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

BY SATYAJIT GHOSH 1, KSHITIJ KHARE 2 AND GEORGE MICHAELIDIS 2,3

1 DEPARTMENT OF STATISTICS & BIOSTATISTICS, RUTGERS UNIVERSITY
2 DEPARTMENT OF STATISTICS, UNIVERSITY OF FLORIDA
3 INFORMATICS INSTITUTE, UNIVERSITY OF FLORIDA

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 regu-

---

Keywords and phrases: Bayesian variable selection, vector auto-regression, pseudo-likelihood, strong selection consistency, high dimensional data

---
latory networks in functional genomics Michailidis and d’Alché Buc (2013) and brain connectivities from neuroimaging data Seth et al. (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 (2007)). 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 $\ell_1$ 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 et al. (2017) and to incorporating exogenous variables Lin and Michailidis (2017). Key issues 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. Sims and Zha (1998); Robertson and Tallman (1999); Mol et al. (2008); Banbura et al. (2010)). 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; e.g. see George and Foster (2000); Ishwaran and Rao (2005); Brown et al. (2002); Kinney and Dunson (2007); Liang et al. (2008); Bondell and Reich (2012); Johnson and Rossell (2012); 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 Casella et al. (2009); Bondell and Reich (2012); Narisetty and He (2014); Castillo et al. (2015); Shin et al. (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 non-true 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 (Narisetty and He (2014); Castillo et al. (2015); Shin et al. (2018)). In all of the above studies, the results hold when the number of coefficients $p$ grows sub-exponentially with the sample size $n$, and the errors are assumed to be independent and identically distributed (i.i.d.) with a common variance parameter $\sigma^2$.

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

$$
X_t = c + \sum_{i=1}^{d} A_i X_{t-i} + \epsilon_t.
$$

The temporal dependence structure of the model is characterized by the $p \times p$ transition matrices $A_1, A_2, \ldots, A_d$, and $c$ is a $p \times 1$ location vector. In the Gaussian VAR model, the errors $\epsilon_t$ are assumed to be independent and identically distributed (i.i.d) $N_p(0, \Sigma_\epsilon)$, with the error covariance matrix $\Sigma_\epsilon$ unknown in most applications. Note that our main parameters of interest are the transition matrices $A_1, A_2, \ldots, A_d$, 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 (2007)) as

$$
X_t - \mu = \sum_{i=1}^{d} A_i (X_{t-i} - \mu) + \epsilon_t,
$$

where $\mu = (I - A_1 - A_2 - \cdots - A_d)^{-1} c$ is known as the process mean. Usually $\mu$ will not be known in advance. In that case, $\mu$ is estimated by the vector of sample means $\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t$. Henceforth, we assume without loss of generality that $\mu = 0$. Based on the data $\{X^0, \cdots, X^T\}$, we define the response matrix $Y$ and design matrix $X$ as follows,

$$
Y = \begin{bmatrix} (X^T)' \\ \vdots \\ (X^d)' \end{bmatrix}_{n \times p} \quad X = \begin{bmatrix} (X^{T-1})' & \cdots & (X^{T-d})' \\ \vdots & \ddots & \vdots \\ (X^{d-1})' & \cdots & (X^0)' \end{bmatrix}_{n \times dp}.
$$
We can then rewrite the above model in a linear regression setup as

\[
Y = X\Phi + E
\]

where

\[
\Phi = \begin{bmatrix}
A'_1 \\
A'_2 \\
\vdots \\
A'_d
\end{bmatrix},
\quad E = \begin{bmatrix}
(\varepsilon^T)' \\
\vdots \\
(\varepsilon^d)'
\end{bmatrix}.
\]

In this formulation, the number of samples is \( n = T - d + 1 \) and the number of unknown parameters in the coefficient matrix \( \Phi \) 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 \( \sigma^2 \) in simple linear regression is replaced by the covariance matrix \( \Sigma_\varepsilon \) 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 \( \Sigma_\varepsilon \), see Section 3). Thus, the presence of temporal dependence introduces significantly non-trivial 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 \( \Phi \). Sparsity in \( \Phi \) is introduced through an ‘activity graph’ \( G \), which is the direct sum of \( dp \times p \) matrices of 1’s and 0’s, identifying which entries of \( \Phi \) are active/inactive. Given \( G \), an appropriate multivariate normal prior is chosen for the active entries of \( \Phi \). 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 sub-exponential rate with \( n \), we establish strong selection consistency of our Bayesian variable selection approach, i.e., 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, i.e., 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 et al. (2018a), 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 non-true activity graph with respect to the true activity graph. Different bounds are obtained for different types of non-true activity graphs, through involved analysis and leveraging selected concentration inequalities from Basu and Michailidis (2015) and Ghosh et al. (2018a) 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 non-true 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 et al. (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 in the Supplement.

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 \( \mathbf{v} \in \mathbb{R}^p \), \( \| \mathbf{v} \| := \sqrt{\sum v_j^2} \) denotes the \( \ell_2 \)-norm. For a matrix \( \mathbf{A} \), \( \| \mathbf{A} \| \) and \( \sigma_{\text{max}}(\mathbf{A}) \) denote spectral norm i.e., \( \| \mathbf{A} \| = \sup_{x \neq 0} \frac{\| \mathbf{A} x \|_2}{\| x \|_2} \) 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 \( \lambda_{\text{max}}(\mathbf{A}) \) and \( \lambda_{\text{min}}(\mathbf{A}) \). The vector \( e_i \) is used for the \( i \)-th 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 \( \Phi \) is a \( p \times p \) matrix then \( \phi = \text{vec}(\Phi) \). Also, \( \mathbf{O} \) represents a zero-matrix of appropriate dimension, and in general vectors are denoted by italicized bold lowercase letters. \( a_n \sim b_n \) implies that \( \frac{a_n}{b_n} \rightarrow c \) for some constant \( c > 0 \) and \( a_n \preceq b_n \) implies \( a_n = O(b_n) \).
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, i.e., the time-series process does not diverge over time. The VAR process \( \{X^t\} \) with lag \( d \) is stable and invertible if \( \det(I_p - \sum_{i=1}^{d} A_i z^i) \neq 0 \) inside the unit circle of the complex plane, i.e., \( \{z \in \mathbb{C} : |z| \leq 1\} \). 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

\[
\tilde{X}^t = \begin{bmatrix} X^t \\ \vdots \\ X^{t-d+1} \end{bmatrix}_{dp \times 1}, \quad \tilde{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & O & \cdots & O & O \\ O & I_p & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & I_p & O \end{bmatrix}_{dp \times dp}
\]

and \( \omega^t = \begin{bmatrix} \epsilon^{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{dp \times 1} \),

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

(3) \[
\tilde{X}^t = \tilde{A} \tilde{X}^{t-1} + \omega^t \quad t = d, \ldots, n + d - 1.
\]

It follows that \( X = [\tilde{X}^n, \tilde{X}^{n+d-2}, \tilde{X}^{n+d-3}, \ldots, \tilde{X}^{d-1}]' \), where the \( i \)-th row of \( X \) is denoted as a \( (dp \times 1) \) vector \( \tilde{X}^n, \tilde{X}^{n+d-2}, \ldots, \tilde{X}^{d-1} \). 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 \( \tilde{A} \) have absolute value strictly less than 1; i.e. \( \max_{1 \leq i \leq p} |\lambda_i(\tilde{A})| < 1 \).

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

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

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

\[
M(f_{\tilde{X}}) := \text{ess sup}_{\theta \in [-\pi, \pi]} \lambda_{\text{max}}(f_{\tilde{X}}(\theta))
\]

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

The largest eigenvalue \( M(f_{\tilde{X}}) \) 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 \( M(f_{\tilde{X}}) \) are...
considered less stable. The smallest eigenvalue \( m(\tilde{f}_X) \) 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 \( \mu_{\max}(\tilde{A}) \) and \( \mu_{\min}(\tilde{A}) \) which are defined as

\[
\mu_{\max}(\tilde{A}) := \max_{\theta \in [-\pi,\pi]} \lambda_{\max} \left( \left( I_p - \tilde{A}' e^{i\theta} \right) \left( I_p - \tilde{A} e^{-i\theta} \right) \right) \leq [1 + (v_{in} + v_{out})/2]^2,
\]

where \( v_{in} = \max_{1 \leq i \leq dp} \sum_{j=1}^{dp} |\tilde{A}_{ij}| \) and \( v_{out} = \max_{1 \leq j \leq dp} \sum_{i=1}^{dp} |\tilde{A}_{ij}| \), and

\[
\mu_{\min}(\tilde{A}) := \min_{\theta \in [-\pi,\pi]} \lambda_{\min} \left( \left( I_p - \tilde{A}' e^{i\theta} \right) \left( I_p - \tilde{A} e^{-i\theta} \right) \right) \geq \left(1 - \rho(\tilde{A})\right)^2 \|P\|^{-2} \|P^{-1}\|^{-2},
\]

and where \( \rho(\tilde{A}) \) is the spectral radius of \( \tilde{A} \) and the columns of \( P \) correspond to eigenvectors of \( \tilde{A} \) (assuming \( \tilde{A} \) 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 Supplemental Section S5 and S6).

**Proposition 2.1.** (a) Let \( \| \cdot \| \) denote the operator norm. Then,

\[
\mu_{\min}(\tilde{A}) \geq 1 - 2\|\tilde{A}\| + \lambda_{\min}(\tilde{A}'\tilde{A}),
\]

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

\[
\mu_{\min}(\tilde{A}) \geq \frac{3 - \sqrt{5}}{2} \left( 2 \sum_{i=1}^d \|A_i\|_1 + \|A_1\|_\infty \right)
\]

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

\[
\mu_{\max}(\tilde{A}) = (1 + \rho(A))^2,
\]

where \( \rho(A) \) is the spectral radius of \( A \). Similarly

\[
\mu_{\min}(\tilde{A}) = (1 - \rho(A))^2.
\]
Since each $\varepsilon^i$ is i.i.d. as $\mathcal{N}_p(0, \Sigma_{\varepsilon})$, each row of $X$ is distributed as $\mathcal{N}_{dp}(0, C_X)$, where $C_X = \Gamma_{\mathcal{X}}(0)$.

The quantities $\mu_{\min}(A)$, $\mu_{\max}(A)$, the eigenvalues of $f_{\mathcal{X}}$, and the eigenvalues of $C_X$ are related by the following chain of inequalities (see Basu and Michailidis (2015))

\[
\frac{\lambda_{\min}(\Sigma_\varepsilon)}{\mu_{\max}(A)} \leq 2\pi m(f_{\mathcal{X}}) \leq \lambda_{\min}(C_X) \leq \lambda_{\max}(C_X) \leq 2\pi M(f_{\mathcal{X}}) \leq \frac{\lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(A)}.
\]

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 $\gamma_{ij}$ to indicate if the $(k, \ell)$-th entry of $\Phi$ is active, i.e., $\gamma_{k\ell} = 1$ ($\Phi_{k\ell} \neq 0$) for $1 \leq k \leq dp, 1 \leq \ell \leq p$. Let $\mathcal{G} := ((\gamma_{k\ell}))_{dp \times p}$ be a matrix of active positions in $\Phi$, henceforth referred to as the “activity graph of $\Phi$.”

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

\[
y_i = X_i\tilde{\phi}_i + \xi_i \quad \text{for } i = 1, \cdots, p,
\]

where $\xi_i$ and $y_i$ correspond to the $i$-th column of $E$ and $Y$, respectively, and $X_i$ is a $n \times \nu_i$ submatrix of $X$ consisting of the columns of $X$ corresponding to the active entries in the $i$-th column of $\Phi$. Let $\sigma_i^2$ be the $i$-th diagonal entry of $\Sigma_{\varepsilon}$. Since the rows of $E$ (i.e. $\varepsilon^i$) are i.i.d. $\mathcal{N}_p(0, \Sigma_{\varepsilon})$, we have that $\xi_i \sim \mathcal{N}_n(0, \sigma_i^2 I_n)$.

Note that the vectors $y_1, y_2, \cdots, y_p$. The joint density of $y_1, y_2, \cdots, y_p$, or equivalently, the likelihood function of $\Phi, \Sigma_{\varepsilon}$ given the data, is given by

\[
L(\Phi, \Sigma_{\varepsilon} \mid y_i, i = 1, \cdots, p) := \frac{1}{\sqrt{(2\pi)^p|\Sigma_{\varepsilon}|}} \exp \left( -\frac{1}{2} tr \left( (Y - X\Phi)\Sigma_{\varepsilon}^{-1}(Y - X\Phi)' \right) \right).
\]

As mentioned earlier, the main parameter of interest is the coefficient matrix $\Phi$, while the error covariance matrix $\Sigma_{\varepsilon}$ is essentially an unknown nuisance parameter. Keeping this in mind, we construct a pseudo-likelihood which is equal to the joint density of $y_1, y_2, \cdots, y_p$ under the assumption that $\Sigma_{\varepsilon}$ is diagonal. In particular, we define

\[
L_{\text{pseudo}}(\Phi, \Sigma_{\varepsilon} \mid y_i, i = 1, \cdots, p) := \frac{1}{\sqrt{(2\pi)^p \prod_{i=1}^{p} \sigma_i^2}} \exp \left( - \frac{1}{2} \sum_{i=1}^{p} \frac{||y_i - X_i\tilde{\phi}_i||^2}{\sigma_i^2} \right).
\]
The function $L_{\text{pseudo}}$ in (13) has a simpler form than the likelihood function $L$ and only involves the diagonal entries of $\Sigma_\epsilon$. 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 $L_{\text{pseudo}}$ 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 $L_{\text{pseudo}}$, we now construct the following hierarchical prior distribution for $(\Phi, \mathcal{G})$.

\[
\phi_{ji} | \mathcal{G}, \sigma^2 \sim (1 - \gamma_{ji})\phi_{ji} = 0 + \gamma_{ji} \mathcal{N}(0, \tau^2 \sigma_i^2) \text{ independently for } 1 \leq i \leq p, 1 \leq j \leq dp
\]
\[
\pi(\mathcal{G}) \propto \prod_{i=1}^{p} \left\{ q_1^{\nu_i(\mathcal{G})}(1 - q_1)^{dp - \nu_i(\mathcal{G})} 1_{\{\nu_i(\mathcal{G}) < M\}} + q_2^{\nu_i(\mathcal{G})}(1 - q_2)^{dp - \nu_i(\mathcal{G})} 1_{\{\nu_i(\mathcal{G}) \geq M\}} \right\}
\]
\[
\sigma_i^2 \sim \text{Inv. Gamma}(\alpha_i, \beta_i/2) \text{ independently for } i = 1, \ldots, p,
\]

where $\sigma^2$ 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 non-zero 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 non-zero 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 non-zero 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 non-zero 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 non-zeros) of a realistic vector, will be referred to as the realistic model cutoff size. Models where the number of non-zero 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, i.e., models where the number of non-zero 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 Narisetty and He (2014); Shin et al. (2018); Cao et al. (2020)) 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

$$
\pi\text{pseudo}(\mathcal{G} | \mathcal{Y}) \propto \pi(\mathcal{G}) \int \int L\text{pseudo}(\Phi, \Sigma_{\epsilon} | y_i, i = 1, \ldots, p)\pi(\Phi, \Sigma_{\epsilon}, \sigma^2)\text{d}\Phi \prod_{i=1}^{p} \pi(\sigma_i^2)\text{d}\sigma_i^2
$$

$$
(15) \quad \propto \prod_{i=1}^{p} q_{\nu_i(\mathcal{G})}(1 - q_{\nu_i(\mathcal{G})})^{d_p - \nu_i(\mathcal{G})} \left( \frac{1}{\tau \sqrt{n}} \right)^{\nu_i(\mathcal{G})} \left| \frac{X'_{i}X_{i}}{n} + \frac{I_{\nu_i}}{n\tau^2} \right|^{-1/2} \left( S_i + \frac{\beta_i}{n} \right)^{-(q_1 + \alpha_i)},
$$

where

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

$$
S_i := \frac{y_i'y_i}{n} - \frac{y_i'X_i}{n} \left( \frac{X'_{i}X_{i}}{n} + \frac{I_{\nu_i}}{n\tau^2} \right)^{-1} \frac{X'_{i}y_i}{n}
$$

Note that the pseudo-posterior probabilities $\pi\text{pseudo}(\mathcal{G} | \mathcal{Y})$ can in principle be used to select the graph by computing the posterior mode defined by $\hat{\mathcal{G}} := \text{arg max}_{\mathcal{G}} \pi\text{pseudo}(\mathcal{G} | \mathcal{Y})$.

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 $Y_n := (X_{1:n,0}, \ldots, X_{n,n+d-1})$ be the set of observations for sample size $n$, which satisfy $X_{n,k} = \sum_{i=1}^{d} A_{i,0:n} X_{n,k-i} + \varepsilon_{n,k}$ for $d \leq k \leq n + d - 1$. The errors $\{\varepsilon_{n,k}\}_{k=d}^{n+d-1}$ are i.i.d. $\mathcal{N}(0, \Sigma_{\varepsilon,0})$. Further, $\{\Phi_{0:n}\}_{n \geq 1}$ denotes the sequence of the true coefficient matrices given by $\Phi_{0:n} := [A_{1,0:n} \ A_{2,0:n} \ldots \ A_{d,0:n}]$, and $\{\Sigma_{\varepsilon,0}\}_{n \geq 1}$ denotes the sequence of the true error covariance matrices. Let $\mathbb{P}_0$ denote the probability measure underlying the true model described above, and $G_0 = G_{0,n}$ the true underlying activity graph for the sparse coefficient matrix $\Phi_0$. The quantities $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ and $C_X$ are as defined in Section 2 with $\Phi_0$ and $\Sigma_{\varepsilon,0}$ as the underlying parameter values. The maximum number of non-null entries within the columns of $\Phi_{0:n}$ and its minimum signal strength are defined respectively by

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

and $s_n := \inf_{(i,j) : \Phi_{ij} \neq 0} |\Phi_{ij}|$. The total number of non-zero entries in $\Phi_{0:n}$ is denoted by $\delta_n$ i.e $\delta_n = \sum_{i=1}^{p} \nu_i$. For ease of exposition, we will henceforth denote $\Phi_{0:n}$ as $\Phi_0$, and $\Sigma_{\varepsilon,0}$ as $\Sigma_{\varepsilon,0}$, 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.** $\frac{1 + \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \sqrt{\frac{k_n \log dp}{n}} = o(1)$.

**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 $\lambda_1 < \lambda_{\min}(C_X)$, the maximum eigenvalue of any principal sub-matrix of $C_X$ of dimension at most $k_n$ is uniformly bounded by $\lambda_2$, and $\sigma_{\min} < \lambda_{\min}(\Sigma_{\varepsilon,0}) \leq \lambda_{\max}(\Sigma_{\varepsilon,0}) < \sigma_{\max}$.

For ease of exposition, we denote $4\pi \lambda_{\max}(\Sigma_{\varepsilon,0})^{1 + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}}$ by $\mathcal{B}_n$, and let $c$ be the universal constant appearing in the Hanson-Wright inequality of (Vershynin, 2020, 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 $M_n = \frac{\min(\sigma_{\min}/4,1)\min(\lambda_1^2/4,1)c}{16B_n} \frac{n}{\log dp}$, and the edge inclusion probabilities satisfy $q_{1,n} = (dp^2)^{-\sigma_{\min}^{-1}}$ and $q_{2,n} = q_{1,n}^{(\log n)/2}$ for large enough $n$, in model prior (14).

**Assumption A4 (Minimum signal strength).** $\frac{B^2_nk_n \log dp + \log n}{n\sigma_n^2} \to 0$.

Note that for our analysis, the slab variance $\tau^2$ 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 et al. (2018a).

- **Assumption A1** states that $p$ can increase at an appropriate sub-exponential 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 et al. (2018a). Note that we assume that the true parameter $\Phi_0$ is sparse and $k_n$, the maximum number (column-wise) of non-zero entries in $\Phi_0$ plays a role in determining the rate at which $p$ can increase with $n$. In Ghosh et al. (2018a), no sparsity is assumed for $\Phi_0$, and assumptions bounding the singular values of $\Phi_0$ are needed to establish (estimation) consistency of the posterior distribution of the coefficient matrix $\Phi$. Assumption A1 also ensures that the true VAR process is well-behaved (stability) and the part regarding $\mu_{\min}(\tilde{A})$ and $\mu_{\max}(\tilde{A})$ allows us to leverage crucial concentration inequalities in Ghosh et al. (2018a) to control the behavior of $X'E$ and $X'X$. 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 $X'X/n$ away from zero and infinity respectively with high probability. This assumption corresponds to Assumption B3 in Ghosh et al. (2018a).

- **Assumption A3** provides the realistic model cutoff size $M_n$ as a multiple of $\frac{n}{n\sigma_n^2 \log p}$. This is very similar to the cutoff sizes used in the context of simple linear regression (for example Narisetty and He (2014); Shin et al. (2018)), wherein the realistic model cutoff sizes are multiples of $\frac{n}{\log p}$. In our more complex setting, the extra factor of $B_n^2$ 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 et al. (2018)). Again, the power of $p$ contains an extra factor of $\mathcal{B}_n^2$, 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, i.e., the smallest entry (in absolute value) of $\Phi_0$. 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 $G_0$ 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 $\Phi, G, \Sigma, \epsilon$ satisfying Assumptions A1-A4, the following holds

$$\pi_{\text{pseudo}}(G = G_0 \mid \mathcal{Y}) \xrightarrow{P} 1 \quad \text{as } n \to \infty,$$

i.e., the probability mass placed by the pseudo posterior on the true activity graph $G_0$ converges to 1 as $n \to \infty$. In particular, the true activity graph $G_0$ 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 $\Phi, G, \Sigma, \epsilon$ satisfying Assumptions A1-A4, there exists a constant $K$ (not depending on $n$) such that

$$\mathbb{E}_0 \left[ \Pi_{\text{pseudo}} \left( \| \Phi - \Phi_0 \|_F > K \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\delta_n \log dp}{n} \mid \mathcal{Y}} \right) \right] \to 0 \quad \text{as } n \to \infty.$$

The estimation rate of $\frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\delta_n \log dp}{n} \mid \mathcal{Y}}$ 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 $e^{n/(\delta_n^2 k_n)}$) as compared to Basu and Michailidis (2015) (smaller order than $e^{n/(\delta_n^2 \delta_n)}$). Note that $\delta_n$ is the total number of non-zero entries in $\Phi_0$, which can be much larger than $k_n$ (the maximum number of entries in any column of $\Phi_0$). For example, if $\Phi_0$ 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 $\Phi$ 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 non-zero entries in $\Phi_0$ grows with $n$, but $k_n$ (the maximum number of non-zero 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 $\Phi, \mathcal{G}, \Sigma_\epsilon$ satisfying Assumptions A1-A3, there exists a constant $K$ (not depending on $n$) such that

$$\mathbb{E}_0 \left[ \Pi_{\text{pseudo}} \left( \left\| \Phi - \Phi_0 \right\|_F > K \frac{1 + \mu_{\max}(\tilde{\mathcal{A}})}{\mu_{\min}(\tilde{\mathcal{A}})} \sqrt{\frac{\delta_n k_n \log dp}{n}} \right) \right] \to 0 \quad \text{as } n \to \infty.$$  

The proof of the main technical results is provided in Appendix A. 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 et al. (2018). The proof strategy first bounds the Bayes factor of any non-true 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 et al. (2018): (a) consider three different cases depending on whether the non-true 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 $i^{th}$ variable in the VAR in a linear regression setup. The vector of observations for the $i^{th}$ variable in the directed acyclic graph model in Cao et al. (2018), 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 Narisetty and He (2014); Cao et al. (2018) either have a non-random 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 $X_i$ in (11) has dependencies across both its rows and columns, and also shares dependencies with the error vector $\xi_i$ 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 non-true 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 Narisetty and He (2014); Shin et al. (2018); Cao et al. (2018).

5. Estimation of Model Parameters. The estimation of the graph $G$ and the parameter matrix $\Phi$ is a two-step procedure. We first estimate $G$ using the mode of the (pseudo) posterior distribution on the activity graphs, i.e.,

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

Note that the space of activity graphs has $2^d p^2$ 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 et al. (2007)). This is because $\pi_{\text{pseudo}}(G \mid Y)$ 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 $G$ of cardinality $k$, let $\text{nbd}(G)$ denote the neighborhood containing the following three elements \{\$G^+\$, $G^0\$, $G^-$\},
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; i.e. \( S(\mathcal{G}) \sim \log \pi_{\text{pseudo}}(\mathcal{G} | \mathcal{Y}) \). At the \( t \)-th iteration we undertake the following steps:

**Step 1** Compute in parallel all \( S(\mathcal{G}) \) for all \( \mathcal{G} \in \text{nbd}(\mathcal{G}[t]) \), 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 \( \hat{\phi}_j \) which is the non-zero component of the \( j \)-th column of \( \Phi \) can easily be estimated using the corresponding posterior mean

\[
\hat{\phi}_j := \left( \mathbf{X}_j' \mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2} \right)^{-1} \mathbf{X}_j' \mathbf{y}_j.
\]

This is because

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

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

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

where \( \mathbf{P}_j := \mathbf{X}_j \left( \mathbf{X}_j' \mathbf{X}_j + \frac{\mathbf{I}_{\nu_j}}{\tau^2} \right)^{-1} \mathbf{X}_j' \), and \( \alpha_j, \beta_j \) are as in (14). Thus, posterior credible intervals can easily be constructed using direct sampling from the non-central matrix variate \( \text{t} \) 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 \( \max_{1 \leq i \leq 50} |\lambda_i(\mathbf{A}_1)| = 0.95 \) 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 $\Sigma_\epsilon = \sigma^2 I_p$ and adjusted so that the signal-to-noise ratio $\text{SNR} = \frac{\|A_1\|_F}{\sigma} = 2$. The following is a $11 \times 6$ submatrix (which we denote by $A_1$ within a parenthesis) of the complete true transition matrix $A_1$,

$$
(A_1) = \begin{pmatrix}
1.38^{\text{(1)}} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -0.99^{\text{(2)}} & 0 & 0 \\
0 & 0 & -1.86^{\text{(3)}} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -0.96^{\text{(4)}} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.33^{\text{(5)}}
\end{pmatrix}.
$$

The 5 circled entries - (1) through (5)- from submatrix $(A_1)$ are selected for subsequent considerations; i.e., 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 $A_1$ 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 non-zero 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 $\hat{A}_1$ by
the posterior mean is given by

\[
\hat{A}_1 = \begin{pmatrix}
1.318 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1.007 & 0 & 0 & 0 & 0 \\
0 & 0 & -1.853 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.978 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -0.336
\end{pmatrix}
\]

It can be seen that the estimates are very close to the true ones. Next, the average estimation error \( \| \hat{A}_1 - A_1 \| \), specificity and sensitivity are reported in the following comparison table, based on these 500 replicates.

<table>
<thead>
<tr>
<th></th>
<th>VAR-SSS</th>
<th>( \ell_1 )-LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Est. Error</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.99</td>
<td>0.97</td>
</tr>
</tbody>
</table>

where

\[
\text{Sensitivity (SN)} = \frac{\text{True Positive (TP)}}{\text{TP} + \text{False Negative (FN)}} \quad \text{Specificity (SP)} = \frac{\text{True Negative (TN)}}{\text{TN} + \text{False Positive (FP)}}
\]

From the above table it is quite clear that the VAR Shotgun Stochastic Search algorithm performs better than the frequentist \( \ell_1 \)-LL in terms of support recovery and the posterior mean is more accurate than the \( \ell_1 \) penalized estimate of \( \hat{A}_1 \).

Next, we examine posterior credible intervals for the 6 circled entries, together with frequentist confidence intervals obtained from an \( \ell_1 \)-penalized likelihood. Since the error process in this setting has covariance \( \Sigma_\epsilon = \sigma^2 I_p \) estimation of the VAR parameters can be
obtained by running \( p \) separate regressions (see Lin and Michailidis (2017)) of the form \( y_j = X \phi_j + \xi_j, \quad 1 \leq j \leq p \), where \( y_j \) and \( \phi_j \) are the \( j \)-th columns of the response matrix \( Y \) and parameter matrix \( \Phi \), 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 \texttt{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.

\begin{figure}[h]
\centering
\subfloat[Iteration no. 51]{
\includegraphics[width=0.4\textwidth]{fig1a.png}
}
\subfloat[Iteration no. 187]{
\includegraphics[width=0.4\textwidth]{fig1b.png}
}
\subfloat[Iteration no. 211]{
\includegraphics[width=0.4\textwidth]{fig1c.png}
}
\subfloat[Iteration no. 305]{
\includegraphics[width=0.4\textwidth]{fig1d.png}
}
\caption{Comparison between posterior credible and confidence intervals of the 5 circled VAR coefficients. The true values are denoted by the red dot.}
\end{figure}
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 we report 95-coverage percentage (i.e. % of how many times the corresponding interval contains the true parameter) using 500 replications.

<table>
<thead>
<tr>
<th>Coef. serial no.</th>
<th>C.I.</th>
<th>Credible region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83.5</td>
<td>88</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
<td>95.5</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>96.5</td>
</tr>
<tr>
<td>4</td>
<td>88.5</td>
<td>90.5</td>
</tr>
<tr>
<td>5</td>
<td>91</td>
<td>94</td>
</tr>
</tbody>
</table>

The results in the table 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 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.

<table>
<thead>
<tr>
<th></th>
<th>No parallelization</th>
<th>With parallelization</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 50$</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>$p = 250$</td>
<td>2.81</td>
<td>1.08</td>
</tr>
<tr>
<td>$p = 500$</td>
<td>11.23</td>
<td>3.59</td>
</tr>
<tr>
<td>$p = 750$</td>
<td>26.74</td>
<td>11.83</td>
</tr>
</tbody>
</table>

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 Erdos-Renyi example in Supplemental Section S6, we generate transition matrices $A_1$ and $A_2$ with non-zero 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 $A_i$’s are set to $0.90$. Further, for each $p$-dimensional VAR model (with lag $d = 1, 2$) the true transition matrices $A_1$ and $A_2$ 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 $X_{t+1} = A_1 X_t + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{Toep}(\rho = 0.5)$. The sample sizes $n_1, n_2$ and $n_3$ corresponding to different VAR models are given in the following table.

<table>
<thead>
<tr>
<th>Sample sizes</th>
<th>Dimension $p$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small VAR</td>
<td>100</td>
<td>25</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>Medium VAR</td>
<td>250</td>
<td>30</td>
<td>75</td>
<td>200</td>
</tr>
<tr>
<td>Big VAR</td>
<td>500</td>
<td>50</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>Large VAR</td>
<td>750</td>
<td>75</td>
<td>200</td>
<td>400</td>
</tr>
</tbody>
</table>

For the VAR-SSS algorithm, the initial activity graph $G_0$ is selected based on an $\ell_1$ penalized least squares estimate ($\ell_1$-LS) which does not use $\Sigma_\epsilon$ and is given by

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

Note that the Bayesian SS algorithm has several hyper-parameters - $\tau^2$, $q$, $\alpha_i$ and $\beta_i$ for $1 \leq i \leq p$. Throughout this analysis we set the slab variance to $\tau^2 = \log n / 2$ and $\alpha_i = 1$, $\beta_i = 2 \forall i$. The prior edge probability $q$ is set to $1/p$. 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 $\ell_1$ penalized log-likelihood estimation ($\ell_1$-LL) (Lin and Michailidis (2017)) in which the estimate of $\Phi$ is given by

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

This gives the maximum likelihood estimate of $\Phi$, based on an estimated sparse $\Sigma_\epsilon$, the latter obtained using the graphical lasso algorithm. The regularization parameter $\lambda_n$ for
both the $\ell_1$-LS and $\ell_1$-LL methods is chosen based on the BIC criterion. 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 $A_i$. Since the exact contemporaneous dependence is not of primary concern, we omit the numerical results for $\Sigma_e$. The results in Table 5 illustrate the selection performance between the VAR-SSS and $\ell_1$-LL methods with a contemporaneous correlated error structure.

Table 5: Sensitivity and Specificity in VAR($d = 1, 2$) with $\Sigma_e = \text{Toep} (\rho = 0.80)$.

<table>
<thead>
<tr>
<th></th>
<th>VAR-SSS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>$n_1$ SN</td>
<td>$n_2$ SN</td>
<td>$n_3$ SN</td>
<td>$n_1$ SP</td>
<td></td>
<td>$n_2$ SP</td>
<td>$n_3$ SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small VAR</td>
<td>$e_1$ 0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.95</td>
<td>0.93</td>
<td>0.96</td>
<td>0.90</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Medium VAR</td>
<td>$e_1$ 0.972</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.96</td>
<td>0.96</td>
<td>0.93</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.92</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Big VAR</td>
<td>$e_1$ 0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.90</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.89</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.90</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Large VAR</td>
<td>$e_1$ 0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.96</td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.91</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.89</td>
<td>0.91</td>
<td>0.90</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\ell_1$-LL</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>$n_1$ SN</td>
<td>$n_2$ SN</td>
<td>$n_3$ SN</td>
<td>$n_1$ SP</td>
<td></td>
<td>$n_2$ SP</td>
<td>$n_3$ SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small VAR</td>
<td>$e_1$ 0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.93</td>
<td>1.00</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.92</td>
<td>0.91</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Medium VAR</td>
<td>$e_1$ 0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.98</td>
<td>0.90</td>
<td>0.95</td>
<td>0.93</td>
<td>0.98</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.96</td>
<td>0.92</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.92</td>
<td>0.91</td>
<td>0.98</td>
</tr>
<tr>
<td>Big VAR</td>
<td>$e_1$ 0.91</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.90</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.96</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.90</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td>0.89</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Large VAR</td>
<td>$e_1$ 0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.96</td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_2$ 0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.95</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$e_3$ 0.89</td>
<td>0.91</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
</tr>
</tbody>
</table>
It can be seen that the VAR-SSS algorithm *matches* the performance of the maximum likelihood method ($\ell_1$-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, i.e. $\| \Phi_0 - \hat{\Phi}_1 \|_F$ for both the VAR-SSS (posterior mean) and $\ell_1$-LL estimates.

Table 6: Estimation error: $\| \hat{\Phi} - \Phi_0 \|_F$ in VAR ($d = 1, 2$) with $\Sigma_t = \text{Toep} (\rho = 0.80)$.

<table>
<thead>
<tr>
<th>Density</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$\ell_1$-LL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lag $d = 1$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small VAR ($p = 100$)</td>
<td>$\epsilon_1$</td>
<td>0.88</td>
<td>0.79</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>0.95</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>Medium VAR ($p = 250$)</td>
<td>$\epsilon_1$</td>
<td>0.9</td>
<td>0.81</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>0.93</td>
<td>0.84</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.1</td>
<td>1</td>
<td>0.89</td>
</tr>
<tr>
<td>Big VAR ($p = 500$)</td>
<td>$\epsilon_1$</td>
<td>0.95</td>
<td>0.85</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>1.1</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.79</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>Large VAR ($p = 750$)</td>
<td>$\epsilon_1$</td>
<td>0.93</td>
<td>0.84</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>1.32</td>
<td>1.23</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.92</td>
<td>1.82</td>
<td>1.72</td>
</tr>
<tr>
<td><strong>Lag $d = 2$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small VAR ($p = 50$)</td>
<td>$\epsilon_1$</td>
<td>0.91</td>
<td>0.82</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>0.93</td>
<td>0.84</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>0.98</td>
<td>0.88</td>
<td>0.79</td>
</tr>
<tr>
<td>Medium VAR ($p = 125$)</td>
<td>$\epsilon_1$</td>
<td>0.93</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>0.96</td>
<td>0.87</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.13</td>
<td>1.03</td>
<td>0.93</td>
</tr>
<tr>
<td>Big VAR ($p = 250$)</td>
<td>$\epsilon_1$</td>
<td>0.98</td>
<td>0.89</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>1.13</td>
<td>1.03</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.81</td>
<td>1.73</td>
<td>1.61</td>
</tr>
<tr>
<td>Large VAR ($p = 400$)</td>
<td>$\epsilon_1$</td>
<td>0.96</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_2$</td>
<td>1.34</td>
<td>1.25</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_3$</td>
<td>1.95</td>
<td>1.86</td>
<td>1.76</td>
</tr>
</tbody>
</table>
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 non-zero entries in $A_1$, the less accurate the posterior mean and $\ell_1$-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 $A_1$ and $A_2$) smaller and setting it equal to 0.70.

Table 7: Sensitivity and Specificity in VAR($d = 1, 2$) with $\Sigma_c = \text{Toep} (\rho = 0.80)$

<table>
<thead>
<tr>
<th>Density</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>Small VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Medium VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Big VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Large VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Density</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>0.92</td>
<td>0.95</td>
<td>0.93</td>
<td>0.96</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.9</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>Small VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.91</td>
<td>0.94</td>
<td>0.92</td>
<td>0.96</td>
<td>0.93</td>
<td>0.98</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.9</td>
<td>0.92</td>
<td>0.91</td>
<td>0.94</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Medium VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.9</td>
<td>0.91</td>
<td>0.92</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Big VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.92</td>
<td>0.95</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.9</td>
<td>0.91</td>
<td>0.93</td>
<td>0.95</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.89</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Large VAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8: Estimation error: $\|\hat{\Phi} - \Phi_0\|_F$ in VAR ($d = 1, 2$) with $\Sigma_\epsilon = \text{Toep} (\rho = 0.80)$.  

<table>
<thead>
<tr>
<th>Lag $d = 1$</th>
<th>Density</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VAR-SSS</td>
<td>$\ell_1$-LL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_3$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_3$</td>
<td></td>
</tr>
<tr>
<td>Small VAR ($p = 100$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.89</td>
<td>0.8</td>
<td>0.71</td>
<td>0.9</td>
<td>0.8</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.92</td>
<td>0.82</td>
<td>0.72</td>
<td>0.93</td>
<td>0.83</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>$e_3$</td>
<td>0.97</td>
<td>0.86</td>
<td>0.78</td>
<td>0.99</td>
<td>0.89</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>Medium VAR ($p = 250$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.92</td>
<td>0.82</td>
<td>0.73</td>
<td>0.93</td>
<td>0.83</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.94</td>
<td>0.86</td>
<td>0.76</td>
<td>0.94</td>
<td>0.85</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>$e_3$</td>
<td>1.12</td>
<td>1.01</td>
<td>0.91</td>
<td>1.12</td>
<td>1.02</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Big VAR ($p = 500$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.97</td>
<td>0.87</td>
<td>0.78</td>
<td>0.99</td>
<td>0.89</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>$e_2$</td>
<td>1.12</td>
<td>1.01</td>
<td>0.92</td>
<td>1.12</td>
<td>1.02</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>$e_3$</td>
<td>1.8</td>
<td>1.71</td>
<td>1.59</td>
<td>1.78</td>
<td>1.68</td>
<td>1.58</td>
<td></td>
</tr>
<tr>
<td>Large VAR ($p = 750$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$e_1$</td>
<td>0.94</td>
<td>0.86</td>
<td>0.75</td>
<td>0.94</td>
<td>0.83</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>$e_2$</td>
<td>1.34</td>
<td>1.24</td>
<td>1.13</td>
<td>1.31</td>
<td>1.21</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>$e_3$</td>
<td>1.93</td>
<td>1.83</td>
<td>1.73</td>
<td>1.92</td>
<td>1.81</td>
<td>1.71</td>
<td></td>
</tr>
</tbody>
</table>

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 |

With reduced spectral radius the entries in $A_1$ and $A_2$ are of smaller magnitude and thus in Tables 7 and 8 the performance of both the VAR-SSS and $\ell_1$-LL deteriorates compared to the results depicted in Tables 5 and 6 in terms of model selection and estimation accuracy.
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 $\{x_{1,i}, x_{2,i}, \cdots, x_{p,i}\}$ and a response $y_i, i = 1, 2, \cdots, n$. The $n \times 1$ response vector $y$ is modeled as

$$y = X\beta + \epsilon \quad \text{where } \epsilon \sim N_n(0, \Sigma_\epsilon),$$

with $X$ being the $n \times p$ design matrix whose $i$-th row $X^i = (x_{1,i}, x_{2,i}, \cdots, x_{p,i})'$ contains the predictors for the $i$-th observation, $\beta = (\beta_1, \cdots, \beta_p)'$ is the vector of corresponding regression coefficients and $\epsilon = (\epsilon_i)_{i=1}^n$ 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 $f_X(\theta)$, i.e. the predictors are identically (but not necessarily independently) distributed. Further, we assume the error process $\{\epsilon_i\}$ to be a univariate centered, covariance-stationary Gaussian process, with finite spectral density $f_\epsilon(\theta)$. In addition, the errors are assumed to be independent of the predictors.

We now consider a high-dimensional setting for this linear regression model, where $p \gg n$. Let $\gamma_i$ indicate whether the $i$-th covariate is active in the model or not, i.e. $\gamma_i = 1(\beta_i \neq 0)$. The vector $\gamma$ 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

$$L(\beta, \Sigma_\epsilon | y_i, i = 1, \cdots, n) := \frac{1}{\sqrt{2\pi |\Sigma_\epsilon|^n}} \exp \left( -\frac{1}{2} (y - X\beta)'\Sigma_\epsilon^{-1} (y - X\beta) \right).$$

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

$$L_{\text{pseudo}}(\beta, \Sigma_\epsilon | y_i, i = 1, \cdots, n) = \frac{1}{\sqrt{2\pi \sigma^2_\epsilon}} \exp \left( -\frac{1}{2\sigma^2_\epsilon} \sum_{i=1}^n (y_i - X_i\beta)^2 \right).$$

The function $L_{\text{pseudo}}$ in (19) has a much simpler form than the likelihood function $L$ and only involves $\sigma^2_\epsilon$ (the diagonal entry of $\Sigma_\epsilon$). Let $\nu_\gamma$ denote the number of non-zero entries
in $\gamma$. We impose the following prior on $\beta, \sigma^2_\varepsilon$ and $\gamma$.

$$
\beta_j \mid \gamma, \sigma^2_\varepsilon \sim (1 - \gamma_j)1_{\beta_j = 0} + \gamma_j \mathcal{N}(0, \sigma^2_\varepsilon \tau^2)
$$

$$
\pi(\gamma) \propto q_1^{\nu_1} (1 - q_1)^{p - \nu_1} 1_{\{\nu_1 < M\}} + q_2^{\nu_2} (1 - q_2)^{p - \nu_2} 1_{\{\nu_2 \geq M\}}
$$

(20) \quad \sigma^2_\varepsilon \sim \text{Inv. Gamma}(\alpha, \beta/2).

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.

$$
\pi_{\text{pseudo}}(\gamma \mid \mathcal{Y}) \propto (1 - q_{\nu_1})^{p - \nu_1} \left( \frac{q_{\nu_1}}{\tau \sqrt{n}} \right)^{\nu_1} \left| \frac{X'_\gamma X\gamma}{n} + \frac{I_{\nu_1}}{n \tau^2} \right|^{1/2} \left( S_\gamma + \frac{\beta}{n} \right)^{-\left( \frac{p + \alpha}{2} \right)},
$$

where $q_{\nu_1} = q_1$ if $\nu_1 \leq M$, and $q_2$ if $\nu_1 > M$. $X_\gamma$ represents the sub-matrix of $X$ including columns corresponding to the active indices in $\gamma$, $S_\gamma := \frac{y'y}{n} - \frac{y'X\gamma}{n} \left( \frac{X'_\gamma X\gamma}{n} + \frac{I_{\nu_1}}{n \tau^2} \right)^{-1} \frac{X'_\gamma y}{n}$.

The posterior probabilities $\pi_{\text{pseudo}}(\gamma \mid \mathcal{Y})$ can be used to select the active covariates by computing the posterior mode defined by $\gamma := \arg\max_\gamma \pi_{\text{pseudo}}(\gamma \mid \mathcal{Y})$.

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 $y_n = (y_{i,n})_{i=1}^n$ is assumed to be generated from a true model with regression coefficient vector $\beta_0 = \beta_{0,n}$, and corresponding active covariate vector $\gamma_0$ such that

$$
y_n = X_n \beta_{0,n} + \varepsilon_n \quad \varepsilon_n \sim \mathcal{N}(0, \Sigma_\varepsilon, 0).
$$

The rows of $X = X_n$ are generated from a $p$-variate covariance-stationary Gaussian process with spectral density $f_X(\theta)$, and the entries of $\varepsilon_n$ are generated from a centered univariate stationary Gaussian process with spectral density $f_\varepsilon(\theta)$. Let $\nu_0 = \|\beta_{0,n}\|_0$ (i.e. the number of non-null entries in $\beta_{0,n}$) 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 $X_n$, and $\sigma^2_{0,n}$ denote the common variance of each coordinate of the stationary error process $\varepsilon_n$. For ease of exposition we will refer to $y_n$, $X_n$, $\beta_{0,n}$, $\varepsilon_n$, $\Omega_{0,n}$, $\sigma^2_{0,n}$ as $y$, $X$, $\beta_0$, $\varepsilon$, $\Omega_0$, $\sigma^2_0$ 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 Supplemental Section S8.

**Theorem 7.1 (Strong Selection Consistency).** For a centered stochastic linear regression model (17) with prior (20) on $\beta, \gamma, \sigma^2$ satisfying Assumptions B1-B5 (provided in Supplemental Section S7) the following holds

$$\pi_{\text{pseudo}}(\gamma = \gamma_0 \mid \mathcal{Y}) \xrightarrow{P} 1 \quad \text{as} \quad n \to \infty.$$

In particular, the true active covariate vector $\gamma_0$ 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 $\beta, \gamma, \Sigma_\epsilon$ satisfying Assumptions B1-B4 (provided in Supplemental Section S7) there exists a constant $K$ (not depending on $n$) such that

$$\mathbb{E}_0 \left[ \Pi_{\text{pseudo}} \left( \|\beta - \beta_0\|_2 > K \sqrt{\frac{\nu_0 \log p}{n}} \mid \mathcal{Y} \right) \right] \to 0 \quad \text{as} \quad n \to \infty.$$

**Remark 7.3.** The estimation consistency rate for stochastic regression (in Theorem 7.2) is of the order of $\sqrt{\nu_0 \log p/n}$. In other words, the radius of the neighborhood around $\beta_0$ on which the posterior asymptotically places all of its mass is of the order of $\sqrt{\nu_0 \log p/n}$, which does not include the stability measures $\mathcal{M}(f_\epsilon)$ and $\mathcal{M}(f_X)$. 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 these stability measures. 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.

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. We illustrate this by focusing on one of the important steps in the estimation consistency proof, which involves bounding $\|e'X_t\|/n$. Using
the independence of $\varepsilon$ and $X_t$ in the stochastic regression model, the term $\varepsilon'X_t(X'X_t)^{-1}X'/\varepsilon$ can be bounded by an appropriate $\chi^2$ random variable, and this can be used to show that $\|\varepsilon'X_t\|/n = o(\sqrt{\nu_0 \log p/n})$. 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 $\|\xi'X_i\|/n$. It can be shown that $\|\xi'X_i\|/n = O\left(\frac{1+\mu_{\text{max}}(\tilde{A})}{\mu_{\text{min}}(\tilde{A})} \sqrt{\frac{\nu_0 \log dp}{n}}\right)$ (see equation (22) in Appendix A).
Appendix A  Proofs for Theorem 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 $t_i = \{j_1, \ldots, j_{\nu t_i}\} \subset \{1, \ldots, dp\}$ be the set of column indices from $X$ corresponding to the non-zero positions of $\Phi_{@n, i}$ in the true model and it represents the neighbors of $i$ in the true activity graph $G_0$.

- Similarly $m_i = \{i_1, \ldots, i_{\nu m_i}\}$ represents the same for any candidate model (distinct from the true model). Given two activity graphs $G_0$ and $G_m$, $t_i = m_i$ implies the neighbors of $i$ are identical in both of the graphs (i.e. $G_0$ and $G_m$ have the same $i$-th column).

Let us define $d(m_i, t_i) := \text{Card}\left(\{t_i \cup m_i\} \setminus \{t_i \cap m_i\}\right)$ - number of disagreements in the $i$-th column between $G_m$ and $G_0$.

- Total number of disagreements is denoted by $D(m, t)$ and is equal to $\sum_{i=1}^p d(m_i, t_i)$.

Let $q_{\nu, n} := q_1\{\nu < M_n\} + q_2\{\nu \geq M_n\}$. Given the true activity graph $G_0$ and another arbitrary graph $G_m$, we have

$$
\frac{\pi(G_m \mid \mathcal{Y})}{\pi(G_0 \mid \mathcal{Y})} = \prod_{i=1}^p \left(1 - \frac{q_1}{\sqrt{n}}\right)^{\nu_{m_i} - \nu_{t_i}} q_{\nu_{m_i}}^{\nu_{m_i} - \nu_{m_i}} \left(1 - q_{\nu_{t_i}}\right)^{dp - \nu_{m_i}} \frac{X_{m_i}^t X_{m_i}}{n \nu_{m_i}} + \frac{1_{\nu_{m_i}}}{n \tau^2} \right)^{-1/2} \frac{S_{\nu_{m_i}} + \frac{\beta}{n}}{S_{\nu_{t_i}} + \frac{\beta}{n}} \left(\nu_{m_i} + \beta_i\right)^{-\frac{(n+1)}{2}}.
$$

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

$$
\left(1 - \frac{q_2}{q_1}\right)^{dp^2} \leq (1 - q_1)^{-dp^2} \leq e^{dp^2 \frac{q_1}{1-\tau^2}} \leq 2
$$
for large enough $n$. Note that $\nu_i \leq k_n < M_n$ for large enough $n$, and $q_2 < q_1$. Hence,

$$\frac{1}{2} \pi(G_m | \mathcal{Y}) \leq \prod_{i=1}^p \left( \frac{2q_{\nu_i}}{\tau\sqrt{n}} \right)^{\nu_{m_i} - \nu_i} \left( \frac{X_m X_m}{n} + \frac{I_{\nu_{m_i}}}{n\tau^2} \right)^{-1/2} \left( \frac{S_{\nu_i} + \frac{\alpha_i}{n}}{S_{\nu_{m_i}} + \frac{\delta_i}{n}} \right)^{\frac{q_2 + \alpha_i}{2}}$$

(*)

$$=: \prod_{i=1}^p B(m_i, t_i)$$

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 $t_i = m_i$, $B(m_i, t_i) = 1$. Recall that for any $m \subseteq \{1, 2, \cdots, dp\}$, $X_m$ denotes the sub-matrix of the columns of $X$ corresponding to the indices in $m$.

Before proving the main result (Theorem 4.1) - a straightforward application of Corollary A.4 - first note that by Assumption A2 there exist $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$ and $0 < \lambda_1 \leq \lambda_2 < \infty$ not depending on $n$ such that $\sigma_{\min} < \lambda_{\min}(\Sigma_{\epsilon, 0}) \leq \lambda_{\max}(\Sigma_{\epsilon, 0}) < \sigma_{\max}$, $\lambda_1 < \lambda_{\min}(C_X)$ and the maximum eigenvalue of any principal submatrix of $C_X$ of dimension at most $k_n$ is bounded above by $\lambda_2$. For ease of presentation, denote $4\pi \lambda_{\max}(\Sigma_{\epsilon, 0})^{1+\mu_{\max}(A)} \mu_{\min}(A)$ by $B_n$. Define next the following events:

$$G_{1,n} := \left\{ \left\| X'X - C_X \right\|_{\max} \leq B_n \sqrt{\frac{2\log dp}{n}} \right\},$$

$$G_{2,n} := \left\{ \left\| X'\mathcal{E} \right\|_{\max} \leq B_n \sqrt{\frac{\log dp^2}{n}} \right\},$$

$$\Xi_{1,n} := \left\{ \frac{7\sigma_{\min}}{8} \leq \frac{\xi'\xi}{n} \leq \frac{3\sigma_{\max}}{2} \right\},$$

$$\Xi_{2,n} := \bigcap_{m:1 \leq |m| < 2M_n} \left\{ \left\| \frac{X_m X_m}{n} - C_{X_m} \right\|_2 \leq 4B_n \sqrt{\nu_m \log dp \frac{\log c}{cn}} \right\},$$

where $P_m$ denotes the projection matrix for the column space of $X_m$, 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

$$\mathbb{P}(G_{1,n} \cap G_{2,n} \cap \Xi_n) \to 1 \quad \text{as} \quad n \to \infty.$$
Note that by Lemma B2, Proposition B1 and Proposition B3 of Ghosh et al. (2018b), with $u = e_i$ and $v = e_j$ - the $i$-th and $j$-th unit vector in $\mathbb{R}^{dp}$ respectively, there exists $c$ (not depending on $n$) such that

$$\mathbb{P}\left( \sup_{1 \leq i, j \leq dp} \left| e_i^t \left( \frac{X^t X}{n} - C_X \right) e_j \right| > \frac{\lambda_{\max}(\Sigma_e) \eta}{\mu_{\min}(\mathcal{A})} \right) \leq 2 \exp \left( -cn \min \left\{ \frac{\eta^2}{4}, \frac{\eta}{2} \right\} + 2 \log dp \right),$$

and again by taking unit vectors $u = e_i \in \mathbb{R}^{dp}$ and $v = e_j \in \mathbb{R}^p$

$$\mathbb{P}\left( \sup_{1 \leq i \leq dp, 1 \leq j \leq p} \left| e_i^t X E_n e_j \right| > 2\pi \lambda_{\max}(\Sigma_e) \left[ 1 + \frac{1 + \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \right] \eta \right) \leq 6 \exp \left( -cn \min \left\{ \frac{\eta^2}{4}, \frac{\eta}{2} \right\} + \log dp^2 \right).$$

Next, by setting an appropriate $\eta \sim \sqrt{\log dp^2 / cn}$ 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 $i^{th}$ diagonal entry of the true error covariance matrix $\Sigma_{\varepsilon,0n}$. Since $\xi_i / \sigma_{i,0} \sim \mathcal{N}_n(0, I_n)$, by applying Hanson-Wright inequality of Rudelson and Vershynin (2013) there exists $K$ such that

$$\mathbb{P}\left( \left| \frac{\xi_i^t \xi_i}{n \sigma_{i,0}^2} - 1 \right| > \frac{1}{8} \right) \leq 2e^{-cn/64K^4}.$$ 

That is $\frac{7\sigma_{i,0}^2}{8} \leq \frac{\xi_i^t \xi_i}{n} \leq \frac{3\sigma_{i,0}^2}{2}$ happens with probability at least $1 - 2e^{-cn/64K^4}$. By the union-sum inequality and Assumption B1, we get $\mathbb{P}(\Xi_{1,n}) \geq 1 - 2pe^{-cn/64K^4} \to 1$ as $n \to \infty$.

Note that for every $m$ such that $1 \leq |m| \leq 2M_n$, by Lemma B2 and Proposition B1 of Ghosh et al. (2018b), there exists $c$ (not depending on $n$) such that

$$(23) \quad \mathbb{P}\left( \left\| \frac{X^t_n X_m}{n} - C_{X_m} \right\|_2 \geq 4\mathcal{B}_n \sqrt{\frac{\nu_m \log dp}{cn}} \right) \leq e^{4\nu_m \log dp + 2\nu_m \log 21} \leq (dp)^{-3\nu_m}$$
for large enough $p$. It follows by the union-sum inequality that

$$
\mathbb{P}\left( \Xi_{2,n}^c \right) \leq \sum_{m:1\leq |m|<2M_n} (dp)^{-3\nu_m}
= \sum_{j=1}^{2M_n} \binom{dp}{j} (dp)^{-3j}
\leq \sum_{j=1}^{\infty} (dp)^{-2j}
= \frac{(dp)^{-2}}{1-(dp)^{-2}}
= \frac{1}{(dp)^{2}-1} \rightarrow 0 \text{ as } n \rightarrow \infty
$$

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

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 $B(m_i, t_i)$ 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 $i (t_i)$ fully contain the neighbors in the candidate model $(m_i)$.

**Lemma A.1.** If $m_i \subset t_i$ then there exists $N_1$ (not depending on $i$ and $\mathcal{Z}_m$) such that for all $n \geq N_1$ we have $B(m_i, t_i) \leq q_{1,n}^{(\nu_{t_i} - \nu_{m_i})/2}$.

Lemma A.2 deals with the case when the true active neighbors of $i (t_i)$ are fully contained in the neighbors of the candidate model $(m_i)$.

**Lemma A.2.** If $t_i \subset m_i$ then there exists $N_2$ (not depending on $i$ and $\mathcal{Z}_m$) such that for all $n \geq N_2$ we have $B(m_i, t_i) \leq q_{1,n}^{(\nu_{m_i} - \nu_{t_i})/2}$.

**Lemma A.3.** If $\mathcal{Z}_m$ is such that

$t_i \neq m_i$
$t_i \not\subset m_i$
$t_i \not\supset m_i$


then there exists \( N_6 \) (not depending on \( \mathcal{G}_m \) and \( i \)) such that for all \( n \geq N_6 \), \( B(m_i, t_i) \leq q_{1,n}^{\nu_{m_i} - \nu_{t_i}/2} \) if \( \nu_{m_i} > \nu_{t_i} \), and \( B(m_i, t_i) \leq q_{1,n}^{1/2} \) if \( \nu_{m_i} \leq \nu_{t_i} \).

The proofs of the above lemmas are given in Section S3 of the Supplemental document. Next, we employ these results to provide a bound on the ratio of the posterior probability for a non-true activity graph \( \mathcal{G}_m \) and the true activity graph \( \mathcal{G}_0 \).

**Corollary A.4.** For any centered Gaussian stable VAR(d) model (11) with prior (14) on \( \Phi, \mathcal{G}, \Sigma_e \) satisfying Assumptions A1-A4, for any “non-true” 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 i.e., \( D(m,t) = \sum_{i=1}^p d(m_i, t_i) \), and \( d(m_i, t_i) \) denotes the number of disagreements between \( m_i \) and \( t_i \).

**Proof.** Note that \( d(m_i, t_i) \leq 2k_n(\nu_{m_i} - \nu_{t_i}) \) for \( \nu_{m_i} > \nu_{t_i} \), and \( d(m_i, t_i) \leq 2k_n \) for \( \nu_{m_i} \leq \nu_{t_i} \). We will assume without loss of generality that the constants \( \sigma_{\text{min}} \) and \( \lambda_1 \) 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 \( t_i \neq m_i \) then for every \( n \geq \max\{N_1, \cdots, N_6\} \)

\[
B(m_i, t_i) \leq (dp^2)^{-2d(m_i, t_i)}.
\]

As noted earlier for \( t_i = m_i \) we have \( B(m_i, t_i) = 1 \). Hence, for all sufficiently large \( n \),

\[
\frac{1}{2} \frac{\pi(\mathcal{G}_m \mid \mathcal{Y})}{\pi(\mathcal{G}_0 \mid \mathcal{Y})} \leq \prod_{i=1}^p B(m_i, t_i) = \prod_{t_i=m_i} B(m_i, t_i) \prod_{t_i \neq m_i} B(m_i, t_i) \leq (dp^2)^{-2D(m,t)}.
\]

\[\square\]

**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 \( n \),
\[
\frac{1 - \pi(G_0 \mid \mathcal{Y})}{\pi(G_0 \mid \mathcal{Y})} = \sum_{G_m \not= G_0} \frac{\pi(G_m \mid \mathcal{Y})}{\pi(G_0 \mid \mathcal{Y})} \\
= \sum_{j=1}^{dp^2} \sum_{G_m \not= G_0} \frac{\pi(G_m \mid \mathcal{Y})}{\pi(G_0 \mid \mathcal{Y})} 1\{D(m,t) = j\} \\
\leq \sum_{j=1}^{dp^2} \left( \frac{dp^2}{j} \right) (dp^2)^{-2j} \\
\leq \sum_{j=1}^{dp^2} (dp^2)^{-j} \\
\leq \frac{(dp^2)^{-1}}{1 - (dp^2)^{-1}} \\
= \frac{1}{dp^2 - 1}.
\]

It follows that \( \frac{1 - \pi(G_0 \mid \mathcal{Y})}{\pi(G_0 \mid \mathcal{Y})} \to 0 \) as \( n \to \infty \). \( \square \)

### A.2 Proof of Theorem 4.2
For notational convenience, denote by \( \Pi_{\text{pseudo}} \) and \( \pi_{\text{pseudo}} \), \( \Pi_n \) and \( \pi_n \), respectively. For any \( n \times s \) submatrix \( X_s \) of \( X \), \( P_\nu \) stands for the projection matrix \( X_s (X_s'X_s)^{-1} X_s' \) onto \( \mathcal{C}(X_s) \) and by \( \tilde{P}_\nu \) is defined to be \( X_s \left( X_s'X_s + \frac{I}{\sigma_i^2} \right)^{-1} X_s' \).

First note that for any \( \eta > 0 \)
\[
E_0(\Pi_n \{ ||\Phi - \Phi_0||_F \geq K\eta \mid \mathcal{Y} \}) = \sum_{\mathcal{G}} E_0(\Pi_n \{ ||\Phi - \Phi_0||_F \geq K\eta \mid \mathcal{Y}, \mathcal{G} \} \pi_n(\mathcal{G} \mid \mathcal{Y})) \\
\leq E_0(\Pi_n \{ ||\Phi - \Phi_0||_F \geq K\eta \mid \mathcal{Y}, \mathcal{G}_0 \}) + E_0 \Pi_n(\mathcal{G} \not= \mathcal{G}_0 \mid \mathcal{Y})
\]

By Theorem 4.1, it is enough to prove that \( E_0(\Pi_n \{ ||\Phi - \Phi_0||_F \geq K\eta \mid \mathcal{Y}, \mathcal{G}_0 \}) \to 0 \) as \( n \to \infty \). Henceforth all the analysis is restricted to the true activity graph \( \mathcal{G}_0 \). Thus for ease of exposition we shall use \( X_t, \nu_t \) and \( \tilde{P}_t \) to denote \( X_t, \nu_t \) and \( \tilde{P}_{\nu_t} \) respectively. Next, under the prior on \( \Phi, \mathcal{G} \) and \( \Sigma_e \) in (14) we have
\[
\tilde{\phi}_i \mid \mathcal{G}_0, \sigma_i^2, \mathcal{Y} \sim \mathcal{N}_{\nu_t} \left( \tilde{\phi}_i, \sigma_i^2 \left( X_t'X_t + \frac{I_{\nu_t}}{\tau_2^2} \right)^{-1} \right)
\]
\[
(24) \quad \sigma_i^2 \mid \mathcal{G}_0, \mathcal{Y} \sim \text{Inv. Gamma} \left( \frac{n}{2} + \alpha_i, \frac{\nu_i + \beta_i}{2} \right).
\]
Also note that
\[
E_0 \left( \Pi_n \left\{ \| \Phi - \Phi_0 \|_F \geq K \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\delta_n \log dp}{n}} \bigg| \mathcal{Y}, \mathcal{G}_0 \right\} \right)
\]
\[
= E_0 \left( \Pi_n \left\{ \| \Phi - \Phi_0 \|_F^2 \geq K^2 \left( \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \right)^2 \frac{\sum_{i=1}^p \nu_i \log dp}{n} \bigg| \mathcal{Y}, \mathcal{G}_0 \right\} \right)
\]
\[
= E_0 \left( \Pi_n \left\{ \sum_{i=1}^p \| \tilde{\Phi}_i - \tilde{\Phi}_0 \|_2^2 \geq \sum_{i=1}^p K^2 \left( \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \right)^2 \frac{\nu_i \log dp}{n} \bigg| \mathcal{Y}, \mathcal{G}_0 \right\} \right)
\]
\[
\leq p \max_{1 \leq i \leq p} E_0 \left( \Pi_n \left\{ \| \tilde{\Phi}_i - \tilde{\Phi}_0 \|_2 \geq K \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\nu_i \log dp}{n}} \bigg| \mathcal{Y}, \mathcal{G}_0 \right\} \right).
\]

The proof will conclude by establishing
\[
\max_{1 \leq i \leq p} E_0 \left( \Pi_n \left\{ \| \tilde{\Phi}_i - \tilde{\Phi}_0 \|_2 \geq K \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\nu_i \log dp}{n}} \bigg| \mathcal{Y}, \mathcal{G}_0 \right\} \right) \leq 5 e^{-2 \log dp}.
\]

For ease of exposition we denote \( \frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\frac{\nu_i \log dp}{n}} \) by \( \eta_{n,i} \). Since we are only dealing with non-zero components of \( \phi_i \) in the true model we simplify further the notation of \( \tilde{\phi}_i \) and \( \tilde{\phi}_0 \), respectively, by \( \phi_i \) and \( \phi_0 \). In order to prove the above claim we first note that
\[
E_0 \Pi_n \left( \| \phi_i - \phi_0 \| \geq K \eta_{n,i} \big| \mathcal{Y}, \mathcal{G}_0 \right)
\]
\[
\leq E_0 \Pi_n \left( \| \phi_i - \hat{\phi}_i \| \geq \frac{K \eta_{n,i}}{2} \big| \mathcal{Y}, \mathcal{G}_0 \right) + P_0 \left( \| \hat{\phi}_i - \phi_0 \| \geq \frac{K \eta_{n,i}}{2} \big| \mathcal{Y}, \mathcal{G}_0 \right)
\]
(25)

First note that,
\[
\| \phi_i - \hat{\phi}_i \| = \sigma_i \left( X_i' X_i + \frac{I_{\nu_i}}{\tau^2} \right)^{-1/2} \left( X_i' X_i + \frac{I_{\nu_i}}{\tau^2} \right)^{1/2} \left( \sigma_i \frac{\zeta_i}{\lambda_{\min}(X_i' X_i)} \right)
\]
\[
\leq \frac{\sigma_i}{\sqrt{\lambda_{\min}(X_i' X_i)}} \| z \|,
\]
where \( z \) is \( \nu_i \times 1 \) standard normal random vector and the last step follows from (24). Thus,
for any $M^* > 0$

$$
E_0 \Pi_n \left( \|\phi_i - \hat{\phi}_i\| \geq \frac{K \eta_{n,i}}{2} \mid \mathcal{Y} \right)
$$

(26)

$$
\leq P \left( \lambda_{\min} \left( \frac{X_i' X_i}{n} \right) < \frac{\lambda_1}{2} \right) + P \left( \|z\| > \sqrt{n} \eta_{n,i} K \sqrt{\frac{\lambda_1}{8(M^*)^2}} \right) + E_0 \Pi_n \left( \sigma_i > M^* \mid \mathcal{Y} \right).
$$

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

$$
\frac{y_i' (I_n - \tilde{P}_i) y_i + \beta_i}{2n} \leq \frac{\xi_i' (I_n - P_i) \xi_i}{2n} + \frac{\phi_0' X_i (I_n - P_i) X_i' \phi_0}{2n} + \frac{2 \phi_0' X_i (I_n - P_i) \xi_i}{2n} + \frac{\beta_i}{2n} + \frac{c_0}{2n}
$$

$$
\leq \frac{\xi_i' (I_n - P_i) \xi_i}{2n} + \frac{\beta_i}{2n} + \frac{c_0}{2n}
$$

$$
\leq 3 \sigma_{\max},
$$

where (i) is a direct application of Lemma S1.3 and (S4) in the Supplemental document. (ii) is obtained by substituting the true model $y_i = X_i \phi_0 + \xi_i$. Therefore the scale parameter $\frac{y_i' (I_n - \tilde{P}_i) y_i + \beta_i}{2n}$ in the distribution of $\sigma_i^2 \mid \mathcal{G}_0, \mathcal{Y}$ is of order $n$. By choosing $M^*$ properly (see Remark S1.1), we can make $E_0 \Pi_n \left( \sigma_i > M^* \mid \mathcal{Y} \right) \leq e^{-2 \log dp}$ 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 (2010) 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 $t > 0$

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

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 $e^{-2 \log dp}$ for a large enough choice of $K$. Moving onto the second
term in (25), we use the following inequalities,
\[
\|\hat{\phi}_i - \phi_0\| = \left\| \left( \mathbf{X}_i' \mathbf{X}_i + \frac{\mathbf{I}_v}{\tau^2} \right)^{-1} \mathbf{X}_i' \mathbf{y}_i - \phi_0 \right\|
\leq \left\| \left( \mathbf{X}_i' \mathbf{X}_i + \frac{\mathbf{I}_v}{\tau^2} \right)^{-1} \left( \mathbf{X}_i' \mathbf{X}_i \phi_0 + \mathbf{X}_i' \mathbf{x}_i - \mathbf{X}_i' \mathbf{X}_i \phi_0 \right) \right\|
\leq \frac{1}{\lambda_{\min}(\mathbf{X}_i^2/n)} \left( \frac{\|\mathbf{X}_i' \mathbf{x}_i\| n}{n} + \frac{\|\phi_0\| n}{n^2} \right).
\]
Therefore
\[
\mathbb{P}_0 \left( \|\hat{\phi}_i - \phi_0\| > \frac{K_{\eta_{n,i}}}{2} \right) \leq \mathbb{P}_0 \left[ \frac{1}{n \lambda_{\min}(\mathbf{X}_i^2/n)} \left( \frac{\|\mathbf{X}_i' \mathbf{x}_i\| n}{n^2} + \frac{\|\phi_0\| n}{n^2} \right) \geq \frac{K_{\eta_{n,i}}}{2} \right]
\leq \mathbb{P}_0 \left[ \frac{1}{\lambda_{\min}(\mathbf{X}_i^2/n)} < \frac{\lambda_1}{2} \right] + \mathbb{P}_0 \left( \frac{\|\mathbf{X}_i' \mathbf{x}_i\| n}{n^2} + \frac{\|\phi_0\| n}{n^2} > \frac{K_{\eta_{n,i}}}{4} \right). \tag{27}
\]
First, we claim that any column of the true parameter matrix $\Phi_0$ is bounded in $\ell_2$ norm by a constant not depending on $n$. This follows from the stationarity and stability of the underlying process $\{X_t\}$. More precisely
\[
X_t = \Phi_0' \tilde{X}_t + \varepsilon_t \text{ where } \tilde{X}_t := [\left( X_{t-1} \right)' \cdots \left( X^{t-d} \right)']' \implies \Gamma_X(0) = \Phi_0' C_X \Phi_0 + \Sigma_{\varepsilon,0}
\]
By pre- and post-multiplication by the unit vector $e_i$ and using Assumption A2 the claim is now established. This implies $\frac{\|\phi_0\| n}{n^2} = o(\eta_{n,i})$ because
\[
\frac{\|\phi_0\| n}{n^2} \leq \frac{1}{\eta_{n,i}} \frac{\|\phi_0\| n}{n^2} \lambda \geq \frac{1}{\eta_{n,i}} \frac{\|\phi_0\| n}{n^2} \lambda \leq \frac{1}{\lambda_{\min}(\mathbf{X}_i^2/n)} \frac{\|\phi_0\| n}{n^2} = o(1).
\]
From (23) and by Assumption A1 there exists $n_4$ such that both the probabilities in (27) are at most $e^{-2 \log p}$ for a large enough choice of $K$. Hence, for appropriate choice of $K > 0$, we have
\[
\mathbb{E}_0 \Pi_n (\|\hat{\phi}_i - \phi_0\| \geq K_{\eta_{n,i}} | \mathcal{Y}) \leq \mathbb{E}_0 \Pi_n (\|\phi_i - \hat{\phi}_i\| \geq \frac{K_{\eta_{n,i}}}{2} | \mathcal{Y}) + \mathbb{P}_0 \left( \|\hat{\phi}_i - \phi_0\| \geq \frac{K_{\eta_{n,i}}}{2} | \mathcal{Y} \right)
\leq 3e^{-2 \log p} + 2e^{-2 \log p}
= \frac{5}{d^2 p^2}.
\]
This completes the proof. \hfill \Box
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 $\Pi_{\text{pseudo}}$ and $\pi_{\text{pseudo}}$, respectively, by $\Pi_n$ and $\pi_n$. For any $n \times s$ submatrix $X_s$ of $X$, $P_{\nu_s}$ stands for the projection matrix $X_s (X_s' X_s)^{-1} X'_s$. Note that $X_s$ is defined to be $X_s (X_s' X_s + \frac{1}{n} I_s)^{-1} X'_s$.

For ease of exposition, we denote $\frac{1 + \mu_{\text{max}}(\mathcal{A})}{\mu_{\text{min}}(\mathcal{A})} \sqrt{\nu_i k_n \log d_p / n}$ by $\eta_i$. Let $\mathcal{G}^i$ denote the $i^{\text{th}}$ column of the activity graph $\mathcal{G}$, for $1 \leq i \leq p$. It follows by (15) that $\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^p$ are mutually independent given $\mathcal{Y}$, and

$$\pi_n(\mathcal{G}^i \mid \mathcal{Y}) \propto \left( \frac{q_{\nu_i}(\mathcal{G})}{\tau \sqrt{n}(1 - q_{\nu_i}(\mathcal{G}))} \right)^{\nu_i(\mathcal{G})} \left| \frac{X'_s X_s}{n} + \frac{I_m}{n \tau^2} \right|^{-1/2} \left( S_i + \frac{\beta_i}{n} \right)^{-(\frac{p}{2} + \alpha_i)}.$$

By the arguments preceding (*), we obtain

$$\frac{1}{2} \frac{\pi_n(\mathcal{G}^i = m_i \mid \mathcal{Y})}{\pi_n(\mathcal{G}^i = t_i \mid \mathcal{Y})} \leq B(m_i, t_i).$$

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 $d(m_i, t_i) \leq 2k_n(\nu_{m_i} - \nu_{t_i})$ for $\nu_{m_i} > \nu_{t_i}$, It follows that

$$\frac{1}{2} \frac{\pi_n(\mathcal{G}^i = m_i \mid \mathcal{Y})}{\pi_n(\mathcal{G}^i = t_i \mid \mathcal{Y})} \leq (d_{m_i, t_i})^{-2d(m_i, t_i)}$$

$\nu_{m_i} > \nu_{t_i}$. 
Next, we focus on the case when $\nu_{m_i} \leq \nu_{t_i}$, with $m_i \neq t_i$. 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 $\|\phi_{t_i \cap m_i^c} \| \geq \nu_{t_i \cap m_i^c} s_i^2$), we get

\[
\frac{1}{2} \frac{\pi_n(\mathcal{F}^i = m_i \mid \mathcal{Y})}{\pi_n(\mathcal{F}^i = t_i \mid \mathcal{Y})} \leq B(m_i, t_i) \leq e^{-\log q_1, \nu_{t_i \cap m_i^c}} e^{-C_1 n \|\phi_{t_i \cap m_i^c}\|^2} \leq e^{C_2 m_i^2 - C_1 n \|\phi_{t_i \cap m_i^c}\|^2}
\]

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

\[
\frac{1}{2} \frac{\pi_n(\mathcal{F}^i = m_i \mid \mathcal{Y})}{\pi_n(\mathcal{F}^i = t_i \mid \mathcal{Y})} \leq e^{-4 \eta_{t_i}^2} \leq (dp)^{-4 \nu_{t_i} \eta_{t_i}} \leq dp^{-4 \eta_{t_i}}.
\]

Hence, we get

\[
\Pi_n \{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}\} = E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i \right\} \right) = \sum_{m_i: \nu_{m_i} \leq \nu_{t_i}} E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i = m_i \right\} \right) \pi_n(\mathcal{F}^i = m_i) + \sum_{m_i: \nu_{m_i} > \nu_{t_i}} E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i = m_i \right\} \right) \pi_n(\mathcal{F}^i = m_i)
\]

\[
\leq \max_{m_i: \nu_{m_i} \leq \nu_{t_i}, \|\phi_{t_i \cap m_i^c}\| \leq C \eta_{t_i}} E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i = m_i \right\} \right) + \pi_n(\mathcal{F}^i = t_i) \left( \sum_{m_i: \nu_{m_i} > \nu_{t_i}} \frac{\pi_n(\mathcal{F}^i = m_i)}{\pi_n(\mathcal{F}^i = t_i)} \right) + \sum_{m_i: \nu_{m_i} > \nu_{t_i}, \|\phi_{t_i \cap m_i^c}\| > C \eta_{t_i}} \pi_n(\mathcal{F}^i = m_i),
\]

\[
\leq \max_{m_i: \nu_{m_i} \leq \nu_{t_i}, \|\phi_{t_i \cap m_i^c}\| \leq C \eta_{t_i}} E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i = m_i \right\} \right) + \frac{dp}{2^j} \left( \frac{(dp)^2}{j} \right)^{-2j} + 2(dp)^{\nu_{t_i}} (dp)^{-4 \eta_{t_i}}
\]

\[
\leq \max_{m_i: \nu_{m_i} \leq \nu_{t_i}, \|\phi_{t_i \cap m_i^c}\| \leq C \eta_{t_i}} E_0 \left( \Pi_n \left\{\|\phi_i - \phi_0\|_2 > K \eta_{t_i} \mid \mathcal{Y}, \mathcal{F}^i = m_i \right\} \right) + \frac{(dp)^{-1}}{1 - (dp)^{-2}} + (dp)^{-1}.
\]
Next, we analyze the terms in (28) separately. For the first term, we note that
\[
\max_{1 \leq i \leq p, \nu_{m_i} \leq \nu_i, \|\phi_{0,t_i}^c \| \leq \tilde{C} \eta_i} E_0 \left( \Pi_n \left\{ \| \phi_i - \phi_0 \|_2 > K \eta_i \mid \mathcal{Y}, \mathcal{G}^i = m_i \right\} \right) \leq 2 e^{-2 \log dp}.
\]

Fix \( i \) and \( m_i \) such that \( \nu_{m_i} \leq \nu_i, \| \phi_{i,c,m_i^c}^c \| \leq \tilde{C} \eta_i \) arbitrarily. Let
\[
\hat{\phi}_{m_i} = \left( X_m' X_m + \frac{I_{\nu_{m_i}}}{\tau^2} \right)^{-1} X_m' y_i.
\]

Using \( \| \phi_{0,t_i,c,m_i^c}^c \| \leq \tilde{C} \eta_i \) and \( K > 6 \tilde{C} \), it follows that
\[
E_0 \Pi_n \left( \| \phi_i - \phi_0 \| \geq K \eta_i \mid \mathcal{Y}, \mathcal{G}^i = m_i \right)
\leq E_0 \Pi_n \left( \| \bar{\phi}_i - \hat{\phi}_{m_i} \| \geq \frac{K \eta_i}{2} \mid \mathcal{Y}, \mathcal{G}^i = m_i \right) +
E_0 \left( \| \hat{\phi}_{m_i} - \phi_{m_i} \| + \| \phi_{0,t_i,c,m_i^c}^c \| \geq \frac{K \eta_i}{2} \mid \mathcal{Y}, \mathcal{G}^i = m_i \right)
\]

(28)
\[
= E_0 \Pi_n \left( \| \bar{\phi}_i - \hat{\phi}_{m_i} \| \geq \frac{K \eta_i}{2} \mid \mathcal{Y}, \mathcal{G}^i = m_i \right) + 1 \left\{ \| \phi_{m_i} - \phi_{0,m_i} \| \geq \frac{K \eta_i}{4} \right\}.
\]

Next, we analyze the terms in (28) separately. For the first term, we note that
\[
\| \bar{\phi}_i - \hat{\phi}_{m_i} \| = \sigma_i \left( X_m' X_m + \frac{I_{\nu_{m_i}}}{\tau^2} \right)^{-1/2} \left( X_m' X_m + \frac{I_{\nu_{m_i}}}{\tau^2} \right)^{1/2} \left( \bar{\phi}_i - \hat{\phi}_i \right)
\leq \frac{\sigma_i}{\sqrt{\lambda_{\min}(X_m' X_m)}} \| z \|,
\]

where \( z \) is \( \nu_{m_i} \times 1 \) standard normal random vector and the last step follows from (16). Note that \( \nu_{m_i} \leq \nu_i \leq k_n \). By Assumption A1, for large enough \( n \), we have that on \( \Xi_{2,n} \)
\[
\left\| \frac{X_m' X_m}{n} - C_{X_m} \right\|_2 \leq \tilde{\mathcal{B}} n \sqrt{\frac{2 \nu_{m_i} \log dp}{n}} \leq \frac{\lambda_1}{2}.
\]

Using \( \nu_{m_i} \leq \nu_i \), it follows that for any \( M^*>0 \)
\[
E_0 \Pi_n \left( \| \bar{\phi}_i - \hat{\phi}_{m_i} \| \geq \frac{K \eta_i}{2} \mid \mathcal{Y}, \mathcal{G}^i = m_i \right)
\leq P \left( \| z \| > \sqrt{n \nu_{m_i}} K \sqrt{\frac{\lambda_1}{8(M^*)^2}} \right) + E_0 \Pi_n ( \sigma_i > M^* \mid \mathcal{Y} ).
\]

(29)
First, we claim that any column of the true parameter matrix $\Phi_0$ is bounded in $\ell_2$ norm by a constant not depending on $n$. This follows from the stationarity and stability of the underlying process $\{X^t\}$. More precisely

$$X^t = \Phi_0' \tilde{X}^t + \varepsilon^t \quad \text{where } \tilde{X}^t := ((X^{t-1})' \cdots (X^{t-d}))'$$

$$\Rightarrow \Gamma_X(0) = \Phi_0'C_X\Phi_0 + \Sigma_{\varepsilon,0}$$

By pre- and post-multiplication by the unit vector $e_i$ and using Assumption A2 the claim is now established. Note that the distribution of $\sigma_i^2 \varepsilon_i^t$ is Inverse-Gamma with shape parameter $\left(\frac{n}{2} + \alpha_i\right) \sim n$ and scale parameter $\frac{y_i'(I_n - \tilde{P}_{nu})y_i + \beta_i}{2}$. Hence,

$$\frac{y_i'(I_n - \tilde{P}_{nu})y_i + \beta_i}{2} \leq \frac{\xi_i'}{2} + \frac{\phi_0'y_i + \beta_i}{2}$$

for large enough $n$. Note that (i) is obtained by substituting the true model $y_i = X_i\phi_0 + \xi_i$, and (ii) follows by restricting to the event $G_{1,n} \cap G_{2,n} \cap \Xi_n$. Therefore, the scale parameter $\frac{y_i'(I_n - \tilde{P}_{nu})y_i + \beta_i}{2}$ in the distribution of $\sigma_i^2 | \mathcal{Y}, \mathcal{G}^i = m_i$ is bounded by a constant multiple of $n$. Hence, by Remark S1.1 in the Supplemental document, there exists a constant $M^*$ such that $\mathbb{E}_0\Pi_n (\sigma_i > M^* | \mathcal{Y}) \leq e^{-2\log dp}$ for all large $n$.

As for the first term in (29), note that $z$ is a $\nu_{m_i} \times 1$ vector with i.i.d. $\mathcal{N}(0, 1)$ entries. It follows from (Vershynin, 2010, Corollary 5.35) that for any $t > 0$

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

By setting $t^2 = 4 \log dp$, it follows that $\frac{\|z\|}{\sqrt{n}} \geq \eta_1 K \sqrt{\frac{\lambda_1}{8(M^*)^2}}$ with probability at most $e^{-2\log dp}$ for a large enough choice of $K$.

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

$$\|\phi_{m_i} - \phi_{0m_i}\| = \left\|\left(X_{m_i}'X_{m_i} + \frac{I_{m_i}}{\tau^2}\right)^{-1}X_{m_i}y_i - \phi_{0m_i}\right\|$$

$$\leq \left\|\left(X_{m_i}'X_{m_i} + \frac{I_{m_i}}{\tau^2}\right)^{-1}\left(X_{m_i}'X_{m_i}\phi_{0i} + X_{m_i}'\xi_i - X_{m_i}'X_{m_i}\phi_{0m_i} - \phi_{0m_i}\right)\right\|.$$

imsart-aos ver. 2014/10/16 file: VAR_model_selection_rev-GM.tex date: June 16, 2020
Since $\phi_{0_t \cap m_i^c} = 0$, $\|\phi_{t_i \cap m_i^c}\| \leq \tilde{C}\eta_{t_i}$, $\|\phi_{0_m_i}\|$ 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

$$\|\hat{\phi}_{m_i} - \phi_{0m_i}\| \leq \frac{2}{\lambda_1} \left( \frac{\|X_m' X_{t_i \cap m_i^c} \phi_{t_i \cap m_i^c}\|}{n} + \frac{\|X_{m_i}' \xi_i\|}{n} + \frac{\|\phi_{0m_i}\|}{n^2} \right) \leq 2 \lambda_2 \left( \frac{\nu_{t_i} \log d_p}{n} \right) \leq \tilde{C}_1 \eta_{t_i}.$$ 

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

$$1 \left\{ \|\hat{\phi}_{m_i} - \phi_{0m_i}\| \geq \frac{\kappa n_{t_i}}{3} \right\} = 0.$$

This completes the proof.

References.


Honoring the research contributions of Charles R. Nelson.


