Achieving Shrinkage and Sparsity in Bayesian Vector Autoregressions with Three-Parameter-Beta-Normal Prior

Rui Meng, Harivallabha Rangarajan, Kristofer Bouchard
Lawrence Berkeley National Laboratory

July 2021

Abstract

Vector autoregressive (VAR) models have been widely used for modeling the temporal dependence of multivariate time series. Global-local priors are widely used to induce shrinkage in such models. In this article, we develop a very flexible model, the three-parameter-beta-normal (TPBN) shrinkage prior based VAR with stochastic volatility considerations. Two efficient inference techniques, Markov Chain Monte Carlo (MCMC) inference and empirical Bayes inference approach are proposed. The empirical Bayes inference approach is achieved by the marginal maximum likelihood (MML), allowing our prior to adapt to both sparse and dense settings. Moreover, the corresponding post-analysis is proposed to induce the true sparsity into model estimates under a fully Bayesian framework. An extensive simulation study demonstrates that our model accurately captures the sparse structure in the data, and significantly outperforms other state-of-the-art models in both parameter estimation and variable selection. Finally, we show that our approach has a promising forecast performance compared to other competing models on the real macroeconomic data.

Keywords: empirical Bayes, high dimensional, scale mixtures of normal distributions.

1 Introduction

Vector autoregressive models (VARs) have been widely used to capture the temporal dependence of multivariate time series, especially in macroeconomic forecasting contexts. [51] is a seminal work in which VAR is applied to macroeconomic research for policy makers. VARs are preferred because of the flexibility and rich parametrization they offer. Their utility can further be enhanced by extensions that incorporate time-varying parameters [46, 32] and stochastic volatility [31] into the framework. However, all of those models suffer from the curse of dimensionality which affects their scalability, and also causes unreliable forecasts.

In the Bayesian literature, several ways have been proposed to alleviate the curse of dimensionality in a VAR [20, 53, 24, 25, 31, 21, 37]. Notably, [20] propose the Minnesota prior, which imposes an informative prior to reduce estimation error. Specifically, the Minnesota prior shrinks the diagonal elements of the coefficient matrix at lag order 1 to ones and shrinks the remaining coefficients toward zeros. [25] extend the work in [53] and study the optimal choice of the informativeness of the prior in the spirit of hierarchical modelling. Other variations include the sum of coefficients [20] and the dummy initial observation prior [52].

Shrinkage priors are categorized into two classes, two-groups priors or spike-and-slab priors [33], and global-local shrinkage priors [11, 5]. The spike-and-slab prior is a discrete mixture of a point mass at zero (the spike) and an absolutely continuous density (the slab) while the global-local prior is considered as the normal scale mixture distribution. Mathematically, those priors for random variables $\{\theta_i\}$ are hierarchically modeled as:

$$\begin{aligned} (\theta_i|\lambda_i,\tau) &\sim \mathcal{N}(0,\lambda_i^2\tau^2)\,, \\ \lambda_i &\sim g_{local}(\lambda_i)\,, \qquad \tau \sim g_{global}(\tau)\,, \end{aligned} \tag{1}$$

where λ_i is the local term characterising the individual behavior and τ is the global term providing substantial shrinkage towards zero. Many variants within the class of global-local shrinkage priors are proposed by differently modeling the local and global distributions. [29] and [2] model λ^2 with the same exponential prior but model τ with the Jeffreys prior and a Gamma prior, respectively. [11] proposed a horseshoe prior by employing a Half Cauchy prior for both λ and τ. [4] extend the horseshoe to the horseshoe+ prior by introducing a hierarchical structure for the local term. Other widely

used priors are the Dirichlet-Laplace prior proposed in [8] and the three-parameter-beta prior in [1]. Such a shrinkage prior setup has at least two convenient features for VAR with exogenous inputs. First, it exerts a strong degree of shrinkage on all elements of $\boldsymbol{C}$ but still provides additional flexibility such that nonzero regression coefficients are permitted. A large class of global-local shrinkage priors share this property [44, 12].

Motivated by the theory introduced in [19], shrinkage priors are studied in VAR models. It states that forecasts are highly correlated to principle components and Bayesian methods can perform equally well with appropriate choice of a shrinkage prior. [40] develop a Bayesian methodology to perform model variable selection using spike and slab priors, called the stochastic search variable selection (SSVS) method. This class of models is flexible in prior specification and performs well in terms of prediction relative to traditional shrinkage priors. The limitation of SSVS is the expensive computation, which requires exploiting 2^N possible combinations of sparse choices in Markov chain Monte Carlo (MCMC), where N is the number of model parameters. So it suffers from slow convergence and is intractable for large datasets. Alternatively, global-local shrinkage priors are getting attention for VARs due to their computational efficiency. In Bayesian VAR literature, [31] study the Normal-Gamma prior, [21] explore the Horseshoe prior, and [32] exploit the Dirichlet-Laplace prior.

This paper proposes a flexible shrinkage prior, the three-parameter-beta (TPB) distribution for a Bayesian VAR model with stochastic volatility, and addresses an important consideration pertaining to the selection of hyperparameters. The question of hyper-parameter selection for global-local priors is also highlighted in [43]. To this end, two approaches are proposed. In the first approach, we place hierarchical priors and conduct inference via an efficient adaptive Metropolis Hasting algorithm [48]. But the hierarchical Bayes approach does not quite address the issue of misspecification of hyper-parameters, resulting from additional specification of priors. Instead of specifying values for hyper-parameters, we propose a Monte Carlo Expectation Maximization (MCEM) algorithm that can learn optimal hyper-parameters from data. Thus it is treated as an Empirical Bayes method [13]. Through several experiments on synthetic data, we illustrate that our model achieves better model fitting performance than other state-of-the-art models.

One caveat of typical global-local shrinkage priors is that they cannot naively perform variable selection, as they shrink the values towards zero, in a continuous spectrum without forcing the values to be exact zeros. Therefore,

those models with global-local priors cannot directly generate true sparse estimates. [32, 30] induce sparsity and shrinkage in both the time-varying parameter model and conjugated VAR respectively. Specifically, they consider a two-stage approach with the signal adaptive variable selector (SAVS). In our work, instead of using SAVS, we minimize posterior risk (posterior expected loss) with respect to parsimonious estimates in [28] via adaptive Lasso and sample parsimonious estimates following the approach in [56]. We conduct this post-analysis on our model and other competitive models and decompose the fitting measures into two - the model selection measure and the bias measure. We find our proposed model still outperforms other models in both fitting measurements, in the synthetic case study.

The remainder of the article is structured as follows. Section 2 discusses the vector autoregressions with stochastic volatility case in the context of the three-parameter-beta-normal prior. Section 3 proposes both MCMC Bayesian inference and MCEM empirical Bayesian inference. In Section 4, we develop the variable selection procedure in the context of VAR models. Section 5 carries out an extensive simulation study and shows the superior estimation performance of our model compared to state-of-the-art models; we also show robust forecasting performance on real macroeconomic data. The conclusions and discussions are presented in Section 6.

2 Model

This section first introduce the general vector autoregressive (VAR) model, and then provide the details on the three-parameter-beta-normal prior in the context of VARs.

2.1 Vector Autoregressive Model (VAR)

Suppose $\boldsymbol{y}_t \in \mathbb{R}^M$ follows a VAR(P) process, which satisfies the following recursion,

$$\boldsymbol{y}_t = \sum_{i=1}^{P} \boldsymbol{A}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t \,, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Omega}_t) \tag{2}$$

where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_P$ are real-valued $M \times M$ matrices of autoregressive coefficients; $\boldsymbol{\epsilon}_t$ is M dimensional Gaussian noise with zero mean and non-

degenerate time-dependent variance-covariance matrix $\boldsymbol{\Omega}_t$. [7, 57] assume $\boldsymbol{\Omega}_t$ is deterministic and diagonal, and use inverse gamma prior to model error variance parameters. Several works assume $\boldsymbol{\Omega}_t$ is deterministic but not diagonal and model it via different approaches. [25, 38] model the coefficients via normal-inverse-Wishart family; [24, 53, 55] consider the Cholesky decomposition of the precision matrix and impose restrictions on the lower triangular matrix. [45, 54] reparametrize Σ via the Cholesky factorization, $\boldsymbol{\Omega}_t = \Gamma'\Lambda\Gamma$ where Γ is a lower triangular matrix with 1's on the diagonal and Λ is a diagonal matrix of the eigenvalues of $\boldsymbol{\Omega}$. Recently, time-varying variance-covariance $\boldsymbol{\Omega}_t$ has garnered more attention. [32] consider a general factorization such that $\boldsymbol{\Omega}_t = \boldsymbol{U}_t\boldsymbol{S}_t\boldsymbol{U}_t'$, where $\boldsymbol{U}_t$ denotes a lower uni-triangular matrix following independent random walks, and $\boldsymbol{S}_t$ is positive diagonal matrix.

Following [17, 31] we consider the stochastic volatility model for time-varying noise. Specifically, We factorize $\boldsymbol{\Omega}_t$ as

$$\boldsymbol{\Omega}_t = \boldsymbol{U}\boldsymbol{H}_t\boldsymbol{U}' \tag{3}$$

where $\boldsymbol{U}$ denotes a lower triangular matrix with unit diagonal and $\boldsymbol{H}_t = \text{diag}(e^{h_{1t}}, \ldots, e^{h_{Mt}})$ is a diagonal matrix. We rewrite the error term as $\boldsymbol{\epsilon}_t = \boldsymbol{U}\boldsymbol{H}_t^{0.5}\boldsymbol{v}_t$ where $\boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Here, $\boldsymbol{U}$ is assumed to be fixed over time, because [46, 9] found little variation in such coefficients. h_{jt}'s are called log-volatilities and commonly modeled by independent AR(1) processes as

$$h_{jt} = \mu_j + \rho_j(h_{jt-1} - \mu_j) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma_{\eta j}^2) \tag{4}$$

where μ_j denotes the unconditional mean, ρ_j the persistence parameter, and $\sigma_{\eta t}$ the error variance of the log volatility process. The prior specification on the parameters of the log volatility equation follows [35]. Recently, [37] propose a factor stochastic volatility model to model the error term for huge-dimensional data.

[16, 15] propose an approach to fast inference coefficients. However, it requires a specific structure of the coefficient prior variances, which cannot be applied to our model. [9, 36, 39] emphasize the computational gains that arise from a transformed system, where the errors in different equations are mutually independent of one another. Assume $\boldsymbol{U}$ is known. Letting $y_{jt}^* = y_{jt} - \sum_{l=1}^{j-1} u_{jl}e^{0.5h_{lt}}v_{lt}$, we have M conditional independent processes and the generic equation for dimension j is written as

$$\boldsymbol{y}_{j\cdot}^* = \boldsymbol{X}\boldsymbol{\alpha}_j + \text{diag}(\sigma_{jj1}^*, \ldots, \sigma_{jjT}^*)\boldsymbol{v}_j\,, \tag{5}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$ and $\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-P})'$, $\boldsymbol{v}_j = (v_{j1}, \ldots, v_{jT})'$ and $\boldsymbol{\alpha}_j$ be the jth row of the coefficient matrix $A = (A_1, \ldots, A_p)$. Also $\sigma^*_{jjt} = u_{jj}e^{0.5h_{jt}} = e^{0.5h_{jt}}$. When we assume $\boldsymbol{U}$ is unknown, we also have M conditional independent processes and the generic equation for variable j is written as

$$\boldsymbol{y}_{j\cdot} = \boldsymbol{X}_{*j}\boldsymbol{\alpha}_{*j} + \text{diag}(\sigma^*_{jj1}, \ldots, \sigma^*_{jjT})\boldsymbol{v}_j\,, \tag{6}$$

where $\boldsymbol{X}_{*j} = (\boldsymbol{x}_{*j1}, \ldots, \boldsymbol{x}_{*jT})'$ and $\boldsymbol{x}_{*jt} = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-P}, e^{0.5h_{1t}}v_{1t}, \ldots, e^{0.5h_{j-1,t}}v_{j-1,t})'$, and $\boldsymbol{\alpha}_{*j} = (\boldsymbol{\alpha}'_j, u_{j1}, \ldots, u_{j,j-1})'$.

For each equation 5, with scale mixture prior on coefficients $\boldsymbol{\alpha}_j$, we allow the fast sampling method in [6] to sample $\boldsymbol{\alpha}_j$ from its posterior distribution. Specifically, to sample any random vector from a structured multivariate Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$, where $\Sigma = (\Phi'\Phi + D^{-1})^{-1}$, $\mu = \Sigma\Phi'\alpha$, $D \in \mathbb{R}^{p\times p}$ is symmetric positive definite, $\Phi \in \mathbb{R}^{n\times p}$, and $\alpha \in \mathbb{R}^{n\times 1}$. The fast inference procedures are designed as

Algorithm 1: Proposed algorithm

(i) Sample $u \sim \mathcal{N}(0, D)$ and $\delta \sim \mathcal{N}(0, I_n)$ independently.

(ii) Set $v = \Phi u + \delta$.

(iii) Solve $(\Phi D\Phi' + I_n)w = (\alpha - v)$.

(iv) $\theta = u + D\Phi' w$.

When D is diagonal, as in the of case of global-local priors in (1), the complexity of the proposed algorithm is $\mathcal{O}(n^2p)$. In comparison to the $\mathcal{O}(p^3)$ complexity of the completing algorithm in [50], when $p \gg n$, this algorithm offers huge computational gains.

2.2 Three-parameter-beta-normal prior

The three-parameter-beta prior (TPB) distribution for a random variable X is defined by the density function

$$f(x; a, b, \phi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\phi^b x^{b-1}(1-x)^{a-1}\{1 + (\phi - 1)x\}^{-(a+b)}\,, \tag{7}$$

for $0 < x < 1$, $a > 0$, $b > 0$ and $\phi > 0$ and is denoted by TPB(a, b, ϕ). It is a subclass of the Gauss hypergeometric (GH) distribution proposed in [3]

and the compound confluent hypergeometric (CCH) distribution proposed in [26]. a and b control the spike at zero and the tail characteristics, respectively. When $0 < a < 1$ and $0 < b < 1$, the density in (7) is bimodal, with modes at 0 and 1. ϕ controls the strength of shrinkage effect. Smaller values of ϕ put greater probabilities on the neighborhood of $X = 1$ while larger values of ϕ move more probabilities towards $X = 0$ [1]. With $\phi = 1$, this distribution is identical to a beta distribution.

The normal scale mixture distribution constructed with three-parameter-beta prior, termed three-beta-parameter-normal (TPBN) prior [1], for parameter β, is defined by

$$\beta|\psi \sim \mathcal{N}(0, \frac{1}{\psi} - 1)\,, \qquad \text{with} \qquad \psi \sim \text{TPB}(a, b, \phi) \tag{8}$$

where the shrinkage parameter ψ follows TPB distribution. The bimodal property of ψ induces the shrinkage behavior: When ψ approaches 0, it creates a diffuse prior on β; when ψ is near 1, it induces strong shrinkage on β. And decreasing ϕ leads ψ close to 1 and then supports stronger shrinkage [1, 57]. Note that the special case for $a = b = 1/2$ gives the horseshoe prior for β [11] and when $a = \phi = 1$ and $b = 1/2$, the distribution of β is the Strawderman-Berger distribution. An alternative representation of the TPBN can be derived via replacing $\frac{1}{\psi} - 1$ by τ. The hierarchical representation is

$$\beta|\tau \sim \mathcal{N}(0, \tau), \tau \sim \mathcal{G}(a, w), \text{and}, w \sim \mathcal{G}(b, \phi) \tag{9}$$

where $\mathcal{G}$ refers to a gamma distribution. Note the TPBN is a hierarchical variant of the Normal-Gamma prior [27].

In the context of VARs, we employ the TPBN prior to autoregressive coefficients $\boldsymbol{A}$ and discuss the column-wise and row-wise layers of shrinkage with scaling parameters. We notice that row-shrinkage and column-shrinkage modifications have been applied to group factor analysis [57] and factor stochastic volatility models [34] and VAR models [31]. The lag-specific shrinkage modification has been studied in a static factor framework in [7].

We place the TPBN priors on coefficient matrix $A \in \mathbb{R}^{T \times MP}$. Each element a_{ij} depends on global shrinkage parameter λ and individual shrinkage parameter ψ_{ij} and the prior is given by

$$a_{ij}|\lambda, \psi_{ij} \sim \mathcal{N}\left(0, \frac{1}{\lambda}(\frac{1}{\psi_{ij}} - 1)\right), \;\; \psi_{ij} \sim \text{TPB}(a, b, \rho)\,, \tag{10}$$

for $i = 1, \ldots, T, j = 1, \ldots, MP$.

The hyper-parameters in the TPB distribution can be chosen depending on the specific dataset. We place Gamma prior on global shrinkage parameter, $\lambda \sim \mathcal{G}(c, d)$. The hyper-parameters are usually set as $c = 0.01$ and $d = 0.01$ for large variance in the prior. Based on the hierarchical representation in (9), by letting $\theta_i = \frac{1}{\psi_i} - 1$ and introducing dummy variables δ_{ij} we have the hierarchical representation as

$$a_{ij}|\lambda, \theta_{ij} \sim \mathcal{N}(0, \frac{1}{\lambda}\theta_{ij}), \qquad \theta_{ij} \sim \mathcal{G}(a, \delta_{ij}), \qquad \delta_{ij} \sim \mathcal{G}(b, \rho). \tag{11}$$

For hyper-parameter ρ, we put a conjugate prior such that $\rho^{\frac{1}{2}} \sim \mathcal{C}^{+}(0, 1)$ where $\mathcal{C}^{+}(0, 1)$ is the standard half-Cauchy distribution.

And in terms of the principle of parsimony, we put the TPBN priors on the free-diagonal elements of $\boldsymbol{U}$ in (3) to put them close to zeros.

3 Inference

This section proposes two efficient inference schemes motivated by the challenge of learning shape parameters a and b in Equation 11. Since TPBN prior is very flexible and there exist no conjugate priors on shape parameters, learning those hyper-parameters is a difficult task. The two inference schemes are detailed below.

In the first scheme, we employ a hierarchical Bayes approach via assuming $a \sim \text{Exp}(1), b \sim \text{Exp}(1)$. Since both priors are non-conjugate, we sample them via an adaptive metropolis-with-Gibbs algorithm [48]. The hierarchical Bayes approach is a fully Bayesian method allowing the uncertainty measures on both parameters. However, the MCMC inference mixes poorly even with the efficient adaptive sampling approach. If the uncertainty quantification of hyper-parameters is not of interest, we advocate an alternative empirical Bayes inference schedule. Without specifying any additional prior, we select the hyper-parameters by maximizing the marginal likelihood. Thus it alleviates the burden of hyper-parameter tuning by the user. Specifically, in accordance with the model specification in (2) and (3), the marginal likelihood, $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{H})\pi(\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{H})d(\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{H})$ measures the performance of the model with respect to its prior. Maximizing the marginal likelihood with respect the hyper-parameters of interest can help us learn the most likely sparsity level from the data. We present MCMC inference in Section 3.1.

Given the derivation of the MCMC inference, we provide the MCEM inference in Section 3.2.

3.1 Posterior inference

This section describes the Gibbs sampling scheme. We derive and illustrate some important conditional posterior distributions in the section and leave other details to the Appendix.

The parameters of $\boldsymbol{A}$ and $\boldsymbol{U}$ are sequentially and conditionally sampled from the posterior distribution according to the M unrelated processes mentioned in (5). Mathematically, we have

$$p(\boldsymbol{A}, \boldsymbol{U}|\boldsymbol{H}, \boldsymbol{y}) \propto p(\boldsymbol{\alpha}_1|\boldsymbol{H}, \boldsymbol{y}) \prod_{j=2}^{M} p(\boldsymbol{\alpha}_j, \boldsymbol{u}_j|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{j-1}, \boldsymbol{H}, \boldsymbol{y}) \tag{12}$$

where $\boldsymbol{H} = (\boldsymbol{H}_1, \ldots, \boldsymbol{H}_T)$ and $\boldsymbol{u}_i$ denote the free parameters on the ith row of $\boldsymbol{U}$.

For the jth term in (12), the conditional posterior can be written as

$$\begin{aligned} & p(\boldsymbol{\alpha}_j, \boldsymbol{u}_j|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{j-1}, \boldsymbol{H}, \boldsymbol{y}) \\ \propto & \mathcal{N}(\boldsymbol{y}_{\cdot j}|\boldsymbol{X}_{*j}\boldsymbol{\alpha}_{*j}, \mathrm{diag}(\sigma^{*2}_{jj1}, \ldots, \sigma^{*2}_{jjT}))\mathcal{N}(\boldsymbol{\alpha}_j|\boldsymbol{\mu}_{A,j}, \mathrm{diag}(\boldsymbol{\sigma}^2_{A,j}))\mathcal{N}(\boldsymbol{u}_j|\boldsymbol{\mu}_{B,j}, \mathrm{diag}(\boldsymbol{\sigma}^2_{U,j})) \\ = & \mathcal{N}([\boldsymbol{\alpha}'_j, \boldsymbol{u}'_j]'|A_j^{-1}(\boldsymbol{C}'_{*j}\boldsymbol{z}_j + \boldsymbol{\mu}'_j\boldsymbol{\Lambda}^{-1}_{*j}), A_j^{-1}) \end{aligned} \tag{13}$$

where $A_j = (\boldsymbol{C}'_{*j}\boldsymbol{C}_{*j} + \Lambda^{-1}_*)$ with $\boldsymbol{C}_{*j} = \mathrm{diag}(\sigma^{*-1}_{jj1}, \ldots, \sigma^{*-1}_{jjT})\boldsymbol{X}_{*j}$, $\boldsymbol{z}_j = \mathrm{diag}(\sigma^{*-1}_{jj1}, \ldots, \sigma^{*-1}_{jjT})\boldsymbol{y}_{\cdot j}$, $\boldsymbol{\mu}_j = (\boldsymbol{\mu}'_{A,j}, \boldsymbol{\mu}'_{B,j})'$ and $\Lambda_{*j} = \mathrm{diag}([\boldsymbol{\sigma}^2_{A_j}, \boldsymbol{\sigma}^2_{B_j}])$. When assuming zero-mean prior $\boldsymbol{\mu}_j = \boldsymbol{0}$, and $MP \gg T$, we would employ the algorithm 1 to fast sample model parameters $\boldsymbol{A}$ and $\boldsymbol{U}$.

The conditional posteriors of auxiliary variables in TPBN for $\boldsymbol{A}$ and $\boldsymbol{U}$ can be derived as closed-form expression. Here, we show those conditional posteriors for coefficient matrix $\boldsymbol{A}$ for example. Given the hierarchical representation in (11), the conditional posteriors of local-shrinkage scale variables θ_{ij} and δ_{ij} follow a generalized inverse Gaussian (GIG) distribution and a Gamma distribution as

$$\begin{aligned} \theta_{ij}|a_{ij}, \lambda, a &\sim \mathcal{GIG}(a - 0.5, 2\delta_{ij}, a^2_{ij}\lambda)\,, \\ \delta_{ij}|a, \theta_{ij} &\sim \mathcal{G}(b + a, \rho + \theta_{ij})\,. \end{aligned} \tag{14}$$

The conditional posterior distribution of global-shrinkage scale variable λ is

$$\lambda|A,\theta,c,d \sim \mathcal{G}(c+0.5k, d+0.5\sum_{i=1}^{T}\sum_{j=1}^{MP}\frac{a_{ij}^2}{\theta_{ij}})\,. \tag{15}$$

The sampling scheme for other hyper-parameters a, b and ρ in the TPB prior are discussed in the Appendix. We adapt the efficient Gibbs sampling algorithm in [35] to simulate the full history of log volatilities $\boldsymbol{h}_m = (h_{m1},\ldots,h_{mT})$ for $m = 1,\ldots,M$.

3.2 Empirical Bayesian approach

We propose an Monte Carlo Expectation Maximization algorithm to obtain the MML estimates of $\boldsymbol{\phi} = (a_A, b_A, a_U, b_U)$. The approach is self-adaptive as the hyper-parameters can be learned from the data.

Given the TPBN based hierarchical representation in (11), the joint likelihood of our model is given by

$$\begin{aligned}
&\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{U},\boldsymbol{H}) + \sum_{S=(A,U)}\big[\log p(\boldsymbol{S}|\lambda_S,\boldsymbol{\theta}_S) + \log p(\boldsymbol{\theta}_S|a_S,\boldsymbol{\delta}_S) + \log p(\boldsymbol{\delta}_S|b_S,\rho_S)\\
&+ p(\rho_S) + p(\lambda_S)\big]\\
&= \sum_{i=1}^{M}\sum_{j=1}^{MP}\big[\log\mathcal{G}(\theta_{Aij}|a_A,\delta_{Aij}) + \log\mathcal{G}(\delta_{Aij}|b_A,\rho_A)\big]\\
&+ \sum_{i=1}^{M}\sum_{j=1}^{i-1}\big[\log\mathcal{G}(\theta_{Uij}|a_U,\delta_{Uij}) + \log\mathcal{G}(\delta_{Uij}|b_U,\rho_U)\big] + \text{ terms not involving } \boldsymbol{\phi}
\end{aligned}$$

Then, at the kth iteration of the EM algorithm, the conditional log-likelihood on $\boldsymbol{\phi}^{(k-1)}$ and $\boldsymbol{y}$ is given by

$$\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(k-1)}) = &\sum_{i=1}^{M}\sum_{j=1}^{MP}\big[-\log(\Gamma(a_A)) + a_A\mathbb{E}_{\phi^{(k-1)}}[\log\theta_{Aij} + \log\delta_{Aij}|\boldsymbol{y}]\\
&- \log(\Gamma(b_A)) + b_A\mathbb{E}_{\phi^{(k-1)}}[\log\delta_{Aij} + \log\rho_A|\boldsymbol{y}])\big]\\
&+ \sum_{i=1}^{M}\sum_{j=1}^{i-1}\big[-\log(\Gamma(a_B)) + a_B\mathbb{E}_{\phi^{(k-1)}}[\log\theta_{Bij} + \log\delta_{Bij}|\boldsymbol{y}]\\
&- \log(\Gamma(b_B)) + b_B\mathbb{E}_{\phi^{(k-1)}}[\log\delta_{Bij} + \log\rho_B|\boldsymbol{y}])\big]
\end{aligned} \tag{16}$$

We maximize $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(k-1)})$ over $\boldsymbol{\phi}$ in the M-step. That is to find $\boldsymbol{\phi} \geq 0$ such that

$$\begin{aligned}
\frac{\partial Q}{\partial a_A} &= -M^2 P\psi(a_A) + \sum_{i=1}^{M}\sum_{j=1}^{MP} \mathbb{E}_{\phi^{(k-1)}}[\log\theta_{Aij} + \log\delta_{Aij}|\boldsymbol{y}] = 0\,, \\
\frac{\partial Q}{\partial b_A} &= -M^2 P\psi(b_A) + \sum_{i=1}^{M}\sum_{j=1}^{MP} \mathbb{E}_{\phi^{(k-1)}}[\log\delta_{Aij} + \log\rho_{A}|\boldsymbol{y}] = 0\,, \\
\frac{\partial Q}{\partial a_U} &= -M(M-1)/2\psi(a_U) + \sum_{i=1}^{M}\sum_{j=1}^{i-1} \mathbb{E}_{\phi^{(k-1)}}[\log\theta_{Uij} + \log\delta_{Uij}|\boldsymbol{y}] = 0\,, \\
\frac{\partial Q}{\partial b_U} &= -M(M-1)/2\psi(b_U) + \sum_{i=1}^{M}\sum_{j=1}^{i-1} \mathbb{E}_{\phi^{(k-1)}}[\log\delta_{Uij} + \log\rho_{U}|\boldsymbol{y}] = 0\,.
\end{aligned} \tag{17}$$

where $\psi = \frac{d(\Gamma(x))}{dx}$ denotes the digamma function. We can solve $\boldsymbol{\phi}$ in (17) using any fast root-finding algorithm. The posterior expectations in (17) are approximated from the mean of N Gibbs samples based on $\boldsymbol{\phi}^{(k-1)}$ where N is the number of Monte Carlo samples. Posterior samples are stimulated according to the MCMC procedures proposed in Section 3.1. This approach is well known as Monte Carlo expectation maximization (MCEM) in [14, 41].

Although MCMC and MCEM provide efficient inference schemes, they do not allow for variable selection. In next section, we discuss the two-stage procedure for variable selection within a fully Bayesian framework.

4 Variable selection

Since the TPBN prior is a global-local shrinkage prior, it cannot assign exact zero mass to sparse coefficients. This implies that this model cannot introduce the true sparsity into parameters directly. However, sparsity is claimed to be important in [32], wherein sparsification reduces the uncertainty of estimates and yields improved forecasting performance in time-varying parameter models. [32] employ the single adaptive variable selection (SAVS) estimate proposed in [47] - this method is treated as a soft-thresholding approach acting on the posterior mean. Alternatively, we conduct a post-analysis method to select variables based on the "decoupled shrinkage and selection" (DSS) method proposed in [28].

Since the coefficients for the auto-regression are of interest, we discuss the variable selection for coefficient matrix $\boldsymbol{A}$, and dismiss the selection for $\boldsymbol{U}$ since those parameters are not of interest. Assume $\hat{\boldsymbol{A}}$ is the posterior mean of $\boldsymbol{A}$. We consider the problem of predicting new observations with new design matrix $\tilde{\boldsymbol{X}}$ such that $\tilde{\boldsymbol{y}}_t \sim \mathcal{N}(\boldsymbol{A}\tilde{\boldsymbol{x}}_t, \boldsymbol{\Omega}_t)$ where $\tilde{\boldsymbol{y}}_t$ and $\tilde{\boldsymbol{x}}_t$ are the tth column of $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}$ respectively. We summarize the original model $\mathbb{E}(\tilde{\boldsymbol{y}}_t|\tilde{\boldsymbol{x}}_t) = \boldsymbol{A}\tilde{\boldsymbol{x}}_t$ by finding a parsimonious coefficient matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}\tilde{\boldsymbol{x}}_t$ closely matches the prediction $\mathbb{E}(\tilde{\boldsymbol{y}}_t|\tilde{\boldsymbol{x}}_t)$. Formally, we define a loss function for the discrepancy in prediction between the full model and parsimonious summary model as

$$\mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\Gamma}) = \lambda\|\boldsymbol{\Gamma}\|_0 + \|\tilde{\boldsymbol{X}}\boldsymbol{\Gamma}' - \tilde{\boldsymbol{X}}\boldsymbol{A}'\|_2^2 \tag{18}$$

where $\|\cdot\|_0 = \sum_{ij} \mathbb{1}(\Gamma_{ij} \neq 0)$. This loss sums two components, one of which is a parsimony on the model parameters Γ and the other of which is the squared prediction loss. By considering $\tilde{\boldsymbol{X}} = \boldsymbol{X}$ as a conventional choice [28] and taking an expectation with respect to posterior distribution of $\boldsymbol{A}$, we write the posterior expected loss as

$$\mathcal{L}(\boldsymbol{\Gamma}) \equiv \mathbb{E}(\mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\Gamma})) = \lambda\|\boldsymbol{\Gamma}\|_0 + T^{-1}\|\boldsymbol{X}\boldsymbol{\Gamma}' - \boldsymbol{X}\hat{\boldsymbol{A}}'\|_2^2 \,. \tag{19}$$

where we drop constant terms with respect Γ.

The equivalent representation under a standard uni-variate regression framework is

$$\mathcal{L}(\boldsymbol{\Gamma}) = \lambda\|\text{vec}[\boldsymbol{\Gamma}]\|_0 + T^{-1}\|(\boldsymbol{I}_M \otimes \boldsymbol{X})\text{vec}[\boldsymbol{\Gamma}'] - (\boldsymbol{I}_M \otimes \boldsymbol{X})\text{vec}[\hat{\boldsymbol{A}}']\|_2^2 \,. \tag{20}$$

Compared with the DSS loss function in (19) of [28], the only difference is the coefficient of the squared prediction loss. However, the solution of Γ does not change as the the solution of weight λ would scale it accordingly. On the other hand, the counting penalty $\lambda\|\Gamma\|_0$ yields an intractable NP-hard combinatorial problem. We apply a similar local linear approximation [28] by solving the following surrogate optimization:

$$\hat{\boldsymbol{\Gamma}} = \arg\min_{\boldsymbol{\Gamma}} T^{-1}\|\boldsymbol{X}\boldsymbol{\Gamma}' - \boldsymbol{X}\hat{\boldsymbol{A}}'\|_2^2 + \lambda \sum_{i=1}^{M}\sum_{j=1}^{MP} \frac{|\boldsymbol{\Gamma}_{ij}|}{|\hat{A}_{ij}|} \,, \tag{21}$$

where $\hat{A}_{ij}$ is the (i,j)th element in the posterior mean $\hat{\boldsymbol{A}}$ and λ is chosen using 10-fold cross-validation. This optimization is essentially an adaptive LASSO regression [58] with weights $\frac{1}{|\hat{\Gamma}_{ij}|}$. An efficient gradient descent algorithm to

find LASSO solutions is proposed in [22] and we use the R package glmnet to solve the adaptive LASSO regression in Equation (21).

One key shortcoming of conducting the DSS method is that uncertainty quantification about the $\hat{\Gamma}$ is not possible as we estimate $\hat{\Gamma}$ by solving the optimization problem in Equation (21). Recently, [56] develop methods to get estimates of posterior uncertainty in both linear and non-parametric regression frameworks. This work propagates posterior uncertainty from the original fitted model though a sparse summary. It projects the full posterior distribution onto the space of the sparse summary model using the restricted set of variables. We adapt it into our self-adaptive VAR model based on the DSS method. Specifically, we calculate the projected posterior for the summary using the posterior samples of $\boldsymbol{A}^{(s)} \sim p(\boldsymbol{A}|\boldsymbol{y})$ and the posterior samples of parsimonious estimates Γ are obtained through $\boldsymbol{\Gamma}^{(s)} = \arg\min_{\boldsymbol{\Gamma}} T^{-1}\|\boldsymbol{X}\boldsymbol{\Gamma}' - \boldsymbol{X}(\boldsymbol{A}^{(s)})'\|_2^2 + \lambda \sum_{i=1}^{M}\sum_{j=1}^{MP} \frac{|\boldsymbol{\Gamma}_{ij}|}{|\hat{A}_{ij}|}$, where s is the index of samples.

5 Experiments

We conduct an extensive simulation study to illustrate the superior performance of our model on parameter estimation and variable selection. We also apply our model to real macroeconomic data and show that our model performs well relative to other commonly used models. We use density predictions as a metric to evaluate the performance of the models on the macroeconomic data.

5.1 Simulation study

We compare our method with a range of commonly used alternatives in this section. We investigate sparse, intermediate and dense data generating processes (DGPs) where the length of time series $T = 50, 150$ or 250 and the dimension of data $m = 20, 50$. The probability of a diagonal entry to be non-zero is 0.8 and the probability of an off-diagonal entry to be non-zero is 0.01, 0.2 and 0.5 in each setting. The non-zero elements are randomly generated from log Gaussian distributions to guarantee the discrepancy between the nonzero value and zero and add a random sign + or − with the equal probability. Moreover, we specify the mean μ and standard deviation σ of the diagonal (D) and the off-diagonal (O) entries are chosen as follows:

- Sparse: $\mu_D = \sigma_D = \mu_O = \sigma_O = 0.3$
- Intermediate: $\mu_D = \sigma_D = 0.15$ and $\mu_O = \sigma_O = 0.1$
- Dense: $\mu_D = \sigma_D = 0.15$ and $\mu_O = \sigma_O = 0.01$.

As for a random variable X following a log Gaussian distribution with parameters μ and σ, given the desired mean μ_X and variance σ_X^2, we can sample it using parameters derived by $\mu = \ln\left(\frac{\mu_X^2}{\sqrt{\mu_X^2+\sigma_X^2}}\right)$ and $\sigma^2 = \ln\left(1+\frac{\sigma_X^2}{\mu_X^2}\right)$. To ensure the stationarity of this dynamic system, we employ rejection sampling to get stationary coefficient matrices A_is, in which all largest eigenvalues of coefficient matrices are within a unit disk.

Concerning the errors, we use a single factor stochastic volatility (SV) specification with factor loadings generated from $\mathcal{N}(0.001, 0.001^2)$ to roughly match the above scaling. The AR(1) process in SV model is assumed to have mean $\mu_{\sigma i} = -12$ with persistences $\rho_i = 0.99$ and innovation standard deviations $\psi_{\sigma i} = 0.1$.

For each of the 252 settings, we simulate 50 independent datasets, and for each of them we run MCMC or MCEM algorithm to obtain 2000 posterior draws after a burn-in of 1000. Our proposed model with both MCMC inference and MCEM inference are compared with the state-of-the-art methods. We considered Bayesian vector autoregression with stochastic volatility model (BVAR-SV) with Normal Gamma (NG) priors with three different shrinkage structures (global, rowwise and columnwise) [31], horseshoe (HS) priors with the same three shrinkage structures [21], stochastic search variable selection (SSVS) priors [24] and Minnesota priors [25]. All MCMC inferences consider total 3000 iterations, the first 1,000 of which is discarded as burn-in. As for the MCEM inference, we also consider total 3000 iterations but we estimate the hyper-parameter in the first 1000 iterations and fix them in the next 2000 iterations. Due to space constraints, we report the results for the $T = 50$, 250 cases, and for models with the global shrinkage structure.

We compare the posterior median to the true values of coefficient matrix $\boldsymbol{A}$ and compute the root mean square errors. For each setting, we report the mean and standard deviation of the corresponding RMSEs in Table 1. The table shows that TPBN models have significantly smaller RMSEs than other models for all scenarios of DGP. Among the two TPBN models, the fitting performance on MCMC and MCEM are very similar. On the other hand, we find that the posterior distribution of shape parameters a and b

in (11) is as flat as our proposed prior, i.e. Exp(1), which implies that it is difficult to learn the sparsity level through MCMC. Additonally, there is a significant discrepancy between the posterior median of hyper-parameters a and b in MCMC and the point estimates in MCEM. So the point estimates from MCEM would be more reasonable. The full results are reported in the Appendix.

Moreover, for all models, through solving the optimization problem proposed in (21), we conduct variable selection and obtain parsimonious coefficient matrix $\boldsymbol{A}$. We then compare the average hit rate that measures the percentage of correctly estimated zeros [32] in Table 2. The higher the probability is, The better fitting performance the model achieves. Moreover, we computed the root mean square error of the nonzero entries on the parsimonious coefficient matrix $\boldsymbol{A}$ in Table 3. Specifically, assuming $\boldsymbol{A}$ is the ground truth and $\hat{\boldsymbol{A}}$ is the estimate, the root mean square error is defined as $\text{RMSE(N)} = \sqrt{\sum_{(i,j)\in\mathcal{I}}(A_{ij} - \hat{A}_{ij})^2}$ where $\mathcal{I}$ is a set of indexes such as $(i,j) \in \mathcal{I}$ if and only if either $A_{ij} \neq 0$ or $\hat{A}_{ij} \neq 0$. Table 2 shows that our TPBN model has higher average hit rate than other models, suggesting that it achieves better variable selection and it is more likely to find the true sparse structure in coefficients. Moreover, within TPBN model, as the length of time series increases, MCEM archives better variable selection than MCMC.

In addition, Table 3 implies that for those non-zero coefficients, TPBN has less biased estimation than other models. We find that NG method results in significantly large bias when $T = 50$. This is because the NG method is very sensitive to the adaptive lasso optimization in the post analysis when T is small, where the optimization generates large biased parsimonious estimation. Moreover, comparing amongst MCMC and MCEM within our TPBN model, we find the same behavior as in variable selection - As T increases, MCEM outperforms MCMC in terms of the fitting performance on nonzero entries.

5.2 Real data analysis

We use the the macroeconomic data constructed in [42]. The data are sampled at a quarterly frequency spanning from 1959Q2 to 2015Q2. All data are transformed to be approximately stationary using the same procedures in [31]. Forecasting is performed on three subsets of the dataset - the small

Table 1: Root mean square errors on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios.

	S20	S50	I20	I50	D20	D50
T = 50						
NG	0.229	0.067	0.235	0.057	0.169	0.065
	(0.542)	(0.032)	(0.762)	(0.009)	(0.486)	(0.076)
HS	0.223	0.044	0.174	0.062	0.324	0.043
	(0.514)	(0.009)	(0.600)	(0.008)	(0.926)	(0.066)
SSVS	0.148	0.210	0.151	0.125	0.185	0.234
	(0.059)	(0.193)	(0.060)	(0.135)	(0.190)	(0.285)
Minnesota	0.069	0.083	0.134	0.114	0.055	0.074
	(0.042)	(0.043)	(0.054)	(0.022)	(0.038)	(0.043)
TPBN_MCMC	**0.032**	**0.029**	0.047	**0.048**	**0.025**	**0.019**
	(0.010)	(0.005)	(0.007)	(0.005)	(0.011)	(0.004)
TPBN_MCEM	**0.032**	**0.029**	**0.046**	**0.048**	**0.025**	**0.019**
	(0.010)	(0.005)	(0.008)	(0.005)	(0.011)	(0.003)
T = 250						
NG	0.198	0.206	0.251	0.093	0.260	0.120
	(0.321)	(0.420)	(0.312)	(0.260)	(0.387)	(0.271)
HS	0.225	0.240	0.248	0.233	0.323	0.179
	(0.290)	(0.456)	(0.263)	(0.351)	(0.398)	(0.335)
SSVS	0.131	0.102	0.139	0.089	0.117	0.126
	(0.125)	(0.109)	(0.121)	(0.037)	(0.104)	(0.122)
Minnesota	0.109	0.071	0.145	0.119	0.075	0.057
	(0.102)	(0.029)	(0.093)	(0.069)	(0.087)	(0.040)
TPBN_MCMC	**0.019**	**0.015**	0.038	**0.032**	**0.019**	**0.015**
	(0.010)	(0.004)	(0.019)	(0.006)	(0.006)	(0.002)
TPBN_MCEM	**0.019**	0.016	**0.037**	**0.032**	**0.019**	**0.015**
	(0.009)	(0.005)	(0.016)	(0.005)	(0.006)	(0.002)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior.

Table 2: Average hit rate on the parsimonious coefficient matrix $\boldsymbol{A}$ for different simulation scenarios..

	S20	S50	I20	I50	D20	D50
T = 50						
NG	0.873	0.987	0.884	0.988	0.884	0.989
	(0.077)	(0.008)	(0.075)	(0.014)	(0.072)	(0.010)
HS	0.827	0.978	0.816	0.972	0.816	0.971
	(0.043)	(0.010)	(0.048)	(0.013)	(0.048)	(0.011)
SSVS	0.803	0.982	0.807	0.975	0.802	0.983
	(0.038)	(0.009)	(0.044)	(0.012)	(0.048)	(0.009)
Minnesota	0.901	0.988	0.835	0.982	0.904	0.989
	(0.044)	(0.005)	(0.043)	(0.007)	(0.032)	(0.005)
TPBN_MCMC	**0.931**	**0.998**	0.905	0.991	**0.918**	**0.995**
	(0.053)	(0.004)	(0.063)	(0.009)	(0.050)	(0.005)
TPBN_MCEM	**0.931**	**0.998**	**0.911**	**0.993**	**0.918**	**0.995**
	(0.049)	(0.003)	(0.058)	(0.009)	(0.044)	(0.005)
T = 250						
NG	0.826	0.914	0.790	0.920	0.790	0.907
	(0.092)	(0.046)	(0.108)	(0.036)	(0.102)	(0.030)
HS	0.803	0.912	0.785	0.893	0.822	0.917
	(0.056)	(0.030)	(0.063)	(0.04)	(0.063)	(0.030)
SSVS	0.848	0.911	0.831	0.898	0.848	0.914
	(0.044)	(0.024)	(0.051)	(0.037)	(0.046)	(0.025)
Minnesota	0.857	0.939	0.789	0.886	0.888	0.956
	(0.055)	(0.017)	(0.038)	(0.022)	(0.064)	(0.015)
TPBN_MCMC	0.924	0.968	0.861	0.938	0.907	**0.961**
	(0.063)	(0.029)	(0.058)	(0.025)	(0.059)	(0.019)
TPBN_MCEM	**0.932**	**0.969**	**0.888**	**0.943**	**0.912**	0.957
	(0.048)	(0.023)	(0.047)	(0.021)	(0.043)	(0.018)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior.

Table 3: Root mean square error on non-zero entries of the parsimonious coefficient matrix $\boldsymbol{A}$ for different simulation scenarios.

	S20	S50	I20	I50	D20	D50
T = 50						
NG	1.240	158.782	68.381	21.421	5.594	134.862
	(2.515)	(380.134)	(450.560)	(67.354)	(35.332)	(294.822)
HS	0.915	0.447	1.649	0.252	1.273	0.172
	(1.957)	(0.210)	(6.929)	(0.183)	(3.455)	(0.183)
SSVS	0.433	2.100	0.335	0.436	0.335	0.630
	(0.230)	(2.358)	(0.193)	(0.425)	(0.345)	(0.707)
Minnesota	0.257	0.856	0.288	0.384	0.116	0.249
	(0.157)	(0.556)	(0.139)	(0.104)	(0.098)	(0.161)
TPBN_MCMC	0.133	**0.210**	0.121	**0.133**	0.047	0.032
	(0.054)	(0.035)	(0.042)	(0.029)	(0.032)	(0.014)
TPBN_MCEM	**0.132**	0.211	**0.115**	**0.133**	**0.046**	**0.031**
	(0.050)	(0.037)	(0.027)	(0.028)	(0.031)	(0.012)
T = 250						
NG	0.948	2.044	0.830	0.503	0.802	0.532
	(1.583)	(4.231)	(1.077)	(1.575)	(1.265)	(1.367)
HS	0.847	2.114	0.713	0.867	0.780	0.622
	(1.349)	(5.000)	(0.879)	(1.583)	(1.082)	(1.332)
SSVS	0.456	0.825	0.341	0.255	0.248	0.354
	(0.400)	(1.362)	(0.275)	(0.111)	(0.206)	(0.361)
Minnesota	0.320	0.418	0.322	0.336	0.144	0.134
	(0.253)	(0.205)	(0.212)	(0.254)	(0.167)	(0.089)
TPBN_MCMC	0.100	0.159	0.119	0.115	0.044	**0.028**
	(0.066)	(0.186)	(0.073)	(0.074)	(0.039)	(0.011)
TPBN_MCEM	**0.097**	**0.145**	**0.101**	**0.106**	**0.040**	0.029
	(0.068)	(0.095)	(0.062)	(0.038)	(0.028)	(0.010)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior.

version evaluates forecasts over the three variables of primary interest: GDP growth, Federal Funds Rate and GDP deflator. The medium and large versions take into consideration a total of 7 and 21 features respectively, following the setup in [31].

We include the same set of models in Section 5.1, namely a normal gamma prior VAR, horseshoe prior VAR, a hierarchical Minnesota prior VAR and the SSVS prior VAR. Following [10], we specify the prior mean with $\boldsymbol{A}_1$ to be a diagonal matrix with entries 0.9, encouraging strong persistence and specify all other prior mean as zeros. And all models consider $P = 4$ lags in this section. For all MCMC inference approaches, we take 3000 iterations wherein 2000 iterations are treated as burnin. As for MCEM inference approach, we also take 3000 iterations in which the first 1000 iterations are used to estimate hyper-parameters and the latest 1000 iterations are used for model forecasting to make a fair comparison with MCMC approaches.

Forecasts are evaluated over a long hold-out sample spanning from 2002Q1 to 2015Q2. We conduct the n-step-ahead forecast for $n = 1, 2, 3, 4$. We recursively forecast and expand the initial estimation window for 50 runs. For each run, we use the log predictive density to evaluate the quality of predictive density via the predictive posterior median, and then we report the trimmed mean over 50 runs by 20%. In addition, we only report the results with global shrinkage structure for NS, HS, TPBN in Table 4. Results for the full set of experiments, and corresponding boxplots for the 50 runs are reported in the Appendix. Table 4 shows that when $M = 3$, SSVS and Minnesota perform better in forecast; one possible explanation for this finding is that, as pointed out in [18], the macroeconomic data is not sparse. In this case, our TPBN model with MCEM inference outperforms other global-local prior based methods in most of versions of datasets. In the medium version, we find that our TPBN model with MECM inference approach dominates other models except for the two-step-ahead forest. Finally, the result in the large version shows that our model with MCMC inference outperforms others. Moreover, we find that among the two inferences of our model, in the small and medium version, MCEM approach consistently outperforms MCMC inference in terms of model forecast while in the large version MCMC performs better than MCEM. It may be because that MCEM effectively learn the hyper-parameters in the small and medium version, which contributes to better forecast on the density distribution at the out-of-sample time steps. In the large version, since the hyper-parameters are learned from more latent coefficients, the better hyper-parameter learning achieved by MCMC

contributes to better posterior sampling and subsequently better forecasting performance.

Table 4: Average log predictive density to a large VAR-SV : 2002Q1 - 2015Q2

	NG	HS	SSVS	Minnesota	TPBN_MCMC	TPBN_MCEM
			Small (M = 3)			
One-step-ahead	12.721	12.927	**13.047**	12.962	12.859	12.970
Two-step-ahead	11.800	12.205	**12.293**	12.241	12.133	12.171
Three-step-ahead	11.198	11.582	**11.658**	11.630	11.515	11.548
Four-step-ahead	10.842	10.201	11.252	**11.272**	11.136	11.175
			Medium (M = 7)			
One-step-ahead	27.028	26.850	27.067	26.910	26.944	**27.127**
Two-step-ahead	26.138	**26.499**	26.492	26.528	26.324	26.492
Three-step-ahead	25.489	25.868	25.752	25.697	25.684	**25.888**
Four-step-ahead	25.118	25.347	25.153	25.264	25.177	**25.361**
			Large (M = 21)			
One-step-ahead	67.288	64.764	64.780	57.180	**68.634**	67.116
Two-step-ahead	66.374	68.843	68.757	67.246	**69.299**	68.837
Three-step-ahead	65.179	68.636	67.888	66.825	**69.371**	67.449
Four-step-ahead	64.661	68.141	67.717	67.717	**68.447**	66.576

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior.

6 Conclusion

We have introduced the TPBN prior to Bayesian vector autoregressive (BVAR) models. To perform effective inference on the flexible model, we propose two efficient inference schemes, Markov Chain Monte Carlo (MCMC) inference and an empirical Bayes approach via Monte Carlo Expectation Maximization (MCEM) to estimate the hyper-parameters based on the marginal maximum likelihood (MML). We also provide the variable selection procedures in the context of BVARs. We find that TPBN shows superior performance on parameter estimation and variable selection in comparison to other existing state-of-the-art models in the synthetic study wherein we evaluate the performance over 50 simulations each of 18 synthetic datasets, with varying levels of size and sparsity.

As for the forecasting performance, our model performs robustly well in different settings. In the macroeconomic data, although our model performs worse than SSVS and Minnesota approaches in the small version of data, our model performs equally well or better than other global-local prior based models. On the other hand, our model achieves notably better forecasting performance in medium and large versions in comparison to other models. In addition, although the MCMC and MCEM inference in our model have similar performance on parameter estimation and variable selection, MCEM consistently outperforms MCMC in terms of forecasting performance on the small and medium subsets of the macroeconomic data but MCMC outperforms MCEM on the large subset. The reason may be that in the small and medium version, MCEM can learn the hyper-parameters effectively whereas MCMC does not.

References

[1] Artin Armagan, David B Dunson, and Merlise Clyde. Generalized beta mixtures of gaussians. *Advances in neural information processing systems*, 24:523, 2011.

[2] Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013.

[3] C Armero and MJ Bayarri. Prior assessments for prediction in queues. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(1):139–153, 1994.

[4] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, Brandon Willard, et al. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017.

[5] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, Brandon Willard, et al. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

[6] Anirban Bhattacharya, Antik Chakraborty, and Bani K Mallick. Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042, 2016.

[7] Anirban Bhattacharya and David B Dunson. Sparse bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

[8] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[9] Andrea Carriero, Todd E Clark, and Massimiliano Marcellino. Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137–154, 2019.

[10] Andrea Carriero, George Kapetanios, and Massimiliano Marcellino. Forecasting government bond yields with large bayesian vector autoregressions. *Journal of Banking & Finance*, 36(7):2026–2047, 2012.

[11] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009.

[12] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[13] George Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[14] George Casella. Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[15] Joshua CC Chan. Minnesota-type adaptive hierarchical priors for large bayesian vars. *International Journal of Forecasting*, 2021.

[16] Joshua CC Chan and Eric Eisenstat. Comparing hybrid time-varying parameter vars. *Economics Letters*, 171:1–5, 2018.

[17] Timothy Cogley and Thomas J Sargent. Drifts and volatilities: monetary policies and outcomes in the post wwii us. *Review of Economic dynamics*, 8(2):262–302, 2005.

[18] Jamie L Cross, Chenghan Hou, and Aubrey Poon. Macroeconomic forecasting with large bayesian vars: Global-local priors and the illusion of sparsity. *International Journal of Forecasting*, 36(3):899–915, 2020.

[19] Christine De Mol, Domenico Giannone, and Lucrezia Reichlin. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328, 2008.

[20] Thomas Doan, Robert Litterman, and Christopher Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100, 1984.

[21] Lendie Follett and Cindy Yu. Achieving parsimony in bayesian vector autoregressions with the horseshoe prior. *Econometrics and Statistics*, 11:130–144, 2019.

[22] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[23] Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[24] Edward I George, Dongchu Sun, and Shawn Ni. Bayesian stochastic search for var model restrictions. *Journal of Econometrics*, 142(1):553–580, 2008.

[25] Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451, 2015.

[26] Michael B Gordy et al. *A generalization of generalized beta distributions*, volume 18. Division of Research and Statistics, Division of Monetary Affairs, Federal . . . , 1998.

[27] Jim E Griffin, Philip J Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis*, 5(1):171–188, 2010.

[28] P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

[29] Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

[30] Niko Hauzenberger, Florian Huber, and Luca Onorante. Combining shrinkage and sparsity in conjugate vector autoregressive models. *Journal of Applied Econometrics*, 2020.

[31] Florian Huber and Martin Feldkircher. Adaptive shrinkage in bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, 37(1):27–39, 2019.

[32] Florian Huber, Gary Koop, and Luca Onorante. Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, pages 1–15, 2020.

[33] Iain M Johnstone, Bernard W Silverman, et al. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649, 2004.

[34] Gregor Kastner. Sparse bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics*, 210(1):98–115, 2019.

[35] Gregor Kastner and Sylvia Frühwirth-Schnatter. Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423, 2014.

[36] Gregor Kastner, Sylvia Frühwirth-Schnatter, and Hedibert Freitas Lopes. Efficient bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26(4):905–917, 2017.

[37] Gregor Kastner and Florian Huber. Sparse bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 39(7):1142–1165, 2020.

[38] Gary Koop and Dimitris Korobilis. *Bayesian multivariate time series methods for empirical macroeconomics*. Now Publishers Inc, 2010.

[39] Gary Koop, Dimitris Korobilis, and Davide Pettenuzzo. Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154, 2019.

[40] Dimitris Korobilis. Var forecasting using bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

[41] Richard A Levine and George Casella. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

[42] Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[43] Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Artificial Intelligence and Statistics*, pages 905–913. PMLR, 2017.

[44] Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010.

[45] Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

[46] Giorgio E Primiceri. Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852, 2005.

[47] Pallavi Ray and Anirban Bhattacharya. Signal adaptive variable selector for the horseshoe prior. *arXiv preprint arXiv:1810.09004*, 2018.

[48] Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[49] Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[50] Håvard Rue. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338, 2001.

[51] Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

[52] Christopher A Sims and Pierre Perron. *4. A Nine-Variable Probabilistic Macroeconomic Forecasting Model*. University of Chicago Press, 2008.

[53] Christopher A Sims and Tao Zha. Bayesian methods for dynamic multivariate models. *International Economic Review*, pages 949–968, 1998.

[54] Michael Smith and Robert Kohn. Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153, 2002.

[55] Daniel F Waggoner and Tao Zha. A gibbs sampler for structural vector autoregressions. *Journal of Economic Dynamics and Control*, 28(2):349–366, 2003.

[56] Spencer Woody, Carlos M Carvalho, and Jared S Murray. Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics*, pages 1–9, 2020.

[57] Shiwen Zhao, Chuan Gao, Sayan Mukherjee, and Barbara E Engelhardt. Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research*, 17(1):6868–6914, 2016.

[58] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A Sampling for the hyper-parameters in TPBN

According to the hierarchical representation in VAR in (11), the sampling procedures for the hyper-parameters a, b and ρ are given as follows:

Because $\rho^{\frac{1}{2}} \sim \mathcal{C}^{+}(0,1)$ has the hierarchical representation $\rho \sim \mathcal{G}(\frac{1}{2}, c)$ and $c \sim \mathcal{G}(\frac{1}{2}, 1)$. The posterior of ρ can be sequentially sampled from the conditional posterior distributions

$$\rho|\boldsymbol{\delta}, c \sim \mathcal{G}(\frac{1}{2} + TMP, c + \sum_{i=1}^{T}\sum_{j=1}^{MP}\delta_{ij})$$
$$c|\rho \sim \mathcal{G}(\frac{1}{2} + 1, 1 + \rho)$$

As for the MCMC inference on shape parameters a and b in (11) Let $\tilde{a} = \ln(a), \tilde{b} = \ln(b)$. The conditional posterior distributions of them are

$$\tilde{a}|\boldsymbol{\theta}, \boldsymbol{\delta} \propto \prod_{i=1}^{T}\prod_{j=1}^{MP}\mathcal{G}(\theta_{ij}|a, \delta_{ij})\mathrm{Exp}(a|1)|\frac{\partial \exp(\tilde{a})}{\partial \tilde{a}}|\,, \tag{22}$$

$$\tilde{b}|\boldsymbol{\delta}, \rho \propto \prod_{i=1}^{T}\prod_{j=1}^{MP}\mathcal{G}(\delta_{ij}|b, \rho)\mathrm{Exp}(b|1)|\frac{\partial \exp(\tilde{b})}{\partial \tilde{b}}|\,. \tag{23}$$

We recursively sample $\tilde{a}$ and $\tilde{b}$ via adaptive Metropolis Hasting [48]. Specifically, for each of them, we assume the proposal distribution given at iteration n is given by $Q_n(x, \cdot) = \mathcal{N}(x, \sigma_n^2)$. And we set $\sigma_n^2 = 1$ when $n \leq 50$, and after 50 iteration we update it by

$$\sigma_n^2 = \begin{cases} \sigma_{n-1}^2 \exp(\delta(n)) & \text{acceptances of the variable is more than } 0.44\,, \\ \sigma_{n-1}^2 \exp(-\delta(n)) & \text{acceptances of the variable is less than } 0.44\,. \end{cases} \tag{24}$$

to make the acceptance rate of proposals for the variable as close as possible to 0.44 [23, 49]. And $\delta(n) = \min(0.01, n^{-1/2})$.

B Synthetic result

B.1 Root mean square error on A in terms of posterior median

Table 5: Root mean square errors on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 50$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.229 (0.542)	0.067 (0.032)	0.235 (0.762)	0.057 (0.009)	0.169 (0.486)	0.065 (0.076)
NG(C)	0.057 (0.042)	0.055 (0.020)	0.071 (0.069)	0.066 (0.030)	0.045 (0.012)	0.066 (0.148)
NG(R)	0.048 (0.012)	0.046 (0.005)	0.056 (0.011)	0.057 (0.005)	0.043 (0.010)	0.046 (0.025)
HS(G)	0.223 (0.514)	0.044 (0.009)	0.174 (0.600)	0.062 (0.008)	0.324 (0.926)	0.043 (0.066)
HS(C)	0.084 (0.043)	0.066 (0.030)	0.093 (0.032)	0.071 (0.016)	0.089 (0.141)	0.084 (0.170)
HS(R)	0.082 (0.035)	0.057 (0.013)	0.091 (0.027)	0.069 (0.012)	0.064 (0.023)	0.053 (0.033)
SSVS	0.148 (0.059)	0.210 (0.193)	0.151 (0.060)	0.125 (0.135)	0.185 (0.190)	0.234 (0.285)
Minnesota	0.069 (0.042)	0.083 (0.043)	0.134 (0.054)	0.114 (0.022)	0.055 (0.038)	0.074 (0.043)
TPBN_MCMC(G)	0.032 (0.010)	0.029 (0.005)	0.047 (0.007)	0.048 (0.005)	0.025 (0.011)	0.019 (0.004)
TPBN_MCMC(C)	0.034 (0.008)	0.030 (0.005)	0.050 (0.009)	0.052 (0.006)	0.026 (0.009)	0.020 (0.003)
TPBN_MCMC(R)	0.047 (0.011)	0.041 (0.006)	0.052 (0.009)	0.051 (0.004)	0.039 (0.010)	0.030 (0.004)
TPBN_MCEM(G)	0.032 (0.010)	0.029 (0.005)	0.046 (0.008)	0.048 (0.005)	0.025 (0.011)	0.019 (0.003)
TPBN_MCEM(C)	0.034 (0.007)	0.030 (0.005)	0.049 (0.009)	0.051 (0.007)	0.026 (0.009)	0.019 (0.003)
TPBN_MCEM(R)	0.039 (0.010)	0.040 (0.005)	0.049 (0.008)	0.051 (0.004)	0.033 (0.010)	0.030 (0.004)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 6: Root mean square errors on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 150$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.324 (0.592)	0.114 (0.410)	0.245 (0.381)	0.050 (0.032)	0.133 (0.366)	0.068 (0.153)
NG(C)	0.057 (0.049)	0.058 (0.072)	0.055 (0.025)	0.054 (0.026)	0.059 (0.080)	0.067 (0.120)
NG(R)	0.038 (0.012)	0.031 (0.006)	0.049 (0.013)	0.043 (0.009)	0.038 (0.016)	0.031 (0.007)
HS(G)	0.332 (0.521)	0.124 (0.338)	0.353 (0.367)	0.219 (0.462)	0.305 (0.448)	0.207 (0.447)
HS(C)	0.134 (0.112)	0.114 (0.202)	0.138 (0.090)	0.079 (0.026)	0.111 (0.083)	0.130 (0.197)
HS(R)	0.131 (0.103)	0.067 (0.042)	0.129 (0.053)	0.074 (0.017)	0.104 (0.060)	0.074 (0.056)
SSVS	0.130 (0.123)	0.125 (0.134)	0.143 (0.069)	0.090 (0.018)	0.104 (0.069)	0.179 (0.234)
Minnesota	0.120 (0.108)	0.089 (0.072)	0.183 (0.140)	0.108 (0.027)	0.060 (0.059)	0.071 (0.046)
TPBN_MCMC(G)	0.022 (0.010)	0.019 (0.004)	0.040 (0.012)	0.037 (0.005)	0.021 (0.007)	0.017 (0.004)
TPBN_MCMC(C)	0.027 (0.019)	0.020 (0.006)	0.042 (0.010)	0.041 (0.009)	0.023 (0.006)	0.018 (0.003)
TPBN_MCMC(R)	0.039 (0.012)	0.032 (0.006)	0.049 (0.013)	0.042 (0.006)	0.035 (0.016)	0.028 (0.004)
TPBN_MCEM(G)	0.022 (0.009)	0.019 (0.004)	0.039 (0.010)	0.038 (0.007)	0.021 (0.006)	0.017 (0.004)
TPBN_MCEM(C)	0.025 (0.016)	0.020 (0.008)	0.040 (0.011)	0.041 (0.010)	0.022 (0.007)	0.017 (0.002)
TPBN_MCEM(R)	0.029 (0.011)	0.028 (0.005)	0.042 (0.012)	0.041 (0.008)	0.029 (0.009)	0.027 (0.004)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 7: Root mean square errors on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 250$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.198 (0.321)	0.206 (0.420)	0.251 (0.312)	0.093 (0.260)	0.260 (0.387)	0.120 (0.271)
NG(C)	0.042 (0.020)	0.046 (0.047)	0.065 (0.049)	0.049 (0.027)	0.043 (0.033)	0.064 (0.083)
NG(R)	0.034 (0.012)	0.027 (0.008)	0.048 (0.020)	0.038 (0.006)	0.034 (0.015)	0.027 (0.004)
HS(G)	0.225 (0.290)	0.240 (0.456)	0.248 (0.263)	0.233 (0.351)	0.323 (0.398)	0.179 (0.335)
HS(C)	0.128 (0.085)	0.070 (0.067)	0.132 (0.086)	0.091 (0.059)	0.120 (0.103)	0.089 (0.090)
HS(R)	0.115 (0.075)	0.057 (0.018)	0.122 (0.063)	0.077 (0.025)	0.107 (0.073)	0.060 (0.029)
SSVS	0.131 (0.125)	0.102 (0.109)	0.139 (0.121)	0.089 (0.037)	0.117 (0.104)	0.126 (0.122)
Minnesota	0.109 (0.102)	0.071 (0.029)	0.145 (0.093)	0.119 (0.069)	0.075 (0.087)	0.057 (0.040)
TPBN_MCMC(G)	0.019 (0.010)	0.015 (0.004)	0.038 (0.019)	0.032 (0.006)	0.019 (0.006)	0.015 (0.002)
TPBN_MCMC(C)	0.019 (0.008)	0.015 (0.003)	0.039 (0.014)	0.035 (0.007)	0.021 (0.007)	0.015 (0.001)
TPBN_MCMC(R)	0.037 (0.014)	0.031 (0.011)	0.045 (0.019)	0.038 (0.007)	0.031 (0.008)	0.025 (0.003)
TPBN_MCEM(G)	0.019 (0.009)	0.016 (0.005)	0.037 (0.016)	0.032 (0.005)	0.019 (0.006)	0.015 (0.002)
TPBN_MCEM(C)	0.019 (0.009)	0.016 (0.005)	0.038 (0.015)	0.034 (0.008)	0.020 (0.007)	0.015 (0.001)
TPBN_MCEM(R)	0.025 (0.011)	0.025 (0.006)	0.040 (0.018)	0.035 (0.005)	0.024 (0.007)	0.024 (0.003)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

B.2 Average hit rate on A in terms of DSS estimates

Table 8: Average hit rate on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 50$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.873 (0.077)	0.987 (0.008)	0.884 (0.075)	0.988 (0.014)	0.884 (0.072)	0.989 (0.010)
NG(C)	0.864 (0.059)	0.990 (0.010)	0.886 (0.078)	0.990 (0.012)	0.856 (0.059)	0.985 (0.014)
NG(R)	0.867 (0.052)	0.986 (0.010)	0.859 (0.069)	0.982 (0.016)	0.854 (0.061)	0.979 (0.013)
HS(G)	0.827 (0.043)	0.978 (0.010)	0.816 (0.048)	0.972 (0.013)	0.816 (0.048)	0.971 (0.011)
HS(C)	0.807 (0.046)	0.974 (0.011)	0.815 (0.039)	0.969 (0.012)	0.820 (0.048)	0.974 (0.012)
HS(R)	0.798 (0.048)	0.964 (0.010)	0.805 (0.041)	0.962 (0.015)	0.810 (0.045)	0.969 (0.012)
SSVS	0.803 (0.038)	0.982 (0.009)	0.807 (0.044)	0.975 (0.012)	0.802 (0.048)	0.983 (0.009)
Minnesota	0.901 (0.044)	0.988 (0.005)	0.835 (0.043)	0.982 (0.007)	0.904 (0.032)	0.989 (0.005)
TPBN_MCMC(G)	0.931 (0.053)	0.998 (0.004)	0.905 (0.063)	0.991 (0.009)	0.918 (0.050)	0.995 (0.005)
TPBN_MCMC(C)	0.944 (0.044)	0.999 (0.002)	0.915 (0.062)	0.995 (0.004)	0.924 (0.044)	0.996 (0.004)
TPBN_MCMC(R)	0.857 (0.055)	0.988 (0.008)	0.864 (0.076)	0.986 (0.013)	0.859 (0.069)	0.986 (0.011)
TPBN_MCEM(G)	0.931 (0.049)	0.998 (0.003)	0.911 (0.058)	0.993 (0.009)	0.918 (0.044)	0.995 (0.005)
TPBN_MCEM(C)	0.936 (0.041)	0.998 (0.003)	0.911 (0.056)	0.995 (0.005)	0.927 (0.045)	0.996 (0.003)
TPBN_MCEM(R)	0.878 (0.049)	0.990 (0.008)	0.877 (0.065)	0.986 (0.013)	0.869 (0.058)	0.984 (0.011)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 9: Average hit rate on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 150$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.821 (0.108)	0.934 (0.034)	0.808 (0.102)	0.922 (0.041)	0.818 (0.073)	0.910 (0.031)
NG(C)	0.837 (0.058)	0.936 (0.032)	0.832 (0.066)	0.911 (0.038)	0.827 (0.068)	0.911 (0.036)
NG(R)	0.841 (0.070)	0.928 (0.035)	0.829 (0.076)	0.900 (0.040)	0.799 (0.071)	0.891 (0.032)
HS(G)	0.821 (0.065)	0.922 (0.026)	0.793 (0.060)	0.902 (0.033)	0.806 (0.055)	0.923 (0.023)
HS(C)	0.827 (0.036)	0.905 (0.030)	0.800 (0.042)	0.892 (0.024)	0.817 (0.050)	0.914 (0.031)
HS(R)	0.818 (0.036)	0.884 (0.025)	0.797 (0.039)	0.879 (0.026)	0.803 (0.054)	0.890 (0.032)
SSVS	0.851 (0.049)	0.904 (0.029)	0.832 (0.047)	0.890 (0.036)	0.828 (0.054)	0.906 (0.032)
Minnesota	0.883 (0.050)	0.943 (0.020)	0.816 (0.039)	0.897 (0.023)	0.898 (0.053)	0.956 (0.014)
TPBN_MCMC(G)	0.932 (0.058)	0.973 (0.020)	0.868 (0.062)	0.931 (0.030)	0.901 (0.060)	0.964 (0.024)
TPBN_MCMC(C)	0.944 (0.042)	0.977 (0.017)	0.903 (0.054)	0.946 (0.023)	0.914 (0.045)	0.969 (0.023)
TPBN_MCMC(R)	0.827 (0.067)	0.912 (0.040)	0.807 (0.068)	0.884 (0.033)	0.797 (0.084)	0.878 (0.031)
TPBN_MCEM(G)	0.931 (0.053)	0.968 (0.019)	0.885 (0.056)	0.934 (0.028)	0.910 (0.050)	0.958 (0.022)
TPBN_MCEM(C)	0.934 (0.044)	0.972 (0.016)	0.900 (0.049)	0.943 (0.022)	0.914 (0.041)	0.960 (0.021)
TPBN_MCEM(R)	0.874 (0.062)	0.928 (0.031)	0.855 (0.057)	0.911 (0.033)	0.835 (0.063)	0.896 (0.026)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 10: Average hit rate on coefficient matrix $\boldsymbol{A}$ for different simulation scenarios with $T = 250$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.826 (0.092)	0.914 (0.046)	0.790 (0.108)	0.920 (0.036)	0.790 (0.102)	0.907 (0.030)
NG(C)	0.841 (0.062)	0.928 (0.033)	0.826 (0.063)	0.916 (0.034)	0.822 (0.064)	0.913 (0.042)
NG(R)	0.841 (0.064)	0.923 (0.034)	0.818 (0.061)	0.905 (0.030)	0.808 (0.069)	0.893 (0.035)
HS(G)	0.803 (0.056)	0.912 (0.030)	0.785 (0.063)	0.893 (0.040)	0.822 (0.063)	0.917 (0.030)
HS(C)	0.804 (0.037)	0.900 (0.026)	0.801 (0.037)	0.890 (0.023)	0.813 (0.053)	0.903 (0.033)
HS(R)	0.795 (0.035)	0.883 (0.027)	0.793 (0.039)	0.878 (0.024)	0.805 (0.049)	0.883 (0.033)
SSVS	0.848 (0.044)	0.911 (0.024)	0.831 (0.051)	0.898 (0.037)	0.848 (0.046)	0.914 (0.025)
Minnesota	0.857 (0.055)	0.939 (0.017)	0.789 (0.038)	0.886 (0.022)	0.888 (0.064)	0.956 (0.015)
TPBN_MCMC(G)	0.924 (0.063)	0.968 (0.029)	0.861 (0.058)	0.938 (0.025)	0.907 (0.059)	0.961 (0.019)
TPBN_MCMC(C)	0.946 (0.042)	0.977 (0.018)	0.898 (0.042)	0.947 (0.024)	0.928 (0.038)	0.967 (0.017)
TPBN_MCMC(R)	0.814 (0.069)	0.898 (0.037)	0.804 (0.065)	0.896 (0.038)	0.792 (0.068)	0.877 (0.038)
TPBN_MCEM(G)	0.932 (0.048)	0.969 (0.023)	0.888 (0.047)	0.943 (0.021)	0.912 (0.043)	0.957 (0.018)
TPBN_MCEM(C)	0.943 (0.040)	0.972 (0.019)	0.901 (0.042)	0.946 (0.022)	0.916 (0.042)	0.960 (0.018)
TPBN_MCEM(R)	0.880 (0.051)	0.926 (0.034)	0.856 (0.053)	0.918 (0.026)	0.848 (0.059)	0.898 (0.028)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

B.3 Root mean square error on A in terms of DSS estimates

Table 11: Root mean square errors on non-zero entries of $\boldsymbol{A}$ for different simulation scenarios with $T = 50$.

	S20	S50	I20	I50	D20	D50
NG(G)	1.240 (2.515)	158.782 (380.134)	68.381 (450.56)	21.421 (67.354)	5.594 (35.332)	134.862 (294.822)
NG(C)	0.283 (0.423)	0.538 (0.672)	0.202 (0.240)	0.199 (0.120)	0.104 (0.036)	0.193 (0.759)
NG(R)	0.197 (0.081)	0.291 (0.065)	0.141 (0.044)	0.153 (0.035)	0.097 (0.034)	0.089 (0.038)
HS(G)	0.915 (1.957)	0.447 (0.210)	1.649 (6.929)	0.252 (0.183)	1.273 (3.455)	0.172 (0.183)
HS(C)	0.366 (0.244)	0.837 (0.667)	0.277 (0.121)	0.377 (0.491)	0.295 (0.449)	0.339 (0.619)
HS(R)	0.310 (0.162)	0.556 (0.281)	0.244 (0.087)	0.267 (0.063)	0.174 (0.091)	0.231 (0.135)
SSVS	0.433 (0.230)	2.100 (2.358)	0.335 (0.193)	0.436 (0.425)	0.335 (0.345)	0.630 (0.707)
Minnesota	0.257 (0.157)	0.856 (0.556)	0.288 (0.139)	0.384 (0.104)	0.116 (0.098)	0.249 (0.161)
TPBN_MCMC(G)	0.133 (0.054)	0.210 (0.035)	0.121 (0.042)	0.133 (0.029)	0.047 (0.032)	0.032 (0.014)
TPBN_MCMC(C)	0.163 (0.063)	0.240 (0.063)	0.128 (0.052)	0.157 (0.058)	0.053 (0.032)	0.042 (0.032)
TPBN_MCMC(R)	0.198 (0.071)	0.299 (0.084)	0.139 (0.034)	0.146 (0.033)	0.099 (0.041)	0.069 (0.030)
TPBN_MCEM(G)	0.132 (0.050)	0.211 (0.037)	0.115 (0.027)	0.133 (0.028)	0.046 (0.031)	0.031 (0.012)
TPBN_MCEM(C)	0.146 (0.049)	0.234 (0.059)	0.129 (0.058)	0.158 (0.069)	0.053 (0.037)	0.039 (0.027)
TPBN_MCEM(R)	0.180 (0.055)	0.288 (0.072)	0.131 (0.032)	0.146 (0.031)	0.084 (0.033)	0.068 (0.025)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 12: Root mean square errors on non-zero entries of $\boldsymbol{A}$ for different simulation scenarios with $T = 150$.

	S20	S50	I20	I50	D20	D50
NG(G)	1.423 (2.456)	1.179 (3.730)	0.859 (1.405)	0.214 (0.219)	0.461 (1.315)	0.340 (1.077)
NG(C)	0.228 (0.186)	0.460 (0.658)	0.158 (0.094)	0.206 (0.164)	0.158 (0.327)	0.213 (0.455)
NG(R)	0.159 (0.068)	0.186 (0.054)	0.132 (0.049)	0.139 (0.065)	0.084 (0.040)	0.087 (0.063)
HS(G)	1.230 (2.061)	0.785 (2.431)	0.961 (1.064)	0.826 (1.977)	0.753 (1.224)	0.735 (1.848)
HS(C)	0.530 (0.358)	0.685 (1.141)	0.366 (0.237)	0.277 (0.192)	0.279 (0.217)	0.383 (0.519)
HS(R)	0.440 (0.263)	0.369 (0.252)	0.307 (0.100)	0.220 (0.085)	0.228 (0.105)	0.236 (0.212)
SSVS	0.468 (0.346)	0.864 (0.942)	0.324 (0.117)	0.259 (0.089)	0.217 (0.134)	0.586 (0.836)
Minnesota	0.425 (0.354)	0.496 (0.426)	0.401 (0.390)	0.291 (0.111)	0.112 (0.103)	0.162 (0.094)
TPBN_MCMC(G)	0.104 (0.059)	0.125 (0.058)	0.115 (0.052)	0.116 (0.042)	0.056 (0.073)	0.039 (0.039)
TPBN_MCMC(C)	0.126 (0.075)	0.159 (0.087)	0.119 (0.045)	0.135 (0.067)	0.044 (0.021)	0.044 (0.035)
TPBN_MCMC(R)	0.166 (0.064)	0.217 (0.076)	0.139 (0.050)	0.138 (0.038)	0.087 (0.043)	0.091 (0.045)
TPBN_MCEM(G)	0.097 (0.044)	0.124 (0.051)	0.104 (0.042)	0.119 (0.051)	0.041 (0.024)	0.034 (0.017)
TPBN_MCEM(C)	0.115 (0.066)	0.151 (0.084)	0.113 (0.052)	0.137 (0.078)	0.044 (0.026)	0.037 (0.022)
TPBN_MCEM(R)	0.143 (0.061)	0.204 (0.068)	0.118 (0.048)	0.141 (0.061)	0.069 (0.031)	0.086 (0.041)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 13: Root mean square errors on non-zero entries of $\boldsymbol{A}$ for different simulation scenarios with $T = 250$.

	S20	S50	I20	I50	D20	D50
NG(G)	0.948 (1.583)	2.044 (4.231)	0.830 (1.077)	0.503 (1.575)	0.802 (1.265)	0.532 (1.367)
NG(C)	0.191 (0.129)	0.365 (0.457)	0.194 (0.166)	0.188 (0.155)	0.107 (0.117)	0.213 (0.315)
NG(R)	0.148 (0.070)	0.199 (0.148)	0.140 (0.075)	0.127 (0.034)	0.086 (0.068)	0.065 (0.026)
HS(G)	0.847 (1.349)	2.114 (5.000)	0.713 (0.879)	0.867 (1.583)	0.780 (1.082)	0.622 (1.332)
HS(C)	0.441 (0.289)	0.616 (1.241)	0.352 (0.230)	0.325 (0.369)	0.293 (0.268)	0.264 (0.279)
HS(R)	0.372 (0.217)	0.344 (0.200)	0.299 (0.158)	0.229 (0.093)	0.231 (0.170)	0.180 (0.107)
SSVS	0.456 (0.400)	0.825 (1.362)	0.341 (0.275)	0.255 (0.111)	0.248 (0.206)	0.354 (0.361)
Minnesota	0.320 (0.253)	0.418 (0.205)	0.322 (0.212)	0.336 (0.254)	0.144 (0.167)	0.134 (0.089)
TPBN_MCMC(G)	0.100 (0.066)	0.159 (0.186)	0.119 (0.073)	0.115 (0.074)	0.044 (0.039)	0.028 (0.011)
TPBN_MCMC(C)	0.102 (0.053)	0.159 (0.088)	0.113 (0.057)	0.120 (0.044)	0.041 (0.028)	0.031 (0.013)
TPBN_MCMC(R)	0.164 (0.086)	0.229 (0.129)	0.135 (0.069)	0.140 (0.045)	0.087 (0.051)	0.081 (0.040)
TPBN_MCEM(G)	0.097 (0.068)	0.145 (0.095)	0.101 (0.062)	0.106 (0.038)	0.040 (0.028)	0.029 (0.010)
TPBN_MCEM(C)	0.104 (0.063)	0.164 (0.099)	0.109 (0.064)	0.117 (0.053)	0.040 (0.028)	0.031 (0.011)
TPBN_MCEM(R)	0.129 (0.073)	0.218 (0.100)	0.115 (0.061)	0.125 (0.033)	0.064 (0.035)	0.072 (0.033)

NOTES: NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

C Forecasting results on macroeconomic data

We provide the full forecasting results on the macroeconomic data in Table 14, Table 15. We haven't completed the experiment for the large version yet.

Table 14: Log Predictive Scores on the macroeconomic datatset for the"small" case.

	FH = 1	FH = 2	FH = 3	FH = 4
NG (G)	12.721	11.800	11.198	10.842
NG (C)	12.706	11.769	11.181	10.841
NG (R)	12.697	11.761	11.167	10.820
HS (G)	12.927	12.205	11.582	11.201
HS (C)	12.869	12.128	11.527	11.150
HS (R)	12.866	12.170	11.553	11.173
SSVS	13.047	12.293	11.658	11.252
Minnesota	12.962	12.241	11.630	11.272
TPBN_MCMC (G)	12.859	12.133	11.515	11.136
TPBN_MCMC (C)	12.978	12.178	11.537	11.157
TPBN_MCMC (R)	12.810	12.038	11.432	11.075
TPBN_MCEM (G)	12.970	12.171	11.548	11.175
TPBN_MCEM (C)	12.981	12.170	11.560	11.171
TPBN_MCEM (R)	12.932	12.184	11.555	11.167

NOTES: FH stands for forecast horizon. NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 15: Log predictive scores on the macroeconomic datatset - for the "medium" case.

	FH = 1	FH = 2	FH = 3	FH = 4
NG (G)	27.028	26.138	25.489	25.118
NG (C)	27.000	26.090	25.490	25.094
NG (R)	27.002	26.118	25.464	25.022
HS (G)	26.850	26.499	25.868	25.347
HS (C)	26.732	26.356	25.778	25.343
HS (R)	26.837	26.435	25.842	25.329
SSVS	27.067	26.492	25.752	25.153
Minnesota	26.910	26.528	25.697	25.264
TPBN_MCMC (G)	26.944	26.324	25.684	25.177
TPBN_MCMC (C)	27.056	26.479	25.894	25.428
TPBN_MCMC (R)	26.488	25.969	25.425	24.967
TPBN_MCEM (G)	27.127	26.492	25.888	25.361
TPBN_MCEM (C)	27.142	26.464	25.897	25.408
TPBN_MCEM (R)	27.001	26.437	25.850	25.391

NOTES: FH stands for forecast horizon. NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

Table 16: Log predictive scores on the macroeconomic datatset - for the "large" case.

	FH = 1	FH = 2	FH = 3	FH = 4
NG (G)	67.288	66.374	65.179	64.661
NG (C)	65.949	65.872	64.755	64.311
NG (R)	50.119	48.812	48.742	49.741
HS (G)	64.764	68.843	68.636	68.141
HS (C)	57.719	65.124	66.216	66.574
HS (R)	61.099	67.148	66.195	70.106
SSVS	64.780	68.757	67.888	67.717
Minnesota	57.180	67.246	66.825	66.750
TPBN_MCMC (G)	68.634	69.299	69.371	68.447
TPBN_MCMC (C)	68.795	70.139	69.801	68.828
TPBN_MCMC (R)	60.720	61.551	63.032	65.690

NOTES: FH stands for forecast horizon. NG refers to a vector autoregressive model with a normal gamma prior, HS to the Horseshoe prior. SSVS to stochastic search variable selection prior, Minnesota to hierarchical Minnesota prior and TPBN to three-parameter-beta-normal prior. G refers to global, C refers to columnwise and R refers to rowwise.

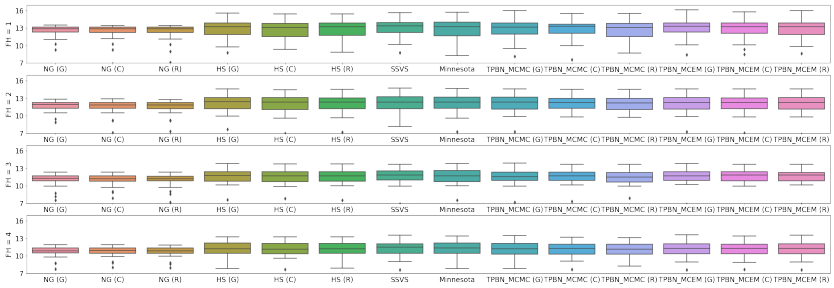

Figure 1: Boxplots of average log predictive density over 50 runs for different models in the small version of macroeconomic dataset.

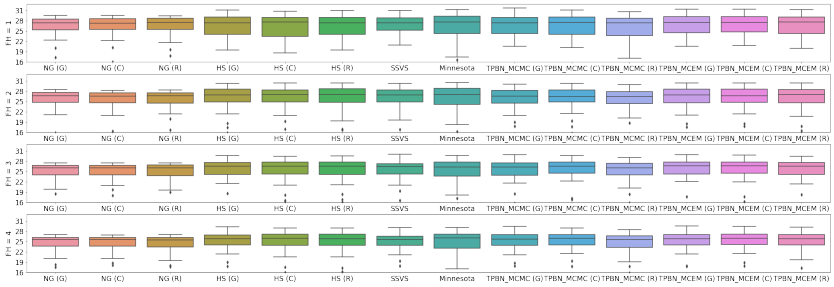

Figure 2: Boxplots of average log predictive density over 50 runs for different models in the medium version of macroeconomic dataset.

We also provide the boxplots over 50 runs in the subsets of the datasets with different dimensions sizes. The boxplots for small, and medium version are displayed in Figure 1, Figure 2. Comparing TPBN with other global-local prior based models, we find that TPBN and HS performs notably better than NG approach. Moreover, comparing TPBN with SSVS and Minnesota prior, TPBN has notably equal or smaller variance than those models.