The Annals of Statistics
2021, Vol. 49, No. 3, 1267–1299
https://doi.org/10.1214/20-AOS1992
© Institute of Mathematical Statistics, 2021

STRONG SELECTION CONSISTENCY OF BAYESIAN VECTOR AUTOREGRESSIVE MODELS BASED ON A PSEUDO-LIKELIHOOD APPROACH

BY SATYAJIT GHOSH[1], KSHITIJ KHARE[2] AND GEORGE MICHAILIDIS[3]

[1]*Department of Statistics and Biostatistics, Rutgers University, satyajitghosh90@ufl.edu*
[2]*Department of Statistics, University of Florida, kdkhare@stat.ufl.edu*
[3]*Informatics Institute, University of Florida, gmichail@ufl.edu*

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 will provide novel insights into the role of temporal dependence for regularized estimates of the models parameters. However, hardly anything is known regarding posterior model selection consistency for Bayesian VAR models in such regimes.

In this work, we develop a pseudo-likelihood based Bayesian approach for consistent variable selection in high-dimensional VAR models by considering hierarchical normal priors on the autoregressive coefficients, as well as on the model space. We establish *strong selection consistency* of the proposed method, namely that the posterior probability assigned to the true underlying VAR model converges to one under high-dimensional scaling where the dimension p of the VAR system grows nearly exponentially with the sample size n.

Further, the result is established under mild regularity conditions on the problem parameters. Finally, as a by-product of these results, we also establish strong selection consistency for the sparse high-dimensional linear regression model with serially correlated regressors and errors.

1. Introduction. Modeling a large panel of time series is an important task in many fields, including macroeconomic modeling Stock and Watson (2005), identification of regulatory networks in functional genomics Michailidis and d'Alché-Buc (2013) and brain connectivities from neuroimaging data Seth, Barrett and Barnett (2015). A popular and flexible model is that of Vector Autoregressions (VAR) that represents the current values of each time series as a linear function of the first d-lags of itself and the other time series under consideration, plus a serially uncorrelated error term. Due to the importance of the VAR model in economic policy analysis (Sims (1980)), its statistical properties have been thoroughly explored in the econometrics literature for low-dimensional settings (Lütkepohl (2005)). However, new applications in genomics and neuroscience, as well as the realization by macroeconomic modelers that VAR models based on a small number of variables (time series) lead to estimates that contradict basic tenets of economic theory, accentuated the need to examine VAR models in high-dimensional settings. An in-depth theoretical analysis of the model for Gaussian data under a sparsity assumption and using an ℓ_1 penalty term was provided in Basu and Michailidis (2015), while follow-up work extended the results to other penalties Melnyk and Banerjee (2016), to strategies for selecting the number of lags Nicholson, Matteson and Bien (2017) and to incorporating exogenous variables Lin and Michailidis (2017). Key is-

Received August 2019; revised June 2020.
MSC2020 subject classifications. 62M10, 62F15, 62J07.
Key words and phrases. Bayesian variable selection, vector autoregression, pseudo-likelihood, strong selection consistency, high-dimensional data.

sues that this line of work addressed was the role of temporal dependence on the estimates of the model parameters, together with providing technical tools (appropriately modified concentration inequalities for the resulting design matrix and with the error term) to handle this dependence (see Proposition 2.4 in Basu and Michailidis (2015)).

On the other hand, Bayesian approaches for analyzing medium and large size VAR models have been quite extensively used in empirical work in macroeconomic modeling and forecasting (e.g. Bańbura, Giannone and Reichlin (2010), De Mol, Giannone and Reichlin (2008), Robertson and Tallman (1999), Sims and Zha (1998)). However, a detailed theoretical investigation of Bayesian variable selection procedures for the VAR parameters in modern high-dimensional settings has not been undertaken. This is in contrast to Bayesian variable selection for high-dimensional linear regression models with independent and identically distributed observations, where a rich literature exists; for example, see Bondell and Reich (2012), Brown, Vannucci and Fearn (2002), George and Foster (2000), Ishwaran and Rao (2005), Johnson and Rossell (2012), Kinney and Dunson (2007), Liang et al. (2008), Narisetty and He (2014), Song and Liang (2015) to name a few. Almost all of these methods consider a hierarchical prior structure, where given a set of active regression coefficients, a concentrated prior around zero (spike) is placed on the set of inactive coefficients, and a diffused prior (slab) is placed on the set of active coefficients. An appropriate prior distribution is then chosen for the activity pattern of coefficients; one common choice is to independently assign each coefficient to be active or inactive with a common probability. The high-dimensional consistency properties of these methods have been extensively studied; see, for example, Bondell and Reich (2012), Casella et al. (2009), Castillo, Schmidt-Hieber and van der Vaart (2015), Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018) and references therein. In this line of work, roughly three major notions of posterior selection consistency emerge. The first is pairwise or posterior ratio consistency (Casella et al. (2009)), which implies that the ratio of the posterior probabilities of any nontrue model and the true model goes to zero. The second notion is model selection consistency, which implies that the chosen model (such as the posterior mode or the sparsest model in a credible region) is equal to the true model with probability converging to one (Bondell and Reich (2012)). The strongest notion of consistency is *strong selection consistency* which implies that the posterior probability of the true model converges to one (Castillo, Schmidt-Hieber and van der Vaart (2015), Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018)). In all of the above studies, the results hold when the number of coefficients p grows subexponentially with the sample size n, and the errors are assumed to be independent and identically distributed (i.i.d.) with a common variance parameter σ^2.

The VAR model, although related to linear regression, leads to a more complex setting. In particular, for a p-dimensional stationary time series $\{X^t\}$, a VAR 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 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. In the Gaussian VAR model, the errors $\boldsymbol{\varepsilon}^t$ are assumed to be independent and identically distributed (i.i.d.) $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, with the error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ unknown in most applications. Note that our main parameters of interest are the transition matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d$, and we treat the error covariance matrix as an unknown nuisance parameter.

The model in (1) can be rewritten in the Yule–Walker representation (Lütkepohl (2005)) as

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

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, $\boldsymbol{\mu}$ is estimated by the vector of sample means $\bar{X} = \sum_1^n X^t$. Henceforth, we assume without loss of generality that $\boldsymbol{\mu} = \mathbf{0}$. Based on the data $\{X^0, \ldots, X^T\}$, we define the response matrix $\mathbf{Y}$ and design matrix $\mathbf{X}$ as follows:

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

We can then 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 in the coefficient matrix $\boldsymbol{\Phi}$ is $q = dp^2$, respectively.

Although the estimation of the VAR model parameters can be formulated as a linear regression problem using (2), there are significant differences from a technical point of view. The most important difference is that the resulting (random) design matrix for the VAR model in (2) exhibits dependencies both between its rows and across its columns, and also with the error term of the model. Another notable difference is that the single variance parameter σ^2 in simple linear regression is replaced by the covariance matrix $\boldsymbol{\Sigma}_\epsilon$ of the (multivariate) errors in the VAR model with p^2 parameters (we partially alleviate this by using a pseudo-likelihood which only involves the p diagonal entries of $\boldsymbol{\Sigma}_\epsilon$; see Section 3). Thus, the presence of temporal dependence introduces significantly nontrivial technical challenges for establishing high-dimensional posterior selection consistency.

In this paper, we first propose a pseudo-likelihood based, fully Bayesian approach for variable selection, and then examine model selection consistency for the VAR model under the assumption of sparsity in $\boldsymbol{\Phi}$. Sparsity in $\boldsymbol{\Phi}$ is introduced through an "activity graph" $\mathcal{G}$, which is the direct sum of d $p \times p$ matrices of 1's and 0's, identifying which entries of $\boldsymbol{\Phi}$ are active/inactive. Given $\mathcal{G}$, an appropriate multivariate normal prior is chosen for the active entries of $\boldsymbol{\Phi}$. Under standard regularity assumptions, which include stability of the true VAR process, uniform boundedness of the eigenvalues of the true error covariance matrix, and letting p increase at an appropriate subexponential rate with n, we establish strong selection consistency of our Bayesian variable selection approach, that is, the posterior probability assigned to the true activity graph converges to one as $n \to \infty$ (Theorem 4.2). As discussed earlier, this is the strongest notion of posterior selection consistency, and implies model selection consistency, that is, the true activity graph will be the mode of the posterior distribution with probability tending to 1 as $n \to \infty$.

To the best of our knowledge, these variable selection results are the first of their kind for Bayesian high-dimensional multivariate models exhibiting temporal dependence. We would like to point out that in our recent work Ghosh, Khare and Michailidis (2019a), the focus was on posterior *estimation* consistency for VAR models in a moderate dimensional setting wherein $p = o(n)$, whereas in this study the focus is on high-dimensional settings with $\log dp^2 = o(n)$. In order to prove our strong selection consistency result, we first obtain bounds for the Bayes factor of any nontrue activity graph with respect to the true activity graph. Different bounds are obtained for different types of nontrue activity graphs, through

involved analysis and leveraging selected concentration inequalities from Basu and Michailidis (2015) and Ghosh, Khare and Michailidis (2019a) to handle the temporal dependence structure within the design matrix and with the design matrix and the error matrix. Careful calculations ensure that the sum of the resulting bounds over all the nontrue activity graphs converges to zero, thereby proving strong selection consistency. Another notable feature of our strong selection consistency result is its application to linear regression with serially *correlated* predictors and errors (as opposed to i.i.d. errors and predictors considered in previous work such as Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018)).

The remainder of the paper is organized as follows. In Section 2, we review notions of stability and interdependence in VAR models. In Section 3, we motivate and describe our pseudo-likelihood based hierarchical model. In Section 4, we provide the necessary assumptions on the underlying true model, and establish our strong selection consistency and estimation consistency results. Based on these results, we provide a framework for estimating the activity graph and the coefficient matrix in Section 5. In Section 6, we illustrate the results in Section 4 and the estimation methodology in Section 5 using synthetic data. In Section 7, as a by-product of our pseudo-likelihood based approach, we establish strong selection consistency for Bayesian linear regression with serially correlated errors. The proofs of the main theorems are given in the Appendix, while those of all supporting lemmas are in the Supplementary Material (Ghosh, Khare and Michailidis (2021)).

Notation. Throughout the paper, $\mathbb{Z}$, $\mathbb{R}$ 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 are 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. $a_n \sim b_n$ if and only if $\frac{a_n}{b_n} \to c$ for some constant $c > 0$ and $a_n \preceq b_n$ if and only if $a_n = O(b_n)$.

2. Model formulation.

2.1. *Preliminaries*: *Stability of VAR models*. Since VAR models are (linear) dynamical systems, for their analysis we need to establish conditions under which the VAR model in (2) is stable, that is, the time-series process does not diverge over time. The VAR process $\{X^t\}$ with lag d is stable and invertible if $\det(\mathbf{I}_p - \sum_{i=1}^d \mathbf{A}_i z^i) \neq 0$ inside the unit circle of the complex plane, that is, $\{z \in \mathbb{C} : |z| \leq 1\}$. It is often convenient to rewrite the VAR model of lag d in (2) as an equivalent dp-variate VAR model of lag 1, $\{\tilde{X}^t\}$, where

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

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

where $\tilde{\mathbf{A}}$ is a $dp \times dp$ transition matrix and the new representation becomes

$$\tilde{X}^t = \tilde{\mathbf{A}}\tilde{X}^{t-1} + \boldsymbol{\omega}^t, \quad t = d, \ldots, n+d-1. \tag{3}$$

It follows that $\mathbf{X} = [\tilde{X}^{n+d-2}\tilde{X}^{n+d-3}\cdots\tilde{X}^{d-1}]'$, where the ith row of $\mathbf{X}$ is denoted as a ($dp \times 1$ vector) $\tilde{X}^{n+d-i-1}$. The VAR process $\{X^t\}$ (with lag d) is stable if and only if the VAR process $\{\tilde{X}^t\}$ is stable (see Basu and Michailidis (2015)), which is true if all the eigenvalues of $\tilde{\mathbf{A}}$ have absolute value strictly less than 1; that is, $\max_{1\le i\le p}|\lambda_i(\tilde{\mathbf{A}})| < 1$.

The autocovariance function of the dp-dimensional centered covariance-stationary time series $\{\tilde{X}^t\}$ (stability guarantees stationarity) is defined as $\Gamma_{\tilde{X}}(h) = \operatorname{Cov}(\tilde{X}^t, \tilde{X}^{t+h})$ $t, h \in \mathbb{Z}$ and it is invariant in t. Also, for the stable VAR process $\{\tilde{X}^t\}$, the (matrix-valued) spectral density $f_{\tilde{X}}$ is defined by

$$f_{\tilde{X}}(\theta) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\operatorname{Cov}(\tilde{X}^t, \tilde{X}^{t+h})\mathrm{e}^{-ih\theta}, \quad \theta \in [-\pi, \pi].$$

The maximum and minimum eigenvalues of the spectral density $f_{\tilde{X}}$ are denoted by

$$\mathscr{M}(f_{\tilde{X}}) := \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \lambda_{\max}\big(f_{\tilde{X}}(\theta)\big),$$

$$m(f_{\tilde{X}}) := \operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda_{\max}\big(f_{\tilde{X}}(\theta)\big).$$

The largest eigenvalue $\mathscr{M}(f_{\tilde{X}})$ captures the "peak" of the spectral density and can be used as a measure of stability of the VAR process. In particular, processes with larger $\mathscr{M}(f_{\tilde{X}})$ are considered less stable. The smallest eigenvalue $m(f_{\tilde{X}})$ captures the dependence among the univariate components of $\{\tilde{X}^t\}$. As pointed out in Basu and Michailidis (2015), Proposition 2.2, it is often easier to work with quantities $\mu_{\max}(\tilde{\mathscr{A}})$ and $\mu_{\min}(\tilde{\mathscr{A}})$ which are defined as

$$\mu_{\max}(\tilde{\mathscr{A}}) := \max_{\theta\in[-\pi,\pi]} \lambda_{\max}\big((\mathbf{I}_p - \tilde{\mathbf{A}}'\mathbf{e}^{i\theta})(\mathbf{I}_p - \tilde{\mathbf{A}}\mathbf{e}^{-i\theta})\big) \le \big[1 + (v_{in} + v_{\text{out}})/2\big]^2, \tag{4}$$

where $v_{in} = \max_{1\le i\le dp}\sum_{j=1}^{dp}|\tilde{\mathbf{A}}_{ij}|$ and $v_{\text{out}} = \max_{1\le j\le dp}\sum_{i=1}^{dp}|\tilde{\mathbf{A}}_{ij}|$, and

$$\begin{aligned}\mu_{\min}(\tilde{\mathscr{A}}) &:= \min_{\theta\in[-\pi,\pi]} \lambda_{\min}\big((\mathbf{I}_p - \tilde{\mathbf{A}}'\mathbf{e}^{i\theta})(\mathbf{I}_p - \tilde{\mathbf{A}}\mathbf{e}^{-i\theta})\big)\\ &\ge \big(1 - \rho(\tilde{\mathbf{A}})\big)^2\|\mathbf{P}\|^{-2}\big\|\mathbf{P}^{-1}\big\|^{-2},\end{aligned} \tag{5}$$

and where $\rho(\tilde{\mathbf{A}})$ is the spectral radius of $\tilde{\mathbf{A}}$ and the columns of $\mathbf{P}$ correspond to eigenvectors of $\tilde{\mathbf{A}}$ (assuming $\tilde{\mathbf{A}}$ is diagonalizable).

We establish additional bounds and properties on $\mu_{\max}$ and $\mu_{\min}$ which will be useful for understanding their behavior, and verifying the validity of our assumptions for consistency in various settings, a novel result of independent interest (see Supplementary Material, Sections S5 and S6).

PROPOSITION 2.1.

(a) *Let $\|\cdot\|$ denote the operator norm. Then*

$$\mu_{\min}(\tilde{\mathscr{A}}) \ge 1 - 2\|\tilde{\mathbf{A}}\| + \lambda_{\min}(\tilde{\mathbf{A}}'\tilde{\mathbf{A}}), \tag{6}$$

(b) *Let $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the maximum row and column sum norms, respectively. Then*

$$\mu_{\min}(\tilde{\mathscr{A}}) \ge \frac{3-\sqrt{5}}{2} - \left(2\sum_{i=1}^{d}\|\mathbf{A}_i\|_1 + \|\mathbf{A}_1\|_\infty\right). \tag{7}$$

(c) *If* $d = 1$ *and* $\tilde{\mathbf{A}} = \mathbf{A}$ *is symmetric, then*

$$\mu_{\max}(\mathscr{A}) = \big(1 + \rho(\mathbf{A})\big)^2, \tag{8}$$

where $\rho(\mathbf{A})$ *is the spectral radius of* $\mathbf{A}$. *Similarly,*

$$\mu_{\min}(\mathscr{A}) = \big(1 - \rho(\mathbf{A})\big)^2. \tag{9}$$

Since each $\boldsymbol{\varepsilon}^t$ is i.i.d. as $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, each row of $\mathbf{X}$ is distributed as $\mathcal{N}_{dp}(\mathbf{0}, \mathbf{C}_X)$, where $\mathbf{C}_X = \Gamma_{\tilde{X}}(0)$. The quantities $\mu_{\min}(\tilde{\mathscr{A}})$, $\mu_{\max}(\tilde{\mathscr{A}})$, the eigenvalues of $f_{\tilde{X}}$ and the eigenvalues of $\mathbf{C}_X$ are related by the following chain of inequalities (see Basu and Michailidis (2015)):

$$\frac{\lambda_{\min}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\max}(\tilde{\mathscr{A}})} \le 2\pi m(f_{\tilde{X}}) \le \lambda_{\min}(\mathbf{C}_X) \le \lambda_{\max}(\mathbf{C}_X) \le 2\pi \mathscr{M}(f_{\tilde{X}}) \le \frac{\lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\tilde{\mathscr{A}})}. \tag{10}$$

3. Bayesian VAR model formulation based on a pseudo-likelihood approach. We consider a high-dimensional setting wherein the dimension p of the VAR model (2) increases with the sample size n. In such settings, a popular and effective method to reduce the dimension of the parameter space is to induce sparsity or zeros in the parameter. Following this approach, we introduce binary variables γ_{ij} to indicate if the (k, ℓ)-th entry of $\boldsymbol{\Phi}$ is active, that is, $\gamma_{k\ell} = \mathbb{1}(\boldsymbol{\Phi}_{k\ell} \neq 0)$ for $1 \le k \le dp, 1 \le \ell \le p$. Let $\mathscr{G} := ((\gamma_{k\ell}))_{dp \times p}$ be a matrix of active positions in $\boldsymbol{\Phi}$, henceforth referred to as the "activity graph of $\boldsymbol{\Phi}$."

Given an activity graph $\mathscr{G}$, let $\nu_i = \nu_i(\mathscr{G}) = \sum_{k=1}^{dp} \gamma_{ki}$ denote the number of nonzero entries in the ith column of $\mathscr{G}$, and $\tilde{\boldsymbol{\phi}}_i = (\phi_{ki})_{k:\gamma_{ki}=1}$ be the ν_i-dimensional vector of active coefficients in the ith column of $\boldsymbol{\Phi}$. The VAR model (2) can now be written as

$$\mathbf{y}_i = \mathbf{X}_i \tilde{\boldsymbol{\phi}}_i + \boldsymbol{\xi}_i \quad \text{for } i = 1, \ldots, p, \tag{11}$$

where $\boldsymbol{\xi}_i$ and $\boldsymbol{y}_i$ correspond to the ith column of $\mathbf{E}$ and $\mathbf{Y}$, respectively, and $\mathbf{X}_i$ is a $n \times \nu_i$ submatrix of $\mathbf{X}$ consisting of the columns of $\mathbf{X}$ corresponding to the active entries in the ith column of $\boldsymbol{\Phi}$. Let σ_i^2 be the ith diagonal entry of $\boldsymbol{\Sigma}_\epsilon$. Since the rows of $\mathbf{E}$ (i.e., $\boldsymbol{\varepsilon}^t$) are i.i.d. $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, we have that $\boldsymbol{\xi}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma_i^2 \mathbf{I}_n)$.

Note that the vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$. The joint density of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$, or equivalently, the likelihood function of $\boldsymbol{\Phi}, \boldsymbol{\Sigma}_\epsilon$ given the data, is given by

$$\begin{aligned} &L(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_\epsilon \mid \mathbf{y}_i, i = 1, \ldots, p) \\ &\quad := \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_\epsilon|^n}} \exp\bigg(-\frac{1}{2} \operatorname{tr}\big((\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi})\boldsymbol{\Sigma}_\epsilon^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi})'\big)\bigg). \end{aligned} \tag{12}$$

As mentioned earlier, the main parameter of interest is the coefficient matrix $\boldsymbol{\Phi}$, while the error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ is essentially an unknown nuisance parameter. Keeping this in mind, we construct a pseudo-likelihood which is equal to the joint density of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$ under the assumption that $\boldsymbol{\Sigma}_\epsilon$ is diagonal. In particular, we define

$$\begin{aligned} &L_{\text{pseudo}}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}_\epsilon \mid \mathbf{y}_i, i = 1, \ldots, p) \\ &\quad := \frac{1}{\sqrt{(2\pi)^p \prod_{i=1}^p \sigma_i^2}^n} \exp\bigg(-\frac{1}{2} \sum_{i=1}^p \frac{\|\mathbf{y}_i - \mathbf{X}_i \tilde{\boldsymbol{\phi}}_i\|^2}{\sigma_i^2}\bigg). \end{aligned} \tag{13}$$

The function L_{pseudo} in (13) has a simpler form than the likelihood function L and only involves the diagonal entries of $\boldsymbol{\Sigma}_\epsilon$. This form leads to significant computational simplifications. For example, under our proposed Bayesian model, the pseudo-posterior probability for any activity graph can be computed in closed form, as shown in (15) below. On the other

hand, it turns out that the regular posterior probabilities computed using the likelihood L in (12) can in general only be expressed as intractable high-dimensional integrals.

The use of L_{pseudo} in place of L, of course, comes with a cost. Ignoring the correlations between the errors leads to a loss of statistical efficiency and raises potential questions regarding the accuracy of the resulting estimates. However, the main result of the paper (Theorem 4.2) establishes under mild regularity conditions (which include the assumption that the eigenvalues of the true error covariance matrix are uniformly bounded) that the pseudo-posterior activity graph density puts all of its mass at the true activity graph generating the data as $n, p \to \infty$. The associated computational gains, along with the consistency result, render the pseudo-likelihood based approach strongly preferable in the high-dimensional setting considered in this paper.

Having made the choice to use L_{pseudo}, we now construct the following hierarchical prior distribution for $(\mathbf{\Phi}, \mathcal{G})$:

$$
\begin{aligned}
&\phi_{ji} \mid \mathcal{G}, \boldsymbol{\sigma^2} \sim (1-\gamma_{ji}) 1_{\phi_{ji}=0} + \gamma_{ji} \mathcal{N}\big(0, \tau^2 \sigma_i^2\big) \\
&\qquad\qquad \text{independently for } 1 \le i \le p, 1 \le j \le dp \\
&\pi(\mathcal{G}) \propto \prod_{i=1}^{p} \big\{ q_1^{v_i(\mathcal{G})} (1-q_1)^{dp - v_i(\mathcal{G})} 1_{\{v_i(\mathcal{G}) < M\}} + q_2^{v_i(\mathcal{G})} (1-q_2)^{dp - v_i(\mathcal{G})} 1_{\{v_i(\mathcal{G}) \ge M\}} \big\} \\
&\sigma_i^2 \sim \text{Inv. Gamma}(\alpha_i, \beta_i/2) \quad \text{independently for } i = 1, \ldots, p,
\end{aligned}
\tag{14}
$$

where $\boldsymbol{\sigma^2}$ denotes $(\sigma_1^2, \ldots, \sigma_p^2)$.

Interpretation of M, q_1 and q_2: Note that under the prior distribution in (14), each column of $\mathcal{G}$ is a priori independent and identically distributed. Next, we comment on how the hyperparameters M, q_1 and q_2 shape this common distribution of the columns of $\mathcal{G}$. For each column of $\mathcal{G}$, the parameter space is the set of dp-dimensional vectors with entries in $\{0, 1\}$. The prior on each column is defined separately on two subsets of the parameter space, as described below:

- The first subset is the set of *realistic vectors*, where the number of nonzero entries in the given column is less than M. On this subset, the probability of each vector is proportional to $q_1^{n_z}(1-q_1)^{n_z}$, where n_z denotes the number of nonzero entries in that vector. In this sense, q_1 will be referred to as the *edge inclusion probability* for realistic vectors.
- The second subset is the set of *unrealistic vectors*, where the number of nonzero entries in the given column is greater than M. On this subset, the probability of each vector is proportional to $q_2^{n_z}(1-q_2)^{n_z}$, where n_z denotes the number of nonzero entries in that vector. In this sense, q_2 will be referred to as the *edge inclusion probability* for unrealistic vectors.

The hyperparameter M, which determines the density (number of nonzeros) of a realistic vector, will be referred to as the *realistic model cutoff size*. Models where the number of nonzero entries in at least one column is larger than M will be referred to as *unrealistically large models*. Note that if $q_2 < q_1$, then we are additionally penalizing unrealistically large models, that is, models where the number of nonzero entries in at least one column is larger than M.

In fact, many recent papers with related priors distributions in the context of simple linear regression and covariance estimation (see Cao, Khare and Ghosh (2020), Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018)) use a similar structure as the prior distribution in (14). In particular, they choose $M = O(n/\log p)$ and make the extreme choice $q_2 = 0$.

The Bayesian hierarchical model defined by (13) and (14) can be used to infer the activity graph as follows: by Bayes' rule, and straightforward computations, the (marginal) pseudo-posterior graph selection probabilities are given by

$$\begin{aligned}
&\pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y}) \\
&\propto \pi(\mathcal{G}) \int\int L_{\text{pseudo}}(\mathbf{\Phi}, \mathbf{\Sigma}_\epsilon \mid \mathbf{y}_i, i = 1, \ldots, p) \pi\big(\mathbf{\Phi} | \mathcal{G}, \boldsymbol{\sigma}^{\mathbf{2}}\big)\, \mathrm{d}\mathbf{\Phi} \prod_{i=1}^{p} \pi\big(\sigma_i^2\big)\, \mathrm{d}\sigma_i^2 \qquad (15) \\
&\propto \prod_{i=1}^{p} q_{\nu_i(\mathcal{G})}^{\nu_i(\mathcal{G})} (1 - q_{\nu_i(\mathcal{G})})^{dp - \nu_i(\mathcal{G})} \left(\frac{1}{\tau\sqrt{n}}\right)^{\nu_i(\mathcal{G})} \left| \frac{\mathbf{X}_i' \mathbf{X}_i}{n} + \frac{\mathbf{I}_{\nu_i}}{n\tau^2} \right|^{-1/2} \left(S_i + \frac{\beta_i}{n} \right)^{-(\frac{n}{2} + \alpha_i)},
\end{aligned}$$

where

$$q_{\nu_i(\mathcal{G})} = \begin{cases} q_1 & \text{if } \nu_i(\mathcal{G}) < M, \\ q_2 & \text{if } \nu_i(\mathcal{G}) \geq M, \end{cases}$$

$$S_i := \frac{\mathbf{y}_i' \mathbf{y}_i}{n} - \frac{\mathbf{y}_i' \mathbf{X}_i}{n} \left(\frac{\mathbf{X}_i' \mathbf{X}_i}{n} + \frac{\mathbf{I}_{\nu_i}}{n\tau^2} \right)^{-1} \frac{\mathbf{X}_i' \mathbf{y}_i}{n}$$

The pseudo-posterior density $\pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y})$ is derived in Section S2 of the Supplementary Material (Ghosh, Khare and Michailidis (2021)). Note that the pseudo-posterior probabilities $\pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y})$ can in principle be used to select the graph by computing the posterior mode defined by $\hat{\mathcal{G}} := \arg\max_{\mathcal{G}} \pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y})$. However, there are 2^{dp^2} possible activity graphs, and searching over such a large space becomes prohibitively expensive. However, standard stochastic search ideas combined with a good starting point can be used to perform a targeted and computationally effective search, as described in detail in Section 5.

4. Theoretical results. In this section, we provide our main consistency result for the pseudo-likelihood based Bayesian model specified in (13) and (14). As previously mentioned, we let the dimension $p = p_n$ of the VAR model vary with n, so that our results are relevant in high-dimensional settings. We assume that our data come from the following true VAR model: 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 i.i.d. $\mathcal{N}_{p_n}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n})$. Further, $\{\mathbf{\Phi}_{0n}\}_{n\geq 1}$ denotes the sequence of the true coefficient matrices given by $\mathbf{\Phi}_{0n}' := [\mathbf{A}_{1,0n} \mathbf{A}_{2,0n} \cdots \mathbf{A}_{d,0n}]$, and $\{\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}\}_{n\geq 1}$ denotes the sequence of the true error covariance matrices. Let $\mathbb{P}_0$ denote the probability measure underlying the true model described above, and $\mathcal{G}_0 = \mathcal{G}_{0,n}$ the true underlying activity graph for the sparse coefficient matrix $\mathbf{\Phi}_0$. The quantities $\mu_{\min}(\tilde{\mathcal{A}})$, $\mu_{\max}(\tilde{\mathcal{A}})$ and $\mathbf{C}_X$ are as defined in Section 2 with $\mathbf{\Phi}_{0n}$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}$ as the underlying parameter values. The maximum number of nonnull entries within the columns of $\mathbf{\Phi}_{0n}$ and its minimum signal strength are defined respectively by

$$k_n := \max_{1 \leq i \leq p} \nu_i + 1$$

and $s_n := \inf_{(i,j):\tilde{\boldsymbol{\phi}}_{ij} \neq 0} |\tilde{\boldsymbol{\phi}}_{ij}|$. The total number of nonzero entries in $\mathbf{\Phi}_{0n}$ is denoted by δ_n, that is, $\delta_n = \sum_{i=1}^{p} \nu_i$. For ease of exposition, we will henceforth denote $\mathbf{\Phi}_{0n}$ as $\mathbf{\Phi}_0$, and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}$ as $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}$, and highlight their dependence on n as needed.

Assumptions for establishing posterior selection consistency: We impose the following regularity assumptions regarding the parameters of the true model.

ASSUMPTION A1. $\frac{1+\mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})} \sqrt{\frac{k_n \log dp}{n}} = o(1)$.

ASSUMPTION A2 (Bounded eigenvalues). There exist $0 < \lambda_1 < \lambda_2 < \infty$ and $0 < \sigma_{\min} < \sigma_{\max} < \infty$ not depending on n such that $\lambda_1 < \lambda_{\min}(\mathbf{C}_X)$, the maximum eigenvalue of any principal submatrix of $\mathbf{C}_X$ of dimension at most k_n is uniformly bounded by λ_2, and $\sigma_{\min} < \lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}) \leq \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}) < \sigma_{\max}$.

For ease of exposition, we denote $4\pi\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0})\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}$ by $\mathscr{B}_n$, and let c be the universal constant appearing in the Hanson–Wright inequality of Vershynin (2018), Theorem 6.2.1. For the Gaussian case, it can be shown by routine computations that the choice $c = 1/256$ works.

ASSUMPTION A3 (Rate of decay of edge probability). The realistic model cutoff size is given by $M_n = \frac{\min(\sigma_{\min}/4,1)\min(\lambda_1^2/4,1)c}{16\mathscr{B}_n^2}\frac{n}{\log dp}$, and the edge inclusion probabilities satisfy $q_{1,n} = (dp^2)^{-\frac{8\mathscr{B}_n^2 k_n}{\sigma_{\min}\lambda_1}}$ and $q_{2,n} = q_{1,n}^{(\log n)^{3/2}}$ for large enough n, in model prior (14).

ASSUMPTION A4 (Minimum signal strength). $\frac{\mathscr{B}_n^2 k_n \log dp + \log n}{n s_n^2} \to 0$.

Note that for our analysis, the slab variance τ^2 is taken to be a fixed positive constant which does not vary with the sample size n. Next, we briefly comment on the assumptions and contrast them with relevant assumptions in Basu and Michailidis (2015) and Ghosh, Khare and Michailidis (2019a).

- Assumption A1 states that p can increase at an appropriate subexponential rate as compared to n. This is a much faster rate than the assumption $p = o(n)$ or $p = o(n/\log n)$ in Ghosh, Khare and Michailidis (2019a). Note that we assume that the true parameter $\mathbf{\Phi}_0$ is sparse and k_n, the maximum number (column-wise) of nonzero entries in $\mathbf{\Phi}_0$ plays a role in determining the rate at which p can increase with n. In Ghosh, Khare and Michailidis (2019a), no sparsity is assumed for $\mathbf{\Phi}_0$, and assumptions bounding the singular values of $\mathbf{\Phi}_0$ are needed to establish (estimation) consistency of the posterior distribution of the coefficient matrix $\mathbf{\Phi}$. Assumption A1 also ensures that the true VAR process is well behaved (stability) and the part regarding $\mu_{\min}(\tilde{\mathscr{A}})$ and $\mu_{\max}(\tilde{\mathscr{A}})$ allows us to leverage crucial concentration inequalities in Ghosh, Khare and Michailidis (2019a) to control the behavior of $\mathbf{X}'\mathbf{E}$ and $\mathbf{X}'\mathbf{X}$. Also note that the rate at which p increases with n is asymptotically equivalent to the lasso-rate established in Proposition 4.1 of Basu and Michailidis (2015).
- Assumption A2 is standard in the literature and allows us to bound the minimum and maximum eigenvalues of the matrix $\mathbf{X}'\mathbf{X}/n$ away from zero and infinity, respectively, with high probability. This assumption corresponds to Assumption B3 in Ghosh, Khare and Michailidis (2019a).
- Assumption A3 provides the realistic model cutoff size M_n as a multiple of $\frac{n}{\mathscr{B}_n^2 \log p}$. This is very similar to the cutoff sizes used in the context of simple linear regression (e.g., Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018)), wherein the realistic model cutoff sizes are multiples of $\frac{n}{\log p}$. In our more complex setting, the extra factor of $\mathscr{B}_n^2$ arises due to the effect of temporal dependence. The rate of decay of the edge inclusion probability $q_{1,n}$ is assumed to be an appropriate power of p. This is again similar to corresponding assumptions in the context of linear regression (Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018)). Again, the power of p contains an extra factor of $\mathscr{B}_n^2$, which is needed to tackle the temporal dependence in the more complex VAR setting.
- Assumption A4 is again a standard assumption and provides a lower bound for the minimum signal strength, that is, the smallest entry (in absolute value) of $\mathbf{\Phi}_0$. Note that we allow the minimum signal strength to converge to zero as $n \to \infty$.

Next, we establish the main posterior consistency result. In particular, we show that the posterior mass assigned to the true activity graph $\mathcal{G}_0$ converges to 1 in probability (under the true model). This notion of consistency is quite powerful and is referred to as strong selection consistency (see Narisetty and He (2014)).

THEOREM 4.1 (Strong selection consistency). *For any centered VAR(d) model (2) with prior (14) on $\mathbf{\Phi}, \mathcal{G}, \mathbf{\Sigma}_\epsilon$ satisfying Assumptions A1–A4, the following holds*:

$$\pi_{\text{pseudo}}(\mathcal{G} = \mathcal{G}_0 \mid \mathcal{Y}) \xrightarrow{\mathbb{P}_0} 1 \quad \text{as } n \to \infty,$$

that is, the probability mass placed by the pseudo-posterior on the true activity graph $\mathcal{G}_0$ converges to 1 *as $n \to \infty$. In particular, the true activity graph $\mathcal{G}_0$ is the mode of the posterior distribution with probability tending to* 1 *as $n \to \infty$.*

The above model selection consistency result can be immediately leveraged to obtain the following estimation consistency result.

THEOREM 4.2 (Estimation consistency rate). *For any centered VAR(d) model (2) with prior (14) on $\mathbf{\Phi}, \mathcal{G}, \mathbf{\Sigma}_\epsilon$ satisfying Assumptions A1–A4, there exists a constant K (not depending on n) such that*

$$\mathbb{E}_0\left[\Pi_{\text{pseudo}}\left(\|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F > K \frac{1 + \mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})} \sqrt{\frac{\delta_n \log dp}{n}} \,\middle|\, \mathcal{Y}\right)\right] \to 0 \quad \text{as } n \to \infty.$$

The estimation rate of $\frac{1+\mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})}\sqrt{\frac{\delta_n \log dp}{n}}$ in the above result is the same as that of the lasso based estimation rate obtained by Basu and Michailidis (2015). However, there is a difference in the assumptions needed to obtain the two results. We allow p to grow at a faster rate (smaller order than $e^{n/(\mathcal{B}_n^2 k_n)}$) as compared to Basu and Michailidis (2015) (smaller order than $e^{n/(\mathcal{B}_n^2 \delta_n)}$). Note that δ_n is the total number of nonzero entries in $\mathbf{\Phi}_0$, which can be much larger than k_n (the maximum number of entries in any column of $\mathbf{\Phi}_0$). For example, if $\mathbf{\Phi}_0$ is banded and $d = 1$, then $\delta_n = pk_n$. On the other hand, Basu and Michailidis (2015) obtain estimation consistency rates for the VAR lasso directly (without establishing model selection consistency) and do not need any assumptions on the minimum signal size analogous to Assumption A4.

REMARK 4.3. Based on a reviewer's comment, we explored the possibility of directly obtaining estimation consistency rates without any assumptions on the minimum signal size for our Bayesian procedure. After some additional work, we were indeed able to *directly* establish estimation consistency rates for Φ only under Assumptions A1–A3. However, the trade-off for removing Assumption A4 is that the estimation consistency rate in Theorem 4.4 grows by a factor of $\sqrt{k_n}$, compared to the rate in Theorem 4.2. As stated earlier, in many settings, the total number of nonzero entries in $\mathbf{\Phi}_0$ grows with n, but k_n (the maximum number of nonzero entries in any column) remains bounded. In these settings, the estimation consistency rate in Theorem 4.4 is the same as the lasso rate. Note that the rate of growth of p in Theorem 4.4 is always faster than the rate of growth of p for the VAR lasso in Basu and Michailidis (2015).

THEOREM 4.4 (Estimation consistency rate without minimum signal size assumption). *For any centered VAR(d) model (2) with prior (14) on $\mathbf{\Phi}, \mathcal{G}, \mathbf{\Sigma}_\epsilon$ satisfying Assumptions A1–A3, there exists a constant K (not depending on n) such that*

$$\mathbb{E}_0\left[\Pi_{\text{pseudo}}\left(\|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F > K \frac{1 + \mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})} \sqrt{\frac{\delta_n k_n \log dp}{n}} \,\middle|\, \mathcal{Y}\right)\right] \to 0 \quad \text{as } n \to \infty.$$

The proof of the main technical results is provided in the Appendix. Next, we provide a brief roadmap of the proof, and compare and contrast it with recent strong selection consistency proofs for linear regression in Narisetty and He (2014), and covariance estimation in Cao, Khare and Ghosh (2019). The proof strategy first bounds the Bayes factor of any nontrue activity graph with respect to the true activity graph $\mathcal{G}_0$ by a product of p terms, each corresponding to one of the p variables. Each of these terms is a product of a "prior part" (depending exclusively on q_1 and q_2) and a "data part" (depending exclusively on $\mathcal{Y}$). The high level conceptual plan of attack for the next step is similar to Narisetty and He (2014) and Cao, Khare and Ghosh (2019): (a) consider three different cases depending on whether the nontrue model (restricted to each specific variable) is contained in the true model, contains the true model or neither of the two, and (b) for each case obtain a uniform bound for the corresponding Bayes factor term over all the variables.

The crucial difference, however lies in the mathematical analysis for implementing this plan. Note that (11) expresses the observations for the ith variable in the VAR in a linear regression setup. The vector of observations for the ith variable in the directed acyclic graph model in Cao, Khare and Ghosh (2019), and all the observations in the simple linear regression model in Narisetty and He (2014), can be similarly expressed in a linear regression setup, albeit with different design matrices, coefficients and errors. The regression models corresponding to Cao, Khare and Ghosh (2019), Narisetty and He (2014) either have a nonrandom design matrix or a design matrix with i.i.d. sub-Gaussian rows, and the design matrix is independent of the error vector. On the other hand, the VAR design matrix $\mathbf{X}_i$ in (11) has dependencies across both its rows and columns, and also shares dependencies with the error vector $\boldsymbol{\xi}_i$ in (11). These dependencies significantly complicate the analysis of the "data part" for all the terms in the Bayes factor, and require a more detailed and involved treatment.

Finally, we show that the sum of bounds over *all* the nontrue activity graph Bayes factors converges to zero in probability under the true model, which implies strong selection consistency. Note that we choose $q_{2,n} > 0$, and do not need to only consider activity graphs wherein the number of active parameters is bounded by an appropriate function of n. Such restrictions have been imposed for the strong selection consistency proofs in Cao, Khare and Ghosh (2019), Narisetty and He (2014), Shin, Bhattacharya and Johnson (2018).

5. Estimation of model parameters. The estimation of the graph $\mathcal{G}$ and the parameter matrix $\boldsymbol{\Phi}$ is a two-step procedure. We first estimate $\mathcal{G}$ using the mode of the (pseudo) posterior distribution on the activity graphs, that is,

$$\hat{\mathcal{G}} := \arg\max_{\mathcal{G}} \pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y}).$$

Note that the space of activity graphs has 2^{dp^2} elements, each one corresponding to a candidate model. To find the model with the highest (pseudo) posterior probability is computationally challenging for large p. Although MCMC techniques based on the Gibbs sampler are available (Narisetty and He (2014), George and McCulloch (1997)), we prefer to use Shotgun Stochastic Search (Hans, Dobra and West (2007)). This is because $\pi_{\text{pseudo}}(\mathcal{G} \mid \mathcal{Y})$ is available in closed form and extensive parallel computation (Step 1 of the algorithm stated below) significantly reduces the run time compared to a slow mixing Markov chain. A brief description of the Bayesian Stochastic Search (Bayesian SSS) algorithm follows.

VAR Shotgun Stochastic Search (VAR-SSS): Given a graph $\mathcal{G}$ of cardinality k, let $\text{nbd}(\mathcal{G})$ denote the neighborhood containing the following three elements $\{\mathcal{G}^+, \mathcal{G}^0, \mathcal{G}^-\}$, where $\mathcal{G}^+$

is a set containing neighboring models of dimension $k+1$; $\mathcal{G}^0$ contains models of dimension k and $\mathcal{G}^-$ is a set containing neighboring models of dimension $k-1$. We define the score of a graph $\mathcal{G}$ to be its posterior probability without the normalizing constant; that is, $S(\mathcal{G}) :\sim \log \pi_{\text{pseudo}}(\mathcal{G}|\mathcal{Y})$. At the tth iteration, we undertake the following steps:

Step 1 Compute in parallel all $S(\mathcal{G})$ for all $\mathcal{G} \in nbd(\mathcal{G}^{[t]})$, construct $\{\mathcal{G}^+, \mathcal{G}^0, \mathcal{G}^-\}$.
Step 2 Sample $\{\mathcal{G}_*^+, \mathcal{G}_*^0, \mathcal{G}_*^-\}$ from $\{\mathcal{G}^+, \mathcal{G}^0, \mathcal{G}^-\}$ with probabilities proportional to $S(\mathcal{G})$, normalized within each set.
Step 3 Sample $\mathcal{G}^{[t+1]}$ from $\{\mathcal{G}_*^+, \mathcal{G}_*^0, \mathcal{G}_*^-\}$ probabilities proportional to $S(\mathcal{G})$.

Once an estimate of $\mathcal{G}$ is obtained, parameter $\tilde{\boldsymbol{\phi}}_j$ which is the nonzero component of the jth column of $\boldsymbol{\Phi}$ can easily be estimated using the corresponding posterior mean $\hat{\boldsymbol{\phi}}_j := (\mathbf{X}_j'\mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2})^{-1}\mathbf{X}_j'\boldsymbol{y}_j$. This is because

$$\tilde{\boldsymbol{\phi}}_j \mid \mathcal{G}, \sigma_j^2, \mathcal{Y} \sim \mathcal{N}_{\nu_j}\left(\hat{\boldsymbol{\phi}}_j, \sigma_j^2\left(\mathbf{X}_j'\mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2}\right)^{-1}\right). \tag{16}$$

Note that $\hat{\boldsymbol{\phi}}_j$ does not depend on σ_j^2, and hence the unconditional mean of $\tilde{\boldsymbol{\phi}}_j$ (given the data $\mathcal{Y}$ and the graph $\mathcal{G}$) is the same as $\hat{\boldsymbol{\phi}}_j$. In addition, the distribution of $\tilde{\boldsymbol{\phi}}_j \mid \mathcal{G}, \mathcal{Y}$ is available in closed form and given by

$$\hat{\boldsymbol{\phi}}_j \mid \mathcal{G}, \mathcal{Y} \sim \mathrm{t}_{n+2\alpha_j}\left(\text{ location } = \hat{\boldsymbol{\phi}}_j, \text{ scale mat. } \hat{\Sigma}_j = \frac{\boldsymbol{y}_j'(\mathbf{I}-\mathbf{P}_j)\boldsymbol{y}_j + \beta_j}{n+2\alpha_j}\left(\mathbf{X}_j'\mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2}\right)^{-1}\right),$$

where $\mathbf{P}_j := \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2})^{-1}\mathbf{X}_j'$, and α_j, β_j are as in (14). Thus, posterior credible intervals can easily be constructed using direct sampling from the noncentral matrix variate t distribution.

6. Performance evaluation. In this section, we evaluate the finite sample performance of our Bayesian modeling framework. The results in Section 4 provide rigorous justification for uncertainty quantification through posterior credible intervals in high-dimensional settings. To the best of our knowledge, asymptotic validity of post-selection inference (confidence intervals) for lasso based estimates has not been established in the high-dimensional VAR setting.

We consider a $p = 50$ dimensional VAR model of lag $d = 1$. The true parameter $\mathbf{A}_1$ has spectral radius $\max_{1\le i\le 50}|\lambda_i(\mathbf{A}_1)| = 0.95$ with entries generated from $U(-10, 10)$ and edge density $\frac{1}{p}$ making it sparse with high signal strength. The error covariance matrix used is $\boldsymbol{\Sigma}_\epsilon = \sigma^2\mathbf{I}_p$ and adjusted so that the signal-to-noise ratio SNR $= \frac{\|\mathbf{A}_1\|_F}{\sigma} = 2$. The following is a 11×6 submatrix (which we denote by $\mathbf{A}_1$ within a parenthesis) of the complete true

transition matrix $\mathbf{A}_1$,

$$(\mathbf{A}_1) = \begin{pmatrix} 1.38\textcircled{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.99\textcircled{2} & 0 & 0 \\ 0 & 0 & -1.86\textcircled{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.96\textcircled{4} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.33\textcircled{5} \end{pmatrix}.$$

The 5 circled entries—① through ⑤—from submatrix $(\mathbf{A}_1)$ are selected for subsequent consideration; that is, calculating their posterior mean, confidence intervals (based on regularized estimation) and so forth. Next, we generate $n = 100$ observations from the above model and first estimate the activity graph $\mathcal{G}$ associated with $\mathbf{A}_1$ by the Bayesian SSS algorithm with $q_1 = q_2 = \frac{1}{p}$ and $\tau^2 = \frac{\log n}{2}$. These two choices are discussed later in this section. The nonzero entries are then estimated via their posterior means, as discussed in Section 5 above. We repeat this process 500 times. In one of 500 repetitions, the estimated $\hat{\mathbf{A}}_1$ by the posterior mean is given by

$$\widehat{(\mathbf{A}_1)} = \begin{pmatrix} 1.318\textcircled{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.007\textcircled{2} & 0 & 0 \\ 0 & 0 & -1.853\textcircled{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.978\textcircled{4} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.336\textcircled{5} \end{pmatrix}.$$

It can be seen that the estimates are very close to the true ones. Next, the average estimation error ($\|\hat{\mathbf{A}}_1 - \mathbf{A}_1\|$), specificity and sensitivity are reported in the comparison table (Table 1), based on these 500 replicates.

TABLE 1
Performance comparison between VAR-SSS and ℓ_1-LL

	VAR-SSS	ℓ_1-LL
Est. Error	0.77	0.81
Sensitivity	0.98	0.97
Specificity	0.99	0.97

TABLE 2
95% coverage of confidence and credible interval

Coef. serial no.	C.I.	Credible region
①	83.5	88
②	91	95.5
③	90	96.5
④	88.5	90.5
⑤	91	94

$$\text{Sensitivity (SN)} = \frac{\text{True Positive (TP)}}{\text{TP} + \text{False Negative (FN)}},$$

$$\text{Specificity (SP)} = \frac{\text{True Negative (TN)}}{\text{TN} + \text{False Positive (FP)}}.$$

From the above table, it is quite clear that the VAR Shotgun Stochastic Search algorithm performs better than the frequentist ℓ_1-LL in terms of support recovery and the posterior mean is more accurate than the ℓ_1 penalized estimate of $\hat{\mathbf{A}}_1$. The ℓ_1 penalized log-likelihood estimation (ℓ_1-LL) (Lin and Michailidis (2017)) of $\mathbf{\Phi}$ is given by

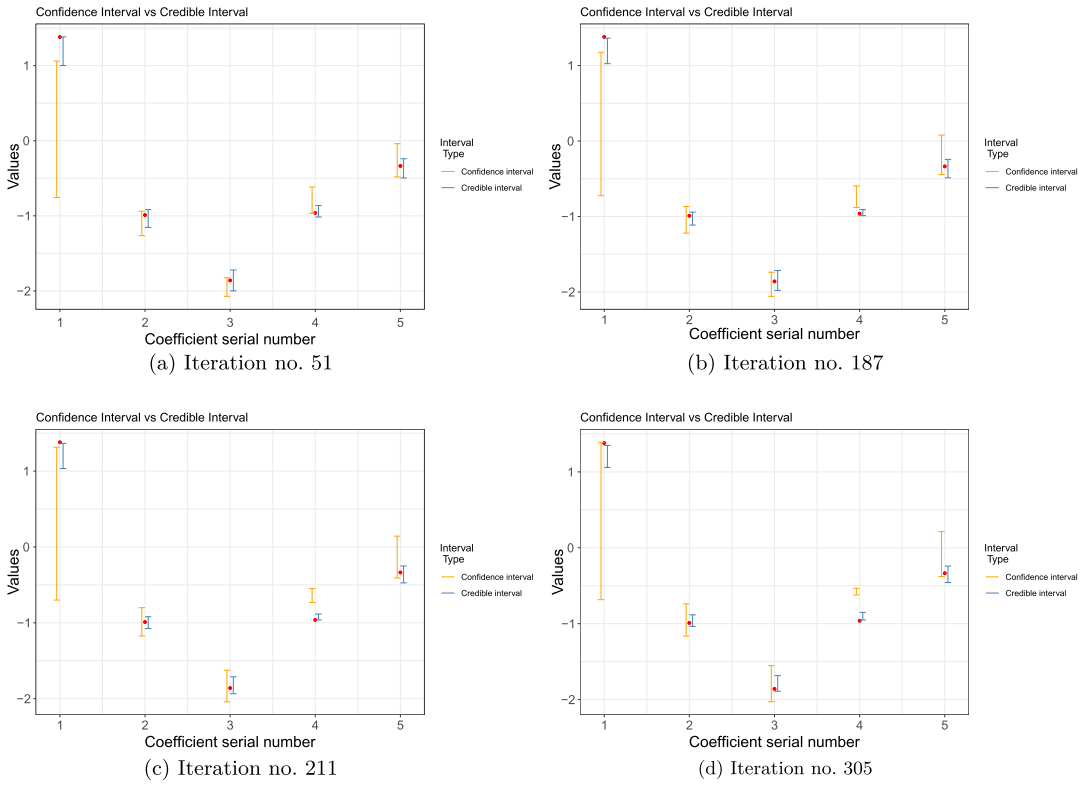
$$\arg\min \frac{1}{n} \big\| \hat{\mathbf{\Sigma}}_\epsilon^{-1/2} (\mathbf{Y} - \mathbf{X}\mathbf{\Phi}) \big\|_F^2 + \lambda_n \big\| \text{vec}(\mathbf{\Phi}) \big\|_1.$$

This gives the maximum likelihood estimate of $\mathbf{\Phi}$, based on an estimated sparse $\mathbf{\Sigma}_\epsilon$, the latter obtained using the graphical lasso algorithm. The regularization parameter λ_n for both the ℓ_1-LS and ℓ_1-LL methods is chosen based on the BIC criterion.

Next, we examine posterior credible intervals for the 6 circled entries, together with frequentist confidence intervals obtained from an ℓ_1-penalized likelihood. Since the error process in this setting has covariance $\mathbf{\Sigma}_\epsilon = \sigma^2 \mathbf{I}_p$ estimation of the VAR parameters can be obtained by running p separate regressions (see Lin and Michailidis (2017)) of the form $\boldsymbol{y}_j = \mathbf{X}\boldsymbol{\phi}_j + \xi_j$, $1 \le j \le p$, where $\boldsymbol{y}_j$ and $\boldsymbol{\phi}_j$ are the j-th columns of the response matrix $\mathbf{Y}$ and parameter matrix $\mathbf{\Phi}$, respectively. The confidence intervals are computed based on the recent paper by Taylor and Tibshirani (2015) in the context of linear regression using their R-package `selectiveInference`. Note that as previously mentioned, the exact form of such confidence intervals for temporally dependent data is not yet available in the literature, but nevertheless provide some rough guidance about uncertainty of the VAR model parameters. We randomly select 4 instances out of 500 iterations and the coverage of both the credible and confidence intervals for the 5-circled entries (95%) is depicted in Figure 1. It can be seen that the true parameters are very close to the center of the credible interval for 4 coefficients and only fail to cover one of them. This is not the case for the confidence intervals, which in addition fail to cover 2 of the coefficients. The upshot is that with current technology, the obtained credible intervals offer a good measure of quantifying uncertainty in sparse high-dimensional VAR models. In the following table (Table 2), we report 95-coverage percentage (i.e., % of how many times the corresponding interval contains the true parameter) using 500 replications.

The results in Table 2 indicate that posterior credible region has significantly better coverage percentage compared to debiased confidence intervals in the VAR setting.

Finally, before proceeding to study the performance of VAR-SSS in detail, we investigate the scalability and computational speed of the algorithm. One of the attractive features of shotgun stochastic search procedure is that Step 1 of the algorithm can be extensively parallelized and there is no sampling involved from conditional distribution. Both of which



(a) Iteration no. 51 (b) Iteration no. 187

(c) Iteration no. 211 (d) Iteration no. 305

FIG. 1. *Comparison between posterior credible and confidence intervals of the* 5 *circled VAR coefficients. The true values are denoted by the red dot.*

will reduce total runtime significantly. In Table 3, we report computational times (averaged over 10 iterations) for both the cases when Step 1 has been executed with parallelization (6 cores) and without parallelization, using the Ubuntu 18.04 OS operating system on a machine equipped with a 2.9 GHz Intel Core i7 8750H processor and 16 GB of total memory.

Model selection performance. Next, we turn our attention to model selection and estimation accuracy of the developed Bayesian procedure. We consider 4 different size VAR models: Small VAR, Medium VAR, Big VAR and Large VAR corresponding to $p = 100, 250, 500, 750$, and lag $d = 1$, and $p = 50, 125, 250, 400$ with lag $d = 2$, respectively. For each setting, motivated by the Erdős–Rényi example in the Supplementary Material, Section S6, we generate transition matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ with nonzero entries drawn from $\text{Unif}(0, 10) \cup \text{Unif}(-10, 0)$ and rescaled to ensure that the process is stable with SNR $= 2$. The spectral radii of the $\mathbf{A}_i$'s are set to 0.90. Further, for each p-dimensional VAR model

TABLE 3
Average runtime of VAR-SSS in minutes

	No parallelization	With parallelization
$p = 50$	0.12	0.04
$p = 100$	0.44	0.13
$p = 250$	2.81	1.08
$p = 500$	11.23	3.59
$p = 750$	26.74	11.83

TABLE 4
Sample sizes n_1, n_2 and n_3 for each VAR models

	Sample sizes			
	Dimension p	n_1	n_2	n_3
Small VAR	100	25	50	75
Medium VAR	250	30	75	200
Big VAR	500	50	150	300
Large VAR	750	75	200	400

(with lag $d = 1, 2$) the true transition matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ have 3 different edge densities—$e_1 = 1/10p$, $e_2 = 1/8p$, $e_3 = 1/6p$. Next, we generate samples with 3 different sample sizes from the model $X^{t+1} = \mathbf{A}_1 X^t + \boldsymbol{\varepsilon}^t$ where $\boldsymbol{\varepsilon}^t \overset{\text{i.i.d}}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with $\boldsymbol{\Sigma}_\epsilon = \text{Toep}(\rho = 0.50)$. The sample sizes n_1, n_2 and n_3 corresponding to different VAR models are given in Table 4.

For the VAR-SSS algorithm, the initial activity graph $\mathcal{G}_0$ is selected based on an ℓ_1 penalized least squares estimate (ℓ_1-LS) which does not use $\boldsymbol{\Sigma}_\epsilon$ and is given by

$$\arg\min \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi}\|_F^2 + \lambda_n \|\text{vec}(\boldsymbol{\Phi})\|_1.$$

Note that the Bayesian SS algorithm has several hyperparameters—τ^2, q, α_i and β_i for $1 \le i \le p$. Throughout this analysis, we set the slab variance to $\tau^2 = \frac{\log n}{2}$ and $\alpha_i = 1, \beta_i = 2 \forall i$. The prior edge probability q is set to $\frac{1}{p}$. This particular choice is not optimized for any given problem, but it does reasonably well and intuitively follows the assumption regarding the magnitude of q that is needed for theoretical guarantees.

We finally compare the performance of VAR-SSS to the ℓ_1-LL (penalized log-likelihood estimation as described earlier). All results reported in the subsequent tables are based on 200 replications. We use sensitivity (SN), specificity (SP) and relative error as the criteria to evaluate the performance of the support recovery and estimation quality of transition matrices $\mathbf{A}_i$. Since the exact contemporaneous dependence is not of primary concern, we omit the numerical results for $\boldsymbol{\Sigma}_\epsilon$. The results in Table 5 illustrate the selection performance between the VAR-SSS and ℓ_1-LL methods with a contemporaneous correlated error structure.

It can be seen that the VAR-SSS algorithm *matches* the performance of the maximum likelihood method (ℓ_1-LL) across all settings. Further, as the sample size increases both the sensitivity and specificity metrics improve. Finally, the estimates of lag $d = 2$ parameters are slightly worse than those of the lag $d = 1$ parameters for both methods.

Estimation consistency. In Table 6, we present the estimation error in Frobenius norm, that is, $\|\boldsymbol{\Phi}_0 - \hat{\boldsymbol{\Phi}}_1\|_F$ for both the VAR-SSS (posterior mean) and ℓ_1-LL estimates.

It can be seen that the estimation error decreases with an increase in the number of time points (sample size) n; further, its values are significantly larger in big and large size VAR models than in small VAR ones. Regarding the level of sparsity in the true transition matrices, the results show that for fixed n and p, the more true nonzero entries in $\mathbf{A}_1$, the less accurate the posterior mean and ℓ_1-LL estimate become. However, both of them perform equally well. Next, similar to Tables 5 and 6, we present model selection and estimation performance of the VAR-SSS by making the spectral radius (of the true transition matrices $\mathbf{A}_1$ and $\mathbf{A}_2$) smaller and setting it equal to 0.70.

With reduced spectral radius the entries in $\mathbf{A}_1$ and $\mathbf{A}_2$ are of smaller magnitude, and thus in Tables 7 and 8 the performance of both the VAR-SSS and ℓ_1-LL deteriorates compared to the results depicted in Tables 5 and 6 in terms of model selection and estimation accuracy.

TABLE 5
Sensitivity and Specificity in VAR$(d = 1, 2)$ *with* $\mathbf{\Sigma}_\epsilon = \text{Toep}(\rho = 0.80)$

		VAR-SSS						ℓ_1-LL					
		n_1		n_2		n_3		n_1		n_2		n_3	
	Density	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP
						lag $d = 1$							
Small VAR	e_1	0.93	0.93	0.93	0.95	0.93	0.96	0.90	0.94	0.92	0.93	0.93	0.96
	e_2	0.93	0.92	0.93	0.92	0.94	0.95	0.91	0.92	0.93	0.92	0.93	0.93
	e_3	0.91	0.92	0.93	0.93	0.93	0.95	0.90	0.92	0.93	0.93	0.93	0.95
Medium VAR	e_1	0.972	0.92	0.93	0.92	0.96	0.96	0.93	0.93	0.91	0.93	0.94	0.95
	e_2	0.92	0.93	0.91	0.92	0.94	0.93	0.91	0.92	0.93	0.92	0.94	0.95
	e_3	0.91	0.92	0.93	0.93	0.94	0.95	0.92	0.92	0.93	0.93	0.93	0.96
Big VAR	e_1	0.91	0.93	0.91	0.93	0.92	0.95	0.91	0.93	0.90	0.93	0.92	0.96
	e_2	0.90	0.92	0.91	0.92	0.94	0.93	0.89	0.92	0.91	0.92	0.91	0.95
	e_3	0.90	0.93	0.90	0.93	0.93	0.94	0.89	0.92	0.90	0.93	0.91	0.95
Large VAR	e_1	0.91	0.93	0.92	0.93	0.93	0.96	0.90	0.93	0.91	0.93	0.93	0.94
	e_2	0.91	0.92	0.92	0.92	0.94	0.95	0.90	0.92	0.91	0.92	0.93	0.93
	e_3	0.89	0.91	0.90	0.93	0.92	0.95	0.91	0.92	0.91	0.93	0.92	0.93
						lag $d = 2$							
Small VAR	e_1	0.93	0.93	0.93	0.94	0.94	1	0.92	0.93	0.92	0.93	0.93	0.96
	e_2	0.92	0.92	0.93	0.93	0.94	0.93	0.92	0.92	0.93	0.92	0.93	0.93
	e_3	0.92	0.91	0.93	0.93	0.92	0.95	0.93	0.92	0.92	0.92	0.91	0.95
Medium VAR	e_1	0.92	0.92	0.93	0.98	0.90	0.95	0.93	0.98	0.90	0.93	0.99	0.95
	e_2	0.92	0.93	0.92	0.92	0.92	0.93	0.96	0.92	0.98	0.92	0.98	0.95
	e_3	0.91	0.92	0.93	0.93	0.93	0.94	0.92	0.91	0.98	0.98	0.91	0.94
Big VAR	e_1	0.91	0.93	0.91	0.92	0.92	0.95	0.90	0.92	0.91	0.92	0.91	0.95
	e_2	0.90	0.92	0.91	0.92	0.94	0.96	0.90	0.92	0.91	0.92	0.90	0.96
	e_3	0.90	0.93	0.90	0.93	0.95	0.94	0.89	0.92	0.90	0.92	0.91	0.95
Large VAR	e_1	0.91	0.93	0.92	0.93	0.93	0.96	0.90	0.93	0.91	0.92	0.93	0.95
	e_2	0.90	0.92	0.92	0.92	0.95	0.95	0.90	0.92	0.91	0.92	0.93	0.93
	e_3	0.89	0.91	0.89	0.93	0.92	0.95	0.90	0.92	0.91	0.93	0.92	0.93

7. Extension to stochastic linear regression. In this section, we consider a linear regression setting where both the predictors and the errors exhibit *temporal dependence* and are generated by a stationary Gaussian process, but remain *independent* of each other. We employ our pseudo-likelihood based methodology to derive high-dimensional consistency results in this setting. Suppose we have data on n time points, each consisting of $p = p_n$ regressors $\{x_{1,i}, x_{2,i}, \ldots, x_{p,i}\}$ and a response y_i, $i = 1, 2, \ldots n$. The $n \times 1$ response vector $\boldsymbol{y}$ is modeled as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where } \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma}_\epsilon), \tag{17}$$

with $\mathbf{X}$ being the $n \times p$ design matrix whose ith row $X^i = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})'$ contains the predictors for the ith observation, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of corresponding regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n$ is the vector of error terms. We assume that the vectors $\{X^i\}$ come from a p-variate covariance-stationary Gaussian process with finite spectral density $f_X(\theta)$, that is, the predictors are identically (but not necessarily independently) distributed. Further, we assume the error process $\{\varepsilon_i\}$ to be a univariate centered, covariance-stationary Gaussian process, with finite spectral density $f_\varepsilon(\theta)$. In addition, the errors are assumed to be *independent* of the predictors.

TABLE 6
Estimation error: $\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_F$ *in VAR* ($d = 1, 2$) *with* $\mathbf{\Sigma}_\epsilon = \text{Toep}(\rho = 0.80)$

		VAR-SSS			ℓ_1-LL		
	Density	n_1	n_2	n_3	n_1	n_2	n_3
		Lag $d = 1$					
Small VAR ($p = 100$)	e_1	0.88	0.79	0.7	0.88	0.78	0.7
	e_2	0.9	0.8	0.7	0.91	0.81	0.71
	e_3	0.95	0.85	0.76	0.96	0.86	0.75
Medium VAR ($p = 250$)	e_1	0.9	0.81	0.71	0.91	0.81	0.71
	e_2	0.93	0.84	0.75	0.92	0.83	0.73
	e_3	1.1	1	0.89	1.1	1	0.9
Big VAR ($p = 500$)	e_1	0.95	0.85	0.77	0.96	0.86	0.75
	e_2	1.1	1	0.91	1.1	1.01	0.9
	e_3	1.79	1.7	1.58	1.77	1.67	1.56
Large VAR ($p = 750$)	e_1	0.93	0.84	0.74	0.92	0.81	0.73
	e_2	1.32	1.23	1.11	1.29	1.19	1.08
	e_3	1.92	1.82	1.72	1.9	1.8	1.69
		Lag $d = 2$					
Small VAR ($p = 50$)	e_1	0.91	0.82	0.73	0.90	0.8	0.72
	e_2	0.93	0.84	0.74	0.94	0.84	0.75
	e_3	0.98	0.88	0.79	0.99	0.9	0.79
Medium VAR ($p = 125$)	e_1	0.93	0.85	0.75	0.94	0.84	0.74
	e_2	0.96	0.87	0.78	0.94	0.85	0.75
	e_3	1.13	1.03	0.93	1.13	1.04	0.94
Big VAR ($p = 250$)	e_1	0.98	0.89	0.8	0.99	0.89	0.79
	e_2	1.13	1.03	0.94	1.13	1.04	0.93
	e_3	1.81	1.73	1.61	1.79	1.7	1.59
Large VAR ($p = 400$)	e_1	0.96	0.88	0.77	0.94	0.84	0.75
	e_2	1.34	1.25	1.13	1.31	1.21	1.11
	e_3	1.95	1.86	1.76	1.92	1.82	1.72

We now consider a high-dimensional setting for this linear regression model, where $p \gg n$. Let γ_i indicate whether the ith covariate is active in the model or not, that is, $\gamma_i = \mathbb{1}(\beta_i \neq 0)$. The vector $\boldsymbol{\gamma}$ indicates which covariates are important in predicting the response, and we refer to it as the active covariate vector. Our goal is to identify the nonzero coefficients to learn about the active covariates, using the Gaussian likelihood

$$L(\boldsymbol{\beta}, \mathbf{\Sigma}_\epsilon \mid y_i, i = 1, \ldots, n) := \frac{1}{\sqrt{2\pi |\mathbf{\Sigma}_\epsilon|}^n} \exp\Big(-\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Sigma}_\epsilon^{-1} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\Big). \tag{18}$$

The parameter of interest is the regression coefficient $\boldsymbol{\beta}$ and we treat the error covariance matrix $\mathbf{\Sigma}_\epsilon$ as an unknown nuisance parameter. Note that all the diagonals of $\mathbf{\Sigma}_\epsilon$ are the same and denoted by $\sigma_{\boldsymbol{\varepsilon}}^2$. As in the case of the VAR model, we construct the following pseudo-likelihood by assuming $\mathbf{\Sigma}_\epsilon$ is diagonal (with all diagonal entries equal to $\sigma_{\boldsymbol{\varepsilon}}^2$ due to stationarity of the error process):

$$L_{\text{pseudo}}(\boldsymbol{\beta}, \mathbf{\Sigma}_\epsilon \mid y_i, i = 1, \ldots, n) = \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{\varepsilon}}^2}^n} \exp\Bigg(-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2} \sum_{i=1}^n (y_i - \mathbf{X}_i'\boldsymbol{\beta})^2\Bigg). \tag{19}$$

The function L_{pseudo} in (19) has a much simpler form than the likelihood function L and only involves $\sigma_{\boldsymbol{\varepsilon}}^2$ (the diagonal entry of $\mathbf{\Sigma}_\epsilon$). Let ν_γ denote the number of nonzero entries in

TABLE 7
Sensitivity and Specificity in VAR($d = 1, 2$) *with* $\mathbf{\Sigma}_\epsilon = \text{Toep}(\rho = 0.80)$

		VAR-SSS						ℓ_1-LL					
		n_1		n_2		n_3		n_1		n_2		n_3	
	Density	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP
						lag $d = 1$							
Small VAR	e_1	0.96	0.98	0.97	0.99	0.98	1	0.97	0.96	0.97	0.97	0.98	1
	e_2	0.96	0.97	0.97	0.98	0.98	1	0.96	0.96	0.97	0.97	0.98	0.98
	e_3	0.94	0.95	0.95	0.96	0.96	0.98	0.96	0.96	0.97	0.97	0.98	0.99
Medium VAR	e_1	0.96	0.97	0.97	0.98	0.98	1	0.96	0.96	0.97	0.97	0.98	1
	e_2	0.95	0.97	0.96	0.99	0.97	1	0.96	0.97	0.98	0.98	1	0.99
	e_3	0.94	0.95	0.95	0.97	0.96	0.99	0.96	0.96	0.97	0.97	0.98	0.99
Big VAR	e_1	0.94	0.95	0.95	0.96	0.96	0.97	0.95	0.95	0.96	0.96	0.97	0.97
	e_2	0.94	0.95	0.95	0.96	0.96	0.97	0.95	0.95	0.96	0.96	0.98	0.98
	e_3	0.94	0.93	0.95	0.94	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.97
Large VAR	e_1	0.95	0.97	0.96	0.99	0.97	1	0.95	0.96	0.97	0.97	0.99	0.98
	e_2	0.93	0.93	0.94	0.95	0.96	0.97	0.95	0.95	0.96	0.96	0.97	0.97
	e_3	0.92	0.91	0.93	0.93	0.94	0.94	0.94	0.93	0.94	0.94	0.95	0.95
						lag $d = 2$							
Small VAR	e_1	0.92	0.95	0.93	0.96	0.94	0.97	0.93	0.91	0.92	0.92	0.93	0.93
	e_2	0.92	0.94	0.93	0.95	0.94	0.97	0.92	0.91	0.92	0.92	0.93	0.94
	e_3	0.9	0.92	0.91	0.93	0.92	0.95	0.92	0.91	0.92	0.92	0.93	0.94
Medium VAR	e_1	0.92	0.94	0.93	0.95	0.94	0.97	0.92	0.91	0.92	0.92	0.93	0.94
	e_2	0.91	0.94	0.92	0.96	0.93	0.98	0.92	0.92	0.93	0.93	0.95	0.94
	e_3	0.9	0.92	0.91	0.94	0.93	0.96	0.92	0.91	0.92	0.92	0.93	0.94
Big VAR	e_1	0.91	0.93	0.92	0.94	0.93	0.95	0.91	0.92	0.93	0.93	0.94	0.94
	e_2	0.91	0.93	0.92	0.94	0.93	0.95	0.91	0.92	0.93	0.93	0.95	0.95
	e_3	0.91	0.91	0.92	0.92	0.93	0.94	0.92	0.92	0.93	0.93	0.94	0.94
Large VAR	e_1	0.92	0.95	0.93	0.97	0.95	0.99	0.91	0.93	0.94	0.94	0.96	0.95
	e_2	0.9	0.91	0.91	0.93	0.93	0.95	0.91	0.92	0.93	0.93	0.94	0.94
	e_3	0.89	0.89	0.90	0.91	0.92	0.92	0.90	0.90	0.91	0.91	0.93	0.92

γ. We impose the following prior on $\boldsymbol{\beta}, \sigma_{\boldsymbol{\varepsilon}}^2$ and $\boldsymbol{\gamma}$:

$$\begin{aligned} \beta_j \mid \boldsymbol{\gamma}, \sigma_{\boldsymbol{\varepsilon}}^2 &\sim (1-\gamma_j)1_{\beta_j=0} + \gamma_j \mathcal{N}\big(0, \sigma_{\boldsymbol{\varepsilon}}^2 \tau^2\big) \quad \text{independently for } j = 1, 2, \ldots, p, \\ \pi(\boldsymbol{\gamma}) &\propto q_1^{\nu_\gamma}(1-q_1)^{p-\nu_\gamma} 1_{\{\nu_\gamma < M\}} + q_2^{\nu_\gamma}(1-q_2)^{p-\nu_\gamma} 1_{\{\nu_\gamma \geq M\}}, \\ \sigma_{\boldsymbol{\varepsilon}}^2 &\sim \text{Inv. Gamma}(\alpha, \beta/2). \end{aligned} \tag{20}$$

Similar to the Bayesian VAR model, we refer to M as the *realistic model cutoff size*, and (q_1, q_2) as the *edge inclusion probabilities*. Based on the pseudo-likelihood in (19) and the prior in (20), the pseudo-posterior can easily be calculated and is available in closed form:

$$\pi_{\text{pseudo}}(\boldsymbol{\gamma} \mid \mathscr{Y}) \propto (1-q_{\nu_\gamma})^{p-\nu_\gamma} \left(\frac{q_{\nu_\gamma}}{\tau\sqrt{n}}\right)^{\nu_\gamma} \left|\frac{\mathbf{X}_\gamma' \mathbf{X}_\gamma}{n} + \frac{\mathbf{I}_{\nu_\gamma}}{n\tau^2}\right|^{-1/2} \left(S_\gamma + \frac{\beta}{n}\right)^{-(\frac{n}{2}+\alpha)}, \tag{21}$$

where $q_{\nu_\gamma} = q_1$ if $\nu_\gamma \leq M$, and q_2 if $\nu_\gamma > M$. $\mathbf{X}_\gamma$ represents the sub-matrix of $\mathbf{X}$ including columns corresponding to the active indices in $\boldsymbol{\gamma}$, $S_\gamma := \frac{\mathbf{y}'\mathbf{y}}{n} - \frac{\mathbf{y}'\mathbf{X}_\gamma}{n}(\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{n} + \frac{\mathbf{I}_{\nu_\gamma}}{n\tau^2})^{-1}\frac{\mathbf{X}_\gamma'\mathbf{y}}{n}$. The posterior probabilities $\pi_{\text{pseudo}}(\boldsymbol{\gamma} \mid \mathscr{Y})$ can be used to select the active covariates by computing the posterior mode defined by $\hat{\boldsymbol{\gamma}} := \arg\max_{\boldsymbol{\gamma}} \pi_{\text{pseudo}}(\boldsymbol{\gamma} \mid \mathscr{Y})$.

TABLE 8
Estimation error: $\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_F$ *in VAR* $(d = 1, 2)$ *with* $\mathbf{\Sigma}_\epsilon = \text{Toep}(\rho = 0.80)$

		VAR-SSS			ℓ_1-LL		
	Density	n_1	n_2	n_3	n_1	n_2	n_3
		Lag $d = 1$					
Small VAR ($p = 100$)	e_1	0.89	0.8	0.71	0.9	0.8	0.72
	e_2	0.92	0.82	0.72	0.93	0.83	0.74
	e_3	0.97	0.86	0.78	0.99	0.89	0.78
Medium VAR ($p = 250$)	e_1	0.92	0.82	0.73	0.93	0.83	0.73
	e_2	0.94	0.86	0.76	0.94	0.85	0.75
	e_3	1.12	1.01	0.91	1.12	1.02	0.92
Big VAR ($p = 500$)	e_1	0.97	0.87	0.78	0.99	0.89	0.78
	e_2	1.12	1.01	0.92	1.12	1.02	0.91
	e_3	1.8	1.71	1.59	1.78	1.68	1.58
Large VAR ($p = 750$)	e_1	0.94	0.86	0.75	0.94	0.83	0.75
	e_2	1.34	1.24	1.13	1.31	1.21	1.1
	e_3	1.93	1.83	1.73	1.92	1.81	1.71
		Lag $d = 2$					
Small VAR ($p = 50$)	e_1	0.91	0.82	0.73	0.91	0.82	0.74
	e_2	0.94	0.84	0.74	0.95	0.84	0.75
	e_3	0.99	0.88	0.8	1	0.9	0.79
Medium VAR ($p = 125$)	e_1	0.94	0.84	0.75	0.95	0.84	0.74
	e_2	0.96	0.88	0.78	0.96	0.87	0.77
	e_3	1.14	1.03	0.93	1.13	1.04	0.93
Big VAR ($p = 250$)	e_1	0.99	0.89	0.8	1	0.9	0.79
	e_2	1.14	1.03	0.94	1.13	1.04	0.93
	e_3	1.82	1.73	1.61	1.8	1.7	1.59
Large VAR ($p = 400$)	e_1	0.96	0.88	0.77	0.96	0.85	0.76
	e_2	1.36	1.26	1.15	1.32	1.22	1.12
	e_3	1.95	1.85	1.75	1.93	1.83	1.72

Next, we examine the selection and estimation consistency for the pseudo-likelihood based approach described above in a high-dimensional setting and allow $p = p_n$ to grow with n. For each n, the data $\boldsymbol{y}_n = (y_{i,n})_{i=1}^n$ is assumed to be generated from a true model with regression coefficient vector $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,n}$, and corresponding active covariate vector $\boldsymbol{\gamma}_0$ such that

$$\boldsymbol{y}_n = \mathbf{X}_n \boldsymbol{\beta}_{0,n} + \boldsymbol{\varepsilon}_n, \qquad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}).$$

The rows of $\mathbf{X} = \mathbf{X}_n$ are generated from a p-variate covariance-stationary Gaussian process with spectral density $f_X(\theta)$, and the entries of $\boldsymbol{\varepsilon}_n$ are generated from a centered univariate stationary Gaussian process with spectral density $f_{\boldsymbol{\varepsilon}}(\theta)$. Let $\nu_0 = \|\boldsymbol{\beta}_{0,n}\|_0$ (i.e., the number of nonnull entries in $\boldsymbol{\beta}_{0,n}$) and the minimum signal strength is represented respectively by $s_n := \inf_{i:\beta_{0_i} \neq 0} |\beta_{0_i}|$. Let $\Omega_{0,n}$ denote the common covariance matrix of the rows of $\mathbf{X}_n$, and $\sigma^2_{0,n}$ denote the common variance of each coordinate of the stationary error process $\boldsymbol{\varepsilon}_n$. For ease of exposition, we will refer to $\boldsymbol{y}_n$, $\mathbf{X}_n$, $\boldsymbol{\beta}_{0,n}$, $\boldsymbol{\varepsilon}_n$, $\Omega_{0,n}$, $\sigma^2_{0,n}$ as $\boldsymbol{y}$, $\mathbf{X}$, $\boldsymbol{\beta}_0$, $\boldsymbol{\varepsilon}$, Ω_0, σ^2_0, respectively.

High-dimensional consistency results for the lasso for linear regression with dependent errors and predictors have been recently established in Basu and Michailidis (2015). To the best of our knowledge, existing work on high-dimensional selection consistency in Bayesian linear regression with spike-and-slab priors (George and McCulloch (1997), Ishwaran and

Rao (2005)), Narisetty and He (2014)) assumes the errors (and also the predictors) to be independent and identically distributed. We consider a more general setting by allowing *both* the predictors and the errors to be drawn from a temporally dependent process. We provide the main consistency results for stochastic regression below. The proofs of these results are delegated to Supplementary Material, Section S8.

THEOREM 7.1 (Strong selection consistency). *For a centered stochastic linear regression model* (17) *with prior* (20) *on* $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_{\boldsymbol{\varepsilon}}^2$ *satisfying Assumptions B*1–*B*5 (*provided in Supplementary Material*, *Section S*7), *the following holds*:

$$\pi_{\text{pseudo}}(\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 \mid \mathscr{Y}) \xrightarrow{\mathbb{P}_0} 1 \quad \text{as } n \to \infty.$$

In particular, the true active covariate vector $\boldsymbol{\gamma}_0$ is the mode of the (pseudo) posterior distribution with probability tending to 1 as $n \to \infty$.

THEOREM 7.2 (Estimation consistency rate without minimum signal size assumption). *For a centered linear stochastic regression model* (17) *with prior* (20) *on* $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\epsilon$ *satisfying Assumptions B*1–*B*4 (*provided in Supplementary Material*, *Section S*7), *there exists a constant* K (*not depending on* n) *such that*

$$\mathbb{E}_0\left[\Pi_{\text{pseudo}}\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > K\sqrt{(1+\mathscr{M}(f_\varepsilon))\frac{v_0 \log p}{n}}\,\middle|\,\mathscr{Y}\right)\right] \to 0$$

$$\text{as } n \to \infty.$$

REMARK 7.3. The estimation consistency rate for stochastic regression (in Theorem 7.2) is of the order of $\sqrt{(1+\mathscr{M}(f_\varepsilon))v_0 \log p/n}$. In other words, the radius of the neighborhood around $\boldsymbol{\beta}_0$ on which the posterior asymptotically places all of its mass is of the order of $\sqrt{(1+\mathscr{M}(f_\varepsilon))v_0 \log p/n}$, which only includes the stability measure $\mathscr{M}(f_\varepsilon)$ of the error process, and does not include the stability measure $\mathscr{M}(f_X)$ of the predictor process. However, from the proof of Theorem 7.2, it follows that the speed with which the posterior probability of this neighborhood approaches 1 depends on this stability measure. Nevertheless, Theorem 7.2 is a stronger result than Theorem 4.2, the corresponding estimation consistency result for the VAR model; in the latter, the estimation consistency rate involves the stability measures of the underlying true VAR process (which are connected to the stability of the predictor process in the regression interpretation of the VAR in (11)).

Note that in the stochastic regression model (17) the error process and the predictor process are independent of each other, while in the VAR model (11) the error process and the predictor process are dependent. This is the key reason for the stronger results that can be obtained for the stochastic regression model under weaker assumptions. We illustrate this by focusing on one of the important steps in the estimation consistency proof, which involves bounding $\|\boldsymbol{\varepsilon}'\mathbf{X}_t\|/n$. Using the independence of $\boldsymbol{\varepsilon}$ and $\mathbf{X}_t$ in the stochastic regression model, the term $\boldsymbol{\varepsilon}'\mathbf{X}_t(\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t'\boldsymbol{\varepsilon}$ can be bounded by an appropriate χ^2 random variable, and this can be used to show that $\|\boldsymbol{\varepsilon}'\mathbf{X}_t\|/n = o(\sqrt{\mathscr{M}(f_\varepsilon)v_0 \log p/n})$. On the other hand, the lack of independence of the error process and the predictor process in the VAR model necessitates the use of more complex concentration inequalities to bound the analogous quantity $\|\boldsymbol{\xi}_i'\mathbf{X}_i\|/n$. it can be shown that $\|\boldsymbol{\xi}_i'\mathbf{X}_i\|/n = O(\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{v_i \log dp}{n}})$ (see equation (22) in the Appendix).

APPENDIX: PROOFS FOR THEOREMS 4.1, 4.2 AND 4.4

Throughout the proof, for ease of presentation we denote $\pi_{\text{pseudo}}(\cdot \mid \mathcal{Y})$ and $\Pi_{\text{pseudo}}(\cdot \mid \mathcal{Y})$ by $\pi(\cdot \mid \mathcal{Y})$ and $\Pi(\cdot \mid \mathcal{Y})$, respectively. Before presenting the proof of the theorems, we first introduce notation needed in subsequent developments:

- Let $\boldsymbol{t}_i = \{j_1, \ldots, j_{\nu_{t_i}}\} \subset \{1, \ldots, dp\}$ be the set of column indices from $\mathbf{X}$ corresponding to the nonzero positions of $\boldsymbol{\Phi}_{0n,\cdot i}$ in the true model and it represents the neighbors of i in the true activity graph $\mathcal{G}_0$.
- Similarly, $\boldsymbol{m}_i = \{i_1, \ldots, i_{\nu_{m_i}}\}$ represents the same for any candidate model (distinct from the true model). Given two activity graphs $\mathcal{G}_0$ and $\mathcal{G}_m$, $\boldsymbol{t}_i = \boldsymbol{m}_i$ implies the neighbors of i are identical in both of the graphs (i.e., $\mathcal{G}_0$ and $\mathcal{G}_m$ have the same ith column).

$$\mathcal{G}_0 = \begin{pmatrix} & & \overset{\boldsymbol{t}_i}{} & & \\ \cdots & \begin{pmatrix} \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \end{pmatrix} & \cdots \end{pmatrix}_{dp \times p}, \qquad \mathcal{G}_m = \begin{pmatrix} & & \overset{\boldsymbol{m}_i}{} & & \\ \cdots & \begin{pmatrix} \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \end{pmatrix} & \cdots \end{pmatrix}_{dp \times p}$$

- Let us define $d(\boldsymbol{m}_i, \boldsymbol{t}_i) := \text{Card}(\{\boldsymbol{t}_i \cup \boldsymbol{m}_i\} \setminus \{\boldsymbol{t}_i \cap \boldsymbol{m}_i\})$—number of disagreements in the ith column between $\mathcal{G}_m$ and $\mathcal{G}_0$.
- Total number of disagreements is denoted by $D(m, t)$ and is equal to $\sum_{i=1}^{p} d(\boldsymbol{m}_i, \boldsymbol{t}_i)$.

Let $q_{\nu,n} := q_1 1_{\{\nu < M_n\}} + q_2 1_{\{\nu \geq M_n\}}$. Given the true activity graph $\mathcal{G}_0$ and another arbitrary graph $\mathcal{G}_m$, we have

$$\frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} = \prod_{i=1}^{p} \left(\frac{1}{\tau\sqrt{n}}\right)^{(\nu_{m_i} - \nu_{t_i})} \frac{q_{\nu_{m_i}}^{\nu_{m_i}} (1 - q_{\nu_{m_i}})^{dp - \nu_{m_i}}}{q_{\nu_{t_i}}^{\nu_{t_i}} (1 - q_{\nu_{t_i}})^{dp - \nu_{t_i}}} \frac{|\frac{\mathbf{X}'_{m_i}\mathbf{X}_{m_i}}{n} + \frac{\mathbf{I}_{\nu_{m_i}}}{n\tau^2}|^{-1/2}}{|\frac{\mathbf{X}'_{t_i}\mathbf{X}_{t_i}}{n} + \frac{\mathbf{I}_{\nu_{t_i}}}{n\tau^2}|^{-1/2}} \left(\frac{S_{\nu_{m_i}} + \frac{\beta_i}{n}}{S_{\nu_{t_i}} + \frac{\beta_i}{n}}\right)^{-(\frac{n}{2} + \alpha_i)}.$$

Since $dp^2 q_1 \to 0$ as $n \to \infty$, it follows that

$$\left(\frac{1 - q_2}{1 - q_1}\right)^{dp^2} \leq (1 - q_1)^{-dp^2} \leq e^{dp^2 \frac{q_1}{1 - q_1}} \leq 2$$

for large enough n. Note that $\nu_{t_i} \leq k_n < M_n$ for large enough n, and $q_2 < q_1$. Hence,

$$\begin{aligned} &\frac{1}{2} \frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} \\ (*) \qquad &\leq \prod_{i=1}^{p} \left(\frac{2q_{\nu_{m_i}}}{\tau\sqrt{n}}\right)^{(\nu_{m_i} - \nu_{t_i})} \frac{|\frac{\mathbf{X}'_{m_i}\mathbf{X}_{m_i}}{n} + \frac{\mathbf{I}_{\nu_{m_i}}}{n\tau^2}|^{-1/2}}{|\frac{\mathbf{X}'_{t_i}\mathbf{X}_{t_i}}{n} + \frac{\mathbf{I}_{\nu_{t_i}}}{n\tau^2}|^{-1/2}} \left(\frac{S_{\nu_{t_i}} + \frac{\beta_i}{n}}{S_{\nu_{m_i}} + \frac{\beta_i}{n}}\right)^{(\frac{n}{2} + \alpha_i)} \\ &=: \prod_{i=1}^{p} B(\boldsymbol{m}_i, \boldsymbol{t}_i) \end{aligned}$$

for large enough n. Here, the inequality follows from the fact that $q_1 \to 0$ and $q_2 \to 0$ as $n \to \infty$ and hence for all large n, $\frac{1}{1-q_1}$ and $\frac{1}{1-q_2}$ are bounded below by $\frac{1}{2}$. Note that for $\boldsymbol{t}_i = \boldsymbol{m}_i$, $B(\boldsymbol{m}_i, \boldsymbol{t}_i) = 1$. Recall that for any $\mathbf{m} \subseteq \{1, 2, \ldots, dp\}$, $\mathbf{X}_m$ denotes the submatrix of the columns of $\mathbf{X}$ corresponding to the indices in $\mathbf{m}$.

Before proving the main result (Theorem 4.1)—a straightforward application of Corollary A.4—first note that by Assumption A2 there exist $0 < \sigma_{\min} \le \sigma_{\max} < \infty$ and $0 < \lambda_1 \le \lambda_2 < \infty$ not depending on n such that $\sigma_{\min} < \lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}) \le \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0}) < \sigma_{\max}$, $\lambda_1 < \lambda_{\min}(\mathbf{C}_X)$ and the maximum eigenvalue of any principal submatrix of $\mathbf{C}_X$ of dimension at most k_n is bounded above by λ_2. For ease of presentation, denote $4\pi\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0})\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}$ by $\mathscr{B}_n$. Define next the following events:

$$G_{1,n} := \left\{ \left\| \frac{\mathbf{X}'\mathbf{X}}{n} - \mathbf{C}_X \right\|_{\max} \le \mathscr{B}_n \sqrt{\frac{2\log dp}{n}} \right\},$$

$$G_{2,n} := \left\{ \left\| \frac{\mathbf{X}'\mathbf{E}}{n} \right\|_{\max} \le \mathscr{B}_n \sqrt{\frac{\log dp^2}{n}} \right\},$$

$$\Xi_{1,n} := \bigcap_{i=1}^{p} \left\{ \frac{7\sigma_{\min}}{8} \le \frac{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}{n} \le \frac{3\sigma_{\max}}{2} \right\} \quad \text{and}$$

$$\Xi_{2,n} := \bigcap_{\mathbf{m}:1\le|\mathbf{m}|<2M_n} \left\{ \left\| \frac{\mathbf{X}_m'\mathbf{X}_m}{n} - \boldsymbol{C}_{\mathbf{X}_m} \right\|_2 \le 4\mathscr{B}_n \sqrt{\frac{\nu_m \log dp}{cn}} \right\},$$

where $P_{\mathbf{m}}$ denotes the projection matrix for the column space of $\mathbf{X}_m$, and c is a constant that does not depend on n.

Let $\Xi_n := \Xi_{1,n} \cap \Xi_{2,n}$. We argue below that

$$\mathbb{P}(G_{1,n} \cap G_{2,n} \cap \Xi_n) \to 1 \quad \text{as } n \to \infty. \tag{22}$$

Note that by *Lemma B2*, *Proposition B1* and *Proposition B3* of Ghosh, Khare and Michailidis (2019b), with $\boldsymbol{u} = \mathbf{e}_i$ and $\boldsymbol{v} = \mathbf{e}_j$—the ith and jth unit vector in $\mathbb{R}^{dp}$, respectively, there exists c (not depending on n) such that

$$\mathbb{P}\left(\sup_{1\le i,j\le dp} \left| \mathbf{e}_i' \left(\frac{\mathbf{X}'\mathbf{X}}{n} - \mathbf{C}_X \right) \mathbf{e}_j \right| > \frac{\lambda_{\max}(\mathbf{\Sigma}_\epsilon)}{\mu_{\min}(\tilde{\mathscr{A}})} \eta \right)$$

$$\le 2\exp\left(-cn \min\left\{ \frac{\eta^2}{4}, \frac{\eta}{2} \right\} + 2\log dp \right),$$

and again by taking unit vectors $\boldsymbol{u} = \mathbf{e}_i \in \mathbb{R}^{dp}$ and $\boldsymbol{v} = \mathbf{e}_j \in \mathbb{R}^p$

$$\mathbb{P}\left(\sup_{\substack{1\le i\le dp,\\ 1\le j\le p}} \left| \mathbf{e}_i' \frac{\mathbf{X}'\mathbf{E}}{n} \mathbf{e}_j \right| > 2\pi\lambda_{\max}(\mathbf{\Sigma}_\epsilon) \left[1 + \frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})} \right] \eta \right)$$

$$\le 6\exp\left(-cn \min\left\{ \frac{\eta^2}{4}, \frac{\eta}{2} \right\} + \log dp^2 \right).$$

Next, by setting an appropriate $\eta \sim \sqrt{\frac{\log dp^2}{cn}}$ and under Assumption A1 we have $\mathbb{P}(G_{1,n}) \to 1$ and $\mathbb{P}(G_{2,n}) \to 1$ as $n \to \infty$.

Let $\sigma_{i,0}^2$ denote the ith diagonal entry of the true error covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\varepsilon},0n}$. Since $\boldsymbol{\xi}_i/\sigma_{0,i} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$, by applying the Hanson–Wright inequality of Rudelson and Vershynin (2013) there exists K such that

$$\mathbb{P}\left(\left| \frac{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}{n\sigma_{0,i}^2} - 1 \right| > \frac{1}{8} \right) \le 2\mathrm{e}^{-cn/64K^4}.$$

That is, $\frac{7\sigma_{0,i}^2}{8} \le \frac{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}{n} \le \frac{3\sigma_{0,i}^2}{2}$ happens with probability at least $1 - 2\mathrm{e}^{-cn/64K^4}$. By the union-sum inequality and Assumption B.1, we get $\mathbb{P}(\Xi_{1,n}) \ge 1 - 2p\mathrm{e}^{-cn/64k^4} \to 1$ as $n \to \infty$.

Note that for every $\mathbf{m}$ such that $1 \le |\mathbf{m}| \le 2M_n$, by *Lemma B2* and *Proposition B1* of Ghosh, Khare and Michailidis (2019b), there exists c (not depending on n) such that

$$\mathbb{P}\left(\left\|\frac{\mathbf{X}_m'\mathbf{X}_m}{n} - \boldsymbol{C}_{\mathbf{X}_m}\right\|_2 \ge 4\mathscr{B}_n\sqrt{\frac{\nu_m \log dp}{cn}}\right) \le \mathrm{e}^{-4\nu_m \log dp + 2\nu_m \log 21} \le (dp)^{-3\nu_m} \tag{23}$$

for large enough p. It follows by the union-sum inequality that

$$\begin{aligned}
\mathbb{P}(\Xi_{2,n}^c) &\le \sum_{\mathbf{m}:1<|\mathbf{m}|<2M_n} (dp)^{-3\nu_m} \\
&= \sum_{j=1}^{2M_n} \binom{dp}{j} (dp)^{-3j} \\
&\le \sum_{j=1}^{\infty} (dp)^{-2j} \\
&= \frac{(dp)^{-2}}{1-(dp)^{-2}} \\
&= \frac{1}{(dp)^2 - 1} \to 0 \quad \text{as } n \to \infty
\end{aligned}$$

Hence, $\mathbb{P}(\Xi_{2,n}) \to 1$ as $n \to \infty$.

From now on (unless otherwise mentioned), we restrict attention to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$ and for ease of exposition we omit n from the notation of these sets. Next, we analyze the behavior of $B(\boldsymbol{m}_i, \boldsymbol{t}_i)$ under different scenarios in a sequence of lemmas (Lemmas A.1–A.3). Lemma A.1 studies the scenario where the true active neighbors of i ($\boldsymbol{t}_i$) fully contain the neighbors in the candidate model ($\boldsymbol{m}_i$).

LEMMA A.1. *If $\boldsymbol{m}_i \subset \boldsymbol{t}_i$ then there exists N_1 (not depending on i and $\mathscr{G}_m$) such that for all $n \ge N_1$ we have $B(\boldsymbol{m}_i, \boldsymbol{t}_i) \le {q_{1,n}}^{(\nu_{t_i} - \nu_{m_i})/2}$.*

Lemma A.2 deals with the case when the true active neighbors of i ($\boldsymbol{t}_i$) are fully contained in the neighbors of the candidate model ($\boldsymbol{m}_i$).

LEMMA A.2. *If $\boldsymbol{t}_i \subset \boldsymbol{m}_i$ then there exists N_2 (not depending on i and $\mathscr{G}_m$) such that for all $n \ge N_2$ we have $B(\boldsymbol{m}_i, \boldsymbol{t}_i) \le q_{1,n}^{(\nu_{m_i} - \nu_{t_i})/2}$.*

LEMMA A.3. *If $\mathscr{G}_m$ is such that*

$$\begin{aligned}
&\boldsymbol{t}_i \ne \boldsymbol{m}_i, \\
&\boldsymbol{t}_i \not\subseteq \boldsymbol{m}_i, \\
&\boldsymbol{t}_i \not\supseteq \boldsymbol{m}_i,
\end{aligned}$$

then there exists N_6 (not depending on $\mathscr{G}_m$ and i) such that for all $n \ge N_6$, $B(\boldsymbol{m}_i, \boldsymbol{t}_i) \le q_{1,n}^{\nu_{m_i} - \nu_{t_i}/2}$ if $\nu_{m_i} > \nu_{t_i}$, and $B(\boldsymbol{m}_i, \boldsymbol{t}_i) \le q_{1,n}^{1/2}$ if $\nu_{m_i} \le \nu_{t_i}$.

The proofs of the above lemmas are given in Section S3 of the Supplementary Material. Next, we employ these results to provide a bound on the ratio of the posterior probability for a nontrue activity graph $\mathscr{G}_m$ and the true activity graph $\mathscr{G}_0$.

COROLLARY A.4. *For any centered Gaussian stable VAR(d) model (11) with prior (14) on $\mathbf{\Phi}, \mathcal{G}, \mathbf{\Sigma}_\epsilon$ satisfying Assumptions A1–A4, for any "nontrue" activity graph $\mathcal{G}_m$ with n sufficiently large the following holds*:

$$\frac{\pi_{\text{pseudo}}(\mathcal{G}_m \mid \mathcal{Y})}{\pi_{\text{pseudo}}(\mathcal{G}_0 \mid \mathcal{Y})} \leq (dp^2)^{-2D(m,t)}.$$

Here, $D(m,t)$ denotes the total number of disagreements between the two activity graphs, that is, $D(m,t) = \sum_{i=1}^{p} d(\boldsymbol{m}_i, \boldsymbol{t}_i)$, and $d(\boldsymbol{m}_i, \boldsymbol{t}_i)$ denotes the number of disagreements between $\boldsymbol{m}_i$ and $\boldsymbol{t}_i$.

PROOF. Note that $d(\boldsymbol{m}_i, \boldsymbol{t}_i) \leq 2k_n(\nu_{m_i} - \nu_{t_i})$ for $\nu_{m_i} > \nu_{t_i}$, and $d(\boldsymbol{m}_i, \boldsymbol{t}_i) \leq 2k_n$ for $\nu_{m_i} \leq \nu_{t_i}$. We will assume without loss of generality that the constants $\sigma_{\min}$ and λ_1 in Assumption A2 are bounded above by 1. It follows by Lemmas A.1–A.3 and Assumption A3 that if we restrict to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$ and $\boldsymbol{t}_i \neq \boldsymbol{m}_i$ then for every $n \geq \max\{N_1, \ldots, N_6\}$

$$B(\boldsymbol{m}_i, \boldsymbol{t}_i) \leq (dp^2)^{-2d(\boldsymbol{m}_i, \boldsymbol{t}_i)}.$$

As noted earlier for $\boldsymbol{t}_i = \boldsymbol{m}_i$ we have $B(\boldsymbol{m}_i, \boldsymbol{t}_i) = 1$. Hence, for all sufficiently large n,

$$\frac{1}{2}\frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} \leq \prod_{i=1}^{p} B(\boldsymbol{m}_i, \boldsymbol{t}_i) = \prod_{\boldsymbol{t}_i = \boldsymbol{m}_i} B(\boldsymbol{m}_i, \boldsymbol{t}_i) \prod_{\boldsymbol{t}_i \neq \boldsymbol{m}_i} B(\boldsymbol{m}_i, \boldsymbol{t}_i) \leq (dp^2)^{-2D(m,t)}. \qquad \square$$

A.1. Proof of Theorem 4.1. The proof is a straightforward application of the above corollary. Given the true activity graph $\mathcal{G}_0$, the total number of graphs $\mathcal{G}_m$ such that it differs from $\mathcal{G}_0$ in exactly j places is $\binom{dp^2}{j}$. This is because $\binom{dp^2}{j}$ can be written as $\sum_k \binom{j}{k}\binom{dp^2 - j}{j - k}$. Now for all sufficiently large n,

$$\begin{aligned}
\frac{1 - \pi(\mathcal{G}_0 \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} &= \sum_{\mathcal{G}_m \neq \mathcal{G}_0} \frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} \\
&= \sum_{j=1}^{dp^2} \sum_{\mathcal{G}_m \neq \mathcal{G}_0} \frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} 1_{\{D(m,t)=j\}} \\
&\leq \sum_{j=1}^{dp^2} \binom{dp^2}{j} (dp^2)^{-2j} \\
&\leq \sum_{j=1}^{dp^2} (dp^2)^{-j} \\
&\leq \frac{(dp^2)^{-1}}{1 - (dp^2)^{-1}} \\
&= \frac{1}{dp^2 - 1}.
\end{aligned}$$

It follows that $\frac{1-\pi(\mathcal{G}_0|\mathcal{Y})}{\pi(\mathcal{G}_0|\mathcal{Y})} \to 0$ as $n \to \infty$. $\square$

A.2. Proof of Theorem 4.2. For notational convenience, denote by Π_{pseudo} and π_{pseudo}, Π_n and π_n, respectively. For any $n \times s$ submatrix $\mathbf{X}_s$ of $\mathbf{X}$, $\mathbf{P}_{\nu_s}$ stands for the projection matrix $\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'$ onto $\mathscr{C}(\mathbf{X}_s)$ and by $\tilde{\mathbf{P}}_{\nu_s}$ is defined to be $\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s + \frac{\mathbf{I}_s}{\tau^2})^{-1}\mathbf{X}_s'$. First, note that for any $\eta > 0$,

$$\begin{aligned}
&\mathbb{E}_0\big(\Pi_n\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F \geq K\eta \mid \mathscr{Y}\}\big) \\
&\quad = \sum_{\mathscr{G}} \mathbb{E}_0\big(\Pi_n\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F \geq K\eta \mid \mathscr{Y}, \mathscr{G}\}\pi_n(\mathscr{G} \mid \mathscr{Y})\big) \\
&\quad \leq \mathbb{E}_0\big(\Pi_n\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F \geq K\eta \mid \mathscr{Y}, \mathscr{G}_0\}\big) + \mathbb{E}_0\Pi_n(\mathscr{G} \neq \mathscr{G}_0 \mid \mathscr{Y})
\end{aligned}$$

By Theorem 4.1, it is enough to prove that $\mathbb{E}_0(\Pi_n\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F \geq K\eta \mid \mathscr{Y}, \mathscr{G}_0\}) \to 0$ as $n \to \infty$. Henceforth all the analysis is restricted to the true activity graph $\mathscr{G}_0$. Thus for ease of exposition we shall use $\mathbf{X}_i$, ν_i and $\tilde{\mathbf{P}}_i$ to denote $\mathbf{X}_{t_i}$, ν_{t_i} and $\tilde{\mathbf{P}}_{\nu_{t_i}}$, respectively. Next, under the prior on $\boldsymbol{\Phi}, \mathscr{G}$ and $\boldsymbol{\Sigma}_\epsilon$ in (14) we have

$$\begin{aligned}
&\tilde{\boldsymbol{\phi}}_i | \mathscr{G}_0, \sigma_i^2, \mathscr{Y} \sim \mathscr{N}_{\nu_i}\left(\hat{\boldsymbol{\phi}}_i, \sigma_i^2\left(\mathbf{X}_i'\mathbf{X}_i + \frac{\mathbf{I}_{\nu_i}}{\tau^2}\right)^{-1}\right), \\
&\sigma_i^2 | \mathscr{G}_0, \mathscr{Y} \sim \text{Inv. Gamma}\left(\frac{n}{2} + \alpha_i, \frac{\mathbf{y}_i'(\mathbf{I}_{\nu_i} - \tilde{\mathbf{P}}_{\nu_i})\mathbf{y}_i + \beta_i}{2}\right).
\end{aligned} \tag{24}$$

Also note that

$$\begin{aligned}
&\mathbb{E}_0\left(\Pi_n\left\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F \geq K\frac{1 + \mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\delta_n \log dp}{n}}\middle| \mathscr{Y}, \mathscr{G}_0\right\}\right) \\
&= \mathbb{E}_0\left(\Pi_n\left\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F^2 \geq K^2\left(\frac{1 + \mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\right)^2 \frac{\sum_{i=1}^p \nu_i \log dp}{n}\middle| \mathscr{Y}, \mathscr{G}_0\right\}\right) \\
&= \mathbb{E}_0\left(\Pi_n\left\{\sum_{i=1}^p \|\tilde{\boldsymbol{\phi}}_i - \tilde{\boldsymbol{\phi}}_{0_i}\|_2^2 \geq \sum_{i=1}^p K^2\left(\frac{1 + \mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\right)^2 \frac{\nu_i \log dp}{n}\middle| \mathscr{Y}, \mathscr{G}_0\right\}\right) \\
&\leq p \max_{1 \leq i \leq p} \mathbb{E}_0\left(\Pi_n\left\{\|\tilde{\boldsymbol{\phi}}_i - \tilde{\boldsymbol{\phi}}_{0_i}\|_2 \geq K\frac{1 + \mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\nu_i \log dp}{n}}\middle| \mathscr{Y}, \mathscr{G}_0\right\}\right).
\end{aligned}$$

The proof will conclude by establishing

$$\max_{1 \leq i \leq p} \mathbb{E}_0\left(\Pi_n\left\{\|\tilde{\boldsymbol{\phi}}_i - \tilde{\boldsymbol{\phi}}_{0_i}\|_2 \geq K\frac{1 + \mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\nu_i \log dp}{n}}\middle| \mathscr{Y}, \mathscr{G}_0\right\}\right) \leq 5\mathrm{e}^{-2\log dp}.$$

For ease of exposition, we denote $\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\nu_i \log dp}{n}}$ by $\eta_{n,i}$. since we are only dealing with nonzero components of $\boldsymbol{\phi}_i$ in the true model we simplify further the notation of $\tilde{\boldsymbol{\phi}}_i$ and $\tilde{\boldsymbol{\phi}}_{0_i}$, respectively, by $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_{0_i}$. In order to prove the above claim, we first note that

$$\begin{aligned}
&\mathbb{E}_0\Pi_n(\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\| \geq K\eta_{n,i} \mid \mathscr{Y}, \mathscr{G}_0) \\
&\quad \leq \mathbb{E}_0\Pi_n\left(\|\boldsymbol{\phi}_i - \hat{\boldsymbol{\phi}}_i\| \geq \frac{K\eta_{n,i}}{2}\middle| \mathscr{Y}, \mathscr{G}_0\right) + \mathbb{P}_0\left(\|\hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_{0_i}\| \geq \frac{K\eta_{n,i}}{2}\middle| \mathscr{Y}, \mathscr{G}_0\right)
\end{aligned} \tag{25}$$

First, note that

$$\begin{aligned}
\|\boldsymbol{\phi}_i - \hat{\boldsymbol{\phi}}_i\| &= \left\|\sigma_i\left(\mathbf{X}_i'\mathbf{X}_i + \frac{\mathbf{I}_{\nu_i}}{\tau^2}\right)^{-1/2} \frac{(\mathbf{X}_i'\mathbf{X}_i + \frac{\mathbf{I}_{\nu_i}}{\tau^2})^{1/2}(\boldsymbol{\phi}_i - \hat{\boldsymbol{\phi}}_i)}{\sigma_i}\right\| \\
&\leq \frac{\sigma_i}{\sqrt{\lambda_{\min}(\mathbf{X}_i'\mathbf{X}_i)}}\|z\|,
\end{aligned}$$

where z is $\nu_i \times 1$ standard normal random vector and the last step follows from (24). Thus, for any $M^* > 0$,

$$
\begin{aligned}
&\mathbb{E}_0 \Pi_n \left(\|\boldsymbol{\phi}_i - \hat{\boldsymbol{\phi}}_i\| \geq \frac{K \eta_{n,i}}{2} \middle| \mathscr{Y} \right) \\
&\leq \mathbb{P}\left(\lambda_{\min}\left(\frac{\mathbf{X}_i' \mathbf{X}_i}{n} \right) < \frac{\lambda_1}{2} \right) + \mathbb{P}\left(\|z\| > \sqrt{n} \eta_{n,i} K \sqrt{\frac{\lambda_1}{8(M^*)^2}} \right) \\
&\quad + \mathbb{E}_0 \Pi_n (\sigma_i > M^* \mid \mathscr{Y}).
\end{aligned} \tag{26}
$$

From (23), there exists n_1 such that for all $n \geq n_1$ the first term above is upper bounded by $\mathrm{e}^{-2 \log dp}$. For the third term, recall from the distribution of $\sigma_i^2 | \mathscr{G}_0, \mathscr{Y}$ in (16) that $(\frac{n}{2} + \alpha_i) \sim n$ and for an appropriate constant $c_0 > 0$,

$$
\begin{aligned}
&\frac{\boldsymbol{y}_i'(\mathbf{I}_n - \tilde{\mathbf{P}}_{\nu_i})\boldsymbol{y}_i + \beta_i}{2n} \overset{(i)}{\leq} \frac{\boldsymbol{y}_i'(\mathbf{I}_n - \mathbf{P}_i)\boldsymbol{y}_i + \beta_i}{2n} + \frac{c_0}{2n} \\
&\overset{(ii)}{=} \frac{\boldsymbol{\xi}_i'(\mathbf{I}_n - \mathbf{P}_i)\boldsymbol{\xi}_i}{2n} + \frac{\boldsymbol{\phi}_{0_i}' \mathbf{X}_i (\mathbf{I}_n - \mathbf{P}_i) \mathbf{X}_i' \boldsymbol{\phi}_{0_i}}{2n} + \frac{2\boldsymbol{\phi}_{0_i}' \mathbf{X}_i (\mathbf{I}_n - \mathbf{P}_i)\boldsymbol{\xi}_i}{2n} + \frac{\beta_i}{2n} + \frac{c_0}{2n} \\
&= \frac{\boldsymbol{\xi}_i'(\mathbf{I}_n - \mathbf{P}_i)\boldsymbol{\xi}_i}{2n} + \frac{\beta_i}{2n} + \frac{c_0}{2n} \\
&\leq \frac{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}{2n} + \frac{\beta_i}{2n} + \frac{c_0}{2n} \\
&\leq 3\sigma_{\max},
\end{aligned}
$$

where (i) is a direct application of Lemma S1.3 and (S4) in the Supplementary Material; (ii) is obtained by substituting the true model $\boldsymbol{y}_i = \mathbf{X}_i \boldsymbol{\phi}_{0_i} + \boldsymbol{\xi}_i$. Therefore, the scale parameter $\frac{\boldsymbol{y}_i'(\mathbf{I}_n - \tilde{\mathbf{P}}_{\nu_i})\boldsymbol{y}_i + \beta_i}{2}$ in the distribution of $\sigma_i^2 | \mathscr{G}_0, \mathscr{Y}$ is of order n. By choosing M^* properly (see Remark S1.1), we can make $\mathbb{E}_0 \Pi_n (\sigma_i > M^* \mid \mathscr{Y}) \leq \mathrm{e}^{-2 \log dp}$ for all large n. As for the second term in (26), first we define z to be a $\nu_i \times 1$ vector with entries i.i.d. $\mathcal{N}(0, 1)$. It follows from (Vershynin (2012) Corollary 5.35) and Assumption A1 that the second term goes to 0 as $n \to \infty$. This is due to the fact that for any $t > 0$,

$$
\mathbb{P}\left(\frac{\|z\|}{\sqrt{n}} > \sqrt{\frac{\nu_i}{n}} + \frac{t}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right) \leq \mathrm{e}^{-t^2/2}.
$$

Set $t^2 = 4 \log dp$. Thus, we can select n_3 so that for all $n \geq n_3$ we get $\frac{\|z\|}{\sqrt{n}} \geq \eta_{n,i} K \sqrt{\frac{\lambda_1}{2(M^*)^2}}$ with probability at most $\mathrm{e}^{-2 \log dp}$ for a large enough choice of K. Moving onto the second term in (25), we use the following inequalities:

$$
\begin{aligned}
\|\hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_{0_i}\| &= \left\| \left(\mathbf{X}_i' \mathbf{X}_i + \frac{\mathbf{I}_{\nu_i}}{\tau^2} \right)^{-1} \mathbf{X}_i' \boldsymbol{y}_i - \boldsymbol{\phi}_{0_i} \right\| \\
&\leq \left\| \left(\mathbf{X}_i' \mathbf{X}_i + \frac{\mathbf{I}_{\nu_i}}{\tau^2} \right)^{-1} \right\| \left\| \mathbf{X}_i' \mathbf{X}_i \boldsymbol{\phi}_{0_i} + \mathbf{X}_i' \boldsymbol{\xi}_i - \mathbf{X}_i' \mathbf{X}_i \boldsymbol{\phi}_{0_i} - \frac{\boldsymbol{\phi}_{0_i}}{n\tau^2} \right\| \\
&\leq \frac{1}{\lambda_{\min}(\mathbf{X}_i' \mathbf{X}_i / n)} \left(\frac{\|\mathbf{X}_i' \boldsymbol{\xi}_i\|}{n} + \frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2} \right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}_0\Big(\|\hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_{0_i}\| > \frac{K\eta_{n,i}}{2}\Big) \\
&\le \mathbb{P}_0\Big[\frac{1}{\lambda_{\min}(\mathbf{X}_i'\mathbf{X}_i/n)}\Big(\frac{\|\mathbf{X}_i'\boldsymbol{\xi}_i\|}{n} + \frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2}\Big) \ge \frac{K\eta_{n,i}}{2}\Big] \qquad (27)\\
&\le \mathbb{P}_0\Big[\lambda_{\min}\Big(\frac{\mathbf{X}_i'\mathbf{X}_i}{n}\Big) < \frac{\lambda_1}{2}\Big] + \mathbb{P}_0\Big(\frac{\|\mathbf{X}_i'\boldsymbol{\xi}_i\|}{n} + \frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2} > \frac{K\eta_{n,i}}{4}\Big).
\end{aligned}
$$

First, we claim that any column of the true parameter matrix $\boldsymbol{\Phi}_0$ is bounded in ℓ_2 norm by a constant not depending on n. This follows from the stationarity and stability of the underlying process $\{X^t\}$. More precisely,

$$
X^t = \boldsymbol{\Phi}_0'\tilde{X}^t + \boldsymbol{\varepsilon}^t \quad \text{where } \tilde{X}^t := \big[(X^{t-1})'\cdots(X^{t-d})'\big]' \quad \Longrightarrow \quad \Gamma_X(0) = \boldsymbol{\Phi}_0'\mathbf{C}_X\boldsymbol{\Phi}_0 + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0}.
$$

By pre- and post-multiplication by the unit vector $\mathbf{e}_i$ and using Assumption A2 the claim is now established. This implies $\frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2} = o(\eta_{n,i})$ because

$$
\frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2} \times \frac{1}{\eta_{n,i}} = \frac{\|\boldsymbol{\phi}_{0_i}\|}{n\tau^2} \times \frac{1}{\mathscr{B}_n}\sqrt{\frac{n}{\nu_i \log dp}} \le \frac{\|\boldsymbol{\phi}_{0_i}\|}{\tau^2\mathscr{B}_n\sqrt{n\log dp}} = o(1).
$$

From (23) and by Assumption A1, there exists n_4 such that both the probabilities in (27) are at most $\mathrm{e}^{-2\log p}$ for a large enough choice of K. Hence, for appropriate choice of $K > 0$, we have

$$
\begin{aligned}
&\mathbb{E}_0\Pi_n\big(\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\| \ge K\eta_{n,i} \mid \mathscr{Y}\big) \\
&\le \mathbb{E}_0\Pi_n\Big(\|\boldsymbol{\phi}_i - \hat{\boldsymbol{\phi}}_i\| \ge \frac{K\eta_{n,i}}{2}\Big|\mathscr{Y}\Big) + \mathbb{P}_0\Big(\|\hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_{0_i}\| \ge \frac{K\eta_{n,i}}{2}\Big|\mathscr{Y}\Big) \\
&\le 3\mathrm{e}^{-2\log dp} + 2\mathrm{e}^{-2\log dp} \\
&= \frac{5}{d^2p^2}.
\end{aligned}
$$

This completes the proof. □

A.3. Proof of Theorem 4.4. Throughout this proof, we will restrict ourselves to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$. For notational convenience, let us denote Π_{pseudo} and π_{pseudo}, respectively, by Π_n and π_n. For any $n \times s$ submatrix $\mathbf{X}_s$ of $\mathbf{X}$, $\mathbf{P}_{\nu_s}$ stands for the projection matrix $\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'$ onto $\mathscr{C}(\mathbf{X}_s)$ and by $\tilde{\mathbf{P}}_{\nu_s}$ is defined to be $\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s + \frac{\mathbf{I}_s}{\tau^2})^{-1}\mathbf{X}_s'$. Note that

$$
\begin{aligned}
&\Pi_n\Big\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F > K\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\delta_n k_n \log dp}{n}}\Big|\mathscr{Y}\Big\} \\
&= \Pi_n\Big\{\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\|_F^2 > K^2\Big(\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\Big)^2\frac{\sum_{i=1}^p \nu_{t_i} k_n \log dp}{n}\Big|\mathscr{Y}\Big\} \\
&= \Pi_n\Big\{\sum_{i=1}^p \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2^2 > \sum_{i=1}^p K^2\Big(\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\Big)^2\frac{\nu_{t_i} k_n \log dp^2}{n}\Big|\mathscr{Y}\Big\} \\
&\le p\max_{1\le i\le p:\nu_{t_i}>0}\Pi_n\Big\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\nu_{t_i} k_n \log dp}{n}}\Big|\mathscr{Y}\Big\}.
\end{aligned}
$$

For ease of exposition, we denote $\frac{1+\mu_{\max}(\tilde{\mathscr{A}})}{\mu_{\min}(\tilde{\mathscr{A}})}\sqrt{\frac{\nu_{t_i} k_n \log dp}{n}}$ by η_{t_i}. Let $\mathscr{G}^i$ denote the ith column of the activity graph $\mathscr{G}$, for $1 \le i \le p$. It follows by (15) that $\mathscr{G}^1, \mathscr{G}^2, \ldots, \mathscr{G}^p$ are mutually independent given $\mathscr{Y}$, and

$$\pi_n(\mathscr{G}^i \mid \mathscr{Y}) \propto \left(\frac{q_{\nu_i(\mathscr{G})}}{\tau\sqrt{n}(1-q_{\nu_i(\mathscr{G})})}\right)^{\nu_i(\mathscr{G})} \left|\frac{\mathbf{X}_i'\mathbf{X}_i}{n} + \frac{\mathbf{I}_{\nu_i}}{n\tau^2}\right|^{-1/2} \left(S_i + \frac{\beta_i}{n}\right)^{-(\frac{n}{2}+\alpha_i)}.$$

By the arguments preceding (*), we obtain

$$\frac{1}{2}\frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i \mid \mathscr{Y})}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i \mid \mathscr{Y})} \le B(\boldsymbol{m}_i, \boldsymbol{t}_i).$$

Note that the proofs of Lemma A.2 and the part of Lemma A.3 for $\nu_{m_i} > \nu_{t_i}$, only use Assumptions A1–A3. Since $d(\boldsymbol{m}_i, \boldsymbol{t}_i) \le 2k_n(\nu_{m_i} - \nu_{t_i})$ for $\nu_{m_i} > \nu_{t_i}$, It follows that

$$\frac{1}{2}\frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i \mid \mathscr{Y})}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i \mid \mathscr{Y})} \le (dp^2)^{-2d(\boldsymbol{m}_i, \boldsymbol{t}_i)}$$

$\nu_{m_i} > \nu_{t_i}$.

Next, we focus on the case when $\nu_{m_i} \le \nu_{t_i}$, with $\boldsymbol{m}_i \ne \boldsymbol{t}_i$. By using arguments at the end of the proof of Lemma A.1 and in the proof of Case III of Lemma A.3 (without Assumption A4 and the bound $\|\boldsymbol{\phi}_{t_i \cap \boldsymbol{m}_i^c}\| \ge \nu_{t_i \cap m_i^c} s_n^2$), we get

$$\begin{aligned}\frac{1}{2}\frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i \mid \mathscr{Y})}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i \mid \mathscr{Y})} &\le B(\boldsymbol{m}_i, \boldsymbol{t}_i) \\ &\le \mathrm{e}^{-\log q_{1,n} \nu_{t_i \cap m_i^c}} \mathrm{e}^{-C_1 n \|\boldsymbol{\phi}_{t_i \cap m_i^c}\|^2} \\ &\le \mathrm{e}^{C_2 n \eta_{t_i}^2 - C_1 n \|\boldsymbol{\phi}_{t_i \cap m_i^c}\|^2}\end{aligned}$$

for appropriate constants $C_1 > 0$ and $C_2 > 0$. It follows that if $\|\boldsymbol{\phi}_{t_i \cap m_i^c}\| > \tilde{C}\eta_{t_i}$, where $\tilde{C} = \sqrt{\frac{C_2}{C_1} + 4}$, then

$$\frac{1}{2}\frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i \mid \mathscr{Y})}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i \mid \mathscr{Y})} \le \mathrm{e}^{-4n\eta_{t_i}^2} \le (dp)^{-4\nu_{t_i} k_n} \le dp^{-4k_n}.$$

Hence, we get

$$\begin{aligned}&\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}\} \\ &= \mathbb{E}_0(\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}, \mathscr{G}^i\}) \\ &= \sum_{\boldsymbol{m}_i : \nu_{m_i} \le \nu_{t_i}} \mathbb{E}_0(\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i\})\pi_n(\mathscr{G}^i = \boldsymbol{m}_i) \\ &\quad + \sum_{\boldsymbol{m}_i : \nu_{m_i} > \nu_{t_i}} \mathbb{E}_0(\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i\})\pi_n(\mathscr{G}^i = \boldsymbol{m}_i) \\ &\le \max_{\boldsymbol{m}_i : \nu_{m_i} \le \nu_{t_i}, \|\boldsymbol{\phi}_{t_i \cap m_i^c}\| \le \tilde{C}\eta_{t_i}} \mathbb{E}_0(\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i\}) \\ &\quad + \pi_n(\mathscr{G}^i = \boldsymbol{t}_i)\left(\sum_{\boldsymbol{m}_i : \nu_{m_i} > \nu_{t_i}} \frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i)}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i)} + \sum_{\boldsymbol{m}_i : \nu_{m_i} \le \nu_{t_i}, \|\boldsymbol{\phi}_{t_i \cap m_i^c}\| > \tilde{C}\eta_{t_i}} \frac{\pi_n(\mathscr{G}^i = \boldsymbol{m}_i)}{\pi_n(\mathscr{G}^i = \boldsymbol{t}_i)}\right) \\ &\le \max_{\boldsymbol{m}_i : \nu_{m_i} \le \nu_{t_i}, \|\boldsymbol{\phi}_{t_i \cap m_i^c}\| \le \tilde{C}\eta_{t_i}} \mathbb{E}_0(\Pi_n\{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_{0_i}\|_2 > K\eta_{t_i} \mid \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i\})\end{aligned}$$

$$+\sum_{j=1}^{dp}\binom{dp}{j}(dp^2)^{-2j}+2(dp)^{\nu_{t_i}}(dp)^{-4k_n}$$

$$\le \max_{\boldsymbol{m}_i:\nu_{m_i}\le\nu_{t_i},\|\boldsymbol{\phi}_{t_i\cap m_i^c}\|\le\tilde{C}\eta_{t_i}} \mathbb{E}_0\big(\Pi_n\{\|\boldsymbol{\phi}_i-\boldsymbol{\phi}_{0_i}\|_2>K\eta_{t_i}\mid\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\}\big)$$

$$+\frac{(dp^2)^{-1}}{1-(dp^2)^{-1}}+(dp^2)^{-1}.$$

Next, we show that there exists N such that for every $n\ge N$, we have

$$\max_{1\le i\le p}\max_{\boldsymbol{m}_i:\nu_{m_i}\le\nu_{t_i},\|\boldsymbol{\phi}_{0,t_i\cap m_i^c}\|\le\tilde{C}\eta_{t_i}} \mathbb{E}_0\big(\Pi_n\{\|\boldsymbol{\phi}_i-\boldsymbol{\phi}_{0_i}\|_2>K\eta_{t_i}\mid\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\}\big)\le 2\mathrm{e}^{-2\log dp}.$$

Fix i and $\boldsymbol{m}_i$ such that $\nu_{m_i}\le\nu_{t_i}$, $\|\boldsymbol{\phi}_{t_i\cap m_i^c}\|\le\tilde{C}\eta_{t_i}$ arbitrarily. Let

$$\hat{\boldsymbol{\phi}}_{m_i}=\left(\mathbf{X}_{m_i}'\mathbf{X}_{m_i}+\frac{\mathbf{I}_{\nu_{m_i}}}{\tau^2}\right)^{-1}\mathbf{X}_{m_i}'\boldsymbol{y}_i.$$

Using $\|\boldsymbol{\phi}_{0,t_i\cap m_i^c}\|\le\tilde{C}\eta_{t_i}$ and $K>6\tilde{C}$, it follows that

$$\begin{aligned}
&\mathbb{E}_0\Pi_n\big(\|\boldsymbol{\phi}_i-\boldsymbol{\phi}_{0_i}\|\ge K\eta_{t_i}\mid\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\big)\\
&\quad\le\mathbb{E}_0\Pi_n\left(\|\tilde{\boldsymbol{\phi}}_i-\hat{\boldsymbol{\phi}}_{m_i}\|\ge\frac{K\eta_{t_i}}{2}\middle|\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\right)\\
&\qquad+\mathbb{P}_0\left(\|\hat{\boldsymbol{\phi}}_{m_i}-\boldsymbol{\phi}_{0_{m_i}}\|+\|\boldsymbol{\phi}_{0,t_i\cap\boldsymbol{m}_i^c}\|\ge\frac{K\eta_{t_i}}{2}\middle|\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\right)\\
&\quad=\mathbb{E}_0\Pi_n\left(\|\tilde{\boldsymbol{\phi}}_i-\hat{\boldsymbol{\phi}}_{m_i}\|\ge\frac{K\eta_{t_i}}{2}\middle|\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\right)+1_{\{\|\hat{\boldsymbol{\phi}}_{m_i}-\boldsymbol{\phi}_{0_{m_i}}\|\ge\frac{K\eta_{t_i}}{3}\}}.
\end{aligned}\tag{28}$$

Next, we analyze the terms in (28) separately. For the first term, we note that

$$\begin{aligned}
\|\tilde{\boldsymbol{\phi}}_i-\hat{\boldsymbol{\phi}}_{m_i}\|&=\left\|\sigma_i\left(\mathbf{X}_{m_i}'\mathbf{X}_{m_i}+\frac{\mathbf{I}_{\nu_{m_i}}}{\tau^2}\right)^{-1/2}\frac{(\mathbf{X}_{m_i}'\mathbf{X}_{m_i}+\frac{\mathbf{I}_{\nu_{m_i}}}{\tau^2})^{1/2}(\tilde{\boldsymbol{\phi}}_i-\hat{\boldsymbol{\phi}}_i)}{\sigma_i}\right\|\\
&\le\frac{\sigma_i}{\sqrt{\lambda_{\min}(\mathbf{X}_{m_i}'\mathbf{X}_{m_i})}}\|z\|,
\end{aligned}$$

where z is $\nu_{m_i}\times 1$ standard normal random vector and the last step follows from (16). Note that $\nu_{m_i}\le\nu_{t_i}\le k_n$. By Assumption A1, for large enough n, we have that on $\Xi_{2,n}$,

$$\left\|\frac{\mathbf{X}_{m_i}'\mathbf{X}_{m_i}}{n}-\boldsymbol{C}_{\mathbf{X}_{m_i}}\right\|_2\le\mathscr{B}_n\sqrt{\frac{2\nu_{m_i}\log dp}{n}}\le\frac{\lambda_1}{2}.$$

Using $\nu_{m_i}\le\nu_{t_i}$, it follows that for any $M^*>0$,

$$\begin{aligned}
&\mathbb{E}_0\Pi_n\left(\|\tilde{\boldsymbol{\phi}}_i-\hat{\boldsymbol{\phi}}_{m_i}\|\ge\frac{K\eta_{t_i}}{2}\mid\mathscr{Y},\mathscr{G}^i=\boldsymbol{m}_i\right)\\
&\quad\le\mathbb{P}\left(\|z\|>\sqrt{n}\eta_{m_i}K\sqrt{\frac{\lambda_1}{8(M^*)^2}}\right)+\mathbb{E}_0\Pi_n(\sigma_i>M^*\mid\mathscr{Y}).
\end{aligned}\tag{29}$$

First, we claim that any column of the true parameter matrix $\boldsymbol{\Phi}_0$ is bounded in ℓ_2 norm by a constant not depending on n. This follows from the stationarity and stability of the underlying process $\{X^t\}$. More precisely,

$$X^t=\boldsymbol{\Phi}_0'\tilde{X}^t+\boldsymbol{\varepsilon}^t\quad\text{where }\tilde{X}^t:=\big[(X^{t-1})'\cdots(X^{t-d})'\big]'\quad\Longrightarrow\quad\Gamma_X(0)=\boldsymbol{\Phi}_0'\mathbf{C}_X\boldsymbol{\Phi}_0+\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},0}.$$

By pre- and post-multiplication by the unit vector $\mathbf{e}_i$ and using Assumption A2, the claim is now established. Note that the distribution of $\sigma_i^2 \mid \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i$ is Inverse-Gamma with shape parameter $(\frac{n}{2} + \alpha_i) \sim n$ and scale parameter $\frac{\boldsymbol{y}_i'(\mathbf{I}_n - \tilde{\mathbf{P}}_{\nu_{m_i}})\boldsymbol{y}_i + \beta_i}{2}$. Hence,

$$\begin{aligned}
&\frac{\boldsymbol{y}_i'(\mathbf{I}_n - \tilde{\mathbf{P}}_{\nu_{m_i}})\boldsymbol{y}_i + \beta_i}{2} \\
&\quad \leq \frac{\boldsymbol{y}_i'\boldsymbol{y}_i + \beta_i}{2} \\
&\quad \overset{(i)}{=} \boldsymbol{\xi}_i'\boldsymbol{\xi}_i + \tilde{\boldsymbol{\phi}}_{0_{t_i}}' \mathbf{X}_{t_i}'\mathbf{X}_{t_i}\tilde{\boldsymbol{\phi}}_{0_{t_i}} + \frac{\beta_i}{2} \\
&\quad \overset{(ii)}{\leq} \frac{3n\sigma_{\max}}{2} + 2n\lambda_2\|\tilde{\boldsymbol{\phi}}_{0_{t_i}}\|_2^2 + \frac{\beta_i}{2}
\end{aligned}$$

for large enough n. Note that (i) is obtained by substituting the true model $\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\phi}_{0_i} + \boldsymbol{\xi}_i$, and (ii) follows by restricting to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$. Therefore, the scale parameter $\frac{\boldsymbol{y}_i'(\mathbf{I}_n - \tilde{\mathbf{P}}_{\nu_i})\boldsymbol{y}_i + \beta_i}{2}$ in the distribution of $\sigma_i^2 | \mathscr{Y}, \mathscr{G}^i = \boldsymbol{m}_i$ is bounded by a constant multiple of n. Hence, by Remark S1.1 in the Supplementary Material, there exists a constant M^* such that $\mathbb{E}_0\Pi_n(\sigma_i > M^* \mid \mathscr{Y}) \leq \mathrm{e}^{-2\log dp}$ for all large n.

As for the first term in (29), note that $\boldsymbol{z}$ is a $\nu_{m_i} \times 1$ vector with i.i.d. $\mathcal{N}(0,1)$ entries. It follows from Vershynin (2012), Corollary 5.35, that for any $t > 0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{z}\|}{\sqrt{n}} > \sqrt{\frac{\nu_{t_i}}{n}} + \frac{t}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) \leq \mathrm{e}^{-t^2/2}.$$

By setting $t^2 = 4\log dp$, it follows that $\frac{\|\boldsymbol{z}\|}{\sqrt{n}} \geq \eta_{t_i} K\sqrt{\frac{\lambda_1}{8(M^*)^2}}$ with probability at most $\mathrm{e}^{-2\log dp}$ for a large enough choice of K.

Moving onto the second term in (28), note that

$$\begin{aligned}
\|\hat{\boldsymbol{\phi}}_{m_i} - \boldsymbol{\phi}_{0_{m_i}}\| &= \left\| \left(\mathbf{X}_{m_i}'\mathbf{X}_{m_i} + \frac{\mathbf{I}_{\nu_{m_i}}}{\tau^2}\right)^{-1} \mathbf{X}_{m_i}\boldsymbol{y}_i - \boldsymbol{\phi}_{0_{m_i}} \right\| \\
&\leq \left\| \left(\mathbf{X}_{m_i}'\mathbf{X}_{m_i} + \frac{\mathbf{I}_{\nu_{m_i}}}{\tau^2}\right)^{-1} \right\| \left\| \mathbf{X}_{m_i}'\mathbf{X}_{t_i}\boldsymbol{\phi}_{0_{t_i}} + \mathbf{X}_{m_i}'\boldsymbol{\xi}_i - \mathbf{X}_{m_i}'\mathbf{X}_{m_i}\boldsymbol{\phi}_{0_{m_i}} - \frac{\boldsymbol{\phi}_{0_{m_i}}}{n\tau^2} \right\|.
\end{aligned}$$

Since $\boldsymbol{\phi}_{0,t_i^c \cap m_i} = 0$, $\|\boldsymbol{\phi}_{t_i \cap m_i^c}\| \leq \tilde{C}\eta_{t_i}$, $\|\boldsymbol{\phi}_{0_{m_i}}\|$ is uniformly bounded in n, $\nu_{m_i} \leq \nu_{t_i}$, and we are restricting to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$, it follows that

$$\begin{aligned}
\|\hat{\boldsymbol{\phi}}_{m_i} - \boldsymbol{\phi}_{0_{m_i}}\| &\leq \frac{2}{\lambda_1}\left(\frac{\|\mathbf{X}_{m_i}'\mathbf{X}_{t_i \cap m_i^c}\boldsymbol{\phi}_{t_i \cap m_i^c}\|}{n} + \frac{\|\mathbf{X}_{m_i}'\boldsymbol{\xi}_i\|}{n} + \frac{\|\boldsymbol{\phi}_{0_{m_i}}\|}{n\tau^2}\right) \\
&\leq \frac{2}{\lambda_1}\left(2\lambda_2\|\boldsymbol{\phi}_{t_i \cap m_i^c}\| + 2\mathscr{B}_n\sqrt{\frac{\nu_{t_i}\log dp}{n}}\right) \\
&\leq \tilde{C}_1\eta_{t_i}
\end{aligned}$$

for an appropriate constant $\tilde{C}_1$. Hence for $K > 6\max(\tilde{C}, \tilde{C}_1)$, we have

$$1_{\{\|\hat{\boldsymbol{\phi}}_{m_i} - \boldsymbol{\phi}_{0_{m_i}}\| \geq \frac{K\eta_{t_i}}{3}\}} = 0.$$

This completes the proof.

Funding. The work of the second and third authors was supported in part by NSF Grant DMS 1821220 and the work of GM was additionally supported in part by NSF Grants DMS 1830175 and IIS 1632730 and NIH 5R01GM11402905.

SUPPLEMENTARY MATERIAL

Supplement to "Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach" (DOI: 10.1214/20-AOS1992SUPP; .pdf). Supplementary information.

REFERENCES

BAŃBURA, M., GIANNONE, D. and REICHLIN, L. (2010). Large Bayesian vector auto regressions. *J. Appl. Econometrics* **25** 71–92. MR2751790 https://doi.org/10.1002/jae.1137

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BONDELL, H. D. and REICH, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107** 1610–1624. MR3036420 https://doi.org/10.1080/01621459.2012.716344

BROWN, P. J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 519–536. MR1924304 https://doi.org/10.1111/1467-9868.00348

CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935 https://doi.org/10.1214/18-AOS1689

CAO, X., KHARE, K. and GHOSH, M. (2020). High-dimensional posterior consistency for hierarchical non-local priors in regression. *Bayesian Anal.* **15** 241–262. MR4050884 https://doi.org/10.1214/19-BA1154

CASELLA, G., GIRÓN, F. J., MARTÍNEZ, M. L. and MORENO, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37** 1207–1228. MR2509072 https://doi.org/10.1214/08-AOS606

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

DE MOL, C., GIANNONE, D. and REICHLIN, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Econometrics* **146** 318–328. MR2465176 https://doi.org/10.1016/j.jeconom.2008.08.011

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972 https://doi.org/10.1093/biomet/87.4.731

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2019a). High-dimensional posterior consistency in Bayesian vector autoregressive models. *J. Amer. Statist. Assoc.* **114** 735–748. MR3963176 https://doi.org/10.1080/01621459.2018.1437043

GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2019b). Supplemental document for "High dimensional posterior consistency in bayesian vector autoregressive models". *J. Amer. Statist. Assoc.*

GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2021). Supplement to "Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach." https://doi.org/10.1214/20-AOS1992SUPP

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849 https://doi.org/10.1198/016214507000000121

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074 https://doi.org/10.1080/01621459.2012.682536

KINNEY, S. K. and DUNSON, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63** 690–698. MR2395705 https://doi.org/10.1111/j.1541-0420.2007.00771.x

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243 https://doi.org/10.1198/016214507000001337

LIN, J. and MICHAILIDIS, G. (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *J. Mach. Learn. Res.* **18** Paper No. 117, 49. MR3725456 https://doi.org/10.1631/jzus.a1500279

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

MELNYK, I. and BANERJEE, A. (2016). Estimating structured vector autoregressive models. In *Proceedings of The 33rd International Conference on Machine Learning* 830–839.

MICHAILIDIS, G. and D'ALCHÉ-BUC, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Math. Biosci.* **246** 326–334. MR3132054 https://doi.org/10.1016/j.mbs.2013.10.003

NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987 https://doi.org/10.1214/14-AOS1207

NICHOLSON, W. B., MATTESON, D. S. and BIEN, J. (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *Int. J. Forecast.* **33** 627–651.

ROBERTSON, J. C. and TALLMAN, E. W. (1999). Vector autoregressions: Forecasting and reality. *Econ. Rev.* **84**.

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258 https://doi.org/10.1214/ECP.v18-2865

SETH, A. K., BARRETT, A. B. and BARNETT, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* **35** 3293–3297.

SHIN, M., BHATTACHARYA, A. and JOHNSON, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* **28** 1053–1078. MR3791100

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.

SIMS, C. A. and ZHA, T. (1998). Bayesian methods for dynamic multivariate models. *Internat. Econom. Rev.* 949–968.

SONG, Q. and LIANG, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 947–972. MR3414135 https://doi.org/10.1111/rssb.12095

STOCK, J. H. and WATSON, M. W. (2005). An empirical comparison of methods for forecasting using many predictors. Princeton Univ. Manuscript.

TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596