Submitted to the Annals of Statistics

SARMA: A COMPUTATIONALLY SCALABLE HIGH-DIMENSIONAL TIME SERIES MODEL

BY FEIQING HUANG[1], YAO ZHENG[2,‡], KEXIN LU[1,*] AND GUODONG LI[1,†]

[1]*Department of Statistics, University of Hong Kong, amieehuang@hku.hk; *neithen@hku.hk; †gdli@hku.hk*

[2]*Department of Statistics, University of Connecticut, ‡yao.zheng@uconn.edu*

This paper introduces a novel parametric infinite-order vector autoregressive model. As a variant of the vector autoregressive moving average (ARMA) model, it not only inherits desirable properties such as parsimony and rich temporal dependence structures, but also avoids two well-known drawbacks of the former: (i) non-identifiability and (ii) computational intractability even for moderate-dimensional data. Moreover, its parameter estimation is scalable with respect to the complexity of temporal dependence, namely the number of decay patterns constituting the autoregressive structure; hence it is called the scalable ARMA (SARMA) model. In the high-dimensional setup, we further impose a low-Tucker-rank assumption on the coefficient tensor of the proposed model. The resulting model has the form of a regression with embedded dynamic factors and hence can be especially suited for financial and economic data. Non-asymptotic error bounds for the proposed estimator are derived, and a tractable alternating least squares algorithm is developed. Theoretical and computational properties of the proposed method are verified by simulation studies, and the advantages over existing methods are illustrated in real applications.

1. Introduction. The advent of big data era has spurred a rapid growth of interest in high-dimensional time series modeling in recent years. The availability of large temporal datasets enables us to better capture the dependence structure of multivariate time series along both time and variable dimensions. Applications can be found in many areas such as economics, finance, environmental science, biology, and neuroscience [17, 18, 14]. Arguably, as the most widely used time series model, the vector autoregressive (AR) model has been a

MSC2020 subject classifications: Primary 62M10; secondary 62H12, 60G10.

Keywords and phrases: high-dimensional time series, identifiability, reduced-rank regression, scalability, tensor decomposition, vector AR(∞), vector ARMA.

primary workhorse for high-dimensional tasks thanks to dimension reduction techniques such as sparsity-inducing methods (see a recent survey by [7]), parameter constraints (see [41] and the references therein), and reduced-rank methods (see, e.g., [29] and [6]). However, the flexibility of the vector AR model can be seriously limited by the finite AR order in practice, since information further in the past is often needed for fully capturing the complex temporal dependence. This will result in poorer forecasting accuracy when the time series is longer. Instead, it is more realistic to assume that the data follow a general infinite-order vector AR (i.e., vector AR(∞)) process; see the literature on multivariate time series, e.g., [26] and [34]. Then the infinite-order process can be estimated consistently by fitting a finite-order vector AR model with a large AR order that grows with the time series length. While this method may improve the forecasting performance, it can make the fitted model unnecessarily complicated and hard to interpret owing to the excessive number of AR coefficient matrices.

Alternatively, as a parsimonious parameterization of vector AR(∞) processes, the vector autoregressive moving average (ARMA) model [26, 34] is widely studied:

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\varepsilon}_{t-j}, \qquad t \in \mathbb{Z} \tag{1.1}$$

where $\boldsymbol{y}_t \in \mathbb{R}^N$ is the observed N-dimensional time series, and $\boldsymbol{\varepsilon}_t \in \mathbb{R}^N$ is the innovation term. When $q \geqslant 1$, this model can be written as a vector AR(∞) process, where the AR coefficient matrices can have a wide range of complex decay patterns even when p and q are small. Thus, information in the infinite past can be incorporated with great flexibility and parsimony, allowing for better interpretability and forecasting performance in practice; see [3] and [10] for empirical evidence. However, the popularity of the vector ARMA model has long been limited mainly due to two challenges, non-identifiability and computational intractability, both caused by the presence of the MA coefficient matrices $\boldsymbol{\Theta}_j$'s; see Section 2.2 for details. There have only been a few existing methods for large vector ARMA models in the literature, such as the EM algorithm in [28], the Bayesian method in [10], the iterative ordinary least squares algorithm in [13], and the optimization-based method in [39]. All these methods tackle the non-identifiability by imposing sophisticated identification constraints, which can render the algorithm time-consuming or inevitably multi-stage.

In this paper, we first investigate the root cause of the aforementioned challenges of the vector ARMA model. As an illustration, consider model (1.1) with orders $p = q = 1$. Once we write it in the vector AR(∞) form with AR coefficient matrices $\boldsymbol{A}_j = \boldsymbol{\Theta}^{j-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$

for $j \geqslant 1$, where $\mathbf{\Phi} = \mathbf{\Phi}_1$ and $\mathbf{\Theta} = \mathbf{\Theta}_1$, it is clear that the MA coefficient matrix $\mathbf{\Theta}$ plays a key role in the corresponding vector ARMA model. On one hand, the multiplicative factor $\mathbf{\Theta}^{j-1}$ makes $\boldsymbol{A}_j$ decay exponentially fast as $j \to \infty$, which agrees with the intuition that, for stationary time series, the dependence between $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-j}$ will diminish quickly as the time lag j increases. The decay property allows the vector ARMA model to capture information in the infinite past while maintaining stationarity. On the other hand, the existence of $\mathbf{\Theta}^{j-1}$ also directly causes the two challenges. In fact, the exponential decay pattern is merely driven by the eigenvalues of $\mathbf{\Theta}$. However, the eigenvectors of $\mathbf{\Theta}$ will interact with $\mathbf{\Phi} - \mathbf{\Theta}$, and this ultimately leads to the non-identifiability of the coefficient matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$. Moreover, the loss function for the parameter estimation will involve a very high-order polynomial in the coefficient matrix $\mathbf{\Theta}$, making the optimization intractable even for a moderately large N. The situation for general vector ARMA models is similar except that $\mathbf{\Theta}$ is generalized to the MA companion matrix; see Section 2.2 for more details. However, for brevity, we will stick to the notation $\mathbf{\Theta}$ in this section.

The above investigation motivates us to design a new parametric vector AR(∞) model, which not only inherits the main advantages of vector ARMA models but avoids the two drawbacks of the latter. Briefly speaking, without loss of much generality, we first assume that $\mathbf{\Theta}$ has distinct eigenvalues, then extract these eigenvalues to generate the temporal decay of $\boldsymbol{A}_j$'s, and finally merge all remaining ingredients of $\boldsymbol{A}_j$ into an $N \times N \times d$ coefficient tensor $\mathcal{G}$, where d is a small number. This leads to a vector AR(∞) model that avoids the problem of non-identifiability. In this model, $\boldsymbol{A}_j$'s are parameterized by $\mathcal{G}$ and a low-dimensional parameter vector $\boldsymbol{\omega}$, and hence the parsimony of the vector ARMA model is also preserved. Meanwhile, the polynomial in $\mathbf{\Theta}$ is decomposed as a weighted sum of one or two-dimensional polynomials, where each polynomial represents a particular decay pattern, and the weights are matrices controlling the relative strength of the decay patterns. As a direct consequence of this decomposable parametric structure, the estimation will be scalable with respect to the number of decay patterns, namely the complexity of temporal dependence. Thus, we call this model the scalable ARMA (SARMA) model.

Similar to the vector ARMA model, the number of parameters in the SARMA model is in the order of N^2, as is determined by the coefficient tensor $\mathcal{G}$. Therefore, when fitting the model to high-dimensional time series, it is necessary to restrict the parameter space of

$\mathcal{G}$ to a reasonable number of degrees of freedom. While following the literature on high-dimensional regression, one may consider sparsity-inducing methods such as the Lasso, this paper advocates a different approach for the following two reasons. First, unlike the independent case, the Lasso may fail to identify the true sparsity patterns of the parameters for time series data, due to the presence of strong temporal dependence (see, e.g., [30]) and the fact that the stationarity of the time series will force the entries of all $N \times N$ coefficient matrices to shrink to zero as $N \to \infty$ (see Remark 1 in [38] and Theorem 1 in Section 2.3). Second, in economic and financial time series, there is typically strong cross-sectional dependence among the N variables. Meanwhile, for such applications N and T are often both large in the high-dimensional setup, which is also known as the data-rich environment in econometrics; see, e.g., [27] and the references therein. This is different from, for example, time-course gene expression data, where usually the dependence among the N genes is believed to be sparse, and T is relatively small compared to N [25]. The strong (i.e., non-sparse) cross-sectional dependence can be better characterized by assuming that the N series are driven by a small number of common latent factors. Indeed, in the econometric literature, high-dimensional data are usually analyzed by factor models, where the estimated factors and loadings can provide insights into the interactive mechanism of large economic and financial systems [4, 24, 5]. Nonetheless, unlike AR and ARMA-type time series models, the factor model by itself cannot be directly used for forecasting.

Motivated by the above considerations, this paper employs a low-Tucker-rank approach that can not only further reduce the dimension of $\mathcal{G}$ but extract common latent factors across the N variables, thereby exploiting the advantages of factor modeling. More importantly, unlike factor models in econometrics, the proposed model can be directly used for forecasting since the factors are embedded in the vector AR(∞) framework. Specifically, let $\mathcal{A} \in \mathbb{R}^{N\times N\times\infty}$ be the tensor formed by stacking all the AR coefficient matrices $\boldsymbol{A}_j$'s. The parametric structure of the proposed vector AR(∞) model satisfies the tensor factorization, $\mathcal{A} = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega})$, where $\mathcal{G} \in \mathbb{R}^{N\times N\times d}$, and $\boldsymbol{L}$ is a known function such that $\boldsymbol{L}(\boldsymbol{\omega}) \in \mathbb{R}^{\infty\times d}$ is of full rank. This directly implies that $\mathcal{A}$ has Tucker rank d at the third mode. In addition, note that the Tucker ranks of $\mathcal{A}$ at the first and second mode correspond to the dimensions of the column and row spaces of $\boldsymbol{A}_j$'s, respectively. Similar to the multivariate reduced-rank regression [36], it is then natural to assume that $\mathcal{G}$ has low Tucker rank $\mathcal{R}_i$ at mode i for

$i = 1$ and 2, so that $\boldsymbol{\mathcal{A}}$ has Tucker ranks $(\mathcal{R}_1, \mathcal{R}_2, d)$ across three modes. As a result, the degrees of freedom of $\mathcal{G}$ can be restricted simultaneously at its first two modes, and $\boldsymbol{\mathcal{A}}$ can be further factorized as $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{L}(\boldsymbol{\omega})$, where $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{\mathcal{R}_1 \times \mathcal{R}_2 \times d}$, and $\boldsymbol{U}_i \in \mathbb{R}^{N \times \mathcal{R}_i}$ for $i = 1, 2$. Interestingly, the low-Tucker-rankness of $\boldsymbol{\mathcal{A}}$ implies different factor structures in three directions: response variables', predictor variables', and predictor time lags' directions; that is, the concept of factor modeling is embedded in the proposed model. Moreover, we can show that the $\mathcal{R}_1$-dimensional space for the response $\boldsymbol{y}_t$ recovers exactly the latent factor space of the factor model [24, 5]. However, the proposed low-Tucker-rank framework provides more flexibility in characterizing the cross-sectional dependence by allowing for a different, $\mathcal{R}_2$-dimensional space for the predictors $\boldsymbol{y}_{t-j}$'s with $j \geqslant 1$. This flexibility is essential for forecasting frameworks, since although $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-j}$'s correspond to the same N variables, the former is the output of the dynamic system, whereas the latter is the input. For a more detailed discussion on the connection with factor models, see Section 2.4.

This paper has three main contributions:

(i) A new multivariate time series model, the SARMA model, is proposed, which preserves the parsimony and exponentially decaying temporal dependence patterns of the vector ARMA model, but avoids the challenges in identification and computation. A sufficient condition for the stationarity of the model is established.

(ii) For high-dimensional modeling and forecasting, a low-Tucker-rank estimation method is introduced to embed factor-based dimensionality reduction in the proposed model, and the corresponding algorithm is scalable with respect to the complexity of temporal dependence. Additionally, a procedure for consistent model selection is developed.

(iii) The nonasymptotic analysis of the proposed estimator involves new techniques for handling nonlinear and nonconvex parametric structure for high-dimensional time series models, and may be of independent interest.

The remainder of this paper is organized as follows. In Section 2, a reparameterization of the vector ARMA model is first derived, which leads to the proposed model, and then the high-dimensional low-Tucker-rank model is introduced. Section 3 presents the proposed estimator, algorithm, and model selection procedure, together with theoretical properties. Simulation studies and empirical examples are provided in Sections 4 and 5, respectively. Section 6 concludes with a discussion on some interesting extensions. We collect the most

important theoretical analysis in the Appendix, while all remaining proofs and some further details for Sections 5 and 6 are given in a separate supplementary file.

2. Scalable autoregressive moving average models .

2.1. *Tensor algebra and notations.* Tensors, a.k.a. multidimensional arrays, are natural high-order extensions of matrices; see [23] for a review of basic tensor algebra. The order of a tensor is known as the dimension, way or mode. This paper will focus on the third-order tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_1 \times d_3}$. Its mode-1 matricization $\mathcal{X}_{(1)}$ is the d_1-by-$(d_2 d_3)$ matrix whose $\{i, (k-1)d_2 + j\}$-th entry is $\mathcal{X}_{ijk}$ for $1 \leqslant i \leqslant d_1, 1 \leqslant j \leqslant d_2$ and $1 \leqslant k \leqslant d_3$. The mode-2 and mode-3 matricizations can be defined similarly. For any sequence of $d_1 \times d_2$ matrices $\{\boldsymbol{X}_i, 1 \leqslant i \leqslant d_3\}$, where d_3 may be infinity, we define the stacking operation $\mathrm{stack}(\cdot)$ such that $\mathcal{X} = \mathrm{stack}(\{\boldsymbol{X}_i\})$ is the $d_1 \times d_2 \times d_3$ tensor obtained by stacking all the matrices along the third mode. Note that $\boldsymbol{X}_i$'s are called the frontal slices of $\mathcal{X}$, and $\mathcal{X}_{(1)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d_3})$. Furthermore, the mode-1 multiplication of a tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ by a matrix $\boldsymbol{Y} \in \mathbb{R}^{q_1 \times d_1}$ leads to a $q_1 \times d_2 \times d_3$ tensor, $\mathcal{X} \times_1 \boldsymbol{Y} = (\sum_{i=1}^{d_1} \mathcal{X}_{ijk} \boldsymbol{Y}_{si})_{1 \leqslant s \leqslant q_1, 1 \leqslant j \leqslant d_2, 1 \leqslant k \leqslant d_3}$; the mode-2 and mode-3 multiplications can be defined similarly. The Tucker rank at mode i of $\mathcal{X}$ is defined as the rank of $\mathcal{X}_{(i)}$ for $1 \leqslant i \leqslant 3$ [35, 12]. Unlike row and column ranks of a matrix, r_1, r_2 and r_3 in general are not identical. Suppose that $\mathcal{X}$ has Tucker ranks $r_i = \mathrm{rank}(\mathcal{X}_{(i)})$ for $1 \leqslant i \leqslant 3$. Then there exists a Tucker decomposition, $\mathcal{X} = \mathcal{Y} \times_1 \boldsymbol{Y}_1 \times_2 \boldsymbol{Y}_2 \times_3 \boldsymbol{Y}_3$, where $\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, and $\boldsymbol{Y}_i \in \mathbb{R}^{d_i \times r_i}$ with $1 \leqslant i \leqslant 3$ are factor matrices.

Throughout the paper, unless otherwise specified, we will denote scalars by lowercase letters $x, y, \ldots$, vectors by boldface lowercase letters $\boldsymbol{x}, \boldsymbol{y}, \ldots$, matrices by boldface capital letters $\boldsymbol{X}, \boldsymbol{Y}, \ldots$, and tensors by calligraphic capital letters $\mathcal{X}, \mathcal{Y}, \ldots$. For any $a, b \in \mathbb{R}$, denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any vector $\boldsymbol{x}$, denote its ℓ_2 norm by $\|\boldsymbol{x}\|_2$. For any matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, let $\sigma_1(\boldsymbol{X}) \geqslant \sigma_2(\boldsymbol{X}) \geqslant \cdots \geqslant \sigma_{d_1 \wedge d_2}(\boldsymbol{X}) \geqslant 0$ be its singular values in descending order. Let $\boldsymbol{X}'$, $\sigma_{\max}(\boldsymbol{X})$ (or $\sigma_{\min}(\boldsymbol{X})$), $\lambda_{\max}(\boldsymbol{X})$ (or $\lambda_{\min}(\boldsymbol{X})$), $\mathrm{colsp}(\boldsymbol{X})$, and $\mathrm{rank}(\boldsymbol{X})$ denote its transpose, largest (or smallest) singular value, largest (or smallest) eigenvalue, column space, and rank, respectively. Its vectorization $\mathrm{vec}(\boldsymbol{X})$ is the long vector obtained by stacking all its columns. In addition, its operator norm, Frobenius norm, and nuclear norm are $\|\boldsymbol{X}\|_{\mathrm{op}} = \sigma_{\max}(\boldsymbol{X})$, $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\sum_{i,j} \boldsymbol{X}_{ij}^2} = \sqrt{\sum_{k=1}^{d_1 \wedge d_2} \sigma_k^2(\boldsymbol{X})}$, and $\|\boldsymbol{X}\|_* = \sum_{k=1}^{d_1 \wedge d_2} \sigma_k(\boldsymbol{X})$, respectively. The Frobenius norm of a tensor $\mathcal{X}$ is $\|\mathcal{X}\|_{\mathrm{F}} = \sqrt{\sum_{i,j,k} \mathcal{X}_{ijk}^2}$.

For any real-valued function f, let $\nabla f(\boldsymbol{x}) \in \mathbb{R}^m$ and $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{m\times m}$ be its first- and second-order derivatives with respect to a scalar or vector $\boldsymbol{x} \in \mathbb{R}^m$, respectively. For any scalar-, vector-, or matrix-valued function $\boldsymbol{f}(\boldsymbol{\eta})$ with $\boldsymbol{\eta} = (\gamma, \theta)' \in \mathbb{R}^2$, let $\nabla_\gamma \boldsymbol{f}(\boldsymbol{\eta}) = \partial \boldsymbol{f}(\boldsymbol{\eta})/\partial\gamma$ and $\nabla_\theta \boldsymbol{f}(\boldsymbol{\eta}) = \partial \boldsymbol{f}(\boldsymbol{\eta})/\partial\theta$ be its first-order partial derivatives, and they both have the same size as $\boldsymbol{f}(\boldsymbol{\eta})$. For any two sequences x_n and y_n, denote $x_n \lesssim y_n$ (or $x_n \gtrsim y_n$) if there exists an absolute constant $C > 0$ such that $x_n \leqslant C y_n$ (or $x_n \geqslant C y_n$). Write $x_n \asymp y_n$ if $x_n \lesssim y_n$ and $x_n \gtrsim y_n$. Let $\mathbb{I}_{\{\cdot\}}$ be the indicator function taking value one when the condition is true and zero otherwise. The capital letters $C, C_{\mathfrak{g}}, \ldots$ and lowercase letters $c, c_{\mathfrak{g}}, \ldots$ represent generic large and small positive absolute constants, respectively, whose values may vary from place to place.

2.2. *Reparameterization of vector ARMA models.* Consider a simple stationary vector ARMA$(1,1)$ model, $\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1}$, for an N-dimensional time series $\{\boldsymbol{y}_t\}_{t=1}^T$. We assume that it is invertible, i.e., all eigenvalues of $\boldsymbol{\Theta}$ are less than one in absolute value, and then it can be written into the following vector AR(∞) form,

$$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \underbrace{\boldsymbol{\Theta}^{j-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})}_{\boldsymbol{A}_j} \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\boldsymbol{A}_j = \boldsymbol{A}_j(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \boldsymbol{\Theta}^{j-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$ with $j \geqslant 1$ are specified in matrix multiplicative forms. Note that the common factor $\boldsymbol{\Phi} - \boldsymbol{\Theta}$ in $\boldsymbol{A}_j$'s can be regarded as a shifted version of the AR coefficient matrix $\boldsymbol{\Phi}$ in the vector ARMA model. It is well known that when fitting the vector ARMA model to real-world datasets, two challenges will be encountered:

(i) *Non-identifiability:* There generally exist many different combinations of $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ that lead to the same values for $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots\}$ and hence the same data generating process.

(ii) *Computational intractability:* When estimating $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ in a sample of size T, for instance, by the least squares method, due to the form in (2.1), the loss function will be a matrix polynomial in $\boldsymbol{\Theta}$ of degree as high as $2(T-1)$. This will make the optimization intractable even for moderate dimension N; see Section 4.2 for numerical evidence.

To overcome these challenges, we aim to block-diagonalize the MA coefficient matrix $\boldsymbol{\Theta}$ such that the high-order term $\boldsymbol{\Theta}^{j-1}$ in (2.1) can be more easily computed.

LEMMA 2.1 (Theorem 1 in [20]). *Real matrices with $\mathcal{R}$ distinct nonzero eigenvalues are dense in the set of all $N \times N$ real matrices with rank at most $\mathcal{R}$, where $0 < \mathcal{R} \leqslant N$.*

By Lemma 2.1, without loss of much generality, we can assume that all of the $\mathcal{R}$ nonzero eigenvalues of $\boldsymbol{\Theta}$ are distinct, where $\mathcal{R} = r + 2s \leqslant N$ is the rank of $\boldsymbol{\Theta}$. Specifically, suppose that $\boldsymbol{\Theta}$ has r distinct nonzero real eigenvalues, $\lambda_1, \ldots, \lambda_r$, and s distinct conjugate pairs of nonzero complex eigenvalues, $\lambda_{r+1}, \ldots, \lambda_{r+2s}$, where $|\lambda_j| \in (0,1)$ for $1 \leqslant j \leqslant r$, $(\lambda_{r+2k-1}, \lambda_{r+2k}) = (\gamma_k e^{i\theta_k}, \gamma_k e^{-i\theta_k})$ with $\gamma_k \in (0,1)$ and $\theta_k \in (-\pi/2, \pi/2)$ for $1 \leqslant k \leqslant s$, and i represents the imaginary unit. Then, the Jordan decomposition $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$ exists such that $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ is invertible and $\boldsymbol{J} \in \mathbb{R}^{N\times N}$ is the real Jordan form; that is, $\boldsymbol{J} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_r, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_s, \boldsymbol{0}\}$ is a real block diagonal matrix, with

$$\boldsymbol{C}_k = \gamma_k \cdot \begin{pmatrix} \cos(\theta_k) & \sin(\theta_k) \\ -\sin(\theta_k) & \cos(\theta_k) \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad 1 \leqslant k \leqslant s; \tag{2.2}$$

see [22, Chap. 3]. We make use of the block-diagonal structure of $\boldsymbol{J}$ to reparameterize the coefficient matrices of the vector AR(∞) form in (2.1) into

$$\begin{aligned} \boldsymbol{A}_1 &= \boldsymbol{\Phi} - \boldsymbol{\Theta} := \boldsymbol{G}_1 \quad \text{and} \\ \boldsymbol{A}_{1+j} &= \boldsymbol{B}\boldsymbol{J}^j\boldsymbol{B}^{-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta}) \\ &:= \sum_{k=1}^{r} \lambda_k^j \boldsymbol{G}_k^I + \sum_{k=1}^{s} \gamma_k^j \left\{ \cos(j\theta_k)\boldsymbol{G}_k^{II,1} + \sin(j\theta_k)\boldsymbol{G}_k^{II,2} \right\}, \quad j \geqslant 1, \end{aligned} \tag{2.3}$$

where $\boldsymbol{G}_k^I$'s, $\boldsymbol{G}_k^{II,1}$'s and $\boldsymbol{G}_k^{II,2}$'s are $N \times N$ real matrices determined jointly by $\boldsymbol{B}$ and $\boldsymbol{B}^{-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta})$ such that

$$\mathrm{rank}(\boldsymbol{G}_j^I) \leqslant 1, \ \ \mathrm{rank}(\boldsymbol{G}_k^{II,1}) \leqslant 2, \ \ \mathrm{rank}(\boldsymbol{G}_k^{II,2}) \leqslant 2, \ \ \forall 1 \leqslant j \leqslant r, 1 \leqslant k \leqslant s, \tag{2.4}$$

and, for each $1 \leqslant k \leqslant s$, $\boldsymbol{G}_k^{II,1}$ and $\boldsymbol{G}_k^{II,2}$ have the same row and column spaces; see the proof of Proposition 2.2 for details. We can view (2.3) as a new parametric vector AR(∞) model, whose unknown parameters are $\boldsymbol{G}_1$, $\boldsymbol{G}_k^I$'s, $\boldsymbol{G}_k^{II,1}$'s, $\boldsymbol{G}_k^{II,2}$'s, λ_k's, γ_k's and θ_k's.

The above reparameterization circumvents the challenges of the vector ARMA model mentioned in (i) and (ii). To better understand how this is possible, consider the MA coefficient matrix $\boldsymbol{\Theta}$. Since $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$, it plays two different roles through $\boldsymbol{J}$ and $\boldsymbol{B}$: while $\boldsymbol{J}$ induces the temporal decay of $\boldsymbol{A}_j$ via the eigenvalues, $\boldsymbol{B}$ is essentially a redundant factor irrelevant to the temporal dynamic. However, $\boldsymbol{B}$ will interact with the shifted AR coefficient matrix $\boldsymbol{\Phi} - \boldsymbol{\Theta}$ through the matrix multiplicative form of $\boldsymbol{A}_j$'s, leading to the identification problem of the vector ARMA model. To avoid this pitfall, we first extract the ingredients of $\boldsymbol{\Theta}$ that generate the temporal decay of $\boldsymbol{A}_{1+j}$, i.e., the scalars λ_k^j's, γ_k^j's and $j\theta_k$'s. Then we

merge all other ingredients of $\boldsymbol{A}_{1+j}$ into the $\boldsymbol{G}$-matrices. Consequently, $\boldsymbol{A}_{1+j}$'s in (2.3) are simply linear combinations of the $\boldsymbol{G}$-matrices. By getting rid of the matrix multiplications, the identification problem will be avoided. As a by-product, since there is no longer any high-order matrix polynomial involved, the computational challenge is also overcome. The cost of these desirable properties is the addition of a small number of unknown parameters: the vector ARMA$(1,1)$ model without any constraint on $\boldsymbol{\Phi}$ has $N^2+2N(r+2s)-(r+2s)^2$ parameters, while model (2.3) has $N^2+(2N-1)(r+2s)$ parameters with the low-rank properties of the $\boldsymbol{G}$-matrices in (2.4) taken into account.

In general, for a stationary and invertible vector ARMA(p,q) model in the form of (1.1), the vector AR(∞) representation can be written as

$$\boldsymbol{y}_t=\sum_{j=1}^{\infty}\underbrace{\left(\sum_{i=0}^{p\wedge j}\boldsymbol{P}\underline{\boldsymbol{\Theta}}^{j-i}\boldsymbol{P}'\boldsymbol{\Phi}_i\right)}_{\boldsymbol{A}_j}\boldsymbol{y}_{t-j}+\boldsymbol{\varepsilon}_t,\qquad \underline{\boldsymbol{\Theta}}=\begin{pmatrix}\boldsymbol{\Theta}_1 & \boldsymbol{\Theta}_2 & \cdots & \boldsymbol{\Theta}_{q-1} & \boldsymbol{\Theta}_q\\ \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I} & \boldsymbol{0}\end{pmatrix}, \tag{2.5}$$

where $\boldsymbol{\Phi}_0=-\boldsymbol{I}$ and $\boldsymbol{P}=(\boldsymbol{I}_N,\boldsymbol{0}_{N\times N(q-1)})$ are known matrices, and $\underline{\boldsymbol{\Theta}}$ is called the MA companion matrix [26]. Similarly, to block-diagonalize $\underline{\boldsymbol{\Theta}}$, we consider its Jordan decomposition: $\underline{\boldsymbol{\Theta}}=\boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$, where $\boldsymbol{B}\in\mathbb{R}^{Nq\times Nq}$ is invertible, and $\boldsymbol{J}$ is the real Jordan form of $\underline{\boldsymbol{\Theta}}$. Let $\widetilde{\boldsymbol{B}}=\boldsymbol{P}\boldsymbol{B}$ and $\breve{\boldsymbol{B}}=\boldsymbol{B}^{-1}\left(\sum_{i=0}^{p}\underline{\boldsymbol{\Theta}}^{p-i}\boldsymbol{P}'\boldsymbol{\Phi}_i\right)$. The following proposition gives the general reparameterization for (2.5) without requiring that all nonzero eigenvalues of $\underline{\boldsymbol{\Theta}}$ are distinct.

PROPOSITION 2.2. Suppose that there are r distinct nonzero real eigenvalues of $\underline{\boldsymbol{\Theta}}$, λ_j for $1\leqslant j\leqslant r$, and s distinct conjugate pairs of nonzero complex eigenvalues of $\underline{\boldsymbol{\Theta}}$, $(\lambda_{r+2k-1},\lambda_{r+2k})=(\gamma_k e^{i\theta_k},\gamma_k e^{-i\theta_k})$ with $\gamma_k\in(0,1)$ and $\theta_k\in(-\pi/2,\pi/2)$ for $1\leqslant k\leqslant s$. Moreover, the algebraic multiplicity of λ_j is n_j for $1\leqslant j\leqslant r$, and that of $(\lambda_{r+2k-1},\lambda_{r+2k})$ is m_k for $1\leqslant k\leqslant s$, so there are $R+2S$ nonzero eigenvalues of $\underline{\boldsymbol{\Theta}}$ in total, where $R=\sum_{k=1}^{r}n_k$ and $S=\sum_{k=1}^{s}m_k$. Assume that the geometric multiplicities of all nonzero eigenvalues are

one. Then for all $j \geqslant 1$, we have

$$\begin{aligned}\boldsymbol{A}_j = &\sum_{k=1}^{p} \mathbb{I}_{\{j=k\}} \boldsymbol{G}_k + \sum_{k=1}^{r} \sum_{i=1}^{n_k} \mathbb{I}_{\{j \geqslant p+(i-1)\vee 1\}} \lambda_k^{j-p-i+1} \binom{j-p}{i-1} \boldsymbol{G}_{k,i}^{I} \\ &+ \sum_{k=1}^{s} \sum_{i=1}^{m_k} \mathbb{I}_{\{j \geqslant p+(i-1)\vee 1\}} \gamma_k^{j-p-i+1} \binom{j-p}{i-1} \\ &\qquad \cdot \Big[\cos\{(j-p-i+1)\theta_k\} \boldsymbol{G}_{k,i}^{II,1} + \sin\{(j-p-i+1)\theta_k\} \boldsymbol{G}_{k,i}^{II,2}\Big], \end{aligned} \tag{2.6}$$

where the first term is suppressed if $p = 0$, and $\boldsymbol{G}_{k,i}^{I}$'s, $\boldsymbol{G}_{k,i}^{II,1}$'s and $\boldsymbol{G}_{k,i}^{II,2}$'s are all determined jointly by $\widetilde{\boldsymbol{B}}$ and $\breve{\boldsymbol{B}}$. Moreover, for any fixed k and i, $\boldsymbol{G}_{k,i}^{II,h}$ for $h = 1, 2$ have the same row and column spaces, and $\operatorname{rank}(\boldsymbol{G}_{j,l}^{I}) \leqslant n_k$ and $\operatorname{rank}(\boldsymbol{G}_{k,i}^{II,h}) \leqslant 2m_k$ for all $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant s$, $1 \leqslant l \leqslant n_k$, $1 \leqslant i \leqslant m_k$, and $h = 1, 2$.

By Proposition 2.2, if all nonzero eigenvalues of $\underline{\boldsymbol{\Theta}}$ are further assumed to be distinct, i.e., $n_1 = \cdots = n_r = m_1 = \cdots = m_s = 1$, as its generality is supported by Lemma 2.1, then (2.6) for the vector AR(∞) form in (2.5) can be reparameterized into

$$\begin{aligned}\boldsymbol{A}_j = &\sum_{k=1}^{p} \mathbb{I}_{\{j=k\}} \boldsymbol{G}_k + \sum_{k=1}^{r} \mathbb{I}_{\{j \geqslant p+1\}} \lambda_k^{j-p} \boldsymbol{G}_k^{I} \\ &+ \sum_{k=1}^{s} \mathbb{I}_{\{j \geqslant p+1\}} \gamma_k^{j-p} \Big[\cos\{(j-p)\theta_k\} \boldsymbol{G}_k^{II,1} + \sin\{(j-p)\theta_k\} \boldsymbol{G}_k^{II,2}\Big] \end{aligned} \tag{2.7}$$

for all $j \geqslant 1$, where the first term is suppressed if $p = 0$.

2.3. *Proposed SARMA model.* We first use tensor notations to obtain a more succinct form of (2.7). Let $d = p + r + 2s$ and $\boldsymbol{\omega} = (\lambda_1, \ldots, \lambda_r, \boldsymbol{\eta}_1', \ldots \boldsymbol{\eta}_s')'$, where $\boldsymbol{\eta}_k = (\gamma_k, \theta_k)'$ for $1 \leqslant k \leqslant s$. Define a matrix-valued function of $\boldsymbol{\omega}$ as follows:

$$\boldsymbol{L}(\boldsymbol{\omega}) = (\ell_{j,k}(\boldsymbol{\omega}))_{j \geqslant 1, 1 \leqslant k \leqslant d} = \begin{pmatrix} \boldsymbol{I}_p & \\ & \boldsymbol{\ell}^I(\lambda_1) \cdots \boldsymbol{\ell}^I(\lambda_r)\, \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_1) \cdots \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_s) \end{pmatrix} \in \mathbb{R}^{\infty \times d},$$

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

$$\boldsymbol{\ell}^I(\lambda) = (\lambda, \lambda^2, \lambda^3, \ldots)' \quad \text{and} \quad \boldsymbol{\ell}^{II}(\boldsymbol{\eta}) = \begin{pmatrix} \gamma\cos(\theta) & \gamma^2\cos(2\theta) & \gamma^3\cos(3\theta) & \cdots \\ \gamma\sin(\theta) & \gamma^2\sin(2\theta) & \gamma^3\sin(3\theta) & \cdots \end{pmatrix}'.$$

For simplicity, we relabel $\boldsymbol{G}_{p+j} = \boldsymbol{G}_j^{I}$ for $1 \leqslant j \leqslant r$, and $\boldsymbol{G}_{p+r+2k-1} = \boldsymbol{G}_k^{II,1}$ and $\boldsymbol{G}_{p+r+2k} = \boldsymbol{G}_k^{II,2}$ for $1 \leqslant k \leqslant s$; the two sets of notations may be used interchangeably when there is no confusion. By stacking all $\boldsymbol{G}_k$'s into the tensor $\boldsymbol{\mathcal{G}} = \operatorname{stack}(\{\boldsymbol{G}_j, 1 \leqslant j \leqslant$

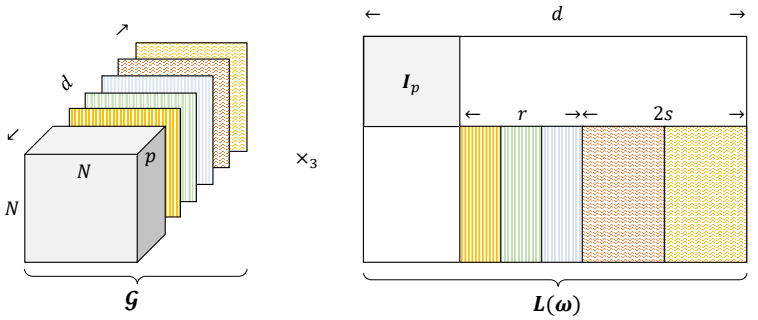

FIGURE 1. *Reparameterization of $\mathcal{A}$ for vector ARMA(p,q) models when all nonzero eigenvalues of $\underline{\Theta}$ are distinct, $r=3$ and $s=2$. In $\boldsymbol{L}(\boldsymbol{\omega})$, rectangles filled with vertical stripes are blocks in Type I parametric form $\boldsymbol{\ell}^I(\cdot)$, those filled with waves are blocks in Type II parametric form $\boldsymbol{\ell}^{II}(\cdot)$, different colors indicate dependence on different subsets of parameters, and empty spaces represent zeros.*

$d\}) \in \mathbb{R}^{N\times N\times d}$ and all $\boldsymbol{A}_j$'s into the tensor $\boldsymbol{\mathcal{A}} = \text{stack}(\{\boldsymbol{A}_j, j \geqslant 1\}) \in \mathbb{R}^{N\times N\times\infty}$, (2.7) can be written in the tensor form:

$$\boldsymbol{\mathcal{A}} := \text{stack}(\{\boldsymbol{A}_j, j \geqslant 1\}) = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega}); \tag{2.8}$$

see Figure 1 for an illustration. In matrix form, this is equivalent to $\boldsymbol{A}_j = \sum_{k=1}^{d} \ell_{j,k}(\boldsymbol{\omega})\boldsymbol{G}_k$ for $j \geqslant 1$, which is exactly (2.7).

Based on the above notation, this paper introduces a new parametric vector AR(∞) model:

$$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \boldsymbol{A}_j(\boldsymbol{\omega}, \mathcal{G})\boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t, \quad \text{or} \quad \boldsymbol{y}_t = \sum_{k=1}^{d} \boldsymbol{G}_k \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega})\boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{2.9}$$

where $\boldsymbol{\varepsilon}_t \in \mathbb{R}^N$ are the innovations, the coefficient tensor $\boldsymbol{\mathcal{A}}(\boldsymbol{\omega}, \mathcal{G}) = \text{stack}(\{\boldsymbol{A}_j(\boldsymbol{\omega}, \mathcal{G}), j \geqslant 1\})$ satisfies (2.8) (i.e., (2.7)), $\boldsymbol{\eta}_k \in \boldsymbol{\Pi} = [0,1) \times (-\pi/2, \pi/2)$ for $1 \leqslant k \leqslant s$, and $\boldsymbol{\omega} \in (-1,1)^r \times \boldsymbol{\Pi}^s \subset \mathbb{R}^{r+2s}$. Model (2.9) does not impose any low-rank constraints on $\mathcal{G}$, and it contains $N^2d + r + 2s$ parameters in total. Note that in the matrix $\boldsymbol{L}(\boldsymbol{\omega})$, the diagonal block $\boldsymbol{I}_p$ implies $\boldsymbol{A}_j = \boldsymbol{G}_j$ for $1 \leqslant j \leqslant p$, while $\boldsymbol{\ell}^I(\cdot)$ and $\boldsymbol{\ell}^{II}(\cdot)$ introduce two types of decay patterns for $\boldsymbol{A}_j$'s with $j \geqslant p+1$: the exponential decay, or a pair of exponentially damped cosine and sine waves. As a result, the magnitude of $\ell_{j,k}(\boldsymbol{\omega})$ will decay exponentially fast as the row index $j \to \infty$ for all k, and hence so will $\boldsymbol{A}_j$.

To better understand the temporal structure induced by the two types of decay patterns, we consider two special vector ARMA models whose MA coefficient matrices are scaled identity matrices: the vector ARMA$(1,1)$ with $\boldsymbol{\Phi} = \boldsymbol{G}^I + \lambda\boldsymbol{I}_N$ and $\boldsymbol{\Theta} = \lambda\boldsymbol{I}_N$, and the vector ARMA$(2,2)$ with $\boldsymbol{\Phi}_1 = \boldsymbol{G}^{II,1} + 2\gamma\cos\theta\boldsymbol{I}_N$, $\boldsymbol{\Phi}_2 = \gamma\sin\theta\boldsymbol{G}^{II,2} - \gamma\cos\theta\boldsymbol{G}^{II,1} - \gamma^2\boldsymbol{I}_N$, $\boldsymbol{\Theta}_1 = 2\gamma\cos\theta\boldsymbol{I}_N$, and $\boldsymbol{\Theta}_2 = -\gamma^2\boldsymbol{I}_N$. Interestingly, their vector AR(∞) representations are

$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \lambda^{j-1} \boldsymbol{G}^I \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t$ (Type I model) and

$$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \gamma^{j-1} \left[\cos\{(j-1)\theta\}\boldsymbol{G}^{II,1} + \sin\{(j-1)\theta\}\boldsymbol{G}^{II,2}\right] \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t \quad \text{(Type II model)},$$

which resemble the two types of decay patterns. As a result, in light of (2.7), model (2.9) in the case of $p = 1$ can be viewed as the superimposition of r Type I AR(∞) models, s Type II AR(∞) models, and a vector AR(1) model, $\boldsymbol{y}_t = (\boldsymbol{G}_1 - \sum_{k=1}^{r} \boldsymbol{G}_k^I - \sum_{k=1}^{s} \boldsymbol{G}_k^{II,1})\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t$; note that the subtracted terms are because j starts with 1 rather than $p+1$ in the above vector AR(∞) models. The richness of the temporal dynamic in the proposed model is thus mainly controlled by r and s, and similar results hold for the model with a general order of p.

Another important finding related to the above discussion is that the proposed model in general admits an additive decomposition in the form of

$$\boldsymbol{y}_t = \sum_{k=1}^{p} \boldsymbol{G}_k \boldsymbol{y}_{t-k} + \sum_{k=1}^{r} f^I(\boldsymbol{x}_t; \lambda_k) + \sum_{k=1}^{s} f^{II}(\boldsymbol{x}_t; \boldsymbol{\eta}_k) + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \boldsymbol{y}'_{t-2}, \ldots)'$, $f^I(\boldsymbol{x}_t; \lambda_k) = \sum_{j=p+1}^{\infty} \lambda^{j-p} \boldsymbol{G}_{p+k} \boldsymbol{y}_{t-j}$ and $f^{II}(\boldsymbol{x}_t; \boldsymbol{\eta}_k) = \sum_{j=p+1}^{\infty} \gamma^{j-p} [\cos\{(j-p)\theta\}\boldsymbol{G}_{p+r+2k-1} + \sin\{(j-p)\theta\}\boldsymbol{G}_{p+r+2k}]\boldsymbol{y}_{t-j}$. Here the dependence on $\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d$ is suppressed for succinctness. It is noteworthy that the parameters pertaining to different decay patterns are separable; that is, each summand $f^I(\cdot)$ or $f^{II}(\cdot)$ involves only a particular λ_k or $\boldsymbol{\eta}_k$. As a result, the parameter estimation can be implemented efficiently via an alternating minimization algorithm, where we iterate through all λ_k's and $\boldsymbol{\eta}_k$'s, updating one at a time while keep other terms constant. Since λ_k's and $\boldsymbol{\eta}_k$'s are only one- or two-dimensional, the computational cost associated with any increase in r or s will be low. In other words, the estimation will be scalable with respect to the complexity of temporal dependence; see the algorithm in Section 3.2 and numerical evidence in Section 4.2. For this reason, we call model (2.9) the scalable ARMA model of orders (p, r, s), or the SARMA(p, r, s) model. The following theorem provides a sufficient condition for the existence of a stationary causal solution of this model.

THEOREM 2.3. *Suppose that $\{\boldsymbol{\varepsilon}_t\}$ is a strictly stationary sequence with $E(\|\boldsymbol{\varepsilon}_t\|_2) < \infty$. If there exists $0 < \rho < 1$ such that*

$$\max\{|\lambda_1|, \ldots, |\lambda_r|, \gamma_1, \ldots, \gamma_s\} \leqslant \rho \quad \textit{and} \quad \sum_{k=1}^{p} \|\boldsymbol{G}_k\|_{\mathrm{op}} + \frac{\rho}{1-\rho} \sum_{k=p+1}^{d} \|\boldsymbol{G}_k\|_{\mathrm{op}} < 1, \tag{2.10}$$

then there exists a unique strictly stationary solution to model (2.9)*, and it has the form of* $\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j}$*, where* $\boldsymbol{\Psi}_j = \sum_{k=1}^{\infty} \sum_{j_1+\cdots+j_k=j} \boldsymbol{A}_{j_1} \cdots \boldsymbol{A}_{j_k}$ *and* $\boldsymbol{A}_j = \sum_{k=1}^{d} \ell_{j,k}(\boldsymbol{\omega}) \boldsymbol{G}_k$ *for all* $j \geqslant 1$.

In addition, $\mathcal{G}$ can be divided into two subtensors $\mathcal{G}^{\mathrm{AR}} := \mathrm{stack}(\{\boldsymbol{G}_k, 1 \leqslant k \leqslant p\}) \in \mathbb{R}^{N \times N \times p}$ and $\mathcal{G}^{\mathrm{MA}} := \mathrm{stack}(\{\boldsymbol{G}_k, p+1 \leqslant k \leqslant d\}) \in \mathbb{R}^{N \times N \times (r+2s)}$, corresponding to the AR and MA parts of the model, respectively. Then we can write (2.8) in the partitioned form:

$$\mathcal{A} = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega}) = \mathrm{stack}\left(\mathcal{G}^{\mathrm{AR}}, \mathcal{G}^{\mathrm{MA}}\right) \times_3 \begin{pmatrix} \boldsymbol{I}_p & \\ & \boldsymbol{L}^{\mathrm{MA}}(\boldsymbol{\omega}) \end{pmatrix}, \tag{2.11}$$

where $\boldsymbol{L}^{\mathrm{MA}}(\boldsymbol{\omega}) = \left(\boldsymbol{\ell}^{I}(\lambda_1), \ldots, \boldsymbol{\ell}^{I}(\lambda_r), \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_1), \ldots, \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_s)\right)$. In Figure 1, the gray box represents $\mathcal{G}^{\mathrm{AR}}$, the colored slices behind it represent $\mathcal{G}^{\mathrm{MA}}$, and the colored rectangles in $\boldsymbol{L}(\boldsymbol{\omega})$ constitute $\boldsymbol{L}^{\mathrm{MA}}(\boldsymbol{\omega})$.

EXAMPLE 1 (SMA model). When $p = 0$, (2.9) reduces to the SMA(r, s) model. The AR parts in $\mathcal{G}$ and $\boldsymbol{L}(\boldsymbol{\omega})$ no longer exist, and (2.11) reduces to $\mathcal{A} = \mathcal{G}^{\mathrm{MA}} \times_3 \boldsymbol{L}^{\mathrm{MA}}(\boldsymbol{\omega})$. This model can be regarded as an analogue of the vector MA model.

EXAMPLE 2 (SAR model). When $r = s = 0$, (2.9) reduces to $\boldsymbol{y}_t = \sum_{j=1}^{p} \boldsymbol{G}_j \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t$ with $\mathcal{A} = \mathcal{G}^{\mathrm{AR}}$, and the MA parts are absent in (2.11). Hence, the SAR(p) model is exactly the vector AR(p) model.

2.4. *Low-Tucker-rank model for high-dimensional time series.* Similar to vector ARMA models, the proposed SARMA model in (2.9) contains $O(N^2)$ unknown parameters and thus requires further dimension reduction when N is large. As shown in Section 2.2, the frontal slices of $\mathcal{G}^{\mathrm{MA}}$ are low-rank matrices if the data follow a vector ARMA process. In light of Example 2 and the literature on high-dimensional low-rank vector AR models [36, 11, 6], it is natural to assume that the frontal slices of $\mathcal{G}^{\mathrm{AR}}$ are also of low rank when N is large. However, this will result in many matrix constraints, making the model hard to interpret.

For better interpretability, we consider the low-Tucker-rank assumption on $\mathcal{G} \in \mathbb{R}^{N \times N \times d}$. This will automatically induce low-rankness uniformly for all its frontal slices. Note that the order d is usually small. Thus, it is only necessary to assume that $\mathcal{G}$ has low Tucker ranks at

its first two modes. We propose the following low-Tucker-rank SARMA model:

$$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \boldsymbol{A}_j(\boldsymbol{\omega}, \mathcal{G})\boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t, \quad \text{with} \quad \operatorname{rank}(\mathcal{G}_{(i)}) \leqslant \mathcal{R}_i, \quad i = 1, 2, \tag{2.12}$$

for some predetermined Tucker ranks $0 < \mathcal{R}_1, \mathcal{R}_2 \leqslant N$. As a result, $\operatorname{rank}(\boldsymbol{G}_k) \leqslant \mathcal{R}_1 \wedge \mathcal{R}_2$ for all $1 \leqslant k \leqslant d$. Moreover, there exists a Tucker decomposition, $\mathcal{G} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2$, where $\mathcal{S} \in \mathbb{R}^{\mathcal{R}_1 \times \mathcal{R}_2 \times d}$ is the core tensor, and $\boldsymbol{U}_i \in \mathbb{R}^{N \times \mathcal{R}_i}$ with $i = 1$ and 2 are factor matrices. Let $\boldsymbol{S}_k \in \mathbb{R}^{\mathcal{R}_1 \times \mathcal{R}_2}$ be the kth frontal slice of $\mathcal{S}$ for $1 \leqslant k \leqslant d$, and then it holds

$$\mathcal{A}(\boldsymbol{\omega}, \mathcal{G}) = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{L}(\boldsymbol{\omega}), \text{ or } \boldsymbol{A}_j(\boldsymbol{\omega}, \mathcal{G}) = \sum_{k=1}^{d} \ell_{j,k}(\boldsymbol{\omega}) \boldsymbol{U}_1 \boldsymbol{S}_k \boldsymbol{U}_2', \ j \geqslant 1. \tag{2.13}$$

Note that the Tucker decomposition is not unique, since $\mathcal{G} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 = (\mathcal{S} \times_1 \boldsymbol{O}_1 \times_2 \boldsymbol{O}_2) \times_1 (\boldsymbol{U}_1 \boldsymbol{O}_1^{-1}) \times_2 (\boldsymbol{U}_2 \boldsymbol{O}_2^{-1})$ for any invertible matrices $\boldsymbol{O}_i \in \mathbb{R}^{\mathcal{R}_i \times \mathcal{R}_i}$ with $i = 1$ and 2. We can consider the higher-order singular value decomposition (HOSVD) of $\mathcal{G}$, namely the special Tucker decomposition uniquely defined by choosing $\boldsymbol{U}_i$ as the tall matrix consisting of the top $\mathcal{R}_i$ left singular vectors of $\mathcal{G}_{(i)}$ and then setting $\mathcal{S} = \mathcal{G} \times_1 \boldsymbol{U}_1' \times_2 \boldsymbol{U}_2'$. Note that $\boldsymbol{U}_1$ is orthonormal, i.e., $\boldsymbol{U}_1' \boldsymbol{U}_1 = \boldsymbol{I}_{\mathcal{R}_1}$. Then (2.12) implies the following dynamic factor structure:

$$\underbrace{\boldsymbol{U}_1' \boldsymbol{y}_t}_{\text{response factor}} = \sum_{k=1}^{d} \boldsymbol{S}_k \underbrace{\Big\{ \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}) \overbrace{\boldsymbol{U}_2' \boldsymbol{y}_{t-j}}^{\text{lag-}j\text{ predictor factor}} \Big\}}_{k\text{th predictor-lag factor}} + \boldsymbol{U}_1' \boldsymbol{\varepsilon}_t. \tag{2.14}$$

The dynamic factor regression form in (2.14) enables us to interpret model (2.12) from the viewpoint of factor modeling [32, 4, 24, 5]. Specifically, $\boldsymbol{U}_1' \boldsymbol{y}_t$ can be viewed as an $\mathcal{R}_1$-dimensional response factor obtained by projecting $\boldsymbol{y}_t$ onto a low-dimensional space. Similarly, $\boldsymbol{U}_2' \boldsymbol{y}_{t-j}$ represents the $\mathcal{R}_2$-dimensional predictor factor at lag $j \geqslant 1$. While there are infinitely many lags, $\boldsymbol{U}_2' \boldsymbol{y}_{t-j}$'s are subsequently reweighted by $\ell_{j,k}(\boldsymbol{\omega})$'s, so that the kth temporal decay pattern is incorporated for $1 \leqslant k \leqslant d$. This eventually results in d predictor-lag factors, with associated weighting matrices $\boldsymbol{S}_k$'s. From (2.14), it is noteworthy that the HOSVD of $\mathcal{G}$ at its first two modes allows for different response and predictor subspaces, since $\mathcal{R}_1 \neq \mathcal{R}_2$ in general and the factor loadings $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ can be different even if $\mathcal{R}_1 = \mathcal{R}_2$. The difference between the two subspaces is intuitive and plausible in practice, since the response factor is the output of the dynamic system, whereas the predictor factors are the inputs.

To understand the connection between our model and the conventional factor model, similarly to (2.14), we can rewrite model (2.12) as

$$\boldsymbol{y}_t = \boldsymbol{U}_1 \sum_{k=1}^{d} \boldsymbol{S}_k \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}) \boldsymbol{U}_2' \boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{2.15}$$

where the right hand side has the form of a factor model with latent factor series $\boldsymbol{f}_t \in \mathbb{R}^{\mathcal{R}_1}$ and loading matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{N \times \mathcal{R}_1}$. There are different versions of normalization restrictions for the identification of $\boldsymbol{\Lambda}$ and $\boldsymbol{f}_t$ in the literature on factor models; see, e.g., [24] and [5]. Nevertheless, it always holds $\operatorname{colsp}(\boldsymbol{\Lambda}) = \operatorname{colsp}(\boldsymbol{U}_1)$. This reveals that the $\mathcal{R}_1$-dimensional response factor space of our model recovers the latent factor space of the factor model. In particular, if the normalization restriction in [24] is used, $\boldsymbol{\Lambda}$ and $\boldsymbol{f}_t$ are identical to $\boldsymbol{U}_1$ and $\sum_{k=1}^{d} \boldsymbol{S}_k \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}) \boldsymbol{U}_2' \boldsymbol{y}_{t-j}$, respectively, up to an orthogonal rotation. Then, suppose that the estimate $\widehat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ is obtained by fitting the factor model in the form of the right hand side of (2.15). The factor series $\boldsymbol{f}_t$ will be estimated by $\widehat{\boldsymbol{f}}_t = \widehat{\boldsymbol{\Lambda}}' \boldsymbol{y}_t$; see [24] for details. Moreover, by the theory for the factor model, without actually fitting our model we can get another estimator of $\boldsymbol{U}_1$, i.e., $\widehat{\boldsymbol{\Lambda}}$, which is consistent up to an orthogonal rotation.

Our model is also related to dynamic factor models in the literature [32, 33]. Consider the special case of model (2.12) where $\boldsymbol{U}_1 = \boldsymbol{U}_2$, i.e., the response and predictor factors are identical. Then it can be seen directly from (2.14) that the response (or predictor) factor series $\{\boldsymbol{U}_1' \boldsymbol{y}_t\}$ follows an $\mathcal{R}_1$-dimensional SARMA model without any low-Tucker-rank structure. Thus, for this special case, a two-step dynamic factor modeling procedure is possible: first obtain the fitted factor series $\widehat{\boldsymbol{f}}_t$ by the factor modeling method in [24], and then fit the low-dimensional SARMA model to $\{\widehat{\boldsymbol{f}}_t\}$. However, it is worth noting that this method will be inconsistent when $\operatorname{colsp}(\boldsymbol{U}_1) \neq \operatorname{colsp}(\boldsymbol{U}_2)$.

3. High-dimensional estimation.

3.1. *Rank-constrained estimation and its nonasymptotic properties.* Suppose that the observed time series $\{\boldsymbol{y}_t\}_{t=1}^{T}$ is generated by model (2.12) with Tucker ranks $(\mathcal{R}_1, \mathcal{R}_2)$ and model orders (p, r, s). Let $\boldsymbol{\Omega} \subset (-1, 1)^r \times \boldsymbol{\Pi}^s$ be the parameter space of $\boldsymbol{\omega} = (\boldsymbol{\lambda}', \boldsymbol{\eta}')'$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)'$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_s')'$, $\boldsymbol{\eta}_k = (\gamma_k, \theta_k)'$ for $1 \leqslant k \leqslant s$, and $\boldsymbol{\Pi} = [0, 1) \times (-\pi/2, \pi/2)$. Denote by $\boldsymbol{\omega}^*$, $\mathcal{G}^*$, and $\boldsymbol{\mathcal{A}}^*$ the true values of $\boldsymbol{\omega}$, $\mathcal{G}$, and $\boldsymbol{\mathcal{A}}$, respectively. Similarly, λ_j^*'s, $\boldsymbol{\eta}_j^*$'s and $\boldsymbol{G}_k^*$'s denote the true values of the corresponding parameters.

The squared error loss function for the proposed model can be written as $\mathbb{L}_T(\boldsymbol{\omega}, \mathcal{G}) = \sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{x}_t\|_2^2$, where $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \boldsymbol{y}_{t-2}', \dots)'$. Since the loss function depends on observations in the infinite past, initial values for $\{\boldsymbol{y}_s, s \leqslant 0\}$ will be needed in practice. We set them to zero for simplicity, and the corresponding effect will be accounted for in our theoretical analysis; see the discussion below Theorem 3.2. Let $\widetilde{\boldsymbol{x}}_t = (\boldsymbol{y}_{t-1}', \dots, \boldsymbol{y}_1', 0, 0, \dots)'$ be the initialized version of $\boldsymbol{x}_t$, and consider the rank-constrained least squares estimator:

$$\widehat{\boldsymbol{\mathcal{A}}} = \widehat{\mathcal{G}} \times_3 \boldsymbol{L}(\widehat{\boldsymbol{\omega}}), \quad \text{with} \quad (\widehat{\boldsymbol{\omega}}, \widehat{\mathcal{G}}) = \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}, \mathcal{G}\in\boldsymbol{\Gamma}(\mathcal{R}_1,\mathcal{R}_2)}{\arg\min} \widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \mathcal{G}), \tag{3.1}$$

where $\boldsymbol{\Gamma}(\mathcal{R}_1, \mathcal{R}_2) = \{\mathcal{G} \in \mathbb{R}^{N\times N\times d} \mid \operatorname{rank}(\mathcal{G}_{(1)}) \leqslant \mathcal{R}_1, \operatorname{rank}(\mathcal{G}_{(2)}) \leqslant \mathcal{R}_2\}$, and

$$\widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \mathcal{G}) = \sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}\widetilde{\boldsymbol{x}}_t\|_2^2 = \sum_{t=1}^{T} \Big\|\boldsymbol{y}_t - \sum_{j=1}^{t-1} \boldsymbol{A}_j(\boldsymbol{\omega}, \mathcal{G})\boldsymbol{y}_{t-j}\Big\|_2^2.$$

After conducting the estimation in (3.1), we can further obtain the estimated orthonormal factor matrices by the HOSVD, $\widehat{\mathcal{G}} = \widehat{\mathcal{S}} \times_1 \widehat{\boldsymbol{U}}_1 \times_2 \widehat{\boldsymbol{U}}_2$. Then the fitted model can be interpreted from the perspective of factor modeling as in (2.14).

ASSUMPTION 1 (Parameter space and stationarity). (i) There exists an absolute constant $0 < \bar{\rho} < 1$ such that for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, $|\lambda_1|, \dots, |\lambda_r|, \gamma_1, \dots, \gamma_s \in \Lambda$, where Λ is a compact subset of $(0, \bar{\rho})$; and (ii) the model corresponding to the true values $\boldsymbol{\omega}^*$ and $\mathcal{G}^*$ is stationary, with $\|\boldsymbol{G}_k^*\|_{\mathrm{op}} \leqslant C_{\mathcal{G}}$ for all $1 \leqslant k \leqslant d$ and some absolute constant $C_{\mathcal{G}} > 0$.

ASSUMPTION 2 (Identifiability). (i) The elements of $\boldsymbol{\omega}^*$ are nonzero, where $\lambda_1^*, \dots, \lambda_r^*$ are distinct, and $\boldsymbol{\eta}_1^*, \dots, \boldsymbol{\eta}_s^*$ are distinct; and (ii) there exist absolute constants $0 < c_{\mathcal{G}} < C_{\mathcal{G}}$ such that $c_{\mathcal{G}}\sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2} \leqslant \|\boldsymbol{G}_k^*\|_{\mathrm{F}} \leqslant C_{\mathcal{G}}\sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2}$ for $p + 1 \leqslant k \leqslant d$.

Theorem 2.3 in the previous section provides a sufficient condition for the strict stationarity. Under Assumption 1(ii), there exists a unique strictly stationary and causal solution of the true model, given by $\boldsymbol{y}_t = \boldsymbol{\Psi}_*(B)\boldsymbol{\varepsilon}_t$, where B is the backshift operator, $\boldsymbol{\Psi}_*(B) = \boldsymbol{I}_N + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j^* B^j$, and $\{\boldsymbol{\Psi}_j^*, j \geqslant 1\}$ are determined by $\boldsymbol{\omega}^*$ and $\mathcal{G}^*$; see Theorem 2.3. Denote

$$\mu_{\min}(\boldsymbol{\Psi}_*) = \min_{|z|=1} \lambda_{\min}(\boldsymbol{\Psi}_*(z)\boldsymbol{\Psi}_*^{\mathsf{H}}(z)) \quad \text{and} \quad \mu_{\max}(\boldsymbol{\Psi}_*) = \max_{|z|=1} \lambda_{\max}(\boldsymbol{\Psi}_*(z)\boldsymbol{\Psi}_*^{\mathsf{H}}(z)),$$

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 [8]. Assumption 2(i) requires that the elements of $\boldsymbol{\omega}^*$ are distinguishable, which is necessary for the identifiability of the SARMA model. For an illustration,

consider model (2.9) with $\lambda_1^* = \lambda_2^*$ and, without loss of generality, assume that $p = 0$. It holds $\ell_{j,1}(\boldsymbol{\omega}^*) = \ell_{j,2}(\boldsymbol{\omega}^*)$ for all $j \geqslant 1$, and then the data generating process has the form of

$$\boldsymbol{y}_t = \sum_{k=1}^{d} \boldsymbol{G}_k^* \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}^*)\boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t = \bar{\boldsymbol{G}}_1^* \sum_{j=1}^{\infty} \ell_{j,1}(\boldsymbol{\omega}^*)\boldsymbol{y}_{t-j} + \sum_{k=3}^{d} \boldsymbol{G}_k^* \sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}^*)\boldsymbol{y}_{t-j} + \boldsymbol{\varepsilon}_t,$$

where $\bar{\boldsymbol{G}}_1^* = \boldsymbol{G}_1^* + \boldsymbol{G}_2^*$. Thus, the model is not identifiable.

To understand Assumption 2(ii), consider the role of $\boldsymbol{G}_{p+1}^*, \ldots, \boldsymbol{G}_d^*$ in model (2.9). We can regard $\sum_{j=1}^{\infty} \ell_{j,k}(\boldsymbol{\omega}^*)\boldsymbol{y}_{t-j}$ as the kth signal with a specific temporal decay pattern. Then $\boldsymbol{G}_k^*$ can be viewed as the energy of the corresponding signal, and its overall strength can be measured by $\|\boldsymbol{G}_k^*\|_{\mathrm{F}}$. Assumption 2(ii) requires that all of the last $d - p = r + 2s$ signals are of comparable strength. In fact, we can relax this assumption to $\|\boldsymbol{G}_k^*\|_{\mathrm{F}} \asymp \|\boldsymbol{G}_j^*\|_{\mathrm{F}}$ for all $p + 1 \leqslant k \neq j \leqslant d$ or replace $\sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2}$ with any smaller rate, and Theorem 3.2 will be unaffected; note that $\sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2}$ is the largest possible rate implied by Assumption 1(ii) since $\mathrm{rank}(\boldsymbol{G}_k^*) \leqslant \mathcal{R}_1 \wedge \mathcal{R}_2$. On the contrary, if there exists $\boldsymbol{G}_{k_0}^*$ that dominates all other $\boldsymbol{G}_k^*$'s with $p+1 \leqslant k \neq k_0 \leqslant d$, such that $\|\boldsymbol{G}_{k_0}^*\|_{\mathrm{F}}/\|\boldsymbol{G}_k^*\|_{\mathrm{F}} \to \infty$ as $N \to \infty$, then only the k_0-th signal (i.e., the strong signal) will be identified, while all other signals (i.e., weak signals) will be ignored by the estimation procedure. For more details about the necessity of Assumption 2(ii), see the proof of Proposition 3.1. This is similar to the concept of strong vs. weak factors in factor models [24], and variants of this assumption may be considered with a more involved technical analysis.

Note that although the proposed model is linear in $\boldsymbol{\mathcal{A}}$, the loss function in (3.1) is nonconvex with respect to $\boldsymbol{\omega}$ and $\mathcal{G}$ jointly. In fact, it is nonconvex with respect to $\boldsymbol{\omega}$ even when fixing $\mathcal{G}$, since the nonlinear function $\boldsymbol{L}(\cdot)$ is nonconvex. This leads to a substantial challenge for the nonasymptotic analysis. To overcome it, a key intermediate step is to "linearize" the tensor mapping $\boldsymbol{\mathcal{A}}(\boldsymbol{\omega}, \mathcal{G})$ within a fixed local neighborhood of $\boldsymbol{\omega}^*$; see (A.3) in the Appendix for details. We achieve this by establishing an equivalence between the perturbation of $\boldsymbol{\mathcal{A}}^*$ and those of $\boldsymbol{\omega}^*$ and $\mathcal{G}^*$, as shown in the proposition below. It will allow us to cover the parameter space of $\boldsymbol{\mathcal{A}}$ by covering those of $\boldsymbol{\omega}$ and $\mathcal{G}$, and subsequently obtain Theorem 3.2.

PROPOSITION 3.1. Under Assumptions 1(i) and 2, for any $\boldsymbol{\mathcal{A}} = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega})$ with $\mathcal{G} \in \mathbb{R}^{N \times N \times d}$ and $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, if $\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2 \leqslant c_{\boldsymbol{\omega}}$, then

$$c_\Delta\left(\|\mathcal{G}-\mathcal{G}^*\|_{\mathrm{F}}+\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\|\boldsymbol{\omega}-\boldsymbol{\omega}^*\|_2\right)\leqslant\|\mathcal{A}-\mathcal{A}^*\|_{\mathrm{F}}$$
$$\leqslant C_\Delta\left(\|\mathcal{G}-\mathcal{G}^*\|_{\mathrm{F}}+\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\|\boldsymbol{\omega}-\boldsymbol{\omega}^*\|_2\right),$$

where $c_{\boldsymbol{\omega}}, c_\Delta, C_\Delta > 0$ are absolute constants defined in (A.2) in the Appendix.

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 $(\boldsymbol{\xi}_{it})_{1\leqslant i\leqslant N}$ within $\boldsymbol{\xi}_t$ are mutually independent and σ^2-sub-Gaussian.

Assumption 3 is weaker than the commonly imposed Gaussian assumption in the literature on high-dimensional time series (see, e.g., [8] and [39]). Let

$$\kappa_1 = \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon)\mu_{\min}(\boldsymbol{\Psi}_*)\min\{1, c_{\bar{\rho}}^2\} \quad \text{and} \quad \kappa_2 = \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\mu_{\max}(\boldsymbol{\Psi}_*)\max\{1, C_{\bar{\rho}}^2\},$$

where $c_{\bar{\rho}}, C_{\bar{\rho}} > 0$ are absolute constants defined in Lemma A.2 in the Appendix.

THEOREM 3.2. *Let $d_{\mathcal{M}} = \mathcal{R}_1\mathcal{R}_2 d + (\mathcal{R}_1 + \mathcal{R}_2)N$. Suppose that $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2 \leqslant c_{\boldsymbol{\omega}}$ and $T \gtrsim (\kappa_2/\kappa_1)^2 d_{\mathcal{M}}\log(\kappa_2/\kappa_1)$. Then under Assumptions 1–3, with probability at least $1 - 4e^{-cd_{\mathcal{M}}\log(\kappa_2/\kappa_1)} - 8e^{-cN} - \{2 + \sqrt{\kappa_2/\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)}\}\sqrt{N/\{(\mathcal{R}_1+\mathcal{R}_2)T\}}$, we have the following estimation and prediction error bounds:*

$$\|\widehat{\mathcal{A}} - \mathcal{A}^*\|_{\mathrm{F}} \lesssim \sqrt{\frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}}{\kappa_1^2 T}} \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T}\left\|(\widehat{\mathcal{A}} - \mathcal{A}^*)_{(1)}\widetilde{\boldsymbol{x}}_t\right\|_2^2 \lesssim \frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}}{\kappa_1 T}.$$

The above theorem, together with Proposition 3.1, implies that, it holds with the same high probability, $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2 \lesssim \sqrt{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}/\{\kappa_1^2(\mathcal{R}_1\wedge\mathcal{R}_2)T\}}$ and $\|\widehat{\mathcal{G}} - \mathcal{G}^*\|_{\mathrm{F}} \lesssim \sqrt{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}/(\kappa_1^2 T)}$. As a result, the consistency of $\widehat{\mathcal{A}}$, $\widehat{\boldsymbol{\omega}}$ and $\widehat{\mathcal{G}}$ can be achieved as $d_{\mathcal{M}}/T \to 0$. Moreover, in the proof of Theorem 3.2 in the Appendix, we show that the effect of initial values for $\{\boldsymbol{y}_s, s \leqslant 0\}$ is dominated by other quantities and hence has no contribution to the final estimation and prediction error rates; see the quantities $|S_i(\widehat{\boldsymbol{\Delta}})|$ for $1 \leqslant i \leqslant 3$ in the proof. The nonexponential tail probability $\{2 + \sqrt{\kappa_2/\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)}\}\sqrt{N/\{(\mathcal{R}_1+\mathcal{R}_2)T\}}$ comes from naively bounding $|S_i(\widehat{\boldsymbol{\Delta}})|$'s by Markov's inequality for technical simplicity; see the proof of Lemma A.5 for details. However, this tail probability may be sharpened by employing more sophisticated concentration inequalities.

3.2. *Alternating least squares algorithm.* Applying the Tucker decomposition for $\mathcal{A}$ in (2.13), the loss function in (3.1) can be equivalently written as

$$\widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \mathcal{S}, \boldsymbol{U}_1, \boldsymbol{U}_2) = \sum_{t=1}^{T} \|\boldsymbol{y}_t - \mathcal{A}_{(1)}\widetilde{\boldsymbol{x}}_t\|_2^2 = \sum_{t=1}^{T} \|\boldsymbol{y}_t - \{\mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{L}(\boldsymbol{\omega})\}_{(1)} \widetilde{\boldsymbol{x}}_t\|_2^2,$$

where $\mathcal{S} \in \mathbb{R}^{\mathcal{R}_1 \times \mathcal{R}_2 \times d}$, and $\boldsymbol{U}_i \in \mathbb{R}^{N \times \mathcal{R}_i}$ for $i = 1, 2$. Thus, the estimation can be implemented as $(\widehat{\boldsymbol{\omega}}, \widehat{\mathcal{S}}, \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2) = \arg\min \widetilde{\mathbb{L}}_T(\boldsymbol{\omega}, \mathcal{S}, \boldsymbol{U}_1, \boldsymbol{U}_2)$, and then the estimator of $\mathcal{A}$ is $\widehat{\mathcal{A}} = \widehat{\mathcal{G}} \times_3 \boldsymbol{L}(\widehat{\boldsymbol{\omega}})$, where $\widehat{\mathcal{G}} = \widehat{\mathcal{S}} \times_1 \widehat{\boldsymbol{U}}_1 \times_2 \widehat{\boldsymbol{U}}_2$. Note that $\boldsymbol{U}_i$'s need not be subject to any orthonormal constraint in this minimization; i.e., the outputs $\widehat{\boldsymbol{U}}_i$'s from the algorithm need not be orthonormal. Instead, to recover the orthonormal factor loading matrices $\boldsymbol{U}_i$'s such that the fitted model can be interpreted via the dynamic factor regression form in (2.14), we will conduct the HOSVD of $\widehat{\mathcal{G}}$ to get orthonormal estimates of $\boldsymbol{U}_i$'s after the optimization procedure.

Let $\boldsymbol{z}_t(\boldsymbol{\omega}) = (\boldsymbol{z}'_{t,1}(\boldsymbol{\omega}), \ldots, \boldsymbol{z}'_{t,d}(\boldsymbol{\omega}))' = \{\boldsymbol{L}'(\boldsymbol{\omega}) \otimes \boldsymbol{I}_N\}\widetilde{\boldsymbol{x}}_t \in \mathbb{R}^{Nd}$, where each N-dimensional vector $\boldsymbol{z}_{t,k}(\boldsymbol{\omega}) = \sum_{j=1}^{t-1} \ell_{j,k}(\boldsymbol{\omega})\boldsymbol{y}_{t-j}$. In addition, denote $\boldsymbol{Z}_t(\boldsymbol{\omega}) = (\boldsymbol{z}_{t,1}(\boldsymbol{\omega}), \ldots, \boldsymbol{z}_{t,d}(\boldsymbol{\omega})) \in \mathbb{R}^{N \times d}$ and $\widetilde{\boldsymbol{X}}_t = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1, \boldsymbol{0}) \in \mathbb{R}^{N \times (T-1)}$. It can be verified that

$$\begin{aligned}\mathcal{A}_{(1)}\widetilde{\boldsymbol{x}}_t &= \Big[\{\boldsymbol{z}'_t(\boldsymbol{\omega})(\boldsymbol{I}_d \otimes \boldsymbol{U}_2)\mathcal{S}'_{(1)}\} \otimes \boldsymbol{I}_N\Big] \operatorname{vec}(\boldsymbol{U}_1) = \boldsymbol{U}_1 \mathcal{S}_{(1)} \{\boldsymbol{Z}'_t(\boldsymbol{\omega}) \otimes \boldsymbol{I}_{R_2}\} \operatorname{vec}(\boldsymbol{U}'_2) \\ &= \big[\{\boldsymbol{z}'_t(\boldsymbol{\omega})(\boldsymbol{I}_d \otimes \boldsymbol{U}_2)\} \otimes \boldsymbol{U}_1\big] \operatorname{vec}(\mathcal{S}_{(1)}) = \boldsymbol{U}_1 \mathcal{G}_{(1)} (\boldsymbol{I}_d \otimes (\boldsymbol{U}'_2 \widetilde{\boldsymbol{X}}_t)) \operatorname{vec}(\boldsymbol{L}(\boldsymbol{\omega})),\end{aligned}$$

which is linear with respect to each of $\boldsymbol{L}(\boldsymbol{\omega})$, $\boldsymbol{U}_1, \boldsymbol{U}_2$ and $\mathcal{S}$, fixing the other three components. This motivates us to update the four components cyclically by an alternating least squares method.

Since $\boldsymbol{L}(\boldsymbol{\omega})$ is nonlinear in $\boldsymbol{\omega}$, let us consider the update of $\boldsymbol{\omega}$ first. Recall that $\boldsymbol{G}_k = \boldsymbol{U}_1 \boldsymbol{S}_k \boldsymbol{U}'_2$ for $1 \leqslant k \leqslant d$, where $\boldsymbol{S}_k$ is the kth frontal slice of $\mathcal{S}$; see (2.13). Let $f^I(\widetilde{\boldsymbol{x}}_t; \lambda_k) = \sum_{j=p+1}^{t-1} \lambda_k^{j-p} \boldsymbol{G}_{p+k} \boldsymbol{y}_{t-j}$ for $1 \leqslant k \leqslant r$, and

$$f^{II}(\widetilde{\boldsymbol{x}}_t; \boldsymbol{\eta}_k) = \sum_{j=p+1}^{t-1} \gamma_k^{j-p} \left[\cos\{(j-p)\theta_k\}\boldsymbol{G}_{p+r+2k-1} + \sin\{(j-p)\theta_k\}\boldsymbol{G}_{p+r+2k}\right] \boldsymbol{y}_{t-j}$$

for $1 \leqslant k \leqslant s$, where we suppress the dependence on $\boldsymbol{G}_{p+1}, \ldots, \boldsymbol{G}_d$ since they will be fixed when updating $\boldsymbol{\omega}$. It can be verified that

$$\mathcal{A}_{(1)}\widetilde{\boldsymbol{x}}_t - \sum_{k=1}^{p} \boldsymbol{G}_k \boldsymbol{y}_{t-k} = \sum_{k=1}^{r} f^I(\widetilde{\boldsymbol{x}}_t; \lambda_k) + \sum_{k=1}^{s} f^{II}(\widetilde{\boldsymbol{x}}_t; \boldsymbol{\eta}_k),$$

where each λ_k or $\boldsymbol{\eta}_k$ appears in one summand only. As a result, the loss function will be block-separable with respect to λ_k's and $\boldsymbol{\eta}_k$'s, when $\mathcal{S}, \boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are fixed. Thus, at each

Algorithm 1: Alternating least squares algorithm

1 **Input:** ranks $(\mathcal{R}_1, \mathcal{R}_2)$, model orders (p, r, s), initialization $\boldsymbol{\omega}^{(0)}, \boldsymbol{U}_1^{(0)}, \boldsymbol{U}_2^{(0)}, \mathcal{S}^{(0)}$ and $\mathcal{G}^{(0)}$.
2 **repeat** $i = 0, 1, 2, \ldots$
3 **for** $k = 1, \ldots, r$:
4 $\lambda_k^{(i+1)} \leftarrow \underset{\lambda \in (-1,1)}{\arg\min}\, \widetilde{\mathbb{L}}_T(\lambda_1^{(i+1)}, \ldots, \lambda_{k-1}^{(i+1)}, \lambda, \lambda_{k+1}^{(i)}, \ldots, \lambda_r^{(i)}, \boldsymbol{\eta}^{(i)}, \mathcal{G}^{(i)})$
5 **for** $k = 1, \ldots, s$:
6 $\boldsymbol{\eta}_k^{(i+1)} \leftarrow \underset{\boldsymbol{\eta} \in [0,1) \times (-\pi/2, \pi/2)}{\arg\min}\, \widetilde{\mathbb{L}}_T(\boldsymbol{\lambda}^{(i+1)}, \boldsymbol{\eta}_1^{(i+1)}, \ldots, \boldsymbol{\eta}_{k-1}^{(i+1)}, \boldsymbol{\eta}, \boldsymbol{\eta}_{k+1}^{(i)}, \ldots, \boldsymbol{\eta}_s^{(i)}, \mathcal{G}^{(i)})$
7 $\boldsymbol{U}_1^{(i+1)} \leftarrow \underset{\boldsymbol{U}_1}{\arg\min} \sum_{t=1}^T \|\boldsymbol{y}_t - [\{\boldsymbol{z}_t'(\boldsymbol{\omega}^{(i+1)})(\boldsymbol{I}_d \otimes \boldsymbol{U}_2^{(i)})\mathcal{S}_{(1)}^{(i)\prime}\} \otimes \boldsymbol{I}_N] \operatorname{vec}(\boldsymbol{U}_1)\|_2^2$
8 $\boldsymbol{U}_2^{(i+1)} \leftarrow \underset{\boldsymbol{U}_2}{\arg\min} \sum_{t=1}^T \|\boldsymbol{y}_t - \boldsymbol{U}_1^{(i+1)} \mathcal{S}_{(1)}^{(i)} \{\boldsymbol{Z}_t'(\boldsymbol{\omega}^{(i+1)}) \otimes \boldsymbol{I}_{R_2}\} \operatorname{vec}(\boldsymbol{U}_2')\|_2^2$
9 $\mathcal{S}^{(i+1)} \leftarrow \underset{\mathcal{S}}{\arg\min} \sum_{t=1}^T \|\boldsymbol{y}_t - [\{\boldsymbol{z}_t'(\boldsymbol{\omega}^{(i+1)})(\boldsymbol{I}_d \otimes \boldsymbol{U}_2^{(i+1)})\} \otimes \boldsymbol{U}_1^{(i+1)}] \operatorname{vec}(\mathcal{S}_{(1)})\|_2^2$
10 $\mathcal{G}^{(i+1)} = \mathcal{S}^{(i+1)} \times_1 \boldsymbol{U}_1^{(i+1)} \times_2 \boldsymbol{U}_2^{(i+1)}$.
11 **until convergence**

λ_k- or $\boldsymbol{\eta}_k$-update step, the irrelevant summands will be treated as the intercept and absorbed into the response. The corresponding optimization subproblem will be only one- or two-dimensional, which can be solved by the Newton-Raphson method. This feature will significantly alleviate the computational burden, since an increase in the orders r and s of the fitted model will have little impact on the cost of $\boldsymbol{\omega}$-updates; see Algorithm 1 and the numerical evidence in Section 4.2.

The alternating updates of $\boldsymbol{U}_1, \boldsymbol{U}_2$ and $\mathcal{S}$ are straightforward. For lines 7–9 in Algorithm 1, we may adopt the closed-form solutions to the corresponding linear least squares problems. Alternatively, when the dimension N is large and the computation of closed-form solutions is time-consuming, the gradient descent method can be used to further speed up the algorithm. Note that Algorithm 1 can also be applied to the SARMA model without any low-Tucker-rank constraint on $\mathcal{G}$. In this case, lines 7, 8 and 10 in the algorithm will be dropped, and line 9 will be the update of $\mathcal{G}^{(i+1)}$, where both $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ in the loss function are set to the $N \times N$ identity matrix.

Algorithm 1 requires predetermined ranks $(\mathcal{R}_1, \mathcal{R}_2)$ and model orders (p, r, s). In practice, we can adopt the data-driven selection procedure to be introduced in the next subsection prior to the parameter estimation. The initialization for Algorithm 1 can be conducted as follows:

We first obtain the initial value $\boldsymbol{\mathcal{A}}^{(0)}$ for $\boldsymbol{\mathcal{A}}$ by the spectral method in [19] or simply set it to $\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}}$, the initial estimator used for the rank selection; see Remark 1 in Section 3.3. To obtain initial values for $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, we conduct the HOSVD of $\boldsymbol{\mathcal{A}}^{(0)}$ for the first two modes, $\boldsymbol{\mathcal{A}}^{(0)} = \boldsymbol{\mathcal{H}}^{(0)} \times_1 \boldsymbol{U}_1^{(0)} \times_2 \boldsymbol{U}_2^{(0)}$, where $\boldsymbol{\mathcal{H}}^{(0)} \in \mathbb{R}^{\mathcal{R}_1 \times \mathcal{R}_2 \times \infty}$. Then we randomly generate the initial value $\boldsymbol{\omega}^{(0)}$ from the uniform distribution on the parameter space $\boldsymbol{\Omega}$. To get $\boldsymbol{\mathcal{S}}^{(0)}$, note that $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{H}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2$, where $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{S}} \times_3 \boldsymbol{L}(\boldsymbol{\omega})$. Since it is infeasible to recover $\boldsymbol{\mathcal{S}}^{(0)}$ exactly from $\boldsymbol{\mathcal{H}}^{(0)}$ and $\boldsymbol{L}(\boldsymbol{\omega}^{(0)})$ based on the latter equation, with a little sacrifice of information, we can first replace $\boldsymbol{\mathcal{H}}^{(0)}$ with its projection, $\boldsymbol{\mathcal{H}}^{(0)} \times_3 [\boldsymbol{L}(\boldsymbol{\omega}^{(0)})\{\boldsymbol{L}'(\boldsymbol{\omega}^{(0)})\boldsymbol{L}(\boldsymbol{\omega}^{(0)})\}^{-1}\boldsymbol{L}'(\boldsymbol{\omega}^{(0)})]$. This allows us to get the feasible solution, $\boldsymbol{\mathcal{S}}^{(0)} = \boldsymbol{\mathcal{H}}^{(0)} \times_3 [\{\boldsymbol{L}'(\boldsymbol{\omega}^{(0)})\boldsymbol{L}(\boldsymbol{\omega}^{(0)})\}^{-1}\boldsymbol{L}'(\boldsymbol{\omega}^{(0)})]$. Finally, we set $\boldsymbol{\mathcal{G}}^{(0)} = \boldsymbol{\mathcal{S}}^{(0)} \times_1 \boldsymbol{U}_1^{(0)} \times_2 \boldsymbol{U}_2^{(0)}$.

3.3. *Selection of ranks and model orders.* While it is assumed in previous subsections that the Tucker ranks and model orders are correctly specified, they are unknown in practice. In the following, we develop a two-stage procedure to select them.

Denote the true values of the ranks and orders by $\mathcal{M}^* := (\mathcal{R}_1^*, \mathcal{R}_2^*, p^*, r^*, s^*)$. Suppose that $\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}}$ is a consistent initial estimator of $\boldsymbol{\mathcal{A}}^*$. Since the Tucker ranks are the same for $\boldsymbol{\mathcal{A}}^*$ and $\boldsymbol{\mathcal{G}}^*$, at the first stage, we can select the ranks by the ridge-type ratio estimator [40]:

$$\widehat{\mathcal{R}}_i = \underset{1 \leqslant j \leqslant N-1}{\arg\min} \frac{\sigma_{j+1}(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}^{\text{init}}) + \tau}{\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}^{\text{init}}) + \tau}, \quad i = 1, 2,$$

where τ is a parameter to be chosen such that Assumption 4 below is satisfied; see similar methods in [40] and [38]. Let

$$\zeta_i = \frac{1}{\sigma_{\mathcal{R}_i^*}(\boldsymbol{\mathcal{A}}_{(i)}^*)} \max_{1 \leqslant j \leqslant \mathcal{R}_i^* - 1} \frac{\sigma_j(\boldsymbol{\mathcal{A}}_{(i)}^*)}{\sigma_{j+1}(\boldsymbol{\mathcal{A}}_{(i)}^*)}, \quad i = 1, 2.$$

ASSUMPTION 4. The parameter $\tau > 0$ is specified such that (i) $\|\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}} - \boldsymbol{\mathcal{A}}^*\|_{\text{F}}/\tau = o_p(1)$ and (ii) $\tau \max\{\zeta_1, \zeta_2\} = o(1)$.

In Assumption 4, Condition (i) requires that the estimation error of $\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}}$ is dominated by τ, and Condition (ii) can be regarded as the minimal signal assumption which will simply reduce to $\tau = o(1)$ if $\sigma_j(\boldsymbol{\mathcal{A}}_{(i)}^*)$ for $1 \leqslant j \leqslant \mathcal{R}_i^*$ and $i = 1, 2$ are bounded above and away from zero by some absolute constants; see Remark 12 in [38] for more details.

REMARK 1. The initial estimator can be obtained based on the vector AR(P) approximation of the vector AR(∞) model, where P grows with T [26]. Let $\boldsymbol{\mathcal{A}}_{1:P} = \text{stack}\{\boldsymbol{\mathcal{A}}_j, 1 \leqslant j \leqslant$

$P\}$. Then we can employ the regularized estimation, $\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}}_{1:P} = \arg\min_{\mathcal{A}\in\mathbb{R}^{N\times N\times P}} \sum_{t=P+1}^{T} \|\boldsymbol{y}_t - \sum_{j=1}^{P} \boldsymbol{A}_j \boldsymbol{y}_{t-j}\|_2^2/(T-P) + \lambda_{\text{nuc}} \sum_{i=1}^{2} \|\boldsymbol{\mathcal{A}}_{(i)}\|_*$, where $\lambda_{\text{nuc}} > 0$, and the nuclear norm regularizer enforces the low-rankness of $\boldsymbol{\mathcal{A}}_{(i)}$ for $i = 1, 2$; see, e.g., [15] and [31]. Let $\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}} = \text{stack}\{\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}}_{1:P}, \mathbf{0}_{N\times N}, \mathbf{0}_{N\times N}, \dots\}$. Along the lines of the proofs of Theorem 2 in [37] and Proposition 4.2 in [39], under some regularity conditions, it can be shown that the approximation error due to the truncation after lag P will be negligible if $P \asymp T^{1/2-\epsilon}$, where $\epsilon \in (0, 1/2)$, and then $\|\widehat{\boldsymbol{\mathcal{A}}}^{\text{init}} - \boldsymbol{\mathcal{A}}^*\|_{\text{F}} = O_p\{\sqrt{(\mathcal{R}_1^* + \mathcal{R}_2^*)NP/(T-P)}\}$. In our simulation studies, we set $\tau = \sqrt{NP\log(T-P)/10(T-P)}$ as $\mathcal{R}_1^*$ and $\mathcal{R}_2^*$ are small.

Fixing $(\mathcal{R}_1, \mathcal{R}_2) = (\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2)$, at the second stage, we obtain $(\widehat{p}, \widehat{r}, \widehat{s})$ by minimizing the high-dimensional Bayesian information criterion (BIC),

$$\text{BIC}(p, r, s; \mathcal{R}_1, \mathcal{R}_2) = \log\left\{\frac{1}{T}\sum_{t=1}^{T}\Big\|\boldsymbol{y}_t - \sum_{j=1}^{t-1}\boldsymbol{A}_j(\breve{\boldsymbol{\omega}}, \breve{\mathcal{G}})\boldsymbol{y}_{t-j}\Big\|_2^2\right\} + \frac{cd_{\mathcal{M}}\log T}{T},$$

where (p, r, s) is searched over the range $0 \leqslant p \leqslant p_{\max}$, $0 \leqslant r \leqslant r_{\max}$, and $0 \leqslant s \leqslant s_{\max}$, for some predetermined upper bounds, and $c > 0$ is an arbitrary constant. Here $\breve{\boldsymbol{\omega}}$ and $\breve{\mathcal{G}}$ represent the estimates obtained from (3.1) by fitting the model with $\mathcal{M} = (\mathcal{R}_1, \mathcal{R}_2, p, r, s)$.

Let $\mathscr{M} = \{\mathcal{M} = (\mathcal{R}_1^*, \mathcal{R}_2^*, p, r, s) \mid 0 \leqslant p \leqslant p_{\max}, 0 \leqslant r \leqslant r_{\max}, 0 \leqslant s \leqslant s_{\max}\}$. For any $\mathcal{M} \in \mathscr{M}$, let $\boldsymbol{\Omega}_{\mathcal{M}} \subset (-1, 1)^r \times \boldsymbol{\Pi}^s$ be the parameter space of $\boldsymbol{\omega}$, and $\boldsymbol{\Gamma}_{\mathcal{M}} = \{\mathcal{G} \in \mathbb{R}^{N\times N\times d} \mid \text{rank}(\mathcal{G}_{(1)}) \leqslant \mathcal{R}_1^*, \text{rank}(\mathcal{G}_{(2)}) \leqslant \mathcal{R}_2^*\}$, where $d = p + r + 2s$. Then define the population minimizer for (3.1) corresponding to $\mathcal{M}$ by

$$(\mathring{\boldsymbol{\omega}}, \mathring{\mathcal{G}}) = \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}_{\mathcal{M}}, \mathcal{G}\in\boldsymbol{\Gamma}_{\mathcal{M}}}{\arg\min}\ \mathbb{E}\left\{\|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{x}_t\|_2^2 \mid \mathscr{F}_{t-1}\right\},$$

where $\mathscr{F}_t$ denotes the σ-field generated by $\{\boldsymbol{\varepsilon}_s\}_{s\leqslant t}$. Note that $\mathcal{M}^* \in \mathscr{M}$, and when $\mathcal{M} = \mathcal{M}^*$, we simply have $(\mathring{\boldsymbol{\omega}}, \mathring{\mathcal{G}}) = (\boldsymbol{\omega}^*, \mathcal{G}^*)$ and $(\breve{\boldsymbol{\omega}}, \breve{\mathcal{G}}) = (\widehat{\boldsymbol{\omega}}, \widehat{\mathcal{G}})$. In addition, let $\mathscr{M}_{\text{mis}} = \{\mathcal{M} \in \mathscr{M} \mid p \neq p^*\} \cup \{\mathcal{M} \in \mathscr{M} \mid r < r^* \text{ or } s < s^*\}$ denote the subset of misspecified models.

ASSUMPTION 5. For every $\mathcal{M} \in \mathscr{M}$, the following conditions hold: (i) $|\lambda_1|, \dots, |\lambda_r|$, $\gamma_1, \dots, \gamma_s \in \Lambda$ for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}_{\mathcal{M}}$, where Λ is the compact subset of $(0, \bar{\rho})$ defined as in Assumption 1; (ii) $\|\mathring{\boldsymbol{G}}_k\|_{\text{op}} \leqslant C_{\mathfrak{g}}$ for $1 \leqslant k \leqslant d$, and $c_{\mathfrak{g}}\sqrt{\mathcal{R}_1^* \wedge \mathcal{R}_2^*} \leqslant \|\mathring{\boldsymbol{G}}_k\|_{\text{F}} \leqslant C_{\mathfrak{g}}\sqrt{\mathcal{R}_1^* \wedge \mathcal{R}_2^*}$ for $p + 1 \leqslant k \leqslant d$, where $0 < c_{\mathfrak{g}} < C_{\mathfrak{g}}$ are absolute constants defined as in Assumption 2; (iii) $\sigma_{\min}(\boldsymbol{L}_{\text{stack}}(\boldsymbol{\omega})) \asymp 1$ and $\sigma_{\max}(\boldsymbol{L}_{\text{stack}}(\boldsymbol{\omega})) \asymp 1$ for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}_{\mathcal{M}}$, where $\boldsymbol{L}_{\text{stack}}(\boldsymbol{\omega})$ is defined as in (A.1) in the Appendix; and (iv) $\|\breve{\boldsymbol{\omega}} - \mathring{\boldsymbol{\omega}}\|_2 \leqslant c_{\mathring{\boldsymbol{\omega}}}$, where $c_{\mathring{\boldsymbol{\omega}}} = \min\{2, c_{\mathfrak{g}}(1 - \bar{\rho})\sigma_{\min}(\boldsymbol{L}_{\text{stack}}(\mathring{\boldsymbol{\omega}}))/(8\sqrt{2}C_L C_{\mathfrak{g}})\}$ is an absolute constant.

ASSUMPTION 6. For all $\mathcal{M} \in \mathscr{M}_{\text{mis}}$, $\mathbb{E}\{\|(\boldsymbol{\mathcal{A}}^* - \mathring{\boldsymbol{\mathcal{A}}})_{(1)}\boldsymbol{x}_t\|_2^2\} \geqslant C_{\text{mis}}N$, where $\mathring{\boldsymbol{\mathcal{A}}} = \mathring{\boldsymbol{\mathcal{G}}} \times_3 \boldsymbol{L}(\mathring{\boldsymbol{\omega}})$, and $C_{\text{mis}} > 0$ is a sufficiently large absolute constant.

Assumption 5 is analogous to Assumptions 1 and 2 for all $\mathcal{M} \in \mathscr{M}$. Assumption 6 guarantees sufficient departure of the misspecified models from the true model.

THEOREM 3.3. *Under Assumption 4,* $\mathbb{P}(\widehat{\mathcal{R}}_1 = \mathcal{R}_1^*, \widehat{\mathcal{R}}_2 = \mathcal{R}_2^*) \to 1$ *as* $T \to \infty$. *Moreover, under Assumptions 1–6, if there exist absolute constants* $c, C > 0$ *such that* $c \leqslant \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon) \leqslant \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) \leqslant C$ *and* $c \leqslant \mu_{\min}(\boldsymbol{\Psi}_*) \leqslant \mu_{\max}(\boldsymbol{\Psi}_*) \leqslant C$, *and* $d_{\mathcal{M},\max}/T \to 0$ *as* $T \to \infty$, *then* $\mathbb{P}(\widehat{\mathcal{M}} = \mathcal{M}^*) \to 1$ *as* $T \to \infty$, *where* $d_{\mathcal{M},\max} = \mathcal{R}_1^*\mathcal{R}_2^*(p_{\max} + r_{\max} + 2s_{\max}) + (\mathcal{R}_1^* + \mathcal{R}_2^*)N$, *and* $\widehat{\mathcal{M}} = (\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2, \widehat{p}, \widehat{r}, \widehat{s})$.

4. Simulation studies.

4.1. *Estimation error rate and model selection consistency.* Since the proposed model is intended to serve as a tractable alternative to the vector ARMA model, we evaluate the performance of our procedure by considering data generated from vector ARMA processes. Specifically, we consider two data generating processes (DGPs),

- DGP1: the vector MA(1) model $\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1}$, and
- DGP2: the vector ARMA(1, 1) model $\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1}$,

which correspond to $p = 0$ and 1, respectively. For both DGPs, $\{\boldsymbol{\varepsilon}_t\}$ are $i.i.d.$ $N(\boldsymbol{0}, \boldsymbol{I}_N)$, and we set $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{-1}$, where $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ is a randomly generated orthogonal matrix, $\boldsymbol{J} = \text{diag}\{\lambda_1, \ldots, \lambda_r, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_s, \boldsymbol{0}\}$ is the real Jordan normal form, with each $\boldsymbol{C}_k$ being the 2×2 block determined by $\boldsymbol{\eta}_k = (\gamma_k, \theta_k)'$, defined as in (2.2). For DGP2, we set $\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{K}\boldsymbol{B}^{-1}$, where $\boldsymbol{K} = \text{diag}\{\delta, 0, \ldots, 0\}$, with the entry $\delta \neq 0$. Note that both DGPs can be written as the proposed SARMA(p, r, s) model, where $\boldsymbol{\omega} = (\lambda_1, \ldots, \lambda_r, \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_s')'$ corresponds to the parameters in $\boldsymbol{J}$. In addition, $\mathcal{R}_1 = \mathcal{R}_2 = r + 2s$ for both DGP1 and DGP2.

In the first experiment, we verify the rate of $\sqrt{d_{\mathcal{M}}/T}$ for the estimation errors $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\|_{\text{F}}$, $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^*\|_2$, and $\|\widehat{\boldsymbol{\mathcal{G}}} - \boldsymbol{\mathcal{G}}^*\|_{\text{F}}$ derived by combining Theorem 3.2 and Proposition 3.1. We set $(r, s) = (1, 0)$ and $\lambda_1 = -0.7$ for both DGPs, and $\delta = 0.5$ for DGP2. We consider $N = 10$, 20 or 40, and T is chosen such that $d_{\mathcal{M}}/T \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$. The estimation is conducted using the algorithm in Section 3.2 given the true ranks and model orders, where

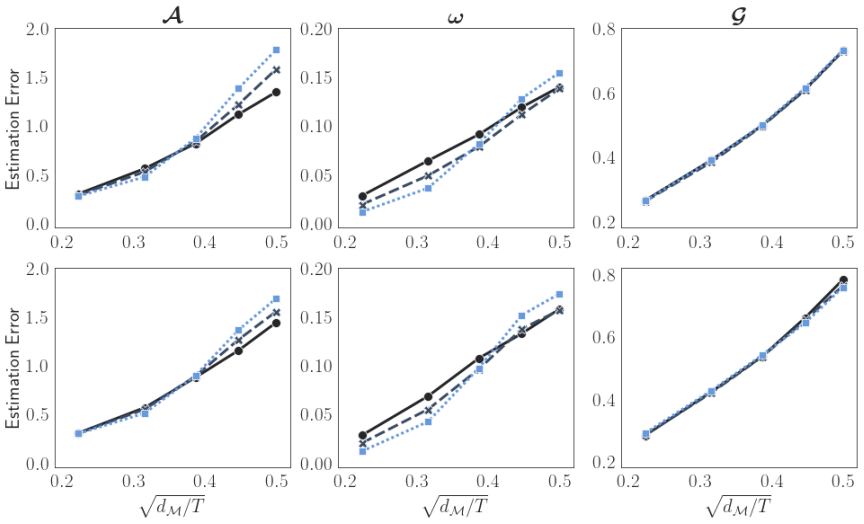

FIGURE 2. *Plots of estimation errors* $\|\widehat{\mathcal{A}} - \mathcal{A}^*\|_{\mathrm{F}}$ *(left panel),* $\|\widehat{\omega} - \omega^*\|_2$ *(middle panel) and* $\|\widehat{\mathcal{G}} - \mathcal{G}^*\|_{\mathrm{F}}$ *(right panel) against* $\sqrt{d_{\mathcal{M}}/T}$*, where* $(\mathcal{R}_1, \mathcal{R}_2, p, r, s) = (1, 1, 0, 1, 0)$ *(top panel) or* $(\mathcal{R}_1, \mathcal{R}_2, p, r, s) = (1, 1, 1, 1, 0)$ *(bottom panel), and* $N = 10$ *(●), 20 (✖) or 40 (■).*

$\boldsymbol{U}_1, \boldsymbol{U}_2$ and $\mathcal{S}$ are updated by the alternating least squares algorithm using their corresponding closed-form solutions. Figure 2 plots the estimation errors averaged over 500 replications against $\sqrt{d_{\mathcal{M}}/T}$. An approximately linear relationship between the estimation error and the theoretical rate can be observed across all settings, which confirms our theoretical results.

The aim of the second experiment is to verify the consistency of the rank and model order selection procedure. We consider three cases under DGP1: $(\mathcal{R}_1, \mathcal{R}_2, r, s)$=(1, 1, 1, 0) (model A), (2, 2, 0, 1) (model B), and (3, 3, 1, 1) (model C). The results for DGP2 are similar and hence are omitted for brevity. For models B and C, we set $\theta_1 = \pi/4$. Note that $\mathcal{A}_{(1)}$ and $\mathcal{A}_{(2)}$ have the same singular values under DGP1. Moreover, when $r \leqslant 1$ and $s \leqslant 1$, the magnitude of the nonzero singular values are directly determined by $|\lambda_1|$ and/or γ_1, which control the signal strength for the rank selection. We consider four levels of signal strength $\{0.6, 0.65, 0.7, 0.75\}$, and set $-\lambda_1$ in model A, γ_1 in model B, and $-\lambda_1 = \gamma_1$ in model C to these values. In addition, we consider $N = 10$ and $T \in [100, 600]$. At the first stage, the initial estimator $\widehat{\mathcal{A}}^{\text{init}}$ is obtained by the method in Remark 1, where $P = T^{1/3}$, and λ_{nuc} is chosen by the cross validation method for time series similar to that in [39]. At the second stage, we minimize the BIC with $c = 0.1$, $p_{\max} = 1$, and $r_{\max} = s_{\max} = 2$. Figure 3 presents the proportion of correct rank selection, $\{(\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2) = (\mathcal{R}_1^*, \mathcal{R}_2^*)\}$, and that of correct rank and model order selection, $\{\widehat{\mathcal{M}} = \mathcal{M}^*\}$, i.e., $\{(\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2, \widehat{p}, \widehat{r}, \widehat{s}) = (\mathcal{R}_1^*, \mathcal{R}_2^*, p^*, r^*, s^*)\}$, based on the two-stage procedure. It can be clearly seen that both proportions increase to one

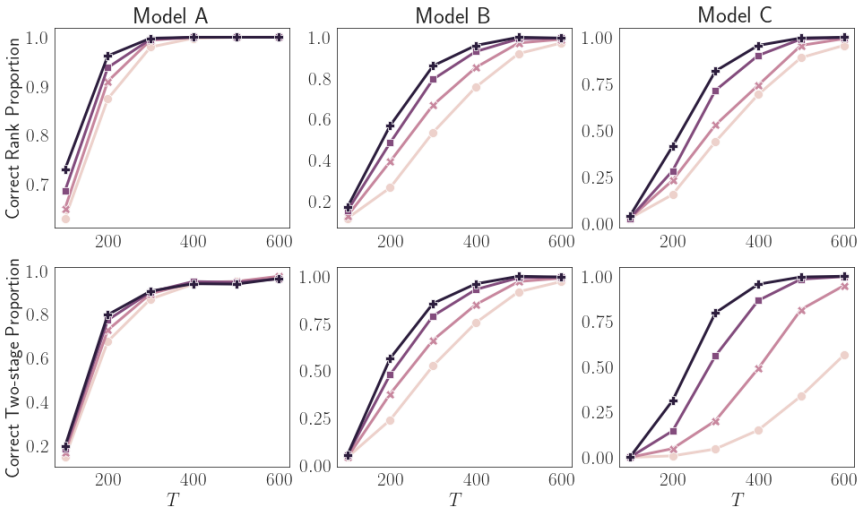

FIGURE 3. *Proportion of correct rank selection (top panel) and two-stage selection (bottom panel) for models A (left panel), B (middle panel) and C (right panel), where the signal strength is 0.6 (●), 0.65 (✖), 0.7 (■) or 0.75 (✚).*

as T and the signal strength increase. For models A and B, the proportion that the ranks and model orders are correctly selected simultaneously is fairly close to one when $T \geqslant 500$ across all settings. For model C, a larger T and signal strength is required to achieve similar performance since the model is more complex. Nevertheless, although not reported in the figure, the proportion of correct selection of $\mathcal{M}^*$ for model C under signal strength 0.6 will be close to one when $T \geqslant 900$.

4.2. *Computational advantage over vector ARMA models.* We conduct two experiments to confirm the computational advantage of the proposed SARMA model over the vector ARMA model. The computation time is reported for a Linux server with Intel(R) Xeon(R) Gold 6246 CPU@3.30GHz and 503GB of RAM, and the algorithms are all implemented by Python.

For a fair comparison, we first consider the case without any low-Tucker-rank structure for $\mathcal{G}$, i.e., $\mathcal{R}_1 = \mathcal{R}_2 = N$. The data are generated from DGP2, where $\boldsymbol{\Theta}$ is defined as in Section 4.1 with $(r, s) = (1, 1)$, $\lambda_1 = -0.8$, $\gamma_1 = 0.8$, and $\theta_1 = \pi/4$. In addition, $\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{K}\boldsymbol{B}^{-1}$, where $\boldsymbol{K} = \text{diag}\{0.8, -0.8, -0.8, 0.8, \ldots, 0.8\}$. We fit both the proposed model and the vector ARMA model to the data assuming that the true model orders are known. The estimation of the vector ARMA model is implemented by the Python function `statsmodels.tsa.statespace.varmax`, which conducts the maximum likelihood

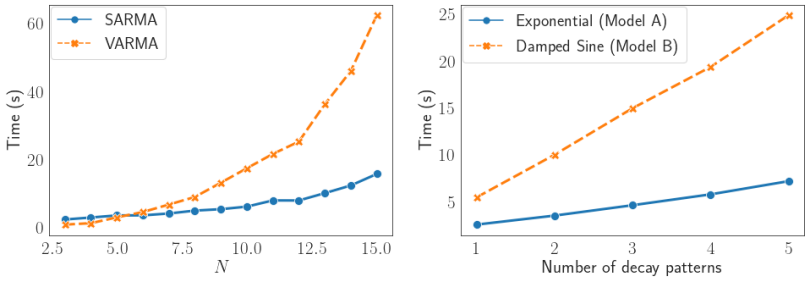

FIGURE 4. *Computation time in seconds against N for two methods (left panel) and that against the number of decay patterns, i.e., r or s, for the proposed method (right panel).*

estimation (MLE) via the Kalman filter [26]. This implementation of the MLE employs the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm by default, which is generally faster than the Newton-Raphson method. We consider $T = 500$ and $N \in [3, 15]$, since the MLE of the vector ARMA model will become extremely slow for larger N. We initialize both algorithms with the true parameter values and record the computation time of 20 iterations, averaged over 500 replications. The left panel of Figure 4 plots the average computation time against N. It can be observed that the computation time for the MLE increases much more rapidly than that of the proposed method as N increases. As discussed in previous sections, unlike the vector ARMA model, the tractable parametric structure of the proposed model allows us to avoid the computation of high-order $N \times N$ matrix polynomials. This results in the significant improvement in the computational speed.

In the next experiment, we verify the computational scalability of the proposed method with respect to increases in r and s. Similarly to the second experiment in Section 4.1, we consider two cases under DGP1: $(\mathcal{R}_1, \mathcal{R}_2, r, s) = (r, r, r, 0)$ (model A), and $(2s, 2s, 0, s)$ (model B), where $r, s \in \{1, \ldots, 5\}$. For model A, we set $-\lambda_1 = 0.8$, $-\lambda_2 = \lambda_5 = 0.5$, $-\lambda_3 = \lambda_4 = 0.2$. For model B, we set $\gamma_k = 0.1 + 0.15(k-1)$ and $\theta_k = (-1)^{k-1}\pi/4$ for $1 \leqslant k \leqslant 5$. We fix $T = 500$ and $N = 40$, and run 5 iterations of the proposed algorithm with rank constraints for 500 replications. The right panel of Figure 4 displays the average computation time against r and s for models A and B, respectively. It shows that an increase in s is more time-consuming than that in r, which is expected since an extra $\boldsymbol{\eta}_k$ involves two parameters, while another λ_k involves only one. Nevertheless, it can be observed that the computation time scales linearly in both r and s. On the other hand, since the MLE is too slow under large N, we do not report its computation time in the figure. Indeed, in most

cases, it takes more than 17 minutes for the MLE to finish 5 iterations, which is about 40 times the computation time of the proposed method.

5. Two empirical examples.

5.1. *Macroeconomic dataset.* This dataset contains observations of 20 quarterly macroeconomic variables from June 1959 to December 2019, with $T = 243$, retrieved from FRED-QD [27]. These variables come from four categories: (i) interest rates, (ii) money and credit, (iii) exchange rates, and (iv) stock market. These categories are usually considered in the construction of financial condition index, since they reflect important factors that can affect the stance of monetary policy and aggregate demand conditions [16, 9, 21]. All series are transformed to be stationary, and standardized to have zero mean and unit variance; see Table S.1 in the supplementary file for more details of the variables and their transformations.

We first explore the factor structures of this dataset. As discussed in Section 2.4, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ capture response factor and predictor factor spaces, respectively. By the rank selection method in Section 3.3, we obtain $(\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2) = (3, 3)$. Note that $\boldsymbol{U}_1$ can be alternatively estimated by $\widehat{\boldsymbol{\Lambda}}$, up to an orthogonal rotation, by fitting the factor model $\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t$ via the method in [24], where $\boldsymbol{f}_t$ is $\mathcal{R}_1$-dimensional; see the discussion below (2.15) in Section 2.4. To verify this, we obtain the estimate $\widehat{\boldsymbol{\Lambda}}$ with $\mathcal{R}_1 = 3$. Figure 5 displays $\widehat{\boldsymbol{\Lambda}}$ based on the factor model, together with $\widehat{\boldsymbol{U}}_1$ and $\widehat{\boldsymbol{U}}_2$ based on the proposed method in Section 3.1; note that all three matrices are orthonormal. Overall, it can be observed that $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{U}}_1$ are mainly influenced by variables in categories (i) and (ii), while the influence from categories (iii) and (iv) is relatively weak. On the other hand, a couple of variables from category (iv) contribute significantly to the first factor of $\widehat{\boldsymbol{U}}_2$. Moreover, the pattern of $\widehat{\boldsymbol{\Lambda}}$ is much similar to $\widehat{\boldsymbol{U}}_1$ than $\widehat{\boldsymbol{U}}_2$. Although $\widehat{\boldsymbol{U}}_1$ and $\widehat{\boldsymbol{\Lambda}}$ do not match exactly due to the rotation and estimation error, their subspace distance $\|\widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1' - \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}'\|_{\mathrm{F}}^2$ is as small as 0.52, whereas that between $\widehat{\boldsymbol{U}}_1$ and $\widehat{\boldsymbol{U}}_2$ is much larger, with $\|\widehat{\boldsymbol{U}}_1\widehat{\boldsymbol{U}}_1' - \widehat{\boldsymbol{U}}_2\widehat{\boldsymbol{U}}_2'\|_{\mathrm{F}}^2 = 3.16$. This confirms empirically the equivalence between $\boldsymbol{U}_1$ and $\boldsymbol{\Lambda}$.

We evaluate the performance of our method based on out-of-sample forecast accuracy. The following rolling forecast procedure is adopted: we first fit the models using historical data with the end point rolling from the fourth quarter of 2015 to the third quarter of 2019, and then conduct one-step-ahead forecasts based on the fitted models. In addition to the proposed

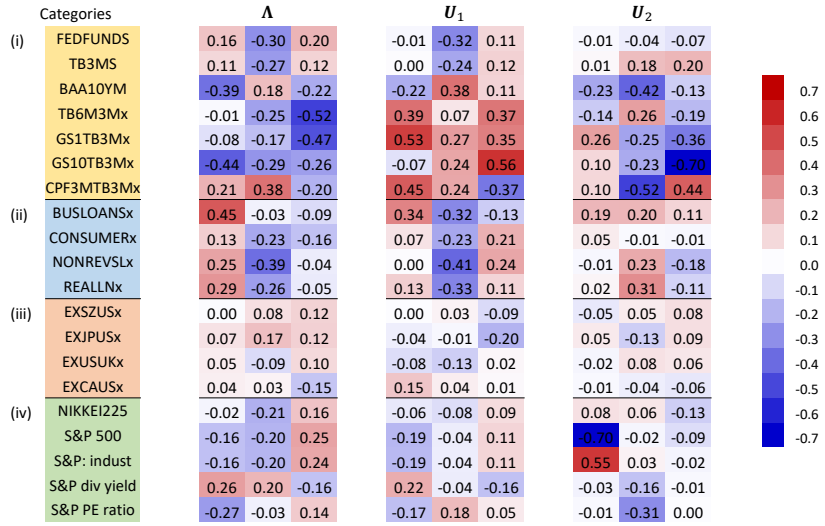

FIGURE 5. *Estimates of factor loadings in the factor model ($\boldsymbol{\Lambda}$) and the proposed model ($\boldsymbol{U}_1$ and $\boldsymbol{U}_2$) for the macroeconomic dataset.*

TABLE 1
Forecast errors for macroeconomic and realized volatility datasets. The smallest numbers in each row are marked in bold.

		Vector AR			Vector ARMA		SARMA
		Lasso	MLR	SHORR	ℓ_1	HLag	
Macroeconomic	MSFE	2.78	2.77	2.71	2.80	2.79	**2.67**
	MAFE	9.26	9.27	8.99	9.28	9.24	**8.75**
Realized Volatility	MSFE	5.17	4.93	4.87	5.19	5.19	**4.78**
	MAFE	21.58	19.02	18.22	21.70	21.70	**16.45**

method, we consider five other existing methods, including three based on the vector AR model and two based on the vector ARMA model. Specifically, for the vector AR model, we consider (a) the Lasso method [8] and two methods in [38]: (b) the multilinear low-rank (MLR) method and (c) the sparse higher-order reduced-rank (SHORR) method, which amounts to further imposing sparsity on the factor matrices in (b). For the vector ARMA model, we apply the method in [39] with (d) the ℓ_1-penalty or (e) the HLag penalty. Note that (a) is used as the Phase-I estimator for the estimators in (d) and (e), and the AR order is selected according to [39]. The AR order for (b) and (c) is chosen as in [38]. For the proposed model, the estimated model orders are $(\widehat{p}, \widehat{r}, \widehat{s}) = (0, 1, 0)$. Throughout the rolling forecast procedure, the same model orders and ranks are used. Table 1 reports the mean squared forecast error (MSFE) and mean absolute forecast error (MAFE) for all methods. It can be observed that the proposed method achieves the smallest forecast errors among all competing ones. Compared to sparsity-inducing estimation methods such as (a), (d) and

(e), the proposed model can better capture the factor structure which is prominent in this dataset. Meanwhile, the proposed model is more flexible than the finite-order vector AR model, resulting in better forecasting performance than (b) and (c).

As mentioned in Section 2.4, when $\boldsymbol{U}_1 = \boldsymbol{U}_2$, the proposed model may also be estimated by a two-step method, where the SARMA model is fitted to the estimated $\mathcal{R}$-dimensional factor series $\widehat{\boldsymbol{f}}_t$ after fitting the factor model, where $\mathcal{R} = \mathcal{R}_1 = \mathcal{R}_2$. We have performed the above rolling forecast procedure using this method and found that MSFE $= 2.80$ and MAFE $= 9.36$, where $\mathcal{R} = 3$ is adopted since $(\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2) = (3, 3)$. The worse forecasting performance of this method suggests that the predictor and response subspaces may be quite different.

5.2. *Realized volatility.* To further demonstrate the forecasting performance of the proposed method, we study daily realized volatilities for 46 stocks from January 2, 2012 to December 31, 2013, with $T = 495$. These are the stocks of top S&P 500 companies ranked by trading volumes on the first day of 2013. Specifically, we obtain the tick-by-tick data from WRDS (https://wrds-www.wharton.upenn.edu) and compute the daily realized volatility from five-minute returns [1]. By examining the sample autocorrelation functions, we have confirmed the stationarity of all series. Each series is then standardized to have zero mean and unit variance. More information about the stocks is given in Table S.2 in the supplementary file. We conduct the same rolling forecast procedure as in Section 5.1, where the last 10% of the sample is used as the forecast period. As shown in Table 1, the proposed method considerably outperforms the other ones in terms of the forecast accuracy.

The estimated ranks and model orders of the proposed model are $(\widehat{\mathcal{R}}_1, \widehat{\mathcal{R}}_2, \widehat{p}, \widehat{r}, \widehat{s}) = (1, 1, 0, 1, 0)$ for this dataset. Under this model, (2.14) can be simply written as the univariate regression, $\boldsymbol{u}_1'\boldsymbol{y}_t = \beta \sum_{j=1}^{\infty} \lambda^j \boldsymbol{u}_2'\boldsymbol{y}_{t-j} + \boldsymbol{u}_1'\boldsymbol{\varepsilon}_t$, where the one-dimensional factor series $\{\boldsymbol{u}_1'\boldsymbol{y}_t\}$ and $\{\boldsymbol{u}_2'\boldsymbol{y}_t\}$ may be regarded as volatilities of two different market indices, signifying how the market responds to and picks up risks across different stocks, respectively. In addition, it is worth noting that $\widehat{\lambda}$ for this dataset is close to 0.95. This lends further support to the well-established fact that the volatility of asset returns is highly persistent, that is, the AR process of the volatility is nearly unit-root; see, e.g., [2].

6. Conclusion and discussion. In this paper, we develop a statistically and computationally attractive alternative to the vector ARMA model. The proposed model mimics the vector ARMA temporal dynamic through a prespecified parametric function $\boldsymbol{L}(\cdot)$ that depicts the temporal decay pattern of the coefficient matrices $\boldsymbol{A}_j$'s in the vector AR(∞) form. As a result, the AR structure over infinitely many time lags can be characterized by an $(r+2s)$-dimensional parameter vector $\boldsymbol{\omega}$; or using tensor notations, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N\times\infty}$ is the mode-3 product of the tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{N\times N\times d}$ and the matrix $\boldsymbol{L}(\boldsymbol{\omega}) \in \mathbb{R}^{\infty\times d}$. In the high-dimensional regime, we further impose a low-Tucker-rank structure on $\boldsymbol{\mathcal{G}}$, leading to the factorization $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{S}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{L}(\boldsymbol{\omega})$, where $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{\mathcal{R}_1\times\mathcal{R}_2\times d}$, and $\boldsymbol{U}_i \in \mathbb{R}^{N\times\mathcal{R}_i}$ for $i = 1, 2$. This enables us to embed the concept of factor modeling in the proposed predictive framework, where the orthonormalized matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ can be interpreted as the loadings of the extracted response and predictor factors, respectively.

We discuss some extensions of the proposed method that are worthwhile to explore further. Firstly, when N is large, the estimated factor loadings may contain many nearly zero entries; see the results of $\widehat{\boldsymbol{U}}_1$ and $\widehat{\boldsymbol{U}}_2$ in Figure 5. To improve the estimation efficiency of the loading matrices, it is natural to further apply sparsity-inducing regularizations to $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$. Following [38], we may assume that these matrices are simultaneously orthonormal and sparse, and apply the vectorized ℓ_1-norm regularization $\|\boldsymbol{U}_1 \otimes \boldsymbol{U}_2\|_1$ together with orthonormality constraints on each $\boldsymbol{U}_i$. Combined with the nonasymptotic techniques in this paper, it can be shown that the estimation error bound for the corresponding estimator of $\boldsymbol{\mathcal{A}}$ will be sharpened to $O_p(\sqrt{K \log N/T})$, where K denotes the number of nonzero entries in $\boldsymbol{U}_1 \otimes \boldsymbol{U}_2$. Thus, a much larger dimension N can be accommodated. Nevertheless, the non-smoothness and non-convexity of the corresponding optimization problem will pose challenges to the development of an efficient algorithm. Thus, we leave this for future research.

Another interesting direction is to consider other forms of $\boldsymbol{L}(\cdot)$. A straightforward generalization is to adopt the more sophisticated decay patterns derived from the general vector ARMA model; see Proposition 2.2, Figure 6 and Section S4 of the supplementary material for details. To allow for even more flexibility in the temporal structure, we can further consider other user-defined decay patterns based on prior knowledge of the data. For instance, economic and financial time series often exhibit seasonality, i.e., similar fluctuations recurring at regular intervals. Denote by h the length of the regular intervals, a.k.a. the sea-

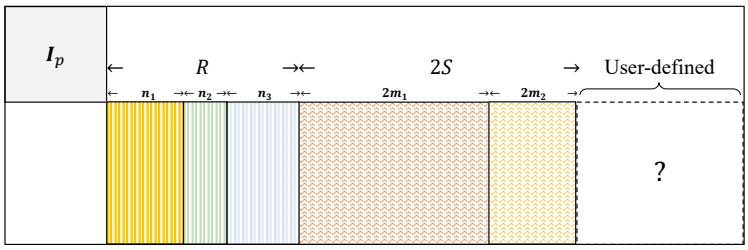

FIGURE 6. *General form of $\boldsymbol{L}(\boldsymbol{\omega})$, where rectangles filled with vertical stripes are blocks in Type I parametric form (exponential decays), those filled with waves are blocks in Type II parametric form (damped sine and cosine waves), different colors indicate dependence on different subsets of parameters λ_k or $\boldsymbol{\eta}_k$; see the supplementary material for detailed expressions. The area with a question mark represents possible extra user-defined parametric blocks.*

sonal period; e.g., $h = 4$ for quarterly data and 12 for monthly data. The seasonality can be induced by incorporating seasonal versions of $\boldsymbol{\ell}^I(\lambda_k)$ and $\boldsymbol{\ell}^{II}(\boldsymbol{\eta}_k)$. For example, we may define $\boldsymbol{\ell}^I_{\text{season}}(\lambda_k) = (\lambda_k, 0, \dots, 0, \lambda_k^2, 0, \dots)'$ and

$$\boldsymbol{\ell}^{II}_{\text{season}}(\boldsymbol{\eta}_k) = \begin{pmatrix} \gamma_k \cos(\theta_k)\ 0 \cdots 0\ \gamma_k^2 \cos(2\theta_k)\ 0 \cdots \\ \gamma_k \sin(\theta_k)\ 0 \cdots 0\ \gamma_k^2 \sin(2\theta_k)\ 0 \cdots \end{pmatrix}',$$

where the nonzero entries occur once every h rows in each column of $\boldsymbol{\ell}^I_{\text{season}}(\lambda_k)$ and $\boldsymbol{\ell}^{II}_{\text{season}}(\boldsymbol{\eta}_k)$. In general, both the regular decay patterns, $\boldsymbol{\ell}^I(\cdot)$ and $\boldsymbol{\ell}^{II}(\cdot)$, and the seasonal decay patterns can be included in $\boldsymbol{L}(\cdot)$; see also Figure 6. Hence, the corresponding seasonal SARMA model will be able to generate various kinds of patterns involving both seasonal and non-seasonal components.

APPENDIX: TECHNICAL DETAILS

A.1. Useful properties of $\boldsymbol{L}(\boldsymbol{\omega})$. According to the definition of $\boldsymbol{L}(\boldsymbol{\omega})$, for $j \geqslant 1$, denote the jth entry of $\boldsymbol{\ell}^I(\lambda_i)$ by $\ell^I_j(\lambda_i) = \lambda_i^j$ with $1 \leqslant i \leqslant r$, and denote the transpose of the jth row of the $\infty \times 2$ matrix $\boldsymbol{\ell}^{II}(\boldsymbol{\eta}_k)$ by $\ell^{II}_j(\boldsymbol{\eta}_k) := (\ell^{II,1}_j(\boldsymbol{\eta}_k), \ell^{II,2}_j(\boldsymbol{\eta}_k))' = (\gamma_k^j \cos(j\theta_k), \gamma_k^j \sin(j\theta_k))'$ with $1 \leqslant k \leqslant s$. Let $\boldsymbol{L}^I(\boldsymbol{\lambda}) = (\boldsymbol{\ell}^I(\lambda_1), \cdots, \boldsymbol{\ell}^I(\lambda_r))$ and $\boldsymbol{L}^{II}(\boldsymbol{\eta}) = (\boldsymbol{\ell}^{II}(\boldsymbol{\eta}_1), \cdots, \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_s))$. In addition, define the following matrix by augmenting $\boldsymbol{L}(\boldsymbol{\omega})$ with $(r + 2s)$ extra columns consisting of first-order derivatives:

$$\boldsymbol{L}_{\text{stack}}(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{I}_p & \\ & \boldsymbol{L}^I(\boldsymbol{\lambda})\ \boldsymbol{L}^{II}(\boldsymbol{\eta})\ \nabla \boldsymbol{L}^I(\boldsymbol{\lambda})\ \nabla_\theta \boldsymbol{L}^{II}(\boldsymbol{\eta}) \end{pmatrix}, \tag{A.1}$$

where $\nabla \boldsymbol{L}^I(\boldsymbol{\lambda}) = (\nabla \boldsymbol{\ell}^I(\lambda_1), \cdots, \nabla \boldsymbol{\ell}^I(\lambda_r))$ and $\nabla_\theta \boldsymbol{L}^{II}(\boldsymbol{\eta}) = (\nabla_\theta \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_1), \cdots, \nabla_\theta \boldsymbol{\ell}^{II}(\boldsymbol{\eta}_s))$. We can similarly define $\nabla_\gamma \boldsymbol{L}^{II}(\boldsymbol{\eta})$. Note that from (A.12), it holds $\text{colsp}\{\nabla_\gamma \boldsymbol{L}^{II}(\boldsymbol{\eta})\} =$

$\mathrm{colsp}\{\nabla_\theta \boldsymbol{L}^{II}(\boldsymbol{\eta})\}$, which is why $\nabla_\gamma \boldsymbol{L}^{II}(\boldsymbol{\eta})$ is not included in $\boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega})$. Denote

$$\sigma_{\min,L} = \sigma_{\min}(\boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*)) \quad \text{and} \quad \sigma_{\max,L} = \sigma_{\max}(\boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*)),$$

where $\boldsymbol{\omega}^*$ is the true value of $\boldsymbol{\omega}$. Lemma A.1 below gives some exponential decay properties induced by the parametric form of $\boldsymbol{L}(\boldsymbol{\omega})$, which will be used repeatedly in our theoretical analysis. Then, based on Lemma A.1(ii), we can show that $\sigma_{\min,L} \asymp 1$ and $\sigma_{\max,L} \asymp 1$ if Assumption 2(i) is further imposed; this is stated in Lemma A.2.

LEMMA A.1. *Suppose that Assumption 1(i) holds. Then (i) there exists an absolute constant $C_L > 0$ such that for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ and $j \geqslant 1$,*

$$\max_{1\leqslant i\leqslant r, 1\leqslant k\leqslant s, 1\leqslant h\leqslant 2} \{|\nabla \ell_j^I(\lambda_i)|, \|\nabla \ell_j^{II,h}(\boldsymbol{\eta}_k)\|_2, |\nabla^2 \ell_j^I(\lambda_i)|, \|\nabla^2 \ell_j^{II,h}(\boldsymbol{\eta}_k)\|_{\mathrm{F}}\} \leqslant C_L \bar{\rho}^j;$$

and (ii) there exists an absolute constant $C_* > 0$ such that $\|\boldsymbol{A}_j^*\|_{\mathrm{op}} \leqslant C_* \bar{\rho}^j$ for all $j \geqslant 1$ if Assumption 1(ii) further holds.

LEMMA A.2. *Let $J = 2(r+2s)$. Denote $x_k^* = \lambda_k^*$ for $1 \leqslant k \leqslant r$ and $x_{r+2k-1}^* = \gamma_k^* e^{i\theta_k^*}, x_{r+2k}^* = \gamma_k^* e^{-i\theta_k^*}$ for $1 \leqslant k \leqslant s$, and let $\nu_1^* = \min\{|x_k^*|, 1 \leqslant k \leqslant r+2s\}$ and $\nu_2^* = \min\{|x_j^* - x_k^*|, 1 \leqslant j < k \leqslant r+2s\}$. Under Assumptions 1(i) and 2(i), the matrix $\boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*)$ has full rank, and its maximum and minimum singular values satisfy*

$$\min\{1, c_{\bar{\rho}}\} \leqslant \sigma_{\min,L} \leqslant \sigma_{\max,L} \leqslant \max\{1, C_{\bar{\rho}}\}.$$

where $C_{\bar{\rho}} = C_1 \bar{\rho} \sqrt{J}(1-\bar{\rho})^{-1} \asymp 1$ and $c_{\bar{\rho}} = 0.25^s (\nu_1^*)^{3J/2} (\nu_2^*)^{J(J/2-1)} / C_{\bar{\rho}}^{J-1} \asymp 1$.

By Lemma A.2, we have $c_{\boldsymbol{\omega}} \asymp 1$, $c_\Delta \asymp 1$, and $C_\Delta \asymp 1$, where

$$c_{\boldsymbol{\omega}} = \min\left\{2, \frac{c_{\mathfrak{g}}(1-\bar{\rho})\sigma_{\min,L}}{8\sqrt{2}C_L C_{\mathfrak{g}}}\right\},$$
$$c_\Delta = c_l \cdot \sigma_{\min,L}, \quad C_\Delta = c_u \cdot \max\left\{\sigma_{\max,L}, (1-\bar{\rho})^{-1}\right\}, \tag{A.2}$$

with $c_l = 0.25\min\{1, \sqrt{2}c_{\mathfrak{g}}\}$ and $c_u = 1 + \sqrt{2}C_{\mathfrak{g}}(\min_{1\leqslant k\leqslant s}\gamma_k^*)^{-1} + 4\sqrt{2}C_L + 2C_L C_{\mathfrak{g}}$; for details about how these constants are chosen, see the proof of Proposition 3.1.

A.2. Notations and main idea: linearization of parametric structure. For simplicity, denote the perturbations of $\boldsymbol{\omega}^*$, $\mathcal{G}^*$ and $\mathcal{A}^*$ by $\delta_{\boldsymbol{\omega}} = \|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_2$, $\delta_{\mathfrak{g}} = \|\mathcal{G} - \mathcal{G}^*\|_{\mathrm{F}}$ and $\delta_{\mathcal{A}} = \|\mathcal{A} - \mathcal{A}^*\|_{\mathrm{F}} = \|\boldsymbol{\Delta}\|_{\mathrm{F}}$, respectively. Let

$$\boldsymbol{\Upsilon} = \left\{\boldsymbol{\Delta} = \mathcal{A} - \mathcal{A}^* \in \mathbb{R}^{N\times N\times\infty} \mid \mathcal{A} = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega}), \mathcal{G} \in \boldsymbol{\Gamma}(\mathcal{R}_1, \mathcal{R}_2), \boldsymbol{\omega} \in \boldsymbol{\Omega}, \delta_{\boldsymbol{\omega}} \leqslant c_{\boldsymbol{\omega}}\right\},$$

where $\mathbf{\Gamma}(\mathcal{R}_1, \mathcal{R}_2) = \{\mathcal{G} \in \mathbb{R}^{N\times N\times d} \mid \mathrm{rank}(\mathcal{G}_{(1)}) \leqslant \mathcal{R}_1, \mathrm{rank}(\mathcal{G}_{(2)}) \leqslant \mathcal{R}_2\}$. It is noteworthy that under the conditions of Theorem 3.2, $\widehat{\mathbf{\Delta}} := \widehat{\mathcal{A}} - \mathcal{A}^* \in \mathbf{\Upsilon}$.

A crucial intermediate step for our theoretical analysis is to establish the following linear approximation within a fixed local neighborhood of $\boldsymbol{\omega}^*$,

$$\mathbf{\Delta}(\boldsymbol{\omega}, \mathcal{G}) = \mathcal{A}(\boldsymbol{\omega}, \mathcal{G}) - \mathcal{A}^* \approx \mathfrak{M}(\boldsymbol{\omega} - \boldsymbol{\omega}^*, \mathcal{G} - \mathcal{G}^*) \times_3 \boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*), \tag{A.3}$$

where $\mathfrak{M} : \mathbb{R}^{r+2s} \times \mathbb{R}^{N\times N\times d} \to \mathbb{R}^{N\times N\times(d+r+2s)}$ is a bilinear function defined as follows:

$$\begin{aligned}\mathfrak{M}(\boldsymbol{a}, \mathcal{B}) = \mathrm{stack}\Bigg(&\mathcal{B}, \ \left\{a_i \boldsymbol{G}_i^{I*}\right\}_{1\leqslant i\leqslant r}, \\ &\left\{a_{r+2k}\boldsymbol{G}_k^{II,1*} - \frac{a_{r+2k-1}}{\gamma_k^*}\boldsymbol{G}_k^{II,2*}, \ a_{r+2k-1}\boldsymbol{G}_k^{II,2*} + \frac{a_{r+2k}}{\gamma_k^*}\boldsymbol{G}_k^{II,1*}\right\}_{1\leqslant k\leqslant s}\Bigg),\end{aligned} \tag{A.4}$$

for any $\boldsymbol{a} = (a_1, \ldots, a_{r+2s})' \in \mathbb{R}^{r+2s}$ and $\mathcal{B} \in \mathbb{R}^{N\times N\times d}$, with the true values $\boldsymbol{\omega}^*$ and $\mathcal{G}^*$ fixed. The linear approximation in (A.3) will be formalized in the proof of Proposition 3.1; in particular, see (A.14) and (A.15) for the linear form and the remainder term, respectively. Note that both the conclusion and proof techniques of Proposition 3.1 are used repeatedly throughout our theoretical analysis, which is detailed in the supplementary material.

In addition, the following notations will be used in the proof of Proposition 3.1. First, for the convenience of notation in the proof, according to the block form of $\boldsymbol{L}(\boldsymbol{\omega})$, we partition $\mathcal{G} \in \mathbb{R}^{N\times N\times d}$ as $\mathcal{G} = \mathrm{stack}(\mathcal{G}^{\mathrm{AR}}, \mathcal{G}^{\mathrm{MA}}) = (\mathcal{G}^{\mathrm{AR}}, \mathcal{G}^{I}, \mathcal{G}^{II})$, where $\mathcal{G}^{\mathrm{AR}} = \mathrm{stack}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p)$, $\mathcal{G}^{I} = \mathrm{stack}(\boldsymbol{G}_1^{I}, \ldots, \boldsymbol{G}_r^{I})$, and $\mathcal{G}^{II} = \mathrm{stack}(\boldsymbol{G}_1^{II,1}, \boldsymbol{G}_1^{II,2}, \ldots, \boldsymbol{G}_s^{II,1}, \boldsymbol{G}_s^{II,2})$ are $N \times N \times p$, $N \times N \times r$, and $N \times N \times 2s$ tensors, respectively. Here, $\boldsymbol{G}_i^{I} = \boldsymbol{G}_{p+i}$ for $1 \leqslant i \leqslant r$, and $\boldsymbol{G}_k^{II,1} = \boldsymbol{G}_{p+r+2k-1}$ and $\boldsymbol{G}_k^{II,2} = \boldsymbol{G}_{p+r+2k}$ for $1 \leqslant k \leqslant s$ according to our relabeling at the beginning of Section 2.3. Then, for any $\mathcal{A} = \mathcal{G} \times_3 \boldsymbol{L}(\boldsymbol{\omega})$, we have $\boldsymbol{A}_k = \boldsymbol{G}_k$ for $1 \leqslant k \leqslant p$, and

$$\boldsymbol{A}_{p+j} = \sum_{i=1}^{r} \ell_j^{I}(\lambda_i)\boldsymbol{G}_i^{I} + \sum_{k=1}^{s}\left\{\ell_j^{II,1}(\boldsymbol{\eta}_k)\boldsymbol{G}_k^{II,1} + \ell_j^{II,2}(\boldsymbol{\eta}_k)\boldsymbol{G}_k^{II,2}\right\}, \ \text{ for } \ j \geqslant 1. \tag{A.5}$$

Moreover, for simplicity, let

$$\mathcal{G}_{\mathrm{stack}} = \mathfrak{M}(\boldsymbol{\omega} - \boldsymbol{\omega}^*, \mathcal{G} - \mathcal{G}^*).$$

Equivalently, we can express $\mathcal{G}_{\mathrm{stack}} = \mathrm{stack}\,(\mathcal{G} - \mathcal{G}^*, \mathcal{D}(\boldsymbol{\omega}))$ as the $N \times N \times (d + r + 2s)$ tensor formed by augmenting $\mathcal{G} - \mathcal{G}^*$ with the $N \times N \times (r + 2s)$ tensor

$$\mathcal{D}(\boldsymbol{\omega}) = \mathrm{stack}\left(\left\{(\lambda_i - \lambda_i^*)\boldsymbol{G}_i^{I*}\right\}_{1\leqslant i\leqslant r},\right.$$

$$\left\{(\theta_k-\theta_k^*)\boldsymbol{G}_k^{II,1*}-\frac{\gamma_k-\gamma_k^*}{\gamma_k^*}\boldsymbol{G}_k^{II,2*},\ (\theta_k-\theta_k^*)\boldsymbol{G}_k^{II,2*}+\frac{\gamma_k-\gamma_k^*}{\gamma_k^*}\boldsymbol{G}_k^{II,1*}\right\}_{1\leqslant k\leqslant s}\Bigg).$$

Lastly, note that for every $\boldsymbol{\Delta}(\boldsymbol{\omega},\mathcal{G})\in\boldsymbol{\Upsilon}$, its corresponding $\mathcal{G}_{\text{stack}}\in\boldsymbol{\Xi}$, where

$$\boldsymbol{\Xi}=\left\{\mathcal{M}(\boldsymbol{a},\mathcal{B})\in\mathbb{R}^{N\times N\times(d+r+2s)}\mid \boldsymbol{a}\in\mathbb{R}^{r+2s},\mathcal{B}\in\boldsymbol{\Gamma}(2\mathcal{R}_1,2\mathcal{R}_2)\right\}. \tag{A.6}$$

A.3. Proof of Proposition 3.1. Let $\boldsymbol{\Delta}=\mathcal{A}-\mathcal{A}^*=\mathcal{G}\times_3\boldsymbol{L}(\boldsymbol{\omega})-\mathcal{G}^*\times_3\boldsymbol{L}(\boldsymbol{\omega}^*)$. Denote by $\boldsymbol{\Delta}_j$ with $j\geqslant 1$ the frontal slices of $\boldsymbol{\Delta}$, i.e. $\boldsymbol{\Delta}_{(1)}=(\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2,\dots)$. Then $\boldsymbol{\Delta}_j=\boldsymbol{G}_j-\boldsymbol{G}_j^*$ for $1\leqslant j\leqslant p$. For $j\geqslant 1$, by (A.5) and the Taylor expansion,

$$\begin{aligned}
\boldsymbol{\Delta}_{p+j}&=\boldsymbol{A}_{p+j}-\boldsymbol{A}_{p+j}^*\\
&=\sum_{k=1}^{r}\left\{\ell_j^I(\lambda_k^*)+\nabla\ell_j^I(\lambda_k^*)(\lambda_k-\lambda_k^*)+\frac{1}{2}\nabla^2\ell_j^I(\widetilde{\lambda}_k)(\lambda_k-\lambda_k^*)^2\right\}\boldsymbol{G}_k^I\\
&\quad+\sum_{k=1}^{s}\left\{\ell_j^{II,1}(\boldsymbol{\eta}_k^*)+(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)'\nabla\ell_j^{II,1}(\boldsymbol{\eta}_k^*)\right.\\
&\qquad\qquad\left.+\frac{1}{2}(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)'\nabla^2\ell_j^{II,1}(\widetilde{\boldsymbol{\eta}}_k)(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)\right\}\boldsymbol{G}_k^{II,1}\\
&\quad+\sum_{k=1}^{s}\left\{\ell_j^{II,2}(\boldsymbol{\eta}_k^*)+(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)'\nabla\ell_j^{II,2}(\boldsymbol{\eta}_k^*)\right.\\
&\qquad\qquad\left.+\frac{1}{2}(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)'\nabla^2\ell_j^{II,2}(\widetilde{\boldsymbol{\eta}}_k)(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)\right\}\boldsymbol{G}_k^{II,2}-\boldsymbol{A}_{p+j}^*\\
&:=\boldsymbol{H}_j+\boldsymbol{R}_j,
\end{aligned}\tag{A.7}$$

where $\widetilde{\lambda}_k$ lies between λ_k^* and λ_k for $1\leqslant k\leqslant r$, $\widetilde{\boldsymbol{\eta}}_k$ lies between $\boldsymbol{\eta}_k^*$ and $\boldsymbol{\eta}_k$ for $1\leqslant k\leqslant s$,

$$\begin{aligned}
\boldsymbol{H}_j&=\sum_{k=1}^{r}\ell_j^I(\lambda_k^*)(\boldsymbol{G}_k^I-\boldsymbol{G}_k^{I*})+\sum_{k=1}^{s}\sum_{h=1}^{2}\ell_j^{II,h}(\boldsymbol{\eta}_k^*)(\boldsymbol{G}_k^{II,h}-\boldsymbol{G}_k^{II,h*})\\
&\quad+\sum_{k=1}^{r}(\lambda_k-\lambda_k^*)\nabla\ell_j^I(\lambda_k^*)\boldsymbol{G}_k^{I*}+\sum_{k=1}^{s}\sum_{h=1}^{2}(\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^*)'\nabla\ell_j^{II,h}(\boldsymbol{\eta}_k^*)\boldsymbol{G}_k^{II,h*},
\end{aligned}\tag{A.8}$$

and

$$\begin{aligned}\boldsymbol{R}_j = &\sum_{i=1}^{r} \nabla \ell_j^I(\lambda_k^*)(\lambda_k - \lambda_k^*)(\boldsymbol{G}_k^I - \boldsymbol{G}_k^{I*}) \\ &+ \sum_{k=1}^{s}\sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla \ell_j^{II,h}(\boldsymbol{\eta}_k^*)(\boldsymbol{G}_k^{II,h} - \boldsymbol{G}_k^{II,h*}) \\ &+ \frac{1}{2}\sum_{k=1}^{r} \nabla^2 \ell_j^I(\widetilde{\lambda}_k)(\lambda_k - \lambda_k^*)^2 \boldsymbol{G}_k^I \\ &+ \frac{1}{2}\sum_{k=1}^{s}\sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla^2 \ell_j^{II,h}(\widetilde{\boldsymbol{\eta}}_k)(\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)\boldsymbol{G}_k^{II,h}.\end{aligned} \tag{A.9}$$

We first handle the terms in $\boldsymbol{R}_j$, and denote $\boldsymbol{R}_j = \boldsymbol{R}_{1j} + \boldsymbol{R}_{2j} + \boldsymbol{R}_{3j}$, where

$$\begin{aligned}\boldsymbol{R}_{1j} = &\sum_{k=1}^{r} \nabla \ell_j^I(\lambda_k^*)(\lambda_k - \lambda_k^*)(\boldsymbol{G}_k^I - \boldsymbol{G}_k^{I*}) \\ &+ \sum_{k=1}^{s}\sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla \ell_j^{II,h}(\boldsymbol{\eta}_k^*)(\boldsymbol{G}_k^{II,h} - \boldsymbol{G}_k^{II,h*}), \\ \boldsymbol{R}_{2j} = &\frac{1}{2}\sum_{k=1}^{r} \nabla^2 \ell_j^I(\widetilde{\lambda}_k)(\lambda_k - \lambda_k^*)^2 (\boldsymbol{G}_k^I - \boldsymbol{G}_k^{I*}) \\ &+ \frac{1}{2}\sum_{k=1}^{s}\sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla^2 \ell_j^{II,h}(\widetilde{\boldsymbol{\eta}}_k)(\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)(\boldsymbol{G}_k^{II,h} - \boldsymbol{G}_k^{II,h*}), \\ \boldsymbol{R}_{3j} = &\frac{1}{2}\sum_{k=1}^{r} \nabla^2 \ell_j^I(\widetilde{\lambda}_k)(\lambda_k - \lambda_k^*)^2 \boldsymbol{G}_k^{I*} \\ &+ \frac{1}{2}\sum_{k=1}^{s}\sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla^2 \ell_j^{II,h}(\widetilde{\boldsymbol{\eta}}_k)(\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)\boldsymbol{G}_k^{II,h*}.\end{aligned} \tag{A.10}$$

Note that for any matrix $\boldsymbol{Y} = \sum_{k=1}^{d} a_k \boldsymbol{X}_k$, it holds

$$\|\boldsymbol{Y}\|_{\mathrm{op}} \leqslant \|\boldsymbol{Y}\|_{\mathrm{F}} \leqslant (\sum_{k=1}^{d} \|\boldsymbol{X}_k\|_{\mathrm{F}}^2)^{1/2} (\sum_{k=1}^{d} a_k^2)^{1/2} = \|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} \|\boldsymbol{a}\|_2,$$

and $\sum_{k=1}^{d} a_k^4 \leqslant (\sum_{k=1}^{d} a_k^2)^2$, where $\boldsymbol{a} = (a_1, \ldots, a_d)' \in \mathbb{R}^d$, and $\boldsymbol{\mathcal{X}}$ is a tensor with frontal slices $\boldsymbol{X}_k$'s such that $\boldsymbol{\mathcal{X}}_{(1)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$. Then, by Lemma A.1(i),

$$\begin{aligned}\|\boldsymbol{R}_{1j}\|_{\mathrm{F}} &\leqslant C_L \bar{\rho}^j \sqrt{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 + 2\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|_2^2} \cdot \|\mathcal{G}^{\mathrm{MA}} - \mathcal{G}^{\mathrm{MA}*}\|_{\mathrm{F}} \\ &\leqslant \sqrt{2} C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}} \cdot \|\mathcal{G}^{\mathrm{MA}} - \mathcal{G}^{\mathrm{MA}*}\|_{\mathrm{F}} \leqslant \sqrt{2} C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}} \delta_{\mathcal{G}},\end{aligned}$$

and

$$\|\boldsymbol{R}_{2j}\|_{\mathrm{F}} \leqslant \frac{\sqrt{2}}{2} C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}}^2 \sqrt{\sum_{k=1}^{r} \|\boldsymbol{G}_k^I - \boldsymbol{G}_k^{I*}\|_{\mathrm{F}}^2 + \sum_{k=1}^{s}\sum_{h=1}^{2} \|\boldsymbol{G}_k^{II,h} - \boldsymbol{G}_k^{II,h*}\|_{\mathrm{F}}^2}$$

$$\leqslant \frac{\sqrt{2}}{2} C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}}^2 \cdot \|\mathcal{G}^{\mathrm{MA}} - \mathcal{G}^{\mathrm{MA}*}\|_{\mathrm{F}} \leqslant \frac{\sqrt{2}}{2} C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}}^2 \delta_{\mathfrak{g}}.$$

Moreover, by Assumption 2(ii) and Lemma A.1(i), we can show that

$$\|\boldsymbol{R}_{3j}\|_{\mathrm{F}} \leqslant C_L C_{\mathfrak{g}} \bar{\rho}^j \sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2} \delta_{\boldsymbol{\omega}}^2.$$

As a result,

$$\begin{aligned} \|\boldsymbol{R}_j\|_{\mathrm{F}} &\leqslant \|\boldsymbol{R}_{1j}\|_{\mathrm{F}} + \|\boldsymbol{R}_{2j}\|_{\mathrm{F}} + \|\boldsymbol{R}_{3j}\|_{\mathrm{F}} \\ &\leqslant C_L \bar{\rho}^j \delta_{\boldsymbol{\omega}} \left(\sqrt{2}\delta_{\mathfrak{g}} + \frac{\sqrt{2}}{2} \delta_{\boldsymbol{\omega}} \delta_{\mathfrak{g}} + C_{\mathfrak{g}} \sqrt{\mathcal{R}_1 \wedge \mathcal{R}_2} \delta_{\boldsymbol{\omega}} \right). \end{aligned} \tag{A.11}$$

Now consider $\boldsymbol{H}_j$ in (A.8). Notice that for any $j \geqslant 1$ and $1 \leqslant k \leqslant s$,

$$\begin{aligned} \nabla_\gamma \ell_j^{II,1}(\boldsymbol{\eta}_k) &= j\gamma_k^{j-1} \cos(j\theta_k) = \frac{1}{\gamma_k} \nabla_\theta \ell_j^{II,2}(\boldsymbol{\eta}_k), \\ \nabla_\gamma \ell_j^{II,2}(\boldsymbol{\eta}_k) &= j\gamma_k^{j-1} \sin(j\theta_k) = -\frac{1}{\gamma_k} \nabla_\theta \ell_j^{II,1}(\boldsymbol{\eta}_k). \end{aligned} \tag{A.12}$$

Thus, the last term on the right side of (A.8) can be simplified to

$$\begin{aligned} &\sum_{k=1}^{s} \sum_{h=1}^{2} (\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*)' \nabla \ell_j^{II,h}(\boldsymbol{\eta}_k^*) \boldsymbol{G}_k^{II,h*} \\ &= \sum_{k=1}^{s} \left[(\theta_k - \theta_k^*) \boldsymbol{G}_k^{II,1*} - \frac{1}{\gamma_k^*} (\gamma_k - \gamma_k^*) \boldsymbol{G}_k^{II,2*} \right] \nabla_\theta \ell_j^{II,1}(\boldsymbol{\eta}_k^*) \\ &\quad + \sum_{k=1}^{s} \left[(\theta_k - \theta_k^*) \boldsymbol{G}_k^{II,2*} + \frac{1}{\gamma_k^*} (\gamma_k - \gamma_k^*) \boldsymbol{G}_k^{II,1*} \right] \nabla_\theta \ell_j^{II,2}(\boldsymbol{\eta}_k^*). \end{aligned} \tag{A.13}$$

Let $\mathcal{H} = \mathrm{stack}(\boldsymbol{H}_1, \boldsymbol{H}_2, \dots)$ and $\mathcal{R} = \mathrm{stack}(\boldsymbol{R}_1, \boldsymbol{R}_2, \dots)$. Then by (A.8) and (A.13), it can be verified that

$$\begin{aligned} \widetilde{\mathcal{H}} &:= \mathrm{stack}(\boldsymbol{G}_1 - \boldsymbol{G}_1^*, \cdots, \boldsymbol{G}_p - \boldsymbol{G}_p^*, \mathcal{H}) \\ &= (\mathcal{G} - \mathcal{G}^*) \times_3 \boldsymbol{L}(\boldsymbol{\omega}^*) + \mathcal{D}(\boldsymbol{\omega}) \times_3 \left(\nabla \boldsymbol{L}^I(\boldsymbol{\lambda}^*), \nabla_\theta \boldsymbol{L}^{II}(\boldsymbol{\eta}^*) \right) \\ &= \mathcal{G}_{\mathrm{stack}} \times_3 \boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*), \end{aligned} \tag{A.14}$$

where $\mathcal{D}(\boldsymbol{\omega}) \in \mathbb{R}^{N \times N \times (r+2s)}$ and $\mathcal{G}_{\mathrm{stack}} \in \mathbb{R}^{N \times N \times (d+r+2s)}$ are defined in Appendix A.2. Note that

$$\boldsymbol{\Delta} = \widetilde{\mathcal{H}} + \mathrm{stack}(\boldsymbol{0}_{N \times N \times p}, \mathcal{R}). \tag{A.15}$$

Moreover,

$$\|\mathcal{D}(\boldsymbol{\omega})\|_{\mathrm{F}}^2 = \sum_{i=1}^{r} (\lambda_i - \lambda_i^*)^2 \|\boldsymbol{G}_i^{I*}\|_{\mathrm{F}}^2 + \sum_{k=1}^{s} \left\| (\theta_k - \theta_k^*) \boldsymbol{G}_k^{II,1*} - \frac{\gamma_k - \gamma_k^*}{\gamma_k^*} \boldsymbol{G}_k^{II,2*} \right\|_{\mathrm{F}}^2$$

$$+\sum_{k=1}^{s}\left\|(\theta_k-\theta_k^*)\boldsymbol{G}_k^{II,2*}+\frac{\gamma_k-\gamma_k^*}{\gamma_k^*}\boldsymbol{G}_k^{II,1*}\right\|_{\mathrm{F}}^2$$

$$=\sum_{i=1}^{r}(\lambda_i-\lambda_i^*)^2\|\boldsymbol{G}_i^{I*}\|_{\mathrm{F}}^2+\sum_{k=1}^{s}(\theta_k-\theta_k^*)^2(\|\boldsymbol{G}_k^{II,1*}\|_{\mathrm{F}}^2+\|\boldsymbol{G}_k^{II,2*}\|_{\mathrm{F}}^2)$$

$$+\sum_{k=1}^{s}\frac{(\gamma_k-\gamma_k^*)^2}{\gamma_k^{*2}}(\|\boldsymbol{G}_k^{II,1*}\|_{\mathrm{F}}^2+\|\boldsymbol{G}_k^{II,2*}\|_{\mathrm{F}}^2), \tag{A.16}$$

which, together with Assumption 2(ii), leads to

$$\sqrt{2}c_{\mathfrak{g}}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\leqslant\|\mathfrak{D}(\boldsymbol{\omega})\|_{\mathrm{F}}\leqslant\frac{\sqrt{2}C_{\mathfrak{g}}}{\min_{1\leqslant k\leqslant s}\gamma_k^*}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}. \tag{A.17}$$

By the simple inequalities $(|x|+|y|)/2\leqslant\sqrt{x^2+y^2}\leqslant|x|+|y|$, we have $0.5(\delta_{\mathfrak{g}}+\|\mathfrak{D}(\boldsymbol{\omega})\|_{\mathrm{F}})\leqslant\|\mathfrak{G}_{\mathrm{stack}}\|_{\mathrm{F}}\leqslant\delta_{\mathfrak{g}}+\|\mathfrak{D}(\boldsymbol{\omega})\|_{\mathrm{F}}$, and thus in view of (A.17) we further have

$$0.5\left(\delta_{\mathfrak{g}}+\sqrt{2}c_{\mathfrak{g}}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\right)\leqslant\|\mathfrak{G}_{\mathrm{stack}}\|_{\mathrm{F}}\leqslant\delta_{\mathfrak{g}}+\frac{\sqrt{2}C_{\mathfrak{g}}}{\min_{1\leqslant k\leqslant s}\gamma_k^*}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}} \tag{A.18}$$

where $\delta_{\mathfrak{g}}=\|\mathfrak{G}-\mathfrak{G}^*\|_{\mathrm{F}}$. By Lemma A.2, $\sigma_{\min,L}=\sigma_{\min}(\boldsymbol{L}_{\mathrm{stack}}(\boldsymbol{\omega}^*))>0$. Then it follows from (A.18) that

$$0.5\sigma_{\min,L}\left(\delta_{\mathfrak{g}}+\sqrt{2}c_{\mathfrak{g}}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\right)\leqslant\|\widetilde{\mathfrak{H}}\|_{\mathrm{F}}\leqslant\sigma_{\max,L}\left(\delta_{\mathfrak{g}}+\frac{\sqrt{2}C_{\mathfrak{g}}\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}}{\min_{1\leqslant k\leqslant s}\gamma_k^*}\right).$$

Combining this with (A.11), (A.15), (A.17), as well as the fact that $\|\mathfrak{G}^{\mathrm{MA}}-\mathfrak{G}^{\mathrm{MA}*}\|_{\mathrm{F}}\leqslant\delta_{\mathfrak{g}}$, we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}&\leqslant\|\widetilde{\mathfrak{H}}\|_{\mathrm{F}}+\|\mathfrak{R}\|_{\mathrm{F}}\\&\leqslant\left\{\sigma_{\max,L}+\frac{\sqrt{2}C_L}{1-\bar{\rho}}\left(\delta_{\boldsymbol{\omega}}+\frac{\delta_{\boldsymbol{\omega}}^2}{2}\right)\right\}\delta_{\mathfrak{g}}+\left(\frac{\sqrt{2}C_{\mathfrak{g}}\sigma_{\max,L}}{\min_{1\leqslant k\leqslant s}\gamma_k^*}+\frac{C_LC_{\mathfrak{g}}}{1-\bar{\rho}}\delta_{\boldsymbol{\omega}}\right)\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\end{aligned}$$

and

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}&\geqslant\|\widetilde{\mathfrak{H}}\|_{\mathrm{F}}-\|\mathfrak{R}\|_{\mathrm{F}}\\&\geqslant\left\{0.5\sigma_{\min,L}-\frac{\sqrt{2}C_L}{1-\bar{\rho}}\left(\delta_{\boldsymbol{\omega}}+\frac{\delta_{\boldsymbol{\omega}}^2}{2}\right)\right\}\delta_{\mathfrak{g}}+\left(\frac{c_{\mathfrak{g}}\sigma_{\min,L}}{\sqrt{2}}-\frac{C_LC_{\mathfrak{g}}}{1-\bar{\rho}}\delta_{\boldsymbol{\omega}}\right)\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}.\end{aligned}$$

Thus, by taking

$$\delta_{\boldsymbol{\omega}}\leqslant c_{\boldsymbol{\omega}}=\min\left\{2,\frac{c_{\mathfrak{g}}(1-\bar{\rho})\sigma_{\min,L}}{8\sqrt{2}C_LC_{\mathfrak{g}}}\right\},$$

we can show that

$$c_\Delta\left(\delta_{\mathfrak{g}}+\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\right)\leqslant\|\boldsymbol{\Delta}\|_{\mathrm{F}}\leqslant C_\Delta\left(\delta_{\mathfrak{g}}+\sqrt{\mathcal{R}_1\wedge\mathcal{R}_2}\delta_{\boldsymbol{\omega}}\right),$$

where

$$c_\Delta=c_l\cdot\sigma_{\min,L}\asymp1\quad\text{and}\quad C_\Delta=c_u\cdot\max\left\{\sigma_{\max,L},(1-\bar{\rho})^{-1}\right\}\asymp1,$$

with $c_l=0.25\min\{1,\sqrt{2}c_{\mathfrak{g}}\}$ and $c_u=1+\sqrt{2}C_{\mathfrak{g}}(\min_{1\leqslant k\leqslant s}\gamma_k^*)^{-1}+4\sqrt{2}C_L+2C_LC_{\mathfrak{g}}$.

A.4. Proof of Theorem 3.2. Let $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \boldsymbol{y}_{t-2}', \dots)'$. Denote by $\boldsymbol{\Delta}_j$ with $j \geqslant 1$ the frontal slices of $\boldsymbol{\Delta}$, i.e., $\boldsymbol{\Delta}_{(1)} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \dots)$. Denote

$$
\begin{aligned}
S_1(\boldsymbol{\Delta}) &= \frac{2}{T}\sum_{t=1}^{T}\langle \sum_{j=1}^{\infty}\boldsymbol{\Delta}_j \boldsymbol{y}_{t-j}, \sum_{k=t}^{\infty}\boldsymbol{\Delta}_k \boldsymbol{y}_{t-k}\rangle, \\
S_2(\boldsymbol{\Delta}) &= \frac{2}{T}\sum_{t=1}^{T}\langle \sum_{j=t}^{\infty}\boldsymbol{A}_j^* \boldsymbol{y}_{t-j}, \sum_{k=1}^{t-1}\boldsymbol{\Delta}_k \boldsymbol{y}_{t-k}\rangle, \qquad \text{(A.19)} \\
S_3(\boldsymbol{\Delta}) &= \frac{2}{T}\sum_{t=1}^{T}\langle \boldsymbol{\varepsilon}_t, \sum_{j=t}^{\infty}\boldsymbol{\Delta}_j \boldsymbol{y}_{t-j}\rangle.
\end{aligned}
$$

The following three lemmas are sufficient for the proof of Theorem 3.2. We state them here and provide detailed proofs in the separate supplementary file.

LEMMA A.3 (Strong convexity and smoothness properties). *Under Assumptions 1–3, if* $T \gtrsim (\kappa_2/\kappa_1)^2 d_{\mathcal{M}} \log(\kappa_2/\kappa_1)$, *then with probability at least* $1 - 2e^{-cd_{\mathcal{M}}\log(\kappa_2/\kappa_1)} - 3e^{-cN}$,

$$
\kappa_1 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \lesssim \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{\Delta}_{(1)}\boldsymbol{x}_t\|_2^2 \lesssim \kappa_2 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{\Delta} \in \boldsymbol{\Upsilon}.
$$

LEMMA A.4 (Deviation bound). *Under the conditions of Lemma A.3, with probability at least* $1 - 2e^{-cd_{\mathcal{M}}\log(\kappa_2/\kappa_1)} - 5e^{-cN}$,

$$
\frac{1}{T}\left|\sum_{t=1}^{T}\langle \boldsymbol{\varepsilon}_t, \boldsymbol{\Delta}_{(1)}\boldsymbol{x}_t\rangle\right| \lesssim \sqrt{\frac{\kappa_2 \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) d_{\mathcal{M}}}{T}}\|\boldsymbol{\Delta}\|_{\mathrm{F}}, \quad \forall \boldsymbol{\Delta} \in \boldsymbol{\Upsilon}.
$$

LEMMA A.5 (Effects of initial values). *Under Assumptions 1–3, if* $T \gtrsim (\kappa_2/\kappa_1) d_{\mathcal{M}}$, *then with probability at least* $1 - \{2 + \sqrt{\kappa_2/\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)}\}\sqrt{N/\{(\mathcal{R}_1 + \mathcal{R}_2)T\}}$,

$$
|S_1(\boldsymbol{\Delta})| \lesssim \kappa_1 \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2, \quad |S_i(\boldsymbol{\Delta})| \lesssim \sqrt{\frac{\kappa_2 \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) d_{\mathcal{M}}}{T}}\|\boldsymbol{\Delta}\|_{\mathrm{F}}, \quad i = 2, 3, \quad \forall \boldsymbol{\Delta} \in \boldsymbol{\Upsilon}.
$$

Now we prove Theorem 3.2. Note that $\sum_{j=1}^{t-1}\boldsymbol{A}_j \boldsymbol{y}_{t-j} = \boldsymbol{\mathcal{A}}_{(1)}\widetilde{\boldsymbol{x}}_t$. Due to the optimality of $\widehat{\boldsymbol{\mathcal{A}}}$, we have

$$
\sum_{t=1}^{T}\|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}^*\widetilde{\boldsymbol{x}}_t - \widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t\|_2^2 \leqslant \sum_{t=1}^{T}\|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}^*\widetilde{\boldsymbol{x}}_t\|_2^2,
$$

Then, since $\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{(1)}^*\widetilde{\boldsymbol{x}}_t = \boldsymbol{\varepsilon}_t + \sum_{j=t}^{\infty}\boldsymbol{A}_j^*\boldsymbol{y}_{t-j}$, it follows that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t\|_2^2 &\leqslant \frac{2}{T}\sum_{t=1}^{T}\langle \boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t\rangle + \underbrace{\frac{2}{T}\sum_{t=1}^{T}\langle \sum_{j=t}^{\infty}\boldsymbol{A}_j^*\boldsymbol{y}_{t-j}, \widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t\rangle}_{S_2(\widehat{\boldsymbol{\Delta}})} \\
&= \frac{2}{T}\sum_{t=1}^{T}\langle \boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{\Delta}}_{(1)}\boldsymbol{x}_t\rangle + S_2(\widehat{\boldsymbol{\Delta}}) - S_3(\widehat{\boldsymbol{\Delta}}), \qquad \text{(A.20)}
\end{aligned}
$$

where $S_2(\cdot)$ and $S_3(\cdot)$ are defined as in (A.19), $\widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t = \sum_{k=1}^{t-1}\widehat{\boldsymbol{\Delta}}_k\boldsymbol{y}_{t-k}$, and $\widehat{\boldsymbol{\Delta}}_{(1)}\boldsymbol{x}_t = \sum_{k=1}^{\infty}\widehat{\boldsymbol{\Delta}}_k\boldsymbol{y}_{t-k}$.

Moreover, applying the inequality $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \geqslant \|\boldsymbol{a}\|_2^2 - 2\langle\boldsymbol{a},\boldsymbol{b}\rangle$ with $\boldsymbol{a} = \widehat{\boldsymbol{\Delta}}_{(1)}\boldsymbol{x}_t = \sum_{j=1}^{\infty}\widehat{\boldsymbol{\Delta}}_j\boldsymbol{y}_{t-j}$ and $\boldsymbol{b} = \sum_{k=t}^{\infty}\widehat{\boldsymbol{\Delta}}_k\boldsymbol{y}_{t-k}$, we can lower bound the left-hand side of (A.20) to further obtain that

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\Delta}}_{(1)}\boldsymbol{x}_t\|_2^2 - S_1(\widehat{\boldsymbol{\Delta}}) \leqslant \frac{2}{T}\sum_{t=1}^{T}\langle\boldsymbol{\varepsilon}_t, \widehat{\boldsymbol{\Delta}}_{(1)}\boldsymbol{x}_t\rangle + S_2(\widehat{\boldsymbol{\Delta}}) - S_3(\widehat{\boldsymbol{\Delta}}), \tag{A.21}$$

where $S_1(\cdot)$ is defined as in (A.19). It is worth pointing out that $S_i(\widehat{\boldsymbol{\Delta}})$ for $1 \leqslant i \leqslant 3$ capture the initialization effect of $\boldsymbol{y}_s = \boldsymbol{0}$ for $s \leqslant 0$ on the estimation.

Note that $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \in \boldsymbol{\Upsilon}$ and $\kappa_2 \geqslant \kappa_1$. Suppose that the high probability events in Lemmas A.3–A.5 hold. Then we can derive the estimation error bound from (A.21):

$$\kappa_1\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 \lesssim \sqrt{\frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}}{T}}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}, \quad \text{or} \quad \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} \lesssim \sqrt{\frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}}{\kappa_1^2 T}}.$$

Furthermore, applying Lemma A.5 again, we can derive the prediction error bound from (A.20) and the above result as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\Delta}}_{(1)}\widetilde{\boldsymbol{x}}_t\|_2^2 \lesssim \frac{\kappa_2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)d_{\mathcal{M}}}{\kappa_1 T}.$$

The proof of this theorem is complete.

SUPPLEMENTARY MATERIAL

Supplement for "SARMA: A Computationally Scalable High-Dimensional Time Series Model" provides all remaining technical proofs and further details for Sections 5 and 6.

REFERENCES

[1] ANDERSEN, T. G., BOLLERSLEV, T., CHRISTOFFERSEN, P. F. and DIEBOLD, F. X. (2006). *Volatility and correlation forecasting* **1**. Elsevier.

[2] ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X. and LABYS, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71** 579-625.

[3] ATHANASOPOULOS, G. and VAHID, F. (2008). VARMA versus VAR for macroeconomic forecasting. *Journal of Business & Economic Statistics* **26** 237–252.

[4] BAI, J. and NG, S. (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics* **3** 89–163.

[5] BAI, J. and WANG, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics* **8** 53–80.

[6] BASU, S., LI, X. and MICHAILIDIS, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing* **67** 1207–1222.

[7] BASU, S. and MATTESON, D. S. (2021). A survey of estimation methods for sparse high-dimensional time series models. *ArXiv preprint arXiv:2107.14754*.

[8] BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* **43** 1535–1567.

[9] BULUT, U. (2016). Do financial conditions have a predictive power on inflation in Turkey? *International Journal of Economics and Financial Issues* **6** 621–628.

[10] CHAN, J. C. C., EISENSTAT, E. and KOOP, G. (2016). Large Bayesian VARMAs. *Journal of Econometrics* **192** 374-390.

[11] CHEN, K., CHAN, K.-S. and STENSETH, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B* **74** 203–221.

[12] DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* **21** 1253–1278.

[13] 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.

[14] 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.

[15] GANDY, S., RECHT, B. and YAMADA, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* **27** 025010.

[16] GOODHART, C., HOFMANN, B. et al. (2001). Asset prices, financial conditions, and the transmission of monetary policy. In *Conference on Asset Prices, Exchange Rates, and Monetary Policy, Stanford University* 2–3. Citeseer.

[17] GORROSTIETA, C., OMBAO, H., BÉDARD, P. and SANES, J. N. (2012). Investigating brain connectivity using mixed effects vector autoregressive models. *NeuroImage* **59** 3347–3355.

[18] HALLIN, M. and LIPPI, M. (2013). Factor models in high-dimensional time series: A time-domain approach. *Stochastic Processes and their Applications* **123** 2678-2695.

[19] HAN, R., WILLETT, R. and ZHANG, A. (2021). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*. To appear.

[20] HARTFIEL, D. J. (1995). Dense sets of diagonalizable matrices. *Proceedings of the American Mathematical Society* **123** 1669–1672.

[21] HATZIUS, J., HOOPER, P., MISHKIN, F. S., SCHOENHOLTZ, K. L. and WATSON, M. W. (2010). Financial conditions indexes: A fresh look after the financial crisis. Working Paper No. 16150, National Bureau of Economic Research.

[22] HORN, R. A. and JOHNSON, C. R. (2012). *Matrix Analysis*, 2nd ed. Cambridge University Press, New York.

[23] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51** 455–500.

[24] LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726.

[25] 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.

[26] LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

[27] MCCRACKEN, M. W. and NG, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* **34** 574-589.

[28] 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.

[29] NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097.

[30] PEÑA, D. and TSAY, R. S. (2021). *Statistical Learning for Big Dependent Data*. John Wiley & Sons, Hoboken, New Jersey.

[31] RASKUTTI, G., YUAN, M. and CHEN, H. (2019). Convex regularization for high-dimensional multi-response tensor regression. *The Annals of Statistics* **47** 1554–1584.

[32] STOCK, J. H. and WATSON, M. W. (2005). Implications of dynamic factor models for VAR analysis. *National Bureau of Economic Research Working Paper No. 11467*.

[33] STOCK, J. H. and WATSON, M. W. (2011). Dynamic factor models. In *Oxford Handbook of Economic Forecasting* (M. P. Clements and D. F. Hendry, eds.) Oxford University Press.

[34] TSAY, R. S. (2014). *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, Hoboken, New Jersey.

[35] TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311.

[36] VELU, R. P., REINSEL, G. C. and WICHERN, D. W. (1986). Reduced rank models for multiple time series. *Biometrika* **73** 105–118.

[37] WANG, D., ZHENG, Y. and LI, G. (2021). High-dimensional low-rank tensor autoregressive time series modelling. *arXiv preprint arXiv:2101.04276*.

[38] WANG, D., ZHENG, Y., LIAN, H. and LI, G. (2021). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association* 1–19.

[39] WILMS, I., BASU, S., BIEN, J. and MATTESON, D. (2021). Sparse identification and estimation of large-scale vector autoregressive moving averages. *Journal of the American Statistical Association*. To appear.

[40] XIA, Q., XU, W. and ZHU, L. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica* **25** 1025–1044.

[41] ZHENG, Y. and CHENG, G. (2021). Finite time analysis of vector autoregressive models under linear restrictions. *Biometrika* **108** 469-489.