Bayesian Vector Autoregressions

Tomasz Woźniak [☆]

Abstract

This article provides an introduction to the burgeoning academic literature on Bayesian Vector Autoregressions, benchmark models for applied macroeconomic research. We first explain Bayes' theorem and the derivation of the closed-form solution for the posterior distribution of the parameters of the model given data. We further consider parameter shrinkage, a distinguishing feature of the prior distributions commonly employed in the analysis of large datasets, as well as an alternative way of specifying the prior distribution using dummy observations. Finally, we describe the mechanisms that enable feasible computations for these linear models that efficiently extract the information content of many variables for economic forecasting and other applications.

Keywords: Bayes' Theorem, Natural-Conjugate Analysis, Dummy Observations Prior, Normal-Wishart Distribution
JEL classification: C11, C32, C53

1. Vector Autoregressions

Vector autoregressions, abbreviated by VAR, are models that capture linear interdependencies between variables over time. They were introduced in economics by Sims (1972, 1980). Despite their simple formulation, VARs are very successful in capturing such stylised facts about economic time series as the decaying pattern in autocorrelations' values with increasing lag order, strong autocorrelations at an annual frequency, as well as dynamic linear interdependencies between time series. Therefore, VARs are well suited to model data that collects vectors of observations of N variables, denoted by y_t, for the time subscript t going from 1 to T.

The basic VAR equation is given by

$$y_t = \mu + A_1 y_{t-1} + \cdots + A_k y_{t-k} + u_t, \tag{1}$$

where μ is a N-vector denoting the constant term of the model, each autoregressive coefficients matrix $A_1, \dots, A_k$ is of dimensions $N \times N$, and u_t is the error component. We follow a common practice and assume that the error term at time t, given the past observations on vector y up to time $t-1$, is normally distributed with the mean set to a vector of zeros and with a covariance matrix Σ, which is denoted by

$$u_t \sim \mathcal{N}_N(\mathbf{0}, \Sigma). \tag{2}$$

VARs are an essential tool for empirical macroeconomic research. The dynamic interdependencies between variables representing the *Granger causality* hypothesis (see Granger, 1969) are captured by these models as showed by Sims (1972). VARs constitute a basis for developing and estimating Structural VAR models. Further, VARs can be shown to approximate other models such as Vector Autoregressive Moving Average models (see for instance Lütkepohl & Poskitt, 1996), State Space, and Dynamic Stochastic General Equilibrium models (Giacomini, 2013). Finally, thanks to the application of an econometric technique called parameter shrinkage, Bayesian VARs can be shown to forecast no worse than the principal component regression using a large number of predictors in both cases.

In this paper, VARs are analysed using Bayesian inference because thanks to an informed use of its elements the models gain features that are useful in applications. This was noticed in the seminal papers by Doan, Litterman & Sims (1984) and Litterman (1986) that established Bayesian VARs as benchmark models for economic forecasting. Doan et al. (1984), for instance, used the parameter shrinkage, that we describe below, in order to fine-tune the forecasting accuracy of the models. Moreover, as argued in Sims & Uhlig (1991), Bayesian models can easily include unit root nonstationary variables without affecting the inference on the parameters of the model and allowing for accurate medium-term forecasting with some caution given to potential explosive roots (see Uhlig, 1994). Also, thanks to feasible and fast computations for the estimation, as well to the flexibility provided by the application of the parameter shrinkage, Bayesian VARs are, amongst other applications, successfully used in macroeconomic forecasting with a large number of variables.

In what follows, we firstly explain the basics of Bayes' rule and its elements, that is the likelihood function, prior

[☆] *Address:* Department of Economics, University of Melbourne, 111 Barry Street, Carlton 3053, Victoria, Australia
Email: tomasz.wozniak@unimelb.edu.au
URL: http://bit.ly/tomaszwozniak

and posterior distributions and provide their characterisation for the VAR models. Further, we derive the posterior distribution for the parameters of the VAR model given the data that is a basis for the Bayesian inference. Finally, we explain the techniques that allow for feasible and fast computations, and that enable accurate forecasting. Throughout the paper, we focus on the Bayesian VAR model with a natural-conjugate prior distribution for which closed-form solutions for the posterior distribution exist. Such analytical derivations for a class of linear models were proposed in a seminal text by Zellner (1971). The exposition of the material is appropriate for a reader with the basic understanding of linear algebra, statistics, and time series analysis. The reader is referred to such textbooks as Hamilton (1994) and Lütkepohl (2005) for more detailed readings on the VAR models, Zellner (1971), Koop (2004), and Greenberg (2007) for the general introduction to Bayesian econometrics, and Bauwens, Richard & Lubrano (1999), Canova (2007), Korobilis (2008), Koop & Korobilis (2010), Del Negro & Schorfheide (2011), and Karlsson (2013) for the up-to-date discussion of macroeconometric methods. Finally, the reader is encourage to solve the two exercises that can be found in the text. In order to facilitate this work, Appendix A defines the statistical distributions that we use, Appendix B lists some useful matrix operations, and in Appendix C the reader is guided through the completing the squares technique that is used for the derivation of the posterior distribution for the parameters of the VAR model.

2. Bayes' Rule

Bayes' theorem was introduced by Thomas Bayes, an English statistician, philosopher and Presbyterian minister, as a solution to the so-called inverse probability problem. It is used, amongst other applications, as a powerful device for statistical inference. Using Bayes' rule we fully represent our knowledge about the parameters of the model having observed the data. In this section, we firstly present Bayes' rule in a form that is suitable for the analysis of the econometric models and, then, we define and characterise each of the elements occurring in the formula. Finally, we interpret the rule as a learning mechanism.

Let θ collect all of the parameters of the model that we want to estimate, and let Y denote the available data that we use for the purpose of the estimation of these parameters. Then, Bayes' rule is simply represented by

$$p\left(\theta|Y\right) = \frac{p\left(Y|\theta\right)p\left(\theta\right)}{p\left(Y\right)}, \tag{3}$$

where p denotes some probability distribution. On the LHS of equation (3) there is a posterior distribution.

A ***posterior distribution*** *fully characterises the knowledge about the parameters of the model having observed the data. It is defined as a conditional distribution of the parameters, θ, given the data, Y.*

The aim of the present paper is to present the derivation of the posterior distribution of the parameters of the VAR model. For that purpose we discuss all of the elements of Bayes' rule.

On the RHS of equation (3), in the numerator, we have a joint distribution of the data and the parameters that is factorised into a product of the conditional distribution of the data given the parameters, and the marginal distribution of parameters

$$p\left(Y, \theta\right) = p\left(Y|\theta\right)p\left(\theta\right). \tag{4}$$

The first element on the RHS of equation (4) is the likelihood function.

A ***likelihood function*** *is equivalent to the conditional distribution of the data given the parameters of the model.*

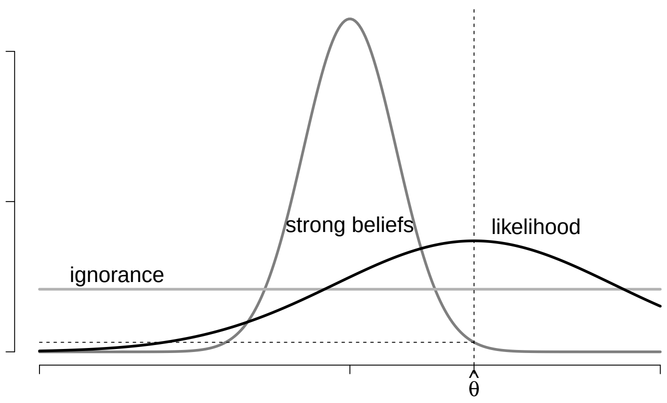
Analysing the likelihood function emphasises the main differences between the frequentist and Bayesian inference rules. In the frequentist approach, the model is a representation of the so-called data generating process. If we know this process and its parameters, θ, then we can randomly draw data from the conditional distribution of the data given the parameters. From the frequentist point of view, therefore, the available data, Y, are just a single realization of the data generating process. Consequently, an estimator that is a function of the data is random. However, the parameters themselves are non-random, although their values are unknown. From a Bayesian perspective the observed data are given and not random. Instead, all of the unknown values, such as the parameters, are considered random and, thus, are characterised by a probability distribution.

This last consideration leads us to the second element on the RHS of equation (4), that is the prior distribution of the parameters.

A ***prior distribution*** *of the parameters characterises the uncertainty about the parameters before observing the data. It is defined as a marginal distribution of the parameters θ.*

The prior distribution needs to be specified by an investigator and it represents the information about the parameters that they want to include in the statistical inference. The investigator may specify the prior distribution to represent her ignorance or to introduce their subjective beliefs about the parameters. Figure 1 illustrates the two cases. Note that imposing strong beliefs might result in giving less chance *a priori* to the value of the parameter preferred by the data and denoted in Figure 1 by $\hat{\theta}$: the distribution representing strong beliefs assigns a much lower value to the point $\hat{\theta}$ than the uninformative prior distribution expressing ignorance.

Figure 1: An example of an informative and uninformative prior distributions



In either case of specifying the informative or uninformative prior distribution, the beliefs represented by them are formally taken into account in the statistical inference. This feature gives rise to the subjective interpretation of probability. In the Bayesian approach, a probability is the degree to which the investigator is convinced about a certain outcome. In the frequentist approach on the other hand, it is a fraction of events that finish with the certain outcome in a sequence of identical independently repeated experiments. Note that the subjective interpretation of probability does not mean that the inference or the results are subjective.

In order to obtain the posterior distribution, the joint distribution of the data and the parameters is divided in equation (3) by a marginal data density. This value is called *marginal likelihood* in statistics, and *model evidence* in machine learning literature.

*A **marginal data density** summarises the evidence in favor of the model contained in the data.*

Firstly, consider the following formulation

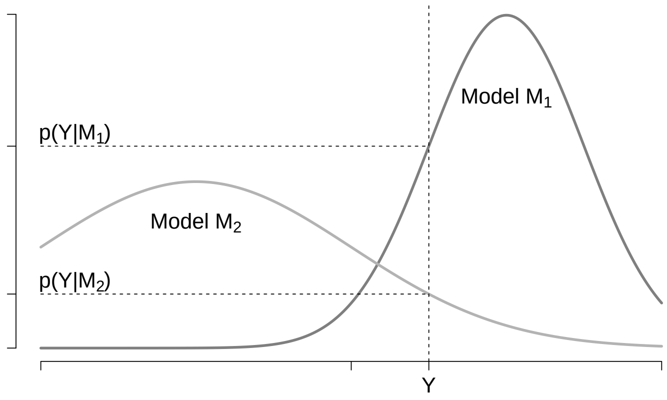
$$p(Y) = \int p(Y|\theta)\, p(\theta)\, d\theta, \tag{5}$$

in which the marginal likelihood is obtained by integrating over the parameter space the joint distribution of the data and the parameters. Notice that according to equation (5), the marginal likelihood is a constant that normalises the kernel of the posterior distribution, that is the product of the likelihood function and the prior distribution. This product fully characterises the posterior distribution, however, it does not integrate to one and, thus, is not a probability density function, but just a kernel. This finding leads to an alternative formulation of Bayes' rule given by

$$p(\theta|Y) \propto L(\theta; Y)\, p(\theta), \tag{6}$$

where $\propto$ denotes the proportionality up to the normalising constant, that is, the marginal data density.

Figure 2: An example of marginal data densities for two models



Secondly, the marginal data density can be interpreted as a predictive density of the data for a particular model, which can be emphasised in the notation by adding the conditioning on the model assumptions:

$$p(Y) = p(Y|\text{Model}) \tag{7}$$

More precisely, the marginal likelihood is an ordinate of the predictive density evaluated at the data point. To see this, consider a situation where the marginal likelihoods from two models, called M_1 and M_2 respectively, are available. Figure 2 illustrates such a situation. In this figure, the marginal data density for model M_1 has a higher value than one for M_2. Therefore, the evidence from the data is in favour of model M_1 over model M_2. The illustration from Figure 2 leads to a more general finding: a better model is capable of predicting the actual data more accurately.

Further, the marginal data density can be used to compute posterior probabilities of models given the data. For that purpose, simply apply Bayes' rule as follows

$$\Pr(M_1|Y) = \frac{p(Y|M_1)\Pr(M_1)}{p(Y|M_1)\Pr(M_1) + p(Y|M_2)\Pr(M_2)}, \tag{8}$$

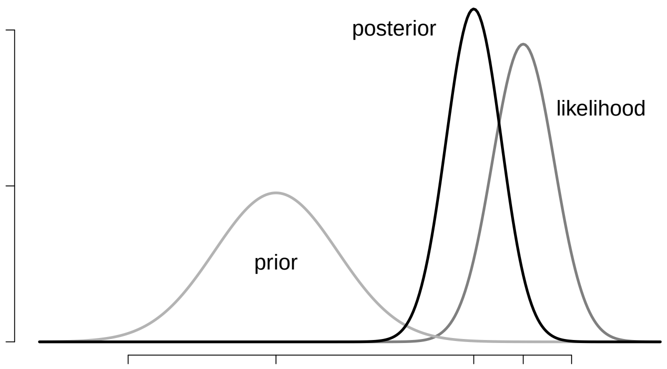
where $\Pr(M_1|Y)$ is the posterior probability of model M_1, while $\Pr(M_1)$ and $\Pr(M_2)$ are the prior probabilities of the models that need to be assumed. Of course, Bayes' rule, as explained in equation (8), can be used for computing the posterior probability of Model M_2, and can easily be generalised to any number of models.

Finally, consider once more the joint distribution of the data and the parameters in the following decompositions

$$p(\theta|Y)\, p(Y) = p(Y, \theta) = p(Y|\theta)\, p(\theta), \tag{9}$$

where on the LHS there are the posterior distribution and the marginal data density, while on the RHS there are the likelihood function and the prior distribution. Equation (9) presents Bayes' rule as a learning mechanism, in which the prior beliefs about the parameters are updated

Figure 3: An example of a prior distribution, likelihood function, and the resulting posterior distribution



by the information from the data in order to deliver the the posterior distribution and the model evidence. The particular role of the data is emphasised by the likelihood principle.

Likelihood principle. *All the information about the parameters of the model included in the data is captured by the likelihood function.*

Figure 3 illustrates the updating mechanism, in which the prior distribution of the parameters is updated by the information from the likelihood function giving the posterior distribution of the parameters given the data.

3. Likelihood function

The importance of the likelihood function is emphasised by the likelihood principle, as a means by which the information about the parameters of the VAR model is extracted from the data. In this section, we firstly focus on an appropriate representation of the VAR model that concisely summarises the information from the whole data sample, and secondly, we use it to write the likelihood function in a form that will greatly facilitate the derivations that will be presented in the subsequent sections.

In order to write the likelihood function in an appropriate form, we firstly introduce some additional notation. Let a $K \times 1$ vector, where $K = 1 + kN$, collect all of the data vectors occurring on the RHS of equation (1), that is $x_t = (1, y_{t-1}', \dots, y_{t-k}')'$, where x' denotes the transpose of x. Further, let a $K \times N$ matrix collect the intercept term and the autoregressive matrices from equation (1), $A = (\mu, A_1, \dots, A_k)'$. Now, equation (1) can be rewritten as

$$y_t' = x_t'A + u_t'. \tag{10}$$

Now, stack all such vectors y_t' for t going from 1 to T one under another to form a $T \times N$ matrix $Y = (y_1, y_2, \dots, y_T)'$, and proceed similarly with vectors x_t' and u_t' forming matrices: $X = (x_1, x_2, \dots, x_T)'$ with dimensions $T \times K$ and $U = (u_1, u_2, \dots, u_T)'$, $(T \times N)$. Using these newly defined matrices, we can write the model in equation (10) as

$$Y = XA + U. \tag{11}$$

Note that this form includes all of the information from the data, where on the LHS there is a matrix collecting the dependent variables, Y, and on the RHS the lagged observations on the variables are summarized in matrix X and multiplied by parameters A.

Further, note that each row of matrix U is conditionally independently normally distributed with the mean being a zero vector and the covariance matrix equal to Σ as in equation (2). These assumptions about the error terms, U, can be concisely summarised by stating that matrix U follows the following matric-variate normal distribution

$$U \sim \mathcal{MN}_{TN}\left(\mathbf{0}_{TN}, \Sigma, I_T\right), \tag{12}$$

where $\mathbf{0}_{TN}$ is a $T \times N$ matrix of zeros, and I_T is an identity matrix of order T. Refer to Appendix A for the definition of the matric-variate normal distribution. Equation (12) states that the error terms, for all time periods t, have zero mean and covariance matrix Σ. The latter is denoted by setting the row-specific covariance matrix to Σ. Finally, the fact that the column-specific covariance matrix is set to I_T, together with the joint normality, implies that the error terms, u_t, are independent over time.

Combining equations (11) and (12) allows us to state that the conditional distribution of the data, Y, given the parameters A and Σ (and past observations X) is given by the following normal-Wishart distribution

$$Y|A, \Sigma \sim \mathcal{MN}_{TN}\left(XA, \Sigma, I_T\right). \tag{13}$$

Therefore, the likelihood function is equal to

$$\begin{aligned} L\left(A, \Sigma; Y\right) &= (2\pi)^{-\frac{TN}{2}} \left|\Sigma\right|^{-\frac{T}{2}} \\ &\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y - XA\right)'\left(Y - XA\right)\right]\right\}. \end{aligned} \tag{14}$$

The likelihood function in equation (14) fully represents the information in the data about the parameters of the VAR model, as stated in the likelihood principle. It is equivalent to the sample distribution of that data, Y, given the parameters, A and Σ, as it is implied by equation (13). For the purpose of the estimation of the parameters, however, we treat it as a function of the parameters given the data. This is made explicit in the notation that we used on the LHS of equations (14). Rewriting the likelihood function from equation (14) in the following appropriate form will facilitate further explanations and derivations

$$\begin{aligned} L\left(A, \Sigma; Y\right) &\propto \left|\Sigma\right|^{-\frac{T}{2}} \\ &\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(A - \hat{A}\right)' X'X\left(A - \hat{A}\right)\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y - X\hat{A}\right)'\left(Y - X\hat{A}\right)\right]\right\}, \end{aligned} \tag{15}$$

where $\hat{A} = (X'X)^{-1}X'Y$ is the least squares (LS) estimator of A.

In equation (15) the likelihood function can be presented as a distribution of the parameters given the data in the form of the normal-Wishart distribution. The distribution has the following parameters

$$\begin{aligned} A|Y,\Sigma, &\sim \mathcal{MN}_{KN}\left(\hat{A},\Sigma,(X'X)^{-1}\right), \\ \Sigma|Y &\sim \mathcal{IW}_{N}\left(\left(Y - X\hat{A}\right)'\left(Y - X\hat{A}\right), T-K-N-1\right). \end{aligned} \tag{16}$$

The analysis above shows that if only data are taken into account, then the expected value of parameter A is $\hat{A}$, that is the LS estimator, and one of Σ is $\left(Y - X\hat{A}\right)'\left(Y - X\hat{A}\right)/(T-K-2N-2)$, which differs from its LS estimator only by term $2N+2$ in the denominator.

This section concludes with a simple exercise for preliminary derivations.

Exercise 1. Derive the normal-Wishart form of the likelihood function from equation (15). As a starting point use equation (14). Determine the parameters of the normal-Wishart distribution consisting of the conditional distribution of A given Σ and Y, as well as of the conditional distribution of Σ given Y.[1]

4. Prior distribution

The choice of the prior distribution for the parameters of the VAR model can supported by a wide range of arguments. In the rich volume of publications on this topic, the prior distributions were chosen so that they: reflect general properties of macroeconomic time series (see e.g. Doan et al., 1984; Litterman, 1986); are supported by theoretical economic models (Kadiyala & Karlsson, 1997; Del Negro & Schorfheide, 2004; Villani, 2009); allow for fast computations (Koop, 2013); assure good forecasting performance (Bańbura, Giannone & Reichlin, 2010; Carriero, Clark & Marcellino, 2015); give sufficient flexibility in the modeling (George, Sun & Ni, 2008); incorporate expert knowledge (Wright, 2013); or allow the data to partially determine them (Giannone, Lenza & Primiceri, 2015). We consider a basic prior distribution that enables analytical derivation of the posterior distribution and, thus, fast computations. This prior distribution is closely related to the distributions assumed in some of the papers mentioned above.

The type of prior distribution that we present is called the natural-conjugate prior distribution. The property of conjugacy means that the prior and posterior distributions are of the same class of probability distributions. In the previous section, the likelihood function was presented as the normal-Wishart distribution of parameters A and Σ given the data. It appears that assuming the normal-Wishart prior distribution results in the posterior distribution of the same form.

Consequently, the natural-conjugate prior distribution is a matric-variate normal conditional prior distribution of A given Σ, and an inverse Wishart marginal prior distribution for Σ, that is

$$\begin{aligned} A|\Sigma, &\sim \mathcal{MN}_{KN}\left(\underline{A},\Sigma,\underline{V}\right), \\ \Sigma &\sim \mathcal{IW}_{N}\left(\underline{S},\underline{\nu}\right). \end{aligned} \tag{17}$$

So-called hyper-parameters $\underline{A}$, $\underline{V}$, $\underline{S}$, and $\underline{\nu}$ fully determine the prior distribution and need to be specified by the investigator. $\underline{A}$ is the prior mean of A, while $\underline{V}$ is proportional to its column-specific covariance matrix. The row-specific prior covariance matrix of A is proportional to Σ which is a parameter of the model. Consequently, the prior distribution of A is conditioned on Σ. Further, $\underline{S}$ is the scale matrix of the inverse Wishart prior distribution for Σ and $\underline{\nu}$ is its degrees of freedom parameter.

The joint prior distribution can be presented as the product of the conditional distribution of A given Σ, and the marginal distribution for Σ

$$p\left(A,\Sigma\right) = p\left(A|\Sigma\right)p\left(\Sigma\right). \tag{18}$$

The exact analytical expression for this prior distribution is given by

$$\begin{aligned} p\left(A,\Sigma\right) &\propto |\Sigma|^{-\frac{\underline{\nu}+N+K+1}{2}} \\ &\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(A-\underline{A}\right)'\underline{V}^{-1}\left(A-\underline{A}\right)\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\underline{S}\right]\right\}. \end{aligned} \tag{19}$$

In the next section, the RHS of equation (19) will be multiplied by the likelihood function from equation (15) in order to derive the posterior distribution. In the remaining part of this section, however, the properties of the introduced prior distribution are considered.

To understand better the ideas behind setting the prior distribution we analyse a commonly used distribution introduced by Doan et al. (1984), here presented in a version proposed by Karlsson (2013). Such a class of prior distributions is often called a *Minnesota prior* reflecting the affiliations of the authors that introduced this idea. Their idea was motivated by the observation that many macroeconomic time series are unit-root nonstationary, that is that the current observation, y_t, is well-approximated by the lagged value of the same variable, y_{t-1}, and a normally distributed shock, u_t, which can be written as

$$y_t = y_{t-1} + u_t. \tag{20}$$

Of course, equation (20) is written in terms of vectors and it is interpreted as a multivariate random walk. Note that

[1] **Hint:** Appropriately add and subtract term $X\hat{A}$, rearrange the obtained terms, use the properties of the exponential function, and those of the transpose of a matrix.

equation (20) is equivalent to equation (1) with $\mu = \mathbf{0}$, $A_1 = I_N$, and $A_2 = \cdots = A_k = \mathbf{0}_{NN}$. Collect these values in one matrix $\underline{A}^* = \left(\mathbf{0}, I_N, \mathbf{0}_{N(k-1)}\right)'$. If we set the mean of the matric-variate normal prior distribution to this value, $\underline{A} = \underline{A}^*$, then we express our belief that the macroeconomic variables included in the VAR model are unit-root nonstationary without seeing the data. This assumption comes only from one's general experience of working with such data.

Consider now the role of matrix $\underline{V}$. Let vec(A) denote a vectorisation of matrix A that stacks its columns one under another in a $NK \times 1$ vector. From the definition of the matric-variate normal distribution in Appendix A it can be read that the prior covariance matrix of vec(A) is given by $\Sigma \otimes \underline{V}$, where $\otimes$ denotes the Kronecker product of matrices. If Σ_{nn} denotes the (n, n)th element of Σ, that is the variance of nth error term from the VAR equation, then $\Sigma \otimes \underline{V}$ implies that the covariance matrix of the nth column of matrix A is equal to $\Sigma_{nn}\underline{V}$. Note that this expression explains what we mean by writing that matrix $\underline{V}$ is proportional to the column-specific covariance matrix of A. It is proportional to $\Sigma_{nn}\underline{V}$ up to the constant Σ_{nn}. Further, note that the prior covariance matrices of the columns of A are simply a rescaled versions of $\underline{V}$. The equation-specific scale is introduced, again, by Σ_{nn}.

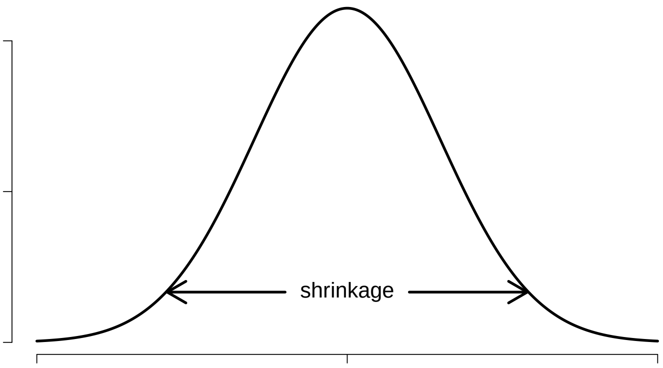
In a vast majority of the articles on Bayesian VARs matrix $\underline{V}$ is assumed to be a diagonal matrix. This assumption is supported by the claim that it is difficult to make *a priori* statements about the correlations between the parameters. Consequently, these correlations are set to zero. The elements on the diagonal of $\underline{V}$ determine the prior variances of the parameters in A. The idea behind the Minnesota prior is to set the dispersion of the prior distributions around the prior mean, $\underline{A}^*$, so that some stylised facts about the role of the variables in the model are reflected. Moreover, in the Minnesota prior a small number of new hyper-parameters together with some additional assumptions, such as $\underline{A} = \underline{A}^*$, determine the original hyper-parameters $\underline{A}$, $\underline{V}$, $\underline{S}$, and $\underline{\nu}$. Therefore, the first element on the diagonal of $\underline{V}$ is set to λ_0, a new strictly positive hyper-parameter that determines the prior variances of the constant terms, μ. Consequently, the prior variance of the constant term in the nth equation, denoted by μ_n, is equal to $\Sigma_{nn}\lambda_0$. The value of λ_0 needs to be specified by the investigator.

The remaining diagonal elements of $\underline{V}$, denoted by $\underline{V}_{ii}$, where i goes from 2 to K are set to

$$\underline{V}_{ii} = \frac{\lambda_1}{l^2\hat{\sigma}_n^2}, \tag{21}$$

where $i = (l-1)N+n+1$ for l going from 1 to k and denoting the subsequent lags in the VAR equation, and n going from 1 to N. λ_1 is the second new hyper-parameter to be determined by the investigator. $\hat{\sigma}_n^2$ denotes the LS estimate of the error term variance from a univariate autoregressive model estimated for each of the N variables in y_t. Using $\hat{\sigma}_n^2$ in the specification of the prior distributions means that we use the information from the data for that purpose. While this is a controversial feature of the Minnesota prior, it occurs to be a practical choice for forecasting. In fact, the Minnesota prior is commonly used in macroeconomic applications.

Figure 4: An illustration of parameter shrinkage



The larger the value of element l^2 from the denominator of equation (21) the smaller the value of the prior variance of a particular parameter. For instance, for the diagonal elements of $\underline{V}$ corresponding to the parameters from matrix A_1, l takes the value of 1. For those corresponding to the parameters from matrix A_2, l takes the value of 2, and so on. Setting a smaller prior variance for the parameters of more distant lags reflects the decaying pattern in the autocorrelation coefficients commonly observed in macroeconomic time series. Again, such a general observation, which is based only on the experience of working with such variables is a motivation for setting the prior distribution.

To understand the role of $\hat{\sigma}_n^2$ in equation (21) consider the prior variances equal to $\Sigma_{mm}\underline{V}_{ii}$, where the subscript mm is used in order to emphasize that m might not be equal n in the determination of the prior variances. Moreover, m takes the values from 1 to N. When the Bayesian estimate of Σ_{nn} approaches $\hat{\sigma}_n^2$, then the variances $\Sigma_{mm}\underline{V}_{ii}$ for which $m = n$ are approximated by $\frac{\lambda_1}{l^2}$. Therefore, the prior variances are determined only by the hyper-parameter λ_1 and the lag in the autoregressive matrices in this case. In the opposite case, when $m \neq n$, the prior variances are equal to $\Sigma_{nn}\frac{\lambda_1}{l^2\hat{\sigma}_m^2}$. Therefore, the prior variances on the effects of other variables in the nth equation are appropriately rescaled.

The only two hyper-parameters in the Minnesota prior to be determined by the investigator are λ_0 and λ_1. λ_0 is responsible for the prior variance of the constant terms, μ, in the VAR equation, whereas λ_1 is responsible for the prior variance of the autoregressive matrices $A_1, \ldots, A_k$. These two hyper-parameters are often called the constant term-related shrinkage parameter and the autoregressive parameters-related shrinkage parameter respectively. The term *shrinkage* refers to the fact that these hyper-parameters determine to what extent the prior distribution of the corresponding parameters of the model is *shrinked* towards the prior mean. This interpretation is illustrated in Figure 4. The shrinkage parameters determine the prior variances in the following way. The larger the shrinkage, the smaller the prior variances and the values of the hyper-parameters λ_0 and λ_1. Later, in Section 7 we will come back to this interpretation and the practical use of the shrinkage technique.

5. Posterior distribution

In this section, we present the posterior distribution of the parameters of the VAR model given the data. As we showed in equation (6) the product of the likelihood function and the prior distribution of the parameters gives the kernel of the posterior distribution. We, therefore, plug the likelihood function in the form presented in equation (15) and the prior distribution presented in equation (19) into equation (6) to obtain

$$
\begin{aligned}
p\left(A,\Sigma|Y\right) \propto\ & |\Sigma|^{-\frac{\nu+T+N+K+1}{2}} \\
& \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(A-\hat{A}\right)' X'X\left(A-\hat{A}\right)\right]\right\} \\
& \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(A-\underline{A}\right)' \underline{V}^{-1}\left(A-\underline{A}\right)\right]\right\} \\
& \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y-X\hat{A}\right)'\left(Y-X\hat{A}\right)\right]\right\} \\
& \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\underline{S}\right]\right\}.
\end{aligned}
\tag{22}
$$

Note that the second and the third line of equation (22) include matrix A and, thus, they determine the parameters of the posterior distribution of A. Further, appropriate rearrangements of the elements in equation (22) lead to presenting the posterior distribution of the parameters given the data in the normal-Wishart form.

Therefore, the joint posterior distribution of A and Σ is factorised into the conditional posterior distribution of A given Σ and the data Y, and the marginal posterior distribution of Σ given data, which can be presented as

$$
p\left(A,\Sigma|Y\right) = p\left(A|Y,\Sigma\right)p\left(\Sigma|Y\right). \tag{23}
$$

The first element on the RHS of equation (23) is the matric-variate distribution, and the second one is the inverse Wishart distribution

$$
\begin{aligned}
A|Y,\Sigma, &\sim \mathcal{MN}_{KN}\left(\overline{A},\Sigma,\overline{V}\right), \\
\Sigma|Y &\sim \mathcal{IW}_{N}\left(\overline{S},\overline{\nu}\right),
\end{aligned}
\tag{24}
$$

where parameters $\overline{A}$, $\overline{V}$, $\overline{S}$, and $\overline{\nu}$ fully characterise the posterior distribution and are given by

$$
\begin{aligned}
\overline{V} &= \left(\underline{V}^{-1} + X'X\right)^{-1}, \\
\overline{A} &= \overline{V}\left(\underline{V}^{-1}\underline{A} + X'Y\right), \\
\overline{\nu} &= \underline{\nu} + T, \\
\overline{S} &= \underline{S} + Y'Y + \underline{A}'\underline{V}^{-1}\underline{A} - \overline{A}'\overline{V}^{-1}\overline{A}.
\end{aligned}
\tag{25}
$$

Different papers present matrix $\overline{S}$ in various forms that can be shown to be equivalent to the one in equation (25). The derivation of the posterior distribution is left as an exercise.

Exercise 2. Derive the normal-Wishart posterior distribution of A and Σ given the data Y presented in equation (24). As a starting point use the kernel of the posterior distribution from equation (22). Determine the parameters of the normal-Wishart distribution consisting of the conditional posterior distribution of A given Σ and Y, as well as of the marginal posterior distribution of Σ given Y.[2]

The analysis of the parameters of the matric-variate normal posterior distribution allows the following interpretations. Firstly, notice that parameters of the posterior distribution intuitively illustrate the updating mechanism. For instance, the prior column-related covariance matrix of A, $\underline{V}$, is updated by the relevant information from the data, represented by the crossproduct $X'X$, to form the posterior column-related covariance matrix $\overline{V}$. The prior degrees of freedom parameter, $\underline{\nu}$, is updated by the sample size, T, in order to obtain the posterior degrees of freedom equal to $\underline{\nu} + T$.

Secondly, the employment of the prior distribution increases the precision of the estimation, that is the inverse of the covariance matrix of the parameters. The covariance matrix of the posterior distribution of vec(A) is equal to $\Sigma \otimes \overline{V}$, whereas the asymptotic covariance matrix of the LS estimator is equal to $\hat{\Sigma}\otimes(X'X)^{-1}$ (compare to Lütkepohl, 2005, Section 3), where $\hat{\Sigma}$ is the LS estimator of Σ. In the Bayesian estimation the LS estimator's precision determined by $X'X$ is simply improved by the precision of the prior distribution $\underline{V}^{-1}$.

Furthermore, the mean of the posterior distribution of A can be presented as

$$
\overline{A} = \overline{V}\underline{V}^{-1}\underline{A} + \overline{V}X'X\hat{A} = \Omega_1\underline{A} + \Omega_2\hat{A}, \tag{26}
$$

[2] **Hint:** Use the properties of the exponential function and the trace of a matrix to rewrite the kernel of the posterior distribution from equation (22) in the following form $|\Sigma|^{-\frac{\nu+T+N+K+1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\bar{S}\right]\right\}$, where $\bar{S}$ collects all of the corresponding elements. Use all of the necessary matrix transformations, including completing the squares for A technique from Appendix C, to rewrite $\bar{S}$ as $\left(A-\overline{A}\right)'\overline{V}^{-1}\left(A-\overline{A}\right)+\underline{S}+Y'Y+\underline{A}'\underline{V}^{-1}\underline{A}-\overline{A}'\overline{V}^{-1}\overline{A}$. Use this form of the kernel of the posterior distribution to determine the parameters of the normal-Wishart distribution from equation (25). Good luck!

represents a linear combination of the prior mean $\underline{A}$ and the LS estimate $\hat{A}$ with weights $\Omega_1 = \overline{V}\underline{V}^{-1}$ and $\Omega_2 = \overline{V}X'X$ respectively. Moreover, if one assumes an uninformative prior distribution by setting the diagonal elements of $\underline{V}$ to infinity and, in effect, those of the prior precision matrix to zeros, then equation (26) implies that the Bayesian estimate of A is equal to the LS estimator $\hat{A}$. Note that in this case, the Bayesian covariance matrix of the posterior distribution of A approaches the covariance matrix of the LS estimator of A. On the other hand, assuming a dogmatic prior distribution that assigns probability mass equal to 1 to the prior mean and probability mass equal to zero for all other points in the parameter space results in the posterior mean equal to $\underline{A}$.

Consider a situation, in which having computed the parameters of the posterior distribution using the available data Y and X, as in equation (24), the investigator receives new information, that is, new observations of vector y collected in $T^* \times N$ matrix Y^* with an appropriately constructed matrix X^*. In this context, the question of the sensitivity of the results to the arrival of new data arises naturally. Of course, due to the small computational burden of the computations for the VARs with the natural-conjugate prior distribution the obvious solution is to stack matrices Y and Y^*, as well as X and X^*, and use the formula from equation (24) in order to compute the parameters of the posterior distribution. It is easy to demonstrate that the solution can be shown equivalent to

$$\begin{aligned} A|Y,\Sigma, &\sim \mathcal{MN}_{KN}\left(\overline{A}^*, \Sigma, \overline{V}^*\right),\\ \Sigma|Y &\sim \mathcal{IW}_N\left(\overline{S}^*, \overline{\nu}^*\right), \end{aligned} \tag{27}$$

where the parameters of the posterior distribution are given by

$$\begin{aligned} \overline{V}^* &= \left(\overline{V}^{-1} + X^{*\prime}X^*\right)^{-1},\\ \overline{A}^* &= \overline{V}^*\left(\overline{V}^{-1}\overline{A} + X^{*\prime}Y^*\right),\\ \overline{\nu}^* &= \overline{\nu} + T^*,\\ \overline{S}^* &= \overline{S} + Y^{*\prime}Y^* + \overline{A}'\overline{V}^{-1}\overline{A} - \overline{A}^{*\prime}\overline{V}^{*-1}\overline{A}^*, \end{aligned} \tag{28}$$

where $\overline{A}$, $\overline{V}$, $\overline{S}$, and $\overline{\nu}$ are the parameters of the posterior computed before the arrival of the new data and given in equation (25). The form of the posterior distribution given in equations (27) and (28) shows how the investigators knowledge about the VARs' parameters is updated upon the arrival of new data. It is an example of a popular statement that *yesterday's posterior is today's prior*. Note that the new parameters of the posterior distribution from equation (28) can be easily expanded to a form that well illustrates the contributions of particular elements. For instance, matrix $\overline{V}^*$ is equal to the inverse of the sum $\underline{V}^{-1} + X'X + X^{*\prime}X^*$, matrix $\overline{A}^*$ is equal to $\overline{V}^*\left(\underline{V}^{-1}\underline{A} + X'Y + X^{*\prime}Y^*\right)$, and $\overline{\nu}^* = \overline{\nu} + T + T^*$.

Further interpretations are presented in the context of the large Bayesian VARs, that is the VAR models for data sets with large number of variables N in Section 7.

6. Dummy observations prior

In this section, an alternative way of specifying the natural-conjugate prior distribution is presented. This originates in the proposal of Doan et al. (1984) and uses the fact that the likelihood function can be represented in a form for the normal-Wishart distribution, as in equation (15). Therefore, consider some appropriately chosen dummy observations collected in matrices Y^+ and X^+ and write the VAR model as

$$Y^+ = X^+A + U^+, \tag{29}$$

where the error term U^+ follows the same matric-variate normal distribution as U in equation (12).

Note that matrices Y^+ and X^+ are not data, but rather they contain some values specified by the investigator. Therefore, the likelihood function written in the normal-Wishart form, as in equation (16), for the model for the dummy observations from equation (29) specifies the natural-conjugate prior distribution given by

$$\begin{aligned} A|\Sigma, &\sim \mathcal{MN}_{KN}\left(\underline{A}^+, \Sigma, \underline{V}^+\right),\\ \Sigma &\sim \mathcal{IW}_N\left(\underline{S}^+, \underline{\nu}^+\right), \end{aligned} \tag{30}$$

the parameters of which are given by

$$\begin{aligned} \underline{A}^+ &= \left(X^{+\prime}X^+\right)^{-1}X^{+\prime}Y^+,\\ \underline{V}^+ &= \left(X^{+\prime}X^+\right)^{-1},\\ \underline{S}^+ &= \left(Y^+ - X^+\underline{A}^+\right)'\left(Y^+ - X^+\underline{A}^+\right),\\ \underline{\nu}^+ &= T^+ - K - N - 1. \end{aligned} \tag{31}$$

An obvious finding from the analysis of the parameters in equation (31) is that these are matrices Y^+ and X^+ that determine the normal-Wishart prior distribution. Doan et al. (1984) proposed such a specification of these matrices that the mean and the covariance matrix of the matric-variate normal distribution correspond to those of the Minnesota prior. This dummy observation prior distribution is often called the sum-of-coefficients prior. Further, Sims (1993) proposes the so-called dummy-initial-observation prior that is consistent with the assumption of cointegration. Finally, Del Negro & Schorfheide (2004) includes in these matrices observations simulated from a theoretical macroeconomic model.

Note that the prior distribution above may not specify a proper statistical distribution. If not enough dummy observations are delivered, then the degrees of freedom parameter does not meet the condition that $\underline{\nu} > N$. This happens if less than $K + 2N + 1$ observations are delivered.

Moreover, if $T^+ < K$ the precision matrix of the matric-variate normal prior distribution for A, $X^{+'}X^+$, is not of full rank and cannot be easily inverted. For instance, the original specifications of Doan et al. (1984) and Sims (1993) give jointly $N+1$ dummy observations, which would not be enough for most of the VAR models. In effect, the prior distribution is not a proper distribution function, that is the area under its probability density function does not integrate to 1. Instead, it depends on some normalising constant. This fact causes problems in the model comparison using the marginal data density because this value depends on the unknown normalising constant of the prior distribution.

A commonly used solution to this problem is to apply a mixed strategy and specify the prior distribution through the dummy observations, as in equation (30), as well as through the prior distribution in the normal-Wishart form, as in equation (17). By analysing the product of these two prior distributions one obtains as new mixed-strategy normal-Wishart prior distribution given by

$$\begin{aligned} A|\Sigma, &\sim \mathcal{MN}_{KN}\left(\widetilde{A}, \Sigma, \widetilde{V}\right), \\ \Sigma &\sim \mathcal{IW}_N\left(\widetilde{S}^+, \widetilde{\nu}\right), \end{aligned} \tag{32}$$

the parameters of which are given by

$$\begin{aligned} \widetilde{V} &= \left(\underline{V}^{-1} + X^{+'}X^+\right)^{-1}, \\ \widetilde{A} &= \widetilde{V}\left(\underline{V}^{-1}\underline{A} + X^{+'}Y^+\right), \\ \widetilde{S} &= \underline{S} + Y^{+'}Y^+ + \underline{A}'\underline{V}^{-1}\underline{A} - \widetilde{A}^{*'}\widetilde{V}^{*-1}\widetilde{A}^*, \\ \widetilde{\nu} &= \underline{\nu} + T^+. \end{aligned} \tag{33}$$

The derivations of the prior distribution above are identical to those for the posterior distribution presented in Section 5. Also in this case, the invertibility of the precision matrix is guaranteed by adding a positive definite matrix $\underline{V}^{-1}$ to a potentially reduced rank matrix $X^{+'}X^+$. Furthermore, the conditions for the degrees of freedom parameter can be met by an appropriate choice of the value of $\underline{\nu}$. Such prior distributions were used, for instance, by Carriero et al. (2015) and Giannone et al. (2015).

Finally, the parameters of the posterior distribution of A and Σ in the normal-Wishart form are given by

$$\begin{aligned} \overline{V} &= \left(\widetilde{V}^{-1} + X'X\right)^{-1}, \\ \overline{A} &= \overline{V}\left(\widetilde{V}^{-1}\widetilde{A} + X'Y\right), \\ \overline{S} &= \widetilde{S} + Y'Y + \widetilde{A}'\widetilde{V}^{-1}\widetilde{A} - \overline{A}^{*'}\overline{V}^{*-1}\overline{A}^*, \\ \overline{\nu} &= \widetilde{\nu} + T. \end{aligned} \tag{34}$$

Also in the case of the parameters of equation (34) the impact of the two forms of the prior distribution and the data can be traced. For instance, the precision matrix of the matric-variate normal distribution can be written as $\underline{V}^{-1} + X^{+'}X^+ + X'X$, the mean of this distribution as $\overline{V}\left(\underline{V}^{-1}\underline{A} + X^{+'}Y^+ + X'Y\right)$, and the degrees of freedom parameter as $\underline{\nu} + T^+ + T$.

7. Large Bayesian VARs

Since the publication of the seminal papers by Doan et al. (1984) and Litterman (1986) Bayesian VARs became benchmark models for economic forecasting. Later, thanks to the developments proposed by De Mol, Giannone & Reichlin (2008), these models were shown to extract the information from a large number of variables at least as efficiently for the predictive purposes as the main competing statistical technique, that is the principal component regression. Finally, Bańbura et al. (2010) proposed a solution to the problem of overfitting of the Bayesian VARs, a feature of richly parametrised models when the excellent in-sample fit coincides with relatively poor forecasting performance of the model. The developments mentioned above, as well as other recent developments, paved the way to a new theme in econometrics called large Bayesian VARs.

Such models are characterised by a large number of variables in the analyzed system and, thus, a large number of parameters. For instance, consider forecasting macroeconomic variables for the U.S. Available data bases include over a hundred monthly and quarterly relevant macro aggregates for the postwar period. Usually, due to high autocorrelations of macro time series at an annual frequency, VARs with $k = 4$ lags are required for the forecasting of quarterly variables, and with $k = 12$ lags for the forecasting of monthly data. With this amount of observations and parameters the LS estimation of VAR models with more than sixty variables becomes impossible. This claim is supported by the observation that in such a case matrix X has more columns than rows, $K > T$, and, thus, its crossproduct $X'X$ is not of full rank. It, therefore, cannot be simply inverted and the estimator $\hat{A} = (X'X)^{-1}X'Y$ which includes this inverse cannot be computed. The employment of the prior distributions in Bayesian inference of the VARs is one of the feasible solutions to this problem. The solution is based on a result that a sum of reduced and full matrices of the same dimensions is a full-rank matrix. Therefore, by adding to $X'X$ a positive definite prior precision matrix $\underline{V}^{-1}$ in the estimation of $\overline{V}$ from equation (25) the problem is solved. This is one of the factors determining the feasibility of the computations in the Bayesian VARs.

The problem of the reduced rank of $X'X$ can be interpreted also as an identification problem. Which means that the available data does not contain enough information to identify and estimate all of the parameters of the model. In this context, consider again the employment of the Minnesota prior in a Bayesian estimate of the A matrix presented in equation (26), with $\underline{A} = \underline{A}^*$. In this case, time series are forecasted based on the parameters estimated as a linear combination of the sample information and the random walk prior assumptions. However, if the data informational content is not sufficient for the estimation, then the variables are estimated on the basis of the random walk process only. The random walk process itself, however, is also a good benchmark forecasting model that is not that easy to beat in forecasting performance comparisons.

The role of the parameter shrinkage is emphasised here as it determines the level to which the posterior distribution is concentrated around the prior mean. The two limiting cases of the uninformative and dogmatic prior distribution analysed at the end of Section 5 illustrate the range of options between which the investigator has when choosing the level of shrinkage. The two extreme cases hardly ever occur optimal from the point of view of the forecasting performance with the Bayesian VARs, which raises the question of the selection of appropriate values for the hyper-parameters such as λ_0 and λ_1. Different approaches to this issue include choosing these values in order to maximise the marginal data density (e.g. Sims & Zha, 1998), to give the balance between the in-sample and out of sample fit of the model (Bańbura et al., 2010), or to optimise the forecasting performance of the model (Carriero et al., 2015). Another proposition of Giannone et al. (2015) is to estimate the hyper-parameters.

Finally, the feasibility of computations for the large Bayesian VARs with the natural-conjugate prior distribution consists not only of the fact that the closed-form solutions for the posterior distribution exist. Another simplification comes from the Kronecker structure of the posterior covariance matrix of A, that is $\Sigma\otimes\overline{V}$. The most demanding computational task for the estimation of these models is inverting of this $KN \times KN$ matrix. The computations are greatly simplified by using the property of the Kronecker product that

$$\left(\Sigma\otimes\overline{V}\right)^{-1} = \Sigma^{-1}\otimes\overline{V}^{-1},$$

which reduces the dimensionality of the problem to inverting two matrices of dimensions $N \times N$ and $K \times K$ respectively. In addition, closed-form solutions for such values as the marginal data density and the predictive density of the one period ahead forecast are available for the Bayesian VARs with the natural-conjugate prior distribution. The reader is referred to Karlsson (2013) for the exact formulas and hints for the derivations.

Such fast computations encouraged research investigating the model space within the class of Bayesian VARs, that is the development of the techniques that verify various forms of the model. For instance, Corander & Villani (2006) apply fractional Bayes techniques in order to determine the lag order of the VAR model as well as the Granger causality structure. Further, Jarociński & Maćkowiak (2013) develop an algorithm that explores the model space in order to determine which variables from a broad data set are useful for forecasting a particular sub-set of variables, that is the Granger priority structure. Finally, Giannone et al. (2015) uses the analytical expression for the marginal data density given the hyper-parameters in order to derive the posterior distribution for the hyper-parameters themselves.

Assuming the natural-conjugate prior leads to the closed-form solutions in the posterior analysis. Such solutions are available for a couple of other classes of prior distributions or simplifications of the model revised lately by Koop (2013). However, deviating from such simplifying assumptions by for example assuming independent prior distribution, that can be presented as

$$p\left(A,\Sigma\right) = p\left(A\right)p\left(\Sigma\right), \tag{35}$$

or assuming a hierarchical prior for the hyper-parameters collected in vector $\lambda = (\lambda_0, \lambda_1)$, that is given by

$$p\left(A,\Sigma,\lambda\right) = p\left(A|\lambda\right)p\left(\Sigma\right)p\left(\lambda\right), \tag{36}$$

leads to the employment of numerical simulation techniques. The computational burden can be partially reduced by applying an approximate inference such as, for instance, the variational Bayes technique used by Hajargasht & Woźniak (2016).

We conclude this section with a remark on the use of the term *Bayesian VARs* that spans a wide range of models. All of them have the common feature that the vector of observations is modeled with some modification of equation (1). Examples of such models include Structural VARs and Vector Error Correction models that allow for the employment of economic theory in the analysis of the variables. Also a broad class of time-varying parameter VARs that allow any of the parameters of the VAR equation to evolve over time according to a feasible process. The last case includes heteroskedastic models in which the covariance matrix of the error term, Σ, changes over time. Bayesian VARs are adjusted to particular data sets that include, amongst others, panel data or variables sampled at different frequencies, for example monthly and quarterly. Finally, the assumption of normality is relaxed in many specifications and replaced by other parametric or nonparametric distributions. For the references or reviews of these models the reader is referred to such textbooks and chapters as Bauwens et al. (1999), Koop & Korobilis (2010), Canova (2007), Del Negro & Schorfheide (2011), and Karlsson (2013), as well as to a volume of *Advances in Econometrics* by Fomby, Kilian & Murphy (2013).

Appendix A. Useful multivariate distributions

This section is based mostly on Karlsson (2013).

Matric-variate normal distribution. The $K \times N$ matrix X is said to have a matric-variate normal distribution

$$X \sim \mathcal{MN}_{KN}\left(A,\Sigma,V\right)$$

where A is the $K \times N$ mean matrix, Σ (with dimensions $N \times N$) and V ($K \times K$) are the positive definite symmetric matrices that are proportional to the variance matrix of the rows and of the columns of X respectively, if vec(X) is multi-variate normal

$$\text{vec}(X) \sim \mathcal{N}\left(\text{vec}(A),\Sigma\otimes V\right).$$

The density of X is, then

$$\begin{aligned}\mathcal{MN}_{KN}\left(X;A,\Sigma,V\right) &= (2\pi)^{-\frac{KN}{2}}\left|\Sigma\right|^{-\frac{K}{2}}\left|V\right|^{-\frac{N}{2}}\\ &\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(X-A\right)'V^{-1}\left(X-A\right)\right]\right\}.\end{aligned}$$

Inverse Wishart distribution. The $N\times N$ positive semi-definite symmetric matrix Σ is said to have a Wishart distribution

$$\Sigma\sim\mathcal{IW}_N\left(S,\nu\right),$$

where S is a $N\times N$ positive definite symmetric matrix called the scale matrix, and $\nu>N$ is the degrees of freedom parameter, if its density is given by

$$\mathcal{IW}\left(\Sigma;S,\nu\right)=c_w^{-1}\left|\Sigma\right|^{-\frac{\nu+N+1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}S\right]\right\},$$

where: $c_w = 2^{\frac{\nu N}{2}}\pi^{\frac{N(N-1)}{4}}\prod_{i=1}^{N}\Gamma\left(\frac{\nu+1-i}{2}\right)|S|^{-\frac{\nu}{2}}$, and Γ is the gamma function.

Normal-Wishart distribution. Let

$$\begin{aligned}X|\Sigma &\sim\mathcal{MN}_{KN}\left(A,\Sigma,V\right), \text{ and}\\ \Sigma &\sim\mathcal{IW}_N\left(S,\nu\right).\end{aligned}$$

Then the joint distribution of X and Σ is said to be normal-Wishart

$$X,\Sigma\sim\mathcal{NW}\left(A,V,S,\nu\right),$$

with the density given by

$$\begin{aligned}\mathcal{NW}\left(X,\Sigma;A,V,S,\nu\right) &= c_{nw}^{-1}\left|\Sigma\right|^{-\frac{\nu+K+N+1}{2}}\\ &\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(X-A\right)'V^{-1}\left(X-A\right)\right]\right\}\\ &\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}S\right]\right\},\end{aligned}$$

where $c_{nw} = 2^{\frac{N(K+\nu)}{2}}\pi^{\frac{N(N+2K-1)}{4}}\prod_{i=1}^{N}\Gamma\left(\frac{\nu+1-1}{2}\right)|V|^{\frac{N}{2}}|S|^{-\frac{\nu}{2}}$.

Appendix B. Useful matrix transformations

This section is based mostly on Lütkepohl (1996).

The Inverse.
For A, B $(a\times a)$ nonsingular: $(AB)^{-1}=B^{-1}A^{-1}$.
For A $(a\times a)$, B $(b\times b)$, A and B nonsingular:
$(A\otimes B)^{-1}=A^{-1}\otimes B^{-1}$.

Matrix multiplication.
For A $(a\times b)$, B and C $(b\times c)$: $A(B\pm C)=AB\pm AC$.
For A and B $(a\times b)$, and C $(b\times c)$: $(A\pm B)C=AC\pm BC$.

The trace.
For A, B $(a\times a)$: $\operatorname{tr}(A\pm B)=\operatorname{tr}(A)\pm\operatorname{tr}(B)$.

The transpose.
For A $(a\times b)$, B $(b\times c)$: $(AB)'=B'A'$.

Appendix C. Completing the squares

In the derivation of the posterior distribution in Section 5 one faces the problem of summing two quadratic forms:

$$\left(X-A\right)'B^{-1}\left(X-A\right),\ \text{and}\ \left(X-C\right)'D^{-1}\left(X-C\right).$$

In what follows, the subsequent steps of completing the squares are presented. Matrices B, D, and F are assumed to be symmetric positive definite matrices.

Firstly, perform all of the multiplications

$$\begin{aligned}&\left(X-A\right)'B^{-1}\left(X-A\right)+\left(X-C\right)'D^{-1}\left(X-C\right)\\ &\quad= X'B^{-1}X-X'B^{-1}A-A'B^{-1}X+A'B^{-1}A+\\ &\qquad+X'D^{-1}X-X'D^{-1}C-C'D^{-1}X+C'D^{-1}C\end{aligned}$$

and group appropriate elements by X

$$\begin{aligned}&X'\left(B^{-1}+D^{-1}\right)X-X'\left(B^{-1}A+D^{-1}C\right)\\ &\qquad-\left(B^{-1}A+D^{-1}C\right)'X+A'B^{-1}A+C'D^{-1}C.\end{aligned}$$

Let $F^{-1}=\left(B^{-1}+D^{-1}\right)$, and write

$$\begin{aligned}&X'F^{-1}X-X'\left(B^{-1}A+D^{-1}C\right)\\ &\qquad-\left(B^{-1}A+D^{-1}C\right)'X+A'B^{-1}A+C'D^{-1}C.\end{aligned}$$

Then, appropriately multiply by $F^{-1}F$ remembering that $F=F'$

$$\begin{aligned}&X'F^{-1}X-X'F^{-1}F\left(B^{-1}A+D^{-1}C\right)\\ &\qquad-\left(B^{-1}A+D^{-1}C\right)'FF^{-1}X+A'B^{-1}A+C'D^{-1}C.\end{aligned}$$

Let $E=F\left(B^{-1}A+D^{-1}C\right)$, and write

$$X'F^{-1}X-X'F^{-1}E-E'F^{-1}X+A'B^{-1}A+C'D^{-1}C.$$

Add and subtract $E'F^{-1}E$

$$\begin{aligned}&X'F^{-1}X-X'F^{-1}E-E'F^{-1}X+E'F^{-1}E\\ &\qquad-E'F^{-1}E+A'B^{-1}A+C'D^{-1}C\end{aligned}$$

in order to obtain

$$\left(X-E\right)'F^{-1}\left(X-E\right)-E'F^{-1}E+A'B^{-1}A+C'D^{-1}C,$$

that is a new quadratic form $\left(X-E\right)'F^{-1}\left(X-E\right)$, where matrices E and F are defined above, and residual terms $-E'F^{-1}E+A'B^{-1}A+C'D^{-1}C$ that need to be taken into account in further derivations of the posterior distribution.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian Vector Auto Regressions. *Journal of Applied Econometrics*, *92*, 71– 92.

Bauwens, L., Richard, J., & Lubrano, M. (1999). *Bayesian inference in dynamic econometric models*. Oxford University Press, USA.

Canova, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton University Press.

Carriero, A., Clark, T. E., & Marcellino, M. (2015). Bayesian VARs: Specification Chioces and Forecast Accuracy. *Journal of Applied Econometrics*, *30*, 46–73.

Corander, J., & Villani, M. (2006). A Bayesian Approach to Modelling Graphical Vector Autoregressions. *Journal of Time Series Analysis*, *53*, 160.

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328.

Del Negro, M., & Schorfheide, F. (2004). Priors from General Equilibrium Models for VARs. *International Economic Review*, *45*, 643–673.

Del Negro, M., & Schorfheide, F. (2011). Bayesian Macroeconometrics. In J. Geweke, G. Koop, & H. K. van Dijk (Eds.), *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press, USA.

Doan, T., Litterman, R. B., & Sims, C. A. (1984). Forecasting and Conditional Projection Using Realistic Prior Distributions. *Econometric Reviews*, *3*, 1–100.

Fomby, T. B., Kilian, L., & Murphy, A. (Eds.) (2013). *VAR Models in Macroeconomics New Developments and Applications: Essays in Honor of Christopher A. Sims* volume 32 of *Advances in Econometrics*. Emerald Insight.

George, E. I., Sun, D., & Ni, S. (2008). Bayesian Stochastic Search for VAR Model Restrictions. *Journal of Econometrics*, *142*, 553–580.

Giacomini, R. (2013). The Relationship Between DSGE and VAR Models. *Advances in Econometrics*, *32*, 1–25.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*, 436–451.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*, 424–438.

Greenberg, E. (2007). *Introduction to Bayesian Econometrics*. Cambridge University Press.

Hajargasht, G., & Woźniak, T. (2016). Variational Bayes Inference for Large Vector Autoregressions. University of Melbourne Working Papers Series.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Jarociński, M., & Maćkowiak, B. (2013). Granger-Causal-Priority and Choice of Variables in Vector Autoregressions. European Central Bank Working Paper Series.

Kadiyala, K. R., & Karlsson, S. (1997). Numerical Methods for Estimation and Inference in Bayesian Var-Models. *Journal of Applied Econometrics*, *12*, 99–132.

Karlsson, S. (2013). Forecasting with Bayesian Vector Autoregression. In G. Elliott, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (pp. 791–897). Elsevier B.V. volume 2.

Koop, G. (2004). *Bayesian Econometrics*. John Wiley & Sons Ltd.

Koop, G. (2013). Forecasting with Medium and Large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.

Koop, G., & Korobilis, D. (2010). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Foundations and Trends in Econometrics*, *3*, 267–358.

Korobilis, D. (2008). Forecasting in Vector Autoregressions with Many Predictors. *Advances in Econometrics*, *23*, 403–431.

Litterman, R. B. (1986). Forecasting with Bayesian Vector Autoregressions: Five Years of Experience. *Journal of Business & Economic Statistics*, *4*, 25–38.

Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons, Ltd.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Lütkepohl, H., & Poskitt, D. S. (1996). Testing for Causation Using Infinite Order Vector Autoregressive Processes. *Econometric Theory*, *12*, 61–87.

Sims, C. A. (1972). Money, Income, and Causality. *The American Economic Review*, *62*, 540 – 552.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*, 1–48.

Sims, C. A. (1993). A Nine-Variable Probabilistic Macroeconomic Forecasting Model. In J. H. Stock, & M. W. Watson (Eds.), *Business Cycles, Indicators and Forecasting* (pp. 179–212). Chicago: University of Chicago Press for NBER.

Sims, C. A., & Uhlig, H. (1991). Understanding Unit Rooters: A Helicopter Tour. *Econometrica*, *59*, 1591–1599.

Sims, C. A., & Zha, T. (1998). Bayesian Methods for Dynamic Multivariate Models. *International Economic Review*, *39*, 949–968.

Uhlig, H. (1994). What Macroeconomists Should Know About Unit Roots. A Bayesian Perspective. *Econometric Theory*, *10*, 645–671.

Villani, M. (2009). Steady-State Priors for Vector Autoregressions. *Journal of Applied Econometrics*, *24*, 630– 650.

Wright, J. H. (2013). Evaluating Real-Time VAR Forecasts with an Informative Democratic Prior. *Journal of Applied Econometrics*, *28*, 762–776.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons.