\alpha^p = (\alpha_{1t}^p,\dots,\alpha_{kt}^p)'$.
*i) Show that the model is equivalent to* $y_t = x_t'\alpha_t^P + \epsilon_t$; $\epsilon_t \sim (0,\sigma^2\Omega(\varrho))$. *Display* $\Omega(\varrho)$.
*ii) Show that, conditional on* $\varrho$, *the minimum MSE estimators for* $(\alpha^p,\sigma^2)$ *are* $\alpha_{ML}^p(\varrho) = (x'\Omega(\varrho)^{-1}x)^{-1}\ (x'\Omega(\varrho)^{-1}y)$ *and* $\sigma_{ML}^2(\varrho) = \frac{1}{T}(y-x\alpha_{ML}^p(\varrho))'\Omega(\varrho)^{-1}(y-x\alpha_{ML}^p(\varrho))$. *Describe a way to maximize the concentrated likelihood as a function of* $\varrho$.
*iii) Obtain posterior estimators for* $(\alpha,\varrho,\sigma^2)$ *when* $g(\alpha,\varrho,\sigma^2)$ *is non-informative. Set up a Gibbs sampler algorithm to compute the joint posterior of the three parameters.*

## 10.5 Panel VAR models

We have extensively discussed macro panel data in chapter 8. Therefore, the focus of this section is narrow. Our attention centers on three problems. First, how to specify Bayesian univariate dynamic panels. Second, how to dynamically group units in the cross section. Third, how to setup panel VAR models with cross sectional interdependencies. Univariate dynamic panels emerge, for example, when estimating steady state income per-capita, or when examining the short and long run effects of oil shocks on output across countries. Grouping is particularly useful, for example, if one is interested in knowing if there are countries which react differently than others after e.g. financial crises. Finally, models with interdependencies are useful to study a variety of transmission issues across countries or sectors which can not be dealt with the models of chapter 8.

### 10.5.1 Univariate dynamic panels

For $i = 1, \dots n$, the model we consider is:

$$y_{it} = A_{1i}(\ell) y_{it-1} + \bar{y}_i + A_{2i}(\ell) Y_t + e_{it} \quad e_{it} \sim (0, \sigma_i^2) \tag{10.50}$$

where $A_{ji}(\ell) = A_{ji1}\ell + \ldots + A_{jiq_j}\ell^{q_j}$ $j = 1, 2$ and $\bar{y}_i$ is the unit specific fixed effect. Here $Y_t$ includes variables which account for cross sectional interdependencies. For example, if $y_{it}$ are regional sales, one element of $Y_t$ could be a national business cycle indicator. Because variables like $Y_t$ are included, $E(e_{it}e_{j\tau}) = 0 \; \forall i \neq j$, all $t, \tau$. We can calculate a number of statistics from (10.50). For example, long run multipliers to shocks are $(1 - A_{1i}(1))^{-1}$ and long run multipliers to changes in $Y_t$ are $(1 - A_{1i}(1))^{-1} A_{2i}(1)$.

Example 10.14 *Let $y_{it}$ be output in Latin American country $i$ and let $Y_t = (x_{1t}, i_t)$, where $i_t$ is US interest rate. Suppose $i_t = A_3(\ell)\epsilon_t$. Then $(1 - A_{1i}(\ell)^{-1} A_{2i}(\ell) A_3(\ell)$ traces out the effect of unitary US interest rate shock at $t$ on the output of country $i$ from $t$ on.*

Stacking the T observations for $(y_{it}, Y_t, e_{it})$ and the fixed effect into the vectors $(y_i, Y, e_i, \mathbf{1})$, letting $\mathsf{X}_i = (y_i, Y, \mathbf{1})$, $\Sigma_i = \sigma_i^2 \times I_T$, $\alpha = [A_1, \ldots A_n]'$, $A_i = (A_{1i1}, \ldots, A_{iq_1}, \bar{y}_i, A_{1i1}, \ldots, A_{2iq_2})$ and setting $y = (y_1, \ldots y_n)', e = (e_1, \ldots, e_n)'$:

$$y = (I_n \otimes \mathsf{X}_i)\alpha + e \quad e \sim (0, \Sigma_i \otimes I_n) \tag{10.51}$$

Clearly, (10.51) has the same format as a VAR, except that $\mathsf{X}_i$ are unit specific and the covariance matrix of the shocks has a diagonal heteroschedastic structure. The first feature is due to the fact that we do not allow for interdependencies across units. The latter is easy to deal with once (10.51) is transformed so that the innovations have spherical disturbances.

If $e$ is normal, the likelihood function of a univariate dynamic panel is therefore the product of a normal for $\alpha$, conditional on $\Sigma_i \otimes I_n$, and $n$ Gamma densities for $\Sigma_i^{-1}$. Since the variance of $e$ is diagonal, $\alpha_{ML}$ can be obtained equation by equation.

Exercise 10.40 *Show that $\alpha_{ML}$ obtained from (10.51) is the same as the estimator obtained by stacking weighted least square estimators obtained from (10.50) for each $i$.*

Conjugate priors for dynamic panels are similar to those described in section 10.2. Since $var(e)$ is diagonal, we can choose $\sigma_i^{-2} \sim \mathbb{G}(a_1, a_2)$, each $i$. Given the panel framework we can use the exchangeability assumption if, a-priori, we expect the $A_i$ to be similar across units. An exchangeable prior on $A_i$ takes the form $A_i \sim \mathbb{N}(\bar{A}, \bar{\sigma}_A^2)$ where $\bar{\sigma}_A^2$ measures the degree of heterogeneity an investigator expects to find in the cross section.

Exercise 10.41 *(Lindlay and Smith) Suppose the model (10.50) has k coefficients in each equation and that $A_i = \bar{A} + v_i$, $i = 1, \ldots n$, $v_i \sim \mathbb{N}(0, \bar{\sigma}_A^2)$, where $\bar{A}, \bar{\sigma}_A^2$ are known. Show the form of the posterior mean for $A_i$. Assuming that $\sigma_i^2$ is fixed, show the form of the posterior variance for $A_i$. Argue that the posterior mean for the stacked vector of $A_i$ is the same as the one obtained by calculating the posterior mean for the system (10.51).*

Exercise 10.41 highlights the importance of exchangeable priors in a model like (10.51). In fact, exchangeability preserves independence across equations and the posterior mean of the coefficients of a dynamic panel can be computed equation by equation.

Exercise 10.42 *(Canova and Marcet) Suppose you want to set up an exchangeable prior on the difference of the coefficients across equations, i.e.* $\alpha_i - \alpha_j \sim \mathbb{N}(0, \Sigma_a)$. *This is advantageous since there is no need to specify the prior mean* $\bar{\alpha}$. *Show the structure of* $\Sigma_a$ *which insures that the ordering of the units in the cross section does not matter.*

We already mentioned the pooling dilemma in section 4 of Chapter 8. We return to this problem in the next exercise which gives conditions under which the posterior distribution for $A_i$ reflects prior, pooled and/or single unit sample information.

Exercise 10.43 *(Zellner and Hong) Let* $y_i = x_i\alpha_i + e_i, \;\; i = 1, \dots n$ *where* $x_i$ *may include lags of* $y_{it}$ *and for each* $i$, $y_i$ *is a* $T \times 1$ *vector,* $x_i$ *a* $T \times k$ *vector and* $\alpha_i$ *a* $k \times 1$ *vector and* $e_i \sim$ iid $\mathbb{N}(0, \sigma_e^2)$. *Assume that* $\alpha_i = \bar{\alpha} + v_i$, *where* $v_i \sim$ iid $\mathbb{N}(0, \kappa^{-1}\sigma_v^2 I_k)$ *with* $0 < \kappa \leq \infty$.
*(i) Show that a conditional point estimate for* $\alpha = (\alpha_1', \dots, \alpha_N')'$ *is the* $Nk \times 1$ *vector* $\tilde{\alpha} = (x'x + \kappa I_{nk})^{-1}(x'x\alpha_{ols} + \kappa \mathbb{I}\alpha_p)$ *where* $x = \text{blockdiag}\{x_i\}$; $\alpha_{ols} = (x'x)^{-1}(x'y)$; $y = (y_1', \dots, y_N')'$, $\alpha_{ols} = (\alpha_{1,ols}', \dots \alpha_{N,ols}')'$, $\alpha_{i,ols} = (x_i'x_i)^{-1}(x_i'y_i)$, $\mathbb{I} = (I_k, \dots I_k)$, $\alpha_p = (\sum_i x_i'x_i)^{-1} (\sum_i x_i'x_i\alpha_{i,ols})$. *Conclude that* $\tilde{\alpha}$ *is a weighted average of individual OLS estimates and of the pooled estimate* $\alpha_p$. *Show that, as* $\kappa \to \infty$, $\tilde{\alpha} = \alpha_p$.
*(ii) (g-prior) Assume that* $v_i \sim$ iid $\mathbb{N}(0, (x_i'x_i)^{-1}\sigma_v^2)$. *Show that* $\tilde{\alpha}_i^1 = (\alpha_{i,ols} + \frac{\sigma_e^2}{\sigma_v^2}\bar{\alpha})/(1 + \frac{\sigma_e^2}{\sigma_v^2})$ *Conclude that* $\tilde{\alpha}_i^1$ *is a weighted average of the OLS estimate and the prior mean* $\bar{\alpha}$.
*(iii) Show that if* $g(\bar{\alpha})$ *is non-informative,* $\tilde{\alpha}_i^2 = (\alpha_{i,ols} + \frac{\sigma_e^2}{\sigma_v^2}\alpha_p)/(1 + \frac{\sigma_e^2}{\sigma_v^2})$. *Conclude that, as* $\frac{\sigma_e^2}{\sigma_v^2} \to \infty$, $\tilde{\alpha}_i = \alpha_p$ *and, as* $\frac{\sigma_e^2}{\sigma_v^2} \to 0$, $\tilde{\alpha}_i = \alpha_{i,ols}$.

Next we describe how dynamic univariate panels can be used to estimate the steady state distribution of income per-capita and of the convergence rates in a panel of EU regions.

Example 10.15 *Here* $A_{1i}(\ell)$ *has only one non-zero element (the first one),* $Y_t$ *is the average EU GDP per-capita and* $A_{2ij} = 1$ *if* $j = 0$ *and zero otherwise. Hence (10.50) is:*

$$\ln(\frac{y_{it}}{Y_t}) = \bar{y}_i + A_i \ln(\frac{y_{it-1}}{Y_{t-1}}) + e_{it} \qquad e_{it} \sim \mathbb{N}(0, \sigma_i^2) \tag{10.52}$$

*We let* $\alpha_i = (\bar{y}_i, A_i)$ *and assume* $\alpha_i = \bar{\alpha} + v_i$, *where* $v_i \sim \mathbb{N}(0, \sigma_a^2 I)$.

*We treat* $\sigma_i^2$ *as known (and estimate it from invidual OLS regressions), assume* $\bar{\alpha}$ *known (estimated averaging individual OLS estimates) and treat* $\bar{\sigma}_a^2$ *as fixed. Let* $\frac{\sigma_i^2}{\sigma_{a_j}^2}$, $j = 1, 2$ *measures the relative importance of prior and sample information: if this ratio goes to infinity sample information does not matter; viceversa, if it is close to zero, prior information is irrelevant. We choose a relative loose prior (*$\frac{\sigma_i^2}{\sigma_{a_j}^2} = 0.5, j = 1, 2$*). Using income per-capita*

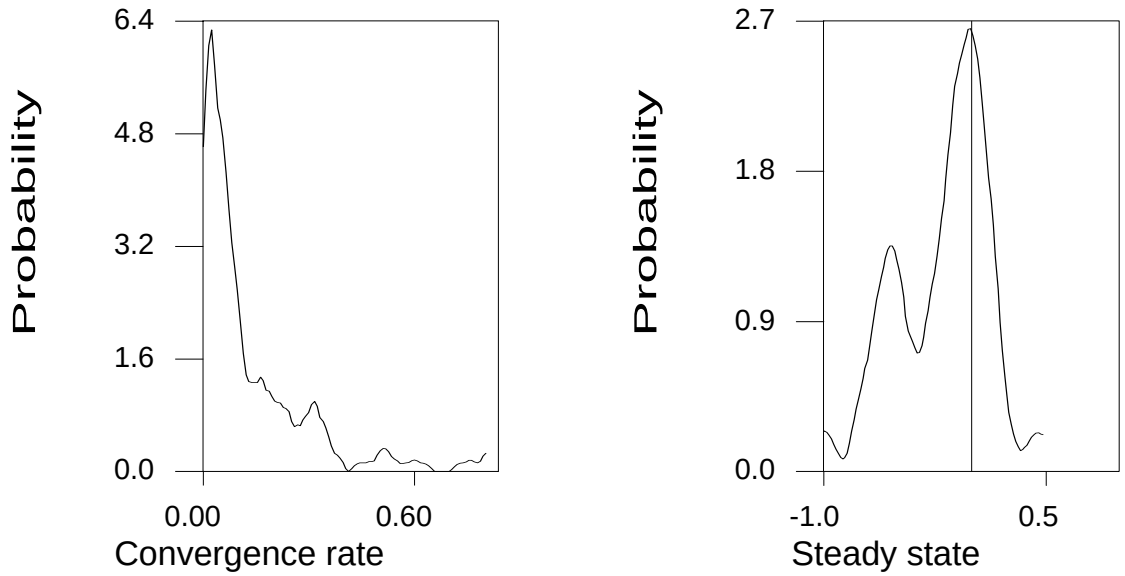


Figure 10.4: Cross sectional distributions.

*for 144 EU regions from 1980 to 1996 we calculate the relative steady state for unit i using* $\tilde{SS}_i = \tilde{\bar{y}}_i \frac{1-\tilde{A}_i^T}{1-\tilde{A}_i} + \tilde{A}_i^{T+1} \frac{y_{i0}}{Y_0}$ *where* $\tilde{\bar{y}}_i, \tilde{A}_i$ *are posterior mean estimates. The rate of convergence to the steady state is* $\tilde{CV}_i = 1 - \tilde{A}_i$ *(If* $\tilde{A} > 1$*, we set* $\tilde{CV} = 0$*). We plot the cross sectional distribution of* $\tilde{CV}$ *and* $\tilde{SS}$ *in figure 10.4. The mode of the convergence rate is 0.09, implying much faster catch up than the literature has found (see e.g. Barro and Sala (1995)). The highest 95% credible set is however large (it goes from 0.03 to 0.55). The cross sectional distribution of relative steady states has at least two modes: one at low relative levels of income and one just below the EU average.*

At times, when the panel is short, one wishes to use cross-sectional information to get better estimates of the parameters of each unit. In other cases, one is interested in estimating the average cross sectional effect. In both situations, the tools of Meta analysis come handy.

Example 10.16 *Continuing with example 10.15, suppose* $g(SS_i) \sim \mathbb{N}(\bar{SS}, \sigma_{SS}^2)$ *where* $\sigma_{SS} = 0.4$ *and assume* $g(\bar{SS}) \propto 1$*. Using the logic of hierarchical models,* $g(\bar{SS}|y)$ *combines prior and data information and* $g(SS_i|y)$ *combines unit specific and pooled information. The posterior mean for* $\bar{SS}$ *is -0.14 indicating that the distribution is highly skewed to the left, the variance is 0.083 and a credible 95 percent interval is (-0.30, 0.02). Since a credible 95 percent posterior interval for* $SS_i$ *is (-0.51, 0.19), this posterior distribution largely overlaps with the one in figure 10.4.*

### 10.5.2 Endogenous grouping

There are many situations when one would like to know whether there are groups in the cross section of a dynamic panel. For example, one type of growth theory predicts the existence of convergence clubs, where clubs are defined by similarities in the features of the various economies or government policies. In monetary economics, one is typically interested in knowing whether regional economies respond differently to union wide monetary policy disturbances or whether the behavioral responses of certain groups of agents (credit constrained vs. credit unconstrained consumers, large vs. small firms, etc.) can be identified. In general, these classifications are exogenously chosen (see for example, Gertler and Gilchrist (1991)) and somewhat arbitrary.

In this subsection we describe a procedure which simultaneously allows for endogenous grouping of cross sectional units and for Bayesian estimation of the parameters of the model. The basic idea is simple: if units $i$ and $i'$ belong to a group, the vector of coefficients will have the same mean and the same dispersion but if the don't, the vector of coefficients of the two units will have different moments.

Let n be the size of the cross section, T the size of the time series, and $\mathcal{O} = 1, 2, \ldots n!$ the ordering of the units of the cross section (the ordering producing a group is unknown). We assume there could be $\psi = 1, 2, \ldots, \bar{\psi}$ break points, $\bar{\psi}$ given. For each group $j = 1, \ldots, \psi+1$ and each unit $i = 1, \ldots, n^j(\mathcal{O})$

$$y_{it} = \bar{y}_i + A_{1i}(\ell) y_{it-1} + A_{2i}(\ell) Y_{t-1} + e_{it} \quad e_{it} \sim (0, \sigma^2_{e_i}) \tag{10.53}$$

$$\alpha_i^j = \bar{\alpha}^j + v_i^j \qquad v_i^j \sim (0, \bar{\Sigma}_j) \tag{10.54}$$

where $\alpha_i = [\bar{y}_i, A_{1i1}, \ldots, A_{1iq_1}, A_{2i1}, \ldots, , A_{2iq_2}]'$ is the $k_i \times 1$ vector of coefficients of unit $i, k_i = q_1 + q_2 + 1$, $n^j(\mathcal{O})$ is the number of units in group $j$, given the $\mathcal{O}$-th ordering, $\sum_j n^j(\mathcal{O}) = n$, for each $\mathcal{O}$. In (10.54), $\alpha_i$ is random but the coefficients of the $n^j(\mathcal{O})$ units belonging to group $j$ have the same mean and same covariance matrix. Since the exchangeable structure may differ across groups, (10.53)-(10.54) capture the idea that there may be clustering of units within groups but that groups may drift apart.

The alternative to (10.53)-(10.54) is a model with homogeneous dynamics in the cross section, that is $\bar{\psi} = 0$, and an exchangeable structure for all units of the cross section, i.e.

$$\alpha_i = \bar{\alpha} + v_i \qquad i = 1, \ldots, n \quad v_i \sim (0, \bar{\Sigma}_i) \tag{10.55}$$

Let $Y$ be a $(nTm) \times 1$ the vector of left hand side variables in (10.53) ordered to have the $n$ cross sections for each $t = 1, \ldots T$, $m$ times, $X$ be a $(nTm) \times (nk)$ matrix of the regressors, $\alpha$ be a $(nk) \times 1$ vector of coefficients, $E$ a $(nTm) \times 1$ vector of disturbances, $\bar{\alpha}$ a $(\psi+1)k \times 1$ vector of means of $\alpha$, $A$ be a $(nk) \times (\psi+1)k$ matrix, $A = diag\{A_j\}$, where $A_j$ has the form $\mathbf{1} \otimes I_k$ where $I_k$ is a $k \times k$ identity matrix and $\mathbf{1}$ is a $n^j(\mathcal{O}) \times 1$ vector of ones. Given an ordering $\mathcal{O}$, the number of groups $\psi$, and the location of the break point $h^j(\mathcal{O})$, we can rewrite (10.53) – (10.54) as:

$$Y = X\alpha + E \qquad E \sim (0, \Sigma_E) \tag{10.56}$$

$$\alpha = \Xi\bar{\alpha} + V \qquad V \sim (0, \Sigma_V) \tag{10.57}$$

where $\Sigma_E$ is $(nTm) \times (nTm)$ and $\Sigma_V = diag\{\Sigma_i\}$ is a $(nk) \times (nk)$ matrix and $\Xi$ is a matrix of zeros and ones. To complete the specification we need priors for $(\bar{\alpha}, \Sigma_E, \Sigma_V)$ and for the submodel characteristics $\mathcal{M}$, indexed by $(\mathcal{O},\ \psi,\ h^j(\mathcal{O}))$. Since the calculation of the posterior distribution is complicated, we take an Empirical Bayes approach.

The approach to group units proceeds in three steps. Given $(\bar{\alpha},\ \Sigma_E,\ \Sigma_V, \mathcal{O})$, we examine how many groups are present. Given $\mathcal{O}$ and $\hat{\psi}$, we check for the location of the break points. Finally we iterate on the first two steps, altering $\mathcal{O}$. The selected submodel is the one that maximizes the predictive density over orderings $\mathcal{O}$, groups $\psi$, and break points $h^j(\mathcal{O})$.

Let $f(Y|H_0)$ be the predictive density of the data under cross sectional homogeneity. Furthermore, let $I^\psi$ be the set of possible break points when there are $\psi$ groups. Let $f(Y^j|H_\psi, h^j(\mathcal{O}), \mathcal{O})$ be the predictive density for group $j$, under the assumption that there are $\psi$ break points with location $h^j(\mathcal{O})$, using ordering $\mathcal{O}$ and let $f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O}) = \prod_{j=1}^{\psi+1} f(Y^j|H_\psi, h^j(\mathcal{O}), \mathcal{O})$. Define the quantities

- $f^-(Y|H_\psi, \mathcal{O}) \equiv \sup_{h^j(\mathcal{O}) \in I^\psi} f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O})$,
- $f^\dagger(Y|H_\psi) \equiv \sup_{\mathcal{O}} f^-(Y|H_\psi, \mathcal{O})$,
- $f^0(Y|H_\psi, \mathcal{O}) \equiv \sum_{h^j(\mathcal{O}) \in I^\psi} g_i^j(\mathcal{O}) f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O})$,

where $g_i^j(\mathcal{O})$ is the prior probability that there is a break at location $h^j(\mathcal{O})$ for group $j$ of ordering $\mathcal{O}$. $f^-$ gives the maximized predictive density with respect to the location of break points, for each $\psi$ and $\mathcal{O}$; $f^\dagger$ the maximized predictive density, for each $\psi$, once the location of the break point and the ordering of the data are chosen optimally. $f^0$ gives the average predictive density with $\psi$ breaks where the average is calculated over all possible locations of the break points, using the prior probability that there is a break point in each location as weight. We choose $g_i^j(\mathcal{O})$ to be uniform over each $(j, \mathcal{O})$ and set $\bar{\psi} << \sqrt{(N/2)}$.

Examining the hypothesis that the dynamics of the cross section are group-based, given $\mathcal{O}$, is equivalent to verifying the hypothesis that there are $\psi$ breaks against the null of no breaks. Such an hypothesis can be examined with a Posterior odds ratio:

$$PO(\mathcal{O}) \quad = \quad \frac{g_0 f(Y|H_0)}{\sum_\psi g_\psi f^0(Y|H_\psi, \mathcal{O}) \mathbb{J}_1(n)} \tag{10.58}$$

where $g_0$ ($g_\psi$) is the prior probability that there are 0 ($\psi$) breaks. Verification of the hypothesis that there are $\psi - 1$ vs. $\psi$ breaks in the cross section can be done using:

$$PO(\mathcal{O}, \psi - 1) \quad = \quad \frac{g_{\psi-1} f^{0(\psi-1)}(Y|H_{\psi-1}, \mathcal{O})}{g_\psi f^{0(\psi)}(Y|H_\psi, \mathcal{O}) \mathbb{J}_2(n)} \tag{10.59}$$

Here $\mathbb{J}_i(n),\ i = 1, 2$ are penalty functions which account for the fact that a model with $\psi$ breaks is more densely parametrized than a model with a smaller number of breaks. Once the number of break points has been found (say, equal to $\hat{\psi}$), we assign units to groups so as to provide the highest total predictive density, i.e. compute $f^-(Y|H_{\hat{\psi}}, \mathcal{O})$. Since there

are $\mathcal{O}$ possible permutations of the cross section over which to search for groups the optimal permutation rule of units in the cross section is the one which achieves $f^{\dagger}(Y|H_{\hat{\psi}})$.

Two interesting questions which emerge are the following. Can we proceed sequentially to test for breaks? Bai (1997) shows that such a procedure produces consistent estimates of the number and the locations of the breaks. However, when there are multiple groups, the estimated break point is consistent for *any* of the existing break points and its location depends on the "strength" of the break. Second, how can we maximize the predictive density over $\mathcal{O}$ when $n$ is large? When no information on the ordering of the units is available and $n$ is moderately large, the approach is computationally demanding. Geographical, economic or sociopolitical factors may help to provide a restricted set of ordering worth examining. But even when economic theory is silent, the maximization does not require n! evaluations, since many orderings give the same predictive density.

Example 10.17 *Suppose n=4, so there are n!=24 possible orderings to examine. Suppose the initial ordering is 1234 and two groups are found: 1 and 234. Then all permutations of 234 with unit 1 coming ahead, i.e. 1243, 1342, etc., give the same predictive density. Similarly permutations which leave unit 1 last need not be examined, i.e. 2341, 2431, etc. This reduces the number of ordering to be examined to 13. By trying another ordering, say 4213, and finding, for example, two groups: 42 and 13, we can further eliminate all the orderings which rotate the elements of each group, i.e. 4132, 2341, etc.. It is easy to verify that once four carefully selected ordering have been tried and, say, two groups found in each trial, we have exhausted all possible combinations.*

Once the submodel characteristics have been determined, we can estimate $[\bar{\alpha}', \ vech(\Sigma_E)', \ vech(\Sigma_V)']'$ using $f^{\dagger}(Y|H_{\psi})$. For example, if $e_{it}$'s and $v_i$ are normally distributed,

$$\begin{aligned}
\hat{\bar{\alpha}}^j &= \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} \alpha^j_{i,ols} \\
\hat{\Sigma}_j &= \frac{1}{n^j(\mathcal{O})-1} \sum_{i=1}^{n^j(\mathcal{O})} (\alpha^j_{i,ols} - \hat{\bar{\alpha}}^j)(\alpha^j_{i,ols} - \hat{\bar{\alpha}}^j)' - \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} (x_i x_i')^{-1} \hat{\sigma}_i^2 \\
\hat{\sigma}_i^2 &= \frac{1}{T-k}(y_i'y_i - y_i'x_i\alpha_{i,ols}) \qquad (10.60)
\end{aligned}$$

where $x_i$ is the matrix of regressors and $y_i$ the vector of dependent variables for unit $i$ and $\alpha^j_{i,ols}$ is the OLS estimator of $\alpha^j$ obtained using the information for unit $i$ (in group $j = 1, \ldots, \psi+1$). Then an Empirical Bayes posterior point estimate for the $\alpha$ vector is $\tilde{\alpha} = (X'\hat{\Sigma}_E^{-1}X + \hat{\Sigma}_V^{-1})^{-1}(X'\hat{\Sigma}_E^{-1}Y + \hat{\Sigma}_V^{-1}A\hat{\bar{\alpha}})$. Alternatively, if the $e_{it}$'s and the $v_i$'s are normal and $g(a_0, \Sigma_E, \Sigma_V)$ is diffuse, we can jointly estimate $(\bar{\alpha}^j, \Sigma_j, \sigma_i^2)$ and the posterior mean for $\alpha$ as follows:

$$\hat{\bar{\alpha}}^j = \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} (\alpha_i^*)^j$$

$$\begin{aligned}
\hat{\Sigma}_j &= \frac{1}{n^j(\mathcal{O}) - k - 1}[\delta * I + \sum_{i=1}^{n^j(\mathcal{O})} ((\alpha_i^*)^j - \hat{\bar{\alpha}}^j)((\alpha_i^*)^j - \hat{\bar{\alpha}}^j)'] \\
\hat{\sigma}_i^2 &= \frac{1}{T+2}(y_i - x_i\alpha_i^*)'(y_i - x_i\alpha_i^*) \\
(\alpha_i^*)^j &= (\frac{1}{\hat{\sigma}_i^2}x_i'x_i + \hat{\Sigma}_j^{-1})^{-1}(\frac{1}{\hat{\sigma}_i^2}x_i'x_i\alpha_{i,ols} + \hat{\Sigma}_j^{-1}\hat{\bar{\alpha}}^j) \qquad (10.61)
\end{aligned}$$

$j = 1, \ldots, \psi + 1; i = 1, \ldots, n^j(\mathcal{O})$; and $\delta > 0$ but small insures that $\hat{\Sigma}_j$ is positive definite.

Exercise 10.44 *Derive (10.60) and (10.61).*

Example 10.18 *(Convergence clubs). The cross sectional posterior distribution of steady states in example 10.15 shows a multimodal shape. One may therefore be interested in knowing whether there are convergence clubs in the data and where the break point is.*

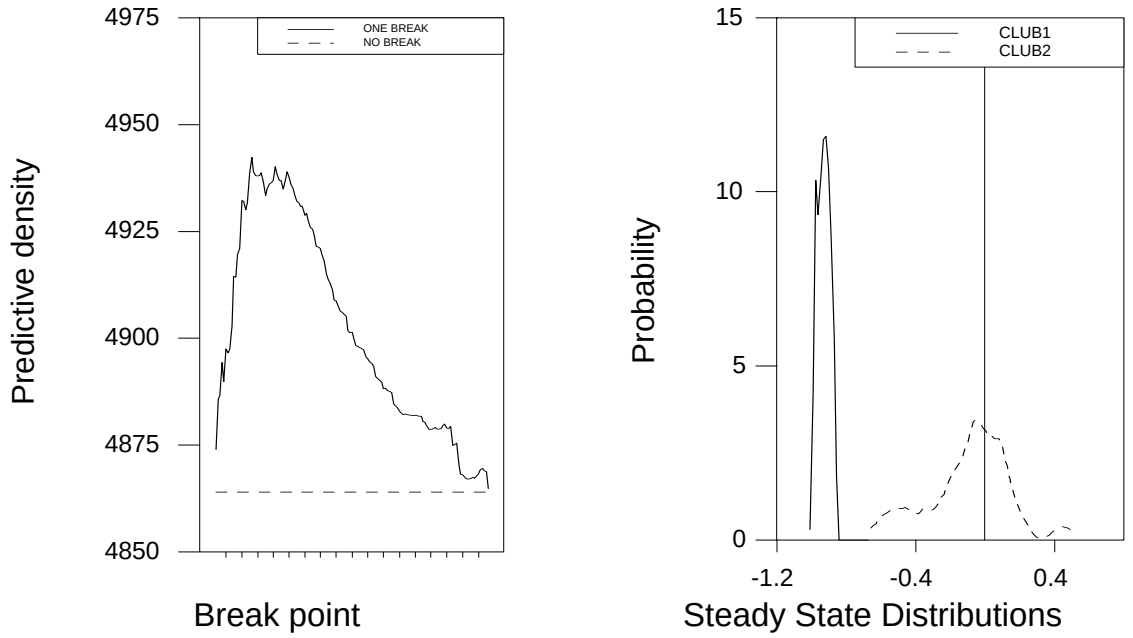


Figure 10.5: Convergence clubs.

*We examined several ordering of cross sectional units based on initial income conditions, growth patterns or geographical characteristics. The one which is optimal orders units using the initial conditions of relative income per-capita. With this ordering, we set $\bar{\psi} = 4$ and sequentially examine $\psi$ against $\psi + 1$ breaks starting from $\psi = 0$. There are up to three breaks in the data with PO ratios of 0.06. 0.52, 0.66 respectively. Conditioning on one*

*break ($\psi = 1$) we plot in the first panel of figure 10.5 the marginal predictive density as a function of the break point, together with the predictive density for $\psi = 0$. Visual inspection indicates that the former is always above the latter and that units up to 23 belong to the first group and from 24 to 144 to the second. The average convergence rates of the two groups are 0.78 and 0.20, suggesting faster convergence to below- average steady states in the first group. The second panel of figure 10.5 suggests that the posterior distributions of the steady states for the two groups are distinct. Not surprisingly, the first 23 units are all poor, Mediterranean and peripheral regions of the EU.*

### 10.5.3 Panel VARs with interdependencies

Neither the panel VAR model studied in chapter 8 nor the specification we have considered so far allow for cross units lagged feedbacks. This may be important e.g. when one is interested in the transmission of shocks across countries. A panel VAR model with interdependencies has the form:

$$y_{it} = A_{1it}(\ell)y_t + A_{2it}(\ell)Y_t + e_{it} \tag{10.62}$$

where $i = 1, ..., n;\ t = 1, ..., T;\ y_{it}$ is a $m_1 \times 1$ vector for each $i$, $y_t = (y'_{1t}, y'_{2t}, \dots y'_{nt})'$, $A^j_{1it}$ are $m_1 \times (nm_1)$ matrices and $A^j_{2it}$ are $m_1 \times m_2$ matrices for each $j$; $Y_t$ is a $m_2 \times 1$ vector of exogenous variables, common to all $i$, $e_{it}$ is a $m_1 \times 1$ vector of disturbances and, for convenience, we have omitted constants and other deterministic components. In (10.62) cross-unit lagged interdependencies appear whenever $A^j_{1it,i'} \neq 0$, for $i' \neq i$ and some $j$, that is, when the matrix of lagged coefficients is not block diagonal at all lags. The presence of lagged cross unit interdependencies adds flexibility to the specification but it is not costless: the number of coefficients is greatly increased (there are $k = nm_1q_1 + m_2q_2$ coefficients in each equation). In (10.62) we allow coefficients to vary over time.

To construct posterior distributions for the unknowns, rewrite (10.62) as:

$$Y_t = X_t\alpha_t + E_t \qquad E_t \sim \mathbb{N}(0, \Sigma_E) \tag{10.63}$$

where $X_t = (I_{nm} \otimes \mathsf{X}_t)$; $\mathsf{X}_t$=$(y'_{t-1},\ y'_{t-2}, \dots, y'_{t-q_1},\ Y'_t, \dots, Y'_{t-q_2})$; $\alpha_t = (\alpha'_{1t}, \dots, \alpha'_{nt})'$ and $\alpha_{it} = (\alpha^{1\prime}_{it}, \dots, \alpha^{m_1\prime}_{it})'$. Here $\alpha^j_{it}$ are $k \times 1$ vectors containing the coefficients for equation $j$ of unit $i$, while $Y_t$ and $E_t$ are $nm \times 1$ vectors containing the endogenous variables and the random disturbances.

Whenever $\alpha_t$ varies with cross–sectional units in different time periods, it is impossible to employ classical methods to estimate it. Two short cuts are typically used: either it is assumed that the coefficient vector does not depend on the unit (apart from a time invariant fixed effect), or that there are no interdependencies (see e.g. Holtz Eakin et al. (1988) or Binder et al (2001)). Neither of these assumptions is appealing in our context. Instead, we assume that $\alpha_t$ can be factored as:

$$\alpha_t = \Xi_1\theta^1_t + \Xi_2\theta^2_t + \sum_{f=3}^{F} \Xi_f\theta^f_t \tag{10.64}$$

where $\Xi_1$ is a vector of ones of dimensions $nmk \times 1$; $\Xi_2$ is a matrix of ones and zeros of dimensions $nmk \times n$, and $\Xi_f$ are conformable matrices. Here $\theta_t^2$ is an $n \times 1$ vector of unit specific factors (the fixed effect), $\theta_t^1$ is the common factor and $\theta_t^f$ is a set of factors which, in principle, is indexed by the unit $i$, the variable $j$, the lag or combinations of all of the above.

Example 10.19 *In a two variable, two lag, two country model with $Y_t = 0$, (10.64) implies*

$$\alpha_t^{i,j,s,\ell} = \theta_t^1 + \theta_t^{2i} + \theta_t^{3j} + \theta_t^{4s} + \theta_t^{5\ell} \tag{10.65}$$

*where $\theta_t^1$ is a common factor, $\theta_t^2 = (\theta_t^{21},\ \theta_t^{22})'$ is a $2 \times 1$ vector of country specific factors, $\theta_t^3 = (\theta_t^{31},\ \theta_t^{32})'$ is a $2 \times 1$ vector of equation specific factors, $\theta_t^{4s} = (\theta_t^{41},\ \theta_t^{42})'$ is a $2 \times 1$ vector of variable specific factors, $\theta_t^{5\ell} = (\theta_t^{51},\ \theta_t^{52})'$ is a $2 \times 1$ vector of lag specific factors.*

All factors in (10.64) are allowed to be time varying; in fact, time invariant structures can be obtained via restrictions on the law of motion of the $\theta_t$. Also, while the factorization in (10.64) is exact, in practice only a few factors will be specified: in that case all the omitted factors will be aggregated into an error term $v_{1t}$. Note also that with (10.64) the over-parametrization of the original model is dramatically reduced because the $nmk \times 1$ vector $\alpha_t$ depends on a much lower dimensional vector of factors.

Let $\theta_t = [\theta_t^1, (\theta_t^2)', (\theta_t^3)', \ldots, (\theta_t^{f_1})',\ f_1 < F]$ and write (10.64) as

$$\alpha_t = \Xi\theta_t + v_{1t} \quad v_{1t} \sim \mathbb{N}(0, \Sigma_E \otimes \Sigma_V) \tag{10.66}$$

where $\Xi = [\Xi_1, \Xi_2, \ldots, \Xi_{f_1}]$ and $V$ is a $k \times k$ matrix. We assume a hierarchical structure on $\theta_t$ which allows for time variations and exchangeability:

$$\theta_t = (I - \mathbb{D}_1)\,\bar{\theta} + \mathbb{D}_1\theta_{t-1} + v_{2t} \qquad v_{2t} \sim \mathbb{N}\left(0, \Sigma_{v_{2t}}\right) \tag{10.67}$$

$$\bar{\theta} = \mathbb{D}_0\theta_0 + v_3 \qquad v_3 \sim \mathbb{N}(0, \Sigma_{v_3}) \tag{10.68}$$

We set $\Sigma_V = \sigma_v^2 I_k$ and, as in section 10.4, we let $\Sigma_{v_{2t}} = \phi_3 * \Sigma_{v_{2t-1}} + \phi_2 * \Sigma_0$ where $\Sigma_0 = diag(\Sigma_{01}, \Sigma_{02}, \ldots \Sigma_{0,f_1})$. We assume that $v_{it},\ i = 1, 2, 3$ and $E_t$ are mutually independent and that $(\sigma_v^2, \phi_3, \phi_2, \mathbb{D}_1, \mathbb{D}_0)$ are known. Here $\mathbb{D}_0$ a matrix which restricts (part of the) means of the factors of the coefficients via an exchangeable prior.

To sum up, the prior for $\alpha_t$ has a multi-step hierarchical structure: with (10.66) we make a large number of coefficients depend on a smaller number of factors. The factors are then allowed to have a general evolving structure (equation (10.67)) and the prior mean of e.g. unit specific factors is potentially linked across units (equation (10.68)). The variance of the innovations in $\theta_t$ is allowed to be time varying to account for heteroschedasticity and other generic volatility clustering that are unit specific or common across units. To complete the specification we need to provide prior densities for $(\Sigma_E^{-1},\ \theta_0,\ \sigma_v^{-2},\ \Sigma_0^{-1},\ \Sigma_{v_3}^{-1})$. Canova and Ciccarelli (2002) study both informative and uninformative priors. Here we consider a special case of the non-informative framework they use.

Since $\alpha_t$ is a $nmk \times 1$ vector, the derivation of its posterior distribution with numerical methods is computationally demanding when $m$ or $n$ are large. To avoid problems rewrite the model as

$$\begin{aligned} y_t &= X_t \Xi \theta_t + e_t \\ \theta_t &= (I - \mathbb{D}_1)\,\bar{\theta} + \mathbb{D}_1 \theta_{t-1} + v_{2t} \\ \bar{\theta} &= \mathbb{D}_0 \theta_0 + v_{3t} \end{aligned} \tag{10.69}$$

where $e_t = E_t + X_t v_{1t}$ has covariance matrix $\sigma_t \Sigma_E = (1 + \sigma^2 X_t' X_t)\Sigma_E$. In (10.69) we have integrated $\alpha_t$ out of the model so that $\theta_t$ becomes the vector of parameters of interest.

We assume $\Sigma_{o1} = \phi_1,\ \Sigma_{0i} = \phi_i * I,\ i = 2, \ldots, f_1$, where $\phi_i$ controls the tightness of factor $i$ of the coefficient vector. Furthermore assume that: $g(\Sigma_E^{-1}, \sigma^{-2}, \theta_0, \sigma_v^{-2}, \Sigma_{v_3}, \phi_i) = g(\Sigma_E^{-1}) g(\sigma^{-2}) g(\sigma_v^{-2}) g(\theta_0, \Sigma_{v_3}) \prod_i g(\phi_i)$ where $g(\Sigma_E^{-1})$ is $\mathbb{W}(\bar{\nu}_1, \bar{\Sigma}_1^{-1})$; $g(\sigma^{-2} \propto$ constant; $g(\sigma_v^{-2}) \propto \sigma_v^{-2}$; $g(\theta_0, \Sigma_{v_3}) \propto \Sigma_{v_3}^{-(\bar{\nu}_2+1)/2}$ where $\bar{\nu}_2 = 1 + N + \sum_{j=1}^{m_1} dim(\theta_{j,t}^f), f > 1$ and $g(\phi_i) \propto (\phi_i)^{-1}$; and the hyperparameters $\bar{\Sigma}_1, \bar{\nu}_1$ are assumed to be known or estimable from the data. The assumptions made imply that the prior for $e_t$ has the form $(e_t | \sigma_t) \sim \mathbb{N}(0, \sigma_t \Sigma_E)$, and $\sigma_t^{-2}$ is Gamma distributed so that $e_t$ is distributed as a multivariate $t$ centered at 0, with scale matrix which depends on $\Sigma_E$ and degrees of freedom equal to $\dim(X_t)$. Since the likelihood of the data is proportional to $\left(\prod_{t=1}^T \sigma_i\right)^{-Nm/2} |\Sigma_E|^{-T/2} \exp\left[-\frac{1}{2} \sum_t (y_t - X_t \Xi \theta_t)' (\sigma_t \Sigma_E)^{-1} (y_t - X_t \Xi \theta_t)\right]$, it is easy to derive the conditional posteriors of the unknowns since the prior is conjugate. In fact, conditional on the other parameters, $\Sigma_E^{-1}$ is Wishart, $\sigma_t^{-2}$ is a Gamma, $\theta_0$ is Normal, $\Sigma_{v_3}^{-1}$ is Wishart and $\phi_i^{-1}$ is Gamma distributed.

Exercise 10.45 *Derive the parameters of the posterior of $\Sigma_E^{-1}$, $\sigma_t^{-1}$, $\Sigma_{v_3}^{-1}$, $\phi_i^{-1}$ and $\theta_0$.*

Finally, the conditional posterior distribution of $(\theta_1, .., \theta_T \mid y^T, \psi_{-\theta_t})$ can be obtained with the Kalman filter/ smoother as described in section 10.4. With these conditional, the Gibbs sampler can be used to draw a sequence of parameters from the joint posterior.

### 10.5.4 Indicators

The panel VAR (10.63) with the hierarchical prior (10.66)- (10.68) provides a framework to recursively construct coincident/leading indicators. In fact, the first equation in (10.69) is

$$y_t = \sum_{f=1}^{f_1} X_{f,t} \theta_t^f + e_t \tag{10.70}$$

where $X_{ft} = X_t \Xi_f$. In (10.70) $y_t$ depends on a common time index $X_{1t}$, on a $n \times 1$ vector of unit specific indices $X_{2t}$, and of a set of indices which depend on variables, lags, units, etc. These indices are particular combinations of lags of the VAR variables, while $\theta_t^j$ measure the impact that different linear combinations of the lags of the right hand side variables have on the current endogenous variables. Hence, it is possible to construct leading indicators

directly from the VAR, without any preliminary distinction between leading, coincident and lagging variables. Also, because the model is recursive, both single-step and multi-step leading indicators can be obtained from the posterior for $\theta_t$. Finally, fan charts can be constructed using the predictive density of future observations and the output of the Gibbs sampler.

Example 10.20 *Suppose we are interested in a model featuring a common, a unit specific and a variable specific indicator. Given (10.70), a leading indicator for $y_t$ based on the common information available at time $t-1$ is $CLI_t = X_{1t}\theta^1_{t|t-1}$; a vector of leading indicators based on the common and unit specific information is $CULI_t = X_{1t}\theta^1_{t|t-1} + X_{2t}\theta^2_{t|t-1}$; a vector of indicators based on the common and variable specific information is $CVLI_t = X_{1t}\theta^1_{t|t-1} + X_{3t}\theta^3_{t|t-1}$; and vector of indicators based on the common, unit specific and variable specific information is $CUVLI_t = X_{1t}\theta^1_{t|t-1} + X_{2t}\theta^2_{t|t-1} + X_{3t}\theta^3_{t|t-1}$.*

While we have derived (10.70) using a prior on the panel VAR, one may want to start the investigation directly from (10.70). In this case, a researcher may be interested in assessing how many indices are necessary to capture the heterogeneities in the coefficients across time, units and variables. We can use Bayes factors to make this choice. A model with $i$ indices is preferable to a model with $i+1$ indices, $i = 1, 2, \ldots, f_1 - 1$, if $\frac{f(y^{t+\tau}|\mathcal{M}_i)}{f(y^{t+\tau}|\mathcal{M}_{i+1})} > 1$ where $f(y^{t+\tau}|\mathcal{M}_i) = \int f(y^{t+\tau}|\theta_{t,i}, \mathcal{M}_i) g(\theta_{t,i}|\mathcal{M}_i) d\theta_{t,i}$ is the predictive density of a model with $i$ indices for $y^{t+\tau} = [y_{t+1}, \ldots, y_{t+\tau}]$, $g(\theta_{t,i}|\mathcal{M}_i)$ is the prior for $\theta$ in model $i$ and $f(Y^{t+\tau}|\theta_{t,i}, \mathcal{M}_i)$ the density of future data, given $\theta_{t,i}$ and $\mathcal{M}_i$. The predictive density for future $y_{t+\tau}$ in model $i$ can be computed with the output of the Gibbs sampler. To do so, draw $\theta^l_t$ from the posterior distribution, construct forecast $y^l_{t+\tau}$ and prediction errors for each $\tau$ and average across draws.

### 10.5.5 Impulse responses

Impulse responses for the model can be computed as posterior revisions of the forecast errors. Since the model is non-linear, forecasts for the vector of endogenous variables may change because the innovations in the model or the innovations in the coefficients are different from zero. Furthermore, because of time variations, revisions depend on the history and the point in time where they are computed.

To see this set $Y_t = 0$, rewrite (10.63) as $\mathbb{Y}_t = \mathbb{A}_t \mathbb{Y}_{t-1} + \mathbb{E}_t$ and let $\alpha_t = vec(\mathbb{A}_{1t})$ where $\mathbb{A}_{1t}$ are the first $m_1$ rows of $\mathbb{A}_t$. Iterating $\tau$ times we have

$$y_{t+\tau} = \mathbb{S}(\prod_{s=0}^{\tau-1} \mathbb{A}_{t+\tau-s})\mathbb{Y}_t + \sum_{i=0}^{\tau-1} \mathbb{A}^*_{i,t+\tau}\, e_{t+\tau-i} \tag{10.71}$$

where $\mathbb{S} = [I, 0, .., 0]$ and $\mathbb{A}^*_{i,t+\tau} = \mathbb{S}(\prod_{s=0}^{i-1} \mathbb{A}_{t+\tau-s})\mathbb{S}'$; $\mathbb{A}^*_{0,t+\tau} = I$. Using (10.67) into (10.66) and iterating gives

$$\alpha_{t+\tau} = \Xi\theta_{t+\tau} + v_{1t+\tau} = \Xi\mathbb{D}_1^{\tau+1}\theta_{t-1} + \Xi\sum_{i=1}^{\tau} \mathbb{D}_1^i(I - \mathbb{D}_1)\bar{\theta} + \Xi\sum_{i=1}^{\tau} \mathbb{D}_1^i v_{2t+\tau-i} + v_{1t+\tau}$$

(10.72)

Define responses at step $j$, given information at $t$ and terminal horizon $\tau$ as $Rev_{t,j}(\tau) = E_{t+j}\mathbb{Y}_{t+\tau} - E_t\mathbb{Y}_{t+\tau},\ \forall \tau \geq j+1$. Using $E_t y_{t+\tau} = \mathbb{S}E_t(\prod_{s=0}^{\tau-1} A_{t+\tau-s})\mathbb{Y}_t$, we have that

$$Rev_{t,j}(\tau) = \sum_{s=0}^{j-1}(E_{t+j}\mathbb{A}^*_{\tau-j+s,t+\tau})e_{t+j-s} + \mathbb{S}[E_{t+j}(\prod_{s=0}^{\tau-j-1}\mathbb{A}_{t+\tau-s})\prod_{s=\tau-j}^{\tau-1}\mathbb{A}_{t+\tau-s} - E_t(\prod_{s=0}^{\tau-1}\mathbb{A}_{t+\tau-s})]\mathbb{Y}_t \tag{10.73}$$

From (10.73) it is clear that forecast revisions can occur because new information present in the innovations of the model, $e_t$, or of the coefficients, $v_{2t}$, alter previous forecasts of $\mathbb{Y}_{t+\tau}$.

Example 10.21 *In equation (10.73) take $j = 1, \tau = 2$. Then $Rev_{t,1}(2) = E_{t+1}\mathbb{Y}_{t+2} - E_t\mathbb{Y}_{t+2} = E_{t+1}\left(\mathbb{A}^*_{1,t+2}\right)e_{t+1} + \mathbb{S}\left[E_{t+1}\left(\mathbb{A}_{t+2}\right)\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)\right]\mathbb{Y}_t$. Similarly, $j = 2, k = 3$, imply $Rev_{t,2}(3) = E_{t+2}\mathbb{Y}_{t+3} - E_t\mathbb{Y}_{t+3} = \sum_{s=0}^{1}\left(E_{t+2}\mathbb{A}^*_{1+s,t+3}\right)e_{t+2-s} + \mathbb{S}[E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+3}\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)]\mathbb{Y}_t$ where $\sum_{s=0}^{1}\left(E_{t+2}\mathbb{A}^*_{1+s,t+3}\right)e_{t+2-s} = \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{S}'e_{t+2} + \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{S}'e_{t+1}$. Hence, changes in $\mathbb{Y}_{t+3}$ due to innovations of the model are $\mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{S}'e_{t+2} + \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{S}'e_{t+1}$ and due to innovations in the coefficients are $\mathbb{S}[E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+3}\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)]\mathbb{Y}_t$. Clearly, responses depend on the time when they are generated (e.g. $t$ vs. $t+1$) and the history of $y_t$.*

The output of the Gibbs sampler can be used to compute the expressions appearing in (10.73). Conditioning on $\mathbb{A}_t$, assuming that $e_t \neq 0$ and that all future innovations in both coefficients and variables are integrated out, $Rev_{t,j}(\tau)$ can be computed as follows:

Algorithm 10.4

1) *Draw $(e_{t+1}, \ldots, e_{t+j})$ and $(\mathbb{A}_{t+1}, \ldots, \mathbb{A}_{t+j})$ from the posterior distribution $L+1$ times.*

2) *For each draw $l = 2, \ldots L+1$, compute $\hat{A}^{*l}_{i,j} = \prod_{s=0}^{j}\mathbb{A}^l_{t+\tau-s}$. Average it $\hat{A}^{*l}_{i,j}$ over $l$.*

3) *For each draw $l = 2, \ldots, L+1$, compute $\hat{e}_{t+\tau} = \sum_{l=2}^{L+1} e^l_{t+\tau}$, $\tau > 1$.*

4) *Given $\mathbb{Y}_t$, $(e^l_{t+j}, \mathbb{A}^l_{t+j})$ from 1), $\hat{A}^{*l}_{i,j}$ from 2), $\hat{e}_{t+\tau}$ from 3), compute $Rev_{t,j}(\tau)$.*

Example 10.22 *We use a VAR model for G-7 countries with GDP growth, inflation, employment growth and the real exchange rate for each country and three indices: a $2 \times 1$ vector of common factors - one for EU and one for non-EU countries, a $7 \times 1$ vector of country specific factors and a $4 \times 1$ vector of variable specific factors.*

*We assume time variations in the factors, use non-informative priors on the hyperparameters but do not impose exchangeability. Figure 10.6 presents 68% bands for the CUVLI indicator for EU GDP growth and inflation, constructed recursively using information available one year in advance. Actual values of EU GDP growth and inflation are superimposed. The model predicts the ups and downs of both series reasonably well using one year ahead information. The Theil-U statistics over the 1996:1-2000:4 and 1991:1-1995:4 sample are 0.87 and 0.66, respectively, much lower than those obtained with a single country VAR (1.25, 1.06) or with a univariate AR (1.04, 0.97).*

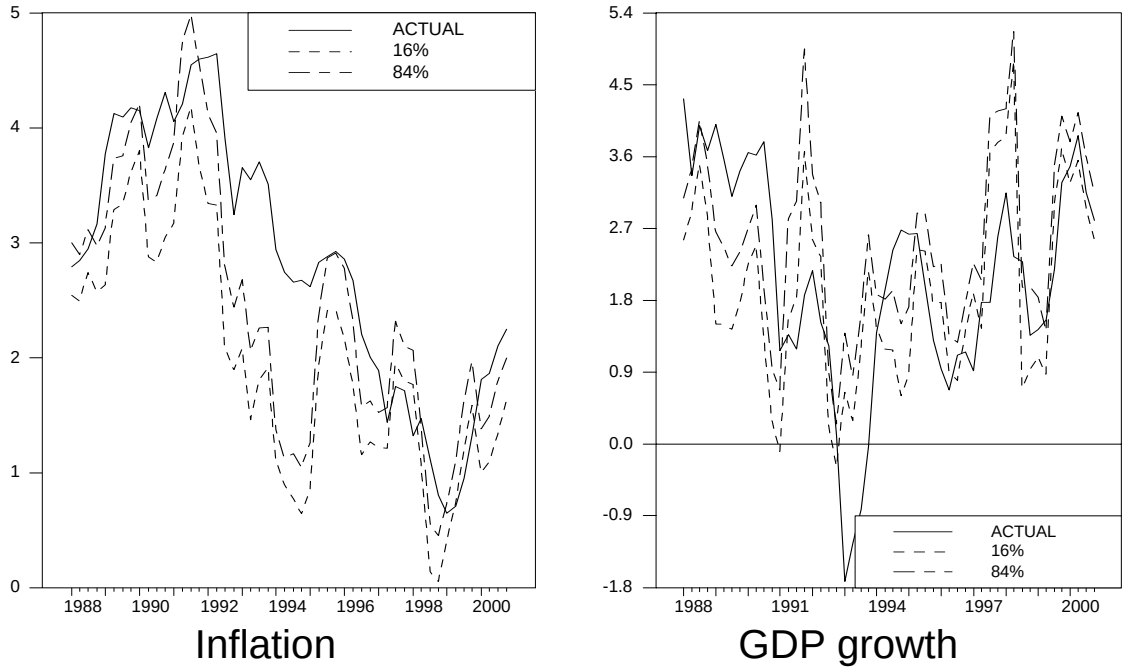


Figure 10.6: One year ahead 68% prediction bands, EU