# Chapter 5

# Bayesian VAR Analysis

Given the fact that VAR models are estimated on comparatively short samples and hence tend to be imprecisely estimated, many users of VAR models are sympathetic to the idea of imposing additional structure in estimation. For example, shrinking parameter estimates toward specific values may help reduce the variance of unrestricted LS estimators. The Bayesian approach provides a formal framework for incorporating such extraneous information in estimation and inference. It also facilitates the inclusion of extraneous economic information about the VAR model parameters that would be difficult to incorporate in frequentist analysis.

The ideas behind the Bayesian approach differ fundamentally from the frequentist approach. Broadly speaking, frequentists condition on the existence of a parameter vector, say $\boldsymbol{\theta}$, that governs the DGP from which the observed sample of data is thought to have been obtained. The objective is to infer the value of $\boldsymbol{\theta}$ from this sample. Whereas the data are stochastic, the parameter vector $\boldsymbol{\theta}$ is nonstochastic. Probability statements refer to the properties of the estimator of $\boldsymbol{\theta}$ in repeated sampling.

In contrast, in Bayesian analysis, the parameter vector $\boldsymbol{\theta}$ is treated as stochastic, whereas the data are considered nonstochastic. Bayesians are concerned with modeling the researcher's beliefs about $\boldsymbol{\theta}$. These beliefs are expressed in the form of a probability distribution. The Bayesian probability concept is a subjective probability concept. It does not require a repeated sampling experiment. Moreover, the nature of the DGP is immaterial for inference on the parameter of interest because inference conditions on the observed data.

The researcher's subjective beliefs about the value of the parameter vector $\boldsymbol{\theta}$ before inspecting the data are summarized in the form of a prior probability distribution (or prior for short) for $\boldsymbol{\theta}$. This prior information is formally combined with the information contained in the data, as captured by the likelihood function of the model, to form the posterior probability distribution (or simply


Prepared for: Kilian, L. and H. Lütkepohl, *Structural Vector Autoregressive Analysis*, Cambridge University Press, Cambridge.

posterior) for $\boldsymbol{\theta}$. This posterior conveys everything the researcher knows about the model parameters after having looked at the data.

Frequentists are aware that they do not know the DGP for a given set of variables, but they evaluate the properties of their methods under the assumption that there exists a true parameter vector $\boldsymbol{\theta}$ that can be objectively defined. They postulate a DGP and then conduct their analysis, as if this model structure including any distributional assumptions were correct. In contrast, Bayesians do not need to take a stand on the DGP. However, their formal framework for deriving and evaluating the posterior requires users to articulate their prior beliefs in the form of a prior probability distribution. It also involves assuming a specific distribution (or family of distributions) for the data, when evaluating the likelihood.

The choice between frequentist and Bayesian methods depends on the preferences of the user. In practice, convenience may also play an important role. In some situations frequentist methods are easier to deal with, and in other cases Bayesian methods are more convenient. It has to be kept in mind, however, that these approaches not only may produce numerically different answers in many cases, but that their interpretation is fundamentally different, even when the estimates coincide numerically. Since Bayesian methods are frequently used in VAR analysis, it is essential to have at least a basic understanding of this approach.

Although Bayesian methods often require extensive computations, they have become quite popular for VAR analysis because the cost of computing has decreased dramatically over the last decades. Moreover, new methods and algorithms have broadened the applicability of Bayesian methods. In Section 5.1, we briefly review some basics of the Bayesian methodology and terminology and then discuss Bayesian methods commonly used in VAR analysis. Section 5.2 reviews some commonly used prior specifications for the reduced-form VAR parameters. There are many good introductory treatments of Bayesian methodology in general and for macroeconometric analysis in particular. Canova (2007) and Koop and Korobilis (2009) fall in the latter category. Recent surveys include, for example, Ciccarelli and Rebucci (2003), Del Negro and Schorfheide (2011), and Karlsson (2013). The present chapter follows in part Lütkepohl (2005, Section 5.4), Geweke and Whiteman (2006), and Canova (2007).

## 5.1 Basic Terms and Notation

### 5.1.1 Prior, Likelihood, Posterior

The Bayesian approach treats the data, $\mathbf{y} = (y_1', \ldots, y_T')'$, as given and the parameter of interest, $\boldsymbol{\theta}$, as unknown. Inference about $\boldsymbol{\theta}$ is conditional on the data. Prior information on $\boldsymbol{\theta}$ is assumed to be available in the form of a density. Suppose the prior information is summarized in the prior probability density function (pdf) $g(\boldsymbol{\theta})$ and the sample pdf conditional on a particular value of the parameter $\boldsymbol{\theta}$ is $f(\mathbf{y}|\boldsymbol{\theta})$. The latter function is algebraically identical to

the likelihood function $l(\boldsymbol{\theta}|\mathbf{y})$. The two types of information are combined by applying Bayes' theorem, which states that the joint density is

$$f(\boldsymbol{\theta}, \mathbf{y}) = g(\boldsymbol{\theta}|\mathbf{y}) f(\mathbf{y}),$$

and, hence,

$$g(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta})}{f(\mathbf{y})}.$$

Here $f(\mathbf{y})$ denotes the unconditional sample density which is just a normalizing constant for a given sample $\mathbf{y}$ and does not depend on $\boldsymbol{\theta}$. Scaling the posterior by $f(\mathbf{y})$ ensures that the posterior density integrates to 1. In other words, the joint sample and prior information can be summarized by a function that is proportional to the likelihood function times the prior density $g(\boldsymbol{\theta})$,

$$g(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{y}) g(\boldsymbol{\theta}), \tag{5.1.1}$$

where $\propto$ indicates that the right-hand side is proportional to the left-hand side. The conditional density $g(\boldsymbol{\theta}|\mathbf{y})$ is the posterior pdf. It contains all the information that we have on the parameter vector $\boldsymbol{\theta}$ after having updated our prior views by looking at the data. Thus, the posterior pdf is the basis for estimation and inference. In the next subsection we discuss Bayesian inference within this general framework in more detail.

### 5.1.2 Bayesian Estimation and Inference

#### Point Estimators

Bayesian inference on the parameter vector $\boldsymbol{\theta}$ is based on the posterior distribution. Often moments of the posterior distribution are of interest. For example, one may be interested in $\mathbb{E}(\boldsymbol{\theta}|\mathbf{y})$. The posterior mean is often used as a point estimator for $\boldsymbol{\theta}$, if the posterior distribution is Gaussian or at least symmetric. More generally one is often interested in expected values of functions of $\boldsymbol{\theta}$. If we are interested in constructing a point estimate of some function $h(\boldsymbol{\theta})$ that may be vector-valued or a scalar, we minimize the expected loss of $h(\boldsymbol{\theta})$ based on some loss function. Denoting this loss function by $\mathcal{L}(h^\dagger, h(\boldsymbol{\theta}))$, the point estimate, say $\widetilde{h}^\dagger$, is chosen such that

$$\widetilde{h}^\dagger = \text{argmin}_{h^\dagger} \int \mathcal{L}(h^\dagger, h(\boldsymbol{\theta})) g(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{5.1.2}$$

where $h^\dagger$ denotes an element of the range of $h(\boldsymbol{\theta})$. For example, if the loss function is quadratic and the first two moments of the posterior distribution exist, the point estimate corresponds to the posterior mean. Under alternative loss functions the median or the mode of the posterior distribution may be used as point estimates. These estimates coincide if the posterior distribution is Gaussian.

### Credible Sets

The Bayesian concept corresponding to a confidence set in classical inference is a credible set. For $0 < \gamma < 1$, a $(1-\gamma)100\%$ credible set for $\boldsymbol{\theta}$ with respect to the posterior $g(\boldsymbol{\theta}|\mathbf{y})$ is a set $\Omega$ such that

$$\mathbb{P}(\boldsymbol{\theta} \in \Omega|\mathbf{y}) = \int_{\Omega} g(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = 1 - \gamma. \tag{5.1.3}$$

This set can be estimated by sampling from the posterior distribution. In general it will not be unique, however. To make the credible set unique, one may choose $\Omega$ such that $g(\boldsymbol{\theta}|\mathbf{y}) \geq c_\gamma$ for all $\boldsymbol{\theta} \in \Omega$, where $c_\gamma$ is the largest number such that $\mathbb{P}(\boldsymbol{\theta} \in \Omega|\mathbf{y}) = 1 - \gamma$. In other words, one may choose the highest posterior density (HPD) set.

Depending on the context, this set may differ numerically from a frequentist confidence interval or region. Given that the posterior density is partly determined by the prior, the posterior for a specific parameter may, for example, be bimodal, in which case an HPD region may consist of two disjoint sets. Even if the HPD set and the frequentist confidence set coincide numerically, their interpretation is quite different. A $(1-\gamma)100\%$ confidence interval estimator asymptotically contains the true parameter value in $(1-\gamma)100\%$ of the cases in repeated sampling. This means that the true parameter is either in a specific interval estimate or it is not in this interval. Frequentists cannot say whether the true parameter is inside or outside a specific confidence interval estimate, but only that an interval constructed by their method will include the true value in repeated sampling with probability $1-\gamma$. In contrast, the Bayesian credible set specifies a region where $(1-\gamma)100\%$ of the probability mass of the posterior distribution is concentrated. This allows them to make probability statements about the parameter of interest, $\boldsymbol{\theta}$, given the data and the stochastic model of the data. There is no true value of $\boldsymbol{\theta}$ for a Bayesian, nor is it considered important what the true value of $\boldsymbol{\theta}$ is. The construction of credible sets easily generalizes to smooth functions $h(\boldsymbol{\theta})$, the posterior of which may be evaluated by simulation.

### Testing Statistical Hypotheses and Model Comparison

Hypothesis testing is another important part of frequentist statistics. In frequentist hypothesis testing, the model specified under the null hypothesis is maintained unless there is overwhelming evidence against it. Bayesians are typically not interested in hypothesis testing, but in quantifying the empirical support for alternative models. Typically, Bayesians choose between models based on their posterior odds ratios. Suppose that we want to compare two models $M_1$ and $M_2$, for example. Then the posterior odds ratio is

$$\frac{g(M_1|\mathbf{y})}{g(M_2|\mathbf{y})} = \frac{g(M_1)}{g(M_2)} \times \frac{f(\mathbf{y}|M_1)}{f(\mathbf{y}|M_2)}, \tag{5.1.4}$$

where $g(M_1)/g(M_2)$ reflects the prior odds and $f(\mathbf{y}|M_1)/f(\mathbf{y}|M_2)$ is the Bayes factor. The Bayes factor is the ratio of the marginal likelihoods or marginal

data densities,

$$f(\mathbf{y}|M_j) = \int f(\mathbf{y}|M_j, \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad j = 1, 2,$$

that are obtained by integrating the parameters out of the likelihood function. If the posterior odds ratio is greater than one, this is evidence in favor of model $M_1$, whereas a posterior odds ratio smaller than one favors $M_2$. Obviously, if the prior for all models is the same, the ratio of the marginal likelihoods, known as the Bayes factor, determines which model is preferred. More generally, if there are $m$ alternative models that are regarded as a priori equally likely, then a Bayesian solution would be to select the model with the highest marginal likelihood. This procedure is also applicable if nested models are considered, as would be the case, for example, in choosing between VAR models of different lag orders.[1]

Note that Bayesian model comparison and frequentist hypothesis testing have quite different interpretations. Frequentist hypothesis testing treats the parameter vector as fixed. A rejection of the null hypothesis in favor of the alternative hypothesis occurs when the sample gives rise to a value of the test statistic that exceeds the critical value at the chosen significance level. The choice of the significance level reflects the user's definition of reasonable doubt. If there is no evidence against the null hypothesis beyond a reasonable doubt, the test is not informative. Thus, the choice of the null hypothesis matters. The null hypothesis is always protected from rejection rather than both hypotheses being treated symmetrically.

Although one could make the case for choosing model $M_1$ over $M_2$ if the posterior odds ratio is much larger than one, implementing this rule in practice would require choosing a threshold value, beyond which a Bayesian would choose $M_1$ for further analysis and discard $M_2$. Choosing such a threshold would be analogous to choosing a significance level in frequentist hypothesis testing because eliminating one model from further consideration corresponds to rejecting the null hypothesis in frequentist testing. Thus, Bayesian model comparison is more akin to model selection procedures in frequentist statistics (such as the information criteria discussed in Chapter 2) than to classical hypothesis testing in that it treats all models under consideration symmetrically.

More generally, many Bayesians view all attempts at choosing between alternative models as misguided. A common alternative is Bayesian model averaging. Consider $m$ candidate models $M_1, \ldots, M_m$. If interest centers on $\mathbb{E}(h(\boldsymbol{\theta}))$, for example, one averages over the candidate models such that

$$\mathbb{E}(h(\boldsymbol{\theta})) = \sum_{i=1}^{m} \mathbb{E}(h(\boldsymbol{\theta})|\mathbf{y}, M_i) g(M_i|\mathbf{y}). \tag{5.1.5}$$

Alternatively, a Bayesian may select the median model. Another choice would

[1] Even some subsets of the parameter space having measure zero is not a problem for Bayesian analysis, as long as the prior probability of the parameters is nonzero, given that models with zero prior probability are of no interest from a Bayesian perspective.

be to select the a posteriori most likely model, i.e., the model with the largest $g(M_i|\mathbf{y})$, $i = 1, \ldots, m$. Each of these choices reflects a different loss function.

### 5.1.3 Simulating the Posterior Distribution

Bayesian inference is based on the posterior distribution. That distribution may not be available in closed form, but often can be simulated using numerical methods. In this subsection some numerical algorithms are discussed that are useful for this purpose.

Typically, the objective is to generate random draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ from the posterior distribution of the parameters or some function of these parameters such as the impulse responses associated with a VAR model. Generating draws from potentially complicated and high-dimensional distributions is an important part of Bayesian analysis. Suppose that we are interested in the expectation of $h(\boldsymbol{\theta})$ which may be a vector or a scalar function. Laws of large numbers suggest that we can obtain a good approximation to this quantity by drawing at random $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ from the posterior distribution and noting that

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}^{(i)}) \overset{a.s.}{\to} \mathbb{E}(h(\boldsymbol{\theta})|\mathbf{y}) = \int h(\boldsymbol{\theta}) g(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{5.1.6}$$

The approximation precision increases with the number $n$ of random draws.

Simulating the posterior distribution is difficult when this distribution is nonstandard or unknown. Below we discuss a range of tools designed for such situations. We only outline the general ideas to enable the reader to determine which algorithm is suitable in a specific situation, without providing the full details. Some of these methods are used in later chapters in the context of specific applications. Further details can be found, for example, in Canova (2007, Section 9.5) or Geweke and Whiteman (2006, Section 3).

#### Direct Sampling

If the joint posterior distribution of all parameters is from a known distribution family, random samples can be generated easily using the random number generators provided in standard software packages. A case in point is Gaussian posterior distributions. Unfortunately, in practice, often the posterior distribution is not from a known family, in which case more sophisticated sampling techniques are required.

#### Acceptance Sampling

Sometimes the posterior density is not available in closed form, but what is known is a function that is proportional to the posterior density. A function that is proportional to a given density function, but does not integrate to one, is called a kernel of this density function. For example, the product of the prior and the likelihood function, $g(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{y})$, is a kernel of the posterior density. It is

not a density because the area under the function usually does not integrate to one. Denote the kernel of the posterior density $g^*(\boldsymbol{\theta}|\mathbf{y})$.

If the underlying distribution is difficult to sample from, we may choose a density $g^{AS}(\boldsymbol{\theta})$ that is easy to sample from and that satisfies $g^{AS}(\boldsymbol{\theta}) > 0$ whenever $g^*(\boldsymbol{\theta}|\mathbf{y}) > 0$. The function $g^{AS}(\boldsymbol{\theta})$ is called the source density. To ensure strictly positive values of the source density whenever $g^*(\boldsymbol{\theta}|\mathbf{y}) > 0$, one may, for example, choose a normal density for $g^{AS}(\boldsymbol{\theta})$ because this density is strictly positive on the whole real line. The maximum (or supremum) of the ratio of the posterior kernel and the source density is defined as

$$\varrho = \sup_{\boldsymbol{\theta} \in \{\boldsymbol{\theta} | g^*(\boldsymbol{\theta}|\mathbf{y}) > 0\}} \frac{g^*(\boldsymbol{\theta}|\mathbf{y})}{g^{AS}(\boldsymbol{\theta})}.$$

Then the $i^{\text{th}}$ sample value, $\boldsymbol{\theta}^{(i)}$, from the posterior can be obtained by proceeding as follows:

**Step 1.** Draw a random number $u$ from $\mathcal{U}(0,1)$, the uniform distribution on the interval $(0,1)$, and draw $\boldsymbol{\theta}^+$ from the distribution corresponding to $g^{AS}(\boldsymbol{\theta})$.

**Step 2.** Retain $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^+$ if $u < g^*(\boldsymbol{\theta}^+|\mathbf{y})/(\varrho \cdot g^{AS}(\boldsymbol{\theta}^+))$. Otherwise return to Step 1.

It can be shown that this algorithm provides a random sample from the posterior distribution. Clearly, if $g^{AS}(\boldsymbol{\theta}|\mathbf{y}) = g^*(\boldsymbol{\theta}|\mathbf{y})$, then $\varrho = 1$,

$$\frac{g^*(\boldsymbol{\theta}^+|\mathbf{y})}{\varrho \cdot g^{AS}(\boldsymbol{\theta}^+)} = 1.$$

Thus, we would retain all draws from the source distribution in Step 2. If, however, the source density is very different from the posterior density such that it assumes very small values when $g(\boldsymbol{\theta}|\mathbf{y})$ is large, then the ratio $g^*(\boldsymbol{\theta}^+|\mathbf{y})/(\varrho \cdot g^{AS}(\boldsymbol{\theta}^+))$ tends to be small, and only very few draws are accepted. Hence, a large number of draws from the source distribution may be necessary for generating a sufficiently large number of draws from the posterior. In other words, although the algorithm is general, it can be computationally costly and other algorithms may be preferable.

### Importance Sampling

Importance sampling avoids the drawback of having to discard many posterior draws. Instead, we retain a suitably weighted average of all posterior draws. This proposal dates back at least to Hammersly and Handscomb (1964) and was first used in the econometrics literature by Kloek and van Dijk (1978). The idea is to generate a sample that facilitates the estimation of the posterior moments of functions of $\boldsymbol{\theta}$. Suppose again that the function of interest is $h(\boldsymbol{\theta})$ and let $g^{IS}(\boldsymbol{\theta})$ be a proper source density, defined as a density that approximates

$g^*(\boldsymbol{\theta}|\mathbf{y})$ and has the same support. Furthermore, define a weighting function as the ratio of $g^*$ and $g^{IS}$,

$$w(\boldsymbol{\theta}) = \frac{g^*(\boldsymbol{\theta}|\mathbf{y})}{g^{IS}(\boldsymbol{\theta})}. \tag{5.1.7}$$

Then, for $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ randomly drawn from the distribution corresponding to $g^{IS}(\boldsymbol{\theta})$, under general conditions, we obtain

$$\frac{\sum_{i=1}^n h(\boldsymbol{\theta}^{(i)}) w(\boldsymbol{\theta}^{(i)})}{\sum_{i=1}^n w(\boldsymbol{\theta}^{(i)})} \xrightarrow{a.s.} \mathbb{E}[h(\boldsymbol{\theta}|\mathbf{y})], \tag{5.1.8}$$

because

$$\int h(\boldsymbol{\theta}) g^*(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int h(\boldsymbol{\theta}) \frac{g^*(\boldsymbol{\theta}|\mathbf{y})}{\int g^*(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}} d\boldsymbol{\theta} = \frac{\int \frac{h(\boldsymbol{\theta}) g^*(\boldsymbol{\theta}|\mathbf{y})}{g^{IS}(\boldsymbol{\theta})} g^{IS}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \frac{g^*(\boldsymbol{\theta}|\mathbf{y})}{g^{IS}(\boldsymbol{\theta})} g^{IS}(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Thus, a properly weighted set of draws from the source distribution can be used to approximate the expectation based on the posterior distribution.

Clearly, the advantage of this algorithm over acceptance sampling is that only a sample of size $n$ has to be drawn. Since the accuracy of the approximation of the expected value of interest improves with the sample size, it is an obvious advantage that all sample values can be used directly. It should be noted, however, that finding a good approximation $g^{IS}(\boldsymbol{\theta})$ to the posterior density may not be easy. The weights may vary dramatically which may undermine the convergence properties of the quantity in expression (5.1.8). Thus, importance sampling may not work well, if the posterior distribution is complicated and difficult to sample from. More recent proposals for generating samples from the posterior therefore involve Markov Chain Monte Carlo methods.

### Markov Chain Monte Carlo (MCMC) Methods

MCMC methods use a Markov chain algorithm to generate one long sample of the parameter vector, the distribution of which converges to the posterior distribution of the parameters, if the chain runs long enough. The draws of the parameter vector are not independent, however, but serially dependent. Laws of large numbers and central limit theorems for dependent samples can be invoked to justify the use of such samples for posterior inference. It should be understood, however, that the approximation precision for the posterior moment of interest tends to be lower, when the posterior is computed from a dependent sample rather than a random sample. Thus, longer samples may be necessary for precise inference. In fact, for the construction of approximately independent samples from the joint posterior, one may want to work only with every $m^{\text{th}}$ sampled vector, where $m$ is a sufficiently large number. Moreover, a large number of initial sample values (also known as transients or the burn-in sample) are usually discarded to ensure a close approximation of the posterior. Diagnostic

tools for assessing the convergence of the posterior to the target distribution are discussed in Chib (2001), among others. Despite their high computational demands, MCMC methods have become quite popular in recent years because they are often simpler and more computationally efficient than other sampling methods. There are several variations of this approach in the literature that differ in the way they choose the $i^{\text{th}}$ draw of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(i)}$, given $\boldsymbol{\theta}^{(i-1)}$.

MCMC methods have been known in the literature for a long time. They have become increasingly popular in Bayesian analysis after the publication of an influential article by Gelfand and Smith (1990). More recent introductory expositions are Chib and Greenberg (1995) and Geweke (2005). In the following, the very general Metropolis-Hastings algorithm and the popular Gibbs sampler are presented. The Gibbs sampler can only be used if the posterior can be broken down in a suitable way. It is very convenient and very efficient if the conditions for its use are satisfied.

**Metropolis-Hastings Algorithm** The Metropolis-Hastings algorithm is based on a conditional density, $\eta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$. For a given $\boldsymbol{\theta}^{(i-1)}$, a candidate $\boldsymbol{\theta}^{+}$ is drawn from the distribution corresponding to $\eta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$, and $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{+}$ is chosen with probability

$$\min\left\{\frac{g(\boldsymbol{\theta}^{+}|\mathbf{y})/\eta(\boldsymbol{\theta}^{+}|\boldsymbol{\theta}^{(i-1)})}{g(\boldsymbol{\theta}^{(i-1)}|\mathbf{y})/\eta(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{+})}, 1\right\}.$$

Otherwise $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$. In other words, a draw is accepted with probability one if it increases the posterior. Otherwise, it is accepted with a probability less than one, the precise value of which depends on how much lower the current posterior value is compared with the previous draw. It can be shown that this algorithm converges to the posterior under general conditions.

There are a number of practical questions related to the Metropolis-Hastings algorithm. Of prime importance is, of course, the choice of the conditional density $\eta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$. There are different proposals in the literature. A review can be found in Chib and Greenberg (1995), for example. Another important issue is the size of the burn-in sample. The question is how many initial draws should be discarded in the construction of the chain. Several criteria are discussed in Cowles and Carlin (1996), among others.

Clearly, one would not want to use this algorithm, if computationally simpler methods exist. In practice, the Metropolis-Hastings sampler is only used for nonstandard problems when there is no alternative. A case in point is the estimation of VAR parameters that cannot be estimated using the Gibbs sampler of either Waggoner and Zha (2003) or Villani (2009). For example, Giannone, Lenza, and Primiceri (2015) use the Metropolis-Hastings algorithm because of the complicated form of the kernel of the posterior distribution of their hyperparameters. Likewise many time-varying coefficient VAR models require Metropolis-Hastings sampling.

**Gibbs Sampler** If the posterior distribution of $\boldsymbol{\theta}$ is difficult to sample from, the problem often may be simplified by partitioning the parameter vector such that the conditional posterior of one of the subvectors, given the remaining elements, has a known, conventional distribution from which samples can be drawn easily. Consider the simplest case where $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}'_{(1)}, \boldsymbol{\theta}'_{(2)})'$ such that $g(\boldsymbol{\theta}_{(1)}|\boldsymbol{\theta}_{(2)}, \mathbf{y})$ and $g(\boldsymbol{\theta}_{(2)}|\boldsymbol{\theta}_{(1)}, \mathbf{y})$ correspond to known distributions that can be simulated easily. This case arises in VAR models, for example, if we partition the parameters into the VAR coefficients and the residual covariance matrix. In this situation, for given $\boldsymbol{\theta}_{(2)}^{(i-1)}$ we may choose $\boldsymbol{\theta}_{(1)}^{(i)}$ by drawing from $g(\boldsymbol{\theta}_{(1)}|\boldsymbol{\theta}_{(2)}^{(i-1)}, \mathbf{y})$ and choose $\boldsymbol{\theta}_{(2)}^{(i)}$ by drawing from $g(\boldsymbol{\theta}_{(2)}|\boldsymbol{\theta}_{(1)}^{(i)}, \mathbf{y})$, starting from some initial value $\boldsymbol{\theta}_{(2)}^{(0)}$. This simple algorithm converges to the joint posterior under general conditions.

The Gibbs sampler can be generalized in several ways. First, a generalization to the case where $\boldsymbol{\theta}$ needs to be partitioned into more than two subvectors to obtain standard conditional posterior distributions is straightforward. It should be noted, however, that the Gibbs sampler works very well only when the posterior distributions of the different subvectors are independent or at least not strongly correlated. This feature should be taken into account in grouping the parameters. Second, it may happen that some of the conditional posterior distributions are not from a standard family. As long as at least one conditional distribution is obtained that is easy to sample from, the Gibbs sampler is worth considering. The other conditional distributions may then be approximated by a Metropolis-Hastings step.

## 5.2 Priors for Reduced-Form VAR Parameters

In Bayesian analysis an important issue is the specification of the prior for the parameters of interest. Often a prior is specified that simplifies the posterior analysis. In particular, it is convenient to specify the prior such that the posterior is from a known family of distributions. If the prior is chosen such that the posterior has the same functional form as the likelihood function, it is called a conjugate prior. If the conjugate prior is from the same distribution family as the likelihood, then it is called a natural conjugate prior. For example, it is shown in Section 5.2.2 that if the likelihood is Gaussian and the prior is also normal, then the posterior again has a normal distribution. The Litterman or Minnesota prior is discussed as a special case. The natural conjugate prior for all the parameters is known as the Gaussian-inverse Wishart prior. It is considered in Section 5.2.4, whereas the more computationally convenient independent Gaussian-inverse Wishart prior is considered in Section 5.2.5.

When using priors from a known family of distributions, it is still necessary to specify at least some of the parameters of the prior distribution. This task is often made easier by imposing additional structure on the prior, reducing the number of parameters to be chosen to a handful of so-called hyperparameters. We discuss this idea in Section 5.2.1. In Sections 5.2.2-5.2.5 we consider popu-

lar choices for the prior distribution of the reduced-form VAR parameters and discuss how these distributions may be parameterized with the help of a small set of hyperparameters.

### 5.2.1 General Procedures for Choosing the Parameters of Prior Densities

In practice, it is often prohibitively difficult to fully specify the prior distribution. It is therefore common to impose additional structure on a given family of prior distributions, so as to reduce the number of parameters to be specified by the user to a small number of so-called hyperparameters. Let $\boldsymbol{\gamma}$ denote the vector of hyperparameters such that $g(\boldsymbol{\theta}) = g_{\boldsymbol{\gamma}}(\boldsymbol{\theta})$. Often $\boldsymbol{\gamma}$ is chosen such that the implied VAR model yields accurate out-of-sample forecasts (see Doan, Litterman, and Sims (1984), Litterman (1986)). Alternatively, Bańbura, Giannone, and Reichlin (2010) suggest to choose these hyperparameters based on the in-sample fit of the model.

Yet another proposal for choosing the hyperparameters was made by Giannone, Lenza, and Primiceri (2015). If one views the prior as being conditioned on the hyperparameters $\boldsymbol{\gamma}$, $g_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|\boldsymbol{\gamma})$, then the prior can be regarded as a hierarchical prior (see Koop (2003)). Suppose that the prior density for $\boldsymbol{\gamma}$ is $g(\boldsymbol{\gamma})$. Then the posterior is

$$g^*(\boldsymbol{\gamma}) \propto h(\mathbf{y}|\boldsymbol{\gamma})g(\boldsymbol{\gamma}),$$

where the sample density with respect to the hyperparameters is obtained as

$$h(\mathbf{y}|\boldsymbol{\gamma}) = \int f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma})g(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta}.$$

This expression is also known as the marginal likelihood because the parameters of interest, $\boldsymbol{\theta}$, are integrated out. If an improper uniform prior, $g(\boldsymbol{\gamma}) = \text{constant}$, is specified, then the posterior of the hyperparameters is equal to the marginal likelihood, and it makes sense to choose the hyperparameters such that $h(\mathbf{y}|\boldsymbol{\gamma})$ is maximized. Of course, strictly speaking, an improper prior does not qualify as a prior density because for an unbounded parameter space a constant prior does not integrate to one.

Giannone, Lenza, and Primiceri (2015) stress two advantages of this approach. First, under certain conditions maximizing the marginal likelihood results in optimal out-of-sample forecasts (also see Geweke (2001) and Geweke and Whiteman (2006)). Second, they point out that their procedure also can be justified from a frequentist point of view.

### 5.2.2 Normal Prior for the VAR Parameters for Given $\Sigma_u$

An early approach to specifying the prior for the VAR slope parameters takes the innovation variance $\Sigma_u$ as given. In practice, we may replace the unknown

$\Sigma_u$ by its LS or ML estimate (see, e.g., Litterman (1986)). Understanding this approach is also useful for expository purposes.

Consider a normally distributed $K$-dimensional VAR($p$) process $y_t$ of the form

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $u_t \sim \mathcal{N}(0, \Sigma_u)$ is a Gaussian white noise error term. For $t = 1, \ldots, T$, the model can be written in matrix notation as

$$Y = AZ + U, \tag{5.2.1}$$

where $Y \equiv [y_1, \ldots, y_T]$, $A \equiv [\nu, A_1, \ldots, A_p]$ and $Z \equiv [Z_0, \ldots, Z_{T-1}]$ with $Z_{t-1} \equiv (1, y'_{t-1}, \ldots, y'_{t-p})'$. Vectorizing the matrix expression (5.2.1), one obtains

$$\mathbf{y} = (Z' \otimes I_K)\boldsymbol{\alpha} + \mathbf{u}, \tag{5.2.2}$$

where $\boldsymbol{\alpha} \equiv \text{vec}(A)$, $\mathbf{y} \equiv \text{vec}(Y)$, and $\mathbf{u} \equiv \text{vec}(U)$. Next we discuss two alternative ways for expressing a normal prior for $\boldsymbol{\alpha}$ under the assumption that the white noise covariance matrix $\Sigma_u$ is known.

### Prior Distribution

Suppose the prior distribution of $\boldsymbol{\alpha}$ is multivariate normal with known mean $\boldsymbol{\alpha}^*$ and covariance matrix $V_{\boldsymbol{\alpha}}$. We write

$$\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\alpha}^*, V_{\boldsymbol{\alpha}})$$

and, hence, the prior density is

$$g(\boldsymbol{\alpha}) = \left(\frac{1}{2\pi}\right)^{K(Kp+1)/2} |V_{\boldsymbol{\alpha}}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)' V_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right]. \tag{5.2.3}$$

Combining this information with the sample information summarized in the Gaussian likelihood function,

$$\begin{aligned} l(\boldsymbol{\alpha}|\mathbf{y}) &= \left(\frac{1}{2\pi}\right)^{KT/2} |I_T \otimes \Sigma_u|^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}[\mathbf{y} - (Z' \otimes I_K)\boldsymbol{\alpha}]'(I_T \otimes \Sigma_u^{-1})[\mathbf{y} - (Z' \otimes I_K)\boldsymbol{\alpha}]\right], \end{aligned}$$

yields the posterior density

$$\begin{aligned} g(\boldsymbol{\alpha}|\mathbf{y}) &\propto g(\boldsymbol{\alpha}) l(\boldsymbol{\alpha}|\mathbf{y}) \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[[V_{\boldsymbol{\alpha}}^{-1/2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)]'[V_{\boldsymbol{\alpha}}^{-1/2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)] \\ &\qquad + \{(I_T \otimes \Sigma_u^{-1/2})\mathbf{y} - (Z' \otimes \Sigma_u^{-1/2})\boldsymbol{\alpha}\}' \\ &\qquad \times \{(I_T \otimes \Sigma_u^{-1/2})\mathbf{y} - (Z' \otimes \Sigma_u^{-1/2})\boldsymbol{\alpha}\}\Big]\Big\}. \end{aligned} \tag{5.2.4}$$

Defining

$$w = \begin{bmatrix} V_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{\alpha}^* \\ (I_T \otimes \Sigma_u^{-1/2})\mathbf{y} \end{bmatrix} \text{ and } W = \begin{bmatrix} V_{\boldsymbol{\alpha}}^{-1/2} \\ Z' \otimes \Sigma_u^{-1/2} \end{bmatrix},$$

the exponent in (5.2.4) can be rewritten as

$$\begin{aligned} &-\tfrac{1}{2}(w - W\boldsymbol{\alpha})'(w - W\boldsymbol{\alpha}) \\ &= -\tfrac{1}{2}[(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})'W'W(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) + (w - W\bar{\boldsymbol{\alpha}})'(w - W\bar{\boldsymbol{\alpha}})], \end{aligned} \tag{5.2.5}$$

where

$$\bar{\boldsymbol{\alpha}} = (W'W)^{-1}W'w = [V_{\boldsymbol{\alpha}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})]^{-1}[V_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}^* + (Z \otimes \Sigma_u^{-1})\mathbf{y}]. \tag{5.2.6}$$

Note that if the precision matrix $V_{\boldsymbol{\alpha}}^{-1}$ were zero, the expression for the posterior mean would simplify to that of the unrestricted LS estimator in equation (2.3.2) by noting that

$$(ZZ' \otimes \Sigma_u^{-1})^{-1}(Z \otimes \Sigma_u^{-1})\mathbf{y} = ((ZZ')^{-1}Z \otimes I_K)\text{vec}(Y) = \text{vec}(YZ'(ZZ')^{-1}).$$

The second term on the right-hand side of (5.2.5) does not contain $\boldsymbol{\alpha}$ and therefore is a constant. Thus,

$$g(\boldsymbol{\alpha}|\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})'\bar{\Sigma}_{\boldsymbol{\alpha}}^{-1}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})\right],$$

with $\bar{\boldsymbol{\alpha}}$ as given in (5.2.6) and

$$\bar{\Sigma}_{\boldsymbol{\alpha}} = (W'W)^{-1} = [V_{\boldsymbol{\alpha}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})]^{-1}. \tag{5.2.7}$$

The posterior density is easily recognizable as the density of a multivariate normal distribution with mean $\bar{\boldsymbol{\alpha}}$ and covariance matrix $\bar{\Sigma}_{\boldsymbol{\alpha}}$. In other words, the posterior distribution of $\boldsymbol{\alpha}$ is $\mathcal{N}(\bar{\boldsymbol{\alpha}}, \bar{\Sigma}_{\boldsymbol{\alpha}})$. This distribution may be used for inference about $\boldsymbol{\alpha}$. Sampling from this posterior distribution is particularly easy because the distribution is of a known form that is easy to draw from.

### An Alternative Representation of the Prior Distribution

The prior information on $\boldsymbol{\alpha}$ can equivalently be written as

$$C\boldsymbol{\alpha} = c + e \quad \text{with} \quad e \sim \mathcal{N}(0, I), \tag{5.2.8}$$

where $C$ is a fixed matrix and $c$ is a fixed vector. If $C$ is a $K(Kp+1) \times K(Kp+1)$ nonsingular matrix, this expression implies

$$\boldsymbol{\alpha} \sim \mathcal{N}(C^{-1}c, C^{-1}C^{-1\prime}), \tag{5.2.9}$$

which is just the normal prior specified earlier with mean $C^{-1}c$ and covariance matrix $(C'C)^{-1}$. Using (5.2.6), the posterior mean can be expressed as

$$\bar{\boldsymbol{\alpha}} = [C'C + (ZZ' \otimes \Sigma_u^{-1})]^{-1}[C'c + (Z \otimes \Sigma_u^{-1})\mathbf{y}]. \tag{5.2.10}$$

Note that under this prior specification there is no need to invert the potentially large covariance matrix $V_{\boldsymbol{\alpha}}$. Another practical advantage of this form is that it does not require the inversion of $C$. In fact, $C$ does not have to be a square matrix.

It is also worth mentioning that the estimator $\bar{\boldsymbol{\alpha}}$ in (5.2.10) is precisely the GLS estimator obtained from a regression model

$$\begin{bmatrix} \mathbf{y} \\ c \end{bmatrix} = \begin{bmatrix} Z' \otimes I_K \\ C \end{bmatrix} \boldsymbol{\alpha} + \varepsilon, \qquad \varepsilon \sim \left(0, \begin{bmatrix} I_T \otimes \Sigma_u & 0 \\ 0 & I \end{bmatrix}\right), \tag{5.2.11}$$

as pointed out by Theil (1963). Thus, in this case imposing the prior amounts to extending the sample. As in the frequentist setting, adding more observations tends to reduce the variance of the GLS estimator compared with the unrestricted LS estimator, and, hence, increases the estimation efficiency.

In order to make these concepts operational, the prior mean $\boldsymbol{\alpha}^*$ and the prior covariance matrix $V_{\boldsymbol{\alpha}}$ (or, equivalently, $C$ and $c$) must be specified. If no specific prior knowledge about the parameters is available, then one may use a diffuse Gaussian prior by choosing the prior variances very large, so that $V_{\boldsymbol{\alpha}}^{-1}$ (also known as the precision matrix) becomes very small. In that case the posterior mean is seen to be close to the LS estimator (see (5.2.6)).

The posterior mean can also be interpreted as a shrinkage estimator of $\boldsymbol{\alpha}$ with the degree of shrinkage determined by $V_{\boldsymbol{\alpha}}^{-1}$. If $V_{\boldsymbol{\alpha}}^{-1} = 0$ is used, the posterior mean reduces to the LS estimator. Of course, in that case Bayesian analysis looses its potential advantage of reducing the variance of the parameter estimates. In practice, there are a number of alternative ways of specifying nontrivial precision matrices. The next subsections discuss the most common choices.

### Practical Considerations

An important concern in practice is that priors may be inadvertently informative about the parameters of interest. Hence, there has been interest in priors that are uninformative. Ideally, assigning equal prior probability to all possible parameter values would avoid such distortions because in that case the prior density is just a constant that cancels from the posterior density due to the requirement that the posterior density integrates to one (see (5.1.1)). Unfortunately, there is no probability density that is constant over the entire Euclidean space. Any nonzero, positive constant would integrate to infinity over the whole space. Hence such a flat prior is not a proper prior. In other words, a truly uninformative prior for the VAR parameters does not exist.

An alternative is a Gaussian prior with $V_{\boldsymbol{\alpha}}^{-1}$ approaching zero. Such a prior is also known as a diffuse prior for the VAR parameters. Note, however, that such priors, although they do not seem to restrict the slope parameters much, may still be unintentionally informative for nonlinear functions of the parameters such as impulse responses. It should always be kept in mind that a prior that seems uninformative in one dimension tends to be informative in other dimensions.

Given the difficulty of agreeing on subjective priors, it has become common practice to choose the priors for the VAR parameters by convention. For example, priors may be chosen such that the associated models have certain desirable properties. Typically, in the VAR literature, priors have been chosen based on the forecast accuracy of the implied estimated models. One such prior that has been quite popular in applied work is discussed in the next subsection.

### 5.2.3 The Original Minnesota Prior

Litterman (1986) and Doan, Litterman, and Sims (1984) propose a specific Gaussian prior for the parameters of a VAR model that is often referred to as the Minnesota prior or the Litterman prior. The original proposal shrinks the VAR estimates toward a multivariate random walk model. This practice has been found useful in forecasting many persistent economic time series. The proposal is to specify the prior mean and covariance matrix as follows:

- In each equation, set the prior mean of the first lag of the dependent variable to one and set the prior mean of all other slope coefficients to zero. In other words, if the prior means were the true parameter values, each variable would follow a random walk.
- Set the prior variances of the intercept terms to infinity and the prior variance of the $ij^{\text{th}}$ element of $A_l$, denoted $a_{ij,l}$, to

$$v_{ij,l} = \begin{cases} (\lambda/l)^2 & \text{if } i = j, \\ (\lambda\theta\sigma_i/l\sigma_j)^2 & \text{if } i \neq j, \end{cases}$$

  where $\lambda$ is the prior standard deviation of $a_{ii,1}$, where $0 < \theta < 1$ controls the relative tightness of the prior variance in the other lags in a given equation compared to the own lags (with a smaller $\theta$ increasing the relative tightness of the other lags), and where $\sigma_i^2$ is the $i^{\text{th}}$ diagonal element of $\Sigma_u$.

For example, in a bivariate VAR(2) model with all slope parameters evaluated at their prior mean, we would have

$$\begin{aligned} y_{1t} = {} & \underset{(\infty)}{0} + \underset{(\lambda)}{1 \cdot y_{1,t-1}} + \underset{(\lambda\theta\sigma_1/\sigma_2)}{0 \cdot y_{2,t-1}} + \underset{(\lambda/2)}{0 \cdot y_{1,t-2}} + \underset{(\lambda\theta\sigma_1/2\sigma_2)}{0 \cdot y_{2,t-2}} + u_{1t}, \\ y_{2t} = {} & \underset{(\infty)}{0} + \underset{(\lambda\theta\sigma_2/\sigma_1)}{0 \cdot y_{1,t-1}} + \underset{(\lambda)}{1 \cdot y_{2,t-1}} + \underset{(\lambda\theta\sigma_2/2\sigma_1)}{0 \cdot y_{1,t-2}} + \underset{(\lambda/2)}{0 \cdot y_{2,t-2}} + u_{2t}. \end{aligned}$$

Here the numbers in parentheses are the prior standard deviations. Each of the two equations specifies a random walk prior mean for the dependent variables. The nonzero prior standard deviations reflect the uncertainty regarding the validity of that model. The standard deviations decline with increasing lag length because more recent lags are assumed to be more likely to have nonzero values. The standard deviations for the intercept terms are set to infinity to

capture our ignorance about the actual values of these parameters. Also, the prior distribution imposes independence across the parameters. Therefore, $V_{\boldsymbol{\alpha}}$ is diagonal. Its inverse is

$$V_{\boldsymbol{\alpha}}^{-1} = \begin{bmatrix} 0 & & & & & & & & & \\ & 0 & & & & & & & & \\ & & \frac{1}{\lambda^2} & & & & & 0 & & \\ & & & \frac{\sigma_1^2}{(\lambda\theta\sigma_2)^2} & & & & & & \\ & & & & \frac{\sigma_2^2}{(\lambda\theta\sigma_1)^2} & & & & & \\ & & & & & \frac{1}{\lambda^2} & & & & \\ & & & 0 & & & \frac{2^2}{\lambda^2} & & & \\ & & & & & & & \frac{2^2\sigma_1^2}{(\lambda\theta\sigma_2)^2} & & \\ & & & & & & & & \frac{2^2\sigma_2^2}{(\lambda\theta\sigma_1)^2} & \\ & & & & & & & & & \frac{2^2}{\lambda^2} \end{bmatrix},$$

where 0 is also substituted for the inverse of the infinite prior standard deviation of the intercepts.

In terms of expression (5.2.9), the prior for the slope parameters may equivalently be specified by choosing

$$C = \begin{bmatrix} 0 & 0 & \frac{1}{\lambda} & & & & & 0 & & \\ 0 & 0 & & \frac{\sigma_1}{\lambda\theta\sigma_2} & & & & & & \\ 0 & 0 & & & \frac{\sigma_2}{\lambda\theta\sigma_1} & & & & & \\ 0 & 0 & & & & \frac{1}{\lambda} & & & & \\ 0 & 0 & & 0 & & & \frac{2}{\lambda} & & & \\ 0 & 0 & & & & & & \frac{2\sigma_1}{\lambda\theta\sigma_2} & & \\ 0 & 0 & & & & & & & \frac{2\sigma_2}{\lambda\theta\sigma_1} & \\ 0 & 0 & & & & & & & & \frac{2}{\lambda} \end{bmatrix} \quad \text{and} \quad c = \frac{1}{\lambda} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where $c = \frac{1}{\lambda}(\text{vec}(I_2)', 0_{1\times 4})'$ and $C$ is an $8 \times 10$ matrix with two leading columns of zeros and the square roots of the reciprocals of the diagonal elements of $V_{\boldsymbol{\alpha}}$ on the main diagonal of the remaining $8 \times 8$ block. Note that in this alternative representation no prior is specified for the intercept.

### Practical Issues

The crucial advantage of the Minnesota prior is that it reduces the problem of specifying a high-dimensional prior distribution to one of selecting two parameters by imposing additional structure on the prior. In specifying the Minnesota prior, the user has to choose only the two hyperparameters $\lambda$ and $\theta$. The parameter $\lambda$ controls the overall prior variance of all VAR coefficients, whereas $\theta$ controls the tightness of the variances of the coefficients of lagged variables other than the dependent variable in a given equation. Roughly speaking, $\theta$ specifies the fraction of the prior standard deviation $\lambda$ attached to the coefficients of other lagged variables. A value of $\theta$ close to one implies that all coefficients of

lag 1 have about the same prior variance except for a scaling factor intended to capture differences in the variability of each variable. For example, Litterman (1986) finds that $\theta = 0.3$ and $\lambda = 0.2$ works well when using a Bayesian VAR model for forecasting U.S. macroeconomic variables. For given $\theta$, the shrinkage is determined by $\lambda$. Therefore $\lambda$ is often referred to as the shrinkage parameter. A smaller $\lambda$ implies a stronger shrinkage towards the prior mean.[2]

There are also a number of other practical problems that have to be addressed in working with the Minnesota prior. For example, the assumption of a known $\Sigma_u$ is unrealistic in practice. In a strict Bayesian approach, a prior pdf has to be specified for the elements of $\Sigma_u$. This possibility is discussed in the next section. A simple alternative is to replace $\Sigma_u$ by its LS estimator or its ML estimator,

$$\widetilde{\Sigma}_u = Y(I_T - Z'(ZZ')^{-1}Z)Y'/T,$$

based on the full sample $T$.

Given that the computation of $\bar{\boldsymbol{\alpha}}$ requires the inversion of the potentially rather large matrix $V_{\boldsymbol{\alpha}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})$, Bayesian estimation based on the Minnesota prior has often been performed for each of the $K$ equations of the system separately. In that case,

$$\bar{a}_k = [V_k^{-1} + \sigma_k^{-2} ZZ']^{-1}(V_k^{-1} a_k^* + \sigma_k^{-2} Z y_{(k)})$$

is the estimator for the parameters $a_k$ of the $k^{\text{th}}$ equation, where $y_{(k)} \equiv (y_{k1}, \ldots, y_{kT})'$. In other words, $a_k'$ is the $k^{\text{th}}$ row of $A = [\nu, A_1, \ldots, A_p]$. Here $V_k$ denotes the prior covariance matrix of $a_k$ and $a_k^*$ is its prior mean. The unknown $\sigma_k^2$ may be replaced by the $k^{\text{th}}$ diagonal element of $\widetilde{\Sigma}_u$.

Shrinking the parameters of models of macroeconomic variables toward a random walk is only plausible for economic time series with stochastic trends. When working with stationary variables, the VAR parameters may be shrunk towards zero instead, as proposed by Lütkepohl (1991, Section 5.4) and implemented, for example, in Baumeister and Kilian (2012). In that case, mean-adjusting the data before fitting a VAR model may be useful to avoid having to specify a prior for the intercept term. Villani (2009) makes the case that a mean-adjusted model form may have advantages if prior information is to be imposed on the steady-state of the variables.

Many other modifications of the Minnesota prior have been proposed, depending on the needs of the researcher (see, for example, Kadiyala and Karlsson (1997), Sims and Zha (1998), Waggoner and Zha (2003), Bańbura, Giannone, and Reichlin (2010), Karlsson (2013)). The main advantage of the Minnesota prior is that it results in a simple analytically tractable normal posterior distribution, which explains why it has remained popular over the years, despite some disadvantages. There are alternatives, however. For example, Sims and

[2] Rather than selecting $\lambda$ and $\theta$ based on rules of thumb, one could instead treat these hyperparameters as endogenous and set them to the values that maximize the marginal likelihood, as proposed by Giannone, Lenza, and Primiceri (2015).

Zha (1998), building on the representation (5.2.11), proposed imposing prior restrictions on the structural VAR parameters rather than the reduced-form VAR parameters (see Chapter 12).

### Cointegration and Near Unit Roots

One potential disadvantage of the Litterman prior is that even if all variables have stochastic trends, it is not clear that shrinking towards a multivariate random walk as in the Litterman prior is optimal because there may be cointegration between the variables (see Chapter 3). This approach can be rationalized on the grounds that exact unit roots are events of probability zero in standard Bayesian analysis. Hence, there is no reason to pay special attention to cointegration relations from a Bayesian point of view. Nevertheless the importance of the concept of cointegration and of VECMs in frequentist analysis has prompted some Bayesians to develop alternative priors that explicitly refer to the parameters of the VECM form of the VAR.

Consider expressing the VAR model in the VEC representation introduced in Chapter 3,

$$\Delta y_t = \nu + \mathbf{\Pi} y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where $\mathbf{\Pi} = -(I_K - A_1 - \cdots - A_p)$. A prior that shrinks $\mathbf{\Pi}$ to zero in the limit reduces the VECM to a VAR model in differences. Such a prior may also be suitable if there are near unit roots (see Chapter 3). A specific prior of this type, referred to as the sum-of-coefficients prior in Doan, Litterman, and Sims (1984), is implemented by augmenting the observations similar to expression (5.2.11). For the sum-of-coefficients prior we augment $Y$ and $Z$ in model (5.2.1) by

$$Y_* = \text{diag}(\mu_1, \ldots, \mu_K)/\tau$$

and

$$Z_* = \begin{bmatrix} & 0_{1\times K} \\ \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} & \otimes \text{diag}(\mu_1, \ldots, \mu_K)/\tau \end{bmatrix},$$

respectively, and consider the vectorized version of the model

$$[Y,\ Y_*] = A[Z,\ Z_*] + [U,\ \mathcal{E}]$$

instead of (5.2.11). If the shrinkage parameter $\tau$ is small, the prior shrinks the posterior mean of $\mathbf{\Pi}$ to zero. If $\tau$ is very large, the posterior mean of $\mathbf{\Pi}$ is close to the LS estimator. The $\mu_k$ are supposed to capture the potentially different levels of the $y_{kt}$. In practice, the sample mean is used as a proxy for $\mu_k$, although that approach is not, strictly speaking, Bayesian (see Bańbura,

Giannone, and Reichlin (2010)). Of course, shrinkage may also be imposed on the $\mathbf{\Gamma}_i$ parameters. Such a prior can be imposed by adding further dummy observations, as described earlier, based on expression (5.2.11). For further motivation and discussion of the sum-of-coefficients prior the reader is referred to Sims and Zha (1998).

Bayesian analysis of VECMs is discussed, for example, in Kleibergen and van Dijk (1994), Kleibergen and Paap (2002), Strachan (2003), and Strachan and Inder (2004). Surveys with many additional references are Koop, Strachan, van Dijk, and Villani (2005) and Karlsson (2013).

### An Empirical Illustration

To illustrate the use of the Minnesota prior, we consider a model including quarterly U.S. GDP deflator inflation ($\Delta\pi_t$), the seasonally adjusted unemployment rate ($ur_t$) and the yield on the 3-month Treasury bills ($r_t$) for the period 1953q1 - 2006q3, as used by Koop and Korobilis (2009).[3] The time series are plotted in Figure 5.1. All three series exhibit considerable persistence, so using the Minnesota prior with shrinkage to a random walk makes sense. To allow for the possibility that the time series have no unit roots, we alternatively consider shrinking all parameters to zero by means of a white noise prior mean.

We consider a VAR(4) model with intercept and impose a conventional Minnesota prior. Following the example of some earlier studies, the unknown error variances are replaced by estimates obtained from fitting univariate AR(4) models to the individual model variables. No estimates of the error covariances are required for the specification of the prior. In constructing the posterior, the error covariance matrix is treated as known and replaced by its LS estimate.

Figure 5.2 illustrates the impact of alternative specifications of the Minnesota prior on the posterior density of selected structural impulse responses (see Chapter 4). The structural responses are obtained by imposing a recursive structure on the impact multiplier matrix with the variables ordered as $y_t = (\Delta\pi_t, ur_t, r_t)'$. In particular, the interest rate is ordered last, so the shock to the interest rate equation may be interpreted as a monetary policy shock with no contemporaneous effect on inflation and unemployment (see Chapter 9). Figure 5.2 focuses on the responses of inflation to an unexpected increase in the interest rate (which represents a contractionary monetary policy shock). Following common practice, the figure plots the 10%, 50%, and 90% quantiles of the draws from the posterior distributions of the individual impulse response coefficients.

Figure 5.2 illustrates how the choice of the prior affects the structural impulse response estimates. The random walk prior mean is used for generating the panels on the left, and the white noise prior mean is used for the panels on the right. Obviously it makes a difference which prior mean specification is used,

[3] The data are available on Gary Koop's webpage at http://personal.strath.ac.uk/gary.koop/bayes_matlab_code_by_koop_and_korobilis.html. This webpage in addition provides a set of Matlab code for BVAR analysis, modified versions of which have been used to produce the results below.

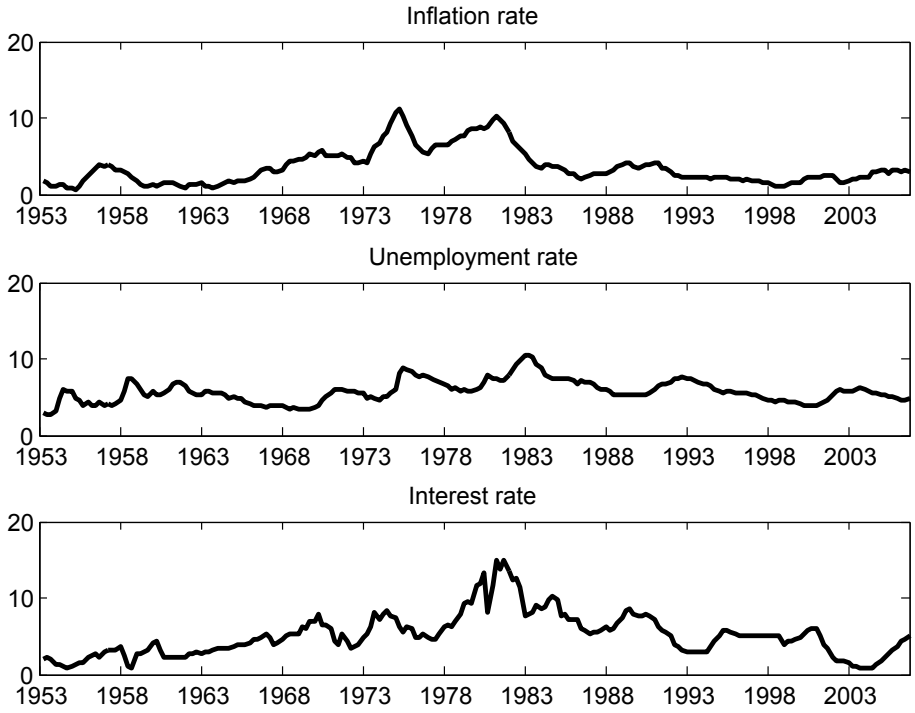


Figure 5.1: Quarterly U.S. inflation, unemployment rate, and interest rate for 1953q1-2006q3.

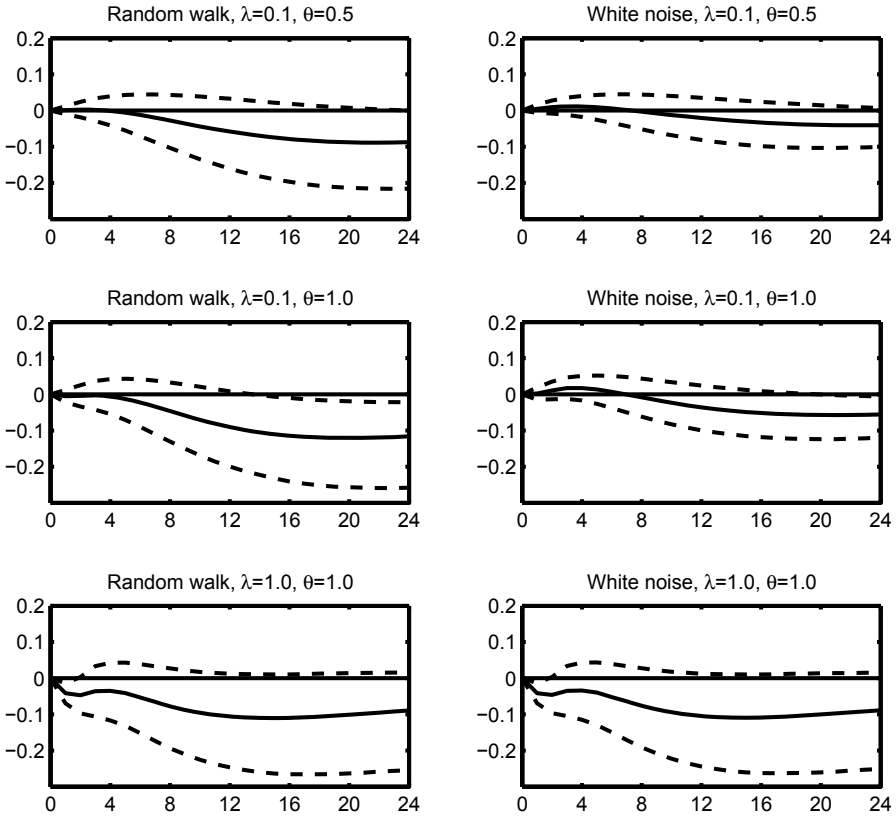


Figure 5.2: Simulated quantiles of inflation responses to monetary policy shocks for different Minnesota priors (pointwise median and 10% and 90% quantiles of the posterior distribution computed from 10,000 draws). The prior standard deviation for the constant terms is set to 1000 throughout; the standard deviations of the innovations are replaced by estimates obtained from fitting univariate AR(4) models to each model variable.

but the choice of the hyperparameters $\lambda$ and $\theta$, which control the prior variance, also is important. A tight prior variance about the random walk prior mean results in smaller bands around the pointwise medians than a prior with larger $\lambda$ and $\theta$ parameters. Shrinkage to a white noise process makes an even larger difference, unless the prior variance is large ($\lambda = 1.0, \theta = 1.0$). In the first two panels on the right, inflation increases in response to a contractionary monetary policy shock. This phenomenon is usually referred to as the price puzzle and has been observed in many structural VAR studies. It is often attributed to omitted variables. This example illustrates that this puzzle can also be an artifact of the choice of the prior. In this example, shrinking the slopes to a white noise mean implies a very different posterior than shrinking to the random walk mean. More generally, one could allow for different prior means in each equation of the VAR model, depending on the order of integration of the individual variables.

### 5.2.4 The Natural Conjugate Gaussian-Inverse Wishart Prior

So far we have replaced the unknown $\Sigma_u$ by an estimate. This approach can be improved upon by specifying a prior not only for the slope parameters, but also for $\Sigma_u$. The prior distribution of the innovation covariance matrix must satisfy the constraint that $\Sigma_u$ is positive definite. This section discusses such a prior and derives the corresponding posterior.

Let $x_i \sim \mathcal{N}(0, \Sigma_x)$, $i = 1, \ldots, n$, be $K$-dimensional independent, identically distributed normal random vectors. Then the distribution of $\sum_{i=1}^{n} x_i x_i'$ is called a ($K$-dimensional) Wishart distribution with parameters $\Sigma_x$ and $n$. We write

$$\sum_{i=1}^{n} x_i x_i' \sim \mathcal{W}_K(\Sigma_x, n). \tag{5.2.12}$$

For univariate standard normal random variables $x_i$, $\sum_{i=1}^{n} x_i^2$ has a $\chi^2(n)$ distribution, which illustrates that the Wishart distribution can be viewed as a multivariate generalization of a $\chi^2$ distribution with $n$ degrees of freedom. If $\Omega \sim \mathcal{W}_K(\Sigma, n)$, then the distribution of $\Omega^{-1}$ depends on $\Sigma^{-1}$ and $n$ only. The latter distribution is an inverted Wishart or inverse Wishart distribution with parameters $\Sigma^{-1}$ and $n$, and is abbreviated as

$$\Omega^{-1} \sim \mathcal{IW}_K(\Sigma^{-1}, n).$$

Suppose that for the VAR($p$) model (5.2.1) with Gaussian innovations, $u_t \sim \mathcal{N}(0, \Sigma_u)$, we specify the priors

$$\boldsymbol{\alpha} | \Sigma_u \sim \mathcal{N}(\boldsymbol{\alpha}^*, V_{\boldsymbol{\alpha}} = V \otimes \Sigma_u) \tag{5.2.13}$$

and

$$\Sigma_u \sim \mathcal{IW}_K(S_*, n), \tag{5.2.14}$$

such that $\Sigma_u^{-1} \sim \mathcal{W}_K(S_*^{-1}, n)$.[4] Expressing the prior covariance matrix of $\boldsymbol{\alpha}$ as a Kronecker product $V \otimes \Sigma_u$ simplifies the posterior distribution, and we obtain a Gaussian-inverse Wishart distribution

$$\boldsymbol{\alpha}|\Sigma_u, \mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\alpha}}, \bar{\Sigma}_{\boldsymbol{\alpha}}), \quad \Sigma_u|\mathbf{y} \sim \mathcal{IW}_K(S, \tau), \tag{5.2.15}$$

where (5.2.7) implies that

$$\bar{\Sigma}_{\boldsymbol{\alpha}} = [(V^{-1} \otimes \Sigma_u^{-1}) + (ZZ' \otimes \Sigma_u^{-1})]^{-1} = (V^{-1} + ZZ')^{-1} \otimes \Sigma_u.$$

Substituting $V \otimes \Sigma_u$ for $V_{\boldsymbol{\alpha}}$ into expression (5.2.6) yields

$$\begin{aligned}
\bar{\boldsymbol{\alpha}} &= [(V^{-1} \otimes \Sigma_u^{-1}) + (ZZ' \otimes \Sigma_u^{-1})]^{-1}[(V^{-1} \otimes \Sigma_u^{-1})\boldsymbol{\alpha}^* + (Z \otimes \Sigma_u^{-1})\mathbf{y}] \\
&= [(V^{-1} + ZZ')^{-1} \otimes \Sigma_u][V^{-1} \otimes \Sigma_u^{-1},\ Z \otimes \Sigma_u^{-1}] \begin{bmatrix} \boldsymbol{\alpha}^* \\ \mathbf{y} \end{bmatrix} \\
&= \left((V^{-1} + ZZ')^{-1}[V^{-1},\ Z] \otimes I_K\right) \mathrm{vec}[A^*,\ Y].
\end{aligned}$$

Hence, the posterior mean can be written in matrix notation as

$$\bar{A} = (A^* V^{-1} + YZ')(V^{-1} + ZZ')^{-1}. \tag{5.2.16}$$

Equation (5.2.16) illustrates why Bayesian estimation methods may be used even when the number of regressors exceeds the sample size. In this case, $ZZ'$ is not invertible and, hence, LS estimation is infeasible. In contrast, Bayesian estimation remains feasible. Adding the invertible precision matrix $V^{-1}$ to $ZZ'$ allows us to invert the sum $V^{-1} + ZZ'$, as required for the construction of the posterior mean $\bar{A}$. Of course, the solution $\bar{A}$ in this case heavily depends on the choice of $V^{-1}$.

The parameters of the inverse Wishart distribution in (5.2.15) are

$$S = T\widetilde{\Sigma}_u + S_* + \hat{A}ZZ'\hat{A}' + A^* V^{-1} A^{*\prime} - \bar{A}(V^{-1} + ZZ')\bar{A}' \tag{5.2.17}$$

and $\tau = T + n$ (see Koop and Korobilis (2009) or Uhlig (1994, 2005)). Here $A^*$ and $\bar{A}$ are $K \times (Kp+1)$ matrices such that $\boldsymbol{\alpha}^* = \mathrm{vec}(A^*)$ and $\bar{\boldsymbol{\alpha}} = \mathrm{vec}(\bar{A})$, $\hat{A} = YZ'(ZZ')^{-1}$, and $\widetilde{\Sigma}_u = (Y - \hat{A}Z)(Y - \hat{A}Z)'/T$. Since the posterior is from the same distributional family as the likelihood function, the prior (5.2.13)-(5.2.14) is a conjugate prior. Given that the prior is also from the same distributional family as the likelihood, it is more specifically a natural conjugate prior.

The advantage of using a natural conjugate prior is that a known posterior distribution is obtained that can be used for inference on $\boldsymbol{\alpha}$ without additional simulations. In fact, the marginal posterior distribution of $\boldsymbol{\alpha}$ is a multivariate $t$-distribution with $\tau = T + n$ degrees of freedom, mean $\bar{\boldsymbol{\alpha}}$ and covariance matrix

$$\Sigma_{\boldsymbol{\alpha}|\mathbf{y}} = \frac{1}{\tau - K - 1}((V^{-1} + ZZ')^{-1} \otimes S) \tag{5.2.18}$$

[4] Sometimes in the literature the inverse Wishart distribution with parameters $S_*^{-1}$ and $n$ is denoted as $\Sigma_u \sim \mathcal{IW}_K(nS_*, n)$ such that $\Sigma_u^{-1} \sim \mathcal{W}_K(S_*^{-1}/n, n)$ (see, e.g., Uhlig (2005)). This difference in notation leaves the definition of the inverse Wishart distribution unaffected.

(e.g., Koop and Korobilis (2009)).

The Gaussian-inverse Wishart posterior distribution can be used as a basis for inference on functions of $\boldsymbol{\alpha}$ and $\Sigma_u$ such as structural impulse responses. If the structural VAR model is just-identified, the structural parameters will be nonlinear functions of the reduced-form parameters considered thus far. In that case, the posterior distribution of the structural impulse responses may be simulated by drawing from the joint posterior distribution of the reduced-form parameters and substituting these draws into the formula of the structural impulse responses. In practice, such draws may be generated by first drawing $\tau$ independent vectors $x_i$, $i = 1, \ldots, \tau$, from a $K$-dimensional normal distribution, $\mathcal{N}(0, S^{-1})$, and then conditioning on $\sum_{i=1}^{\tau} x_i x_i'$ for $\Sigma_u^{-1}$ in simulating a draw from the posterior of $\boldsymbol{\alpha}$ in (5.2.15). Note that, if the prior parameters are specified, $S$ in (5.2.17) does not involve any unknown quantities, and hence can be computed when the prior is specified and the data are available. This means that when draws from the posterior are required, they can be obtained quickly and easily. Of course, the choice of the prior determines to some extent the posterior. If we choose $V^{-1} = 0$, for example, (5.2.16) implies that the posterior mean reduces to the LS estimator. More detailed discussion of this case can be found in Chapter 13.

### An Empirical Illustration

We illustrate the use of the Gaussian-inverse Wishart prior based on the same empirical example already employed for the Minnesota prior. Figure 5.3 plots the inflation responses to a contractionary monetary policy shock for different specifications of the prior. Our analysis is based on one of many possible configurations of this prior. The prior mean of the VAR parameters is either a random walk or white noise. The covariance matrix $V$ is specified as $\eta I$, where $\eta$ is a prespecified constant. By varying $\eta$ we can examine the effect of changing the prior variances. The hyperparameter $\eta$ takes the place of $\lambda$ in the Minnesota prior and determines the amount of shrinkage. A smaller value of $\eta$ implies a smaller prior variance and, hence, more shrinkage whereas a larger $\eta$ implies less shrinkage of the parameter estimates. The prior for $\Sigma_u$ is chosen arbitrarily to be $S_* = I_K$ and $n = K + 1$.

Figure 5.3 illustrates that a small value of $\eta$ (implying a tight prior variance) has a substantial effect on the estimated impulse responses. In fact, for a very small $\eta = 0.01$, inflation is estimated to respond positively to an interest rate shock. Thus, there is again a price puzzle. If $\eta$ is increased, the posterior mean approaches the LS estimator as expected because all terms involving $V^{-1}$ in the formulas for the posterior moments disappear if $\eta \to \infty$ and $V^{-1} \to 0$ .

### Extensions of a Gaussian-Inverse Wishart Prior

Giannone, Lenza, and Primiceri (2015) show that the Gaussian-inverse Wishart prior (5.2.13)-(5.2.14) implies a closed-form expression for the marginal likelihood that is easy to maximize with respect to the hyperparameters. They pro-

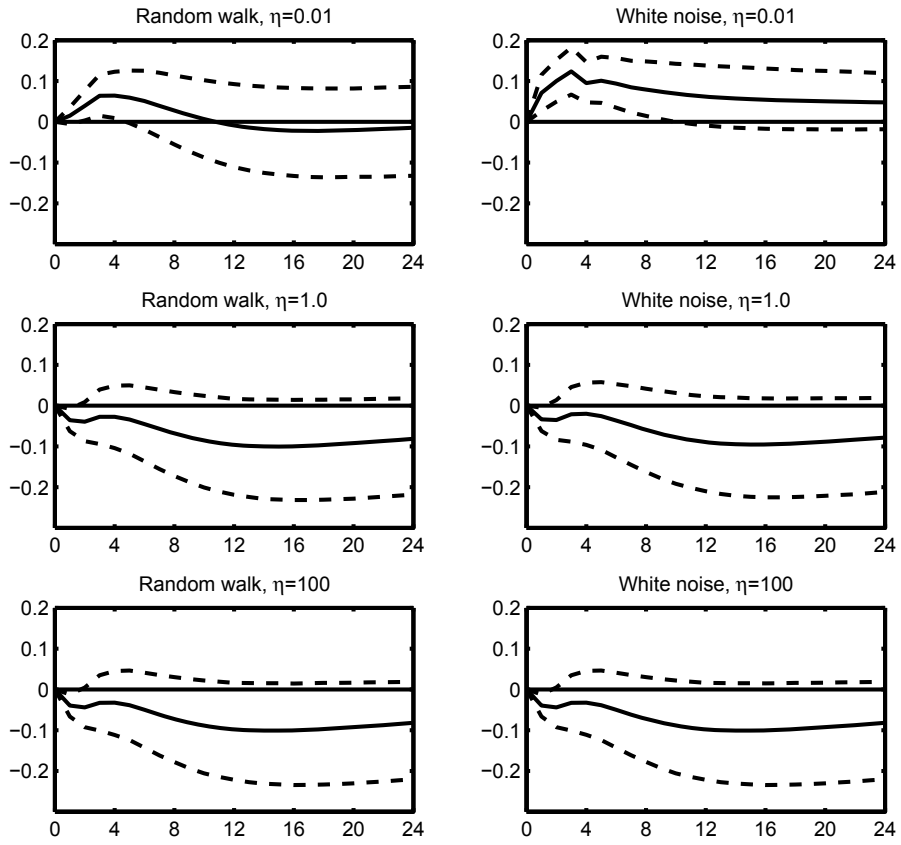


Figure 5.3: Simulated quantiles of inflation responses to monetary policy shocks for different Gaussian-inverse Wishart priors (pointwise median and 10% and 90% quantiles of the posterior distribution computed from 10,000 draws, given the prior parameters $V = \eta I$, $S_* = I_K$, $n = K + 1 = 4$).

vide evidence that choosing the hyperparameters in this way results in models that tend to forecast more accurately and imply economically plausible impulse responses. This finding is based on quarterly VAR models. For monthly VAR models, there is evidence that Bayesian shrinkage estimation along these lines may actually worsen the accuracy of the forecasts compared with unrestricted LS estimation, as illustrated in Baumeister and Kilian (2012). Intuitively, this difference arises because forecasts from quarterly models tend to be smoother than forecasts from monthly models. Thus priors that smooth the dynamics of the VAR model are less likely to oversmooth in quarterly models.

Alternatively, one could also specify a proper prior for the hyperparameters and determine the posterior. Then a Gibbs sampler could be used for simulating draws from the joint posterior of $\boldsymbol{\alpha}$, $\Sigma_u$, and the hyperparameters. Because, for a given set of hyperparameters, the posterior of $\boldsymbol{\alpha}$ and $\Sigma_u$ is from a known distribution, it is easy to draw from the conditional posterior of the latter parameters (see Giannone, Lenza, and Primiceri (2015) for details).

**Drawbacks of the Gaussian-Inverse Wishart Prior**

A drawback of the natural conjugate prior is that it hinges on the regression matrix being $Z' \otimes I_K$. In other words, it requires that the regressors in each of the $K$ equations of a $K$-dimensional VAR model be the same. In some applications, this condition is problematic because one may want to drop lags of some variable in one equation, but not in others (see Chapter 2). Even though the Bayesian approach can be viewed as an alternative to subset VAR models because it reduces the parameter variability by smoothing, there can be arguments for eliminating lags of one variable from some equation even in Bayesian analysis. This situation arises, for example, when one variable is specified to be Granger non-causal for some other variable. In such a case the posterior will no longer be Gaussian-inverse Wishart and we have to revert to simulation methods for generating draws from the posterior.

Another undesirable feature of the natural conjugate prior is the multiplicative covariance structure $V_{\boldsymbol{\alpha}} = V \otimes \Sigma_u$. Notice that the Minnesota prior has a more general covariance that is not encompassed by this expression unless $\theta = 1$. Thus, the Kronecker product form in (5.2.13) is clearly restrictive. It implies that the prior covariance matrices for the lags of the $k^{\text{th}}$ variable in different equations are proportional. More precisely, denoting the $ij^{\text{th}}$ element of $\Sigma_u$ by $\sigma_{ij}$, the lags of variable $y_{kt}$ in equations $i$ and $j$ have prior covariances $\sigma_{ik}V$ and $\sigma_{jk}V$, respectively. Put differently, they differ by a multiplicative factor.

If these features are deemed too restrictive, one may, of course, specify a Gaussian-inverse Wishart prior of a more general type. This approach entails the loss of the known closed form distribution of the posterior, however, and, hence, makes it necessary to use computationally more costly simulation techniques for inference.

In the next subsection a prior is discussed that is less restrictive than the natural conjugate prior and still makes it easy to sample from the posterior because it provides a natural basis for employing a Gibbs sampler.

### 5.2.5 The Independent Gaussian-Inverse Wishart Prior

In the natural conjugate Gaussian-inverse Wishart prior the distributions of the parameters $\boldsymbol{\alpha}$ and $\Sigma_u$ are not independent. The prior for $\boldsymbol{\alpha}$ in (5.2.13) obviously depends on $\Sigma_u$. Alternatively, one may explicitly impose independence of the priors of $\boldsymbol{\alpha}$ and $\Sigma_u$ by specifying the joint prior pdf to be of the form

$$g(\boldsymbol{\alpha}, \Sigma_u) = g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) g_{\Sigma_u}(\Sigma_u). \tag{5.2.19}$$

This approach facilitates the use of a Gibbs sampler. The prior resulting from the marginal priors

$$\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\alpha}^*, V_{\boldsymbol{\alpha}}) \tag{5.2.20}$$

and

$$\Sigma_u \sim \mathcal{IW}_K(S_*, n), \tag{5.2.21}$$

is called independent Gaussian-inverse Wishart because of the independence assumption for the marginal prior distributions of $\boldsymbol{\alpha}$ and $\Sigma_u$.

Assuming a Gaussian VAR process to start with, we know from Section 5.2.1 that the posterior of $\boldsymbol{\alpha}$ given $\Sigma_u$ is normal,

$$\boldsymbol{\alpha}|\Sigma_u, \mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\alpha}}, \bar{\Sigma}_{\boldsymbol{\alpha}}), \tag{5.2.22}$$

where

$$\begin{aligned} \bar{\boldsymbol{\alpha}} &= [V_{\boldsymbol{\alpha}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})]^{-1} [V_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}^* + (Z \otimes \Sigma_u^{-1})\mathbf{y}] \\ &= \left[V_{\boldsymbol{\alpha}}^{-1} + \sum_{t=1}^{T} \mathbf{Z}_t' \Sigma_u^{-1} \mathbf{Z}_t\right]^{-1} \left[V_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}^* + \sum_{t=1}^{T} \mathbf{Z}_t' \Sigma_u^{-1} y_t\right]. \end{aligned} \tag{5.2.23}$$

and

$$\bar{\Sigma}_{\boldsymbol{\alpha}} = [V_{\boldsymbol{\alpha}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})]^{-1} = \left[V_{\boldsymbol{\alpha}}^{-1} + \sum_{t=1}^{T} \mathbf{Z}_t' \Sigma_u^{-1} \mathbf{Z}_t\right]^{-1}. \tag{5.2.24}$$

Here $\mathbf{Z}_t$ is $Z_t' \otimes I_K$ if the same lagged variables appear in all equations. If some lags are removed from some of the equations, these expressions may still be used after removing the corresponding element from $\boldsymbol{\alpha}$ and redefining the rows of $\mathbf{Z}_t$ accordingly. Thus, the expressions in terms of $\mathbf{Z}_t$ are, in fact, more general than the expressions involving $Z$.

The conditional posterior of $\Sigma_u$, given $\boldsymbol{\alpha}$, is an inverse Wishart distribution,

$$\Sigma_u|\boldsymbol{\alpha}, \mathbf{y} \sim \mathcal{IW}_K(S, \tau) \tag{5.2.25}$$

with

$$S = S_* + \sum_{t=1}^{T} (y_t - \mathbf{Z}_t\boldsymbol{\alpha})(y_t - \mathbf{Z}_t\boldsymbol{\alpha})'$$

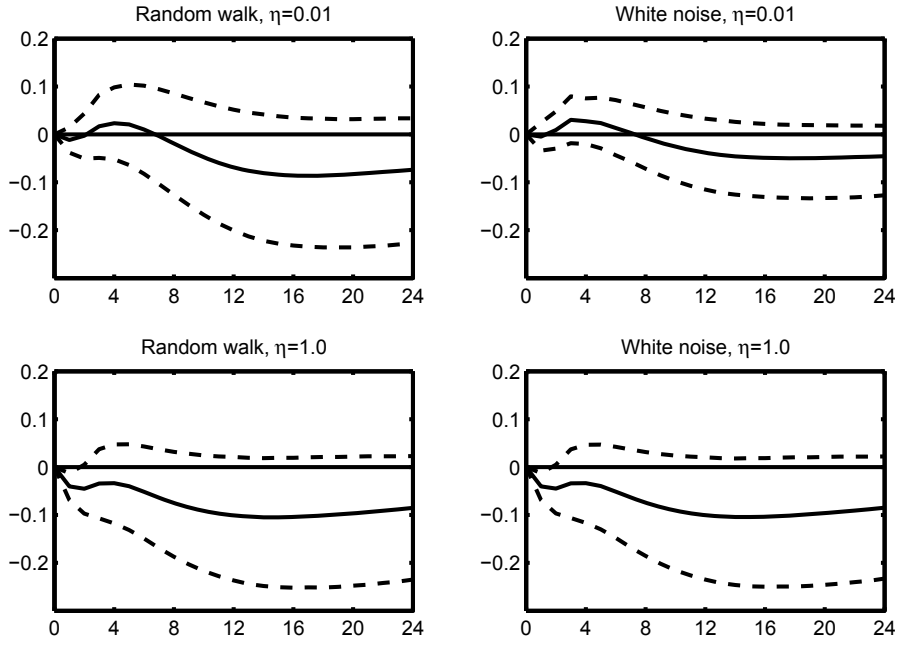


Figure 5.4: Simulated quantiles of inflation responses to monetary policy shocks for different independent Gaussian-inverse Wishart priors (pointwise median and 10% and 90% quantiles of the posterior distribution computed from 10,000 draws, given the prior parameters $V_{\boldsymbol{\alpha}} = \eta I$, $S_* = I_K$, $n = K + 1 = 4$).

and

$$\tau = T + n.$$

Both conditional posteriors are from known distribution families and therefore easy to sample from, facilitating the use of the Gibbs sampler for drawing samples from the joint posterior distribution.

### Empirical Illustration

We now reexamine our empirical example using the independent Gaussian-inverse Wishart prior. The responses of inflation to a contractionary interest rate shock are depicted in Figure 5.4. They are computed by using a Gibbs sampler to draw from the posterior. The $i^{\text{th}}$ iteration is based on the conditional distributions

$$\boldsymbol{\alpha}|\Sigma_u^{(i-1)}, \mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\alpha}}^{(i-1)}, \bar{\Sigma}_{\boldsymbol{\alpha}}^{(i-1)}) \quad \text{and} \quad \Sigma_u|\boldsymbol{\alpha}^{(i)}, \mathbf{y} \sim \mathcal{IW}_K(S^{(i)}, \tau),$$

where

$$\bar{\boldsymbol{\alpha}}^{(i-1)} = [V_{\boldsymbol{\alpha}}^{-1} + ZZ' \otimes (\Sigma_u^{(i-1)})^{-1}]^{-1}[V_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}^* + (Z \otimes (\Sigma_u^{(i-1)})^{-1})\mathbf{y}],$$

$$\bar{\Sigma}_{\boldsymbol{\alpha}}^{(i-1)} = [V_{\boldsymbol{\alpha}}^{-1} + ZZ' \otimes (\Sigma_u^{(i-1)})^{-1}]^{-1},$$

and

$$S^{(i)} = S_* + \sum_{t=1}^{T} (y_t - \mathbf{Z}_t \boldsymbol{\alpha}^{(i)})(y_t - \mathbf{Z}_t \boldsymbol{\alpha}^{(i)})'.$$

A burn-in sample of 20,000 draws is discarded, and then 10,000 draws are computed to determine the quantiles of the pointwise distributions of the structural impulse responses. We use parameter settings for $\boldsymbol{\alpha}^*$, $V_{\boldsymbol{\alpha}}$, $S_*$, and $n$ in the prior distributions similar to those for the natural conjugate Gaussian-inverse Wishart prior. Figure 5.4 illustrates once again that the posterior and, hence, the estimated impulse responses depend on the prior. The posterior quantiles look a little different from those for the other priors if the hyperparameter $\eta$ is small and, hence, the prior is tight, whereas they are similar to those obtained with other priors when $\eta$ is larger ($\eta = 1.0$).

Although this example only serves as an illustration, it should alert the reader to the fact that the choice of priors in Bayesian estimation is not innocuous. It may substantially affect the estimates. Just how large this impact is, may be difficult to determine in practice, especially when dealing with nonlinear functions of model parameters.

## 5.3 Extensions and Related Issues

So far we have primarily considered Gaussian likelihood functions in estimating the VAR model. Although this specification is commonly used in applied work, the assumption of unconditional normality is problematic in many macroeconomic applications (see, e.g., Kilian (1998b)). For example, models with volatility clustering necessarily give rise to non-Gaussian unconditional distributions. Examples of such models are discusssed in Chapters 14 and 18. While alternative distributions can be accommodated by the Bayesian framework, this generality usually increases the computational cost of Bayesian methods. More detailed discussions of how to use Bayesian analysis in specific settings can, for example, be found in Chapters 12, 13, and 18.

In this chapter, we have discussed priors for the parameters of the reduced-form VAR model. This approach continues to be a widely used approach in applied work. An alternative approach in structural VAR analysis is to impose priors on the parameters of the structural VAR representation. The latter approach is briefly discussed in Chapters 12 and 13 in the context of the question of how to conduct inference about structural impulse responses and related statistics.

Our analysis in this chapter has taken the lag order of the VAR model as given. Bayesians typically avoid the question of lag order selection by choosing a conservative large order $p$, but incorporate the prior belief that we are increasingly confident that the lagged coefficients are zero, the longer the lag length is. For example, the Minnesota prior postulates that the prior standard deviation

shrinks by a factor of $1/l$ for $l = 1, 2, \ldots, p$, as we have seen in this chapter. This device avoids having to truncate the lag structure at some order lower than $p$ at the cost of imposing additional structure on the prior variances of all lagged coefficients. Although this approach is intuitively appealing, it is ad hoc. There is no guarantee that this prior will result in more accurate forecasts or impulse response estimates than estimating an unrestricted VAR($p$) model or for that matter estimating a VAR($\widehat{p}$) model obtained by conventional lag order selection methods.

Although this approach is not common in applied work, it is also possible to consider restricted VAR models within the Bayesian framework. For example, Koop and Korobilis (2009) describe a so-called stochastic search variable selection (SSVS) prior that may be useful in reducing the curse of dimensionality by eliminating some lags from some equations of a VAR model based on Bayesian procedures. Details on such priors can be found in the related Bayesian literature.

Finally, whereas in this chapter we have focused on priors motivated by the improved forecast accuracy of the estimated VAR model, there also have been efforts to construct priors for VAR model parameters that incorporate restrictions implied by dynamic macroeconomic models. For example, Ingram and Whiteman (1994), Del Negro and Schorfheide (2004, 2011), and Del Negro, Schorfheide, Smets, and Wouters (2007) discuss priors for VAR models derived from specific DSGE models.