Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

An Interpretable and Efficient Infinite-Order Vector Autoregressive Model for High-Dimensional Time Series

Yao Zheng

To cite this article: Yao Zheng (11 Mar 2024): An Interpretable and Efficient Infinite-Order Vector Autoregressive Model for High-Dimensional Time Series, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2311365

To link to this article: https://doi.org/10.1080/01621459.2024.2311365

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2024.2311365

An Interpretable and Efficient Infinite-Order Vector Autoregressive Model for High-Dimensional Time Series

Yao Zheng

Department of Statistics, University of Connecticut, Storrs, CT

ABSTRACT

As a special infinite-order vector autoregressive (VAR) model, the vector autoregressive moving average (VARMA) model can capture much richer temporal patterns than the widely used finite-order VAR model. However, its practicality has long been hindered by its non-identifiability, computational intractability, and difficulty of interpretation, especially for high-dimensional time series. This article proposes a novel sparse infinite-order VAR model for high-dimensional time series, which avoids all above drawbacks while inheriting essential temporal patterns of the VARMA model. As another attractive feature, the temporal and cross-sectional structures of the VARMA-type dynamics captured by this model can be interpreted separately, since they are characterized by different sets of parameters. This separation naturally motivates the sparsity assumption on the parameters determining the cross-sectional dependence. As a result, greater statistical efficiency and interpretability can be achieved with little loss of temporal information. We introduce two ℓ_1-regularized estimation methods for the proposed model, which can be efficiently implemented via block coordinate descent algorithms, and derive the corresponding nonasymptotic error bounds. A consistent model order selection method based on the Bayesian information criteria is also developed. The merit of the proposed approach is supported by simulation studies and a real-world macroeconomic data analysis. Supplementary materials for this article are available online including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received September 2022
Accepted January 2024

KEYWORDS
Granger causality; High-dimensional time series; Infinite-order vector autoregression; Sparse estimation; VARMA

1. Introduction

Let $\boldsymbol{y}_t \in \mathbb{R}^N$ be the observation of an N-dimensional time series at time t. The need for modeling $\boldsymbol{y}_t$ with a large dimension N is ubiquitous, ranging from economics and finance (Nicholson et al. 2020; Wilms et al. 2023) to biology and neuroscience (Lozano et al. 2009; Gorrostieta et al. 2012), and to environmental and health sciences (Dowell and Pinson 2016; Davis, Zang, and Zheng 2016). For modeling $\boldsymbol{y}_t$, three issues are of particular importance:

(I1) Flexibility of temporal dynamics: As N increases, it is more likely that $\boldsymbol{y}_t$ contains component series with complex temporal dependence structures. Then information further in the past may be needed to generate more flexible temporal dynamics.
(I2) Efficiency: It is important that the estimation is efficient both statistically and computationally under large N, so that accurate forecasts can be obtained.
(I3) Interpretability: Ideally, the model should have easy interpretations, such as direct implications of Granger causality (Granger 1969) among the N component series.

The finite-order vector autoregressive (VAR) model, coupled with dimension reduction techniques such as sparse (Basu and Matteson 2021) and low-rank (Wang et al. 2022) methods, has been widely studied for high-dimensional time series. This model is highly popular due to its theoretical and computational tractability, and the coefficient matrices have intuitive interpretations analogous to those in the multivariate linear regression. However, in practice, a large lag order is often required for the VAR model to adequately fit the data (Chan, Eisenstat, and Koop 2016; Nicholson et al. 2020). Thus, it is more realistic to assume that the data follow the more general, infinite-order VAR (VAR(∞)) process:

$$\boldsymbol{y}_t = \sum_{h=1}^{\infty} \boldsymbol{A}_h \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

where $\boldsymbol{\varepsilon}_t$ are the innovations, and $\boldsymbol{A}_h \in \mathbb{R}^{N\times N}$ are the AR coefficient matrices; in particular, it reduces to the VAR(P) model when $\boldsymbol{A}_h = \boldsymbol{0}$ for $h > P$. In fact, if a sample $\{\boldsymbol{y}_t\}_{t=1}^T$ is generated from (1.1), we can approximate it by a VAR(P) model provided that $P \to \infty$ at an appropriate rate as the sample size $T \to \infty$ (Lütkepohl 2005), which in turn explains the practical need for a large P. Nonetheless, for $\boldsymbol{y}_t$ in (1.1) to be stationary, $\boldsymbol{A}_h$ must diminish quickly as $h \to \infty$; otherwise, the infinite sum will be ill-defined. The decay property of $\boldsymbol{A}_h$, coupled with a large P, will not only pose difficulties in high-dimensional estimation, but make the fitted VAR(P) model hard to interpret. Take the Lasso estimator of the VAR(P) model with sparse $\boldsymbol{A}_h$'s. Since all entries of $\boldsymbol{A}_h$ must be small at even moderately large h, the Lasso may fail to capture the significant

CONTACT Yao Zheng yao.zheng@uconn.edu Department of Statistics, University of Connecticut, Storrs, CT.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

yet small entries. Moreover, the sparsity pattern of $\boldsymbol{A}_h$ for the fitted model generally varies substantially across h, making it even more difficult to interpret $\boldsymbol{A}_h$'s simultaneously (Shojaie, Basu, and Michailidis 2012; Nicholson et al. 2020).

In the literature on multivariate time series, an alternative approach to infinite-order VAR modeling is to consider the vector autoregressive moving average (VARMA) model. For example, the VARMA(1, 1) model is

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1}, \tag{1.2}$$

where $\boldsymbol{\Phi}, \boldsymbol{\Theta} \in \mathbb{R}^{N\times N}$ are the AR and MA coefficient matrices. Assuming that (1.2) is invertible, that is, all eigenvalues of $\boldsymbol{\Theta}$ are less than one in absolute value, (1.2) can be written as the VAR(∞) process in (1.1) with $\boldsymbol{A}_h = \boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \boldsymbol{\Theta}^{h-1}(\boldsymbol{\Phi}-\boldsymbol{\Theta})$ for $h \geq 1$. Note that $\boldsymbol{A}_h$ diminishes quickly as $h \to \infty$ due to the exponential factor $\boldsymbol{\Theta}^{h-1}$, so the VAR($\infty$) process is well defined. Hence, the MA part of the model is the key to parsimoniously generating VAR(∞)-type temporal dynamics. For the general VARMA(p, q) model, $\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\varepsilon}_{t-j}$, the richness of temporal patterns will increase with p and q, but with only small orders p and q, the VARMA model can usually provide more accurate forecasts than large-order VAR models in practice (Athanasopoulos and Vahid 2008; Chan, Eisenstat, and Koop 2016). Compared with finite-order VAR models, the VARMA model is more favorable in terms of (I1) but suffers from severe drawbacks regarding (I2), as its computation is generally complicated due to the following two problems:

(P1) Non-identifiability: For example, in the VARMA(1, 1) case, there are multiple pairs of $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ corresponding to the same process. The root cause of this problem is the matrix multiplications in the parametric form of $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \boldsymbol{\Theta}^{h-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$.

(P2) High-order matrix polynomials: Consider as an example the ordinary least squares (OLS) estimation of the VARMA(1, 1) model. For a sample $\{\boldsymbol{y}_t\}_{t=1}^{T}$, since $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta})$ is an hth-order matrix polynomial for $1 \leq h \leq T$, the loss function will have a computational complexity of $O(T^2N^3)$[1], hence, unscalable under large N.

While recent attempts have been made to improve the feasibility of VARMA models (Metaxoglou and Smith 2007; Chan, Eisenstat, and Koop 2016; Dias and Kapetanios 2018; Wilms et al. 2023), they do not tackle (P1) and (P2) directly, but rather resort to sophisticated identification constraints and optimization methods. Moreover, high-dimensional VARMA models can be difficult to interpret due to their latent MA structures. Particularly, while it may be natural to assume that $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ in (1.2) are sparse under large N (Wilms et al. 2023), this does not necessarily result in a sparse VAR(∞) model; that is, $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta})$'s may not be sparse. Thus, the sparse VARMA model is not particularly attractive in terms of (I3).

[1]The computational complexity in this article is calculated in a model of computation where field operations (addition and multiplication) take constant time.

For high-dimensional time series, we aim to develop a sparse VAR(∞) model that is favorable in all of (I1)–(I3). The proposed approach is motivated by reparameterizing the VAR(∞) form of the VARMA(p, q) model into formulation (1.1) with

$$\boldsymbol{A}_h = \sum_{k=1}^{d} \ell_{h,k}(\boldsymbol{\omega})\boldsymbol{G}_k \quad \text{for} \quad h \geq 1, \tag{1.3}$$

where $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_d \in \mathbb{R}^{N\times N}$ are unknown coefficient matrices, $\{\ell_{h,k}(\cdot)\}_{h=1}^{\infty}$ for $1 \leq k \leq d$ are different sequences of real-valued functions characterizing the exponential decay pattern of $\boldsymbol{A}_h$, with $\ell_{h,k}(\boldsymbol{\omega}) \to 0$ as $h \to \infty$ for each k, and $\boldsymbol{\omega}$ is an unknown low-dimensional parameter vector; see also Huang, Lu, and Zheng (2023) for a high-dimensional Tucker-low-rank time series model concurrently developed from (1.3) with different techniques and interpretations. Similar to the orders (p, q) of the VARMA model, d can be viewed as the overall order that controls the complexity of temporal patterns of the VAR(∞) model; see Section 2 for the detailed model formulation. Note that (1.3) preserves the essential temporal patterns of the VARMA process, since it is derived directly from the former with little loss of generality. Thus, it is fundamentally more flexible than finite-order VAR models, that is, more desirable regarding (I1). Moreover, each $\boldsymbol{A}_h = \boldsymbol{A}_h(\boldsymbol{\omega}, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_d)$ in (1.3) is a linear combination of matrices. Hence, unlike $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta})$ mentioned above, this form of $\boldsymbol{A}_h$ gets rid of all matrix multiplications. As a result, both problems (P1) and (P2) are eliminated, and then (I2) can be achieved. To tackle the high dimensionality, we assume that $\boldsymbol{G}_k$'s are sparse, leading to the proposed sparse parametric VAR(∞) (SPVAR(∞)) model. In addition to improving the estimation efficiency as required by (I2), the sparsity assumption enables greater interpretability, that is, (I3), thanks to the novel separation of temporal and cross-sectional dependence in parameterizing the VARMA-type dynamic structure:

(D1) Temporal dependence: In (1.3), the decay pattern of $\boldsymbol{A}_h$ as $h \to \infty$ is fully characterized by the scalar weights $\ell_{h,k}(\boldsymbol{\omega})$'s.

(D2) Cross-sectional dependence: The $\boldsymbol{G}_k$'s, independent of the above decay pattern as $h \to \infty$, fully capture the cross-sectional dependence.

As a result of (D2), the Granger causal network of the N component series of $\boldsymbol{y}_t$ is directly linked to the aggregate sparsity pattern of $\boldsymbol{G}_k$'s. Moreover, as detailed in Section 2.1, $\{\ell_{h,k}(\boldsymbol{\omega})\}_{h=1}^{\infty}$'s in (1.3) are specifically defined such that $\boldsymbol{A}_k = \boldsymbol{G}_k$ for $1 \leq k \leq p$, whereas $\boldsymbol{A}_{p+j}$ for $j \geq 1$ are expressed as linear combinations of $\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d$, where p is the AR order of the VARMA(p, q) model from which (1.3) originates. Consequently, there is an interesting dichotomy in the interpretations of different $\boldsymbol{G}_k$'s: On the one hand, each $\boldsymbol{G}_k$ with $1 \leq k \leq p$ has the same interpretation as the lag-k AR coefficient matrix of the VAR(p) model, capturing the short-term cross-sectional dependence. On the other hand, the "MA" coefficient matrices $\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d$ encapsulate the cross-sectional dependence associated with the VARMA-type temporal structure, that is, the long-term influence among the component series that extends into high lags. It is worth noting that the Granger causal network each $\boldsymbol{G}_k$

individually captures is specific to a particular temporal pattern characterized by $\{\ell_{h,k}(\boldsymbol{\omega})\}_{h=1}^{\infty}$. This granularity provides a more detailed perspective on Granger causality from a temporal standpoint; see Section 2.2 for details. Additionally, in view of (D1), the sparsity of $\boldsymbol{G}_k$'s incurs little loss of temporal information, so the essential VARMA-type temporal pattern is well preserved. This is a distinct advantage over regularized VARMA models (Chan, Eisenstat, and Koop 2016; Wilms et al. 2023).

In fact, even compared to sparse finite-order VAR models, the proposed model can be more interpretable for the following two reasons. First, while the AR coefficient matrices $\boldsymbol{A}_h$ must diminish quickly as $h \to \infty$ to ensure stationarity of $\boldsymbol{y}_t$, $\boldsymbol{G}_k$'s do not need to decay thanks to the diminishing $\ell_{h,k}(\boldsymbol{\omega})$'s. Consequently, $\boldsymbol{G}_k$'s, which have relatively strong signals, can be easier to interpret than the diminishing $\boldsymbol{A}_h$'s. Second, similar to the orders (p, q) of VARMA models, the required d is generally small in practice. For example, $d = 2$ works well for the macroeconomic data in Section 6, so we only need to interpret two adjacency matrices $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$. However, if the VAR(P) model were fitted, we would have to interpret P adjacency matrices, where the required P would be much larger.

We summarize the main contributions of this article as follows:

(i) A sparse parametric VAR(∞) model is introduced for high-dimensional time series, which is favorable regarding (I1)–(I3), while avoiding problems (P1) and (P2).
(ii) We develop two ℓ_1-regularized estimators, which can be implemented via efficient block coordinate descent algorithms, and derive their nonasymptotic error bounds under weak sparsity; particularly, our theory takes into account the effect of initializing $\boldsymbol{y}_t = \boldsymbol{0}$ for $t \leq 0$, which is needed for feasible estimation of VAR(∞) models.
(iii) A high-dimensional Bayesian information criterion (BIC) is proposed for model order selection, and its consistency is established.

The remainder of this article is organized as follows. Section 2 introduces the proposed model and its interpretation. Section 3 presents two ℓ_1-regularized estimators and their nonasymptotic theory. Section 4 introduces the proposed BIC. Sections 5 and 6 provide simulation and empirical studies. Section 7 concludes with a brief discussion. The block coordinate descent algorithms for implementing the estimation, additional simulation and empirical results, and all technical proofs are provided in a separate supplementary file.

Unless otherwise specified, we denote scalars, vectors and matrices by lowercase letters (e.g., x), boldface lowercase letters (e.g., $\boldsymbol{x}$), and boldface capital letters (e.g., $\boldsymbol{X}$), respectively. Let $\mathbb{I}_{\{\cdot\}}$ be the indicator function taking value one when the condition is true and zero otherwise. For any $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_q-norm of any $\boldsymbol{x} \in \mathbb{R}^p$ is denoted by $\|\boldsymbol{x}\|_q = (\sum_{j=1}^{p} |x_j|^q)^{1/q}$ for $q > 0$. For any $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, let $\boldsymbol{X}^\top, \sigma_{\max}(\boldsymbol{X})$ (or $\sigma_{\min}(\boldsymbol{X})$), $\lambda_{\max}(\boldsymbol{X})$ (or $\lambda_{\min}(\boldsymbol{X})$), $\mathrm{vec}(\boldsymbol{X})$, $\|\boldsymbol{X}\|_{\mathrm{op}}$, and $\|\boldsymbol{X}\|_{\mathrm{F}}$ be its transpose, largest (or smallest) singular value, largest (or smallest) eigenvalue, vectorization, operator norm $\|\boldsymbol{X}\|_{\mathrm{op}} = \sigma_{\max}(\boldsymbol{X})$, and Frobenius norm $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$, respectively. We use $C > 0$ (or $c > 0$) to denote generic large (or small) absolute constants. For any sequences x_n and y_n, denote $x_n \lesssim y_n$ (or $x_n \gtrsim y_n$) if there is $C > 0$ such that $x_n \leq C y_n$ (or $x_n \geq C y_n$). We write $x_n \asymp y_n$ if $x_n \lesssim y_n$ and $x_n \gtrsim y_n$. In addition, $x_n \gg y_n$ if $y_n / x_n \to 0$ as $n \to \infty$.

2. Proposed Model

2.1. Motivation: Reparameterization of VARMA Models

This section introduces the motivation behind the proposed model. Recall that the shared root cause of problems (P1) and (P2) of the VARMA(1, 1) model, as discussed in Section 1, lies in the matrix multiplications involved in computing the AR coefficient matrices $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \boldsymbol{\Theta}^{h-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$ in the VAR(∞) form of the model. Thus, the key to overcoming both problems is to eliminate the matrix multiplications in the parameterization of $\boldsymbol{A}_h$.

To this end, we show that a reparameterization of $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta})$ free of matrix multiplications can be derived via the following two main steps: (a) Block-diagonalize $\boldsymbol{\Theta}$ via the Jordan decomposition, $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$, where $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ is an invertible matrix, and $\boldsymbol{J} \in \mathbb{R}^{N \times N}$ is the real Jordan form containing eigenvalues of $\boldsymbol{\Theta}$; see (2.1) for details. (b) Then, merge $\boldsymbol{B}$ with all remaining components in the expression of $\boldsymbol{A}_h(\boldsymbol{\Phi}, \boldsymbol{\Theta})$.

Specifically, by Theorem 1 in Hartfiel (1995), for any $0 < n \leq N$, real matrices with n distinct nonzero eigenvalues are dense in the set of all $N \times N$ real matrices with rank at most n. Thus, with only a little loss of generality, we can assume that $\boldsymbol{\Theta}$ is a real matrix with n distinct nonzero eigenvalues, where $n = \mathrm{rank}(\boldsymbol{\Theta})$; a more general result allowing repeated eigenvalues is derived in the technical appendix of Huang, Lu, and Zheng (2023). Then suppose that $\boldsymbol{\Theta}$ has r nonzero real eigenvalues, $\lambda_1, \ldots, \lambda_r$, and s conjugate pairs of nonzero complex eigenvalues, $(\lambda_{r+2m-1}, \lambda_{r+2m}) = (\gamma_m e^{i\theta_m}, \gamma_m e^{-i\theta_m})$ for $1 \leq m \leq s$, where $|\lambda_j| \in (0, 1)$ for $1 \leq j \leq r$, $\gamma_m \in (0, 1)$ and $\theta_m \in (0, \pi)$ for $1 \leq m \leq s$, and i represents the imaginary unit. Therefore, $n = r + 2s$, and the real Jordan form of $\boldsymbol{\Theta}$ is a real block diagonal matrix:

$$\boldsymbol{J} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_r, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_s, \boldsymbol{0}\}, \qquad (2.1)$$
$$\boldsymbol{C}_m = \gamma_m \cdot \begin{pmatrix} \cos\theta_m & \sin\theta_m \\ -\sin\theta_m & \cos\theta_m \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

where $1 \leq m \leq s$; see chap. 3 in Horn and Johnson (2012).

Let $\boldsymbol{A}_1 = \boldsymbol{\Phi} - \boldsymbol{\Theta} := \boldsymbol{G}_1$. Substituting the Jordan decomposition $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$ into the expression of $\boldsymbol{A}_h$, we can show that for all $h \geq 2$, $\boldsymbol{A}_h = \boldsymbol{B}\boldsymbol{J}^{h-1}\boldsymbol{B}^{-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta}) = \sum_{j=1}^{r} \lambda_j^{h-1} \boldsymbol{G}_{1+j} + \sum_{m=1}^{s} \gamma_m^{h-1} [\cos\{(h-1)\theta_m\}\boldsymbol{G}_{1+r+2m-1} + \sin\{(h-1)\theta_m\}\boldsymbol{G}_{1+r+2m}]$, where $\boldsymbol{G}_2, \ldots, \boldsymbol{G}_{1+r+2s} \in \mathbb{R}^{N\times N}$ are determined jointly by $\boldsymbol{B}$ and $\boldsymbol{B}^{-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$; see the proof of Proposition 1 in the supplementary file for details. This result is a reparameterization of $\boldsymbol{A}_h$'s in terms of the scalars λ_j's, γ_m's, θ_m's, and matrices $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{1+r+2s}$. As each $\boldsymbol{A}_h$ is a linear combination of $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{1+r+2s}$, problems (P1) and (P2) are tackled at their root: It not only ensures the identifiability of the parameters λ_j's, γ_m's, θ_m's, and the $\boldsymbol{G}$-matrices, up to a permutation in the indices j and m, but also leads to a significantly reduced computational complexity, such as $O(TN^2 + T^2N)$ for the squared loss function.

In general, the VARMA(p,q) model is given by $\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\varepsilon}_{t-j}$, where $\boldsymbol{\Phi}_i, \boldsymbol{\Theta}_j \in \mathbb{R}^{N\times N}$ for $1 \le i \le p$ and $1 \le j \le q$. Assuming invertibility, it has the following VAR(∞) representation:

$$\boldsymbol{y}_t = \sum_{h=1}^{\infty} \underbrace{\left(\sum_{i=0}^{p\wedge h} \boldsymbol{P} \underline{\boldsymbol{\Theta}}^{h-i} \boldsymbol{P}^\top \boldsymbol{\Phi}_i \right)}_{\boldsymbol{A}_h} \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t, \tag{2.2}$$

$$\underline{\boldsymbol{\Theta}} = \begin{pmatrix} \boldsymbol{\Theta}_1 & \boldsymbol{\Theta}_2 & \cdots & \boldsymbol{\Theta}_{q-1} & \boldsymbol{\Theta}_q \\ \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I} & \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{\Phi}_0 = -\boldsymbol{I}$ and $\boldsymbol{P} = (\boldsymbol{I}_N, \boldsymbol{0}_{N\times N(q-1)})$ are constant matrices, $\underline{\boldsymbol{\Theta}}$ is called the MA companion matrix, and all eigenvalues of $\underline{\boldsymbol{\Theta}}$ are less than one in absolute value; see Lütkepohl (2005). Similar to the VARMA(1, 1) case, the following reparameterization can be derived.

Proposition 1. Suppose that all nonzero eigenvalues of $\underline{\boldsymbol{\Theta}}$ are distinct, and there are r distinct nonzero real eigenvalues of $\underline{\boldsymbol{\Theta}}$, $\lambda_j \in (-1,0)\cup(0,1)$ for $1 \le j \le r$, and s distinct conjugate pairs of nonzero complex eigenvalues of $\underline{\boldsymbol{\Theta}}$, $(\lambda_{r+2m-1}, \lambda_{r+2m}) = (\gamma_m e^{i\theta_m}, \gamma_m e^{-i\theta_m})$ with $\gamma_m \in (0,1)$ and $\theta_m \in (0,\pi)$ for $1 \le m \le s$. Then for all $h \ge 1$, we have

$$\begin{aligned} \boldsymbol{A}_h = & \sum_{k=1}^{p} \mathbb{I}_{\{h=k\}} \boldsymbol{G}_k + \sum_{j=1}^{r} \mathbb{I}_{\{h\ge p+1\}} \lambda_j^{h-p} \boldsymbol{G}_{p+j} \\ & + \sum_{m=1}^{s} \mathbb{I}_{\{h\ge p+1\}} \gamma_m^{h-p} \big[\cos\{(h-p)\theta_m\} \boldsymbol{G}_{p+r+2m-1} \\ & + \sin\{(h-p)\theta_m\} \boldsymbol{G}_{p+r+2m} \big], \end{aligned} \tag{2.3}$$

where $\boldsymbol{G}_k = \boldsymbol{A}_k$ for $1 \le k \le p$, and $\{\boldsymbol{G}_k\}_{k=p+1}^{p+r+2s}$ are determined jointly by $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{B}}_-$, with $\widetilde{\boldsymbol{B}} = \boldsymbol{P}\boldsymbol{B}$ and $\widetilde{\boldsymbol{B}}_- = \boldsymbol{B}^{-1}\left(\sum_{i=0}^{p} \underline{\boldsymbol{\Theta}}^{p-i} \boldsymbol{P}^\top \boldsymbol{\Phi}_i\right)$. In addition, the corresponding term in (2.3) is suppressed if p, r or s is zero.

Throughout this article, we denote $d = p + r + 2s$. Let $\boldsymbol{\omega} = (\lambda_1, \ldots, \lambda_r, \boldsymbol{\eta}_1^\top, \ldots \boldsymbol{\eta}_s^\top)^\top \in \mathbb{R}^{r+2s}$, where $\boldsymbol{\eta}_m = (\gamma_m, \theta_m)^\top$ for $1 \le m \le s$, and $\boldsymbol{g} = \mathrm{vec}(\boldsymbol{G}) \in \mathbb{R}^{N^2 d}$, where $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_d) \in \mathbb{R}^{N\times Nd}$. Then, we can succinctly write (2.3) in the parametric form of $\boldsymbol{A}_h = \boldsymbol{A}_h(\boldsymbol{\omega}, \boldsymbol{g}) = \sum_{k=1}^{d} \ell_{h,k}(\boldsymbol{\omega}) \boldsymbol{G}_k$ for all $h \ge 1$. Here $\ell_{h,k}(\cdot)$'s are real-valued functions predetermined according to (2.3), which can be defined conveniently through a matrix as follows: for any $h \ge 1$ and $1 \le k \le d$, $\ell_{h,k}(\boldsymbol{\omega})$ is the (h,k)th entry of the $\infty \times d$ matrix,

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\omega}) &= \big(\ell_{h,k}(\boldsymbol{\omega})\big)_{h\ge 1, 1\le k\le d} \\ &= \begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{0}_{p\times 1} & \cdots & \boldsymbol{0}_{p\times 1} & \boldsymbol{0}_{p\times 2} & \cdots & \boldsymbol{0}_{p\times 2} \\ \boldsymbol{0}_{\infty\times p} & \boldsymbol{\ell}^I(\lambda_1) & \cdots & \boldsymbol{\ell}^I(\lambda_r) & \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_1) & \cdots & \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_s) \end{pmatrix} \\ &\in \mathbb{R}^{\infty\times d}, \end{aligned}$$

where, for any λ and $\boldsymbol{\eta} = (\gamma, \theta)^\top$, the blocks $\boldsymbol{\ell}^I(\lambda)$ and $\boldsymbol{\ell}^{II}(\boldsymbol{\eta})$ are defined as

$$\boldsymbol{\ell}^I(\lambda) = (\lambda, \lambda^2, \lambda^3, \ldots)^\top \in \mathbb{R}^\infty,$$

$$\boldsymbol{\ell}^{II}(\boldsymbol{\eta}) = \begin{pmatrix} \gamma\cos(\theta) & \gamma^2\cos(2\theta) & \gamma^3\cos(3\theta) & \cdots \\ \gamma\sin(\theta) & \gamma^2\sin(2\theta) & \gamma^3\sin(3\theta) & \cdots \end{pmatrix}^\top \in \mathbb{R}^{\infty\times 2}.$$

2.2. Proposed Sparse Parametric VAR(∞) Model

Motivated by the discussion in Section 2.1, we propose the following VAR(∞) model for high-dimensional time series:

$$\begin{aligned} \boldsymbol{y}_t &= \sum_{h=1}^{\infty} \boldsymbol{A}_h(\boldsymbol{\omega}, \boldsymbol{g}) \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t \\ &= \sum_{k=1}^{d} \boldsymbol{G}_k \sum_{h=1}^{\infty} \ell_{h,k}(\boldsymbol{\omega}) \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\omega} \in (-1,1)^r \times \boldsymbol{\Pi}^s \subset \mathbb{R}^{r+2s}$ is a parameter vector, with $\boldsymbol{\Pi} = [0,1)\times(0,\pi)$, $\ell_{h,k}(\cdot)$'s are known real-valued functions defined as in Section 2.1, $\boldsymbol{G}_k \in \mathbb{R}^{N\times N}$ for $1 \le k \le d$ are parameter matrices with $d = p + r + 2s$. To handle the high-dimensionality, we assume that $\boldsymbol{G}_k$'s are sparse matrices. In this section, we will focus on the exact sparsity as it is instrumental for model interpretability. However, it will be relaxed to weak sparsity in our theoretical analysis; see Assumptions 4 and 4′ in Section 3. We call model (2.4) with exactly or weakly sparse $\boldsymbol{G}_k$'s the Sparse Parametric VAR(∞) (SPVAR(∞)) model.

Note that if no sparsity assumption is imposed on $\boldsymbol{G}_k$'s, then (2.4) provides an alternative low-dimensional time series model comparable to the VARMA model; see Section 2.3 for its stationarity condition. While formulation (2.4) is derived from the VARMA model, it is worth clarifying that it relaxes the restrictions on $\boldsymbol{G}_{p+j}$ for $1 \le j \le r + 2s$. Specifically, by Proposition 1, if $\{\boldsymbol{y}_t\}$ is indeed generated from a VARMA model, then $\boldsymbol{G}_{p+j}$'s would fulfill certain restrictions as determined by the Jordan decomposition of the MA companion matrix $\underline{\boldsymbol{\Theta}}$. By contrast, (2.4) treats these matrices as free parameters.

The resemblance between (2.4) and the VARMA model is mainly achieved by $\ell_{h,k}(\cdot)$'s, which yield VARMA-type decay patterns of $\boldsymbol{A}_h$ as $h \to \infty$. According to (2.3), $\ell_{h,k}(\cdot)$'s implicitly depend on the orders (p,r,s). Note that p and (r,s) are counterparts of the AR and MA orders of the VARMA model, respectively. In fact, when $r = s = 0$, (2.4) reduces to the VAR(p) model, $\boldsymbol{y}_t = \sum_{h=1}^{p} \boldsymbol{G}_h \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t$. For this reason, we call $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p$ and $\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d$ the AR and MA coefficient matrices of the model, respectively. While larger (p,r,s) allow for more complex temporal patterns, similar to the VARMA model, usually it suffices to use small orders in practice; see Section 6 for empirical evidence.

The proposed model can be directly used to infer the multivariate Granger causality (MGC), which concerns Granger causal (GC) relations (Granger 1969) between any pair of component series in $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{N,t})^\top$; see Shojaie and Fox (2021) for an excellent review. By definition, $\{y_{j,t}\}$ is GC for $\{y_{i,t}\}$ if the past information of $y_{j,t}$ can improve the forecast of $y_{i,t}$, where $1 \le i \ne j \le N$. Most existing works study the MGC under the finite-order VAR for its convenience: Under the model $\boldsymbol{y}_t = \sum_{h=1}^{P} \boldsymbol{A}_h \boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t$, $\{y_{j,t}\}$ is GC for $\{y_{i,t}\}$ if $a_{i,j,h} \ne 0$ for some $h \in \{1, \ldots, P\}$, where $a_{i,j,h}$ is the (i,j)th entry of $\boldsymbol{A}_h$, for $1 \le i \ne j \le N$. Notably, while working with $\boldsymbol{A}_h$'s would be

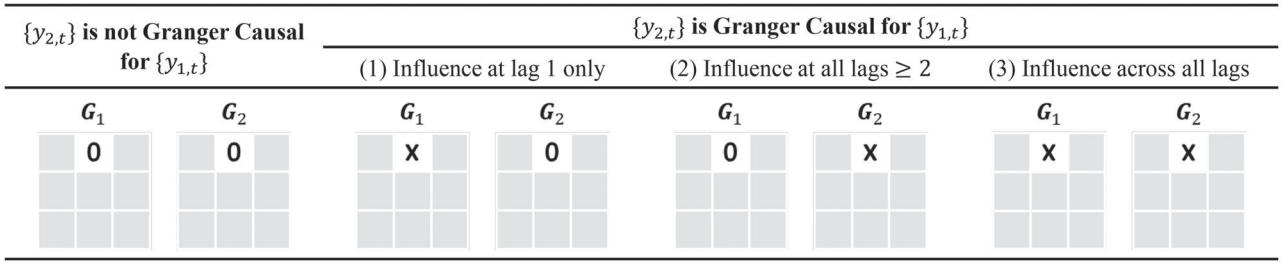

Figure 1. Illustration for different scenarios of Granger causality of $\{y_{2,t}\}$ for $\{y_{1,t}\}$ when $(p, r, s) = (1, 1, 0)$ and $N = 3$, as determined by the $(1, 2)$th entry of $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$. Cell $(1, 2)$ of $\boldsymbol{G}_k$ is marked with "0" when $g_{1,2,k} = 0$, and "X" when $g_{1,2,k} \neq 0$.

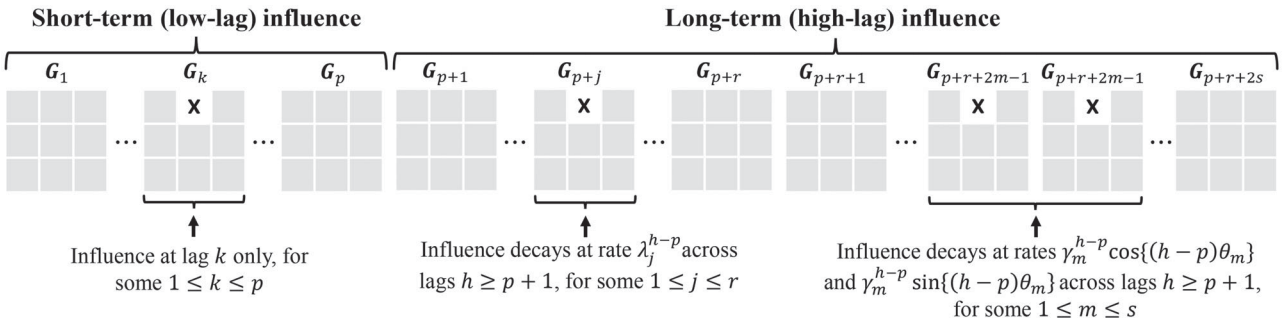

Figure 2. Illustration for different types of lagged influence of $\{y_{2,t}\}$ on $\{y_{1,t}\}$ under general orders (p, r, s) and $N = 3$. Cell $(1, 2)$ of $\boldsymbol{G}_k$ is marked with "X" when $g_{1,2,k} \neq 0$.

infeasible when $P = \infty$, we can directly infer the MGC through $\boldsymbol{G}_k$'s: By (2.4), we have that $\{y_{j,t}\}$ is GC for $\{y_{i,t}\}$ if $g_{i,j,k} \neq 0$ for some $k \in \{1, \ldots, d\}$, where $g_{i,j,k}$ is the (i, j)th entry of $\boldsymbol{G}_k$, for $1 \le i \neq j \le N$; see Figure 1 for an illustration with $(i, j) = (1, 2)$, $(p, r, s) = (1, 1, 0)$, and $N = 3$.

More interestingly, since each $\boldsymbol{G}_k$ captures a piece of cross-sectional information associated with a particular sequence $\{\ell_{h,k}(\boldsymbol{\omega})\}_{h=1}^{\infty}$, we can discern the decay pattern of any GC relations over time, achieving a more granular understanding of the MGC. For simplicity, consider the model for $y_{1,t}$ when $(p, r, s) = (1, 1, 0)$: $y_{1,t} = \sum_{j=1}^{N} g_{1,j,1} y_{j,t-1} + \sum_{j=1}^{N} g_{1,j,2} \sum_{h=2}^{\infty} \lambda^{h-1} y_{j,t-h} + \varepsilon_{1,t}$, where $g_{i,j,k}$ denotes the (i, j)th entry of $\boldsymbol{G}_k$. First, it is clear that $\{y_{j,t}\}$ is GC for $\{y_{1,t}\}$ if $g_{1,j,1}$ and $g_{1,j,2}$ are not both zero. Second, if this GC relation exists, the lagged influence of $\{y_{j,t}\}$ on $\{y_{1,t}\}$ can be classified into the following three scenarios: (a) *lag-one only*, if $g_{1,j,1} \neq 0$ and $g_{1,j,2} = 0$; (b) *all lags beyond lag one*, if $g_{1,j,1} = 0$ and $g_{1,j,2} \neq 0$; and (c) *all lags*, if $g_{1,j,1} \neq 0$ and $g_{1,j,2} \neq 0$. In scenarios (b) and (c), the exponential decay of the influence over time is determined by λ; see Figure 1 for an illustration for $j = 2$.

In general, with orders (p, r, s), the model equation for $y_{1,t}$ will consist of two conditional mean terms: The first term involves the sum of $g_{1,j,k} y_{j,t-k}$ for lags $1 \le k \le p$, whereas the second term captures the influence beyond lag p. The latter involves a weighted mixture of r distinct exponential decay rates and s distinct pairs of damped cosine and sine waves. Then the lagged influence of $\{y_{j,t}\}$ on $\{y_{1,t}\}$ can be generalized to the following three scenarios, if the GC relation exists: (1) *short-term only*, if $g_{1,j,k} \neq 0$ for some $1 \le k \le p$, while $g_{1,j,p+1} = \cdots = g_{1,j,d} = 0$; (2) *long-term only*, if $g_{1,j,1} = \cdots = g_{1,j,p} = 0$, while $g_{1,j,k} \neq 0$ for some $p + 1 \le k \le d$; and (3) *both short-term and long-term influences*, if $g_{1,j,k} \neq 0$ for some $1 \le k \le p$ and some $p + 1 \le k \le d$. A more detailed illustration is given in Figure 2.

Remark 1. In many applications, the cross-sectional dependence may not be time-invariant; for example, Barigozzi and Brownlees (2017) found that the estimated Granger causal network in a sparse VAR system for stock volatilities may be time-varying. Time-varying cross-sectional dependence is also common in behavioral and neural studies: for example, different segments of video time series of freely moving animals may correspond to distinct behaviors (Costacurta et al. 2022), and discrete shifts in the dynamics of neural activity may reflect changes in underlying brain state (Fiecas et al. 2023). To accommodate such applications, the proposed model can be extended to allow $\boldsymbol{G}_k$'s to be time varying; for example, a Markov-switching SPVAR(∞) model may be developed along the lines of Li, Safikhani, and Shojaie (2022).

Remark 2. In VAR models, the GC relations as captured by the coefficient matrices $\boldsymbol{A}_h$'s correspond to lagged cross-sectional dependence, whereas the instantaneous cross-sectional dependence is captured by the variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ of $\boldsymbol{\varepsilon}_t$. While this section focuses on the former, $\boldsymbol{\Sigma}_\varepsilon$ can also be estimated based on residuals from the fitted SPVAR(∞) model; see Remark 5 in Section 3.1.

Remark 3. We can also conduct impulse response analysis based on the VMA(∞) form of the proposed model; see Theorem 1 in Section 2.3 for the VMA(∞) representation. For example, when $(p, r, s) = (1, 1, 0)$, the corresponding MA coefficient matrices are $\boldsymbol{\Psi}_1 = \boldsymbol{G}_1$, $\boldsymbol{\Psi}_2 = \boldsymbol{G}_1^2 + \lambda \boldsymbol{G}_2$, $\boldsymbol{\Psi}_3 = \boldsymbol{G}_1^3 + \lambda \boldsymbol{G}_1 \boldsymbol{G}_2 + \lambda \boldsymbol{G}_2 \boldsymbol{G}_1 + \lambda^2 \boldsymbol{G}_2$, etc. When $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ are both sparse with their nonzero entries in sufficiently different positions, all $\boldsymbol{\Psi}_j$'s will also tend to be sparse; this is indeed the case for the empirical example in Section 6. Thus, we can alternatively interpret the high-dimensional time series via the impulse response analysis.

2.3. Stationarity Condition

We provide a sufficient condition on $\boldsymbol{\omega}$ and $\boldsymbol{G}_k$'s for the existence of a unique strictly stationary solution for (2.4) in the following theorem, which is valid whether $\boldsymbol{G}_k$'s are sparse or not. Similar to the AR companion matrix of a VARMA(p,q) model, denote

$$\underline{\boldsymbol{G}}_1 = \begin{pmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 & \cdots & \boldsymbol{G}_{p-1} & \boldsymbol{G}_p \\ \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I} & \boldsymbol{0} \end{pmatrix}.$$

Theorem 1. Suppose that there exists $0 < \bar{\rho} < 1$ such that

$$\max\{|\lambda_1|, \ldots, |\lambda_r|, \gamma_1, \ldots, \gamma_s\} \le \bar{\rho} \quad \text{and}$$
$$\rho(\underline{\boldsymbol{G}}_1) + \frac{\bar{\rho}}{1-\bar{\rho}} \sum_{k=1}^{r+2s} \rho(\boldsymbol{G}_{p+k}) < 1,$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix, and $\rho(\underline{\boldsymbol{G}}_1)$ disappears when $p = 0$. Moreover, $\{\boldsymbol{\varepsilon}_t\}$ is a strictly stationary sequence. Then there exists a unique strictly stationary solution to the model equation in (2.4), given by $\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j}$, where $\boldsymbol{\Psi}_j = \sum_{k=1}^{\infty} \sum_{j_1+\cdots+j_k=j} \boldsymbol{A}_{j_1} \cdots \boldsymbol{A}_{j_k}$ for $j \ge 1$, with $\boldsymbol{A}_h = \sum_{k=1}^{d} \ell_{h,k}(\boldsymbol{\omega}) \boldsymbol{G}_k$ for $h \ge 1$.

When $r = s = 0$, the condition in Theorem 1 reduces to $\rho(\underline{\boldsymbol{G}}_1) < 1$, which coincides with the necessary and sufficient condition for the strict stationarity of the VAR(p) model. When r and s are not both zero, the stationarity region for $\boldsymbol{G}_k$'s in Theorem 1 will be larger if $\bar{\rho}$ becomes smaller, that is, if $\boldsymbol{A}_h$ diminishes more quickly as $h \to \infty$.

Remark 4. If $\{\boldsymbol{y}_t\}$ is a VARMA(p,q) process fulfilling the representation in (2.4), it is known that the necessary and sufficient condition for its strict stationarity is simply $\rho(\underline{\boldsymbol{G}}_1) < 1$; see Lütkepohl (2005). This suggests that the sufficient condition in Theorem 1 could sometimes be restrictive. Indeed, the condition on $\boldsymbol{\omega}$ and $\boldsymbol{G}_k$'s in Theorem 1 is derived from the necessary and sufficient condition: $\sum_{j=1}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$, where $\boldsymbol{\Psi}_j$'s are functions of $\boldsymbol{A}_h$'s as defined in the VMA(∞) form of $\{\boldsymbol{y}_t\}$ in Theorem 1, and $\|\cdot\|$ is any submultiplicative matrix norm. This motivates us to recommend a more general numerical method to check stationarity for practical use: first compute the sequence $\{\boldsymbol{\Psi}_j\}$ using the parameters $\boldsymbol{\omega}$ and $\boldsymbol{G}_k$'s, and then numerically check whether the partial sum $\sum_{j=1}^{J} \|\boldsymbol{\Psi}_j\|$ converges as $J \to \infty$. This method is applied in Section 6 to check the stationarity of the fitted model.

3. High-Dimensional Estimation

3.1. ℓ_1-Regularized Joint Estimator

We first propose an ℓ_1-regularized estimator for the SPVAR(∞) model via jointly fitting all component series of $\boldsymbol{y}_t$. An alternative estimator will be introduced in the next section.

For $\{\boldsymbol{y}_t\}_{t=1}^{T}$ generated from (2.4) with orders (p,r,s), the squared loss is $\mathbb{L}_T(\boldsymbol{\omega}, \boldsymbol{g}) = T^{-1} \sum_{t=1}^{T} \|\boldsymbol{y}_t - \sum_{h=1}^{\infty} \boldsymbol{A}_h(\boldsymbol{\omega}, \boldsymbol{g}) \boldsymbol{y}_{t-h}\|_2^2 = T^{-1} \sum_{t=1}^{T} \|\boldsymbol{y}_t - \sum_{k=1}^{d} \boldsymbol{G}_k \sum_{h=1}^{\infty} \ell_{h,k}(\boldsymbol{\omega}) \boldsymbol{y}_{t-h}\|_2^2$. Here $\boldsymbol{g} = \mathrm{vec}(\boldsymbol{G})$, where $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_d) \in \mathbb{R}^{N \times Nd}$. Since the loss function depends on observations in the infinite past, initial values for $\{\boldsymbol{y}_t, t \le 0\}$ will be needed in practice. We set them to zero as $\mathbb{E}(\boldsymbol{y}_t) = \boldsymbol{0}$, and then the corresponding loss becomes

$$\widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \boldsymbol{g}) = \frac{1}{T} \sum_{t=1}^{T} \Big\| \boldsymbol{y}_t - \sum_{h=1}^{t-1} \boldsymbol{A}_h(\boldsymbol{\omega}, \boldsymbol{g}) \boldsymbol{y}_{t-h} \Big\|_2^2 \tag{3.1}$$
$$= \frac{1}{T} \sum_{t=1}^{T} \Big\| \boldsymbol{y}_t - \sum_{k=1}^{d} \boldsymbol{G}_k \sum_{h=1}^{t-1} \ell_{h,k}(\boldsymbol{\omega}) \boldsymbol{y}_{t-h} \Big\|_2^2.$$

The initialization effect will be taken into account in our theoretical analysis, and its negligibility is confirmed by our simulation study; see Lemmas S6–S8 and Section S2 in the supplementary file. We propose the ℓ_1-regularized joint estimator (JE) as follows:

$$(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{g}}) = \arg\min_{\boldsymbol{\omega} \in \boldsymbol{\Omega}, \boldsymbol{g} \in \mathbb{R}^{N^2 d}} \left\{ \widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \boldsymbol{g}) + \lambda_g \|\boldsymbol{g}\|_1 \right\}, \tag{3.2}$$

where $\lambda_g > 0$ is the regularization parameter, and $\boldsymbol{\Omega} \subset (-1,1)^r \times \boldsymbol{\Pi}^s$ denotes the parameter space of $\boldsymbol{\omega}$. Let $\boldsymbol{a} = \mathrm{vec}(\boldsymbol{A})$, where $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots)$ is the horizontal concatenation of $\{\boldsymbol{A}_h\}_{h=1}^{\infty}$. Note that $\boldsymbol{a} = (\boldsymbol{L}(\boldsymbol{\omega}) \otimes \boldsymbol{I}_{N^2}) \boldsymbol{g}$. Based on (3.2), the estimator of $\boldsymbol{A}_h$ is $\widehat{\boldsymbol{A}}_h = \sum_{k=1}^{d} \ell_{h,k}(\widehat{\boldsymbol{\omega}}) \widehat{\boldsymbol{G}}_k$ for $h \ge 1$. Then, $\widehat{\boldsymbol{a}} = \mathrm{vec}(\widehat{\boldsymbol{A}}) = (\boldsymbol{L}(\widehat{\boldsymbol{\omega}}) \otimes \boldsymbol{I}_{N^2}) \widehat{\boldsymbol{g}}$, where $\widehat{\boldsymbol{A}} = (\widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \ldots)$.

Denote the true value of any parameter with the superscript "$*$", for example, $\boldsymbol{g}^*$, $\boldsymbol{\omega}^*$, and $\boldsymbol{a}^*$. For $\boldsymbol{\omega}^* \in \boldsymbol{\Omega}$, let $\nu_{\mathrm{lower}}^* = (\min_{1 \le j \le r} |\lambda_j^*|) \wedge (\min_{1 \le m \le s} |\gamma_m^*|)$ and $\nu_{\mathrm{gap}}^* = \min_{1 \le j \ne k \le r+2s} |x_j^* - x_k^*|$, where $x_j^* = \lambda_j^*$ for $1 \le j \le r$ and $(x_{r+2m-1}^*, x_{r+2m}^*) = (\gamma_m^* e^{i\theta_m^*}, \gamma_m^* e^{-i\theta_m^*})$ for $1 \le m \le s$. The assumptions for our theoretical analysis are presented as follows.

Assumption 1 (Parameter space and stationarity). (i) There exists an absolute constant $0 < \bar{\rho} < 1$ such that $|\lambda_1|, \ldots, |\lambda_r|, \gamma_1, \ldots, \gamma_s \le \bar{\rho}$ for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}$; and (ii) the time series $\{\boldsymbol{y}_t\}$ is stationary.

Assumption 2 (Separability). (i) There exists an absolute constant $c_\nu > 0$ such that $\nu_{\mathrm{lower}}^* \ge c_\nu$ and $\nu_{\mathrm{gap}}^* \ge c_\nu$; and (ii) r and s are fixed.

Assumption 3 (Sub-Gaussian errors). Let $\boldsymbol{\varepsilon}_t = \boldsymbol{\Sigma}_\varepsilon^{1/2} \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ is a sequence of i.i.d. random vectors with zero mean and $\mathrm{var}(\boldsymbol{\xi}_t) = \boldsymbol{I}_N$, and $\boldsymbol{\Sigma}_\varepsilon$ is a positive definite covariance matrix. In addition, the coordinates $(\xi_{it})_{1 \le i \le N}$ within $\boldsymbol{\xi}_t$ are mutually independent and σ^2-sub-Gaussian.

Assumption 1(i) ensures that $|\lambda_j|$'s and γ_m's are bounded away from one. A sufficient condition for Assumption 1(ii) is given in Theorem 1. Under stationarity, $\{\boldsymbol{y}_t\}$ has the VMA(∞) form $\boldsymbol{y}_t = \boldsymbol{\Psi}_*(B) \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\Psi}_*(B) = \boldsymbol{I}_N + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j^* B^j$, and B is the backshift operator; see Theorem 1. Let $\mu_{\min}(\boldsymbol{\Psi}_*) = \min_{|z|=1} \lambda_{\min}(\boldsymbol{\Psi}_*(z) \boldsymbol{\Psi}_*^{\mathsf{H}}(z))$ and $\mu_{\max}(\boldsymbol{\Psi}_*) = \max_{|z|=1} \lambda_{\max}(\boldsymbol{\Psi}_*(z) \boldsymbol{\Psi}_*^{\mathsf{H}}(z))$, where $\boldsymbol{\Psi}_*^{\mathsf{H}}(z)$ is the conjugate transpose of $\boldsymbol{\Psi}_*(z)$ for $z \in \mathbb{C}$. It can be verified that $\mu_{\min}(\boldsymbol{\Psi}_*) > 0$; see also Basu and Michailidis (2015). Then we define the positive constants $\kappa_1 = \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon) \mu_{\min}(\boldsymbol{\Psi}_*)$ and $\kappa_2 = \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) \mu_{\max}(\boldsymbol{\Psi}_*)$. Assumption 2(i) requires that

different λ_j^*'s or $\boldsymbol{\eta}_m^*$'s are bounded away from zero and from each other. Since these parameters lie in bounded parameter spaces, this also entails that r and s must be fixed; see Assumption 2(ii). Assumption 3 relaxes the Gaussian assumption commonly used in the literature on high-dimensional time series models (e.g., Basu and Michailidis 2015) to sub-Gaussianity.

Let $\boldsymbol{g}_{\mathrm{AR}} = \mathrm{vec}(\boldsymbol{G}_{\mathrm{AR}})$ and $\boldsymbol{g}_{\mathrm{MA}} = \mathrm{vec}(\boldsymbol{G}_{\mathrm{MA}})$, where $\boldsymbol{G}_{\mathrm{AR}} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p) \in \mathbb{R}^{N\times Np}$ and $\boldsymbol{G}_{\mathrm{MA}} = (\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d) \in \mathbb{R}^{N\times N(r+2s)}$. Let $g_{i,j,k}$ be the (i,j)th entry of $\boldsymbol{G}_k$. Then, we define the weak sparsity of $\boldsymbol{g}_{\mathrm{AR}}^*$ and $\boldsymbol{g}_{\mathrm{MA}}^*$ by restricting them into the ℓ_q-"balls", $\mathbb{B}_q(R_q^{\mathrm{AR}}) := \{\boldsymbol{g}_{\mathrm{AR}} \in \mathbb{R}^{N^2p} \mid \sum_{k=1}^{p}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_q^{\mathrm{AR}}\}$ and $\mathbb{B}_q(R_q^{\mathrm{MA}}) := \{\boldsymbol{g}_{\mathrm{MA}} \in \mathbb{R}^{N^2(r+2s)} \mid \sum_{k=p+1}^{d}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_q^{\mathrm{MA}}\}$, respectively, which is a more general assumption than exact sparsity.

Assumption 4 (Weak sparsity). There exists $q \in [0,1]$ such that $\boldsymbol{g}_{\mathrm{AR}}^* \in \mathbb{B}_q(R_q^{\mathrm{AR}})$ and $\boldsymbol{g}_{\mathrm{MA}}^* \in \mathbb{B}_q(R_q^{\mathrm{MA}})$ for some radii $R_q^{\mathrm{AR}}, R_q^{\mathrm{MA}} > 0$.

Assumption 4 implies that $\boldsymbol{g}^* \in \mathbb{B}_q(R_q)$, where $R_q := R_q^{\mathrm{AR}} + R_q^{\mathrm{MA}}$ and $\mathbb{B}_q(R_q) := \{\boldsymbol{g} \in \mathbb{R}^{N^2d} \mid \sum_{k=1}^{d}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_q\}$. If $q = 0$, Assumption 4 becomes the exact sparsity constraints—$\boldsymbol{g}_{\mathrm{AR}}^*$ and $\boldsymbol{g}_{\mathrm{MA}}^*$ have at most R_q^{AR} and R_q^{MA} nonzero entries, respectively. If $q \in (0,1]$, the ℓ_q-"balls" enforce a certain decay rate on the absolute values of the entries in $\boldsymbol{g}^*$ as the dimension N grows. Note that we do not require R_q^{AR} and R_q^{MA} to be fixed.

A main theoretical challenge is that the loss function $\widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \boldsymbol{g})$ is highly nonconvex with respect to $\boldsymbol{\omega}$. Consequently, the global statistical consistency commonly established for high-dimensional convex M-estimators is not available. However, if the nonconvex loss function exhibits a benign convex curvature over local regions, then a form of local statistical consistency can be established; see, for example, Loh (2017). For many nonconvex M-estimators, certain convexity holds within a constant-radius neighborhood of the true parameter value; for the high-dimensional setup, this is termed as local restricted strong convexity in Loh (2017). Then it can be shown that all local optima within this region can enjoy the same convergence rate as the ℓ_1-regularized least squared estimator for linear regression; see also Janková and van de Geer (2021) and Wang and He (2022) for other works on local statistical guarantees for estimators with nonconvex losses or regularizers. Our method is reminiscent of that for high-dimensional nonconvex M-estimators in the literature. However, our setting is special in that $\widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \boldsymbol{g})$ is only partially nonconvex, as it is convex with respect to $\boldsymbol{g}$, for any fixed $\boldsymbol{\omega}$. Thus, unlike Loh (2017), we only need to restrict $\boldsymbol{\omega}$ within a local region of restricted curvature around $\boldsymbol{\omega}^*$, while $\boldsymbol{g}$ can be free.

Let $\underline{\alpha}_{\mathrm{MA}} = \min_{1\le j\le r+2s}\|\boldsymbol{G}_{p+j}^*\|_{\mathrm{F}}$ and $\overline{\alpha}_{\mathrm{MA}} = \max_{1\le j\le r+2s}\|\boldsymbol{G}_{p+j}^*\|_{\mathrm{F}}$, which are both allowed to grow with N. Then let $\alpha = \overline{\alpha}_{\mathrm{MA}}/\underline{\alpha}_{\mathrm{MA}}$. The local convexity of our loss function around $\boldsymbol{\omega}^*$ is an immediate consequence of the following proposition.

Proposition 2. Suppose that $\underline{\alpha}_{\mathrm{MA}} > 0$. Then under Assumptions 1(i) and 2, there exists a constant $c_{\boldsymbol{\omega}} = \min(2, c/\alpha) > 0$ such that for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ with $\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2 \le c_{\boldsymbol{\omega}}$, it holds $\|\boldsymbol{g} - \boldsymbol{g}^*\|_2 + \underline{\alpha}_{\mathrm{MA}}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2 \lesssim \|\boldsymbol{a} - \boldsymbol{a}^*\|_2^2 \lesssim \|\boldsymbol{g} - \boldsymbol{g}^*\|_2 + \overline{\alpha}_{\mathrm{MA}}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2$, where $\boldsymbol{a} = (\boldsymbol{L}(\boldsymbol{\omega}) \otimes \boldsymbol{I}_{N^2})\boldsymbol{g}$.

Proposition 2 shows that the mapping $(\boldsymbol{\omega}, \boldsymbol{g}) \to \boldsymbol{a}$ is linear within a constant-radius neighborhood of $\boldsymbol{\omega}^*$. Then, since the squared loss of our model is convex with respect to $\boldsymbol{a}$, it is also convex with respect to $(\boldsymbol{\omega}, \boldsymbol{g})$ jointly within the local region of $\boldsymbol{\omega}^*$. Note that the radius $c_{\boldsymbol{\omega}}$ is a constant independent of N and T under the mild condition that $\underline{\alpha}_{\mathrm{MA}} \asymp \overline{\alpha}_{\mathrm{MA}}$, in which case $\{\|\boldsymbol{G}_{p+j}^*\|_{\mathrm{F}}\}_{j=1}^{r+2s}$ are of the same order of magnitude.

Since Proposition 2 relies on confining $\boldsymbol{\omega}$ to a local neighborhood of $\boldsymbol{\omega}^*$, the theoretical guarantees derived in this article are applicable to local estimators. That is, to derive nonasymptotic error bounds, we need to assume that the estimator $\widehat{\boldsymbol{\omega}}$ obtained from (3.2) lies within the local region of $\boldsymbol{\omega}^*$ defined in Proposition 2. We will discuss the practical aspect of this assumption after stating the main result. For simplicity, denote

$$\eta_T = \sqrt{\frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\log\{N(p\vee 1)\}}{\kappa_1^2 T}} \quad \text{and}$$
$$\varpi = \frac{\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)}{\kappa_2(p\vee 1)}.$$

Theorem 2. Suppose that Assumptions 1–4 hold with $\sum_{j=0}^{\infty}\|\boldsymbol{\Psi}_j^*\|_{\mathrm{op}}^2 < \infty$, $R_q \lesssim \varpi/\eta_T^{2-q}$, $\alpha^2 \lesssim R_q/R_q^{\mathrm{MA}}$, $\varpi \lesssim \overline{\alpha}_{\mathrm{MA}}^2 R_q/R_q^{\mathrm{MA}}$, and $\underline{\alpha}_{\mathrm{MA}} > 0$. In addition, assume that $\log N \gtrsim (\kappa_2/\kappa_1)^2$, $T \gtrsim \max\{\kappa_2(p\vee 1)^4, (\kappa_2/\kappa_1)^2(p\vee 1)\log\{(\kappa_2/\kappa_1)\alpha N(p\vee 1)\}\}$, and we solve (3.2) with $\lambda_g \asymp \sqrt{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\log\{N(p\vee 1)\}/T}$. If $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2 \le c_{\boldsymbol{\omega}}$, then with probability at least $1 - C(p\vee 1)e^{-c(\kappa_1/\kappa_2)^2\log N}$,

$$\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_q} \quad \text{and}$$
$$\frac{1}{T}\sum_{t=1}^{T}\Big\|\sum_{h=1}^{t-1}(\widehat{\boldsymbol{A}}_h - \boldsymbol{A}_h^*)\boldsymbol{y}_{t-h}\Big\|_2^2 \lesssim \frac{\eta_T^{2-q}R_q}{\kappa_1^{1-q}}.$$

Combining Theorem 2 with Proposition 2, we immediately have the estimation error bounds $\|\widehat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_q}$ and $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2 \lesssim \underline{\alpha}_{\mathrm{MA}}^{-1}\eta_T^{1-q/2}\sqrt{R_q}$. In particular, under exact sparsity, when $r = s = 0$, the bound for $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2$ in Theorem 2 matches that for the Lasso estimator of VAR(p) models in Basu and Michailidis (2015), while the Gaussian assumption is relaxed. Also note that we do not require the uniqueness of the optimal solution to (3.2), that is, Theorem 2 is valid for all local optima within the constant-radius neighborhood of $\boldsymbol{\omega}^*$.

The JE can be efficiently implemented via the block coordinate descent algorithm; see Section S1.1 of the supplementary file for details. While the value of $c_{\boldsymbol{\omega}}$ is unknown in practice, it is known to be independent of N and T under the mild condition that $\underline{\alpha}_{\mathrm{MA}} \asymp \overline{\alpha}_{\mathrm{MA}}$. The practical implication of the condition $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2 \le c_{\boldsymbol{\omega}}$ is that a reasonably good initialization for $\boldsymbol{\omega}$ will be needed for the optimization algorithm of (3.2). For nonconvex estimators, to meet such requirements, commonly a convex preliminary estimator is used to initialize the algorithm (e.g., Janková and van de Geer 2021). However, for our model, the initialization task can be simplified, because the r values

$\lambda_1, \ldots, \lambda_r \in (-1, 1)$ and the s values $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s \in [0, 1) \times (0, \pi)$ are restricted to bounded spaces and must be well separated from one another; see Assumptions 1(i) and 2(i). In fact, when r and s are larger, the initialization of $\boldsymbol{\omega}$ will be even easier, as the selected r and s values will be denser on the bounded space and hence naturally tend to be closer to the true values. In practice, we recommend considering several different initial values for $\boldsymbol{\omega}$ and selecting the solution of the optimization with minimum in-sample squared loss; see Section S1.2 of the supplementary file for details.

Remark 5. Following the method for sparse VAR(P) models in Krampe and Paparoditis (2021), under a weak sparsity assumption on $\boldsymbol{\Sigma}_\varepsilon$, we can construct a high-dimensional estimator of $\boldsymbol{\Sigma}_\varepsilon$ as $\widehat{\boldsymbol{\Sigma}}_\varepsilon = \mathrm{THR}_{\lambda_\varepsilon}(T^{-1}\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t^\top)$, where the residuals $\widehat{\boldsymbol{\varepsilon}}_t$ are obtained based on $\widehat{\boldsymbol{A}}_h$'s, and $\mathrm{THR}_{\lambda_\varepsilon}(\cdot)$ is the entrywise thresholding function with a chosen threshold parameter $\lambda_\varepsilon > 0$; see Krampe and Paparoditis (2021) for details. Then, based on $\widehat{\boldsymbol{\Sigma}}_\varepsilon$ and $\widehat{\boldsymbol{A}}_h$'s, we can estimate $\mathrm{var}(\boldsymbol{y}_t)$, so the instantaneous cross-sectional dependence can be interpreted. We leave a rigorous theoretical study of this estimation for future research.

Remark 6. While Theorem 2 establishes statistical error bounds, an interesting avenue for future research is to develop a more comprehensive estimation theory that integrates both statistical and algorithmic convergence analyses; see similar works such as Agarwal, Negahban, and Wainwright (2012) and Loh (2017). To tackle the theoretical challenges arising from the nonconvexity of the loss function, Proposition 2 may be leveraged to transform the problem into a convex one within a local region around $\boldsymbol{\omega}^*$.

3.2. ℓ_1-Regularized Rowwise Estimator

While Theorem 2 allows R_q to grow with N, it requires $R_q \lesssim \varpi/\eta_T^{2-q}$; for example, if $q = 0$, then this essentially will become $R_0 \lesssim T/\log\{N(p \vee 1)\}$. However, this requirement could be stringent when T is relatively small. To relax the sparsity requirement, we further introduce a rowwise estimator (RE) based on separately fitting each row of the proposed model.

For $1 \le i \le N$, the ith row of model (2.4) is $y_{i,t} = \sum_{h=1}^{\infty}\boldsymbol{a}_{i,h}^\top\boldsymbol{y}_{t-h} + \varepsilon_{i,t}$, where $\boldsymbol{a}_{i,h} = \sum_{k=1}^{d}\ell_{h,k}(\boldsymbol{\omega})\boldsymbol{g}_{i,k} \in \mathbb{R}^N$ is the ith row of $\boldsymbol{A}_h$, and $\boldsymbol{g}_{i,k} \in \mathbb{R}^N$ is the ith row of $\boldsymbol{G}_k$. Then, the squared loss for the ith row is $\mathbb{L}_{i,T}(\boldsymbol{\omega}, \boldsymbol{g}_i) = T^{-1}\sum_{t=1}^{T}(y_{i,t} - \sum_{h=1}^{\infty}\boldsymbol{a}_{i,h}^\top\boldsymbol{y}_{t-h})^2 = T^{-1}\sum_{t=1}^{T}\{y_{i,t} - \sum_{k=1}^{d}\boldsymbol{g}_{i,k}^\top\sum_{h=1}^{\infty}\ell_{h,k}(\boldsymbol{\omega})\boldsymbol{y}_{t-h}\}^2$, where $\boldsymbol{g}_i = (\boldsymbol{g}_{i,1}^\top, \ldots, \boldsymbol{g}_{i,d}^\top)^\top \in \mathbb{R}^{Nd}$ is the ith row of $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_d)$. Note that joint loss function as defined in the previous section can be decomposed as $\mathbb{L}_T(\boldsymbol{\omega}, \boldsymbol{g}) = \sum_{i=1}^{N}\mathbb{L}_{i,T}(\boldsymbol{\omega}, \boldsymbol{g}_i)$. Thus, the rowwise losses $\mathbb{L}_{i,T}(\cdot)$'s can be minimized separately with respect to $\boldsymbol{g}_i$ for $1 \le i \le N$. Meanwhile, since $\boldsymbol{\omega}$ is shared by all $\mathbb{L}_{i,T}(\cdot)$'s, each rowwise minimization can yield a consistent estimator of $\boldsymbol{\omega}$. This motivates us to consider the following ℓ_1-regularized RE for $1 \le i \le N$:

$$(\widehat{\boldsymbol{\omega}}_i, \widehat{\boldsymbol{g}}_i) = \arg\min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}, \boldsymbol{g}_i\in\mathbb{R}^{Nd}}\left\{\widetilde{\mathbb{L}}_{i,T}(\boldsymbol{\omega}, \boldsymbol{g}_i) + \lambda_g\|\boldsymbol{g}_i\|_1\right\}, \quad (3.3)$$

where $\lambda_g > 0$ is the regularization parameter, and $\widetilde{\mathbb{L}}_{i,T}(\boldsymbol{\omega}, \boldsymbol{g}_i)$ is defined by setting the initial values $\{y_{i,s}, s \le 0\}$ to zero, that is, $\widetilde{\mathbb{L}}_{i,T}(\boldsymbol{\omega}, \boldsymbol{g}_i) = T^{-1}\sum_{t=1}^{T}(y_{i,t} - \sum_{h=1}^{t-1}\boldsymbol{a}_{i,h}^\top\boldsymbol{y}_{t-h})^2 = T^{-1}\sum_{t=1}^{T}\{y_{i,t} - \sum_{k=1}^{d}\boldsymbol{g}_{i,k}^\top\sum_{h=1}^{t-1}\ell_{h,k}(\boldsymbol{\omega})\boldsymbol{y}_{t-h}\}^2$. Let $\boldsymbol{a}_i = (\boldsymbol{a}_{i,1}^\top, \boldsymbol{a}_{i,2}^\top, \ldots)^\top \in \mathbb{R}^\infty$ be the ith row of $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots)$ for $1 \le i \le N$. Note that $\boldsymbol{a}_i = (\boldsymbol{L}(\boldsymbol{\omega}) \otimes \boldsymbol{I}_N)\boldsymbol{g}_i$. Based on (3.3), we have $\widehat{\boldsymbol{a}}_i = (\widehat{\boldsymbol{a}}_{i,1}^\top, \widehat{\boldsymbol{a}}_{i,2}^\top, \ldots)^\top = (\boldsymbol{L}(\widehat{\boldsymbol{\omega}}) \otimes \boldsymbol{I}_N)\widehat{\boldsymbol{g}}_i$, where $\widehat{\boldsymbol{g}}_i = (\widehat{\boldsymbol{g}}_{i,1}^\top, \ldots, \widehat{\boldsymbol{g}}_{i,d}^\top)^\top$, and $\widehat{\boldsymbol{a}}_{i,h} = \sum_{k=1}^{d}\ell_{h,k}(\widehat{\boldsymbol{\omega}}_i)\widehat{\boldsymbol{g}}_{i,k}$. The algorithm for the RE is provided in Section S1.1 of the supplementary file.

Similar to the previous section, we can derive the nonasymptotic error bounds for the RE. For $1 \le i \le N$, let $\boldsymbol{g}_{i,\mathrm{AR}} = (\boldsymbol{g}_{i,1}^\top, \ldots, \boldsymbol{g}_{i,p}^\top)^\top \in \mathbb{R}^{Np}$ and $\boldsymbol{g}_{i,\mathrm{MA}} = (\boldsymbol{g}_{i,p+1}^\top, \ldots, \boldsymbol{g}_{i,d}^\top)^\top \in \mathbb{R}^{N(r+2s)}$. To define the weak sparsity of $\boldsymbol{g}_{i,\mathrm{AR}}^*$ and $\boldsymbol{g}_{i,\mathrm{MA}}^*$, we consider the ℓ_q-"balls", $\mathbb{B}_q(R_{i,q}^{\mathrm{AR}}) := \{\boldsymbol{g}_{i,\mathrm{AR}} \in \mathbb{R}^{Np} \mid \sum_{k=1}^{p}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_{i,q}^{\mathrm{AR}}\}$ and $\mathbb{B}_q(R_{i,q}^{\mathrm{MA}}) := \{\boldsymbol{g}_{i,\mathrm{MA}} \in \mathbb{R}^{N(r+2s)} \mid \sum_{k=p+1}^{d}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_{i,q}^{\mathrm{MA}}\}$. The following is the row-wise counterpart of Assumption 4.

Assumption 4′ (Rowwise weak sparsity). For $1 \le i \le N$, there exists $q \in [0, 1]$ such that $\boldsymbol{g}_{i,\mathrm{AR}}^* \in \mathbb{B}_q(R_{i,q}^{\mathrm{AR}})$ and $\boldsymbol{g}_{i,\mathrm{MA}}^* \in \mathbb{B}_q(R_{i,q}^{\mathrm{MA}})$ for some radii $R_{i,q}^{\mathrm{AR}}, R_{i,q}^{\mathrm{MA}} > 0$.

Let $R_{i,q} = R_{i,q}^{\mathrm{AR}} + R_{i,q}^{\mathrm{MA}}$, and then by Assumption 4′, $\boldsymbol{g}_i^* \in \mathbb{B}_q(R_{i,q}) := \{\boldsymbol{g}_i \in \mathbb{R}^{Nd} \mid \sum_{k=1}^{d}\sum_{j=1}^{N}|g_{i,j,k}|^q \le R_{i,q}\}$. Moreover, Assumption 4′ implies the overall sparsity level in Assumption 4, since it leads to $\boldsymbol{g}_{\mathrm{AR}}^* \in \mathbb{B}_q(R_q^{\mathrm{AR}})$, $\boldsymbol{g}_{\mathrm{MA}}^* \in \mathbb{B}_q(R_q^{\mathrm{MA}})$, and consequently $\boldsymbol{g}^* \in \mathbb{B}_q(R_q)$, where $R_q^{\mathrm{AR}} = \sum_{i=1}^{N}R_{i,q}^{\mathrm{AR}}$, $R_q^{\mathrm{MA}} = \sum_{i=1}^{N}R_{i,q}^{\mathrm{MA}}$, and $R_q = R_q^{\mathrm{MA}} + R_q^{\mathrm{AR}} = \sum_{i=1}^{N}R_{i,q}$.

For $1 \le i \le N$, let $\underline{\alpha}_{i,\mathrm{MA}} = \min_{1\le j\le r+2s}\|\boldsymbol{g}_{i,p+j}^*\|_2$ and $\overline{\alpha}_{i,\mathrm{MA}} = \max_{1\le j\le r+2s}\|\boldsymbol{g}_{i,p+j}^*\|_2$, which are both allowed to grow with N. Denote $\alpha_i = \overline{\alpha}_{i,\mathrm{MA}}/\underline{\alpha}_{i,\mathrm{MA}}$. The rowwise counterparts of Proposition 2 and Theorem 2 are established as follows.

Proposition 3. Fix $1 \le i \le N$. Suppose that $\underline{\alpha}_{i,\mathrm{MA}} > 0$. Then under Assumptions 1(i) and 2, there exists a constant $c_{i,\boldsymbol{\omega}} = \min(2, c/\alpha_i) > 0$ such that for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ with $\|\boldsymbol{\omega}-\boldsymbol{\omega}^*\|_2 \le c_{i,\boldsymbol{\omega}}$, it holds $\|\boldsymbol{g}_i - \boldsymbol{g}_i^*\|_2 + \underline{\alpha}_{i,\mathrm{MA}}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2 \lesssim \|\boldsymbol{a}_i - \boldsymbol{a}_i^*\|_2^2 \lesssim \|\boldsymbol{g}_i - \boldsymbol{g}_i^*\|_2 + \overline{\alpha}_{i,\mathrm{MA}}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2$, where $\boldsymbol{a}_i = (\boldsymbol{L}(\boldsymbol{\omega}) \otimes \boldsymbol{I}_N)\boldsymbol{g}_i$.

Theorem 3. Suppose that Assumptions 1–3 and 4′ hold with $\sum_{j=0}^{\infty}\|\boldsymbol{\Psi}_j^*\|_{\mathrm{op}}^2 < \infty$, $R_{i,q} \lesssim \varpi/\eta_T^{2-q}$, $\alpha_i^2 \lesssim R_{i,q}/R_{i,q}^{\mathrm{MA}}$, $\varpi \lesssim \overline{\alpha}_{i,\mathrm{MA}}^2 R_{i,q}/R_{i,q}^{\mathrm{MA}}$, and $\underline{\alpha}_{i,\mathrm{MA}} > 0$, for $1 \le i \le N$. In addition, assume that $\log N \gtrsim (\kappa_2/\kappa_1)^2$, $T \gtrsim \max\{\kappa_2(p \vee 1)^4, (\kappa_2/\kappa_1)^2(p \vee 1)\log\{(\kappa_2/\kappa_1)\alpha_{\max}N(p \vee 1)\}\}$, with $\alpha_{\max} = \max_{1\le i\le N}\alpha_i$, and we solve (3.3) with $\lambda_g \asymp \sqrt{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\log\{N(p \vee 1)\}/T}$. For $1 \le i \le N$, if $\|\widehat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^*\|_2 \le c_{i,\boldsymbol{\omega}}$, then with probability at least $1 - C(p \vee 1)e^{-c(\kappa_1/\kappa_2)^2\log N}$,

$$\|\widehat{\boldsymbol{a}}_i - \boldsymbol{a}_i^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_{i,q}} \quad \text{and}$$

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\sum_{h=1}^{t-1}(\widehat{\boldsymbol{a}}_{i,h} - \boldsymbol{a}_{i,h}^*)^\top\boldsymbol{y}_{t-h}\right\|_2^2 \lesssim \frac{\eta_T^{2-q}R_{i,q}}{\kappa_1^{1-q}}.$$

Compared to Theorem 3, the sparsity condition in Theorem 3 is much weaker, that is, $R_{i,q} \lesssim \varpi/\eta_T^{2-q}$ for $1 \le i \le N$; or

essentially, $R_{i,0} \lesssim T/\log\{N(p \vee 1)\}$ when $q = 0$. Thus, the RE may be preferred in practice when T is relatively small.

Moreover, by Theorem 3 and Proposition 3, we have $\|\widehat{\boldsymbol{g}}_i - \boldsymbol{g}_i^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_{i,q}}$ and $\|\widehat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^*\|_2 \lesssim \underline{\alpha}_{i,\mathrm{MA}}^{-1}\eta_T^{1-q/2}\sqrt{R_{i,q}}$ for $1 \le i \le N$. Note that each RE $\widehat{\boldsymbol{\omega}}_i$ is a consistent estimator of $\boldsymbol{\omega}^*$, and the estimation error is proportional to $\underline{\alpha}_{i,\mathrm{MA}}^{-1}\sqrt{R_{i,q}}$. On the other hand, as implied by Theorem 2, the estimation error of the JE for $\boldsymbol{\omega}^*$ is proportional to $\underline{\alpha}_{\mathrm{MA}}^{-1}\sqrt{R_q}$. For example, if $R_{i,q} \asymp R_q/N$ and $\underline{\alpha}_{i,\mathrm{MA}}^2 \asymp \underline{\alpha}_{\mathrm{MA}}^2/N$, then the two bounds will be comparable. However, intuitively, allowing different estimators $\widehat{\boldsymbol{\omega}}_i$ for different rows may enhance the flexibility in practice, although it may also increase the risk of overfitting. In addition, combining the results for $\widehat{\boldsymbol{a}}_i, \widehat{\boldsymbol{g}}_i$ and the prediction error across all rows, we have $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_q}$, $\|\widehat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \lesssim \eta_T^{1-q/2}\sqrt{R_q}$, and $T^{-1}\sum_{t=1}^{T}\|\sum_{h=1}^{t-1}(\widehat{\boldsymbol{A}}_h - \boldsymbol{A}_h^*)\boldsymbol{y}_{t-h}\|_2^2 \lesssim \eta_T^{2-q}R_q/\kappa_1^{1-q}$. Here, with a slight abuse of notation, $\widehat{\boldsymbol{a}}$, $\widehat{\boldsymbol{g}}$, and $\widehat{\boldsymbol{A}}_h$'s represent the estimates obtained based on merging the RE $\widehat{\boldsymbol{a}}_i$ or $\widehat{\boldsymbol{g}}_i$ for $1 \le i \le N$. Note that these bounds match exactly those of the JE in the previous section.

In addition to the above upper bounds analysis, we numerically assess the actual comparative performance of RE and JE via simulations in Section S2.2 of the supplementary file. It is shown that they can perform very similarly for the estimation of $\boldsymbol{g}^*$, while RE may outperform JE for the estimation of $\boldsymbol{\omega}^*$, resulting in an overall advantage for the estimation of $\boldsymbol{a}^*$. However, as long as T is not too small compared to R_q, JE and RE tend to have similar out-of-sample forecast accuracy; see the empirical analysis in Section 6 and the simulation study in Section S2.4 of the supplementary file for details. Furthermore, as commented by one referee, the competitive numerical performance of the JE might hint that its more stringent sparsity condition could be an artifact of the proof technique.

4. Model Order Selection

In this section, we introduce a Bayesian information criterion (BIC) based approach to selecting the model orders for the proposed high-dimensional SPVAR(∞) model.

Let $\mathcal{M}^* = (p^*, r^*, s^*)$ denote the true orders. For the feasibility of order selection, it is crucial to ensure that $\mathcal{M}^*$ is irreducible; that is, if $\{\boldsymbol{y}_t\}$ is generated with orders $\mathcal{M}^*$, there is no alternative parameterization with reduced orders. As established in Lemma S14 in the supplementary file, the irreducibility of r^* and s^* is guaranteed if λ_j^*'s, γ_m^*'s, and $\underline{\alpha}_{\mathrm{MA}}$ are nonzero. On the other hand, p^* is irreducible under the following assumption.

Assumption 5 (Irreducibility). $\boldsymbol{G}_{p^*} \neq \sum_{j=1}^{r^*}\boldsymbol{G}_{p^*+j} + \sum_{m=1}^{s^*}\boldsymbol{G}_{p^*+r^*+2m-1}$.

To select the model orders, for any $\mathcal{M} = (p, r, s)$, we define the high-dimensional BIC,

$$\mathrm{BIC}(\mathcal{M}) = \log\widetilde{\mathbb{L}}_T(\widehat{\boldsymbol{\omega}}_{\mathcal{M}}, \widehat{\boldsymbol{g}}_{\mathcal{M}}) + \tau_N d\left[\frac{\log\{N(p \vee 1)\}}{T}\right]^{1-q/2}\log T, \tag{4.1}$$

where $\widehat{\boldsymbol{\omega}}_{\mathcal{M}}$ and $\widehat{\boldsymbol{g}}_{\mathcal{M}}$ denote estimates obtained by fitting the model with orders $\mathcal{M}$ using either the JE in (3.2) or the RE in (3.3). In particular, if the RE is employed, then $\widetilde{\mathbb{L}}_T(\widehat{\boldsymbol{\omega}}_{\mathcal{M}}, \widehat{\boldsymbol{g}}_{\mathcal{M}}) = \sum_{i=1}^{N}\mathbb{L}_{i,T}(\widehat{\boldsymbol{\omega}}_{i,\mathcal{M}}, \widehat{\boldsymbol{g}}_{i,\mathcal{M}})$, where $\widehat{\boldsymbol{\omega}}_{\mathcal{M}}$ and $\widehat{\boldsymbol{g}}_{\mathcal{M}}$ denote collections of $\widehat{\boldsymbol{\omega}}_{i,\mathcal{M}}$'s and $\widehat{\boldsymbol{g}}_{i,\mathcal{M}}$'s, respectively. Note that for notational simplicity, we suppress the dependence of $\widetilde{\mathbb{L}}_T(\cdot)$ and $\mathbb{L}_T(\cdot)$ on $\mathcal{M}$ in this section. Additionally, $\tau_N > 0$ is a sequence possibly dependent on N satisfying the following condition.

Assumption 6 (Penalty parameter). $\tau_N \gtrsim N^{-1}R_q\{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\}^{1-q/2}/\kappa_1^{3-2q}$.

Assumption 6 ensures that the proposed BIC can rule out any overspecified model, $\mathcal{M} \in \mathscr{M}_{\mathrm{over}} = \{\mathcal{M} \in \mathscr{M} \mid p \ge p^*, r \ge r^* \text{ and } s \ge s^*\} \setminus \mathcal{M}^*$. When the constants κ_1, κ_2 and $\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)$ are fixed, Assumption 6 can be simplified to $\tau_N \gtrsim N^{-1}R_q$. While R_q is unknown in practice, to set a reasonable τ_N, we may assume that $R_q \lesssim N$; for example, this will hold if $\boldsymbol{G}_k^*$'s are (weakly) row-sparse. Then it would suffice to fix $\tau_N \equiv \tau > 0$. In practice, we may simply set $q = 0$. We recommend $\tau = 0.05$, which performs well in our simulations.

Based on (4.1), we estimate the model orders by

$$\widehat{\mathcal{M}} = (\widehat{p}, \widehat{r}, \widehat{s}) = \arg\min_{\mathcal{M} \in \mathscr{M}}\mathrm{BIC}(\mathcal{M}),$$

where $\mathscr{M} = \{(p, r, s) \mid 0 \le p \le \overline{p}, 0 \le r \le \overline{r}, 0 \le s \le \overline{s}\}$, with $\overline{\mathcal{M}} := (\overline{p}, \overline{r}, \overline{s})$ being predetermined maximum orders. Since the true orders are usually small in practice, $\overline{\mathcal{M}}$ need not be large; for example $\overline{p} = \overline{r} = \overline{s} = 6$ may be sufficient for most applications. Our simulations show that $\widehat{\mathcal{M}}$ is insensitive to the choice of $\overline{\mathcal{M}}$ as long as it is large enough compared to $\mathcal{M}^*$.

Let $\mathscr{M}_{\mathrm{mis}} = \{\mathcal{M} \in \mathscr{M} \mid p < p^*, r < r^* \text{ or } s < s^*\}$. To establish the conditions that prevent the proposed BIC from selecting any misspecified model, we need to accurately quantify the minimum difference between any $\mathcal{M} \in \mathscr{M}_{\mathrm{mis}}$ and $\mathcal{M}^*$. This analysis is challenging since there is no monotonic nested ordering over $\mathscr{M}$ due to the involvement of three different orders, p, r and s. Particularly, $\mathcal{M} \in \mathscr{M}_{\mathrm{mis}}$ may not be nested within $\mathcal{M}^*$ regarding all three orders. For instance, if $\mathcal{M}^* = (1, 1, 0)$, then a misspecified model may be $\mathcal{M}_1 = (\overline{p}, 0, 0)$ or $\mathcal{M}_2 = (0, \overline{r}, \overline{s})$, where, for example, $\overline{p} = \overline{r} = \overline{s} = 6$. Clearly, we cannot simply treat $\mathcal{M}_1$ or $\mathcal{M}_2$ as a smaller model than $\mathcal{M}^*$, as they possess orders as large as $\overline{p}$, $\overline{r}$, or $\overline{s}$.

To uniformly accommodate the possibly nonnested relationship between $\mathcal{M} \in \mathscr{M}_{\mathrm{mis}}$ and $\mathcal{M}^*$, we leverage their connections with a common model, $\overline{\mathcal{M}} = (\overline{p}, \overline{r}, \overline{s})$. Specifically, we can show that model (2.4) with any orders $\mathcal{M} = (p, r, s) \in \mathscr{M}$ can be reparameterized as the model with $\overline{\mathcal{M}} = (\overline{p}, \overline{r}, \overline{s})$. In addition, the corresponding parameter vectors, denoted $\overline{\boldsymbol{\omega}} \in (-1, 1)^{\overline{r}} \times \boldsymbol{\Pi}^{\overline{s}}$ and $\overline{\boldsymbol{g}} \in \mathbb{R}^{N \times N\overline{d}}$, satisfy the following equality constraints:

$$\overline{\boldsymbol{C}}_1^{\mathcal{M}}\overline{\boldsymbol{\omega}} = \boldsymbol{0} \quad \text{and} \quad \left(\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\overline{\boldsymbol{\omega}}) \otimes \boldsymbol{I}_{N^2}\right)\overline{\boldsymbol{g}} = \boldsymbol{0}, \tag{4.2}$$

where $\overline{\boldsymbol{C}}_1^{\mathcal{M}} \in \mathbb{R}^{(\delta_r + 2\delta_s) \times (\overline{r} + 2\overline{s})}$ is a constant matrix encoding $(\delta_r + 2\delta_s)$ constraints on $\overline{\boldsymbol{\omega}}$, specifying which elements are restricted to zero, and the matrix function $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\overline{\boldsymbol{\omega}}) \in \mathbb{R}^{\delta_d \times \overline{d}}$ encodes δ_d equality constraints on $\overline{\boldsymbol{g}}$ for any given $\overline{\boldsymbol{\omega}}$, with $\delta_r = \overline{r} - r$, $\delta_s = \overline{s} - s$, and $\delta_d = \overline{d} - d$; see Section S7.3 in the supplementary file for detailed definitions of $\overline{\boldsymbol{C}}_1^{\mathcal{M}}$ and $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\cdot)$. In particular, increasing p by one amounts to deleting a particular

row from the constraint matrix $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\cdot)$. On the other hand, increasing r (or s) by one is equivalent to deleting a particular row (or a pair of rows) from both $\overline{\boldsymbol{C}}_1^{\mathcal{M}}$ and $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\cdot)$.

Note that $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\cdot)$ cannot reduce to a constant matrix independent of $\overline{\boldsymbol{\omega}}$ except in the special cases where $p = \overline{p} - 1$ or $r = s = 0$. In particular, when $p = \overline{p} - 1$, the second equation in (4.2) is essentially the reducibility condition of $\overline{p}$, which resembles that for p^* in Assumption 5(i). However, in general, this equation represents much more intricate constraints, since $\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\cdot)$ is a nonlinear function. The complexity of this form can be understood from two perspectives. First, due to the nonlinearity of model (2.4) in $\boldsymbol{\omega}$, the effect of any underspecification in r or s will be highly nonlinear. Second, the order p plays a special role in the definition of $\ell_{h,k}(\cdot)$'s as it is involved in $\mathbb{I}_{\{h\geq p+1\}}\lambda_j^{h-p}$ and $\mathbb{I}_{\{h\geq p+1\}}\gamma_m^{h-p}$; see (2.3). Then, whenever $p \neq p^*$, the exponent $h - p$ will differ from that under $\mathcal{M}^*$ for all lags $h \geq p + 1$, thereby affecting all $\ell_{h,k}(\cdot)$'s. Consequently, due to the interplay between p and $\ell_{h,k}(\cdot)$'s, an underspecification in p generally will also have a nonlinear effect.

Let $\boldsymbol{\Gamma}_{\mathcal{M}} = \{\overline{\boldsymbol{\omega}} \in (-1,1)^{\overline{r}} \times \boldsymbol{\Pi}^{\overline{s}},\ \overline{\boldsymbol{g}} \in \mathbb{R}^{N^2\overline{d}} : \overline{\boldsymbol{C}}_1^{\mathcal{M}}\overline{\boldsymbol{\omega}} = \boldsymbol{0}$ and $(\overline{\boldsymbol{C}}_2^{\mathcal{M}}(\overline{\boldsymbol{\omega}}) \otimes \boldsymbol{I}_{N^2})\overline{\boldsymbol{g}} = \boldsymbol{0}\}$ denote the restricted parameter space for any candidate model $\mathcal{M}$. By leveraging (4.2), we can characterize the minimum difference between the true model and the approximated model of orders $\mathcal{M} \in \mathscr{M}_{\text{mis}}$ via the quantity $\delta_{\mathcal{M}} := \kappa_1 \inf_{(\boldsymbol{\omega},\boldsymbol{g})\in\boldsymbol{\Gamma}_{\mathcal{M}}} \|(\boldsymbol{L}(\boldsymbol{\omega}) \otimes \boldsymbol{I}_{N^2})\boldsymbol{g} - \boldsymbol{a}^*\|_2^2$; see Proposition S1 and the proof of Theorem 4 in Section S7 of the supplementary file for details. We may regard $\delta_{\mathcal{M}}$ as the signal strength of the misspecification. The following assumption guarantees that $\delta_{\mathcal{M}}$ is large enough for the BIC to detect the misspecification.

Assumption 7 (Minimum signal strength).
(i) $\min_{\mathcal{M}\in\mathscr{M}_{\text{mis}}} \delta_{\mathcal{M}}/N \gg (T^{-1}\log N)^{1-q/2}\tau_N \log T$; and (ii) $\max_{\mathcal{M}\in\mathscr{M}_{\text{mis}}} \delta_{\mathcal{M}}^{-1}|\widetilde{\mathbb{L}}_T(\widehat{\boldsymbol{\omega}}_{\mathcal{M}},\widehat{\boldsymbol{g}}_{\mathcal{M}}) - \mathbb{E}\{\mathbb{L}_T(\boldsymbol{\omega}^\circ_{\mathcal{M}},\boldsymbol{g}^\circ_{\mathcal{M}})\}| = o_p(1)$, where $(\boldsymbol{\omega}^\circ_{\mathcal{M}},\boldsymbol{g}^\circ_{\mathcal{M}})$ is the minima of $\mathbb{E}\{\mathbb{L}_T(\boldsymbol{\omega}_{\mathcal{M}},\boldsymbol{g}_{\mathcal{M}})\}$ over the parameter space $\boldsymbol{\omega}_{\mathcal{M}} \in (-1,1)^r \times \boldsymbol{\Pi}^s$ and $\boldsymbol{g}_{\mathcal{M}} \in \mathbb{R}^{N^2 d}$.

Note that $\delta_{\mathcal{M}}/N$ can be viewed as the average level of misspecification across N rows of the model equation. As mentioned earlier, we may let $\tau_N \equiv \tau$ under mild condition. Thus, the lower bound in Assumption 7(i) tends to zero as $T \to \infty$. Assumption 7(ii) requires that the empirical loss for any fitted misspecified model converges to some population loss at a rate faster than $\delta_{\mathcal{M}}$ as $T \to \infty$. Here the mispecified model with parameters $(\boldsymbol{\omega}^\circ_{\mathcal{M}},\boldsymbol{g}^\circ_{\mathcal{M}})$ can be understood as the best approximation of the process $\{\boldsymbol{y}_t\}$ under the misspecification. Now we are ready to establish the consistency of the estimator $\widehat{\mathcal{M}}$.

Theorem 4. If the JE (or the RE) is used, suppose that for any $\mathcal{M} \in \mathscr{M}_{\text{over}}$, there is a subvector $\widehat{\boldsymbol{\omega}}_{\mathcal{M}^*} \in (-1,1)^{r^*} \times \boldsymbol{\Pi}^{s^*}$ of $\widehat{\boldsymbol{\omega}}_{\mathcal{M}}$ (or $\widehat{\boldsymbol{\omega}}_{i,\mathcal{M}^*} \in (-1,1)^{r^*} \times \boldsymbol{\Pi}^{s^*}$ of $\widehat{\boldsymbol{\omega}}_{i,\mathcal{M}}$ with $1 \leq i \leq N$) such that $\|\widehat{\boldsymbol{\omega}}_{\mathcal{M}^*} - \boldsymbol{\omega}^*\|_2 \leq c_{\boldsymbol{\omega}}$ (or $\|\widehat{\boldsymbol{\omega}}_{i,\mathcal{M}^*} - \boldsymbol{\omega}^*\|_2 \leq c_{i,\boldsymbol{\omega}}$ with $1 \leq i \leq N$), and the conditions in Theorem 2 (or 3) hold with $\mathcal{M} = \mathcal{M}^*$. In addition, suppose that $\overline{\mathcal{M}}$ is fixed, with $\overline{p} \geq p^*, \overline{r} \geq r^*$ and $\overline{s} \geq s^*$. Under Assumptions 5–7, $\mathbb{P}(\widehat{\mathcal{M}} = \mathcal{M}^*) \to 1$ as $N, T \to \infty$.

5. Simulation Experiments

In this section, we present two simulation experiments to verify the estimation error rates of the JE and the consistency of the BIC. Four additional experiments on the estimation error of the RE, its comparison with the JE, sensitivity analysis of the initialization for $\{\boldsymbol{y}_t, t \leq 0\}$, and comparison of the proposed estimators with competing approaches are provided in Section S2 of the supplementary file.

Throughout this section, we generate $\{\boldsymbol{y}_t\}$ from model (2.4), where $\{\boldsymbol{\varepsilon}_t\}$ are generated independently from $N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ with $\sigma = 0.2$, and each $\boldsymbol{G}_k$ is exactly sparse with cN nonzero entries for $1 \leq k \leq d$, so the overall sparsity level is $R_0 = cdN$. We generate $\{\boldsymbol{G}_k\}_{k=1}^d$ by drawing their nonzero entries independently from the uniform distribution on $[-0.5, 0.5]$. Then, to ensure the stationarity of $\{\boldsymbol{y}_t\}$, after setting $\boldsymbol{\omega}$, we rescale all $\boldsymbol{G}_k$'s by a common factor such that $\rho(\underline{\boldsymbol{G}}_1) + \bar{\rho}\sum_{k=1}^{r+2s}\rho(\boldsymbol{G}_{p+k})/(1-\bar{\rho}) = 0.8$; see Theorem 1.

In the first experiment, we examine the estimation error rates for the JE. Two data generating processes are considered: $(p,r,s) = (1,1,0)$ (DGP1) and $(1,0,1)$ (DGP2), where $\lambda_1 = -0.6$ for DGP1, and $(\gamma_1,\theta_1) = (0.6,\pi/4)$ for DGP2. We let all $\boldsymbol{G}_k$'s be row-sparse matrices with three nonzero entries in each row, that is, $R_0 = 3dN$, where $N = 10, 20, 40$, or 80. Note that by Theorem 2, we have $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2/\sqrt{N} \lesssim \eta_T\sqrt{R_0/N}$, $\|\widehat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2/\sqrt{N} \lesssim \eta_T\sqrt{R_0/N}$, and $\underline{\alpha}_{\text{MA}}\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2/\sqrt{N} \lesssim \eta_T\sqrt{R_0/N}$, where $\eta_T = \sqrt{T^{-1}\log N}$. To verify these bounds, we choose a grid of equally spaced values for the theoretical rate $\eta_T\sqrt{R_0/N} = \sqrt{3T^{-1}d\log N}$ within the range of $\mathscr{I}_1 = [0.3756, 0.4981]$ for DGP1 and $\mathscr{I}_2 = [0.46, 0.61]$ for DGP2. Then we compute T given the theoretical rate, N and d. The selected ranges $\mathscr{I}_1$ and $\mathscr{I}_2$ lead to the same range of T for both DGPs under any N; that is, the ranges of the x-axis in Figure 3 are set such that the corresponding points in upper and lower panels share the same T. Across all settings, T falls in the range of $[55, 186]$. Figure 3 plots the scaled estimation errors $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2/\sqrt{N}$, $\|\widehat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2/\sqrt{N}$, and $\underline{\alpha}_{\text{MA}}\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2/\sqrt{N}$, averaged over 500 replications, against the theoretical rate $\eta_T\sqrt{R_0/N}$. An approximately linear relationship can be observed across all settings, confirming our theoretical results.

In the second experiment, we verify the consistency of the proposed BIC. Three cases of true model orders are considered: $(p^*, r^*, s^*) = (0,0,1)$, $(0,1,1)$, and $(1,0,1)$, referred to as DGPs 1, 2, and 3, respectively. We set $N = 40$, $\theta_1 = \pi/4$, and $\lambda_1 = -\gamma_1 = \bar{\rho}$, where three choices of the decay rate are considered: $\bar{\rho} \in \{0.45, 0.5, 0.5\}$. For $1 \leq k \leq d$, each $\boldsymbol{G}_k$ contains $3N$ nonzero entries, so $R_0 = 3dN$, but unlike the first experiment, we do not restrict each row of $\boldsymbol{G}_k$ to have exactly three nonzero entries. We set $\tau = 0.05$ and $\overline{p} = \overline{r} = \overline{s} = 9$; the results are found to be unchanged if the maximum orders are 3. Figure 4 displays the proportion of correct order selection based on 500 replications for each setting, with the models fitted by the JE; the results for the RE are very similar and hence omitted. It shows that the BIC generally performs better as T or $\bar{\rho}$ increases, and the proportion of correct order selection eventually becomes close to one with sufficiently large T. Thus, the consistency of the BIC is verified. Additionally, the required sample size for achieving accurate order selection follows this order among the three DGPs: DGP1 < DGP3 < DGP2. To understand this, first note that $R_0 =$

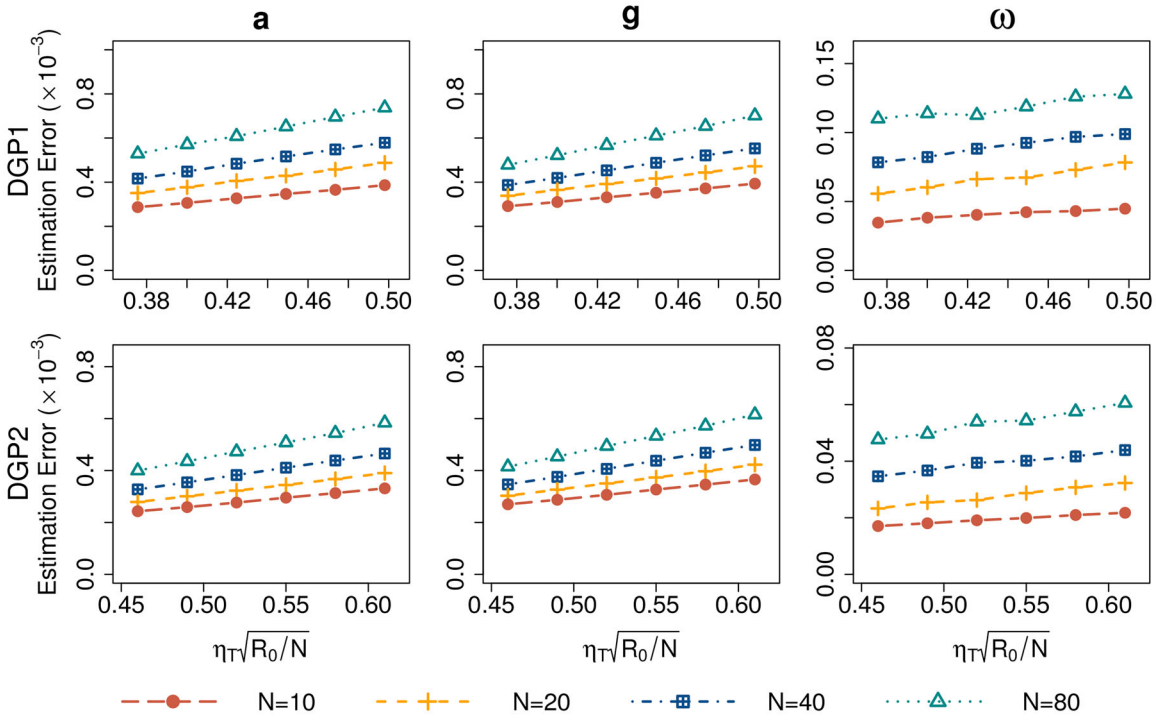

Figure 3. Plots of scaled estimation errors $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}^*\|_2/\sqrt{N}$ (left panel), $\|\widehat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2/\sqrt{N}$ (middle panel), and $\underline{\alpha}_{\text{MA}}\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2/\sqrt{N}$ (right panel) against theoretical rate $\eta_T\sqrt{R_0/N}$ for JE.

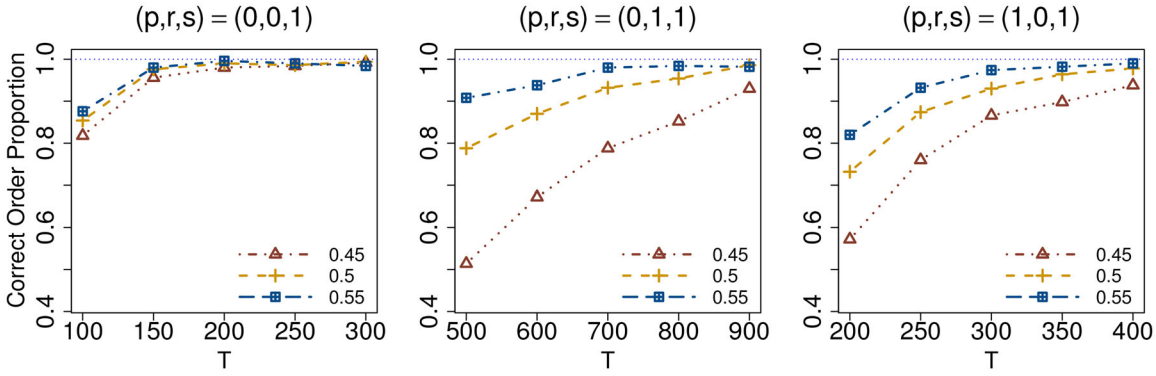

Figure 4. Proportion of correct model order selection for three DGPs and three choices of decay rates, $\bar{\rho} \in \{0.45, 0.5, 0.55\}$.

$6N, 9N$, and $9N$ for DGPs 1, 2, and 3, respectively. Thus, the estimation accuracy is highest for DGP1, and so is the order selection accuracy. Moreover, since DGP2 has a more complex temporal structure than DGP3, it leads to greater challenges in estimating $\boldsymbol{\omega}$ and, consequently, in order selection.

6. Empirical Analysis

We analyze $N = 20$ quarterly macroeconomic variables of the United States from the first quarter of 1969 to the fourth quarter of 2007. These are key economic and financial indicators collected by Koop (2013), seasonally adjusted as needed. We conduct the transformations following Koop (2013) to make all series stationary, resulting in a sample of length $T = 194$. Then each series is normalized to have zero mean and unit variance; see Table S1 in the supplementary file for detailed descriptions of the 20 variables.

We first fit the proposed model to the entire dataset. Using the JE and the proposed BIC, we select $(p, r, s) = (1, 1, 0)$, so $d = 2$, and the fitted model is $\boldsymbol{y}_t = \widehat{\boldsymbol{G}}_1\boldsymbol{y}_{t-1} + \sum_{h=2}^{\infty}(-0.45)^{h-1}\widehat{\boldsymbol{G}}_2\boldsymbol{y}_{t-h} + \boldsymbol{\varepsilon}_t$, where $\widehat{\boldsymbol{G}}_1$ and $\widehat{\boldsymbol{G}}_2$ are displayed in Figure 5; the estimation results based on the RE are roughly similar and provided in the supplementary file. The stationarity of the model is confirmed by the method in Remark 4. As discussed in Section 2.2, $\widehat{\boldsymbol{G}}_1$ and $\widehat{\boldsymbol{G}}_2$ captures lag-one (or short-term) and higher-lag (or long-term) dependence, respectively. Note that $\widehat{\boldsymbol{G}}_1$ is much denser than $\widehat{\boldsymbol{G}}_2$, suggesting that many dynamic interactions are short-term. However, most of the nonzero entries in $\widehat{\boldsymbol{G}}_2$ are fairly large in absolute value, supporting the necessity of a VARMA-type model. For the Granger causal (GC) interpretation, take the model equation for real GDP (RGDP) as an example:

$$y_{\text{RGDP},t} = 0.17y_{\text{Cons},t-1} + 0.11y_{\text{IP:total},t-1} + 0.07y_{\text{HStarts:total},t-1} + 0.12y_{\text{S\&P:indust},t-1}$$

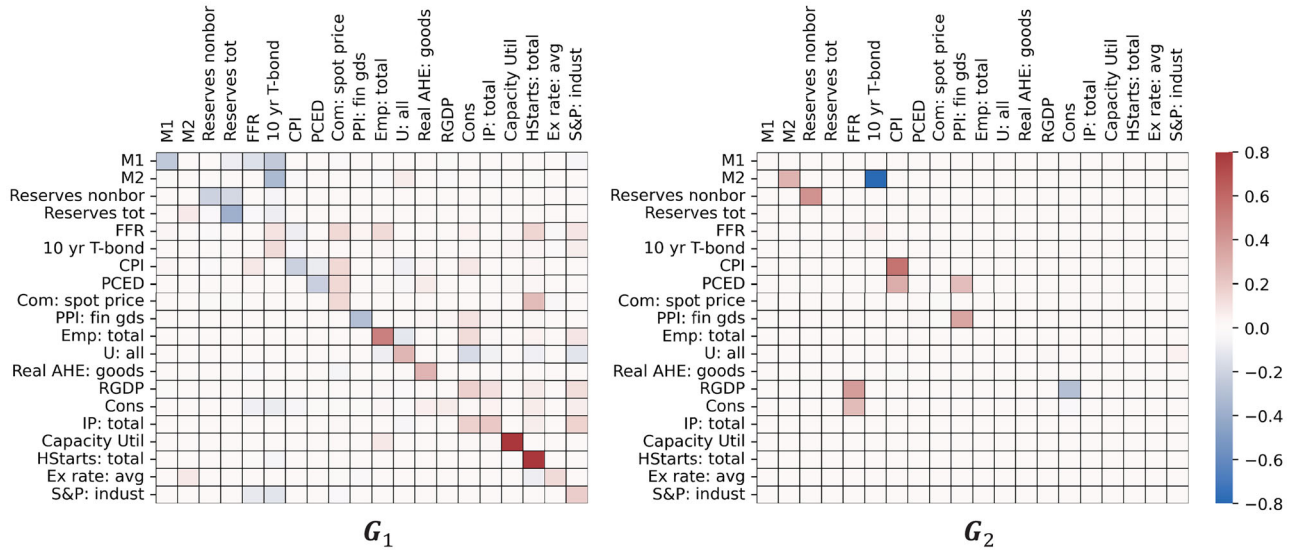

Figure 5. Estimates of G_1 and G_2 for the proposed model based on JE.

$$+\sum_{h=2}^{\infty}(-0.45)^{h-1}(0.39y_{\mathrm{FFR},t-h}-0.30y_{\mathrm{Cons},t-h})$$
$$+\varepsilon_{\mathrm{RGDP},t},$$

suppressing other lag-one terms with coefficients less than 0.014 in absolute value for brevity. The above equation indicates that five time series are GC for RGDP and can be categorized as follows: (a) the industrial production index (IP: total), housing starts (HStarts: total), and S&P stock price index (S&P: indust) only have short-term influence on RGDP; (b) the federal funds rate (FFR) only has long-term influence on RGDP; (c) the real personal consumption expenditures (Cons) has both short-term and long-term influence on RGDP. For other insights from the estimation results, see Section S3 in the supplementary file for more discussions.

Next we evaluate the forecasting performance via a rolling procedure: First set the forecast origin to $t = 166$ (Q4-2000). For each $k = 1,\ldots,28$, fit the model using the data of $1 \le t \le T_{\mathrm{train}} = 165 + k$, and then compute the one-step ahead forecast for $t = 166 + k$. Thus, rolling forecasts over the period of Q1-2001 to Q4-2007 are obtained. We measure the forecast error by $\|\widehat{\boldsymbol{y}}_t - \boldsymbol{y}_t\|_2$; our findings based on the ℓ_1-norm are similar and hence are omitted. For the proposed model, we consider both JE and RE, and implement them using a fixed regularization parameter λ_g throughout the forecasting period. Five other competing approaches are considered as follows:

(i) VAR OLS: As a low-dimensional baseline, we consider the VAR(4) model fitted via the OLS method, where the lag order 4 is employed following Koop (2013).
(ii) VAR Lasso: Since the VAR(∞) model can be approximated by the VAR(P) with $P \to \infty$ as $T \to \infty$, we fit the sparse VAR(P) model via the Lasso with $P = \lfloor 1.5\sqrt{T_{\mathrm{train}}} \rfloor$ following the first-stage estimation in Wilms et al. (2023).
(iii) VAR HLag: Same as (ii) except that the hierarchical lag (HLag) regularization in Nicholson et al. (2020) is used instead of the ℓ_1-regularization.
(iv) VARMA ℓ_1: Sparse VARMA(p,q) (Wilms et al. 2023) with the ℓ_1-regularization for the second stage and $p = q = \lfloor 0.75\sqrt{T_{\mathrm{train}}} \rfloor$ as in the above article.
(v) VARMA HLag: Same as (iv) except that the HLag regularization is used at the second stage.

We implement (ii)–(v) by the R package `bigtime` which offers two regularization parameter selection methods, cross validation (CV) and BIC. We observe that neither one of these two methods uniformly outperforms the other throughout the forecasting period. To better ensure the competitiveness of (ii)–(v), we obtain the forecast errors under both CV and BIC and only report the smaller value for each rolling step.

The average forecast error over the entire forecast period is 5.367, 4.307, 4.069, 4.318, 4.144, 3.971, and 3.968 for VAR OLS, VAR Lasso, VAR HLag, VARMA ℓ_1, VARMA HLag, SPVAR(∞) JE, and SPVAR(∞) RE, respectively. Among the 28 rolling steps, each of these approaches performs best 4, 4, 0, 2, 2, 10, and 6 times, respectively. Thus, based on these measures, SPVAR(∞) has the highest overall forecast accuracy among all models, and the performance of JE and RE are very similar; see Table S2 in the supplementary file for the forecast errors of all seven methods for each rolling step. Moreover, to check whether the advantage of the SPVAR(∞)-based forecasts is statistically significant, we conduct the model confidence set (MCS) procedure of Hansen, Lunde, and Nason (2011) implemented by the R package `MCS`. We find that based on either the Tmax or TR statistic, the 97.5% MCS only includes SPVAR(∞) JE and SPVAR(∞) RE, confirming that the proposed model indeed outperforms the competing ones in terms of forecasting for the data.

7. Conclusion and Discussion

This article develops the SPVAR(∞) model as a tractable variant of the VARMA model for high-dimensional time series. It overcomes the drawbacks in identification, computation, and interpretation of the latter, while greater statistical efficiency and Granger causal interpretations are achieved by imposing

sparsity on the parameter matrices capturing the cross-sectional dependence. To the best of our knowledge, it is the first high-dimensional sparse VARMA- or VAR(∞)-type model with all of the above advantages.

There is a vast literature on nonlinear and nonstationary VAR models (e.g., Kalliovirta, Meitz, and Saikkonen 2016; Zhang and Wu 2021), factor-augmented VAR (Miao, Phillips, and Su 2022), and other extensions. The method in this article can be extended to develop corresponding VAR(∞) counterparts; for example, (2.4) can be extended to the nonlinear model: $\boldsymbol{y}_t = f(\boldsymbol{x}_t^{[1]}, \ldots, \boldsymbol{x}_t^{[d]}) + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{x}_t^{[k]} = \sum_{h=1}^{\infty} \ell_{h,k}(\boldsymbol{\omega})\boldsymbol{y}_{t-h}$ for $1 \le k \le d$ parsimoniously summarize the temporal information over all lags into d predictors. Other interesting extensions include imposing group sparsity on $\boldsymbol{G}_k$'s to capture group-wise homogeneity (Basu, Shojaie, and Michailidis 2015), extending $\ell_{h,k}(\boldsymbol{\omega})$'s to polynomial decay functions for long-memory time series (Chung 2002), and incorporating dynamic factor structures (Wang et al. 2022). Lastly, it is important to study the high-dimensional statistical inference under the proposed model, for example, hypothesis testing for Granger causality (Chernozhukov et al. 2021; Babii, Ghysels, and Striaukas 2022).

Supplementary Materials

Online supplementary materials contain the block coordinate descent algorithms for implementing the proposed estimation, additional simulation studies and empirical results, and all technical proofs for the main paper.

Acknowledgments

I am grateful to the Editor, Associate Editor, and two anonymous referees for their valuable comments which led to substantial improvement of this article.

Disclosure Statement

There are no competing interests to declare.

Funding

This research was partially supported by NSF grant DMS-2311178.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012), "Fast Global Convergence of Gradient Methods for High-Dimensional Statistical Recovery," *The Annals of Statistics*, 40, 2452–2482. [8]

Athanasopoulos, G., and Vahid, F. (2008), "VARMA versus VAR for Macroeconomic Forecasting," *Journal of Business & Economic Statistics*, 26, 237–252. [2]

Babii, A., Ghysels, E., and Striaukas, J. (2022), "High-Dimensional Granger Causality Tests with an Application to VIX and News," *Journal of Financial Econometrics*, to appear. [13]

Barigozzi, M., and Brownlees, C. (2017), NETS: Network Estimation for Time Series," *Journal of Applied Econometrics*, 34, 347–364. [5]

Basu, S., and Matteson, D. S. (2021), "A Survey of Estimation Methods for Sparse High-Dimensional Time Series Models," ArXiv preprint arXiv:2107.14754. [1]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-Dimensional Time Series Models," *The Annals of Statistics*, 43, 1535–1567. [6,7]

Basu, S., Shojaie, A., and Michailidis, G. (2015), "Network Granger Causality with Inherent Grouping Structure," *Journal of Machine Learning Research*, 16, 417–453. [13]

Chan, J. C., Eisenstat, E., and Koop, G. (2016), "Large Bayesian VARMAs," *Journal of Econometrics*, 192, 374–390. [1,2,3]

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2021), "Lasso-Driven Inference in Time and Space," *The Annals of Statistics*, 49, 1702–1735. [13]

Chung, C.-F. (2002), "Sample Means, Sample Autocovariances, and Linear Regression of Stationary Multivariate Long Memory Processes," *Econometric Theory*, 18, 51–78. [13]

Costacurta, J., Duncker, L., Sheffer, B., Gillis, W., Weinreb, C., Markowitz, J., Datta, S. R., Williams, A., and Linderman., S. (2022), "Distinguishing Discrete and Continuousbehavioral Variability Using Warped Autoregressive HMMs," in *Advances in Neural Information Processing Systems* (Vol. 35), pp. 23838–23850. [5]

Davis, R. A., Zang, P., and Zheng, T. (2016), "Sparse Vector Autoregressive Modeling," *Journal of Computational and Graphical Statistics*, 25, 1077–1096. [1]

Dias, G. F., and Kapetanios, G. (2018), "Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets," *Journal of Econometrics*, 202, 75–91. [2]

Dowell, J., and Pinson, P. (2016), "Very-Short-Term Probabilistic Wind Power Forecasts by Sparse Vector Autoregression," *IEEE Transactions on Smart Grid*, 7, 763–770. [1]

Fiecas, M. B., Coffman, C., Xu, M., Hendrickson, T. J., Mueller, B. A., Klimes-Dougan, B., and Cullen, K. R. (2023), "Approximate Hidden Semi-Markov Models for Dynamic Connectivity Analysis in Resting-State FMRi," *Statistics and Its Interface*, 16, 259–277. [5]

Gorrostieta, C., Ombao, H., Bédard, P., and Sanes, J. N. (2012), "Investigating Brain Connectivity Using Mixed Effects Vector Autoregressive Models," *NeuroImage*, 59, 3347–3355. [1]

Granger, C. W. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37, 424–438. [1,4]

Hansen, P. R., Lunde, A., and Nason, J. M. (2011), "The Model Condence Set," *Econometrica*, 79, 453–497. [12]

Hartfiel, D. J. (1995), "Dense Sets of Diagonalizable Matrices," *Proceedings of the American Mathematical Society*, 123, 1669–1672. [3]

Horn, R. A., and Johnson, C. R. (2012), *Matrix Analysis* (2nd ed.), New York: Cambridge University Press. [3]

Huang, F., Lu, K., and Zheng, Y. (2023). "SARMA: Scalable Low-Rank High-Dimensional Autoregressive Moving Averages via Tensor Decomposition," working paper. [2,3]

Janková, J., and van de Geer, S. (2021), "De-Biased Sparse PCA: Inference and Testing for Eigenstructures of Large Covariance Matrices," *IEEE Transactions on Information Theory*, 67, 2507–2527. [7]

Kalliovirta, L., Meitz, M., and Saikkonen, P. (2016), "Gaussian Mixture Vector Autoregression," *Journal of Econometrics*, 192, 485–498. [13]

Koop, G. M. (2013), "Forecasting with Medium and Large Bayesian VARs," *Journal of Applied Econometrics*, 28, 177–203. [11,12]

Krampe, J., and Paparoditis, E. (2021), "Sparsity Concepts and Estimation Procedures for High-Dimensional Vector Autoregressive Models," *Journal Time Series Analysis*, 42, 554–579. [8]

Li, X., Safikhani, A., and Shojaie, A. (2022), "Estimation of High-Dimensional Markov-Switching VAR Models with an Approximate EM Algorithm," arXiv preprint arXiv:2210.07456. [5]

Loh, P.-L. (2017), "Statistical Consistency and Asymptotic Normality for High-Dimensional Robust M-estimators," *The Annals of Statistics*, 45, 866–896. [7,8]

Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009), "Grouped Graphical Granger Modeling for Gene Expression Regulatory Networks Discovery," *Bioinformatics*, 25, i110–i118. [1]

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Berlin: Springer. [1,4,6]

Metaxoglou, K., and Smith, A. (2007), "Maximum Likelihood Estimation of VARMA Models Using a State-Space EM Algorithm," *Journal of Time Series Analysis*, 28, 666–685. [2]

Miao, K., Phillips, P. C., and Su, L. (2022), "High-Dimensional VARs with Common Factors," *Journal of Econometrics*. to appear. [13]

Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020), "High Dimensional Forecasting via Interpretable Vector Autoregression," *Journal of Machine Learning Research*, 21, 1–52. [1,2,12]

Shojaie, A., Basu, S., and Michailidis, G. (2012), "Adaptive Thresholding for Reconstructing Regulatory Networks from Time-Course Gene Expression Data," *Statistics in Biosciences*, 4, 66–83. [2]

Shojaie, A., and Fox, E. B. (2021), "Granger Causality: A Review and Recent Advances," arXiv preprint arXiv:2105.02675. [4]

Wang, D., Zheng, Y., Lian, H., and Li, G. (2022), "High-Dimensional Vector Autoregressive Time Series Modeling via Tensor Decomposition," *Journal of the American Statistical Association*, 117, 1338–1356. [1,13]

Wang, L., and He, X. (2022), "Analysis of Global and Local Optima of Regularized Quantile Regression in High Dimensions: A Subgradient Approach," *Econometric Theory*, 1–45. [7]

Wilms, I., Basu, S., Bien, J., and Matteson, D. (2023), "Sparse Identification and Estimation of Large-Scale Vector Autoregressive Moving Averages," *Journal of the American Statistical Association*, 118, 571–582. [1,2,3,12]

Zhang, D., and Wu, W. B. (2021), "Convergence of Covariance and Spectral Density Estimates for High-Dimensional Locally Stationary Processes," *The Annals of Statistics*, 49, 233–254. [13]