

Statistica Sinica Preprint No: SS-2021-0151

Title	Statistical Inference for High-Dimensional Vector Autoregression with Measurement Error
Manuscript ID	SS-2021-0151
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0151
Complete List of Authors	Xiang Lyu, Jian Kang and Lexin Li
Corresponding Author	Lexin Li
E-mail	lexinli@berkeley.edu
Notice: Accepted version subject to English editing.	

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL VECTOR AUTOREGRESSION WITH MEASUREMENT ERROR

Xiang Lyu[†], Jian Kang[‡], and Lexin Li[†]

[†] *University of California at Berkeley, and* [‡] *University of Michigan*

Abstract: High-dimensional vector autoregression with measurement error is frequently encountered in a large variety of scientific and business applications. In this article, we study statistical inference of the transition matrix under this model. While there has been a large body of literature studying sparse estimation of the transition matrix, there is a relative paucity of inference solutions, especially in the high-dimensional setting. We study both global and simultaneous testing of the transition matrix. We first develop a new sparse expectation-maximization algorithm to estimate the model parameters, and carefully characterize the estimation precision. We then construct a Gaussian matrix, after proper bias and variance corrections, from which we derive the test statistics. Finally, we develop the testing procedures and establish their asymptotic guarantees. We study the finite-sample performance of our tests through intensive simulations, and illustrate with a neuroimaging based brain connectivity analysis.

Key words and phrases: Brain network analysis; covariance inference; expectation-maximization; simultaneous testing; global testing; vector autoregression.

1. Introduction

In this article, we study statistical inference for high-dimensional vector autoregression (VAR) with measurement error. We consider the model,

$$\begin{aligned}\mathbf{y}_t &= \mathbf{x}_t + \boldsymbol{\epsilon}_t, \\ \mathbf{x}_{t+1} &= \mathbf{A}_* \mathbf{x}_t + \boldsymbol{\eta}_t,\end{aligned}\tag{1.1}$$

where $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,p})^\top \in \mathbb{R}^p$ is the observed multivariate time series, $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,p})^\top \in \mathbb{R}^p$ is the multivariate latent signal that admits an autoregressive structure, $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \dots, \epsilon_{t,p})^\top \in \mathbb{R}^p$ is the measurement error for the observed time series, $\boldsymbol{\eta}_t = (\eta_{t,1}, \dots, \eta_{t,p})^\top \in \mathbb{R}^p$ is the white noise of the latent signal, and $\mathbf{A}_* = (A_{*,ij}) \in \mathbb{R}^{p \times p}$ is the sparse transition matrix that encodes the directional relations among the variables of \mathbf{x}_t . Furthermore, we focus on the scenario $\|\mathbf{A}_*\|_2 < 1$ such that the VAR model of \mathbf{x}_t is stationary. The error terms $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are i.i.d. multivariate normal with mean zero and covariance $\sigma_{\epsilon,*}^2 \mathbf{I}_p$ and $\sigma_{\eta,*}^2 \mathbf{I}_p$, respectively, and are independent of \mathbf{x}_t . Here we focus on the lag-1 autoregressive structure and homoscedastic errors. We later discuss potential extensions in Section 7.

Model like (1.1) is frequently employed in a variety of scientific and business applications, e.g., finance, engineering, and neuroscience. Our motivation is brain effective connectivity analysis based on functional magnetic resonance imaging (fMRI). The brain is a highly interconnected dynamic system, in which the activity and temporal evolution of neural elements are triggered and influenced by the activities of other elements (Garg et al., 2011). Of great interest in neuroscience is to understand

the directional relations among the neural elements through fMRI, which measures synchronized blood oxygen dependent signals at different brain locations. The VAR model is an important tool to model such relations, which are encoded by the transition matrix \mathbf{A}_* , while the stationarity is often assumed (Bullmore and Sporns, 2009; Chen et al., 2011). However, unlike a typical VAR model, the observed time series \mathbf{y}_t is the contaminated version of the true signal \mathbf{x}_t , added with a measurement error $\boldsymbol{\epsilon}_t$ (Zhang et al., 2015; Cao et al., 2019).

We address the statistical inference problem of the transition matrix \mathbf{A}_* under model (1.1), and we aim at a high-dimensional setting where p^2 exceeds the length of series T . We first test the global hypotheses,

$$\begin{aligned} H_0 : A_{*,ij} &= A_{0,ij}, \quad \text{for all } (i, j) \in \mathcal{S} && \text{versus} \\ H_1 : A_{*,ij} &\neq A_{0,ij}, \quad \text{for some } (i, j) \in \mathcal{S}, \end{aligned} \tag{1.2}$$

for a given $\mathbf{A}_0 = (A_{0,ij}) \in \mathbb{R}^{p \times p}$ and $\mathcal{S} \subseteq [p] \times [p]$, where $[p] = \{1, \dots, p\}$. A common choice is $\mathbf{A}_0 = \mathbf{0}_{p \times p}$ and $\mathcal{S} = [p] \times [p]$. We next test the simultaneous hypotheses,

$$H_{0;ij} : A_{*,ij} = A_{0,ij}, \quad \text{versus} \quad H_{1;ij} : A_{*,ij} \neq A_{0,ij}, \quad \text{for all } (i, j) \in \mathcal{S}. \tag{1.3}$$

There has been a large body of literature studying sparse estimation of \mathbf{A}_* in VAR models (Hsu et al., 2008; Song and Bickel, 2011; Negahban and Wainwright, 2011; Han et al., 2015, among many others). However, they all assumed that there is no measurement error $\boldsymbol{\epsilon}_t$, i.e., \mathbf{x}_t is fully observed. Besides, while both estimation and inference can produce a sparse representation of \mathbf{A}_* , they are utterly different problems.

Sparse estimation usually does not explicitly control the false discovery rate (type I error), and does not produce an explicit significance quantification (p -value). There has been a relative paucity of inference methods for \mathbf{A}_* in VAR models. Existing inference solutions mostly focused on the low-dimensional VAR setting. For instance, Staudenmayer and Buonaccorsi (2005) studied the inference for a one-dimensional VAR model. See also Tsay and Chen (2018) for a review. More recently, for the high-dimensional VAR setting, Krampe et al. (2018) proposed to bootstrap the de-biased Lasso estimator, while Zheng and Raskutti (2019) extended the de-correlated score test of Ning and Liu (2017). However, they only addressed the global testing problem (1.2), but not the simultaneous testing problem (1.3). Besides, it is unclear how to adapt their tests for additional measurement error. To our knowledge, there is no existing solution to directly address global and simultaneous testing problems under the high-dimensional VAR setting with error.

Our proposal is built on two key ingredients: a sparse expectation-maximization (EM) algorithm, and the high-dimensional covariance inference. The first ingredient, the EM algorithm, offers a way to estimate model parameters in the presence of measurement error. Early EM methods, however, only justified the convergence to a local optimum and did not consider sparsity. Recently, Balakrishnan et al. (2017) provided sufficient conditions to guarantee the convergence of the standard EM to the global optimum but only in a low-dimensional setting, while Cai et al. (2019) extended the guarantee to a high-dimensional sparse Gaussian mixture model. See also Wang et al. (2015); Yi and Caramanis (2015). On the other hand, they all worked with

i.i.d. observations, whereas our problem involves temporally highly dependent data. Extension from independent to dependent observations is not trivial. The second ingredient, the high-dimensional covariance inference, has been intensively studied in recent years, including both global testing (Chen et al., 2010; Cai and Jiang, 2011; Xiao and Wu, 2013) and simultaneous testing (Liu, 2013; Cai et al., 2013). However, they all assumed the data which the covariance is constructed from are fully observed. By contrast, our inference is about the transition matrix \mathbf{A}_* of the latent unobserved \mathbf{x}_t , and the covariance of the observed \mathbf{y}_t is a nonlinear transformation of \mathbf{A}_* , making it difficult to trace back to \mathbf{A}_* . Consequently, there is a considerable gap before we can apply the existing covariance inference tools to our setting.

In this article, we develop inferential procedures for both the global and simultaneous testing problems (1.2) and (1.3) under the high-dimensional VAR model with measurement error. Our proposal includes three main steps. First, we develop a new sparse EM algorithm to estimate relevant model parameters. Next, we construct a Gaussian matrix on the domain of transition matrix, from which we derive the test statistics. Finally, we develop the global and simultaneous testing procedures with proper theoretical guarantees.

In the first step, we develop a new sparse EM algorithm to estimate both the transition matrix \mathbf{A}_* and the error variances $\sigma_{\epsilon,*}^2$ and $\sigma_{\eta,*}^2$. In particular, the maximization step is done via a generalized Dantzig selector for the Yule-Walker equation, which can be efficiently solved by parallel linear programming (Candes and Tao, 2007; Han et al., 2015). We then establish the convergence of our sparse EM estimators

to the true parameters, within the statistical precision required for the test statistics and the transition matrix inferences in later steps. We note that, the existing EM theory adopts the log-likelihood in an infinite-sample scheme as the key analytical tool, which becomes an expectation at a single observation given i.i.d. observations (Balakrishnan et al., 2017; Cai et al., 2019). However, the temporal dependence in our model makes the expectation of the log-likelihood change with the sample size. To tackle the issue, we consider the expectation in a finite-sample scheme instead, which introduces additional technical difficulty. We then derive several new concentration inequalities to establish the statistical error under some weak sparsity assumptions.

In the second step, we construct a Gaussian matrix for test statistics for the transition matrix inference. This is built on a key observation that the inference on \mathbf{A}_* is equivalent to that on the lagged auto-covariance of some noise term. Since this noise is not directly observed, we employ the sparse EM algorithm in the first step to reconstruct the noise. We then study the non-asymptotic behavior of the sample lagged auto-covariance of the reconstructed noise, and explicitly characterize its bias and variance. This in turn leads to the construction of the test statistic matrix whose entries marginally follow a standard Gaussian distribution under the null.

In the third step, we develop a global testing procedure based on the extreme distribution of the maximal entry of the test statistic matrix from the second step, and develop a simultaneous testing procedure by thresholding at a level that controls false discovery rate (FDR). Theoretically, we obtain the asymptotic size and power of the global test, which together establish the consistency of our test. We also

show that our simultaneous test achieves a consistent FDR control. Our testing procedures are extensions of the covariance inference methods such as Cai and Jiang (2011); Liu (2013); Cai et al. (2013). But unlike the existing methods that are built on the sample covariance of fully observed data, our tests are obtained from the sample lagged auto-covariance of the reconstructed noise. This difference requires us to derive new concentration inequalities and Gaussian approximations to disentangle the reconstruction error, lag effect, and temporal dependence. These new theoretical results themselves are of independent interest.

We employ the following notation throughout the article. Let $|\mathcal{S}|$ denote the cardinality of a set \mathcal{S} . For a scalar $a \in \mathbb{R}$, let $\lceil a \rceil$ and $\lfloor a \rfloor$ denote the smallest and largest integer greater than or smaller than a . For two scalars $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maxima and minima. For a vector $\mathbf{a} = (a_1, \dots, a_p)^\top \in \mathbb{R}^p$, define $\|\mathbf{a}\|_1 = \sum_{i=1}^p |a_i|$, $\|\mathbf{a}\|_2 = (\sum_{i=1}^p |a_i|^2)^{1/2}$, and $\|\mathbf{a}\|_\infty = \max_{1 \leq i \leq p} |a_i|$. For an index set $\mathcal{S} \subseteq [p]$, let $\mathbf{a}_{\mathcal{S}}$ denote the sub-vector of \mathbf{a} containing only the coordinates indexed by \mathcal{S} . For a matrix $\mathbf{M} = (M_{ij}) \in \mathbb{R}^{p_1 \times p_2}$, define $\|\mathbf{M}\|_1 = \sum_{ij} |M_{ij}|$, $\|\mathbf{M}\|_2 = \lambda_{\max}^{1/2}(\mathbf{M}^\top \mathbf{M})$, $\|\mathbf{M}\|_F = (\sum_{ij} M_{ij}^2)^{1/2}$, $\|\mathbf{M}\|_{\max} = \max_{ij} |M_{ij}|$, $\|\mathbf{M}\|_{l_1} = \max_{j \in [p_2]} \sum_{i=1}^{p_1} |M_{ij}|$, $\|\mathbf{M}\|_{l_\infty} = \max_{i \in [p_1]} \sum_{j=1}^{p_2} |M_{ij}|$, and $\|\mathbf{M}\|_{r,2} = \max_{i \in [p_1]} \sqrt{\sum_{j=1}^{p_2} |M_{ij}|^2}$ to be its element-wise ℓ_1 norm, spectral norm, Frobenius norm, max norm, maximum absolute column sum, maximum absolute row sum, and maximal row-wise Euclidean norm, respectively. Let $\mathbf{M}_{i\cdot}$ and $\mathbf{M}_{\cdot j}$ denote the i th row and j th column. Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote its smallest and largest eigenvalue, $\text{tr}(\mathbf{M})$ the trace, and $|\mathbf{M}|$ the determinant. Define $D(\mathbf{M})$ as a diagonal matrix whose diagonal elements are those of \mathbf{M} .

The rest of the article is organized as follows. Section 2 presents the sparse EM algorithm and the estimation accuracy guarantees. Section 3 constructs the test statistic matrix. Section 4 develops the testing procedures and their theoretical guarantees. Section 5 presents simulations, and Section 6 illustrates with a brain functional network example. Section 7 concludes with a discussion. All proofs and additional numerical results are relegated to the Supplementary Appendix.

2. Sparse EM Estimation

2.1 Sparse EM algorithm

Let $\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^T$ denote the complete data, where T is the total number of observations, \mathbf{y}_t is observed but \mathbf{x}_t is latent. Let $\Theta = \{\mathbf{A}, \sigma_\eta^2, \sigma_\epsilon^2\}$ collect all the parameters of interest in model (1.1), and $\Theta_* = \{\mathbf{A}_*, \sigma_{\eta,*}^2, \sigma_{\epsilon,*}^2\}$ denote the true parameters. The goal is to estimate Θ_* by maximizing the log-likelihood function of the observed data, $\ell(\Theta|\{\mathbf{y}_t\}_{t=1}^T)$, with respect to Θ . The computation of $\ell(\Theta|\{\mathbf{y}_t\}_{t=1}^T)$, however, is highly nontrivial. The standard EM algorithm then turns to an auxiliary function, named the finite-sample Q -function,

$$Q_y(\Theta|\Theta') = \mathbb{E} \left[\ell(\Theta|\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^T) \mid \{\mathbf{y}_t\}_{t=1}^T, \Theta' \right],$$

which is defined as the expectation of the log-likelihood function for the complete data $\ell(\Theta|\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^T)$, conditioning on a parameter set Θ' and the observed data \mathbf{y}_t , and the expectation is taken with respect to the latent data \mathbf{x}_t . The Q -function can be computed efficiently, and provides a lower bound of the target log-likelihood

function $\ell(\Theta|\{\mathbf{y}_t\}_{t=1}^T)$ for any Θ , with equality if $\Theta = \Theta'$. Maximizing Q-function provides an uphill step of the likelihood. Starting from an initial set of parameters $\hat{\Theta}_0$, the EM algorithm then alternates between the expectation step (E-step), where the Q-function $Q_y(\Theta|\hat{\Theta}_k)$ conditioning on the parameters $\hat{\Theta}_k$ of the k th iteration is computed, and the maximization step (M-step), where the parameters are updated by maximizing the Q-function $\hat{\Theta}_{k+1} = \operatorname{argmax}_{\Theta} Q_y(\Theta|\hat{\Theta}_k)$.

For our problem, we carry out the E-step via the standard Kalman filter and smoother (Ghahramani and Hinton, 1996). For the M-step, the maximizer \mathbf{A} of $Q_y(\Theta|\hat{\Theta}_k)$ satisfies that $(T-1)^{-1} \sum_{t=1}^{T-1} \mathbf{E}_{t,t+1;k} = \{(T-1)^{-1} \sum_{t=1}^{T-1} \mathbf{E}_{t,t;k}\} \mathbf{A}^\top$, where $\mathbf{E}_{t,s;k} = \mathbb{E} \left\{ \mathbf{x}_t \mathbf{x}_s^\top | \{\mathbf{y}_{t'}\}_{t'=1}^T, \hat{\Theta}_{k-1} \right\}$ for $s, t \in [T]$ is obtained from the E-step. The standard EM algorithm directly inverts the matrix involving $\mathbf{E}_{t,t;k}$'s, which is computationally challenging when the dimension p is high. In addition, it yields a dense estimator of \mathbf{A}_* , leading to a divergent statistical error. To overcome these issues, we propose a sparse EM algorithm to deal with the high dimensionality and to produce a sparse estimate of the transition matrix. Specifically, we consider a generalized Dantzig selector for the Yule-Walker equation (Candes and Tao, 2007),

$$\begin{aligned} \hat{\mathbf{A}}_k &= \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{p \times p}} \|\mathbf{A}\|_1, \\ \text{such that } &\left\| \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E}_{t,t+1;k} - \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E}_{t,t;k} \mathbf{A}^\top \right\|_{\max} \leq \tau_k, \end{aligned} \quad (2.4)$$

where $\tau_k \geq 0$ is the tolerance parameter that is tuned in a data-driven manner. The optimization problem (2.4) is solved using linear programming in a row-by-row parallel fashion. We next update the variance estimates as,

Algorithm 1 Sparse EM algorithm for model (1.1).

Initialization: $\hat{\Theta}_0 = \{\hat{\mathbf{A}}_0, \hat{\sigma}_{\eta,0}^2, \hat{\sigma}_{\epsilon,0}^2\}$, and the iteration number $k = 1$.

repeat

1. E-step: Obtain $\mathbf{E}_{t;k}$, $\mathbf{E}_{t,t;k}$, and $\mathbf{E}_{t,t+1;k}$ via Kalman filter and smoothing, conditional on $\{\mathbf{y}_t\}_{t \in [T]}$ and $\hat{\Theta}_{k-1}$.

2. M-step:

2.1. Compute $\hat{\mathbf{A}}_k$ by (2.4).

2.2. Compute $\hat{\sigma}_{\eta,k}^2$ and $\hat{\sigma}_{\epsilon,k}^2$ by (2.5).

3. Collect $\hat{\Theta}_k = \{\hat{\mathbf{A}}_k, \hat{\sigma}_{\eta,k}^2, \hat{\sigma}_{\epsilon,k}^2\}$, and set $k = k + 1$.

until the stopping criterion is met.

$$\begin{aligned} \hat{\sigma}_{\eta,k}^2 &= \frac{1}{p(T-1)} \sum_{t=1}^{T-1} \left\{ \text{tr}(\mathbf{E}_{t+1,t+1;k}) - \text{tr}(\hat{\mathbf{A}}_k \mathbf{E}_{t,t+1;k}) \right\}, \\ \hat{\sigma}_{\epsilon,k}^2 &= \frac{1}{pT} \sum_{t=1}^T \left\{ \mathbf{y}_t^\top \mathbf{y}_t - 2\mathbf{y}_t^\top \mathbf{E}_{t;k} + \text{tr}(\mathbf{E}_{t,t;k}) \right\}, \end{aligned} \quad (2.5)$$

where $\mathbf{E}_{t;k} = \mathbb{E}\{\mathbf{x}_t | \{\mathbf{y}_{t'}\}_{t'=1}^T, \hat{\Theta}_{k-1}\}$ for $t \in [T]$, and (2.5) comes from taking derivative on $Q_y(\Theta | \hat{\Theta}_k)$. We terminate the algorithm when the estimates are close enough in two consecutive iterations, e.g., $\min \{\|\hat{\mathbf{A}}_k - \hat{\mathbf{A}}_{k-1}\|_F, |\hat{\sigma}_{\eta,k} - \hat{\sigma}_{\eta,k-1}|, |\hat{\sigma}_{\epsilon,k} - \hat{\sigma}_{\epsilon,k-1}|\} \leq 10^{-3}$.

We summarize our sparse EM procedure in Algorithm 1.

2.2 Estimation consistency

We next establish the estimation precision for our sparse EM estimators that is required for subsequent global and simultaneous testing. The technical assumptions and the proof are relegated to the Supplementary Appendix.

Theorem 1. *Suppose the following conditions hold.*

(a) *The initial parameter set $\hat{\Theta}_0 = \{\hat{\mathbf{A}}_0, \hat{\sigma}_{\eta,0}^2, \hat{\sigma}_{\epsilon,0}^2\}$ lies in a local neighborhood of Θ_* ,*

satisfying Assumptions S1-S4 in the Supplementary Appendix.

- (b) The number of iterations K satisfies $K \geq c \lceil \log(Tp) \rceil$ for some constant $c > 0$.
- (c) The tolerance parameter τ_k in (2.4) satisfies that $\tau_k = c_k \sqrt{\log(p)/T}$ for some positive constant c_k for all $k \leq K$.
- (d) The dimension of time series p and the length of series T satisfy that $C \log p \leq T$ for some positive constant C .

Then, the sparse EM estimator $\hat{\Theta}_K = \{\hat{\mathbf{A}}_K, \hat{\sigma}_{\eta,K}^2, \hat{\sigma}_{\epsilon,K}^2\}$ at the K th iteration satisfies that, for any constant $c_0 > 0$, there exist positive constants c_1 to c_5 , such that the following events occur with probability at least $1 - p^{-c_0}$:

$$\begin{aligned} |\hat{\sigma}_{\epsilon,K}^2 - \sigma_{\epsilon,*}^2| &\leq c_1 \sqrt{\frac{\log p}{Tp}}, & |\hat{\sigma}_{\eta,K}^2 - \sigma_{\eta,*}^2| &\leq c_2 \sqrt{\frac{\log p}{Tp}}, \\ \|\hat{\mathbf{A}}_K - \mathbf{A}_*\|_{\max} &\leq c_3 \sqrt{\frac{\log p}{T}}, & \|\hat{\mathbf{A}}_K - \mathbf{A}_*\|_{l_\infty} &\leq c_4 \left(\frac{\log p}{T}\right)^{1/4}, \\ \|\hat{\mathbf{A}}_K - \mathbf{A}_*\|_{r,2} &\leq c_5 \left(\frac{\log p}{T}\right)^{3/8}. \end{aligned}$$

We make some remarks. First, we require the initialization to be reasonably close to the true parameter as stated in condition (a). See Section S1 of the Supplementary Appendix for more discussion. Second, after a sufficient number of iterations k , we see that the errors of the sparse EM estimators are to be dominated by the statistical error, which decays fast in terms of p and T . Finally, we observe the phenomenon of ‘‘blessing of dimensionality’’, in that the errors of $\sigma_{\epsilon,*}^2$ and $\sigma_{\eta,*}^2$ decrease when the dimension p grows under a fixed sample size T .

3. Test Statistic

We next construct a Gaussian matrix as our test statistic for the transition matrix inference in our high-dimensional VAR with measurement error. From model (1.1), we observe a time series of \mathbf{y}_t that follows an autoregressive structure, $\mathbf{y}_{t+1} = \mathbf{A}_* \mathbf{y}_t + \mathbf{e}_t$, with the error term $\mathbf{e}_t = -\mathbf{A}_* \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_{t+1} + \boldsymbol{\eta}_t$. Then the lag-1 auto-covariance of the error \mathbf{e}_t is of the form,

$$\boldsymbol{\Sigma}_e = \text{Cov}(\mathbf{e}_t, \mathbf{e}_{t-1}) = -\sigma_{\epsilon,*}^2 \mathbf{A}_*.$$

This equality leads to a key observation that, one can apply the covariance testing methods on $\boldsymbol{\Sigma}_e$ to infer the transition matrix \mathbf{A}_* . However, \mathbf{e}_t is not directly observed. Define the generic estimators of Θ_* by $\{\hat{\mathbf{A}}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_\eta^2\}$. We use them to reconstruct this error, and obtain the sample lag-1 auto-covariance estimator,

$$\hat{\boldsymbol{\Sigma}}_e = \frac{1}{T-2} \sum_{t=2}^{T-1} \hat{\mathbf{e}}_t \hat{\mathbf{e}}_{t-1}^\top, \quad \text{where} \quad \hat{\mathbf{e}}_t = \mathbf{y}_{t+1} - \hat{\mathbf{A}} \mathbf{y}_t - \frac{1}{T-1} \sum_{t'=1}^{T-1} (\mathbf{y}_{t'+1} - \hat{\mathbf{A}} \mathbf{y}_{t'}).$$

This sample estimator $\hat{\boldsymbol{\Sigma}}_e$, nevertheless, involves some bias due to the reconstruction of the error term, and also an inflated variance due to the temporal dependence of the time series data. We next explicitly quantify such bias and variance, by characterizing the non-asymptotic behavior of $\hat{\boldsymbol{\Sigma}}_e$, which eventually leads to our Gaussian matrix test statistic.

Denote the maximal row-wise ℓ_1 estimation error as $\Delta_1 = \|\mathbf{A}_* - \hat{\mathbf{A}}\|_{\ell_\infty}$, and the maximal row-wise Euclidean estimation error as $\Delta_2 = \|\mathbf{A}_* - \hat{\mathbf{A}}\|_{r,2}$. The next proposition characterizes the non-asymptotic behavior of $\hat{\boldsymbol{\Sigma}}_e$.

Proposition 1. For any constant $c > 0$, there exist positive constants c_1, c_2 , and c_3 , such that, when $T \geq c_1 \log p$,

$$\mathbb{P} \left\{ \left\| \widehat{\Sigma}_e + (\sigma_{\eta,*}^2 + \sigma_{\epsilon,*}^2) \widehat{\mathbf{A}} - \sigma_{\eta,*}^2 \mathbf{A}_* - \frac{1}{T-2} \sum_{t=2}^{T-1} (\mathbf{e}_t \mathbf{e}_{t-1}^\top - \mathbb{E} \mathbf{e}_t \mathbf{e}_{t-1}^\top) \right\|_{\max} \leq c_2 \left(\Delta_1 s_r \sqrt{\frac{\log p}{T}} + \Delta_2^2 + \frac{\log p}{T} \right) \right\} \geq 1 - c_3 p^{-c},$$

where $s_r = \max_{i \in [p]} |\{j : A_{*,ij} \neq 0\}|$ is the maximal row-wise sparsity of \mathbf{A}_* .

This proposition suggests using $(\sqrt{T-2})\widehat{\Sigma}_e$ to construct the test statistic, since $(T-2)^{-1/2} \sum_{t=2}^{T-1} (\mathbf{e}_t \mathbf{e}_{t-1}^\top - \mathbb{E} \mathbf{e}_t \mathbf{e}_{t-1}^\top)$ converges to a zero-mean Gaussian matrix by the central limit theorem. The max norm error of sparse EM estimator $\widehat{\mathbf{A}}_K$ implies that the non-vanishing bias of $(\sqrt{T-2})\widehat{\Sigma}_e$ is $\sqrt{T-2}\{-(\sigma_{\eta,*}^2 + \sigma_{\epsilon,*}^2)\widehat{\mathbf{A}} + \sigma_{\eta,*}^2 \mathbf{A}_*\}$, which can be estimated by $\sqrt{T-2}\{-(\widehat{\sigma}_\eta^2 + \widehat{\sigma}_\epsilon^2)\widehat{\mathbf{A}} + \widehat{\sigma}_\eta^2 \mathbf{A}_0\}$ under the null hypothesis. Meanwhile, after the bias correction and some direct calculation of the entry-wise variance of $(T-2)^{-1/2} \sum_{t=2}^{T-1} (\mathbf{e}_t \mathbf{e}_{t-1}^\top - \mathbb{E} \mathbf{e}_t \mathbf{e}_{t-1}^\top)$, the entry-wise limit variance of $(\sqrt{T-2})\widehat{\Sigma}_e$ is,

$$\begin{aligned} \sigma_{*,ij}^2 &= (\sigma_{\epsilon,*}^2 + \sigma_{\eta,*}^2)^2 + \sigma_{\epsilon,*}^4 A_{*,ij}^2 + 2\sigma_{\epsilon,*}^4 A_{*,ii} A_{*,jj} + \sigma_{\epsilon,*}^4 \|\mathbf{A}_{*,i:}\|_2^2 \|\mathbf{A}_{*,j:}\|_2^2 \\ &\quad + (\sigma_{\epsilon,*}^4 + \sigma_{\epsilon,*}^2 \sigma_{\eta,*}^2) (\|\mathbf{A}_{*,i:}\|_2^2 + \|\mathbf{A}_{*,j:}\|_2^2), \quad i, j \in [p]. \end{aligned}$$

Plugging the estimators $\{\widehat{\mathbf{A}}, \widehat{\sigma}_\epsilon^2, \widehat{\sigma}_\eta^2\}$ into the above equation, we obtain the corresponding estimator $\widehat{\sigma}_{ij}^2$. We also comment that, one can use any generic estimators $\{\widehat{\mathbf{A}}, \widehat{\sigma}_\epsilon^2, \widehat{\sigma}_\eta^2\}$ to estimate the bias and variance of $(\sqrt{T-2})\widehat{\Sigma}_e$. Later, we present the sufficient conditions on the estimation precision of the generic estimators, so to achieve the desired theoretical properties of inference. We then show that the sparse

EM estimators satisfy those conditions.

We now construct the Gaussian matrix test statistic \mathbf{H} , with the entry,

$$H_{ij} = \frac{\sum_{t=2}^{T-1} \{\hat{e}_{t,i}\hat{e}_{t-1,j} + (\hat{\sigma}_\eta^2 + \hat{\sigma}_\epsilon^2) \hat{A}_{ij} - \hat{\sigma}_\eta^2 A_{0,ij}\}}{\sqrt{T-2} \hat{\sigma}_{ij}}, \quad i, j \in [p]. \quad (3.6)$$

Denote the estimation errors, $\Delta_\epsilon = |\hat{\sigma}_\epsilon^2 - \sigma_{\epsilon,*}^2|$, $\Delta_\eta = |\hat{\sigma}_\eta^2 - \sigma_{\eta,*}^2|$, and $\Delta_\sigma = \max_{i,j \in [p]} |\hat{\sigma}_{ij}^2 - \sigma_{*,ij}^2|$. The next theorem provides the sufficient conditions to guarantee the asymptotic standard normality of H_{ij} 's under the null.

Theorem 2. *Suppose the following conditions hold.*

- (a) *The estimation errors satisfy that $\Delta_1 = o_p\{s_r^{-1}(\log p)^{-1/2}\}$, $\Delta_2 = o_p(T^{-1/4})$, $\Delta_\epsilon = o_p(T^{-1/2})$, $\Delta_\eta = o_p(T^{-1/2})$, and $\Delta_\sigma = o_p(1)$.*
- (b) *The dimension and length of time series satisfy that $\log p = o(T^{1/2})$.*

Then, for all $i, j \in [p]$, as $p, T \rightarrow \infty$, we have that,

$$\frac{\sum_{t=2}^{T-1} \{\hat{e}_{t,i}\hat{e}_{t-1,j} + (\hat{\sigma}_\eta^2 + \hat{\sigma}_\epsilon^2) \hat{A}_{ij} - \hat{\sigma}_\eta^2 A_{*,ij}\}}{\sqrt{T-2} \hat{\sigma}_{ij}} \xrightarrow{d} \mathbf{N}(0, 1).$$

Consequently, \mathbf{H} is to serve as the test statistic for our inference procedures.

4. Transition Matrix Inference

4.1 Global inference

We first develop a testing procedure for the global hypotheses (1.2). The key observation is that the squared maximum entry of a zero mean normal vector converges

to a Gumbel distribution (Cai and Jiang, 2011). Specifically, we construct the global test statistic as,

$$G_{\mathcal{S}} = \max_{(i,j) \in \mathcal{S}} H_{ij}^2.$$

The next theorem states that the asymptotic null distribution of $G_{\mathcal{S}}$ is Gumbel. We again state the sufficient conditions required for the generic estimators $\{\hat{\mathbf{A}}, \hat{\sigma}_{\epsilon}^2, \hat{\sigma}_{\eta}^2\}$ first, and show later that the sparse EM estimators satisfy these conditions.

Theorem 3. *Suppose the following conditions hold.*

- (a) *The estimation errors satisfy $\Delta_1 = o_p\{(s_r \log p)^{-1}\}$, $\Delta_2 = o_p\{(T \log p)^{-1/4}\}$, $\Delta_{\epsilon} = o_p\{(T \log p)^{-1/2}\}$, $\Delta_{\eta} = o_p\{(T \log p)^{-1/2}\}$, and $\Delta_{\sigma} = o_p\{(\log p)^{-1}\}$.*
- (b) *The dimension and length of time series satisfy that $\log p = o(T^{1/7})$.*

Then, under the global null hypothesis (1.2), for any $\mathcal{S} \subseteq [p] \times [p]$, $x \in \mathbb{R}$,

$$\lim_{|\mathcal{S}| \rightarrow \infty} \mathbb{P}\left(G_{\mathcal{S}} - 2 \log |\mathcal{S}| + \log \log |\mathcal{S}| \leq x\right) = \exp\{-\exp(-x/2)/\sqrt{\pi}\}.$$

We note that condition (a) about the estimation consistency in Theorem 3 is stronger than that in Theorem 2 for the asymptotic normality. This is because the Gumbel convergence is built upon the normality property that needs to be guaranteed first. Moreover, the rate $\log p = o(T^{1/7})$ in condition (c) is needed to establish the Gaussian approximation of the test statistic H_{ij} while dealing with the temporal dependence.

Based on Theorem 3, we define the asymptotic α -level test as,

$$\Psi_{\alpha} = \mathbf{1}\left[G_{\mathcal{S}} > 2 \log |\mathcal{S}| - \log \log |\mathcal{S}| - \log \pi - 2 \log\{-\log(1 - \alpha)\}\right].$$

We reject the global null if $\Psi_{\alpha} = 1$.

Next, we study the asymptotic power of the test Ψ_α . Toward that end, we introduce a parameter class of alternatives,

$$\mathcal{A}(c, \mathcal{S}) = \left\{ \{ \mathbf{A}_*, \sigma_{\eta,*}^2, \sigma_{\epsilon,*}^2 \} : \max_{(i,j) \in \mathcal{S}} \frac{\sigma_{\eta,*}^2 \delta_{ij}}{\sigma_{*,ij}} \geq c \sqrt{\frac{\log |\mathcal{S}|}{T}} \right\}, \quad (4.7)$$

where $\delta_{ij} = |A_{*,ij} - A_{0,ij}|$ is the distance between the null and the true transition matrix. The class $\mathcal{A}(c, \mathcal{S})$ requires at least one entry in \mathcal{S} has a proper signal-to-noise ratio against the null. Note that this is a very large class, because the imposed magnitude $\sqrt{\log |\mathcal{S}|/T}$ is vanishing, and it only requires one entry to satisfy. The next theorem shows that Ψ_α has the power converging to one uniformly over $\mathcal{A}(2\sqrt{2}, \mathcal{S})$. Together, Theorems 3 and 4 establish the asymptotic size and power, and thus the consistency of the global test Ψ_α .

Theorem 4. *Suppose the same conditions in Theorem 3 hold. Then,*

$$\inf_{\{ \mathbf{A}_*, \sigma_{\eta,*}^2, \sigma_{\epsilon,*}^2 \} \in \mathcal{A}(2\sqrt{2}, \mathcal{S})} \mathbb{P}(\Psi_\alpha = 1) \rightarrow 1, \quad \text{as } |\mathcal{S}| \rightarrow \infty.$$

Next, we show that adopting the sparse EM estimators developed in Section 2 to construct tests yields the same desired results as in Theorems 3 and 4. Recall the sparse EM estimators at iteration K are denoted as $\{ \hat{\mathbf{A}}_K, \hat{\sigma}_{\eta,K}^2, \hat{\sigma}_{\epsilon,K}^2 \}$. Plugging in these estimators yields the corresponding sparse EM estimator $\hat{\sigma}_{ij,K}^2$ of $\sigma_{*,ij}^2$. Denote the global test statistic and the α -level test based on these estimators as $G_{\mathcal{S},\text{sEM}}$ and $\Psi_{\alpha,\text{sEM}}$, respectively. The next proposition establishes their size and power properties. The conditions for this proposition essentially combine those of Theorems 1 and 3. The new condition on the row-wise sparsity allows s_r to diverge with T .

Proposition 2. *Suppose the following conditions hold.*

(a) *Suppose conditions (a), (b), and (c) of Theorem 1 hold.*

(b) *Suppose $\log p = o(T^{1/7})$, and $s_r^4 \log^5 p = o(T)$.*

Then, under the global null hypothesis (1.2), for any $\mathcal{S} \subseteq [p] \times [p]$, $x \in \mathbb{R}$,

$$\lim_{|\mathcal{S}| \rightarrow \infty} \mathbb{P} \left(G_{\mathcal{S}, \text{sEM}} - 2 \log |\mathcal{S}| + \log \log |\mathcal{S}| \leq x \right) = \exp \left\{ - \exp(-x/2) / \sqrt{\pi} \right\},$$

$$\inf_{\{\mathbf{A}_*, \sigma_{\eta, *}, \sigma_{\epsilon, *}^2\} \in \mathcal{A}(2\sqrt{2}, \mathcal{S})} \mathbb{P}(\Psi_{\alpha, \text{sEM}} = 1) \rightarrow 1, \quad \text{as } |\mathcal{S}| \rightarrow \infty.$$

4.2 Simultaneous inference with FDR control

We next develop a testing procedure for the simultaneous hypotheses (1.3) with a proper FDR control. Let $\mathcal{H}_0 = \{(i, j) : A_{*, ij} = A_{0, ij}, (i, j) \in \mathcal{S}\}$ denote the set of true null hypotheses, and $\mathcal{H}_1 = \{(i, j) : (i, j) \in \mathcal{S}, (i, j) \notin \mathcal{H}_0\}$ denote the set of true alternatives. The test statistic H_{ij} follows a standard normal distribution when $H_{0, ij}$ holds, and as such, we reject $H_{0, ij}$ if $|H_{ij}| > t$ for some thresholding value $t > 0$. Let $R_{\mathcal{S}}(t) = \sum_{(i, j) \in \mathcal{S}} \mathbb{1}\{|H_{ij}| > t\}$ denote the number of rejections at t . Then the false discovery proportion (FDP) and the false discovery rate (FDR) in our simultaneous testing problem are,

$$\text{FDP}_{\mathcal{S}}(t) = \frac{\sum_{(i, j) \in \mathcal{H}_0} \mathbb{1}\{|H_{ij}| > t\}}{R_{\mathcal{S}}(t) \vee 1}, \quad \text{and} \quad \text{FDR}_{\mathcal{S}}(t) = \mathbb{E} \{ \text{FDP}_{\mathcal{S}}(t) \}.$$

An ideal choice of the threshold t is to reject as many true positives as possible, while controlling the false discovery rate at the pre-specified level β . That is, we choose

Algorithm 2 Simultaneous inference with FDR control.

1. Calculate H_{ij} for all $(i, j) \in \mathcal{S}$.
2. Compute the thresholding value,

$$\hat{t} = \inf \left\{ 0 < t \leq \sqrt{2 \log |\mathcal{S}|} : \frac{\{2 - 2\Phi(t)\} |\mathcal{S}|}{R_{\mathcal{S}}(t) \vee 1} \leq \beta \right\}.$$

If \hat{t} does not exist, set $\hat{t} = \sqrt{2 \log |\mathcal{S}|}$.

3. For all $(i, j) \in \mathcal{S}$, reject $H_{0;ij}$ if $|H_{ij}| > \hat{t}$.
-

$\inf\{t > 0 : \text{FDP}_{\mathcal{S}}(t) \leq \beta\}$ as the threshold. However, \mathcal{H}_0 in $\text{FDP}_{\mathcal{S}}(t)$ is unknown. Observing that $\mathbb{P}(|H_{ij}| > t) \approx 2\{1 - \Phi(t)\}$ by Theorem 2, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, we estimate the false rejections $\sum_{(i,j) \in \mathcal{H}_0} \mathbb{1}\{|H_{ij}| > t\}$ in $\text{FDP}_{\mathcal{S}}(t)$ using $\{2 - 2\Phi(t)\} |\mathcal{S}|$. Moreover, we restrict the search of t to the range $(0, \sqrt{2 \log |\mathcal{S}|}]$, since $\mathbb{P}(\hat{t} \text{ exists in } (0, \sqrt{2 \log |\mathcal{S}|}]) \rightarrow 1$ as we show later in the proof of Theorem 5. We summarize our simultaneous testing procedure in Algorithm 2.

Next, we study the asymptotic FDR control of Algorithm 2.

Assumption 1. *Suppose there exist positive constants u_1 and u_2 , such that,*

$$\left| \left\{ (i, j) : (i, j) \in \mathcal{H}_1, \frac{\sigma_{n,*}^2 \delta_{ij}}{\sigma_{*,ij}} > (4 + u_1) \sqrt{\frac{\log p}{T}} \right\} \right| \geq u_2 \sqrt{\log \log |\mathcal{S}|}.$$

In addition, suppose $|\mathcal{H}_0| \geq c_1 |\mathcal{S}|$ for some positive constant c_1 .

This assumption requires the number of alternatives cannot be too small. Otherwise, $\sum_{(i,j) \in \mathcal{H}_0} \mathbb{1}\{|H_{ij}| > t\} \approx R_{\mathcal{S}}(t)$ for any t , and the resulting FDR is close to one regardless the thresholding value. This requirement is rather mild, since the required number is logarithm of logarithm of $|\mathcal{S}|$. Liu and Shao (2014) showed this is nearly

necessary in that the FDR control for large-scale simultaneous testing would fail if the number of true alternatives is fixed. In addition, this assumption also requires the number of nulls cannot be too small, which again is a mild condition.

Assumption 2. For some constants $0 < v < (1 - \bar{\sigma})/(1 + \bar{\sigma})$, $\gamma > 0$, and $u > 0$, suppose $|\{(i_1, j_1), (i_2, j_2)\} : |\tilde{\sigma}_{i_1 j_1, i_2 j_2}| > (\log |\mathcal{S}|)^{-2-\gamma}; (i_1, j_1) \neq (i_2, j_2); (i_1, j_1), (i_2, j_2) \in \mathcal{H}_0\}| \leq u |\mathcal{S}|^{1+v}$, where $\tilde{\sigma}_{i_1 j_1, i_2 j_2}$ is the limit covariance between $H_{i_1 j_1}$ and $H_{i_2 j_2}$, for $(i_1, j_1) \neq (i_2, j_2) \in \mathcal{S}$, and $\bar{\sigma} = \max_{(i_1, j_1) \neq (i_2, j_2); (i_1, j_1), (i_2, j_2) \in \mathcal{H}_0} |\tilde{\sigma}_{i_1 j_1, i_2 j_2}|$.

This assumption bounds the number of strongly correlated entries in the null hypotheses. The bound, $|\mathcal{S}|^{1+v}$, is weak, since there are $|\mathcal{S}|^2$ pairs in total and the majority of them are allowed to be strongly correlated. A similar assumption was adopted in Xia et al. (2018) to ensure the FDR control consistency. The explicit expression of $\tilde{\sigma}_{i_1 j_1, i_2 j_2}$ is given in the proof of Theorem 5.

The next theorem shows that the simultaneous testing procedure in Algorithm 2 controls both FDR and FDP. We again state the sufficient conditions required for any estimators $\{\hat{\mathbf{A}}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_\eta^2\}$ first, then show that the sparse EM estimators satisfy these conditions.

Theorem 5. Suppose the following conditions hold.

- (a) Suppose Assumptions 1 and 2 hold.
- (b) The estimation errors satisfy the precisions in (a) of Theorem 3.
- (c) The dimension and length of time series satisfy $p \leq T^{c_2}$ for some $c_2 > 0$.

Then, for simultaneous hypotheses (1.3), for any $\mathcal{S} \subseteq [p] \times [p]$,

$$\lim_{|\mathcal{S}| \rightarrow \infty} \frac{FDR_{\mathcal{S}}(\hat{t})}{\beta|\mathcal{H}_0|/|\mathcal{S}|} = 1, \quad \text{and} \quad \frac{FDP_{\mathcal{S}}(\hat{t})}{\beta|\mathcal{H}_0|/|\mathcal{S}|} \xrightarrow{p} 1 \quad \text{as } |\mathcal{S}| \rightarrow \infty.$$

Compared to the global testing, the estimation consistency condition (b) remains the same for the simultaneous testing. Meanwhile, the simultaneous testing places some additional requirements on the numbers of nulls and alternatives as in Assumption 1, the entry dependence as in Assumption 2, and the trade-off between p and T as in (c). These requirements are reasonable because, intuitively, the global testing only deals with the maximum entry, whereas the simultaneous testing tackles every individual entry. As such, the simultaneous testing relies more on the dependence structure among the entries, and needs a larger sample size than the global testing. Finally, the slight deflation $\beta|\mathcal{H}_0|/|\mathcal{S}|$ in the limiting FDR comes from substituting $|\mathcal{H}_0|$ with $|\mathcal{S}|$ in the false rejection approximation.

Next, we show that, when employing the sparse EM estimators in Section 2, we can obtain the same properties as in Theorem 5.

Proposition 3. *Suppose the following conditions hold.*

- (a) *Suppose Assumptions 1 and 2 hold.*
- (b) *Suppose conditions (a), (b), and (c) of Theorem 1 hold.*
- (c) *Suppose $|\mathcal{H}_0| \geq c_1|\mathcal{S}|$ for some positive constant c_1 .*
- (d) *Suppose $s_r^4 \log^5 p = o(T)$, and $p \leq T^{c_2}$ for some positive constant c_2 .*

Then, for simultaneous hypotheses (1.3), for any $\mathcal{S} \subseteq [p] \times [p]$,

$$\lim_{|\mathcal{S}| \rightarrow \infty} \frac{FDR_{\mathcal{S}}(\hat{t})}{\beta |\mathcal{H}_0| / |\mathcal{S}|} = 1, \quad \text{and} \quad \frac{FDP_{\mathcal{S}}(\hat{t})}{\beta |\mathcal{H}_0| / |\mathcal{S}|} \xrightarrow{p} 1 \quad \text{as } |\mathcal{S}| \rightarrow \infty.$$

The conditions for this proposition combine those of Theorems 1 and 5. The requirement on the sparse EM algorithm is the same as that for global testing.

5. Simulations

5.1 Setup

We carry out intensive simulations to study the finite-sample performance of our proposed method. We generate the data following model (1.1). We consider four common network structures for the transition matrix \mathbf{A}_* : banded, Erdős-Rényi, stochastic block, and hub, as shown in Figure 1. We first fix $\sigma_{\epsilon,*} = \sigma_{\eta,*} = 0.2$, $\|\mathbf{A}_*\|_2 = 0.97$, and vary the dimension and sample size $(p, T) = (30, 500), (50, 500), (50, 1000), (70, 1000)$, where $p^2 > T$. Next, we fix $p = 50, T = 1000, \sigma_{\epsilon,*} = \sigma_{\eta,*} = 0.2$, and vary the signal strength $\|\mathbf{A}_*\|_2 = 0.7, 0.8, 0.9, 0.97$. Finally, we fix $p = 50, T = 1000, \|\mathbf{A}_*\|_2 = 0.97$, and vary the noise level $(\sigma_{\epsilon,*}, \sigma_{\eta,*}) = (0.2, 0.2), (0.3, 0.3), (0.4, 0.4), (0.5, 0.5)$.

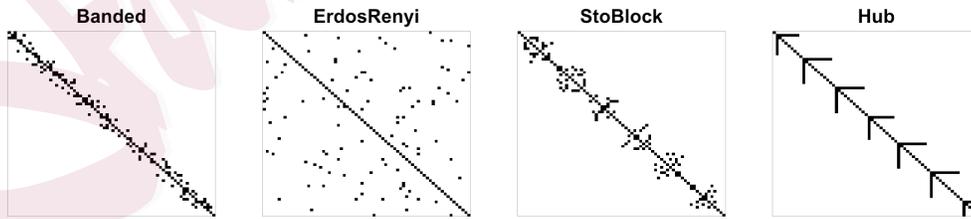


Figure 1: Structures of the transition matrix \mathbf{A}_* . Dots represent the nonzero entries.

5.2 Parameter estimation

We first report the estimation accuracy of our sparse EM. The tuning of the tolerance parameter τ_k in (2.4) is done by minimizing the average prediction error of the testing samples, where we use the first 25% of data points for testing, the last 60% for training, and the middle 15% discarded to reduce the temporal dependence between the training and testing samples. We initialize the transition matrix at $0.1\mathbf{I}_p$, and the error variances at 1×10^{-5} . We have experimented with some other initializations in Section S4.2 of the Appendix, and found that the results are relatively stable. We have also found that our algorithm converges fast, usually within 10 iterations.

We compare our method with three alternative solutions, including the standard EM without sparsity constraint, the Lasso estimator (Hsu et al., 2008), and the Dantzig estimator (Han et al., 2015). The last two methods were designed for VAR without measurement error. We evaluate the estimation accuracy by the Frobenius error $\|\hat{\mathbf{A}} - \mathbf{A}_*\|_F$. Figure 2 reports the average estimation accuracy out of 200 data replications for the varying (p, T) , signal strength $\|\mathbf{A}_*\|_2$, and noise level $(\sigma_{\epsilon,*}, \sigma_{\eta,*})$, respectively. It is seen that our proposed sparse EM achieves the smallest estimation error across all settings. Moreover, our method shows a similar and relatively robust performance across different network structures.

5.3 Global and simultaneous inferences

We next evaluate the performance of our global and simultaneous inference procedures. The alternative matrix is of the form, $\mathbf{A}_* + \mathbf{C}$, where the (i, j) th entry of \mathbf{C} is

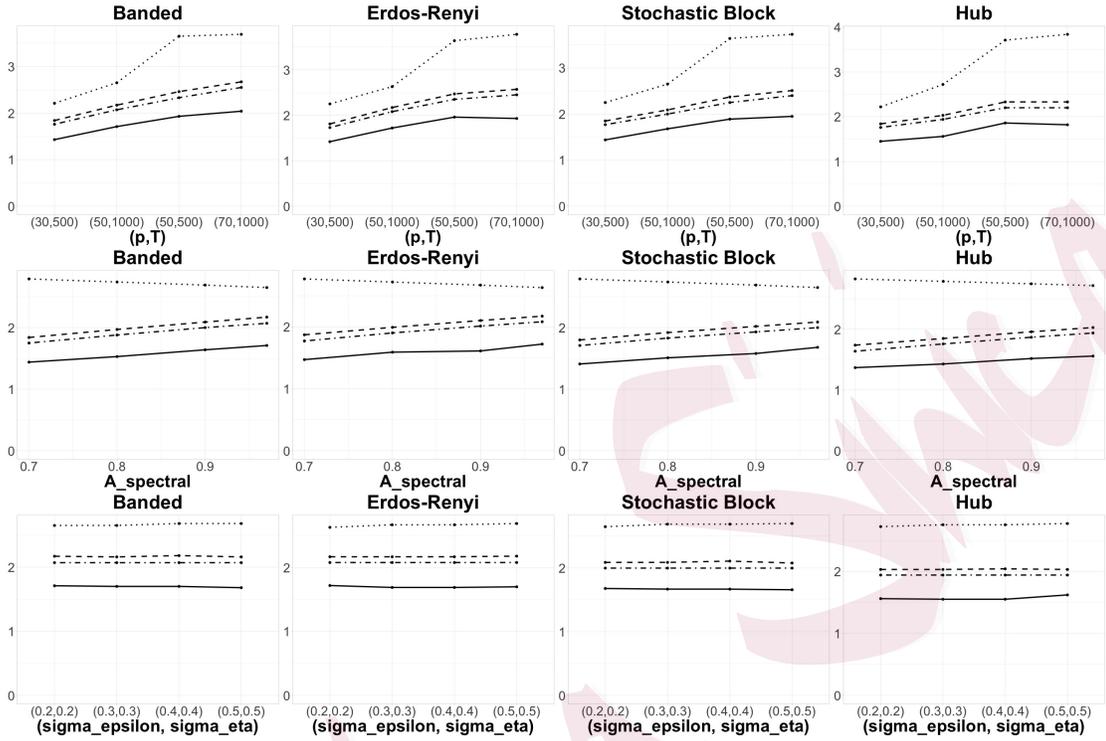


Figure 2: Frobenius estimation error of the transition matrix \mathbf{A}_* for four network structures, and the varying (p, T) (top row), signal strength $\|\mathbf{A}_*\|_2$ (middle row), and noise level $(\sigma_{\epsilon,*}, \sigma_{\eta,*})$ (bottom row). Four methods are compared: the proposed sparse EM (solid line), the standard EM (dotted line), the Lasso estimator (dot-dashed line), and the Dantzig estimator (dashed line).

$20(\sigma_{*,ij}/\sigma_{\eta,*}^2)\sqrt{\log |\mathcal{S}|/T}$, and the rest are zero. The positions of the nonzero entries are randomly sampled in each data replication. Table 1 reports the empirical size and power of global inference, based on 200 data replications, with the significant level set at $\alpha = 5\%$. It is seen that our global test generally maintains a reasonable control of the size, and at the same time achieves a good power. Table 2 reports the average false discovery proportion and the average true positive rate for simultaneous inference, based on 200 data replications, with the FDR level set at 5% . It is seen

that our simultaneous test achieves both a high true positive rate and a low false discovery proportion in all cases. We further report the comparison of our simultaneous inference with some sparse estimation solutions in Section S4.1 of the Supplementary Appendix, and show that our method achieves a competitive performance. We also present an additional simulation example to investigate the empirical size of our test in Section S4.3 of the Supplementary Appendix.

6. Brain Connectivity Analysis

We illustrate the proposed method with a brain connectivity study based on task-evoked fMRI. The data is part of the Human Connectome Project (HCP, Van Essen et al., 2013), where the overarching objective is to understand brain connectivity patterns of healthy adults. We study the fMRI scans of two individual subjects of the same age and sex and both participating in the same story-math task. The task consists of blocks of auditory stories and addition-subtraction calculations, and requires the participant to answer a series of questions. An accuracy score is given at the end based on the participant's answers. The performance of the two subjects differs considerably, with one achieving the perfect score and the other getting only about half correct. We aim to estimate and infer the brain connectivity networks of the two subjects and compare them. We have pre-processed the fMRI data following the pipeline of Glasser et al. (2013). The resulting data for each subject is $p = 264$ time series, corresponding to 264 brain regions-of-interest following the brain atlas of Power et al. (2011). The length of each time series is

	(p, T)	Size	Power	$\ \mathbf{A}_*\ _2$	Size	Power	$(\sigma_{\epsilon,*}, \sigma_{\eta,*})$	Size	Power
banded	(30,500)	2.5 (0.16)	88.5 (0.32)	0.7	7 (0.26)	79.5 (0.40)	(0.2,0.2)	5.5 (0.23)	91.0 (0.29)
	(50,500)	3.5 (0.18)	86.5 (0.34)	0.8	6.5 (0.25)	84.0 (0.37)	(0.3,0.3)	5.5 (0.23)	100 (0)
	(50,1000)	5.5 (0.23)	91.0 (0.29)	0.9	5.5 (0.23)	89.0 (0.31)	(0.4,0.4)	5 (0.22)	100 (0)
	(70,1000)	1.5 (0.12)	92.5 (0.26)	0.97	5.5 (0.23)	91.0 (0.29)	(0.5,0.5)	5.5 (0.23)	100 (0)
Erdős-Rényi	(30,500)	3.5 (0.18)	89.0 (0.31)	0.7	1.5 (0.12)	80.5 (0.40)	(0.2,0.2)	1.5 (0.12)	92.5 (0.26)
	(50,500)	4 (0.2)	93.5 (0.25)	0.8	1.5 (0.12)	87.0 (0.34)	(0.3,0.3)	1 (0.1)	100 (0)
	(50,1000)	1.5 (0.12)	92.5 (0.26)	0.9	1 (0.1)	91.0 (0.29)	(0.4,0.4)	1 (0.1)	100 (0)
	(70,1000)	3.5 (0.18)	91.0 (0.29)	0.97	1.5 (0.12)	92.5 (0.26)	(0.5,0.5)	1.5 (0.12)	100 (0)
stochastic block	(30,500)	3 (0.17)	90.0 (0.30)	0.7	3 (0.17)	81.5 (0.39)	(0.2,0.2)	3.5 (0.18)	92.0 (0.27)
	(50,500)	4 (0.2)	86.0 (0.35)	0.8	3.5 (0.18)	86.0 (0.35)	(0.3,0.3)	3.5 (0.18)	100 (0)
	(50,1000)	3.5 (0.18)	92.0 (0.27)	0.9	3.5 (0.18)	89.0 (0.31)	(0.4,0.4)	3 (0.17)	100 (0)
	(70,1000)	3 (0.17)	91.0 (0.29)	0.97	3.5 (0.18)	92.0 (0.27)	(0.5,0.5)	3.5 (0.18)	100 (0)
hub	(30,500)	3 (0.17)	86.5 (0.34)	0.7	5.5 (0.23)	76.5 (0.43)	(0.2,0.2)	4 (0.2)	87.0 (0.34)
	(50,500)	2.5 (0.16)	88.5 (0.32)	0.8	5 (0.22)	80.5 (0.40)	(0.3,0.3)	4 (0.2)	100 (0)
	(50,1000)	4 (0.2)	87.0 (0.34)	0.9	5 (0.22)	84.5 (0.36)	(0.4,0.4)	4 (0.2)	100 (0)
	(70,1000)	3.5 (0.18)	87.0 (0.34)	0.97	4 (0.2)	87.0 (0.34)	(0.5,0.5)	5.5 (0.23)	100 (0)

Table 1: Empirical size and power, in percentage, of the global test for four network structures, varying (p, T) (left column), signal strength $\|\mathbf{A}_*\|_2$ (middle column), and noise level $(\sigma_{\epsilon,*}, \sigma_{\eta,*})$ (right column). The standard errors are in the parentheses.

$T = 316$. The 264 brain regions have been further grouped into 14 functional modules (Smith et al., 2009): auditory (AD), cerebellar (CR), cingulo-opercular task control (CO), default mode network (DMN), dorsal attention (DAT), fronto-parietal task control (FP), memory retrieval (MR), salience (SA), sensory/somatomotor hand

	(p, T)	FDR	TPR	$\ \mathbf{A}_*\ _2$	FDR	TPR	$(\sigma_{\epsilon,*}, \sigma_{\eta,*})$	FDR	TPR
banded	(30,500)	4.44 (0.03)	73.72 (0.05)	0.7	4.96 (0.03)	71.85 (0.05)	(0.2,0.2)	4.22 (0.02)	92.27 (0.02)
	(50,500)	3.83 (0.02)	67.48 (0.05)	0.8	4.78 (0.02)	82.58 (0.04)	(0.3,0.3)	4.21 (0.02)	92.24 (0.02)
	(50,1000)	4.22 (0.02)	92.27 (0.02)	0.9	4.44 (0.02)	88.91 (0.03)	(0.4,0.4)	4.19 (0.02)	92.25 (0.02)
	(70,1000)	3.78 (0.01)	88.54 (0.02)	0.97	4.22 (0.02)	92.27 (0.02)	(0.5,0.5)	4.2 (0.02)	92.19 (0.02)
Erdős-Rényi	(30,500)	4.05 (0.03)	75.66 (0.07)	0.7	4.69 (0.02)	70.94 (0.05)	(0.2,0.2)	3.93 (0.02)	97.21 (0.02)
	(50,500)	3.9 (0.02)	65.83 (0.05)	0.8	4.51 (0.02)	87.21 (0.03)	(0.3,0.3)	3.94 (0.02)	97.19 (0.02)
	(50,1000)	3.93 (0.02)	97.21 (0.02)	0.9	4.14 (0.02)	94.5 (0.02)	(0.4,0.4)	3.92 (0.02)	97.16 (0.02)
	(70,1000)	4.18 (0.01)	91.42 (0.02)	0.97	3.93 (0.02)	97.21 (0.02)	(0.5,0.5)	3.99 (0.02)	97.19 (0.02)
stochastic block	(30,500)	4.16 (0.03)	74.2 (0.05)	0.7	4.88 (0.03)	66.87 (0.04)	(0.2,0.2)	4.3 (0.02)	89.56 (0.03)
	(50,500)	3.68 (0.02)	61.16 (0.05)	0.8	4.81 (0.02)	79.01 (0.04)	(0.3,0.3)	4.27 (0.02)	89.53 (0.02)
	(50,1000)	4.3 (0.02)	89.56 (0.03)	0.9	4.54 (0.02)	86.17 (0.03)	(0.4,0.4)	4.27 (0.02)	89.54 (0.02)
	(70,1000)	3.94 (0.02)	84.6 (0.02)	0.97	4.3 (0.02)	89.56 (0.03)	(0.5,0.5)	4.26 (0.02)	89.47 (0.02)
hub	(30,500)	4.43 (0.03)	80.93 (0.05)	0.7	4.73 (0.03)	64.95 (0.06)	(0.2,0.2)	4.34 (0.02)	95.37 (0.02)
	(50,500)	3.6 (0.02)	59.05 (0.06)	0.8	4.68 (0.02)	81.64 (0.04)	(0.3,0.3)	4.36 (0.02)	95.38 (0.02)
	(50,1000)	4.34 (0.02)	95.37 (0.02)	0.9	4.53 (0.02)	91.66 (0.03)	(0.4,0.4)	4.32 (0.02)	95.34 (0.02)
	(70,1000)	4.48 (0.02)	77.16 (0.03)	0.97	4.34 (0.02)	95.37 (0.02)	(0.5,0.5)	3.56 (0.02)	95.46 (0.02)

Table 2: Average false discovery proportion (FDP) and true positive rate (TPR), in percentage, of the simultaneous test for four network structures, varying (p, T) (left column), signal strength $\|\mathbf{A}_*\|_2$ (middle column), and noise level $(\sigma_{\epsilon,*}, \sigma_{\eta,*})$ (right column). The standard errors are in the parentheses.

(SMH), sensory/somatomotor mouth (SMM), subcortical (SC), uncertain (UN), ventral attention (VA), and visual (VS). Each module possesses a relatively autonomous functionality, and complex brain tasks are believed to perform through coordinated

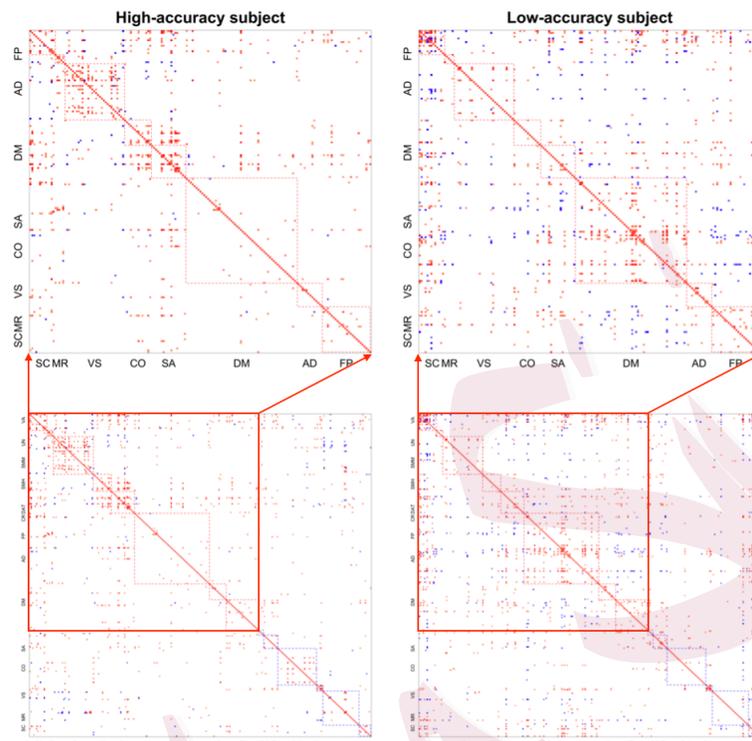


Figure 3: Heatmaps of the identified brain connectivity patterns for the high-accuracy subject (left column) and low-accuracy subject (right column). The 14 functional modules are indicated by the blocks (bottom row), and the 8 modules that demonstrate the most within-module connections are highlighted and amplified (top row). The colored entries indicate the selected connections, and the color scale ranges from blue (negative statistics) to red (positive statistics).

collaborations among the modules.

We apply the proposed tests, and verify that, for this data, the key model assumptions hold reasonably well. See Section S4.4 of the Supplementary Appendix for more details. We begin with the global test for each subject separately. The p -values for the global test for both subjects are smaller than 10^{-15} , indicating that at least one pair of brain regions have statistically significant connectivity. We then

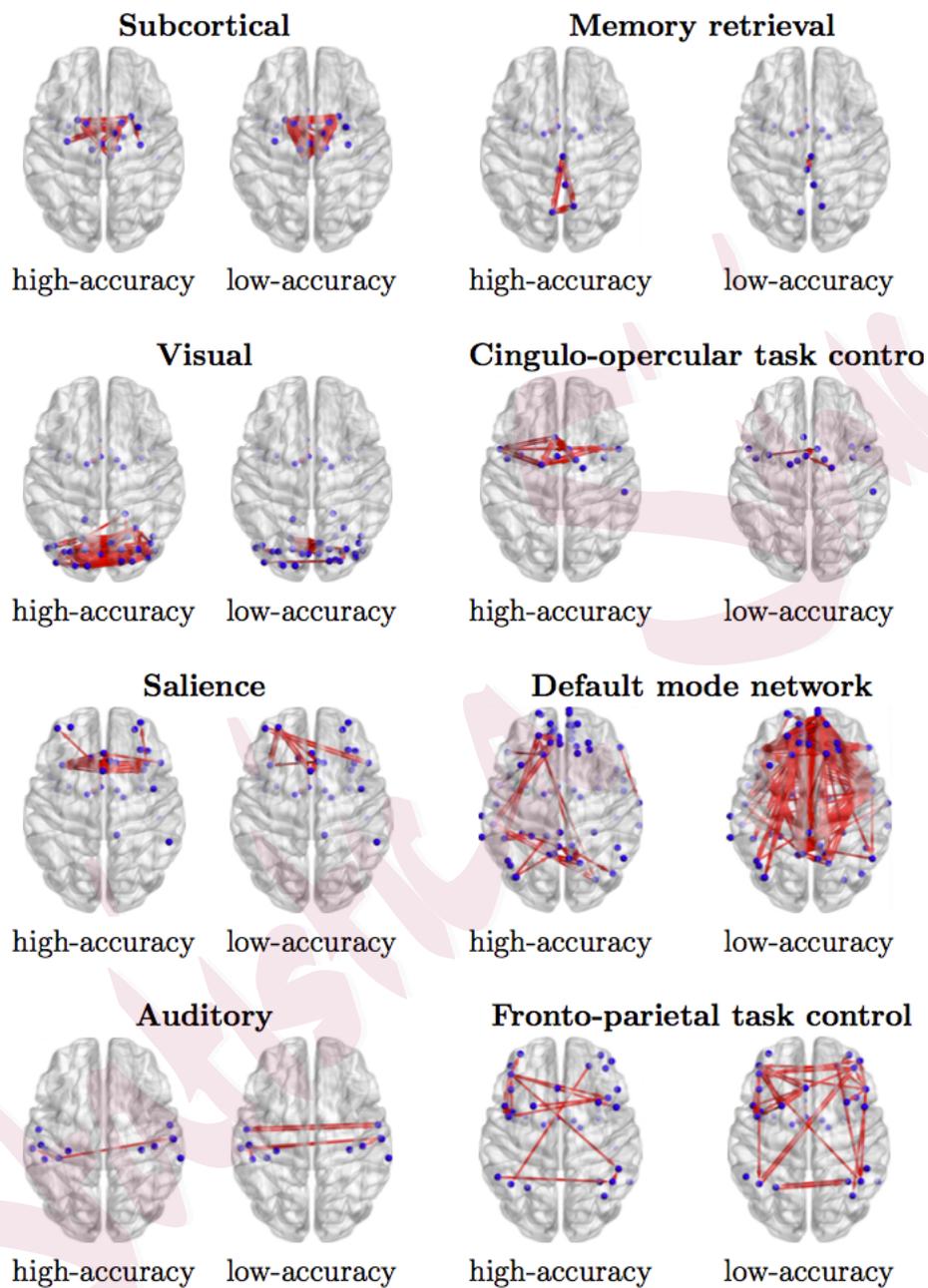


Figure 4: Visualization of the identified brain regions and their within-module connections of the high-accuracy and low-accuracy subjects for the eight functional modules.

apply the simultaneous test, with FDR set at 0.001. First of all, we have identified more within-module connections than the between-module connections (294 out of 7700 or 3.8% versus 961 out of 61936 or 1.6% for the high-accuracy subject, and 376 out of 7700 or 4.9% versus 1350 out of 61936 or 2.2% for the low-accuracy subject). The partition of the brain regions to the functional modules has been fully based on biological knowledge, and our finding lends some numerical support to this partition. Second, the majority of within-module connections are concentrated on eight functional modules. Moreover, when comparing between the two subjects among those modules, we find that the high-accuracy subject has more within-module connections than the low-accuracy subject for the following functional modules: visual (118 versus 27 out of 961), salience (29 versus 11 out of 324), cingulo-opercular task control (17 versus 3 out of 196), and memory retrieval (6 versus 2 out of 25) modules. Such findings suggest that the high-accuracy subject has exhibited more intensive neural activities for processing visual imagery, memory retrieval, tonic alertness and executive control when performing the story-math task, which agrees with the literature (Sadaghiani and D'Esposito, 2015; Luo et al., 2014). On the other hand, we find that the high-accuracy subject has fewer connections than the low-accuracy subject for the following functional modules: default mode network (25 versus 200 out of 3364), fronto-parietal task control (15 versus 37 out of 625), auditory (2 versus 8 out of 169), and subcortical (19 versus 49 out of 169) modules. These findings again agree with the literature, in that these modules have been found strongly associated with the language and reasoning type tasks (Schultz and Cole, 2016), and the high-accuracy

subject has exhibited less brain activity interplay related to auditory processing and mind wandering (Van Praag et al., 2017). Figure 3 shows the identified connectivity patterns for the two subjects, and Figure 4 shows the corresponding brain regions visualized using BrainNet Viewer (Xia et al., 2013).

7. Discussion

In this article, we study both global and simultaneous inferences of the transition matrix under the high-dimensional vector autoregression model with measurement error. There is no existing solution for this type of problem, and our proposal makes a useful contribution for scientific applications such as brain connectivity analysis. The technical tools we develop are also of independent interest, and can facilitate general inferences for other models involving latent variables or correlated observations.

In Theorem 2, we establish the marginal asymptotic normal distribution for each element of the transition matrix estimator. Meanwhile, our model and regularity conditions ensure that the correlations among the entries of transition matrix are not overly strong, so that we can carry out both global and simultaneous inferences for the entire transition matrix based on the marginal characterization in Theorem 2. More specifically, for the global test, the spectral condition that $\|\mathbf{A}_*\|_2 < 1$ ensures the correlations among the entries of transition matrix satisfy the eigenvalue condition of the Gumbel convergence for the maxima of joint distribution (Cai et al., 2013). For the simultaneous test, in addition to the spectral condition, Assumption 1 regulates the size of the null and alternative hypotheses, whereas Assumption 2

bounds the number of strongly correlated entries in the nulls. Together, these conditions guarantee a sufficient number of weakly dependent entries to ensure an accurate FDR cutoff estimation. We note that, Krampe et al. (2018) proposed a bootstrap method for global inference of a sparse VAR model without measurement error. They imposed a similar condition as $\|\mathbf{A}_*\|_2 < 1$, along with some other stronger conditions, to control the correlations among the entries of the transition matrix.

Our numerical experiments have found that our inference procedures work well across a range of network structures. Such robustness can be explained by noting that, the inferential guarantees mainly rely on the row-wise sparsity, the estimation accuracy of the transition matrix and error variances, along with the spectral condition of \mathbf{A}_* and the regularity conditions on the null and alternative hypotheses as in Assumptions 1 and 2. Recall that the test statistic H_{ij} is constructed from the sample auto-covariance $\hat{\Sigma}_e$ of the residual $\hat{\mathbf{e}}_t$. Its deviation from the auto-covariance Σ_e of the true error \mathbf{e}_t is controlled by the row-wise sparsity s_r and the sparse EM estimation errors, as specified in Proposition 1. As long as the relations among the row-wise sparsity s_r , the dimension p , and the number of time points T are properly specified, such as in condition (b) of Proposition 2, or condition (c) of Proposition 3, and the sparse EM estimators of $\{\mathbf{A}, \sigma_\epsilon^2, \sigma_\eta^2\}$ are reasonably accurate, we can obtain the asymptotic normality of the test statistic, then the subsequent asymptotic guarantees of the tests, regardless of any particular network structure for \mathbf{A}_* .

We have primarily focused on a lag-1 autoregressive structure in this article. Meanwhile, our proposal can be extended in a relatively straightforward fashion to a

more general lag structure. Specifically, suppose the number of lags is d . Then the latent process in model (1.1) becomes $\mathbf{x}_t = \sum_{l=1}^d \mathbf{A}_{l,*} \mathbf{x}_{t-l} + \boldsymbol{\eta}_{t-1}$, and the problem of interest becomes testing $\mathbf{A}_{1,*}, \dots, \mathbf{A}_{d,*}$. This lag- d VAR model can be equivalently rewritten as a lag-1 model, such that $\tilde{\mathbf{x}}_t = \tilde{\mathbf{A}}_* \tilde{\mathbf{x}}_{t-1} + \tilde{\boldsymbol{\eta}}_{t-1}$, $\tilde{\mathbf{x}}_t = (\mathbf{x}_t^\top, \dots, \mathbf{x}_{t-d+1}^\top)^\top \in \mathbb{R}^{pd}$, $\tilde{\boldsymbol{\eta}}_t = (\boldsymbol{\eta}_t^\top, \mathbf{0}_p^\top, \dots, \mathbf{0}_p^\top)^\top \in \mathbb{R}^{pd}$, and

$$\tilde{\mathbf{A}}_* = \begin{pmatrix} \mathbf{A}_{1,*} & \mathbf{A}_{2,*} & \dots & \mathbf{A}_{d,*} \\ \mathbf{I}_p & \mathbf{0}_{p \times p} & \dots & \mathbf{0}_{p \times p} \\ \mathbf{0}_{p \times p} & \mathbf{I}_p & \dots & \mathbf{0}_{p \times p} \\ \mathbf{0}_{p \times p} & \mathbf{0}_{p \times p} & \dots & \mathbf{I}_p \end{pmatrix}_{pd \times pd}.$$

We can then apply our test to the first block row of $\tilde{\mathbf{A}}_*$, which in turn tests $\mathbf{A}_{1,*}, \dots, \mathbf{A}_{d,*}$.

We have assumed a homoscedastic and independent error structure for both error terms $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$. This is essentially a tradeoff. Under such an error structure, the individual variables in \mathbf{x}_t are still non-identically distributed and highly correlated given the autoregressive structure of the model. In applications such as brain connectivity analysis, it is often reasonable to keep a simplified error structure (Zhang et al., 2015), and our real data analysis yields reasonable findings. In the VAR literature, more general error structures have been considered. However, when estimating the transition matrix, none of the existing methods directly estimated this error structure. By contrast, our inference hinges on a good estimate of the error terms. A more general form of the error structure would introduce more unknown parameters, and requires a considerable amount of extra work to characterize the estimation precision.

We thus keep a simple error structure in this first work on statistical inference, and leave the more general form of the error terms as future research.

In brain connectivity analysis, the early experiments usually focus on a single subject (Friston, 2011). More recently, data involving multiple subjects are emerging. It is of interest to extend our modeling framework to multiple subjects. The key is to capture the subject-to-subject variability by incorporating the subject-specific covariates, meanwhile integrating common information shared across different subjects. A full pursuit of this topic is beyond the scope of this article, and we leave it as future research.

Acknowledgements

Kang's research was partially supported by NSF grant IIS-2123777, and NIH grants R01DA048993, R01MH105561 and R01GM124061. Li's research was partially supported by NSF grant CIF-2102227, and NIH grants R01AG061303, R01AG062542 and R01AG034570.

References

- Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.
- Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical anal-

ysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–198.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525.

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.

Cao, X., Sandstede, B., and Luo, X. (2019). A functional data method for causal dynamic network modeling of task-related fmri. *Frontiers in Neuroscience*, 13:127.

Chen, G., Glen, D., Saad, Z., Hamilton, J. P., Thomason, M., Gotlib, I., and Cox, R. (2011). Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis. *Computers in biology and medicine*, 41:1142–55.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional

covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36.

Garg, R., Cecchi, G., and Rao, R. (2011). Full-brain auto-regressive modeling (farm) using fmri. *NeuroImage*, 58:416–41.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. Technical report, University of Toronto.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Krampe, J., Kreiss, J., and Paparoditis, E. (2018). Bootstrap based inference for

sparse high-dimensional time series models. *arXiv preprint arXiv:1806.11083*, 1(1):1–36.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978.

Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t -tests with false discovery rate control. *The Annals of Statistics*, 42(5):2003–2025.

Luo, Y., Qin, S., Fernandez, G., Zhang, Y., Klumpers, F., and Li, H. (2014). Emotion perception and executive control interact in the salience network during emotionally charged working memory processing. *Human Brain Mapping*, 35(11):5606–5616.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

-
- Sadaghiani, S. and D’Esposito, M. (2015). Functional characterization of the cingulo-opercular network in the maintenance of tonic alertness. *Cerebral Cortex*, 25(9):2763–2773.
- Schultz, D. H. and Cole, M. W. (2016). Higher intelligence is associated with less task-related brain network reconfiguration. *Journal of neuroscience*, 36(33):8551–8561.
- Smith, S. D., Fox, P. T., Miller, K., Glahn, D., Fox, P., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A., and Beckmann, C. F. (2009). Correspondence of the brain; functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, 106:13040–5.
- Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 1(1):1–28.
- Staudenmayer, J. and Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association*, 100(471):841–852.
- Tsay, R. S. and Chen, R. (2018). *Nonlinear time series analysis*, volume 891. John Wiley & Sons.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

- Van Praag, C. D. G., Garfinkel, S. N., Sparasci, O., Mees, A., Philippides, A. O., Ware, M., Ottaviani, C., and Critchley, H. D. (2017). Mind-wandering and alterations to default mode network connectivity when listening to naturalistic versus artificial sounds. *Scientific Reports*, 7:45273.
- Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Advances in neural information processing systems*, pages 2521–2529.
- Xia, M., Wang, J., and He, Y. (2013). Brainnet viewer: A network visualization tool for human brain connectomics. *PLOS ONE*, 8(7):1–15.
- Xia, Y., Cai, T., and Cai, T. T. (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *Journal of the American Statistical Association*, 113(521):328–339.
- Xiao, H. and Wu, W. B. (2013). Asymptotic theory for maximum deviations of sample covariance matrix estimates. *Stochastic Processes and their Applications*, 123(7):2899–2920.
- Yi, X. and Caramanis, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575.
- Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015). A dynamic di-

rectional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110(509):93–106.

Zheng, L. and Raskutti, G. (2019). Testing for high-dimensional network parameters in auto-regressive models. *Electronic Journal of Statistics*, 13(2):4977–5043.

Statistica Sinica