Spike and Slab Priors on Variable Orderings in VARs*

Ping Wu
University of Strathclyde

Gary Koop
University of Strathclyde

May 2022

Abstract

It is increasingly common to estimate Bayesian Vector Autoregressions (VARs) in a structural form involving the Cholesky decomposition of the reduced form error covariance matrix. The resulting structural form has an error covariance matrix which is diagonal, allowing for equation by equation estimation of the VAR, leading to a huge reduction in the computational burden. However, this leads to order dependence. Posterior and predictive results differ depending on the way the variables are ordered in the VAR. In this paper, we propose the use of spike and slab priors over different variable orderings and allow the data to select the optimal ordering. We develop two models and Markov Chain Monte Carlo (MCMC) methods for posterior sampling over orderings based on the Plackett-Luce and Bradley-Terry models. In a macroeconomic exercise involving VARs with 20 variables we demonstate the effectiveness of our two approaches in choosing the optimal ordering and find substantive forecasting improvements relative to a strategy of subjectively selecting a single ordering.

*We would like to thank participants at the NBER-NSF SBIES conference 2022 for useful comments.

1 Introduction

Bayesian VARs have traditionally been estimated in reduced form, where the right hand side of the VAR equation involves an $n \times 1$ vector of dependent variables, $\mathbf{y}_t$, and the error covariance matrix for the VAR, $\mathbf{\Sigma_t}$, is unrestricted (apart from being positive definite). However, as VARs have become larger and larger, researchers have increasingly been estimating VARs in a structural form where the right hand side of the VAR is $\mathbf{B}_0\mathbf{y}_t$ and the error covariance matrix is diagonal. $\mathbf{B}_0$ is a lower triangular matrix based on the Cholesky decomposition of $\mathbf{\Sigma_t}$. The fact that the error covariance matrix of the structural VAR is diagonal means that Bayesian estimation can proceed one equation at a time. As shown, e.g., in Carriero et al. (2019) the MCMC algorithm based on the reduced form VAR requires $O(n^6)$ elementary operations to take one draw of the VAR coefficients. This is reduced to $O(n^4)$ with the structural VAR. Thus, in larger VARs, it can be computationally impractical to work in the reduced form in cases where MCMC analysis is feasible when working in the structural form.[1]

However, the use of the Cholesky decomposition of the reduced form error covariance matrix leads to order dependence. That is, posterior and predictive results will vary depending on the way that the variables are ordered in the VAR. To clarify the precise nature of this order dependence, we highlight the discussion of sub-section 3.1 of Carriero et al. (2019) who demonstrate that the posterior of the structural form VAR coefficients, conditional on $\mathbf{\Sigma_t}$ is invariant to ordering. The lack of order invariance arises due to the fact that the implied prior on $\mathbf{\Sigma_t}$ is not order invariant. Arias et al. (2021) demonstrates the importance of the ordering issue in VARs both theoretically and empirically. The authors show that, although point forecasts are not sensitive to the way variables are ordered, predictive standard deviations can be subtantially affected by the way that the variables are ordered. The VARs considered in Arias et al. (2021) are all low dimensional. Chan et al. (2021) consider ordering issues in high dimensional VARs and demonstrate that the theoretical and empirical findings of Arias et al. (2021) hold with additional force in higher dimensions. Thus, there is growing theoretical and empirical evidence that ordering issues are important, particularly in the large VARs that cannot easily be estimated in reduced form.

These considerations have stimulated interest in order invariant approaches. Reduced form estimation of VARs is order invariant with commonly-used priors, but is not scaleable to large VARs. Chan et al. (2021) critiques various order invariant approaches and proposes a new order invariant approach which avoids the use of the Cholesky decomposition relies on stochastic volatility to identify the model and is scaleable. The present paper adopts a different strategy to address the ordering issue. We retain the Cholesky decom-

[1]Numerical instability issues can also plague the Bayesian estimation of reduced form VARs due to the need to invert and/or take Cholesky dcompositions of enormous posterior covariance matrices for the VAR coefficients.

position of the reduced form error covariance but develop methods for finding the optimal ordering of the variables. An alternative way of viewing what we do is that we treat the ordering of variables in the VAR as unknown and estimate it using Bayesian methods.

2 Bayesian Inference on Orderings in VARs

In this section, we develop Bayesian methods to carrying out inference on ways of ordering the variables in VARs. After defining a standard VAR likelihood function, we develop a spike and slab prior and allow for inference on variable ordering. Subsequently, we develop MCMC methods which allow for posterior inference and prediction.

2.1 The Likelihood Function

We work with VARs with stochastic volatility (SV) involving $\mathbf{y}_t = (y_{1,t}, \ldots, y_{n,t})'$, an $n \times 1$ vector, observed over the periods $t = 1, \ldots, T$:

$$\mathbf{y}_t = \mathbf{a} + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{B}_0^{-1} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_t), \tag{1}$$

where $\mathbf{a}$ is an $n \times 1$ vector of intercepts, $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are $n \times n$ matrices of VAR coefficients, $\mathbf{B}_0$ is an $n \times n$ matrix, and $\mathbf{D}_t = \text{diag}\left(\mathrm{e}^{h_{1,t}}, \ldots, \mathrm{e}^{h_{n,t}}\right)$ is diagonal. Finally, each of the log-volatility $h_{i,t}$ follows the stationary AR(1) process:

$$h_{i,t} = \phi_i h_{i,t-1} + u_{i,t}^h, \quad u_{i,t}^h \sim \mathcal{N}\left(0, \omega_i^2\right)$$

where $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian distribution.[2]

$\mathbf{B}_0$ has ones on the diagonal. That is, in the likelihood function we model $\mathbf{B}_0$ as

$$\mathbf{B}_0 = \begin{pmatrix} 1 & b_{1,2} & \ldots & b_{1,n} \\ b_{2,1} & 1 & & b_{2,n} \\ \vdots & & \ddots & \vdots \\ b_{n,1} & \ldots & b_{n,n-1} & 1 \end{pmatrix}. \tag{2}$$

These assumptions suffice to define a likelihood function which is the same as that used in Chan et al. (2021) which shows that, although the homoskedastic version of the model leads to a lack of identification, the presence of stochastic volatility means that the $\mathbf{B}_0$ is identified. In this paper, we develop a prior which identifies even the homoskedastic version of the VAR.

[2]It is simple to extend our methods to the homoskedastic VAR or to allow for $\mathbf{A}_i$ for $i = 1, \ldots, p$ to be time varying. Allowing for $\mathbf{B}_0$ to be time varying would be a greater challenge, but could be done by replacing our spike and slab prior with a dynamic spike and slab prior.

2.2 Spike and Slab Priors over Variable Orderings

The parameters in our model include VAR coefficients and those in the SV processess as well as $\mathbf{B}_0$. The developments in this paper relate to the latter and, hence, this is what we discuss in this sub-section. For the other parameters, our prior choices are standard and, thus, we relegate discussion of them to the appendix.

A common assumption is that $\mathbf{B}_0$ is a lower triangular matrix with ones on the diagonal. It is this which leads to order dependence as shown, e.g., in Chan et al. (2021). But the lower triangularity assumption has the advantage that it leads to simple and fast Bayesian computation. Hence, in this paper, we retain the lower triangularity assumption but express uncertainty over which lower triangular form is appropriate. To be precise, we use priors which rule out all forms for $\mathbf{B}_0$ except those that can be, via permutations of the columns of the matrix, put into a lower triangular form with ones on the diagonal. Each valid ordering is defined through a row vector $\boldsymbol{\rho}^l$ for $l = 1 \dots L$ which are stored in an $L \times n$ matrix P. For instance, in a VAR with 3 variables there are 6 possible orderings (e.g. 1,2,3; 1,3,2; 2,1,3; etc.) and, thus, $L = 6$. Each $\boldsymbol{\rho}^l$ denotes a unique ordering and is an $1 \times n$ vector $\boldsymbol{\rho}^l = \left(\rho_1^l, \dots, \rho_n^l\right)$, where ρ_1^l is the variable which takes the first place under ordering l, ρ_2^l the second, etc.

We stress that the variables in $\mathbf{y}_t$ are ordered in a particular way (from variable 1 through variable n) and that they always appear in this order in the model. When we refer to variable orderings this relates to $\mathbf{B}_0$ with the idea being that, through appropriate permutations of its columns, it becomes lower triangular. That is, formally the variables in $\mathbf{y}_t$ never change order, nor is $\mathbf{B}_0$ itself necessarily lower triangular. However, we consider a prior which only allows for choices of $\mathbf{B}_0$ which can be transformed into lower triangular form after appropriate switching of its columns. This is equivalent to considering all possible structural VARs with lower triangular triangular impact matrices for different variable orderings. When we use phrases below which refer to different orderings of the variables, they should be interpreted in this context. For each choice of l we restrict $\mathbf{B}_0$ to lower triangular form (after permutations) through a Dirac spike-and-slab prior. Lower triangularity is obtained if, for $i < j$, $i, j = 1, \dots, n$, variable ρ_i^l is ordered before variable ρ_j^l, $i \neq j$, then is is non-zero, otherwise it is zero. Thus,

$$p\left(b_{\rho_i^l,\rho_j^l} \mid \boldsymbol{\rho}^l\right) = \mathbb{1}_{(i<j)}\Delta_0\left(b_{\rho_i^l,\rho_j^l}\right) + \mathbb{1}_{(i>j)}\mathcal{N}\left(a_b, V_b\right), \quad i \neq j. \tag{3}$$

This defines the prior for $\mathbf{B}_0$ conditional on a specific ordering.

We now need a prior over orderings. When dealing with ordered or ranked data, two classic models have shown great popularity over the years: the Plackett-Luce model (Luce, 1959; Plackett, 1975) and the Bradley-Terry model (Bradley and Terry, 1952). The idea of our approach is to use these models, not for the data itself, but to design priors for P. The basic idea of the Plackett-Luce model is that it provides a distribution

for rank ordered data.[3] The Plackett-Luce model is referred to as a multiple comparisons model. In contrast, the Bradley-Terry model is for pairwise comparisons. For instance, it is commonly used with sporting data and builds a model for whether player i beats player j. In our context, it it a model for whether variable i is ordered ahead of variable j in the VAR. Knowledge of all pairwise orderings like this suffices to order all the variables in the VAR. We focus on the Plackett-Luce model and provide details about the Bradley-Terry model in the appendix.

2.2.1 A Plackett-Luce Prior over Orderings

To explain the Plackett-Luce prior over orderings, we first introduce parameters for each variable, $\lambda_j > 0$ for $j = 1 \dots n$ which represent the so-called skill rating or ability of each variable (i.e. λ_j determines the probability that variable $y_{j,t}$ is ordered first in the VAR) and we denote $\lambda := \{\lambda_i\}_{i=1}^n$. The Plackett-Luce prior over orderings is given by:

$$p\left(\boldsymbol{\rho}^l \mid \lambda\right) = \prod_{i=1}^{n} \frac{\lambda_{\rho_i^l}}{\sum_{k=i}^{n} \lambda_{\rho_k^l}}. \tag{4}$$

It is a hierarchical prior in that λ is treated as unknown and estimated.

There is an alternative way of writing the Plackett-Luce model which involves intoducing a latent variable z_i for each variable

$$z_i \mid \lambda_i \sim \mathcal{E}\left(\lambda_i\right), \quad i = 1, 2, \dots, n \tag{5}$$

where $\mathcal{E}$ denote the exponential distribution.

If we define

$$p\left(\boldsymbol{\rho}^l \mid \mathbf{z}\right) \equiv \Pr\left(z_{\rho_1^l} < z_{\rho_2^l} < \cdots < z_{\rho_n^l}\right) \tag{6}$$

and calculate

$$p\left(\boldsymbol{\rho}^l \mid \lambda\right) = \int p\left(\boldsymbol{\rho}^l \mid \mathbf{z}\right) p\left(\mathbf{z} \mid \lambda\right) \mathrm{d}\mathbf{z}, \tag{7}$$

it can be shown to be the same prior as in (4).

[3]The Plackett-Luce model assumes the ranking is complete, an assumption we maintain for our variable orderings. However, it is worth noting that Bayesian methods for the extended Plackett-Luce model are available (Johnson et al., 2021) and could easily be used in our approach. This model allows for the ranking to be incomplete which, in some cases, could be useful for VAR ordering problems. For instance, a researcher may wish to know whether a block of macroeconomic variables is ordered before or after a block of financial variables but does not care about the ordering of variables within each block.

Another equivalent latent variable representation of the Plackett-Luce model uses the latent variables in (5) and writes

$$p\left(\boldsymbol{\rho} \mid \mathbf{z}\right) = \Pr\left(\rho_1^l = \min\{z_{\rho_1^l}, \cdots, z_{\rho_n^l}\}\right) \times \cdots \times \Pr\left(\rho_{n-1}^l = \min\{z_{\rho_{n-1}^l}, z_{\rho_n^l}\} | z_{\rho_1^l}, \cdots, z_{\rho_{n-2}^l}\right). \tag{8}$$

The equivalence of 8 and 6 arises since $z_1, \ldots, z_n$ are independent exponentially distributed random variables with rate parameters $\lambda_1, \ldots, \lambda_n$. A property of the exponential distribution is that $\min\{z_1, \ldots, z_n\}$ is also exponentially distributed, with parameter $\lambda_1 + \cdots + \lambda_n$ which provides us with the following distribution:

$$\Pr\left(\min\{z_{\rho_1^l}, z_{\rho_2^l}, \cdots, z_{\rho_n^l}\}\right) \sim \mathcal{E}\left(\lambda_{\rho_1^l} + \lambda_{\rho_2^l} + \cdots + \lambda_{\rho_n^l}\right). \tag{9}$$

This latent variable representation is important in our MCMC algorithm. Intuitively, this algorithm will first search over all n variables to draw the variable to be ordered first, then search over the remaining $n-1$ variables to find the variable ordered second, etc.

Finally, we require a prior for λ, which is the vector of abilities for the variables. For these, we adopt independent gamma prior distributions:

$$p(\lambda) = \prod_{i=1}^{n} \mathcal{G}\left(\lambda_i; a_{\lambda,i}, b_{\lambda,i}\right).$$

As explained in Caron and Doucet (2012), the hyperparameters $b_{\lambda,i}$ are just scaling parameters on λ_i. As the likelihood is invariant to a rescaling of the λ_i, these hyperparameters can be fixed without influencing inference. Following Henderson and Kirrane (2018) we set $b_{\lambda,i} = 1$.

Following Caron and Doucet (2012) we estimate the other set of prior hyperparameters, $a_{\lambda,i}$. We begin by setting $a_{\lambda,i} = a_\lambda$ for $i = 1, \ldots, n$ which leads to an exchangeable prior specification in which we believe that some variables are better than others but we have no beliefs about which the stronger and the weaker variables are (Henderson and Kirrane, 2018). Thus, we assume $\lambda_i \sim \mathcal{G}(a_\lambda, 1)$. To estimate a_λ, we use the Metropolis-Hastings step proposed in Caron and Doucet (2012).

2.3 Posterior of ordering

We propose two steps to sample the ordering. One is the forward step, another is the backward step. In the forward step, we first decide which variable will take the first place. After deciding the first place, we next decide the second place. We repeat this process

until the last place. Then, in the backward step, we first decide which variable will take the last place. After deciding the last place, we next decide the second last place. We repeat this process until the first place. Finally, we will get the optimal ordering $\boldsymbol{\rho}^*$. This method is inspired by rewriting Equation (6) as

$$\begin{aligned} p\left(\boldsymbol{\rho}^* \mid \mathbf{z}\right) \equiv & \Pr\left(z_{\rho_1^*} < z_{\rho_2^*} < \cdots < z_{\rho_n^*}\right) \\ = & \Pr\left(\rho_1^* = \min\{z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_n^*}\}\right) \times \Pr\left(\rho_2^* = \min\{z_{\rho_2^*}, \cdots, z_{\rho_n^*}\} | \rho_1^*\right) \times \cdots \times \\ & \Pr\left(\rho_{n-1}^* = \min\{z_{\rho_{n-1}^*}, z_{\rho_n^*}\} | z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_{n-2}^*}\right). \end{aligned} \tag{10}$$

This is because: Since $z_1, \ldots, z_n$ are independent exponentially distributed random variables with rate parameters $\lambda_1, \ldots, \lambda_n$. Then $\min\{z_1, \ldots, z_n\}$ is also exponentially distributed, with parameter $\lambda_1 + \cdots + \lambda_n$. So:

$$\Pr\left(\min\{z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_n^*}\}\right) \sim \mathcal{E}\left(\lambda_{\rho_1^*} + \lambda_{\rho_2^*} + \cdots + \lambda_{\rho_n^*}\right). \tag{11}$$

Then the selected variable is distributed according to the categorical distribution

$$\Pr\left(z_{\rho_1^*} = \min\left\{z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_n^*}\right\}\right) = \frac{\lambda_{\rho_1^*}}{\lambda_{\rho_1^*} + \cdots + \lambda_{\rho_n^*}}. \tag{12}$$

So

$$\begin{aligned} p\left(\rho^* \mid \lambda\right) = & \prod_{i=1}^{n} \frac{\lambda_{\rho_i^*}}{\sum_{k=i}^{n} \lambda_{\rho_k^*}}, \\ = & \frac{\lambda_{\rho_1^*}}{\lambda_{\rho_1^*} + \lambda_{\rho_2^*} + \cdots + \lambda_{\rho_n^*}} \times \frac{\lambda_{\rho_2^*}}{\lambda_{\rho_2^*} + \lambda_{\rho_3^*} + \cdots + \lambda_{\rho_n^*}} \times \cdots \times \frac{\lambda_{\rho_{n-1}^*}}{\lambda_{\rho_{n-1}^*} + \lambda_{\rho_n^*}} \times \frac{\lambda_{\rho_n^*}}{\lambda_{\rho_n^*}}, \\ \\ = & \Pr\left(\rho_1^* = \min\{z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_n^*}\}\right) \times \Pr\left(\rho_2^* = \min\{z_{\rho_2^*}, \cdots, z_{\rho_n^*}\} | \rho_1^*\right) \times \cdots \times \\ & \Pr\left(\rho_{n-1}^* = \min\{z_{\rho_{n-1}^*}, z_{\rho_n^*}\} | z_{\rho_1^*}, z_{\rho_2^*}, \cdots, z_{\rho_{n-2}^*}\right). \end{aligned}$$

This method can be thought as: we first decide which variable will take the first place. There are n variables in total, so we compare the n variables. The winner needs to beat all other $n-1$ variables. After deciding the first place, we next decide the second place. The winner needs to beat all other $n-2$ variables. We repeat this process until the last place.

3 Bayesian Computation

Before discussing the MCMC algorithm, we describe the prior on λ, which is the ability of variables. Beliefs about the relative abilities of the variables are expressed through

independent gamma prior distributions:

$$p(\lambda) = \prod_{i=1}^{n} \mathcal{G}\left(\lambda_i; a_{\lambda,i}, b_{\lambda,i}\right).$$

As explained in Caron and Doucet (2012), the parameter $b_{\lambda,i}$ is just a scaling parameter on λ_i. As the likelihood is invariant to a rescaling of the λ_i, this parameter does not have any influence on inference. Hence, to ensure that the maximum a posteriori estimate satisfies $\sum_{i=1}^{n} \lambda_i = 1$, they set $b_{\lambda,i} = n a_{\lambda,i} - 1$. Henderson and Kirrane (2018) suggested that $b_{\lambda,i}$ can be set equal to 1, due to the scale invariance of λ. We can follow their choice, setting $b_{\lambda,i} = 1$.

For $a_{\lambda,i}$, taking $a_{\lambda,i} = a_\lambda$ for $i = 1, \ldots, n$ leads to an exchangeable prior specification in which we believe that some variables are better than others but we have no beliefs about who the stronger and the weaker variables are (Henderson and Kirrane, 2018). This choice is the same as the choice in Caron and Doucet (2012). We follow their choice, setting $a_{\lambda,i} = a_\lambda$. Then our choice of $a_{\lambda,i}$ and $b_{\lambda,i}$ leads to the specification $\lambda_i \sim \mathcal{G}(a_\lambda, 1)$. For a_λ, we use the Metropolis-Hastings step proposed in Caron and Doucet (2012) to sample it.

Next we develop a posterior sampler which allows for Bayesian estimation of the multiple comparisons and pair comparisons. Below we discuss sampling of:

Step 1: $p\left(\mathbf{B}_0 \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \boldsymbol{\rho}^*, \mathbf{z}, \lambda\right) = p\left(\mathbf{B}_0 \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \boldsymbol{\rho}^*\right)$;

Step 2: $p\left(\boldsymbol{\rho}^* \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \mathbf{z}, \lambda, a_\lambda\right) = p\left(\boldsymbol{\rho}^* \mid \mathbf{B}_0, \mathbf{z}, \lambda\right)$;

Step 3: $p\left(\mathbf{z} \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \lambda, a_\lambda\right) = p\left(\mathbf{z} \mid \boldsymbol{\rho}^*, \lambda, a_\lambda\right)$;

Step 4: $p\left(\lambda \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda\right) = p\left(\lambda \mid \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda\right)$.

Step 5: $p\left(a_\lambda \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \mathbf{z}, \lambda\right) = p\left(a_\lambda \mid \lambda\right)$.

The details of the rest of the posterior sampler are the same as Chan et al. (2021).

Step 1: $p\left(\mathbf{B}_0 \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \boldsymbol{\rho}^*, \mathbf{z}, \lambda\right) = p\left(\mathbf{B}_0 \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \boldsymbol{\rho}^*\right)$.

We first rewrite our model as:

$$(\mathbf{Y} - \mathbf{X}\mathbf{A})\mathbf{B}_0' = \mathbf{E}$$

where $\mathbf{Y}$ is the $T \times n$ matrix of dependent variables, $\mathbf{X}$ is the $T \times k$ matrix of lagged dependent variables with $k = 1 + np$, $\mathbf{A} = (\mathbf{a}, \mathbf{A}_1, \ldots, \mathbf{A}_n)'$ is the $k \times n$ matrix of VAR coefficients and $\mathbf{E}$ is the $T \times n$ matrix of errors. Then, for $i = 1, \ldots, n$, we have

$$(\mathbf{Y} - \mathbf{X}\mathbf{A})\mathbf{b}_i = \mathbf{E}_i, \quad \mathbf{E}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_{\mathbf{h}_i,}\right)$$

where $\mathbf{E}_i$ is the i-th column of $\mathbf{E}$ and $\boldsymbol{\Omega}_{\mathbf{h}_i,} = \operatorname{diag}\left(\mathrm{e}^{h_{i,1}}, \ldots, \mathrm{e}^{h_{i,T}}\right)$.

The free elements in $\mathbf{b}_i$ depend on the ordering $\boldsymbol{\rho}^*$: we first find the ordering of variable i, those variables whose ordering is before variable i will take the slab $\mathcal{N}(a_b, V_b)$ as prior, whiel those variables whose ordering is after variable i will take the Dirac-spike as prior (that is, they are equal to zero). And for the variable i itself, the associated element in $\mathbf{b}_i$ will be one.

Hence, the full conditional distribution of $\mathbf{b}_i$ is given by

$$\begin{aligned} p(\mathbf{b}_i \mid \mathbf{y}, \mathbf{A}, \boldsymbol{\rho}^*, \mathbf{h}_{i,\cdot}) &\propto \mathrm{e}^{-\frac{1}{2}\mathbf{b}_i'(\mathbf{Y}-\mathbf{XA})'\boldsymbol{\Omega}_{\mathbf{h}_{i,}}^{-1}(\mathbf{Y}-\mathbf{XA})\mathbf{b}_i} \times \mathrm{e}^{-\frac{1}{2}(\mathbf{b}_i-\mathbf{b}_{0,i})'\mathbf{V}_{\mathbf{b}_i}^{-1}(\mathbf{b}_i-\mathbf{b}_{0,i})} \\ &\propto \mathrm{e}^{-\frac{1}{2}\left(\mathbf{b}_i-\widehat{\mathbf{b}}_i\right)'\mathbf{K}_{\mathbf{b}_i}\left(\mathbf{b}_i-\widehat{\mathbf{b}}_i\right)} \end{aligned}$$

where

$$\mathbf{K}_{\mathbf{b}_i} = \mathbf{V}_{\mathbf{b}_i}^{-1} + (\mathbf{Y}-\mathbf{XA})'\boldsymbol{\Omega}_{\mathbf{h}_{i,}}^{-1}(\mathbf{Y}-\mathbf{XA}), \quad \widehat{\mathbf{b}}_i = \mathbf{K}_{\mathbf{b}_i}^{-1}\mathbf{V}_{\mathbf{b}_i}^{-1}\mathbf{b}_{0,i}.$$

Step 2: $p(\boldsymbol{\rho}^* \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \mathbf{z}, \lambda) = p(\boldsymbol{\rho}^* \mid \mathbf{B}_0, \mathbf{z}, \lambda)$.

Step 2 consists of two steps: the forward step and the backward step.

The forward step:

1) we first decide which variable will take the first place. There are n variables in total, so we compare the n variables. The winner needs to beat all other $n-1$ variables. The success of variable i means all other variables will not appear in this equation, that is, the i-th row in $\mathbf{B}_0$ will be zero and the prior will come from the Dirac spike component (Here we use a small variance $c = 0.0001$). The success of variable i also means this variable will appear in all other equations, that is, the i-th column in $\mathbf{B}_0$ will be non-zero and the prior will come from the slab component (except itself, because it always appears in its own equation. This is not influenced by the ordering).

Let $b_{i\cdot}$ denote the i-th row in $\mathbf{B}_0$ and delete the i-th element (which is the variable itself), $b_{\cdot i}$ denote the i-th column in $\mathbf{B}_0$ and delete the i-th element. Let p_i denote the success probability of variable i. Then:

$$p_i = p(i = \rho_1^* \mid \mathbf{B}_0) \propto \frac{\lambda_i}{\sum_{k=1}^n \lambda_k} \phi(b_{i\cdot}; 0, c)\, \phi(b_{\cdot i}; a_b, V_b),$$

Hence, after normalization, we obtain

$$p_i = \frac{p_i}{\sum_{k=1}^n p_k}.$$

We can sample from the multinomial distribution as follows:

a) Create an array p containing the cumulative probabilities of p_i, $i = 1, 2, \cdots, n$;

b) Generate U, a uniform(0,1) random value;

c) Select the first index such that $p_i > U$.

2) conditioning on the first place, we can decide which variable will take the second place. There are $n-1$ variables in total, so we compare the $n-1$ variables. The winner needs to beat all other $n-2$ variables. The success of variable j means all other variables will not appear in this equation, that is, the j-th row in $\mathbf{B}_0$ will be zero and the prior will come from the Dirac spike component (except the variable which takes the first place). The success of variable j also means this variable will appear in all other equations, that is, the j-th column in $\mathbf{B}_0$ will be non-zero and the prior will come from the slab component (also, except the variable which takes the first place).

Let $b_{j\cdot}$ denote the j-th row in $\mathbf{B}_0$, delete the j-th element (which is the variable itself) and delete the ρ_1^*-th element (which takes the first place), $b_{\cdot j}$ denote the j-th column in $\mathbf{B}_0$, delete the j-th element and delete the ρ_1^*-th element. Let p_j denote the success probability of variable j. Then:

$$p_j = p\left(j = \rho_2^* \mid \mathbf{B}_0\right) \propto \frac{\lambda_j}{\sum_{k=1,k\neq\rho_1^*}^{n} \lambda_k} \phi\left(b_{j\cdot}; 0, c\right) \phi\left(b_{\cdot j}; a_b, V_b\right),$$

Hence, after normalization, we obtain

$$p_j = \frac{p_j}{\sum_{k=1,k\neq\rho_1^*}^{n} p_k}.$$

$\vdots$

We can proceed until the $(n-1)$-th place. The remaining variable will automatically take the last place.

The backward step:

This is almost the same as the forward step, except that we first decide which variable will take the last place, then decide which variable will take the second last place.

To implement Step 3 and Step 4, we follow the sampling approach in Caron and Doucet (2012):

Step 3: $p\left(\mathbf{z} \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \lambda, a_\lambda\right) = p\left(\mathbf{z} \mid \boldsymbol{\rho}^*, \lambda, a_\lambda\right)$.

We get an optimal ordering ρ^* in Step 2. To sample $\mathbf{z}$, we define a matrix Γ. The matrix stores the comparison result in the forward and backward step. More specifically, it stores the ordering of variables in the forward step, followed by the ordering in the backward step. Then:

For $i = 1, 2$ (denoting the forward step and backward step) and $j = 1, \ldots, n-1$, sample

$$z_{ij} \mid \Gamma, \lambda \sim \mathcal{E}\left(\sum_{m=j}^{n} \lambda_{\rho_{im}^*}\right).$$

Step 4: $p\left(\lambda \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda\right) = p\left(\lambda \mid \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda\right)$.

For $i = 1, \ldots, n$, sample

$$\lambda_i \mid \mathbf{z}, a_\lambda \sim \mathcal{G}\left(a_\lambda + w_i, b_\lambda + \sum_{i=1}^{2}\sum_{j=1}^{n} z_{ij}\right).$$

where w_i denotes the total number of wins of variable i. If we let w_{ij} denote the number of comparisons where i beats j, then $w_i = \sum_{j=1, j\neq i}^{n} w_{ij}$. The total number of comparisons between i and j, N_{ij}, has the relationship that $N_{ij} = w_{ij} + w_{ji}$.

Step 5: $p\left(a_\lambda \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \mathbf{z}, \lambda\right) = p\left(a_\lambda \mid \lambda\right)$.

We use a random walk sampler in which candidate draws, denoted $\log\left(a_\lambda^c\right)$, is obtained from $\mathcal{N}\left(\ln\left(a_\lambda\right), \omega^2\right)$, where ω^2 is a tuning parameter and we set $\omega^2 = 0.1^2$ in our application. The acceptance probability for a_λ^c is given by

$$\min\left\{1, b_\lambda^{n(a_\lambda^c - a_\lambda)}\left(\frac{\Gamma\left(a_\lambda\right)}{\Gamma\left(a_\lambda^c\right)}\right)^n \left(\lambda\right)^{a_\lambda^c - a_\lambda}\right\},$$

where $\Gamma(\cdot)$ denotes the Gamma function.

4 Application

We use the data in Chan (2021). All definitions and priors are the same, except that matrix $\mathbf{B}_0$ can be (via permutations of the columns of the matrix) put into a lower triangular form with ones on the diagonal.

4.1 Posterior Ordering

Posterior ordering is calculated as: For each model (BVARSV-FBPL, or BVARSV-FBBT), we store the ordering in each iteration (after burn-in) (denoted as $\rho^{*,t}$. Note that $\rho^{*,t}$ is an $n \times 1$ vector. The first element in $\rho^{*,t}$ is the variable which takes the first place). Then we will have a matrix P^*. Each row in P^* denotes the ordering in that iteration. For instance, the t-th row in P^* is $\rho^{*,t}$. Defined in this way, the first column

in P^* will be the variable which takes the first place. We can compute the frequency of each variable. This is shown in the first column. Similarly, we compute the frequency for other columns.

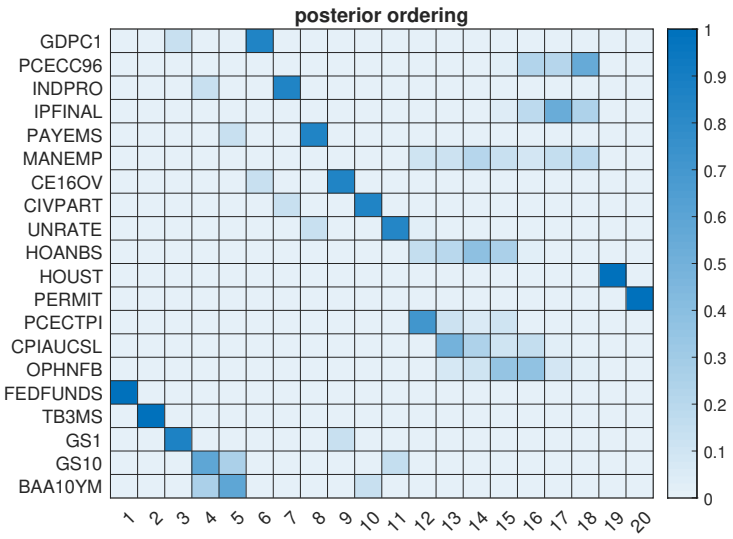

Figure 1: Posterior Ordering: BVARSV-FBPL (Multiple Comparisons)

4.2 Forecasting Results

The forecasting design adopted is iterative forecasting. We consider an initial estimation period from 1960Q1 to 1987Q4. The remaining observations (1988Q1 to 2021Q3) are used as a hold-out period to evaluate our forecasting methods. After obtaining $h \in \{1, 2, 3, 4\}$-step-ahead predictive distributions for a given period in the hold-out, we include this period in the estimation sample and repeat this procedure until we reach the end of the sample. To assess forecasting accuracy, we use root mean square forecast errors (RMSFEs) for point forecasts and average log predictive likelihoods (ALPLs) for density forecasts. To compare each model M against the benchmark B, we therefore consider the percentage gains in terms of RMSFE, defined as

$$(1 - \text{RMSFE}_{i,h}^{M}/\text{RMSFE}_{i,h}^{B}) \times 100$$

and the percentage gain in terms of ALPL, which is

$$(\text{ALPL}_{i,h}^{M} - \text{ALPL}_{i,h}^{B}) \times 100$$

Table 1 reports the percentage gains in RMSFE relative to the standard BVARSV model, and the percentage gains in ALPL relative to the standard BVARSV model. We find that

there is not much difference between their RMSFEs, but find substantive forecasting improvements relative to a strategy of subjectively selecting a single ordering. Table 2 reports the forecasting result about joint forecasting. It also confirms that our approach provides substantive forecasting improvements relative to a strategy of subjectively selecting a single ordering.

Table 1: Forecasting Results

	% gains in RMSFE		% gains in ALPL	
	h=1	h=4	h=1	h=4
GDPC1	-5.51	-0.49	13.51	51.20
INDPRO	0.34	-0.83	14.70	2.79
UNRATE	-9.50	-18.65	97.68	265.95
CPIAUCSL	0.23	1.32	3.15	16.98

Table 2: Joint Forecasting (% gains in ALPL)

	h=1	h=4
All	4530.41	7097.65
First 4	946.00	1260.42
Last 4	159.34	490.81

References

J. E. Arias, J. F. Rubio-Ramirez, and M. Shin. Macroeconomic forecasting and variable ordering in multivariate stochastic volatility models. *Federal Reserve Bank of Philadelphia Working Papers*, 2021.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

F. Caron and A. Doucet. Efficient bayesian inference for generalized bradley–terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

A. Carriero, T. E. Clark, and M. Marcellino. Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212:137–154, 2019.

J. C. Chan. Minnesota-type adaptive hierarchical priors for large bayesian vars. *International Journal of Forecasting*, 37(3):1212–1226, 2021.

J. C. Chan, G. Koop, and X. Yu. Large order-invariant bayesian vars with stochastic volatility. *arXiv preprint arXiv:2111.07225*, 2021.

D. A. Henderson and L. J. Kirrane. A comparison of truncated and time-weighted plackett–luce models for probabilistic forecasting of formula one results. *Bayesian Analysis*, 13(2):335–358, 2018.

S. Johnson, D. Henderson, and R. Boys. On bayesian ifnerence for the extended plackett-luce model. *Bayesian Analysis*, 17:1–26, 2021.

R. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

R. Plackett. The analysis of permutations. *Applied Statistics*, 24:193–202, 1975.

Appendices

A Data

Table 3: Description of variables used in the forecasting application

Variable	Mnemonic	Transformation
Real Gross Domestic Product	GDPC1	400Δ log
Personal Consumption Expenditures	PCECC96	400Δ log
Industrial Production Index	INDPRO	400Δ log
Industrial Production: Final Products	IPFINAL	400Δ log
All Employees: Total nonfarm	PAYEMS	400Δ log
All Employees: Manufacturing	MANEMP	400Δ log
Civilian Employment	CE16OV	400Δ log
Civilian Labor Force Participation Rate	CIVPART	no transformation
Civilian Unemployment Rate	UNRATE	no transformation
Nonfarm Business Section: Hours of All Persons	HOANBS	400Δ log
Housing Starts: Total	HOUST	400Δ log
New Private Housing Units Authorized by Building Permits	PERMIT	400Δ log
Personal Consumption Expenditures: Chain-type Price index	PCECTPI	400Δ log
Consumer Price Index for All Urban Consumers: All Items	CPIAUCSL	400Δ log
Nonfarm Business Section: Real Output Per Hour of All Persons	OPHNFB	400Δ log
Effective Federal Funds Rate	FEDFUNDS	no transformation
3-Month Treasury Bill: Secondary Market Rate	TB3MS	no transformation
1-Year Treasury Constant Maturity Rate	GS1	no transformation
10-Year Treasury Constant Maturity Rate	GS10	no transformation
Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity	BAA10YM	no transformation

B The Bradley-Terry model

B.1 A Bradley-Terry Prior over Orderings

The Bradley-Terry prior over ordering is

$$\Pr\left(\min\{z_{\rho_i^*}, z_{\rho_j^*}\}|\mathbf{z}_{-\rho_i^*,-\rho_j^*}\right) \tag{13}$$

This means that the variable that given the remaining $(n-2)$ places, we can compare the i-th place and $(i+1)$-th place. This is pair comparison and equation (**??**) is the Bradley Terry model. The next question is which two places to compare.

We want to mention that to bring a sequence into order, successively applying adjacent transpositions, is always possible. Moreover, any reasonable choice for the adjacent transpositions will work. In this paper, we consider the choice as: from the first to the last, then from the last to the first. We denote this method as the forward and backward Bradley-Terry (hereafter FBBT) model.

Like the FBPL method, it also consists of two steps. One is the forward step, another is the backward step. In the forward step, we repeatedly step through the sequence (from the first place to the last place), compares adjacent elements and swaps them if they are in the wrong order. The pass through the sequence is repeated until there is no swaps (this procedure is known as Bubble sort). Then, in the backward step, we repeatedly step through the sequence (from the last place to the first place), compares adjacent elements and swaps them if they are in the wrong order. The pass through the sequence is repeated until there is no swaps. Finally, we will get the optimal ordering $\boldsymbol{\rho}^*$.

Here we describe details in the forward step. For example, we randomly select a ordering to start as $\boldsymbol{\rho}^0$:

1) the first pass is from $i = 1$ to $i = n$: compare ρ_i^0 and ρ_{i+1}^0, swap them if they are in the wrong order. From the first pass, we can get a new ordering say $\boldsymbol{\rho}^1$;

2) the second pass is from $i = 1$ to $i = n$: compare ρ_i^1 and ρ_{i+1}^1, swap them if they are in the wrong order. From the second pass, we can get a new ordering say $\boldsymbol{\rho}^2$;

$\vdots$

repeat this procedure until there is no swap.

To complement the Bayesian estimation, we need a prior on the pair comparisons. There is a simple form for the prior distribution (4)

$$\Pr\left(\min\{z_{\rho_i^*}, z_{\rho_j^*}\}|\mathbf{z}_{-\rho_i^*,-\rho_j^*}\right) \sim \mathcal{E}\left(\lambda_{\rho_i^*} + \lambda_{\rho_j^*}\right). \tag{14}$$

The selected variable is distributed as

$$\Pr\left(z_{\rho_i^*} = \min\{z_{\rho_i^*}, z_{\rho_j^*}\}|\mathbf{z}_{-\rho_i^*,-\rho_j^*}\right) = \frac{\lambda_{\rho_i^*}}{\lambda_{\rho_i^*} + \lambda_{\rho_j^*}}. \tag{15}$$

This is: given any remaining $(n-2)$ places, we can compare the i-th place and j-th place. This is pair comparison and equation (15) is the Bradley Terry model.

B.2 Bayesian Computation

Step 1 and Step 5 are the same as FBPL. So here we describe Step 2 to Step 4.

Step 2: $p\left(\boldsymbol{\rho}^* \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \mathbf{z}, \lambda\right) = p\left(\boldsymbol{\rho}^* \mid \mathbf{B}_0, \mathbf{z}, \lambda\right)$.

Step 2 consists of two steps: the forward step and the backward step.

The forward step:

Suppose that the ordering before updating is $\boldsymbol{\rho}^0$:

We create a matrix W. Matrix W is an $n \times n$ matrix. Let w_{ij} denote the number of comparisons where i beats j, $w_i = \sum_{j=1, j\neq i}^{n} w_{ij}$, the total number of wins of element i, and $N_{ij} = w_{ij} + w_{ji}$, the total number of comparisons between i and j.

1) the first pass is from $i = 1$ to $i = n$: compare ρ_i^0 and ρ_{i+1}^0, swap them if they are in the wrong order, and update matrix W.

The comparison involves $b_{\rho_i^0,\rho_{i+1}^0}$ and $b_{\rho_{i+1}^0,\rho_i^0}$. The success probability is to assess whether the coefficient $b_{\rho_i^0,\rho_{i+1}^0}$ is generated from the Dirac spike component (Here we use a small variance c), and the coefficient $b_{\rho_{i+1}^0,\rho_i^0}$ is generated from the slab component. Let p_s denote the success probability. Then:

$$p_s = p\left(\rho_i^0 \text{ beats } \rho_{i+1}^0 \mid b_{\rho_i^0,\rho_{i+1}^0}, b_{\rho_{i+1}^0,\rho_i^0}\right) \propto \frac{\lambda_{\rho_i^0}}{\lambda_{\rho_i^0} + \lambda_{\rho_{i+1}^0}} \phi\left(b_{\rho_i,\rho_{i+1}}; 0, c\right) \phi\left(b_{\rho_{i+1},\rho_i}; a_b, V_b\right),$$

The failure probability is to assess whether the coefficient $b_{\rho_i^0,\rho_{i+1}^0}$ is generated from the slab component, and the coefficient $b_{\rho_{i+1}^0,\rho_i^0}$ is generated from the spike component. Let p_f denote the success probability. Then:

$$p_f = p\left(\rho_{i+1}^0 \text{ beats } \rho_i^0 \mid b_{\rho_i^0,\rho_{i+1}^0}, b_{\rho_{i+1}^0,\rho_i^0}\right) \propto \frac{\lambda_{\rho_{i+1}^0}}{\lambda_{\rho_i^0} + \lambda_{\rho_{i+1}^0}} \phi\left(b_{\rho_i,\rho_{i+1}}; a_b, V_b\right) \phi\left(b_{\rho_{i+1},\rho_i}; 0, c\right),$$

Hence, after normalization, we obtain

$$p_s = \frac{p_s}{p_s + p_f}.$$

From the first pass, we can get a new ordering say $\boldsymbol{\rho}^1$.

2) the second pass is from $i = 1$ to $i = n$: compare ρ_i^1 and ρ_{i+1}^1, swap them if they are in the wrong order, and update matrix W. From the second pass, we can get a new ordering say $\boldsymbol{\rho}^2$;

$\vdots$

Repeat this procedure until there is no swap. This will be the optimal ordering $\boldsymbol{\rho}^*$.

The backward step:

This is almost the same as the forward step, except that the pass is from $i = n$ to $i = 1$.

To implement Step 3 and Step 4, we follow the sampling approach in Caron and Doucet (2012):

Step 3: $p(\mathbf{z} \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \lambda, a_\lambda) = p(\mathbf{z} \mid \boldsymbol{\rho}^*, \lambda, a_\lambda)$.

For $1 \leqslant i < j \leqslant n$, sample

$$z_{ij} \mid \boldsymbol{\rho}^*, \lambda \sim \mathcal{G}(N_{ij}, \lambda_i + \lambda_j),$$

where N_{ij} is the total number of comparisons between variable i and variable j, and $N_{ij} > 0$. We can get this from Step 2.

Step 4: $p(\lambda \mid \mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{B}_0, \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda) = p(\lambda \mid \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda)$.

For $i = 1, \ldots, n$, sample

$$\lambda_i \mid \boldsymbol{\rho}^*, \mathbf{z}, a_\lambda \sim \mathcal{G}\left(a_\lambda + w_i, b_\lambda + \sum_{i<j} z_{ij} + \sum_{i>j} z_{ji}\right).$$

where w_i denotes the total number of wins of variable i. If we let w_{ij} denote the number of comparisons where i beats j, then $w_i = \sum_{j=1, j\neq i}^{n} w_{ij}$. The total number of comparisons between i and j, N_{ij}, has the relationship that $N_{ij} = w_{ij} + w_{ji}$.