Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models

Satyajit Ghosh, Kshitij Khare & George Michailidis

To cite this article: Satyajit Ghosh, Kshitij Khare & George Michailidis (2018): High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1437043

To link to this article: https://doi.org/10.1080/01621459.2018.1437043

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 0, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2018.1437043

Taylor & Francis
Taylor & Francis Group

High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models

Satyajit Ghosh, Kshitij Khare, and George Michailidis

Department of Statistics and the Informatics Institute, University of Florida, Gainesville, FL

ABSTRACT

Vector autoregressive (VAR) models aim to capture linear temporal interdependencies among multiple time series. They have been widely used in macroeconomics and financial econometrics and more recently have found novel applications in functional genomics and neuroscience. These applications have also accentuated the need to investigate the behavior of the VAR model in a high-dimensional regime, which provides novel insights into the role of temporal dependence for regularized estimates of the model's parameters. However, hardly anything is known regarding properties of the posterior distribution for Bayesian VAR models in such regimes. In this work, we consider a VAR model with two prior choices for the autoregressive coefficient matrix: a nonhierarchical matrix-normal prior and a hierarchical prior, which corresponds to an *arbitrary* scale mixture of normals. We establish posterior consistency for both these priors under standard regularity assumptions, when the dimension p of the VAR model grows with the sample size n (but still remains smaller than n). A special case corresponds to a shrinkage prior that introduces (group) sparsity in the columns of the model coefficient matrices. The performance of the model estimates are illustrated on synthetic and real macroeconomic datasets. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2017
Revised January 2018

KEYWORDS
Bayesian lasso; Posterior consistency; Shrinkage prior; Vector autoregressive models

1. Introduction

There has been recent interest in modeling high-dimensional time series datasets. In macroeconomics, Mol, Giannone, and Reichlin (2008) advocated the need to include a large number of variables in econometric models to improve forecastability, while Billio et al. (2012) examined stock returns of many financial institutions to assess systemic risk of the financial system. Similar modeling challenges arise in functional genomics for the reconstruction of regulatory networks as discussed in Basu, Shojaie, and Michailidis (2015), while in neuroscience one is interested in understanding functional connectivity between brain regions (Seth, Barrett, and Barnett 2015).

A popular and informative model has been vector autoregressions (VAR), that captures linear temporal dependencies between time series. The VAR model and its properties have been thoroughly explored in low-dimensional settings both from a frequentist (for a comprehensive overview see Lütkepohl 2007) and a Bayesian perspective (Bańbura, Giannone, and Reichlin 2010).

More recently, Basu and Michailidis (2015) provided an in-depth analysis of the model for Gaussian data in a *high-dimensional setting under sparsity assumptions*, while Melnyk and Banerjee (2016) extended the results to other regularizers (e.g., group lasso, sparse group lasso, etc.). The results of Basu and Michailidis (2015); Melnyk and Banerjee (2016) and related follow-up work (Raskutti, Yuan, and Chen (2018); Schweinberger, Babkin, and Ensor (2017); Lin and Michailidis (2017)) indicate that the resulting estimation error rates are those obtained for independent and identically distributed data times a factor that captures the temporal dependence in the data.

On the Bayesian front, there has been primarily methodological/computational work for low-dimensional VAR models. The so-called Minnesota prior (Litterman 1979; Doan, Litterman, and Sims 1984) has been a staple of applied econometric work involving VAR models. This is a normal prior distribution on the elements of the transition matrix that puts stronger weights on the "own" lags of each time series, since they are considered more informative for forecasting purposes than lags from "other" time series. For large size VAR models, Bańbura, Giannone, and Reichlin (2010) advocated normal-inverted Wishart distribution that leads to a posterior mean that can be interpreted as a ridge shrinkage estimator, suitable for such models. A first attempt for Bayesian estimation of VAR models combined with variable selection is presented in Korobilis (2013), where an indicator variable is specified for each parameter in the transition matrix that indicates whether the cross-autocorrelation coefficient is included or set to zero. A prior is specified for the indicator variables that in principle can also be combined with the Minnesota prior.

On the other hand, Bayesian investigations into high-dimensional asymptotics of statistical models that *incorporate sparsity* with *temporally dependent data* are not in general available to the best of our knowledge. Hence, the main objective of this work is to study posterior (estimation) consistency for a VAR model, which asserts that the posterior concentrates around the "true" parameter value (in an appropriate norm)

CONTACT George Michailidis gmichail@ufl.edu Department of Statistics and the Informatics Institute, University of Florida, Gainesville, FL 32611.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

as the sample size increases. There is a rich literature on high-dimensional posterior estimation consistency for linear regression models for independent and identically distributed data. Ghoshal (1999) established posterior consistency and asymptotic normality with a general prior on the p-vector of regression coefficients (with appropriate positivity and Lipschitz assumptions) when $p^3 \log p/n \to 0$ and $p^4 \log p/n \to 0$, respectively. Bontemps (2011) extended the work of Ghoshal (1999) by permitting the model to be misspecified and the number of predictor variables to grow proportionally to the sample size. Armagan et al. (2013) focused on shrinkage priors, which are appropriate scale mixtures of normal priors and induced weak sparsity in the vector of regression coefficients (see Carvalho, Polson, and Scott 2010; Griffin and Brown 2010; Armagan, Clyde, and Dunson 2011; Armagan, Dunson, and Lee 2013). They established posterior consistency under a simple sufficient condition on prior concentration when $p = o(n)$. Lee and Oh (2013) established posterior consistency under a high-dimensional Bayesian principal component analysis (PCA) regression setup with $p > n$ under appropriate assumptions on the rank of the design matrix. Posterior estimation consistency in linear regression models with g-priors has also been addressed in Sparks, Khare, and Ghosh (2015).

A crucial difference between the linear regression models considered in the above work and VAR models (expressed as a linear model) is that the design matrix in the latter case is random, and exhibits dependencies both between its rows and across its columns, and also with the error term in the model (see Section 2). This leads to a significantly more involved and challenging theoretical analysis that we successfully resolve. In this article, we investigate high-dimensional posterior consistency for Bayesian VAR models in two natural and relevant settings: (a) with a nonhierarchical matrix normal prior on the $dp \times p$ autoregressive parameter matrix and (b) a hierarchical prior which corresponds to a general scale mixture of normals. In particular, this includes spherically symmetric priors such as the multivariate-t and standard shrinkage priors which induce (group) sparsity in the columns of the coefficient matrices, such as the group structure in Basu, Shojaie, and Michailidis (2015). Further, we employ a flat (uniform) prior distribution for the error term. Note that the joint maximum likelihood estimation problem for a sparse VAR model, with a sparse error covariance matrix is investigated in Lin and Michailidis (2017). The posterior consistency results are established under mild regularity assumptions on the underlying spectral density and with $p = o(n/\log n)$. The key to handling the dependencies, within the design matrix and also between the design matrix and the error term, is a pair of high-dimensional concentration inequalities established in the supplementary material (Propositions B1 and B3). Note that we are considering the "large p large n" setting with $p = o(n)$. However, we make no assumption reducing the effective dimension of the "true parameter matrix." We only assume that the matrix norm of the true parameter matrix is of the order p in the nonhierarchical prior setting and bounded by a constant in the hierarchical prior setting. The large p small n situation, where p is allowed to grow at a much faster rate than n is also of interest, but assumptions such as sparsity/restricted eigenvalue type conditions are required, which in turn reduce the effective dimension of the true parameters. General posterior consistency results for VAR models in the large p small n setting are also not available to the best of our knowledge and are topics of future discussion/research.

The remainder of the article is organized as follows. In Section 2, we introduce the VAR model and necessary notions of posterior consistency. We consider the nonhierarchical matrix normal prior on the coefficient matrix in Section 3.1 and establish posterior consistency under suitable regularity assumptions. In Section 3.3, we prove posterior consistency considering a hierarchical prior corresponding to a scale-mixture of matrix normals. In Sections 4 and 5, the methodology/results of this article are illustrated on simulated and real datasets, respectively. Finally, we conclude with a discussion in Section 6.

1.1. Notation

Throughout this article, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ denote the sets of integers, real numbers, and complex numbers, respectively. We denote the cardinality of a set J by $|J|$. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\| := \sqrt{\sum v_j^2}$ denotes the ℓ_2-norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ and $\sigma_{\max}(\mathbf{A})$ denote spectral norm, that is, $\|\mathbf{A}\| = \sup_{x \neq 0} \frac{\|\mathbf{A}x\|_2}{\|x\|_2}$ and the largest singular value of $\mathbf{A}$, respectively. For a symmetric or Hermitian matrix $\mathbf{A}$, we denote its maximum and minimum eigenvalues by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$. The vector e_i is used for the ith unit vector in $\mathbb{R}^p$. Bold uppercase letters are only used to denote matrices, and vectorized form of such matrices is represented by corresponding lower cases. For example, if $\boldsymbol{\Phi}$ is a $p \times p$ matrix then $\boldsymbol{\phi}$ is $\text{vec}(\boldsymbol{\Phi})$. Also, $\mathbf{O}$ represents a zero-matrix of appropriate dimension, and in general vectors are denoted by italicized bold lowercase letters.

2. Model Formulation

For a p-dimensional stationary time series $\{X^t\}$, a vector autoregressive model of lag-d is given by

$$X^t = \boldsymbol{c} + \sum_{i=1}^{d} \mathbf{A}_i X^{t-i} + \boldsymbol{\varepsilon}^t. \tag{1}$$

The temporal dependence structure of the VAR model is characterized by the $p \times p$ transition matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d$ and $\boldsymbol{c}$ is a $p \times 1$ location vector which we choose to be $\mathbf{0}$. In the Gaussian VAR, the errors $\boldsymbol{\varepsilon}^t$ are iid $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is a $p \times p$ unknown error covariance matrix. The model in (1) can be rewritten in the Yule–Walker representation (Lütkepohl 2007) as

$$X^t - \boldsymbol{\mu} = \sum_{i=1}^{d} \mathbf{A}_i (X^{t-i} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}^t,$$

where $\boldsymbol{\mu} = (\mathbf{I} - \mathbf{A}_1 - \mathbf{A}_2 - \cdots - \mathbf{A}_d)^{-1}\boldsymbol{c}$ is known as the process mean. Usually $\boldsymbol{\mu}$ will not be known in advance. In that case, $\boldsymbol{\mu}$ may be estimated by the vector of sample means $\bar{X} = \sum_1^n X^t$. An alternative estimator is $\hat{\boldsymbol{\mu}} = (\mathbf{I} - \hat{\mathbf{A}}_1 - \hat{\mathbf{A}}_2 - \cdots - \hat{\mathbf{A}}_d)^{-1}\hat{\boldsymbol{c}}$ in which $\hat{\boldsymbol{c}}$ and $\hat{\mathbf{A}}_i$'s are the least-square estimator. Henceforth we assume without loss of generality $\boldsymbol{\mu} = \mathbf{0}$. Based on the data $\{X^0, \ldots, X^T\}$, we define the response matrix $\mathbf{Y}$ and design

matrix $\mathbf{X}$ as follows,

$$\mathbf{Y} = \begin{bmatrix} (X^T)' \\ \vdots \\ (X^d)' \end{bmatrix}_{n\times p} \quad \mathbf{X} = \begin{bmatrix} (X^{T-1})' & \cdots & (X^{T-d})' \\ \vdots & \ddots & \vdots \\ (X^{d-1})' & \cdots & (X^0)' \end{bmatrix}_{n\times dp}.$$

We can now rewrite the above model in a linear regression setup as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Phi} + \mathbf{E}, \tag{2}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \\ \vdots \\ \mathbf{A}_d' \end{bmatrix} \qquad \mathbf{E} = \begin{bmatrix} (\boldsymbol{\varepsilon}^T)' \\ \vdots \\ (\boldsymbol{\varepsilon}^d)' \end{bmatrix}.$$

In this formulation, the number of samples is $n = T - d + 1$ and the number of unknown parameters is $q = dp^2$, respectively. Vectorizing (column-wise) each matrix, we get

$$\boldsymbol{y} := \text{vec}(\mathbf{Y}) = \mathbf{Z}\boldsymbol{\phi} + \boldsymbol{\varepsilon},$$

where $\mathbf{Z} := (\mathbf{I}_p \otimes \mathbf{X})$, $\boldsymbol{\phi} = \text{vec}(\boldsymbol{\Phi})$, and $\boldsymbol{\varepsilon} = \text{vec}(\mathbf{E})$. In this article, we consider a high-dimensional setting where the dimension p of the VAR model increases with the sample size n. However, we assume that the lag d does not vary with n. This basic formulation of regression lends itself easily to a Bayesian analysis in which priors are placed on the unknown parameter matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

As previously mentioned, we let the dimension $p = p_n$ of the VAR model vary with n, so that our results are relevant to high-dimensional settings. We assume that our data come from a true VAR model described as follows: for every, $n \geq 1$, let $\mathcal{Y}_n := (X^{n,0}, \ldots, X^{n,n+d-1})$ be the set of observations for sample size n, which satisfy $X^{n,k} = \sum_{i=1}^{d} \mathbf{A}_{i,0n} X^{n,k-i} + \boldsymbol{\varepsilon}^{n,k}$ for $d \leq k \leq n + d - 1$. The errors $\{\boldsymbol{\varepsilon}^{n,k}\}_{k=d}^{n+d-1}$ are iid $\mathcal{N}_{p_n}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0n})$. Here $\{\boldsymbol{\Phi}_{0n}\}_{n\geq 1}$ denotes the sequence of the true coefficient matrices given by $\boldsymbol{\Phi}_{0n}' := [\mathbf{A}_{1,0n}\ \mathbf{A}_{2,0n}\ \ldots\ \mathbf{A}_{d,0n}]$, and $\{\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0n}\}_{n\geq 1}$ denotes the sequence of the true error covariance matrices. Let $\mathbb{P}_0$ denote the probability measure underlying the true model described above.

Next, consider a Bayesian model which builds on (2) by placing priors on the parameters $(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. In particular, let $\{\pi_n(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})\}_{n\geq 1}$ and $\{\pi_n(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \mid \mathcal{Y}_n)\}_{n\geq 1}$ denote the sequences of the corresponding (joint) prior and posterior densities. Analogously, $\{\Pi_n(\cdot)\}_{n\geq 1}$ and $\{\Pi_n(\cdot \mid \mathcal{Y}_n)\}_{n\geq 1}$ denote the corresponding sequences of (joint) prior and posterior distributions. We will also use the notation π_n an Π_n to denote the marginal prior and posterior densities/distributions for $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ as appropriate.

Note that our main parameter of interest is $\boldsymbol{\Phi}$, while the error covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is more of an unknown nuisance parameter that we need to deal with. One would hope that as the sample size n tends to infinity, the posterior probability assigned to any ε neighborhood of $\boldsymbol{\Phi}_{0n}$ converges to 1 almost surely under $\mathbb{P}_0$. We now formally define a notion of posterior consistency that formalizes this.

Definition 1. The sequence of posterior distributions $\Pi_n(\cdot \mid \mathcal{Y}_n)$ is said to be consistent at $\{\boldsymbol{\Phi}_{0n}\}_{n\geq 1}$, if for every $\varepsilon > 0$, $\Pi_n(\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_{0n}\| > \varepsilon \mid \mathcal{Y}) \to 0$ as $n \to \infty$ a.s. $\mathbb{P}_0$.

For ease of exposition, we will henceforth denote $\boldsymbol{\Phi}_{0n}$ as $\boldsymbol{\Phi}_0$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0n}$ as $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0}$, and highlight their dependence on n as needed.

2.1. Stability of VAR(d) Process

Since VAR models are dynamical systems, the notion of "stability" plays an important role in their analysis and asymptotic properties.

Definition 2. A VAR(d) process defined in (1) is said to be *stable* if the matrix valued polynomials $\mathcal{A}(z) := \mathbf{I}_p - \sum_{i=1}^{d} \mathbf{A}_i z^i$ satisfies $\det(\mathcal{A}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathbb{C} : |z| = 1\}$.

The autocovariance function of a p-dimensional centered covariance-stationary time series $\{X^t\}$ is defined as $\Gamma_X(h) = \text{cov}(X^t, X^{t+h})\ \ t, h \in \mathbb{Z}$ and the corresponding spectral density is given by $f_X(\theta) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_X(h) e^{-ih\theta}, \quad \theta \in [-\pi, \pi]$. For a Gaussian stable VAR(d) model, the spectral density has a closed-form expression,

$$f_X(\theta) = \frac{1}{2\pi} \left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j e^{-ij\theta} \right)^{-1} \times \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \left[\left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j e^{-ij\theta} \right)^{-1} \right]^*,$$

where * denotes the Hermitian conjugate of a matrix and $i \equiv \sqrt{-1}$. The autocovariance function which characterizes a centered Gaussian process, can be used to quantify the temporal and cross-sectional dependence for VAR(d) models. The peak of the spectral density, measured by its maximum eigenvalue $\mathcal{M}(f_X) := \max_{\theta\in[-\pi,\pi]} \lambda_{\max}(f_X(\theta))$ can be used as a measure of stability of the process. Also the minimum eigenvalue $\mathfrak{m}(f_X) := \min_{\theta\in[-\pi,\pi]} \lambda_{\max}(f_X(\theta))$ captures cross-dependence among its components. However, as mentioned in Basu and Michailidis (2015) instead of working with $\mathcal{M}(f_X)$ and $\mathfrak{m}(f_X)$ it is often easier to work with the eigenvalues of $\mathcal{A}^*(z)\mathcal{A}(z)$ over the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. Let

$$\mu_{\min}(\mathcal{A}) := \min_{|z|=1} \lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z))$$
$$= \min_{\theta\in[-\pi,\pi]} \lambda_{\min} \left(\left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j' e^{ij\theta} \right) \left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j e^{-ij\theta} \right) \right)$$
$$\mu_{\max}(\mathcal{A}) := \min_{|z|=1} \lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))$$
$$= \max_{\theta\in[-\pi,\pi]} \lambda_{\max} \left(\left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j' e^{ij\theta} \right) \left(\mathbf{I}_p - \sum_{j=1}^{d} \mathbf{A}_j e^{-ij\theta} \right) \right).$$

For stable VAR(d) process, $0 < \mu_{\min}(\mathcal{A}) \leq \mu_{\max}(\mathcal{A}) < \infty$. Since each $\boldsymbol{\varepsilon}^t$ is iid as $\mathcal{N}_p(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, each row of $\mathbf{X}$ is distributed

as $\mathcal{N}_{dp}(0, \mathbf{C}_X)$, where the covariance matrix $\mathbf{C}_X$ has the following structure,

$$\mathbf{C}_X = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \cdots & \Gamma(d-1) \\ \Gamma(1)' & \Gamma(0) & \cdots & \Gamma(d-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(d-1)' & \Gamma(d-2)' & \cdots & \Gamma(0) \end{bmatrix}_{dp \times dp}. \quad (3)$$

The quantities $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ provide a useful bound for the eigenvalues of $\mathbf{C}_X$. As mentioned in Melnyk and Banerjee (2016) and from Proposition 2.3 and eq. (2.6) of Basu and Michailidis (2015), we have the following chain of inequalities,

$$\begin{aligned} \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})}{\mu_{\max}(\mathcal{A})} &\leq 2\pi \mathfrak{m}(f_X) \leq \lambda_{\min}(\mathbf{C}_X) \leq \lambda_{\max}(\mathbf{C}_X) \\ &\leq 2\pi \mathcal{M}(f_X) \leq \frac{\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})}{\mu_{\min}(\mathcal{A})}. \end{aligned} \quad (4)$$

We finally note that the p-dimensional VAR(d) model in (2) can be equivalently written as a dp-dimensional VAR(1) process. Let

$$\tilde{X}^t = \begin{bmatrix} X^t \\ \vdots \\ X^{t-d+1} \end{bmatrix}_{dp \times 1}, \quad \tilde{\mathbf{A}}_1 = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{d-1} & \mathbf{A}_d \\ \mathbf{I}_p & \mathbf{O} & \cdots & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_p & \cdots & \mathbf{O} & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{I}_p & \mathbf{O} \end{bmatrix}_{dp \times dp},$$

and

$$\boldsymbol{\omega}^t = \begin{bmatrix} \boldsymbol{\varepsilon}^{t+1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{dp \times 1}.$$

Then the new representation becomes

$$\tilde{X}^t = \tilde{\mathbf{A}}_1 \tilde{X}^{t-1} + \boldsymbol{\omega}^t \qquad t = d, \ldots, n+d-1. \quad (5)$$

It follows that $\mathbf{X} = [\tilde{X}^{n+d-2}\ \tilde{X}^{n+d-3} \cdots \tilde{X}^{d-1}]'$, that is, ith row of $\mathbf{X}$ is denoted as ($dp \times 1$ vector) $\tilde{X}^{n+d-i-1}$. Note that if the underlying VAR(d) process $\{X^t\}$ is stable then the process $\tilde{X}^t$ with the characteristic polynomial, $\tilde{\mathcal{A}}(z) := \mathbf{I}_{dp} - \tilde{\mathbf{A}}_1 z$ is also stable. This is because $\{\tilde{X}^t\}$ can be viewed as generated according to a stable VAR(1) process with transition matrix $\tilde{\mathbf{A}}_1$ and $\{\tilde{X}^t\}$ is stable if and only if $\{X^t\}$ is stable (Lütkepohl 2007). Based on $\tilde{\mathcal{A}}(z)$, we define

$$\begin{aligned} \mu_{\min}(\tilde{\mathcal{A}}) &:= \min_{\theta \in [-\pi, \pi]} \lambda_{\min}\left(\left(\mathbf{I}_p - \tilde{\mathbf{A}}_1' e^{i\theta}\right)\left(\mathbf{I}_p - \tilde{\mathbf{A}}_1 e^{-i\theta}\right)\right) \\ \mu_{\max}(\tilde{\mathcal{A}}) &:= \max_{\theta \in [-\pi, \pi]} \lambda_{\max}\left(\left(\mathbf{I}_p - \tilde{\mathbf{A}}_1' e^{i\theta}\right)\left(\mathbf{I}_p - \tilde{\mathbf{A}}_1 e^{-i\theta}\right)\right). \end{aligned} \quad (6)$$

While $\mu_{\min}(\tilde{\mathcal{A}})$ and $\mu_{\max}(\tilde{\mathcal{A}})$ are not necessarily the same as $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$, the inequalities in (4) still hold with $\mu_{\min}(\mathcal{A})$ and $\mu_{\min}(\mathcal{A})$ replaced by $\mu_{\min}(\tilde{\mathcal{A}})$ and $\mu_{\max}(\tilde{\mathcal{A}})$, respectively.

3. Bayesian Estimation and Posterior Consistency

In this section, we first discuss Bayesian estimation of VAR models with nonhierarchical and hierarchical scale mixture matrix normal prior distributions on the parameter matrix $\mathbf{\Phi}$ (conditioned on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$) and subsequently establish high-dimensional posterior consistency in this setting under mild regularity assumptions. We start by introducing the necessary notation for the *matrix-variate normal distribution*. Let $M_{a,b}$ denote the space of $a \times b$ matrices.

Definition 3. An $a \times b$ random matrix $\mathbf{X}$ is defined to follow a matrix-variate normal distribution ($\mathcal{MN}_{a\times b}(\mathbf{M}, \mathbf{B}_1, \mathbf{B}_2)$) if its density function (on the space $M_{a,b}$) is given by

$$\frac{|\mathbf{B}_1|^{-b/2}|\mathbf{B}_2|^{-a/2}}{(2\pi)^{ab/2}} e^{-\frac{1}{2}\mathrm{tr}\{\mathbf{B}_1^{-1}(\mathbf{X}-\mathbf{M})\mathbf{B}_2^{-1}(\mathbf{X}-\mathbf{M})'\}}.$$

Here $\mathbf{M} \in M_{a,b}$. $\mathbf{B}_1 \in M_{a,a}$ and $\mathbf{B}_2 \in M_{b,b}$ which are both positive definite matrices corresponding to the variances among the rows and columns of $\mathbf{X}$, respectively. Note that the matrix normal distribution is related to the multivariate normal distribution in the following way: $\mathbf{X} \sim \mathcal{MN}_{n\times p}(\mathbf{M}, \mathbf{B}_1, \mathbf{B}_2)$, if and only if $\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}_{np}(\mathrm{vec}(\mathbf{M}), \mathbf{B}_2 \otimes \mathbf{B}_1)$.

3.1. Nonhierarchical Matrix Normal Prior

We consider a matrix normal prior for $\mathbf{\Phi}$ conditional on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$, and a flat (uniform) prior on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$. In particular,

$$\mathbf{\Phi} \mid \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \sim \mathcal{MN}_{dp\times p}(\mathbf{O}, \mathbf{U}^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \text{ and } \pi(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \propto 1, \quad (7)$$

where $\mathbf{U}$ is a $dp \times dp$ known positive definite matrix. Note that under this matrix normal prior $\mathbf{U}^{-1}$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ are the covariance matrices corresponding to the columns and rows of $\mathbf{\Phi}$, respectively. The posterior distribution of $\mathbf{\Phi}$ (conditional on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$) can easily be shown to be $\mathcal{MN}_{dp\times p}(\mathbf{\Phi}_{\mathrm{PM}}, (\mathbf{X}'\mathbf{X} + \mathbf{U})^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$, where $\mathbf{\Phi}_{\mathrm{PM}} := (\mathbf{X}'\mathbf{X} + \mathbf{U})^{-1}\mathbf{X}'\mathbf{Y}$ is the (conditional) posterior mean which does not depend on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$. Hence, the unconditional posterior mean of $\mathbf{\Phi}$ is available in closed form and is given by $\mathbf{\Phi}_{\mathrm{PM}}$. It follows by standard computations using the multivariate normal density that the marginal posterior density of $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ is proportional to

$$|\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}|^{-n/2}\, e^{-tr(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\widehat{\mathbf{\Sigma}}_{res})},$$

where $\widehat{\mathbf{\Sigma}}_{res} = \mathbf{Y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{U})^{-1}\mathbf{X}^T)\mathbf{Y}$. This density is proper if and only if $n > 2p$. In this case, the marginal posterior density of $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ corresponds to the Inverse-Wishart density with scale parameter $\widehat{\mathbf{\Sigma}}_{res}$ and shape parameter $n - p - 1$. We summarize the above observations in the lemma below.

Lemma 1. Under the nonhierarchical prior in (7), the posterior density of $(\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$ is proper if and only if $n > 2p$. In this case

$$\begin{aligned} \mathbf{\Phi} \mid \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathcal{Y} &\sim \mathcal{MN}_{dp\times p}\big(\mathbf{\Phi}_{\mathrm{PM}}, (\mathbf{X}'\mathbf{X} + \mathbf{U})^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\big) \\ \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \mid \mathcal{Y} &\sim \text{Inverse-Wishart}(\widehat{\mathbf{\Sigma}}_{res}, n - p - 1). \end{aligned}$$

3.2. Assumptions for Posterior Consistency

We will establish our results under the high-dimensional setting from Section 2. Recall that $\mathbf{\Phi}_0 = \mathbf{\Phi}_{0n}$ denotes the true underlying parameter matrix, and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0} = \mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}$ denotes the true underlying error covariance matrix in this setting. The quantities $\mu_{\min}(\tilde{\mathcal{A}})$, $\mu_{\max}(\tilde{\mathcal{A}})$, and $\mathbf{C}_X$ are as defined in (6) and (3), but with $\mathbf{\Phi}_0$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}$ as the underlying parameter values. We assume the following:

Assumption A1. The VAR(d) model given in (1) is stable.

Assumption A2. $\frac{1+\mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})}$ is $\mathrm{O}\left(\sqrt[5]{\frac{n}{p}}\right)$ as $n \to \infty$.

Assumption A3. $0 < \inf_{n\geq 1} \lambda_{\min}(\mathbf{C}_X) < \infty$ and $\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}) = \mathrm{O}(1)$.

Assumption A4. The true parameter matrix of the VAR model (2), $\mathbf{\Phi}_0$ and the hyperparameter $\mathbf{U}$ of (7) are such that $\left\|\mathbf{\Phi}_0^T\mathbf{U}\mathbf{\Phi}_0\right\| = \mathrm{o}(n)$ and $\|\mathbf{U}\mathbf{\Phi}_0\| = \mathrm{o}(n)$.

Assumption A5. $p = \mathrm{o}(n)$.

When $d = 1$, we deal with a VAR(1) model and $\mathbf{C}_X$ becomes $\Gamma_X(0)$, while $\mu_{\min}(\tilde{\mathcal{A}})$ is the same as $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\tilde{\mathcal{A}})$ is also equal to $\mu_{\max}(\mathcal{A})$. Assumption A1 is a standard assumption which ensures that the underlying VAR process is well-behaved. Assumption A2 plays an important role in deriving high-dimensional concentration bounds for $\mathbf{X}'\mathbf{X}/n$ and $\mathbf{X}'\mathbf{E}/n$ around $\mathbf{C}_X$ and $\mathbf{O}$, respectively (see Propositions B.1 and B.3 in the supplementary material). Assumption A3 is needed to ensure that $\lambda_{\min}(\mathbf{X}'\mathbf{X}/n)$ is bounded away from 0 with high probability. Further, if we consider that each column of $\mathbf{\Phi}$ is independently and identically distributed according to a normal prior distribution, that is, $\mathbf{U} = \mathbf{I}_{dp}$, Assumption A4 reduces to $\|\mathbf{\Phi}_0'\mathbf{\Phi}_0\|_2 = \mathrm{o}(n)$ and $\|\mathbf{\Phi}_0\| = \mathrm{o}(n)$.

We now state the main theoretical result of posterior consistency with a nonhierarchical matrix normal prior distribution on $\mathbf{\Phi}$. The proof is given in Appendix C.1 of the supplementary material.

Theorem 1 (Posterior consistency for nonhierarchical prior). For any centered VAR(d) model (2) with nonhierarchical prior (7) on $\mathbf{\Phi}$ satisfying Assumptions A1–A5, the posterior consistency of the parameter matrix can be achieved, that is, for every fixed every $\varepsilon > 0$

$$\mathbb{E}_0\left[\Pi_n(\|\mathbf{\Phi} - \mathbf{\Phi}_0\| > \varepsilon | \mathcal{Y} = (X^0, \ldots, X^n))\right] \to 0 \quad \text{as } n \to \infty$$

where $\mathbf{\Phi}_0$ is the true parameter matrix under the model (2).

A natural question to ask is whether the assumption $p = o(n)$ can be relaxed for posterior consistency. In the lemma below, we consider a situation in which Assumptions A1–A4 are satisfied and p is the same order as n, and prove that the resulting posterior is *not consistent*. The proof is given in Appendix C.2 of the supplementary material.

Lemma 2. Consider a (sequence of) VAR(1) model with $p_n = \gamma n$, $\mathbf{\Phi}_{0n} = \alpha \mathbf{I}_{p_n}$, and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n} = \mathbf{I}_{p_n}$, where $\gamma \in (0, \frac{1}{2})$, $\alpha \in (0, 1)$ do not depend on n. If we use the nonhierarchical prior (7) on $\mathbf{\Phi}$ with $\|\mathbf{U}\| = o(n)$, then there exists $\varepsilon > 0$ such that

$$\liminf_{n\to\infty} \mathbb{E}_0\left[\Pi_n(\|\mathbf{\Phi} - \mathbf{\Phi}_0\| > \varepsilon | \mathcal{Y} = (X^0, \ldots, X^n))\right] > 0.$$

Remark. Note that the condition $\|\mathbf{U}\| = o(n)$ assumed in Lemma 2 corresponds to Assumption A4 in the setting of the lemma. The reason for making this assumption is that we want to show violating Assumption A5 ($p = o(n)$) can lead to posterior inconsistency, even if all of Assumptions A1–A4 hold. If we decide to violate Assumption A4 too by assuming $\|\mathbf{U}\| = O(n)$ or $\|\mathbf{U}\| \gg n$ (goes to ∞ at the same rate or faster than n), then the posterior inconsistency proof becomes comparatively easier. We have provided the corresponding proofs in supplemental Section C.3 and supplemental Section C.4, respectively.

3.3. Hierarchical Normal-Mixture Prior

Next, we study the posterior consistency of the parameter matrix in model (2) in which $\mathbf{\Phi}$ has the following hierarchical prior:

$$\mathbf{\Phi}|\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathbf{U} \sim \mathcal{MN}_{dp\times p}(\mathbf{O}, \mathbf{U}^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}),$$
$$\pi(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \propto 1,$$

and

$$\mathbf{U} \sim \pi_{\mathrm{scl}}(.), \tag{8}$$

where $\mathbf{U}$ is the $dp \times dp$ matrix having probability density $\pi_{\mathrm{scl}}(\cdot)$ over the space of $dp \times dp$ positive definite matrices, $\mathbb{M}^+_{dp}$. As shown below, the group lasso and multivariate t distribution prior on $\mathbf{\Phi}$ can be obtained from (8) using appropriate choices of $\pi_{\mathrm{scl}}(\cdot)$. The lemma below shows that the posterior is proper if $n > (d+1)p$, and provides the form of various conditional and marginal posterior densities. The proof is given in Appendix C.5 of the supplementary material.

Lemma 3. Under the hierarchical normal-mixture prior in (8), the posterior density of $(\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathbf{U})$ is proper if $n > (d+1)p$. In this case,

$$\mathbf{\Phi} \mid \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathbf{U}, \mathcal{Y} \sim \mathcal{MN}_{dp\times p}\left(\mathbf{\Phi}_{\mathrm{PM}}, (\mathbf{X}'\mathbf{X} + \mathbf{U})^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)$$
$$\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \mid \mathbf{U}, \mathcal{Y} \sim \text{Inverse-Wishart}(\widehat{\mathbf{\Sigma}}_{res}, n - p - 1)$$
$$\pi(\mathbf{U}|\mathcal{Y}) \propto \frac{|\mathbf{U}|^{dp/2}}{|\mathbf{X}'\mathbf{X} + \mathbf{U}|^{dp/2}|\widehat{\mathbf{\Sigma}}_{res}|^{(n-p-1)/2}}\pi_{\mathrm{scl}}(\mathbf{U}).$$

The Bayesian group lasso prior was proposed by Kyung et al. (2010) in the context of linear regression. We adapt it to the VAR setting as follows. Suppose the rows of $\mathbf{\Phi}$ are divided in G groups $\mathbf{\Phi}_{[1]}, \ldots, \mathbf{\Phi}_{[G]}$, where each $\mathbf{\Phi}_{[g]}$ is an $m_g \times p$ submatrix of $\mathbf{\Phi}$ (hence $\sum m_g = dp$) and $\mathbf{X}_g$ is the submatrix of $\mathbf{X}$ of order $n \times m_g$ corresponding to the group $\mathbf{\Phi}_{[g]}$. The frequentist group lasso estimator (conditional on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$) is obtained by solving

$$\min_{\mathbf{\Phi}_{[1]},\ldots,\mathbf{\Phi}_{[G]}} \left\| \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\left(\mathbf{Y} - \sum_{g=1}^{G}\mathbf{X}_g\mathbf{\Phi}_{[g]}\right)\right\|_F^2 + \sum_{g=1}^{G}\lambda_g \left\| \mathbf{\Phi}_{[g]}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\right\|_F,$$

where λ_g is a tuning parameter corresponding to the group g. The group lasso estimator (conditional on $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$) can also be expressed as the maximum a posteriori probability (MAP) estimate under model (2) with the prior

$$\pi(\mathbf{\Phi} \mid \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \propto \exp\left(-\sum_{g=1}^{G}\lambda_g \left\|\mathbf{\Phi}_{[g]}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\right\|_F\right),$$

which is a multivariate generalization of the double exponential prior and can also be expressed as a scale mixture of normals with Gamma hyperpriors (Park and Casella 2008; Kyung et al. 2010) leading to the group lasso hierarchy,

$$\mathbf{\Phi}_{[g]}|\tau_g, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \overset{\text{ind}}{\sim} \mathcal{MN}_{m_g\times p}(\mathbf{O}, \tau_g\mathbf{I}_{m_g}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$$

and

$$\tau_g \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda_g^2}{2}\right), \quad g = 1, \ldots, G.$$

Here $\text{Gamma}(\alpha, \lambda)$ denotes the Gamma distribution with shape parameter α and rate parameter λ. This can be alternatively presented as $\mathbf{\Phi}|\boldsymbol{\tau}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \sim \mathcal{MN}_{dp\times p}(\mathbf{O}, \text{BDiag}(\tau_1, \ldots, \tau_G), \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$ and $\tau_g \overset{\text{ind}}{\sim} \text{Gamma}(\frac{m_g+1}{2}, \frac{\lambda_g^2}{2})$ where $\text{BDiag}(\tau_1, \ldots, \tau_G)$ denotes a block-diagonal matrix with gth block to be $\tau_g \mathbf{I}_{m_g}$. Note that under the above hierarchical prior, conditionally on $(\tau_1, \ldots, \tau_G)$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$, the columns of $\mathbf{\Phi}$ are independent. If $m_g = 1 \ \forall\, g = 1, \ldots, dp$ we get the ordinary Bayesian lasso.

Under the specification given in (8), suppose we assume $\mathbf{U} = \text{Diag}(\tau_1, \ldots, \tau_{dp})$ and $1/\tau_i \overset{\text{ind}}{\sim} \text{Gamma}(\alpha_i, \lambda_i/2)$ then it can be shown that the prior density for $\mathbf{\Phi}$ given only $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ is proportional to

$$\prod_{i=1}^{dp} \left(\left\|\mathbf{\Phi}_{i\cdot}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\right\|_2^2 + \lambda_i\right)^{-(\alpha_i+\frac{1}{2})},$$

which corresponds to the multivariate t-distribution.

3.4. Estimation

For the hierarchical model given in (8), the posterior density of $\mathbf{\Phi}$ is intractable and quantities such as the posterior mean are not available in closed form. Hence, we develop a Markov chain Monte Carlo algorithm to generate values from the posterior density. It follows by straightforward calculations that

$$\begin{aligned}
\mathbf{\Phi} \mid \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathbf{U}, \mathcal{Y} &\sim \mathcal{MN}_{dp\times p}(\mathbf{\Phi}_{\text{PM}}, (\mathbf{X}'\mathbf{X}+\mathbf{U})^{-1}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \\
\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \mid \mathbf{U}, \mathcal{Y} &\sim \text{Inverse-Wishart}\left(\widehat{\mathbf{\Sigma}}_{res}, n-p-1\right) \\
\pi(\mathbf{U}|\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathcal{Y}) &\propto |\mathbf{U}|^{dp/2} \exp\left[-\frac{1}{2}\text{tr}\left\{\mathbf{\Phi}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{\Phi}'\mathbf{U}\right\}\right] \pi_{scl}(\mathbf{U}),
\end{aligned}$$

where $\widehat{\mathbf{\Sigma}}_{res} = \mathbf{Y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X}+\mathbf{U})^{-1}\mathbf{X}^T)\mathbf{Y}$. While the conditional posterior distribution of $\mathbf{\Phi}$ given $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$, $\mathbf{U}$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ given $\mathbf{U}$ are easy to simulate from (being Matrix-normal and Inverse-Wishart), the tractability of the conditional posterior density of $\mathbf{U}$ given $\mathbf{\Phi}$, $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ depends on the form of the prior $\pi_{scl}(\mathbf{U})$. We show below that for three standard choices of $\pi_{scl}(\mathbf{U})$ corresponding to the Wishart prior, the group lasso prior, and the multivariate t-prior $\pi(\mathbf{U}|\mathbf{\Phi})$ becomes a tractable density and easy to simulate from.

Case 1: Wishart Prior

For a $dp \times dp$ positive definite matrix $\mathbf{V}$, let $\mathbf{U} \sim \text{Wishart}_{dp}(\mathbf{V}, df = \nu + dp)$, that is, $\pi(\mathbf{U}) \propto |\mathbf{U}|^{\frac{\nu-1}{2}} \exp[-\frac{1}{2}\text{tr}\{\mathbf{V}^{-1}\mathbf{U}\}]$. In this case,

$$\pi(\mathbf{U}|\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathcal{Y}) \propto |\mathbf{U}|^{\frac{\nu+dp-1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left\{(\mathbf{\Phi}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{\Phi}' + \mathbf{V}^{-1})\mathbf{U}\right\}\right],$$

which is $\text{Wishart}_{dp}((\mathbf{\Phi}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{\Phi}' + \mathbf{V}^{-1})^{-1}, \ df = \nu + 2dp)$. Note that as long as we have $\nu > -(dp+1)$ the posterior of $\mathbf{U}$ given $\mathbf{\Phi}$, $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$, $\mathcal{Y}$ is proper.

Case 2: Bayesian Group Lasso

In this case as already discussed in Section 3.3, $\mathbf{U}^{-1}$ has a block diagonal form $\text{BDiag}(\tau_1, \ldots, \tau_G)$ and τ_g's are a priori independently distributed as Gamma with scale $(m_g+1)/2$ and rate $\lambda_g^2/2$. Hence, the conditional distribution of τ_g has the following form,

$$\frac{1}{\tau_g}\Big|\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\mathcal{Y} \overset{\text{ind}}{\sim} \text{Inverse} - \text{Gaussian}\left(\mu_g = \frac{\lambda_g}{\|\mathbf{\Phi}_{[g]}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\|_F}, \lambda_g^2\right).$$

Case 3: Multivariate t-Distribution

By taking $\mathbf{U}^{-1} = \text{Diag}(\tau_1, \ldots, \tau_{dp})$ and $1/\tau_i$ to be independently distributed as Gamma with shape α_i and rate $\lambda_i/2$, we have the multivariate t-distribution as the prior on $\mathbf{\Phi}$. In this case, the conditional distribution of τ_i has the following form:

$$\frac{1}{\tau_i}\Big|\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}, \mathcal{Y} \overset{\text{ind}}{\sim} \text{Gamma}\left(\alpha_i + \frac{dp}{2}, \frac{\|\mathbf{\Phi}_{i\cdot}'\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1/2}\|_2^2 + \lambda_i}{2}\right).$$

3.5. Assumptions for Posterior Consistency

We now introduce regularity conditions to establish posterior consistency under the hierarchical prior model.

Assumption B1. The VAR(d) model given in (1) is stable.

Assumption B2. $\frac{1+\mu_{\max}(\tilde{A})}{\mu_{\min}(\tilde{A})}$ is $\text{O}\left(\sqrt[5]{\frac{n}{p}}\right)$ as $n \to \infty$.

Assumption B3. $0 < \inf_{n\geq 1} \lambda_{\min}(\mathbf{C}_X) \leq \sup_{n\geq 1} \lambda_{\max}(\mathbf{C}_X) < \infty$ and $0 < \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}) = \text{O}(1)$.

Assumption B4. The singular values of the true parameter matrices $\{\mathbf{\Phi}_{0n}\}_{n\geq 1}$ are uniformly bounded. Equivalently, the eigenvalues of $\{\mathbf{\Phi}_{0n}'\mathbf{\Phi}_{0n}\}_{n\geq 1}$ are uniformly bounded.

Assumption B5. $p = \text{o}\left(\frac{n}{\log n}\right)$.

Assumption B6. There exists (fixed and not-depending on n) $\alpha > 0$ such that

$\liminf_{n\to\infty} \pi_{\text{scl},n}(\lambda_{\max}(\mathbf{U}) > \alpha) > 0$ and for every $\beta > 0$ we have $\lim_{n\to\infty} \pi_{\text{scl},n}(\lambda_{\max}(\mathbf{U}) > \beta n) = 0$.

We now discuss these assumptions and compare them to the assumptions for the nonhierarchical prior model.

- Assumptions B1 and B2 are identical to A1 and A2, while B3 is fairly similar to A3.
- One key difference is the permissible scaling of p as a function of the sample size n in Assumption B5, which is slightly more stringent than the permissible scaling for the nonhierarchical matrix normal prior in Assumption A5.
- Note that Assumption B6 is a mild one. For example, a sufficient condition for this assumption to be satisfied is that $\limsup_{n\to\infty} \max_{1\leq i\leq p} \mathbb{E}_{\pi_{\text{scl},n}}[\mathbf{U}_{ii}^{\delta}] < \infty$ for some $\delta > 0$ and $\liminf_{n\to\infty} \pi_{\text{scl},n}(\mathbf{U}_{11} > \alpha) > 0$. It can be easily checked that this condition, and hence Assumption B6, is satisfied in the case of Wishart, Inverse-Wishart, Bayesian group lasso, multivariate t-distribution, Horseshoe (Carvalho, Polson, and Scott 2010), Strawderman-Berger and generalized double Pareto (Armagan et al. 2013) priors as long as the prior parameters do not depend on n.

- In the nonhierarchical prior case, assumptions regarding $\mathbf{\Phi}_0$ and $\mathbf{U}$ (nonrandom) are simultaneously and exclusively provided in Assumption A4 through the conditions $\left\|\mathbf{\Phi}_0^T\mathbf{U}\mathbf{\Phi}_0\right\| = o(n)$ and $\|\mathbf{U}\mathbf{\Phi}_0\| = o(n)$. For the hierarchical prior case, for clarity of exposition, we provide the assumptions regarding $\mathbf{\Phi}_0$ in Assumption B4, and those regarding the distribution of $\mathbf{U}$ (random) in Assumption B6. Combining these two assumptions, it can be easily shown that a priori $\left\|\mathbf{\Phi}_0^T\mathbf{U}\mathbf{\Phi}_0\right\|$ and $\|\mathbf{U}\mathbf{\Phi}_0\|$ converge to zero in $\Pi_{\text{scl},n}$-probability as $n \to \infty$. In that sense, the assumptions on $(\mathbf{\Phi}_0, \mathbf{U})$ in the hierarchical model are stronger than in the nonhierarchical model case.

With these assumptions in hand, we state our key consistency result, whose proof is delegated to Appendix C.6 of the supplementary material.

Theorem 2 (Posterior consistency for hierarchical prior). For any centered VAR(d) model with the hierarchical prior (8) on the transition matrix satisfying Assumptions B1–B6, the posterior consistency of the transition matrix can be achieved, that is, for every fixed $c\varepsilon > 0$

$$\mathbb{E}_0\left[\Pi_n(\|\mathbf{\Phi} - \mathbf{\Phi}_0\| > \varepsilon | \mathcal{Y} = (X^0, \ldots, X^n))\right] \to 0 \quad \text{as } n \to \infty$$

where $\mathbf{\Phi}_0$ is the true parameter matrix under the model (2).

4. Performance Evaluation

To illustrate the performance of our Bayesian modeling framework for VAR processes, we design three sets of numerical experiments involving: (a) Small VAR ($p = 10$), (b) Medium VAR ($p = 100$), and (c) Large VAR ($p = 500$) models, each with two lags—(i) $d = 1$ and (ii) $d = 2$.

In each setting, we use transition matrices $\mathbf{A}_i$'s with 10%–30% nonzero entries that are generated from $U(0, 2) \cup U(-2, 0)$ selected at random and rescaled to ensure that the process is stable with SNR $= 2$. For small VAR models, we generate $n = 40, 80, 120$ time points, for medium VAR models, $n = 400, 800, 1200$, while for large VAR models we use $n = 2000, 4000, 6000$. The hyper-parameters for the prior distributions are selected using the deviance information criterion (DIC). Note that $\text{DIC} = 2\bar{D} - D(\bar{\mathbf{\Phi}}, \bar{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}})$, where $D(\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) := -2\log L(\mathcal{Y}|\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) = n\log|\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}| + \text{tr}\{\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{\Phi}'\mathbf{X}'\mathbf{X}\mathbf{\Phi} - 2\mathbf{\Phi}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{Y})\}$, $\bar{D}$ is the posterior expectation of $D(\mathbf{\Phi}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$, and $\bar{\mathbf{\Phi}}$ and $\bar{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}$ are the posterior expectation of $\mathbf{\Phi}$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$, respectively.

4.1. Nonhierarchical Prior

We generate two different error processes using $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}_p$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\left((\rho^{|i-j|})\right)_{ij}$ (Toeplitz form). For each of the small, medium, and large VAR models, $\mathbf{U}$ is taken to be a diagonal matrix, cI_{dp} where c is chosen according to the minimum DIC value over the interval $[0, 10]$. In Table 1, for both the posterior mean (PM) and least-square estimator (LS), we report their relative estimation error ($\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_2/\|\mathbf{\Phi}_0\|_2$) and the standard error of $\|\hat{\mathbf{\Phi}}\|_2$ within parenthesis averaged over $10 \times p$ replicates for small and medium VAR and 100 replications for large VAR ($p = 500$). Since the true parameter matrix $\mathbf{\Phi}_0$ is sparse, we identify entries whose 95% posterior credible intervals contain zero, and set them to zero in both parameter matrix estimates (PM and LS).

First, we assume the true error covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ is diagonal, that is, $\sigma^2\mathbf{I}_p$. Here % denotes percentage of nonzero entries in $\mathbf{\Phi}$ and d represents lag length of the underlying VAR process. Recall that the sample sizes used for small VAR models are $n_1 = 40$, $n_2 = 80$, $n_3 = 120$, for medium VAR ones $n_1 = 400$, $n_2 = 800$, $n_3 = 1200$, and for large VAR ones $n_1 = 2000$, $n_2 = 4000$, $n_3 = 6000$.

It can be seen that the relative estimation error decreases with an increase in the number of time points (sample size) n for both lags $d = 1, 2$; further, its values are significantly larger in medium and large size VAR models than in small VAR ones. Moreover, the estimation error for lag 1 is uniformly smaller than that for lag 2, and the same holds true for their respective standard errors. Regarding the percentage of nonzero entries in

Table 1. Relative error in VAR ($d = 1, 2$) with $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}_p$ where % denotes percentage of nonzero entries in $\mathbf{\Phi}_0$.

		Lag $d = 1$						Lag $d = 2$					
		n_1		n_2		n_3		n_1		n_2		n_3	
	%	LS	PM	LS	PM	LS	PM	LS	PM	LS	PM	LS	PM
Small VAR	10	0.93	0.83	0.82	0.70	0.70	0.55	1.26	1.09	1.11	0.96	1.00	0.76
		(0.21)	(0.12)	(0.12)	(0.08)	(0.07)	(0.05)	(0.24)	(0.17)	(0.17)	(0.07)	(0.10)	(0.08)
	20	1.06	1.00	1.00	0.87	0.83	0.71	1.39	1.22	1.28	1.07	1.14	0.94
		(0.34)	(0.24)	(0.26)	(0.16)	(0.16)	(0.11)	(0.35)	(0.30)	(0.27)	(0.21)	(0.18)	(0.15)
	30	1.23	1.15	1.12	0.97	0.99	0.81	1.57	1.40	1.45	1.28	1.30	1.15
		(0.45)	(0.37)	(0.37)	(0.27)	(0.27)	(0.19)	(0.47)	(0.40)	(0.37)	(0.32)	(0.34)	(0.22)
Medium VAR	10	1.81	1.64	1.69	1.53	1.56	1.40	2.45	2.12	2.31	2.01	2.24	1.85
		(0.40)	(0.21)	(0.31)	(0.12)	(0.23)	(0.07)	(0.46)	(0.32)	(0.37)	(0.24)	(0.31)	(0.14)
	20	1.95	1.81	1.85	1.70	1.74	1.56	2.60	2.26	2.50	2.15	2.43	2.01
		(0.50)	(0.32)	(0.42)	(0.25)	(0.35)	(0.17)	(0.57)	(0.42)	(0.51)	(0.33)	(0.40)	(0.29)
	30	2.10	1.94	1.98	1.84	1.87	1.74	2.74	2.46	2.62	2.31	2.52	2.20
		(0.64)	(0.43)	(0.52)	(0.38)	(0.45)	(0.27)	(0.69)	(0.54)	(0.60)	(0.46)	(0.56)	(0.39)
Large VAR	10	2.70	2.47	2.55	2.34	2.46	2.22	3.66	3.18	3.50	3.04	3.38	2.92
		(0.58)	(0.30)	(0.50)	(0.23)	(0.40)	(0.16)	(0.67)	(0.45)	(0.57)	(0.35)	(0.49)	(0.28)
	20	2.84	2.62	2.75	2.52	2.60	2.35	3.81	3.33	3.69	3.23	3.54	3.05
		(0.70)	(0.42)	(0.60)	(0.34)	(0.54)	(0.25)	(0.78)	(0.56)	(0.70)	(0.50)	(0.61)	(0.40)
	30	3.03	2.78	2.90	2.65	2.71	2.49	3.96	3.49	3.89	3.32	3.78	3.25
		(0.82)	(0.54)	(0.70)	(0.44)	(0.63)	(0.35)	(0.88)	(0.69)	(0.84)	(0.63)	(0.76)	(0.51)

Table 2. Relative error in VAR ($d = 1, 2$) with $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ = Toeplitz ($\rho = 0.80$) where % denotes percentage of nonzero entries in $\mathbf{\Phi}_0$.

		Lag $d = 1$						Lag $d = 2$					
		n_1		n_2		n_3		n_1		n_2		n_3	
	%	LS	PM	LS	PM	LS	PM	LS	PM	LS	PM	LS	PM
Small VAR	10	1.03	0.95	0.95	0.82	0.82	0.64	1.45	1.24	1.31	1.08	1.19	0.98
		(0.27)	(0.18)	(0.20)	(0.09)	(0.10)	(0.05)	(0.37)	(0.27)	(0.30)	(0.19)	(0.18)	(0.09)
	20	1.20	1.10	1.06	0.94	0.98	0.83	1.60	1.41	1.49	1.30	1.35	1.17
		(0.40)	(0.30)	(0.29)	(0.21)	(0.25)	(0.17)	(0.48)	(0.37)	(0.40)	(0.29)	(0.31	(0.19)
	30	1.36	1.27	1.22	1.11	1.12	0.99	1.76	1.61	1.66	1.45	1.50	1.26
		(0.50)	(0.41)	(0.44)	(0.32)	(0.32)	(0.25)	(0.60)	(0.51)	(0.50)	(0.40)	(0.46)	(0.30)
Medium VAR	10	2.02	1.87	1.94	1.72	1.75	1.61	2.80	2.46	2.70	2.35	2.53	2.21
		(0.52)	(0.34)	(0.41)	(0.27)	(0.37)	(0.16)	(0.69)	(0.50)	(0.62)	(0.43)	(0.54)	(0.33)
	20	2.20	2.01	2.10	1.87	1.92	1.80	2.95	2.62	2.86	2.52	2.74	2.34
		(0.63)	(0.45)	(0.55)	(0.38)	(0.45)	(0.33)	(0.81)	(0.62)	(0.71)	(0.52)	(0.65)	(0.42)
	30	2.36	2.20	2.19	2.06	2.06	1.91	3.12	2.78	2.99	2.67	2.91	2.50
		(0.73)	(0.56)	(0.67)	(0.48)	(0.56)	(0.41)	(0.94)	(0.72)	(0.83)	(0.67)	(0.77)	(0.57)
Large VAR	10	3.03	2.80	2.88	2.63	2.83	2.55	4.19	3.67	4.09	3.55	3.93	3.44
		(0.75)	(0.49)	(0.67)	(0.41)	(0.60)	(0.33)	(1.03)	(0.73)	(0.97)	(0.62)	(0.87)	(0.58)
	20	3.17	2.94	3.07	2.79	2.89	2.69	4.34	3.85	4.25	3.70	4.11	3.594
		(0.86)	(0.60)	(0.79)	(0.51)	(0.69)	(0.45)	(1.14)	(0.84)	(1.06)	(0.76)	(1.00)	(0.66)
	30	3.33	3.10	3.27	3.00	3.08	2.80	4.53	4.00	4.37	3.89	4.26	3.70
		(0.98)	(0.73)	(0.90)	(0.63)	(0.82)	(0.56)	(1.26)	(0.95)	(1.17)	(0.87)	(1.11)	(0.79)

the true transition matrices, the results show that for fixed n and p, the more true nonzero entries in $\mathbf{A}_1, \mathbf{A}_2$, the less accurate the posterior mean and the LS estimator are, while their variability as indicated by their standard errors also follows the same pattern. However, the posterior mean clearly outperforms the LS estimates, especially in settings with large p. This is to a large extent because the true transition matrices $\mathbf{A}_1, \mathbf{A}_2$ exhibit weaker signal as p or the number of nonzero edges increases (this is to ensure stability of the underlying VAR model) and due to our choice of $\mathbf{U} = c\,\mathbf{I}_{dp}$ the posterior mean is the ridge regression estimator which applies shrinkage on the coefficients.

Next, we introduce correlation in the error components by specifying $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ to be of Toeplitz form. As discussed in sec. 3.2.1 of Lütkepohl (2007) the generalized least-square estimate in this multivariate regression set-up is the same as the ordinary one, that is, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, a result due to Zellner (1962). In Table 2, we compare the performance of least squares and posterior (ridge) estimates with noise covariance $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ = Toeplitz ($\rho = 0.8$).

In this setting, the relative estimation error of both the least squares and ridge estimators increases compared to that with an uncorrelated error structure given in Table 1; in particular, the performance of the LS estimator deteriorates even further. However, with an increase in sample size, the accuracy of both estimates significantly improves. Further, as gleaned from the entries of the table corresponding to lag 2, the relative error exhibits a further increase, a pattern consistent with the results in Table 1. This is quite expected as we not only have Toeplitz type error covariance structure, but also the total number of unknown parameters has increased by p^2.

Finally, we study the support recovery under both error processes. In Table 3, we provide the percentage of true positives recovered by using 95% posterior credible intervals based on the same sample sizes n_1, n_2, and n_3 as used previously.

Table 3 indicates that support recovery is not sensitive to the sample size, or to the lag; however, it deteriorates for all VAR models and error covariance settings, as the density of nonzero entries increases and exhibits a small increase with model dimension.

4.2. Hierarchical Priors

As discussed in Section 3.4, three types of hierarchical priors (Wishart, group-lasso, and multivariate t) are studied. Analogously to the nonhierarchical prior case, the performance of the LS estimator is not at all satisfactory in this setup as well. Thus, we only compare the relative accuracy of the three prior choices in this setting. We select $\mathbf{V} = cI_{dp}$ and $df = \nu = dp$ for the Wishart prior, $\lambda_i = \lambda$ for all $1 \le i \le dp$ for the

Table 3. Percentage of true positive nonzero entries recovered in $\mathbf{\Phi}$.

		Lag $d = 1$						Lag $d = 2$					
		$\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}_p$			$\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ = Toep			$\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}_p$			$\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ = Toep		
	%	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
Small VAR	10	85.0	85.0	82.0	85.0	85.0	84.0	84.7	84.5	81.6	84.6	84.6	83.6
	20	80.0	80.0	81.0	80.0	78.0	80.0	79.7	79.6	80.5	79.5	77.6	79.8
	30	77.0	75.0	77.0	77.0	73.0	73.0	76.6	74.6	76.5	76.7	72.6	72.6
Medium VAR	10	89.3	90.0	89.3	89.3	89.8	88.3	88.8	89.5	89.0	88.9	89.4	87.9
	20	85.3	85.5	85.5	85.3	84.5	85.0	84.9	85.3	85.3	84.8	84.1	84.8
	30	81.0	80.8	81.3	81.0	79.8	78.8	80.8	80.5	81.0	80.6	79.3	78.3
Large VAR	10	92.0	92.1	92.2	92.0	92.1	91.8	91.8	91.6	91.8	91.7	91.7	91.4
	20	88.1	88.1	88.1	88.1	87.9	87.9	87.7	87.8	87.7	87.7	87.5	87.5
	30	83.1	83.3	83.4	83.1	82.8	82.9	82.7	82.8	82.9	82.7	82.4	82.6

Table 4. Relative error in VAR ($d = 1, 2$) with $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}_p$, where % denotes percentage of nonzero entries in $\mathbf{\Phi}_0$.

		Wishart			Group lasso			Multivariate t		
Lag $d = 1$	%	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
Small VAR	10	0.84	0.75	0.65	0.84	0.75	0.64	0.85	0.74	0.58
	20	0.99	0.91	0.73	1.00	0.89	0.79	1.00	0.90	0.76
	30	1.14	1.09	0.92	1.16	1.02	0.94	1.15	1.04	0.93
Medium VAR	10	1.63	1.53	1.42	1.62	1.54	1.43	1.63	1.53	1.39
	20	1.78	1.71	1.53	1.76	1.69	1.56	1.79	1.66	1.55
	30	1.93	1.83	1.70	1.95	1.86	1.72	1.95	1.82	1.69
Large VAR	10	2.39	2.26	2.16	2.41	2.31	2.22	2.42	2.34	2.22
	20	2.59	2.42	2.31	2.58	2.47	2.39	2.59	2.46	2.32
	30	2.77	2.61	2.52	2.70	2.60	2.53	2.74	2.63	2.54
Lag $d = 2$										
Small VAR	10	0.88	0.74	0.60	0.87	0.80	0.71	0.89	0.78	0.63
	20	1.05	0.91	0.83	1.04	0.94	0.85	1.08	0.94	0.86
	30	1.25	1.13	0.97	1.23	1.10	0.99	1.27	1.13	1.03
Medium VAR	10	1.71	1.58	1.46	1.71	1.58	1.53	1.70	1.60	1.50
	20	1.85	1.72	1.60	1.85	1.76	1.71	1.89	1.82	1.66
	30	2.05	1.96	1.77	2.01	1.95	1.82	2.04	1.97	1.83
Large VAR	10	2.52	2.37	2.26	2.51	2.41	2.29	2.53	2.44	2.25
	20	2.69	2.58	2.47	2.67	2.59	2.47	2.69	2.59	2.51
	30	2.88	2.75	2.60	2.86	2.75	2.62	2.87	2.74	2.64

group-lasso prior and $\alpha_i = 1,\ \lambda_i = \lambda$ for all $1 \le i \le dp$ for the multivariate t prior. The hyperparameters c and λ are chosen using DIC. In Table 4, we report the relative estimation error ($\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_2 / \|\mathbf{\Phi}_0\|_2$, $d = 1, 2$) of the three hierarchical estimators when the error process covariance is set to $\sigma^2 \mathbf{I}_p$ and d represents lag length in the underlying VAR model.

Next in Table 5, we present relative estimation errors with the same three hierarchical priors when $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} =$ Toeplitz ($\rho = 0.8$).

All of our hierarchical estimates outperform the ridge estimator (Tables 1 and 2) across all settings considered. This is again expected, since the $\mathbf{A}_i$'s have sparse structure by construction and the group lasso prior favors sparsity. However, the above results are not conclusive whether the group lasso estimate exhibits better accuracy than the Wishart or multivariate t estimates.

To gain some insight into this issue, we use a VAR(1) model with $p = 9$ and transition matrix $\mathbf{A}_1$ in which the columns form three groups each containing three columns. The sparsity increases as we move from group 1 to group 3. Finally, we rescale the coefficient matrix so that the corresponding VAR process is stable with SNR $= 2$. The structure of the resulting $\mathbf{A}_1$ transition matrix is depicted next, where * indicates nonzero entries.

$$\mathbf{A}_1 = \begin{pmatrix} * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & 0 & * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}_{9\times 9}$$

Table 5. Relative error in VAR ($d = 1, 2$) with $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} =$ Toeplitz ($\rho = 0.80$), where % denotes percentage of nonzero entries in $\mathbf{\Phi}_0$.

		Wishart			Group lasso			Multivariate t		
Lag $d = 1$	%	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
Small VAR	10	0.90	0.76	0.65	0.90	0.83	0.69	0.89	0.79	0.70
	20	1.06	0.94	0.77	1.03	0.92	0.82	1.07	0.96	0.81
	30	1.23	1.13	0.99	1.21	1.09	0.99	1.23	1.14	0.97
Medium VAR	10	1.75	1.61	1.50	1.73	1.63	1.53	1.74	1.61	1.56
	20	1.92	1.80	1.67	1.92	1.78	1.73	1.91	1.79	1.72
	30	2.08	1.98	1.82	2.02	1.98	1.88	2.07	1.94	1.84
Large VAR	10	2.58	2.49	2.33	2.60	2.50	2.38	2.60	2.49	2.33
	20	2.76	2.65	2.51	2.75	2.65	2.54	2.77	2.62	2.57
	30	2.93	2.81	2.71	2.90	2.81	2.74	2.93	2.82	2.74
Lag $d = 2$										
Small VAR	10	1.16	1.02	0.96	1.15	1.06	0.98	1.18	1.06	0.89
	20	1.31	1.24	1.12	1.31	1.22	1.11	1.34	1.24	1.10
	30	1.49	1.41	1.23	1.49	1.41	1.27	1.49	1.40	1.26
Medium VAR	10	2.25	2.13	2.04	2.25	2.17	2.05	2.27	2.16	2.01
	20	2.43	2.30	2.16	2.41	2.32	2.24	2.45	2.32	2.26
	30	2.59	2.47	2.33	2.59	2.49	2.38	2.59	2.48	2.36
Large VAR	10	3.37	3.24	3.10	3.36	3.25	3.18	3.36	3.24	3.11
	20	3.53	3.40	3.29	3.51	3.43	3.31	3.54	3.46	3.36
	30	3.72	3.62	3.41	3.65	3.57	3.49	3.74	3.59	3.48

Table 6. Relative error.

Estimator	$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}_p$	$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} =$ Toeplitz
LS	0.604	1.384
Non H	0.583	0.614
W	0.462	0.544
Mult. t	0.430	0.414
GL	0.321	0.362

We generate $n = 100$ observations from the above VAR(1) model using two white noise variances (1) $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}_p$ and (2) $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} =$ Toeplitz ($\rho = 0.80$) and report the relative estimation error ($\|\hat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 / \|\mathbf{A}_1\|_2$) of five different estimates—least squares (LS), posterior mean for nonhierarchical normal prior (Non H), hierarchical Wishart (W), group lasso (GL), and multivariate t prior (Mult. t), in Table 6.

The results show that the group lasso estimator exhibits the best performance, followed by the multivariate t one, whereas the LS estimator is the least accurate. The result is consistent with the structure of the underlying transition matrix, since the group lasso prior can capitalize on it.

In Appendix A.1 of the supplementary document, we illustrate the posterior estimate of a VAR(1) model transition matrix $\mathbf{A}_1$, 95% credible intervals and estimated posterior densities of several entries of $\mathbf{A}_1$. We also look into the performance of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ when the true error covariance is Toeplitz($\rho = 0.8$). The relative error of $\hat{\boldsymbol{\Phi}}$ under all the four different priors using a new noise covariance matrix which is generated from a Wishart distribution with degrees of freedom $\nu = p$ and scale matrix $\mathbf{I}_p$ is also given.

5. Application to Macroeconomic Data

We use the proposed Bayesian framework to understand the lead-lag relationships in the FRED-MD dataset containing $p = 137$ key macroeconomic variables for the period January 1973 to December 2014. VAR modeling for this task was strongly advocated by Sims (1980) and since then has become a standard tool for it, although usually the focus is on small models involving few macroeconomic indicators (e.g., consumer price index, employment index, and the federal funds rate). However, recent work has advocated for larger VAR models (see Bernanke, Boivin, and Eliasz 2005; Bańbura, Giannone, and Reichlin 2010, and references therein), to improve forecastability and also avoid the presence of hard to interpret or even contradictory to economic theory relationships, because of not including an adequate number of variables for properly modeling the economic phenomenon under consideration. Before centering the data and estimating $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ as discussed earlier in Section 3.4, we ensure stationarity by processing the variables according to the recommendations in Stock and Watson (2012). The specific transformations used for each time series are given in the supplementary documents. Analogously to Bańbura, Giannone, and Reichlin (2010), we consider the following three different size VAR models:

- SMALL: This model contains $p = 4$ key variables—CPI, number of employees nonfarm (PAYEMS), Federal Funds Rate (FEDFUNDS), and Unemployment Rate (UNRATE).
- MEDIUM: In addition to the four variables in the SMALL VAR model, this one contains an additional 16 variables (total $p = 20$) listed next—Reserves Of Depository Institutions (NONBORRES), Total Reserves of Depository Institutions (TOTRESNS), M2 (M2REAL), Real Personal Income (RPI), Real personal consumption expenditures (DPCERA), IP Index (INDPRO), Capacity Utilization: Manufacturing (CAPT), Housing Starts: Total New Privately Owned (HOUST), Avg Hourly Earnings : Goods-Producing (CES), M1 (M1), S & P's Common Stock Price Index: Composite (S.P.), 10-Year Treasury Rate (GS10), Personal Cons. Expend.: Chain Index (PCEPI), Foreign Exchange Rate (EXS), Crude Oil, spliced WTI, and Cushing (OIL), and Retail and Food Services Sales (RETAILx).
- LARGE: This specification has all $p = 134$ macroeconomic indicators (three were excluded from further analysis due to the presence of a large number of missing values).

Based on initial exploratory work, we choose lag $d = 6$ according to the Bayesian information criterion (BIC) and the following distributions were used for prior specification to obtain the estimated parameter matrix $\boldsymbol{\Phi}$: (i) nonhierarchical normal (Non H), (ii) hierarchical Wishart (W), (iii) group lasso (GL), and (iv) multivariate t prior (Mult. t). Since with an increase in the lag length d the number of parameters increases linearly, we suggest using BIC over the Akaike information criterion (AIC). For the nonhierarchical prior, we use $\mathbf{U} = \text{BDiag}(\lambda_1, \ldots, \lambda_d)$, while for the hierarchical Wishart, group-lasso, and multivariate t priors on $\boldsymbol{\Phi}$, we use $\mathbf{V} = c_1 \mathbf{I}_{dp}$ and $\alpha_i = \alpha$ for all $1 \le i \le dp$. The values of c_1 and λ are chosen using the deviance information criterion (DIC) which, as explained previously, is a hierarchical Bayesian modeling generalization of BIC. The respective posterior means were compared to the least-square (LS) estimate. For each of the estimates $\hat{\boldsymbol{\Phi}}$, the residual norm ratio ($\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Phi}}\|_F / \|\mathbf{Y}\|_F$) which measures the in-sample fit, is reported in Table 7.

Note that since the LS estimator is obtained by minimizing $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi}\|_F$, it will always result in minimum relative residual norm as observed in Table 7, that is, the LS estimator is always the best one in terms of in-sample prediction accuracy.

Next, we investigate the four different Bayesian estimates based on their out-of-sample prediction performance with respect to the benchmark prior, analogously to the evaluation strategies discussed in Bańbura, Giannone, and Reichlin (2010) and Stock and Watson (2012). We consider the following two benchmark priors:

1. Prior information is imposed exactly by setting $\mathbf{U}^{-1} = \mathbf{O}$ matrix (the zero matrix) and it corresponds to $\lambda = 0$ in the Minnesota prior. Bańbura, Giannone, and Reichlin (2010) used this specification as the benchmark prior, in which case the corresponding benchmark model becomes a random walk with drift, that is, $X^t = \boldsymbol{\alpha} + X^{t-1} + \boldsymbol{\varepsilon}^t$.

Table 7. In-sample prediction error.

	SMALL ($p = 4$)	MEDIUM ($p = 20$)	LARGE ($p = 134$)
LS	0.844	0.863	0.673
Non H	0.852	0.864	0.674
W	0.845	0.863	0.675
Mult. t	0.877	0.885	0.685
GL	0.847	0.873	0.674

Table 8. Out-of-sample relative prediction error.

		Uniform prior			Random walk		
		SMALL $p = 4$	MEDIUM $p = 20$	LARGE $p = 134$	SMALL $p = 4$	MEDIUM $p = 20$	LARGE $p = 134$
Nonhierarchical	$h = 1$	0.88	0.72	0.62	0.49	0.40	0.33
	$h = 6$	0.90	0.78	0.62	0.43	0.42	0.37
	$h = 12$	0.95	0.84	0.74	0.43	0.41	0.37
Wishart	$h = 1$	0.88	0.81	0.68	0.49	0.45	0.32
	$h = 6$	0.86	0.86	0.68	0.42	0.40	0.36
	$h = 12$	0.93	0.92	0.71	0.43	0.41	0.38
Group Lasso	$h = 1$	0.90	0.86	0.60	0.51	0.45	0.30
	$h = 6$	0.88	0.88	0.63	0.46	0.42	0.34
	$h = 12$	0.93	0.92	0.67	0.47	0.45	0.37
Mult. t	$h = 1$	0.91	0.89	0.71	0.50	0.46	0.31
	$h = 6$	0.87	0.93	0.77	0.49	0.44	0.34
	$h = 12$	0.92	0.94	0.81	0.45	0.41	0.38

2. A uniform prior on $\mathbf{\Phi}$ by setting $\mathbf{U} = \mathbf{O}$ which corresponds to $\lambda = \infty$ in the Minnesota prior. The posterior mean coincides with the least-squared estimate (LS).

Let $\hat{X}^{t+h}$ be the h-step ahead predicted value for X^{t+h} based on our posited Bayesian model and using the data up to time t. The corresponding forecast under the benchmark prior is $\hat{X}_{\mathbf{O}}^{t+h}$. The mean squared forecast error relative to the benchmark (RMSFE) is defined to be $\frac{\sum_{t=T_0}^{T_1} \|X^{t+h} - \hat{X}^{t+h}\|_2}{\sum_{t=T_0}^{T_1} \|X^{t+h} - \hat{X}_{\mathbf{O}}^{t+h}\|_2}$. Table 8 gives the RMFSE results for three different choices of forecasting horizons, $h = 1, 6, 12$, for the two benchmark priors considered, over the period $T_0 =$ January 1978 to $T_1 =$ December 2006. An RMSFE value smaller than 1 implies the VAR model with the corresponding prior outperforms that with the naive/benchmark prior.

It can easily be seen that all four Bayesian methods not only outperform the LS estimate (uniform prior on $\mathbf{\Phi}$), but also exhibit substantially smaller relative error compared to the random walk with drift process (point-mass prior on $\mathbf{\Phi}$). Further, increasing the number of predictor variables improves forecasting performance, a point argued forcefully in favor of large VAR models by Sims Sims (1980). On the other hand, forecasting performance deteriorates for larger values of h, an expected result. Nevertheless, even for $h = 12$ (1 year ahead), the results are still very satisfactory. Further, for the SMALL and MEDIUM VAR models, the nonhierarchical normal and hierarchical Wishart priors result in better prediction, whereas for the LARGE VAR model the group lasso prior outperforms other forecasts.

Next, in Table 9, we examine closely the out-of-sample prediction performance of the following three macroeconomic variables—CPI, PAYEMS, and FEDFUND under the hierarchical Wishart prior.

Note that Bańbura, Giannone, and Reichlin (2010) only considered a random walk process as the naive prior. From Table 9, it can be seen that although for CPI and PAYEMS the LS estimate performs better than the Bayesian estimates in SMALL and MEDIUM VARs, overall the Wishart prior has better forecasting accuracy than both of the benchmark priors. As previously observed, adding information (i.e., including more variables) improves the accuracy of forecasts for all three variables. The fourth column (LARGE BGR) provides the

Table 9. Out-of-sample relative prediction error for CPI, PAYEMS, and FedFund for the three VAR model specifications considered. The column LARGE BVAR corresponds to the entries of Table III in Bańbura, Giannone, and Reichlin (2010) for a Bayesian VAR model with a normal-inverted Wishart prior distribution and $d = 13$ lags, based on the same set of variables, but covering the period 1971–2003.

Uniform prior		SMALL $p = 4$	MEDIUM $p = 20$	LARGE $p = 134$	LARGE BVAR $p = 134$
$h = 1$	CPI	1.05	0.91	0.44	—
	PAYEMS	1.21	1.04	0.91	—
	FFUND	0.78	0.75	0.68	—
$h = 6$	CPI	1.03	0.97	0.38	—
	PAYEMS	1.08	1.06	0.48	—
	FFUND	0.92	0.82	0.68	—
$h = 12$	CPI	0.98	0.96	0.42	—
	PAYEMS	0.93	0.91	0.73	—
	FFUND	0.92	0.88	0.72	—
Random walk					
$h = 1$	CPI	0.43	0.41	0.34	0.50
	PAYEMS	0.45	0.43	0.39	0.46
	FFUND	0.50	0.45	0.36	0.75
$h = 6$	CPI	0.38	0.34	0.28	0.40
	PAYEMS	0.53	0.48	0.39	0.50
	FFUND	0.41	0.37	0.36	1.29
$h = 12$	CPI	0.51	0.45	0.42	0.44
	PAYEMS	0.51	0.88	0.73	0.78
	FFUND	0.33	0.31	0.28	1.93

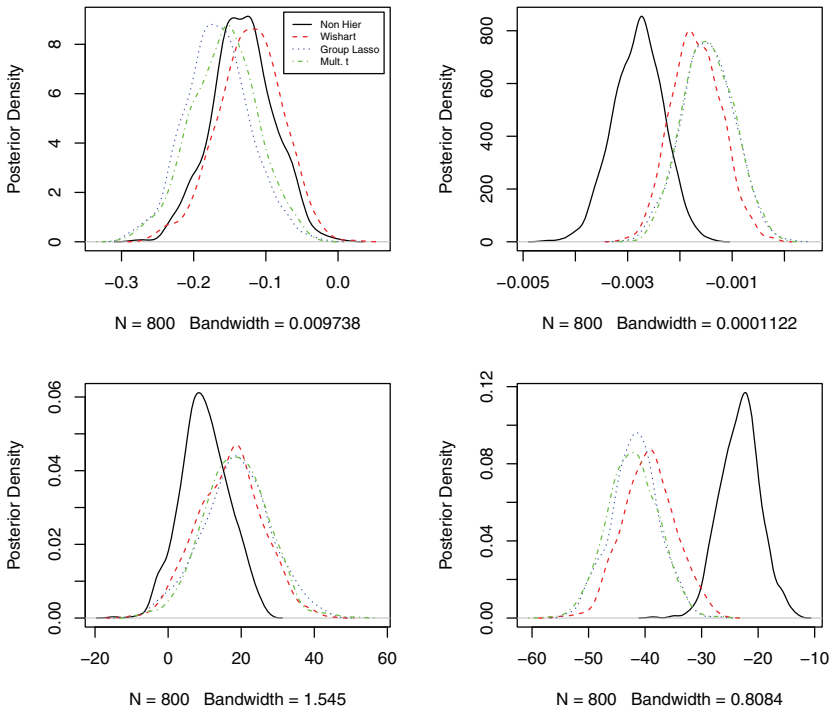


Figure 1. Posterior densities of entries (11), (24), (32), and (42) in $\mathbf{A}_1$ under four different priors.

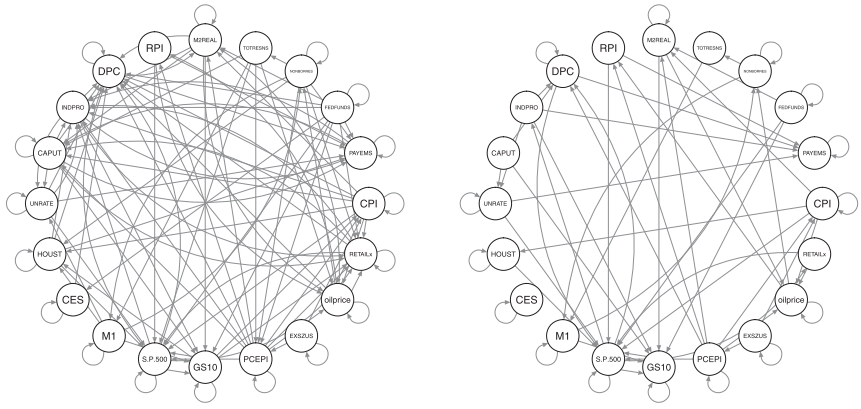
numbers reported in Table III in Bańbura, Giannone, and Reichlin (2010), where a Bayesian VAR model on the same 134 variables, with $d = 13$ lags was estimated using a normal-inverted Wishart distribution that leads to a ridge type posterior mean estimate for the parameters in $\mathbf{\Phi}$ and based on data covering the period 1971–2003. Although the results are not directly comparable to those obtained by our methodology, they nevertheless provide a certain degree of calibration. It can be seen that our model is more parsimonious using only $d = 6$ lags and provides better forecasting performance for all three variables at all forecasting horizons.

Next, we examine in more detail the estimated transition matrix $\mathbf{A}_1$ for the SMALL VAR model ($p = 4$) and the nonhierarchical normal and group lasso priors. Estimated posterior densities of the bold marked entries are shown in Figure 1. It is worth noting that the nonhierarchical prior centers around a different value and exhibits a less smooth behavior than the hierarchical one. This smoothness should be expected given the specification of the latter.

$$\hat{\mathbf{A}}_1^{\text{NonH}} = \begin{array}{c} \\ \text{CPI} \\ \text{PAYEMS} \\ \text{FEDFund} \\ \text{UnRate} \end{array} \begin{array}{c} \begin{array}{cccc} \text{CPI} & \text{PAYEMS} & \text{FEDFund} & \text{UnRate} \end{array} \\ \begin{pmatrix} \mathbf{-0.133} & -0.001 & 0.001 & 0.001 \\ -0.016 & 0.311 & 0.001 & \mathbf{-0.002} \\ -1.000 & \mathbf{10.200} & 0.498 & -0.185 \\ 1.217 & \mathbf{-23.300} & -0.035 & -0.105 \end{pmatrix} \end{array}$$

$$\hat{\mathbf{A}}_1^{GL} = \begin{array}{c} \\ \text{CPI} \\ \text{PAYEMS} \\ \text{FEDFund} \\ \text{UnRate} \end{array} \begin{array}{c} \begin{array}{cccc} \text{CPI} & \text{PAYEMS} & \text{FEDFund} & \text{UnRate} \end{array} \\ \begin{pmatrix} \mathbf{-0.167} & -0.005 & 0.001 & 0.001 \\ -0.021 & 0.560 & 0.001 & \mathbf{-0.001} \\ -1.614 & \mathbf{18.607} & 0.486 & -0.134 \\ 1.609 & \mathbf{-41.818} & -0.015 & -0.107 \end{pmatrix} \end{array}$$

Further, we present the 95% posterior credible intervals (PCI) of $\mathbf{A}_1$ under the above two priors.



(a) Least squares estimate (LS) (b) Nonhierarchial normal estimate (NonH)

Figure 2. Network representation of the transition matrix ($\mathbf{A}_1$).

Nonhierarchical:

$$= \begin{pmatrix} \mathbf{(-0.19, -0.07)} & (-0.08, 0.07) & (+0, +0) & (-0, +0) \\ (-0.05, 0.01) & (0.27, 0.35) & (+0, +0) & \mathbf{(-0, -0)} \\ (-7, 5.25) & \mathbf{(1.86, 19)} & (0.44, 0.55) & (-0.3, -0.06) \\ (-1.98, 4.45) & \mathbf{(-27.8, -18.7)} & (-0.06, -0.01) & (-0.07, 0.06) \end{pmatrix}$$

Group lasso:

$$= \begin{pmatrix} \mathbf{(-0.22, -0.11)} & (-0.10, 0.1) & (+0, +0) & (-0, +0) \\ (-0.05, 0.01) & (0.51, 0.61) & (+0, +0) & \mathbf{(-0, -0)} \\ (-8.58, 5.7) & \mathbf{(6.5, 31.01)} & (0.44, 0.55) & (-0.26, -0) \\ (-1.75, 4.86) & \mathbf{(-47.72, -36.37)} & (-0.04, 0.01) & (-0.07, 0.06) \end{pmatrix}$$

Next, in Figure 2 we depict the estimated networks for the MEDIUM VAR based on the first lag transition matrix produced by: (a) least squares and (b) a nonhierarchical normal prior, where for ease of representation the nodes of the network are abbreviated; the full list of the variable names is given in Table A1 of Appendix A in the supplementary material.

As expected, for most variables their previous lag value influences the current value. Further, for the LS-based network, there is a high degree of connectivity, whereas the nonhierarchical-based one exhibits a sparser structure. For the latter, of interest is that the employment index (PAYEMS), the personal consumer expenditures (GS10) and CPI are influenced by many other variables. On the other hand, the Federal Funds Rate influences the broad stock market (SP500) as expected based on finance theory and GS10. In general, the sparser result provided by the nonhierarchical prior, in addition to better forecasting also aids in interpretation, vis-a-vis the LS estimate.

6. Discussion

In this article, we investigate posterior consistency in Bayesian VAR(d) models with both nonhierarchical and hierarchical matrix normal prior distributions on the transition matrices under a Gaussian assumption for the temporal evolution of the time series under consideration and in the presence of a general covariance matrix that captures additional contemporary dependence between them. We establish posterior consistency for both of these priors under high-dimensional scaling. To obtain the desired results, some novel concentration inequalities are provided that are of independent interest. The methodology is illustrated on synthetic and real macroeconomic data. The proposed priors provide better forecasts than the LS estimates for periods up to 1 year ahead, while leading to sparser and potentially easier to interpret relationships, especially for large-scale models.

Supplementary Materials

For the sake of brevity we move additional simulation and real data study, proofs of Lemmas 1–3, Theorems 1 and 2 and useful high-dimensional results of VAR models to the supplementary document.

Funding

The authors gratefully acknowledge support from NSF grants DMS-1511945 (KK) and IIS-1632730 and CNS-1422078 (GM)and NIH grant R01 5R01GM11402902.

References

Armagan, A., Clyde, M., and Dunson, D. B. (2011), "Generalized Beta Mixtures of Gaussians," in *Advances in Neural Information Processing Systems 24*, eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Available at *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4214276/*. [2]

Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [2]

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013), "Posterior Consistency in Linear Models under Shrinkage Priors," *Biometrika*, 100, 1011–1018. [2]

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian Vector Auto Regressions," *Journal of Applied Econometrics*, 25, 71–92. [1,10,11]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-Dimensional Time Series Models," *Annals of Statistics*, 43, 1535–1567. [1,3,4]

Basu, S., Shojaie, A., and Michailidis, G. (2015), "Network Granger Causality with Inherent Grouping Structure," *Journal of Machine Learning Research*, 16, 417–453. [1,2]

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005), "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120, 387–422. [10]

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012), "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," *Journal of Financial Economics*, 104, 535–559. [1]

Bontemps, D. (2011), "Bernstein-von Mises Theorems for Gaussian Regression with Increasing Number of Regressors," *Annals of Statistics*, 39, 2557–2584. [2]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [2]

Doan, T., Litterman, R., and Sims, C. (1984), "Forecasting and Conditional Projection using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100. [1]

Ghoshal, S. (1999), "Asymptotic Normality of Posterior Distributions in High-Dimensional Linear Models," *Bernoulli*, 5, 315–331. [2]

Griffin, J. E., and Brown, P. J. (2010), "Inference with Normal-gamma Prior Distributions in Regression Problems," *Bayesian Analysis*, 5, 171–188. [2]

Korobilis, D. (2013), "Var Forecasting Using Bayesian Variable Selection," *Journal of Applied Econometrics*, 28, 204–230. [1]

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–411. [5]

Lee, J., and Oh, H.-S. (2013), "Bayesian Regression Based on Principal Components for High-dimensional Data," *Journal of Multivariate Analysis*, 117, 175–192. [2]

Lin, J., and Michailidis, G. (2017), "Regularized Estimation and Testing for High-Dimensional Multi-Block Vector-Autoregressive Models," *Journal of Machine Learning Research*, 18, 1–49. [1,2]

Litterman, R. B. (1979), "Techniques of Forecasting Using Vector Autoregressions," Technical Report, Working Papers 115, Federal Reserve Bank of Minneapolis. Available at *https://www.minneapolisfed.org/research/wp/wp115.pdf*. [1]

Lütkepohl, H. (2007), *New Introduction to Multiple Time Series Analysis*, New York: Springer. [1,2,4,8]

Melnyk, I., and Banerjee, A. (2016), "Estimating Structured Vector Autoregressive Models," in *Proceedings of The 33rd International Conference on Machine Learning*. Available at *http://proceedings.mlr.press/v48/melnyk16.html*. [1,4]

Mol, C. D., Giannone, D., and Reichlin, L. (2008), "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics*, 146, 318–328. Honoring the research contributions of Charles R. Nelson. [1]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [5]

Raskutti, G., Yuan, M., and Chen, H. (2018), "Convex Regularization for High-Dimensional Tensor Regression," *The Annals of Statistics*. [1]

Schweinberger, M., Babkin, S., and Ensor, K. B. (2017), "High-dimensional Multivariate Time Series with Additional Structure," *Journal of Computational and Graphical Statistics*, 26, 610–622. [1]

Seth, A. K., Barrett, A. B., and Barnett, L. (2015), "Granger Causality Analysis in Neuroscience and Neuroimaging," *Journal of Neuroscience*, 35, 3293–3297. [1]

Sims, C. A. (1980), "Macroeconomics and Reality," *Econometrica: Journal of the Econometric Society*, 48, 1–48. [10,11]

Sparks, D. K., Khare, K., and Ghosh, M. (2015), "Necessary and Sufficient Conditions for High-dimensional Posterior Consistency Under *g*-Priors," *Bayesian Analysis*, 10, 627–664. [2]

Stock, J. H., and Watson, M. W. (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493. [10]

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368. [8]